Policy Zooming: Adaptive Discretization-based Infinite-Horizon Average-Reward Reinforcement Learning

Avik Kar[1], **Rahul Singh**[1]
[1]Indian Institute of Science
avikkar@iisc.ac.in, rahulsingh0188@gmail.com

Abstract

We study infinite-horizon average-reward reinforcement learning for continuous space Lipschitz Markov decision processes (MDPs) in which an agent can play policies from a given set Φ. The proposed algorithms efficiently explore the policy space by "zooming" into the "promising regions" of Φ, thereby achieving adaptivity gains in the performance. We upper bound the regret as $\tilde{\mathcal{O}}\big(T^{1-d_{\text{eff.}}^{-1}}\big)$, where $d_{\text{eff.}} = d_z^\Phi + 2$ for our model-free algorithm *PZRL-MF* and $d_{\text{eff.}} = 2d_{\mathcal{S}} + d_z^\Phi + 3$ for our model-based algorithm *PZRL-MB*. Here, $d_{\mathcal{S}}$ is the dimension of the state space, and d_z^Φ is the zooming dimension given a set of policies Φ. d_z^Φ is an alternative measure of the complexity of the problem, and it depends on the underlying MDP as well as on Φ. Hence, the proposed algorithms exhibit low regret in case the problem instance is benign and/or the agent competes against a low-complexity Φ (that has a small d_z^Φ). When specialized to the case of finite-dimensional policy space, we obtain that $d_{\text{eff.}}$ scales as the dimension of this space under mild technical conditions; and also obtain $d_{\text{eff.}} = 2$, or equivalently $\tilde{\mathcal{O}}(\sqrt{T})$ regret for *PZRL-MF*, under a curvature condition on the average reward function that is commonly used in the multi-armed bandit (MAB) literature.

Code — https://github.com/avik-kar/Policy_zooming
Extended version — https://arxiv.org/pdf/2405.18793

1 Introduction

Reinforcement Learning (RL) (Sutton and Barto 2018) is a popular framework in which an agent repeatedly interacts with an unknown environment modeled by an MDP (Puterman 2014) and the goal is to choose actions sequentially in order to maximize the cumulative rewards earned by the agent. We study infinite-horizon average reward MDPs in continuous state and action spaces endowed with a metric, in which the transition kernel and reward functions are Lipschitz (Assumption 2.1). The class of Lipschitz MDPs covers a broad class of problems, such as the class of linear MDPs (Jin et al. 2020), RKHS MDPs (Chowdhury and Gopalan 2019), linear mixture models, RKHS approximation, and the nonlinear function approximation framework considered in Osband and Van Roy (2014) and Kakade et al. (2020). See Maran et al. (2024a,b) for more details, or refer to Figure 1. Even though discrete and linear MDPs have been extensively studied in the literature, they might not be suitable for many real-world applications since it is becoming increasingly common to deploy RL and control algorithms in systems that are nonlinear and continuous (Nair et al. 2023; Kumar et al. 2021).

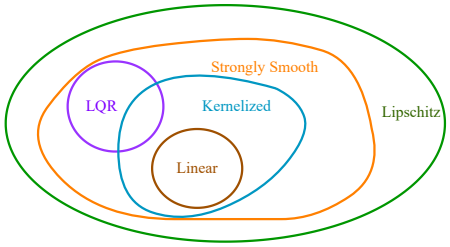
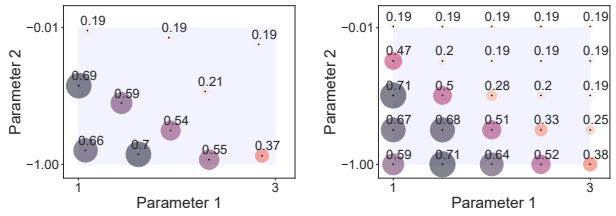


Figure 1: Relations among families of continuous space RL problems. LQR stands for Linear Quadratic Regulator (Abbasi-Yadkori and Szepesvári 2011). Our assumptions correspond to the green set. Diagram is taken from Maran et al. (2024a).

Let $d_{\mathcal{S}}$ and $d_{\mathcal{A}}$ denote the dimensions of the state and action spaces, respectively, and define $d := d_{\mathcal{S}} + d_{\mathcal{A}}$. For episodic Lipschitz MDPs, the regret scales as $\tilde{\mathcal{O}}\big(K^{1-d_{\text{eff.}}^{-1}}\big)$[1], where K is the number of episodes and $d_{\text{eff.}}$ is the *effective dimension*, which depends on both the underlying MDP and the algorithm. For instance, using a *fixed discretization* yields $d_{\text{eff.}} = d + 2$ (Song and Sun 2019). In contrast, adaptive algorithms can exploit MDP structure to reduce $d_{\text{eff.}}$. Prior works (Cao and Krishnamurthy 2020; Sinclair, Banerjee, and Yu 2023) employ adaptive discretization and a technique called "zooming," which reduces $d_{\text{eff.}}$ to $d_z + 2$, where d_z is the *zooming dimension*. However, this notion of d_z, designed for episodic settings, fails to capture adaptivity in average reward problems. Specifically, as the horizon grows, $d_z \to d$, making adaptive methods no better than fixed discretization. To address this, Kar and Singh (2025) introduced a new definition of zooming dimension tailored for average reward RL, and achieved $d_{\text{eff.}} = 2d_{\mathcal{S}} + d_z + 3$, with $d_z \leq d$. However, their methods assume compactness of the state-action space and focus solely on the complexity of the MDP.

In this work, we propose adaptive discretization-based al-

[1]$\tilde{\mathcal{O}}$ suppresses poly-logarithmic dependence in K or T.



(a) *PZRL-MF* (b) Policy UCB

Figure 2: We show the policies activated by different algorithms for one single trajectory of the transmission scheduling example (See Section 5). The radius of the balls around an active policy is proportional to its average reward. Uniform discretization-based algorithms waste resources to learn a larger number of policies, whereas adaptive algorithms activate more policies from the near-optimal regions.

gorithms that (i) handle non-compact spaces and (ii) apply to the setups efficiently where the performance is to be compared against a known class of policies. Specifically, we propose the *zooming dimension* given a policy class Φ, denoted by d_z^Φ, that captures the joint complexity of the MDP and the comparator policy class. Thus, we refine the idea of d_z to depend on a policy class as well. If the optimal policy belongs to a "simple" class, this refinement enables significantly smaller d_z^Φ, yielding $d_z^\Phi \ll d$. We analyze regret with respect to a given policy class Φ (1), a widely accepted approach in complex systems (Hazan and Singh 2022; Rakhlin and Sridharan 2014). Our model-free algorithm *PZRL-MF* and model-based algorithm *PZRL-MB* achieve regret bounds with effective dimensions $d_{\text{eff.}} = d_z^\Phi + 2$ and $d_{\text{eff.}} = 2d_{\mathcal{S}} + d_z^\Phi + 3$, respectively. It turns out that our algorithms activate policies from the given policy class in an efficient way as compared to an algorithm that uses a uniform grid for policy search. Figure 2 depicts that *PZRL-MF* activates fewer policies from suboptimal regions and more from near-optimal regions as compared to a naive uniform discretization.

To illustrate the intuition behind zooming, we revisit its origin in the simpler setting of Lipschitz MABs (Kleinberg, Slivkins, and Upfal 2008).

Lipschitz MABs: The Zooming Algorithm. The agent maintains a set of "active arms," and their "confidence balls" whose radii are equal to the confidence radii associated with the corresponding arm's estimated reward. Thanks to Lipschitz continuity, rewards of nearby arms can be inferred from the active ones.[2] New arms are activated only if not covered by existing confidence balls. The agent selects the arm with the highest upper confidence bound (UCB) index. Because the confidence radius shrinks with the number of plays, the algorithm "zooms in" on promising regions, those with a high UCB index. This adaptive behavior yields an effective dimension of $d_z + 2$. A similar idea has been explored in (Bubeck et al. 2011).

[2]Let a be an active arm with confidence radius η and empirical mean $\hat{\mu}_a$. Then, for any arm a' in its confidence ball, the mean reward lies in $[\hat{\mu}_a - (1+L)\eta, \hat{\mu}_a + (1+L)\eta]$ with high probability.

1.1 Challenges

For MABs, the zooming algorithm plays only from amongst the active arms since the UCB index of an active arm turns out to be an optimistic estimate of the mean reward of each arm lying inside its confidence ball. For policies, however, rewards are not unbiased estimates of the long-term average unless the controlled Markov process (CMP) is at stationarity, making confidence radius design and optimism proofs more involved. Moreover, to our knowledge, model-free UCB indices have not been explored for average reward RL, not even in tabular MDPs. Another challenge lies in selecting an appropriate norm for measuring distances between policies (note that since the policy space is not finite-dimensional, all norms are not equivalent).

1.2 Contributions

1. To the best of our knowledge, this is the first work to provide finite-time regret bounds for average reward RL in general state-action space MDPs with $d > 1$. Prior works (Ortner and Ryabko 2012; Qian et al. 2019; Wei et al. 2021; He, Zhong, and Yang 2023) are either limited to finite action spaces or assume $d_{\mathcal{S}} = 1$.

2. We develop two algorithms, *PZRL-MF* (model-free) and *PZRL-MB* (model-based), that use policy-based zooming and UCB methods (Lattimore and Szepesvári 2020). Our main novelty is a new complexity measure for average reward RL: the zooming dimension d_z^Φ, defined via policy covers of Φ. We show regret bounds of $\tilde{\mathcal{O}}\big(T^{1-d_{\text{eff.}}^{-1}}\big)$ with $d_{\text{eff.}} = d_z^\Phi + 2$ for *PZRL-MF*, and $d_{\text{eff.}} = 2d_{\mathcal{S}} + d_z^\Phi + 3$ for *PZRL-MB*. Importantly, a small d_z^Φ does not imply the MDP belongs to nice class of MDPs, such as linear MDPs or tabular MDPs. When Φ is parameterized over $W \subset \mathbb{R}^{d_w}$, we show $d_z^\Phi \le d_w$ under mild assumptions. For MDPs with bi-Lipschitz average reward functions, we get $d_{\text{eff.}} = 2$ and hence an $\mathcal{O}(\sqrt{T})$ regret for *PZRL-MF*.

3. Along the way, we prove a novel sensitivity result (Theorem 4.1) for Markov processes on general state spaces (Meyn and Tweedie 2012). We bound the distance between the stationary distributions of Markov chains in terms of a weighted distance measure (13) between the transition kernels, improving over the existing results (Mitrophanov 2005; Mouhoubi 2021) that bounds the same quantity in terms of the sup distance between the transition kernels.

4. Existing algorithms for general state spaces are often computationally intractable without linearity (Ayoub et al. 2020) or deterministic dynamics (Wu et al. 2024). In contrast, our methods are efficient, and in fact, *PZRL-MF* has the same computational complexity as zooming for MABs.

5. In Section 5, we demonstrate the applicability of our framework via a transmission scheduling problem. This MDP is neither tabular nor linear; however, the optimal policy is known to belong to a known class of policies that can be described by finitely many parameters. Simulation results show the practical relevance of our algorithms.

1.3 Past Works

Lipschitz episodic MDPs: Domingues et al. (2021) uses smoothing kernels to estimate the transition kernel, and ob-

tains a regret upper bound with $d_{\text{eff.}} = 2d + 1$. Cao and Krishnamurthy (2020) performs adaptive discretization and zooming and achieves regret upper bound with $d_{\text{eff.}} = d_z + 2$, where d_z is the zooming dimension defined specifically for the episodic case. Sinclair, Banerjee, and Yu (2023) also obtains adaptivity gains with $d_{\text{eff.}} = d_z + d_{\mathcal{S}}$ for a model-based algorithm. The same work also shows a regret lower bound of $\Omega(K^{1-(d_z+2)^{-1}})$. This lower bound is worse than the $\mathcal{O}(\sqrt{K})$ dependence that is achievable for the tabular case or the function approximation techniques. This is not surprising since Lipschitz MDPs are a broader class of MDPs (Maran et al. 2024a,b); See Figure 1. Even though function approximation techniques yield an $\mathcal{O}(\sqrt{K})$ regret, this comes at the expense of a larger prefactor in the regret bound as compared to Lipschitz MDPs. Moreover, function approximation techniques are computationally inefficient unless the underlying MDP is linear, and the feature maps are *known*. The knowledge of feature maps seems to be a restrictive assumption since learning features efficiently is an active topic in itself (Modi et al. 2024).

Average reward RL: Tabular MDPs are well-studied by now, and popular algorithms with a tight $\tilde{\mathcal{O}}(\sqrt{DSAT})$ regret bound exist (Jaksch, Ortner, and Auer 2010; Tossou, Basu, and Dimitrakakis 2019); where D is the MDP diameter. In contrast, continuous MDPs have only recently gained attention. Wei et al. (2021) uses function approximation, in which the relative value function is a linear function of the features, and obtains a $\tilde{\mathcal{O}}(\sqrt{T})$ regret. He, Zhong, and Yang (2023) uses function approximation techniques and obtains a regret bound of $\tilde{\mathcal{O}}(poly(d_E, B)\sqrt{d_F T})$, where B is the span of the relative value function. When the transition kernel of the underlying MDP is α-Hölder continuous and infinitely often smoothly differentiable, then Ortner and Ryabko (2012) obtains a regret upper bound with $d_{\text{eff.}} = (2d + 2\alpha)/\alpha$. Kar and Singh (2025) performs adaptive discretization and zooming and achieves regret upper-bound with $d_{\text{eff.}} = 2d_{\mathcal{S}} + d_z + 3$ for Lipschitz MDPs with compact state-action spaces.

2 Problem Setup

Notation. $\mathbb{N}$ denotes the set of natural numbers. Let $(\Omega, \mathcal{F})$ be a measurable space, and let $\mu : \mathcal{F} \mapsto \mathbb{R}$ be a signed measure, then we denote total variation norm (Folland 2013) of μ by $\|\mu\|_{TV}$, i.e., $\|\mu\|_{TV} := \sup\{\sum_i |\mu(B_i)| : \{B_i\}_i \subset \mathcal{F} \text{ partitions } \Omega\}$. For $B \subseteq \mathcal{S}$, $\text{diam}\,(B) := \sup_{s,s' \in B} \rho(s, s')$. We denote $a \wedge b$ the minimum, and $a \vee b$ the maximum of $a, b \in \mathbb{R}$. $\lceil a \rceil$ denotes the smallest integer that is larger than a for $a \in \mathbb{R}$. In general, we use a superscript (f) $((b))$ to indicate that an object is associated with algorithm *PZRL-MF* (*PZRL-MB*).

Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r)$ be an MDP, where the dimensions of the state-space $\mathcal{S}$ and action-space $\mathcal{A}$ are $d_{\mathcal{S}}$ and $d_{\mathcal{A}}$, respectively. The spaces $\mathcal{S}, \mathcal{A}$ are endowed with metrics $\rho_{\mathcal{S}}$ and $\rho_{\mathcal{A}}$, respectively. The space $\mathcal{S} \times \mathcal{A}$ is endowed with a metric ρ that is sub-additive, i.e., we have, $\rho\left((s, a), (s', a')\right) \leq \rho_{\mathcal{S}}(s, s') + \rho_{\mathcal{A}}(a, a')$, for all $(s, a), (s', a') \in \mathcal{S} \times \mathcal{A}$. We let $\mathcal{S}$ be endowed with Borel σ-algebra $\mathcal{B}_{\mathcal{S}}$. The state and the action taken at time t are denoted by s_t, a_t, respectively. The transition kernel is $p : \mathcal{S} \times \mathcal{A} \times \mathcal{B}_{\mathcal{S}} \to [0, 1]$, i.e., $\mathbb{P}\left(s_{t+1} \in B | s_t = s, a_t = a\right) = p(s, a, B)$, a.s., for all $(s, a, B) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}_{\mathcal{S}}$, $t \in \{0\} \cup \mathbb{N}$, and is not known to the agent. The reward function $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a measurable map, and the reward earned by the agent at time t is equal to $r(s_t, a_t)$. A stationary deterministic policy is a measurable map $\phi : \mathcal{S} \to \mathcal{A}$ that implements the action $\phi(s)$ when the system state is s. Let Φ_{SD} be the set of all such policies. The infinite horizon average reward for the MDP $\mathcal{M}$ under a policy ϕ is denoted by $J_{\mathcal{M}}(\phi)$, and is defined as,

$$J_{\mathcal{M}}(\phi) := \liminf_{t \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T-1} r(s_t, \phi(s_t))\right].$$

The maximum average reward attainable with a set of policies $\Phi \subseteq \Phi_{SD}$ is denoted by $J^{\star}_{\mathcal{M},\Phi}$. The regret of a learning algorithm ψ w.r.t. a class of *comparator* policies Φ until T is defined as (Rakhlin and Sridharan 2014),

$$\mathcal{R}_{\Phi}(T; \psi) := T J^{\star}_{\mathcal{M},\Phi} - \sum_{t=0}^{T-1} r(s_t, a_t). \tag{1}$$

The current work derives an upper bound on $\mathcal{R}_{\Phi}(T; \psi)$ in terms of the *zooming dimension*, a joint complexity measure of class Φ and the MDP $\mathcal{M}$ when the algorithms are only allowed to play policies from Φ. Note that if Φ contains an optimal policy, then $\mathcal{R}_{\Phi}(T; \psi)$ is the usual regret. An MDP is Lipschitz if it satisfies the following properties.

Assumption 2.1 (Lipschitz continuity)**.** (i) The reward function r is L_r-Lipschitz, i.e., $\forall\, s, s' \in \mathcal{S}, a, a' \in \mathcal{A}$,

$$|r(s, a) - r(s', a')| \leq L_r \rho\left((s, a), (s', a')\right).$$

(ii) The transition kernel p is L_p-Lipschitz, i.e., $\forall\, s, s' \in \mathcal{S}, a, a' \in \mathcal{A}$,

$$\|p(s, a, \cdot) - p(s', a', \cdot)\|_{TV} \leq L_p \rho\left((s, a), (s', a')\right).$$

While studying infinite-horizon average reward MDPs, some sort of ergodicity assumption is required. In fact, uniform ergodicity is the weakest known sufficient condition that ensures efficient computation of an optimal policy even when the MDP is known (Arapostathis et al. 1993).

Assumption 2.2 (Ergodicity)**.** Let $\Phi \subseteq \Phi_{SD}$ be the comparator class of policies. The CMP $\{s_t\}_t$ that is induced by transition kernel p under application of any $\phi \in \Phi$ is uniformly ergodic (Douc et al. 2018), that is, there exist two constants, $C \in (0, \infty)$ and $\alpha \in (0, 1)$ and for every $\phi \in \Phi$, there exists a unique distribution $\mu^{(\infty)}_{\phi,p}$ such that

$$\left\|\mu^{(t)}_{\phi,p,s} - \mu^{(\infty)}_{\phi,p}\right\|_{TV} \leq C\alpha^t,\ \forall s \in \mathcal{S}, t \in \{0\} \cup \mathbb{N}, \tag{2}$$

where $\mu^{(t)}_{\phi,p,s}$ denotes the distribution of s_t given $s_0 = s$.

We call $\mu^{(\infty)}_{\phi,p}$ as the stationary distribution of the CMP induced by p under the application of policy ϕ. We need Φ to be endowed with an appropriate metric ρ_{Φ} such that Φ is bounded and $J_{\mathcal{M}}$ is a Lipschitz function on Φ. In fact, under Assumption 2.1 and Assumption 2.2, $J_{\mathcal{M}}$ is Lipschitz w.r.t. the metric ρ_{Φ} set equal to the metric induced by ∞ norm.

Further, $J_{\mathcal{M}}$ is Lipschitz w.r.t. a weighted distance measure too under a mild technical condition. We show these two results in Theorem 4.1. For a discussion on the metric spaces, see Appendix C of the extended version. We next define the zooming dimension d_z^{Φ}, a joint complexity measure for the problem instance and the comparator policy class Φ.

Zooming dimension. Suboptimality of a policy ϕ w.r.t. Φ is defined as $\Delta_{\Phi}(\phi) := J^{\star}_{\mathcal{M},\Phi} - J_{\mathcal{M}}(\phi)$. Define the sets of policies $\Phi_{\gamma} := \{\phi \in \Phi \mid \Delta_{\Phi}(\phi) \in (\gamma, 2\gamma]\}$, and $\Phi_{\leq\gamma} := \{\phi \in \Phi \mid \Delta_{\Phi}(\phi) \leq \gamma\}$. Then, the zooming dimension of the problem given the policy space Φ is defined as

$$d_z^{\Phi} := \inf \Big\{ d' > 0 \mid \mathcal{N}_{\frac{\gamma}{c_z}}(\Phi_{\gamma}) \leq c_{z_1}\gamma^{-d'}, \text{ and } \mathcal{N}_{\frac{\gamma}{c_z}}(\Phi_{\leq\gamma}) \leq c_{z_2}\gamma^{-d'}, \ \forall\gamma \geq 0 \Big\}, \quad (3)$$

where $\mathcal{N}_{\gamma}(\Phi')$ denotes the γ-covering number of $\Phi' \subseteq \Phi$ w.r.t. metric ρ_{Φ}, c_{z_1} and c_{z_2} are problem-dependent constants, $c_z := 2(\max\{2, C_{ub}\} + L_J)$. C_{ub} is a problem-dependent constant.

Remark 2.3. We note that even if the policy class is high-dimensional, the zooming dimension could be small, as it is a measure of the size of the set of near-optimal policies.

3 Algorithm

We propose a model-free algorithm *PZRL-MF* and a model-based algorithm *PZRL-MB*. Both the algorithms combine policy-based zooming with the principle of optimism in the face of uncertainty (Lattimore and Szepesvári 2020). They maintain a set of active policies, compute their UCB indices, and then play an active policy with the highest UCB index in the current episode. Its zooming component zooms in and activates only those policies from Φ, for which it is not possible to generate a good estimate of its performance using the performance estimates of nearby active policies. However, they differ in the way in which they compute UCB indices and activate new policies and hence are discussed separately. The algorithms are summarized in Algorithm 1.

3.1 PZRL-MF

Policy Diameter. Diameter at time t is defined as,

$$\text{diam}_t^{(f)}(\phi) := \frac{C}{1-\alpha}\left(\sqrt{\frac{c_d^{(f)}\log\left(\frac{T}{\delta}\right)}{1 \vee N_t(\phi)}} + \frac{1+K_t(\phi)}{1 \vee N_t(\phi)}\right), \phi \in \Phi, \quad (4)$$

where $N_t(\phi)$ is the number of plays of the policy ϕ until time t, $K_t(\phi)$ is the number of episodes that began before time t, and in which ϕ was played, while $c_d^{(f)}$ is an appropriate constant.

Active Policies. $\Phi_t^{act.}$ denotes the set of policies active at time t. Define the following ball in the policy space,

$$B_{\phi,t} := \left\{\phi' \in \Phi : \rho_{\Phi}(\phi', \phi) \leq \text{diam}_t^{(f)}(\phi)\right\}. \quad (5)$$

Since the confidence ball associated with a policy shrinks when it is played, in the possible event that $\cup_{\phi\in\Phi_t^{act.}} B_t(\phi)$ does not cover the set Φ anymore, the proposed algorithm activates a new policy to ensure that the union of confidence balls of the active policies covers Φ. Thus, *PZRL-MF* possesses the *covering invariance* property, i.e., $\cup_{\phi\in\Phi_t^{act.}} B_{\phi,t}$ covers Φ at all times.

Algorithm 1: Policy Zooming for RL (*PZRL-MF*/*PZRL-MB*)

Input Horizon T, confidence parameter δ, ergodicity coefficient α and policy class Φ
Initialize $h = 0$, $k = 0$, $\Phi_0^{act.} = \{\}$.
for $t = 0$ to $T-1$ **do**
  if $h \geq H_k$ **then**
    $k \leftarrow k+1$, $h \leftarrow 0$
    Update the set of active policies $\Phi_t^{act.} \subset \Phi$
    For every $\phi \in \Phi_t^{act.}$ compute $\text{Index}_t(\phi)$, where
    $\text{Index}_t(\phi) = \text{Index}_t^{(f)}(\phi)$ (6) if *PZRL-MF*,
    $\text{Index}_t(\phi) = \text{Index}_t^{(b)}(\phi)$ (12) if *PZRL-MB*.
    Choose $\phi^{(k)} \in \arg\max_{\phi\in\Phi_t^{act.}} \text{Index}_t(\phi)$.
    $H_k = 1 \vee N_t(\phi^{(k)})$
  end if
  $h \leftarrow h+1$
  Play $a_t = \phi^{(k)}(s_t)$, observe s_{t+1} and receive $r(s_t, a_t)$.
end for

Model-free UCB Index. Let ϕ_t denote the policy played at time t. The UCB index at time t is defined as

$$\text{Index}_t^{(f)}(\phi) := \frac{1}{N_t(\phi)}\sum_{i=0}^{t-1}\mathbb{I}_{\{\phi_i=\phi\}} r(s_i, \phi(s_i)) + (1+L_J)\,\text{diam}_t^{(f)}(\phi), \ \phi \in \Phi_t^{act.}, \quad (6)$$

where L_J is the Lipschitz constant associated with $J_{\mathcal{M}}$.

3.2 PZRL-MB

We assume $\mathcal{S}$ to be bounded for *PZRL-MB*. *PZRL-MB* maintains an adaptive partition of the state space $\mathcal{S}$ for each active policy. We use $\mathcal{P}_{\phi,t}$ to denote the state partition corresponding to policy ϕ at time t; see Appendix A.1 in the extended version for more details on the procedure to create these partitions. Loosely speaking, as time progresses, $\mathcal{P}_{\phi,t}$ is finer in those regions of $\mathcal{S}$ that have been visited relatively more number of times while playing ϕ. $\mathcal{P}_{\phi,t}$ consists of a certain type of subsets of $\mathcal{S}$ called cells. The cells comprising $\mathcal{P}_{\phi,t}$ are called active cells at time t corresponding to policy ϕ. Let $q_{\phi,t}^{-1}(s)$ be the active cell corresponding to ϕ at time t that contains the state s.

Policy Diameter. The model-based policy diameter for $\phi \in \Phi$ is defined as follows:

$$\text{diam}_t^{(b)}(\phi) := \int_{\mathcal{S}} \text{diam}\left(q_{\phi,t}^{-1}(s)\right)\mu_{\phi,p}^{(\infty)}(ds). \quad (7)$$

Policy balls for *PZRL-MB* are defined similar to (5), replacing $\text{diam}_t^{(f)}(\phi)$ with $\text{diam}_t^{(b)}(\phi)$. Similar to *PZRL-MF*, *PZRL-MB* maintains a set of active policies, $\Phi_t^{act.}$ that satisfies the covering invariance property.

Approximate Diameter: The agent cannot compute $\text{diam}_t^{(b)}(\phi)$ since it does not know $\mu_{\phi,p}^{(\infty)}$. However, the agent

can compute a "tight" lower bound of $\text{diam}_t^{(b)}(\phi)$, which can then be used in (5) in lieu of $\text{diam}_t^{(b)}(\phi)$. This approximation causes the regret upper bound to increase only by a constant factor (See Appendix A.3 and A.6 in the extended version).

Model-based UCB Index. *PZRL-MB* evaluates the UCB indices of the active policies by using an estimate of the transition kernel. The algorithm constructs a set of plausible discretized transition kernels using its estimate (10). A curated bias term is added to the discretized reward function in order to overcome the discretization error. Then, the UCB indices are computed using an iterative algorithm (11), similar to the policy evaluation algorithm. Computation of the UCB index involves the following three steps:

(i) Estimating the Transition Kernel: Denote $S_{\phi,t}$ to be the set of representative points of the cells in $\mathcal{P}_{\phi,t}$, and denote $\bar{S}_{\phi,t}$ to be the set of representative points of all the cells of size of the smallest cell in $\mathcal{P}_{\phi,t}$. At time t, for every active policy ϕ, *PZRL-MB* constructs the empirical transition distribution, $\hat{p}_{\phi,t}^{(d)}(s,\cdot)$ with the width of the bins set equal to the diameter of $q_{\phi,t}^{-1}(s)$ for every $s \in S_{\phi,t}$. Then a continuous extension of $\hat{p}_{\phi,t}^{(d)}$, $\hat{p}_{\phi,t}$ is computed. $\hat{p}_{\phi,t}$ is again discretized with the width of the bins set equal to the diameter of the smallest active cell. This discrete estimate is denoted by $\wp_{S_{\phi,t}\to\bar{S}_{\phi,t},\hat{p}_{\phi,t}}$. For a detailed discussion on the estimation of the transition kernel, see Appendix A.2 in the extended version of the paper.

(ii) Confidence Ball: For a policy ϕ and a representative state $s \in S_{\phi,t}$, the confidence radius associated with the estimate $\wp_{S_{\phi,t}\to\bar{S}_{\phi,t},\hat{p}_{\phi,t}}$ is defined as follows,

$$\eta_{\phi,t}(s) := 3\left(\frac{c_d^{(b)}\log\left(T\delta^{-1}\right)}{\sum_{i=0}^{t-1}\mathbb{I}_{\left\{s_i\in q_{\phi,t}^{-1}(s)_i\right\}}}\right)^{\frac{1}{d_{\mathcal{S}}+2}} + \left(3(1+L_\phi)L_p + C_p\right)\ \text{diam}\left(q_{\phi,t}^{-1}(s)\right), \quad (8)$$

where $c_d^{(b)} > 0$ is a constant that is discussed in Lemma D.2 in the extended version, C_p is as described in Assumption 4.3, and L_ϕ is the Lipschitz constant associated with ϕ, i.e., for all $s, s' \in \mathcal{S}$, $\rho_{\mathcal{A}}(\phi(s),\phi(s')) \le L_\phi \rho_{\mathcal{S}}(s,s')$. It follows from the rule used for activating a new cell that we have,

$$\eta_{\phi,t}(s) \le C_{\eta,\phi}\ \text{diam}\left(q_{\phi,t}^{-1}(s)\right), \quad (9)$$

for every $s \in S_{\phi,t}$, where $C_{\eta,\phi} := 3\left(1+(1+L_\phi)L_p\right) + C_p$. Let $\Theta_{\phi,t}$ denote the set of all possible discretized transition kernels that describe outgoing transition probabilities from points in $\bar{S}_{\phi,t}$, with a support on the discrete state space $\bar{S}_{\phi,t}$. We define a set of transition probability kernels associated with $\wp_{S_{\phi,t}\to\bar{S}_{\phi,t},\hat{p}_{\phi,t}}$ as follows,

$$\mathcal{C}_{\phi,t} := \Big\{\theta\in\Theta_{\phi,t} \mid \Big\|\theta(\bar{s},\cdot) - \wp_{S_{\phi,t}\to\bar{S}_{\phi,t},\hat{p}_{\phi,t}}(s,\cdot)\Big\|_1 \le \eta_{\phi,t}(s) \text{ for every } s\in S_{\phi,t}, \bar{s}\in\bar{S}_{\phi,t}\cap q_t^{-1}(s)\Big\}, \quad (10)$$

(iii) Computing the UCB Indices of Active Policies: Let us fix a time t. To obtain the UCB index of a policy $\phi \in \Phi$, we perform the following iterations,

$$\overline{V}_0^{\phi,t}(s) = 0,$$
$$\overline{V}_{i+1}^{\phi,t}(s) = r(s,\phi(s)) + (1+L_\phi)L_r\ \text{diam}\left(q_{\phi,t}^{-1}(s)\right) + \max_{\theta\in\mathcal{C}_{\phi,t}}\sum_{s'\in\bar{S}_{\phi,t}}\theta(s,s')\overline{V}_i^{\phi,t}(s'), \quad (11)$$

$s \in \bar{S}_{\phi,t}$, $i \in \mathbb{Z}_+$. The difference of two consecutive iterates of (11) is shown to converge in Lemma A.4 in the extended version. We define the UCB indices as follows,

$$\text{Index}_t^{(b)}(\phi) := \lim_{i\to\infty}\left(\overline{V}_{i+1}^{\phi,t}(s) - \overline{V}_i^{\phi,t}(s)\right) + L_J\ \text{diam}_t^{(b)}(\phi), \quad (12)$$

for any $s \in \bar{S}_{\phi,t}$.

Remark 3.1. Similar to the zooming algorithm for bandits (Kleinberg, Slivkins, and Upfal 2019), we assume access to an oracle that takes as input a finite collection of open balls, and then either declares that they cover Φ, or outputs a point that is uncovered. In general, such an oracle may not be computationally efficient. However, when Φ has a finite-dimensional parameterization, we can perform a grid search.

4 Regret Analysis

In this section, we present our main results, Theorem 4.2 and Theorem 4.5, that yield upper bounds on the regret of *PZRL-MF* and *PZRL-MB*, respectively. Before presenting these, we first show that the average reward function $J_{\mathcal{M}}(\cdot)$ is a Lipschitz function of the policies w.r.t. the sup-norm distance. Furthermore, under a mild assumption, it is also a Lipschitz function of the policies w.r.t. a weighted distance measure. This result provides an important insight into selecting an appropriate norm for defining the zooming dimension. Define

$$\rho_{\Phi,\infty}(\phi,\phi') := \sup_{s\in\mathcal{S}}\rho_{\mathcal{A}}(\phi(s),\phi'(s)),\ \forall\phi,\phi'\in\Phi.$$

Consider a probability measure ν on $(\mathcal{S},\mathcal{B}_{\mathcal{S}})$. Define the metric,

$$\rho_{\Phi,\nu}(\phi,\phi') := \int_{\mathcal{S}}\rho_{\mathcal{A}}(\phi(s),\phi'(s))\,d\nu(s), \quad (13)$$

Theorem 4.1. *Let the MDP $\mathcal{M}$ satisfy Assumption 2.1 and 2.2. (i) Then, the infinite horizon average reward is $L_{J,\infty}$-Lipschitz w.r.t. the metric $\rho_{\Phi,\infty}$, i.e., for $\phi,\phi'\in\Phi$ we have,*

$$|J_{\mathcal{M}}(\phi) - J_{\mathcal{M}}(\phi')| \le L_{J,\infty}\rho_{\Phi,\infty}(\phi,\phi'),$$

where,

$$L_{J,\infty} := L_r + \frac{L_p}{2(1-\alpha)}\left(\left\lceil\log_{\frac{1}{\alpha}}(C)\right\rceil + 1\right). \quad (14)$$

(ii) Furthermore, if $\mu_{\phi,p}^{(\infty)}(\xi) \le \kappa\nu(\xi)$, $\forall\xi\in\mathcal{B}_{\mathcal{S}}, \phi\in\Phi$, for some probability measure ν and a constant $\kappa > 0$, then $J_{\mathcal{M}}(\cdot)$ is $L_{J,\nu}$-Lipschitz w.r.t. the metric $\rho_{\Phi,\nu}$, i.e., for $\phi,\phi'\in\Phi$ we have,

$$|J_{\mathcal{M}}(\phi) - J_{\mathcal{M}}(\phi')| \le L_{J,\nu}\rho_{\Phi,\nu}(\phi,\phi'),$$

where, $L_{J,\nu} := \kappa L_{J,\infty}$.

The above theorem is proved in Appendix B of the extended version. The next two theorems are the main results of this work, and bound the regrets of *PZRL-MF* and *PZRL-MB*.

Theorem 4.2. *If the MDP $\mathcal{M}$ satisfies Assumptions 2.1 and 2.2, then with a probability at least $1-\delta$, the regret of PZRL-MF, i.e. $\mathcal{R}_\Phi(T;$ PZRL-MF$)$, is bounded above as $\tilde{\mathcal{O}}(T^{1-d_{\text{eff.}}^{-1}})$ where $d_{\text{eff.}} = d_z^\Phi + 2$.*

The following assumptions are required for the analysis of *PZRL-MB*.

Assumption 4.3 (Bounded Radon-Nikodym derivative)**.** The probability measures $\{p(s,\phi(s),\cdot)\}_s$ are absolutely-continuous w.r.t. the Lebesgue measure on $(\mathcal{S},\mathcal{B}_\mathcal{S})$, with density functions given by $\{f_{\phi,s}\}_s$ for every $\phi\in\Phi$. These densities satisfy

$$\left\|\frac{\partial f_{\phi,s}(s_+)}{\partial s_+(i)}\right\|_\infty \le C_p, \forall s\in\mathcal{S}, i=1,2,\ldots,d_\mathcal{S},$$

where $s_+ = (s_+(1), s_+(2), \cdots, s_+(d_\mathcal{S}))$.

Assumption 4.4. There exists $\kappa'>0$ such that for every $\zeta\subseteq\mathcal{B}_\mathcal{S}$, $\mu_{\phi,p}^{(\infty)}(\zeta)\ge\kappa'\lambda(\zeta)$, where λ is the Lebesgue measure (Billingsley 2017) on $(\mathcal{S},\mathcal{B}_\mathcal{S})$.

The above two assumptions are not restrictive. See Remark 4.7 in the extended version for more details on this. We use $\Phi_{\text{Lip.}}$ to denote the class of Lipschitz policies.

Theorem 4.5. *Let $\Phi\subseteq\Phi_{Lip.}$. If the MDP $\mathcal{M}$ satisfies Assumption 2.1, 2.2, 4.3 and 4.4, then with a probability at least $1-\delta$, the regret of PZRL-MB, i.e. $\mathcal{R}_\Phi(T;$ PZRL-MB$)$, is upper-bounded as $\tilde{\mathcal{O}}(T^{1-d_{\text{eff.}}^{-1}})$ where $d_{\text{eff.}} = 2d_\mathcal{S} + d_z^\Phi + 3$.*

Refer to Appendix G in the extended version for detailed proofs of the above two results. Here, we provide a generic *proof sketch*.

Proof sketch. We decompose the regret as follows,

$$\mathcal{R}_\Phi(T;\psi) = \underbrace{\sum_{k=1}^{K(T)}\sum_{t=\tau_k}^{\tau_{k+1}-1} J_{\mathcal{M},\Phi}^\star - J_\mathcal{M}(\phi^{(k)})}_{(a)} + \underbrace{\sum_{k=1}^{K(T)}\sum_{t=\tau_k}^{\tau_{k+1}-1} J_\mathcal{M}(\phi^{(k)}) - r(s_t,\phi^{(k)}(s_t))}_{(b)},$$

where $\phi^{(k)}$ denotes the policy played in the k-th episode, τ_k denotes the time when the k-th episode starts, and $K(T)$ denotes the total number of episodes till time T. We bound the terms (a) and (b) separately.

Bounding (a): This term is further decomposed into the sum of the regrets arising due to playing policies from the sets Φ_γ, where γ assumes the values 2^{-i}, $i=1,2,\ldots,\lceil\log(1/\epsilon)\rceil$, and $\epsilon = T^{-d_{\text{eff.}}^{-1}}$. Cumulative regret arising from playing policies not in the set $\cup_{i=1}^{\lceil\log(1/\epsilon)\rceil}\Phi_{2^{-i}}$ is bounded by ϵT. Regret due to playing policies from Φ_γ is bounded in the following three steps:

(1) First, we derive a condition under which a γ-suboptimal policy is no longer played.

(2) Then, we deduce an upper bound of the number of plays of a policy ϕ in terms of its suboptimality gap by concluding that the condition stated in (1) holds when ϕ has been played sufficiently many times.

(3) Then, we establish an upper bound on the number of policies that are activated by the algorithms from Φ_γ.

The product of two upper bounds discussed in (2) and (3), when multiplied with 2γ, yields regret from playing policies in Φ_γ. We then add these regret terms corresponding to different sets Φ_γ, where $\gamma=2^{-i}$ and $i=1,2,\ldots,\lceil\log(1/\epsilon)\rceil$; to this we add the regret arising due to playing policies with suboptimality less than ϵ, which is bounded by ϵT.

Bounding (b): We show that term (b) is bounded as $\mathcal{O}(K(T))$ using uniform ergodicity. Then we show that the total number of episodes for *PZRL-MF* as well as *PZRL-MB* are bounded above by $\mathcal{O}\left(T^{d_z^\Phi/d_{\text{eff.}}}\right)$, where $d_{\text{eff.}} = d_z^\Phi+2$ and $d_{\text{eff.}} = 2d_\mathcal{S}+d_z^\Phi+3$, respectively.

We obtain the desired regret bound after summing the upper bounds on (a) and (b). □

Remark 4.6 (Discontinuous Policies). The regret analysis of *PZRL-MB* extends to policies with discontinuities, provided all discontinuities lie on the boundaries of active cells. This condition can be enforced by redefining the cells (Definition A.1 in the extended version) accordingly. For simplicity, we define the cells using dyadic cubes.

The next result quantifies an upper bound on $d_{\text{eff.}}$ for an important class of policies; See Appendix C in the extended version for the proof.

Corollary 4.7 (Finite parameterization)**.** *We now consider a set Φ that consists of policies that have been parameterized by finitely many parameters from the set $W\subset\mathbb{R}^{d_W}$. For each $w\in W$, let $\phi(\cdot;w):\mathcal{S}\to\mathcal{A}$ be the policy parameterized by w. Assume that the policies satisfy $L_W\rho_\Phi(\phi(\cdot;w),\phi(\cdot;w'))\ge\|w-w'\|_2$ for all $w,w'\in W$. We have $d_{\text{eff.}}\le d_W+2$ for PZRL-MF and $d_{\text{eff.}}\le 2d_\mathcal{S}+d_W+3$ for PZRL-MB.*

Corollary 4.8 (Bi-Lipschitz MDPs)**.** *Consider a bi-Lipschitz MDP, i.e., the average reward function $J_\mathcal{M}:\Phi\to\mathbb{R}$ satisfies the following properties: there exist two constants $\bar{L}\ge\underline{L}>0$ such that for every ϕ,ϕ'*

$$\underline{L}\rho_\Phi(\phi,\phi') \le |J_\mathcal{M}(\phi)-J_\mathcal{M}(\phi')| \le \bar{L}\rho_\Phi(\phi,\phi').$$

Then, the regret of PZRL-MF w.r.t. the policy class scales as $\tilde{\mathcal{O}}(\sqrt{T})$ on a high probability set.

We note that the assumption made in Corollary 4.8 commonly made in continuum bandits literature such as Cope (2009); Yu and Mannor (2011); Combes, Proutière, and Fauquette (2020).

5 Simulations

We evaluate the proposed algorithms on the following two systems.

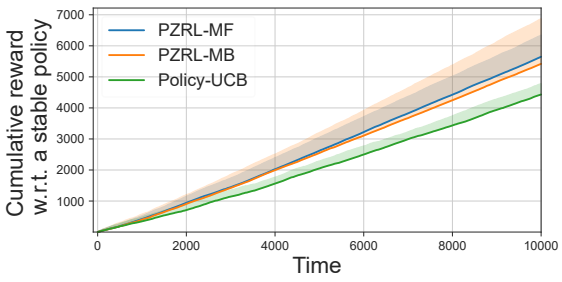
Transmission scheduling for remote estimation of a stochastic dynamic process. Consider a process $\{x_t\}$ that

evolves as $x_{t+1} = \beta x_t + w_t$, where $|\beta| < 1$ and $\{w_t\}$ is i.i.d., $w_t \sim \mathcal{N}(0,1)$ for all t. A sensor observes $\{x_t\}$, encodes it into data packets, and transmits them to a remote estimator across an unreliable wireless channel. $c_t \in \{0,1\}$ denotes the channel state at time t. $c_t = 1$ (0) denotes that the channel state is good (bad). $\{c_t\}$ is a Markov process with transition probabilities $p_{ij} := \mathbb{P}(c_{t+1} = j \mid c_t = i)$, $i, j \in \{0,1\}$, where $p_{01}, p_{11} > 0$. $a_t \in \{0,1\}$ denotes the decision made by the sensor regarding whether or not a packet transmission is attempted at time t; $a_t = 1$ denotes that transmission is attempted. The estimator state $\hat{x}_t$ evolves as $\hat{x}_{t+1} = x_{t+1}c_t a_t + (1 - c_t a_t)\beta \hat{x}_t$. The estimation error $\{e_t\}$ evolves as $e_{t+1} = (\beta e_t + w_t) - \beta c_t a_t e_t$. The agent's estimate of c_t, denoted by b_t can be updated recursively. The actions $\{a_t\}$ are to be chosen so as to minimize the error with a minimal amount of transmission power. The agent earns a reward $r_t := -e_t^2 - \lambda a_t$, where $\lambda > 0$ is the number of units of resource required for transmission. Dutta and Singh (2023) shows that a threshold policy is optimal in this setup, one which transmits only when b_t exceeds a certain threshold (that is allowed to depend upon e_t). Hence, the optimal policy can be described by specifying the threshold curve, which in turn can be approximated by a curve with finitely many parameters. This problem does not fit into the class of Linear MDPs or Tabular MDPs. However, it can be shown that the average reward function is Lipschitz when the comparator policy class consists only of stable policies, and hence fits within our framework. We compare the empirical performance of the proposed algorithms, *PZRL-MF* and *PZRL-MB* (Algorithm 1) with that of a heuristic algorithm Policy UCB (Algorithm 2 in the extended version), which discretizes the policy space uniformly at time $t = 0$ and plays the policy with the highest model-free UCB index from the set of finite set of policies in every episode. For both *PZRL-MF* and *PZRL-MB*, we use the following parameterization: $\phi(s; w) = \mathbb{I}_{\{w(1)+w(2)e_t < b_t\}}$, $w = (w(1), w(2)) \in [1,3] \times [-1, -0.01]$. We plot the cumulative reward minus the average performance of the policy that suggests transmission irrespective of the system state, averaged over 50 runs in Figure 3(a). Both *PZRL-MF* and *PZRL-MB* outperform the fixed discretization-based algorithm, Policy UCB.

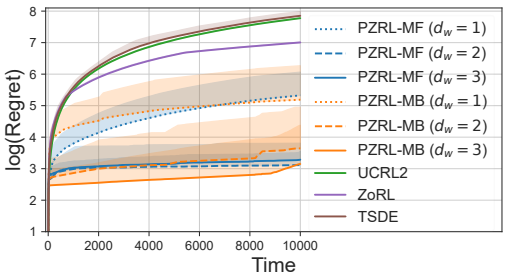
Continuous RiverSwim. We modify the RiverSwim MDP (Strehl and Littman 2008) to obtain its continuous version. The state s_t describes the location of the agent in the river and evolves as follows upon the application of action a_t at time t:

$$s_{t+1} = \begin{cases} (0 \vee (s_t - \frac{1}{2}(1 + \frac{w_t}{2}))) \wedge 6 & \text{w.p. } \frac{2(1-a_t)}{5} \\ s_t & \text{w.p. } 0.2 \\ (0 \vee (s_t + \frac{1}{2}(1 + \frac{w_t}{2}))) \wedge 6 & \text{w.p. } \frac{2(1+a_t)}{5}, \end{cases}$$

where $\{w_t\}$ is i.i.d. and $w_t \sim \mathcal{N}(0, 0.5)$, and $t \in \{0\} \cup \mathbb{N}$. Here, $\mathcal{S} = [0,6]$ and $\mathcal{A} = [0,1]$. The reward function is given by $r(s,a) = 0.005(((s-6)/6)^4 + ((a-1)/2)^4) + 0.5((s/6)^4 + ((a+1)/2)^4)$. Note that the policy that chooses the action 1 at all times, irrespective of the current state, is optimal. For both *PZRL-MF* and *PZRL-MB*, we use the following parameterizations:



(a) Transmission Scheduling



(b) Continuous RiverSwim

Figure 3: Simulation results.

1. 1 parameter: $\phi(s; w) = w$, $w \in [-1, 1]$.
2. 2 parameters: $\phi(s; w) = w(1) + w(2)s$, $w = (w(1), w(2)) \in [-1, 1] \times [-0.5, 0.5]$.
3. 3 parameters: $\phi(s; w) = w(1) + w(2)s + w(3)s^2$, $w = (w(1), w(2), w(3)) \in [-1, 1] \times [-0.5, 0.5]^2$.

We compare the empirical performance of *PZRL-MF* and *PZRL-MB* (Algorithm 1) with that of ZoRL (Kar and Singh 2025), UCRL2 (Jaksch, Ortner, and Auer 2010) and TSDE (Ouyang et al. 2017). Since these competitor policies are designed for finite state-action spaces, we apply them on a uniform discretization of $\mathcal{S} \times \mathcal{A}$. We plot the logarithm of the cumulative regret averaged over 50 runs for the Continuous RiverSwim environment in Figure 3(b), and observe that *PZRL-MF* and *PZRL-MB* outperforms every other algorithm, and amongst *PZRL-MF* and *PZRL-MB*, *PZRL-MB* has the edge over *PZRL-MF*. Policy classes with 2 and 3 parameters outperform the single-parameter policy class for both proposed algorithms.

6 Conclusion

The central idea of zooming-based algorithms is to capitalize on its adaptive nature. The adaptivity is captured via the zooming dimension. We identify the absence of policy class dependence in the existing definition of the zooming dimension. To rectify this, we define the zooming dimension in terms of coverings of the policy space, allowing it to depend on the comparator policy class. We propose zooming-based algorithms *PZRL-MF* and *PZRL-MB*, and prove that their regret can be bounded as $\tilde{\mathcal{O}}(T^{1-d_{\text{eff.}}^{-1}})$ where $d_{\text{eff.}} = d_z^{\Phi} + 2$ and $d_{\text{eff.}} = 2d_{\mathcal{S}} + d_z^{\Phi} + 3$, respectively. Simulation results support our theoretical findings.

References

Abbasi-Yadkori, Y.; and Szepesvári, C. 2011. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, 1–26. JMLR Workshop and Conference Proceedings.

Arapostathis, A.; Borkar, V. S.; Fernández-Gaucherand, E.; Ghosh, M. K.; and Marcus, S. I. 1993. Discrete-time controlled Markov processes with average cost criterion: a survey. *SIAM Journal on Control and Optimization*, 31(2): 282–344.

Ayoub, A.; Jia, Z.; Szepesvari, C.; Wang, M.; and Yang, L. 2020. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, 463–474. PMLR.

Billingsley, P. 2017. *Probability and measure*. John Wiley & Sons.

Bubeck, S.; Munos, R.; Stoltz, G.; and Szepesvári, C. 2011. X-Armed Bandits. *Journal of Machine Learning Research*, 12(5).

Cao, T.; and Krishnamurthy, A. 2020. Provably adaptive reinforcement learning in metric spaces. *Advances in Neural Information Processing Systems*, 33: 9736–9744.

Chowdhury, S. R.; and Gopalan, A. 2019. Online learning in kernelized Markov decision processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 3197–3205. PMLR.

Combes, R.; Proutière, A.; and Fauquette, A. 2020. Unimodal bandits with continuous arms: Order-optimal regret without smoothness. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 4(1): 1–28.

Cope, E. W. 2009. Regret and convergence bounds for a class of continuum-armed bandit problems. *IEEE Transactions on Automatic Control*, 54(6): 1243–1253.

Domingues, O. D.; Menard, P.; Pirotta, M.; Kaufmann, E.; and Valko, M. 2021. Kernel-Based Reinforcement Learning: A Finite-Time Analysis. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 2783–2792. PMLR.

Douc, R.; Moulines, E.; Priouret, P.; and Soulier, P. 2018. *Markov chains: Basic definitions*. Springer.

Dutta, M.; and Singh, R. 2023. Optimal scheduling policies for remote estimation of autoregressive Markov processes over time-correlated fading channel. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, 6455–6462. IEEE.

Folland, G. B. 2013. *Real analysis: modern techniques and their applications*. John Wiley & Sons.

Hazan, E.; and Singh, K. 2022. Introduction to online non-stochastic control. *arXiv preprint arXiv:2211.09619*.

He, J.; Zhong, H.; and Yang, Z. 2023. Sample-efficient Learning of Infinite-horizon Average-reward MDPs with General Function Approximation. In *The Twelfth International Conference on Learning Representations*.

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.; Yang, Z.; Wang, Z.; and Jordan, M. I. 2020. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, 2137–2143. PMLR.

Kakade, S.; Krishnamurthy, A.; Lowrey, K.; Ohnishi, M.; and Sun, W. 2020. Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems*, 33: 15312–15325.

Kar, A.; and Singh, R. 2025. Provably Adaptive Average Reward Reinforcement Learning for Metric Spaces. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, 1924–1964. PMLR.

Kleinberg, R.; Slivkins, A.; and Upfal, E. 2008. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, 681–690.

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

Kumar, A.; Fu, Z.; Pathak, D.; and Malik, J. 2021. Rma: Rapid motor adaptation for legged robots. *arXiv preprint arXiv:2107.04034*.

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

Maran, D.; Metelli, A. M.; Papini, M.; and Restell, M. 2024a. No-Regret Reinforcement Learning in Smooth MDPs. *The Forty-first International Conference on Machine Learning*.

Maran, D.; Metelli, A. M.; Papini, M.; and Restelli, M. 2024b. Projection by Convolution: Optimal Sample Complexity for Reinforcement Learning in Continuous-Space MDPs. *The 37th Annual Conference on Learning Theory*.

Meyn, S. P.; and Tweedie, R. L. 2012. *Markov chains and stochastic stability*. Springer Science & Business Media.

Mitrophanov, A. Y. 2005. Sensitivity and convergence of uniformly ergodic Markov chains. *Journal of Applied Probability*, 42(4): 1003–1014.

Modi, A.; Chen, J.; Krishnamurthy, A.; Jiang, N.; and Agarwal, A. 2024. Model-free representation learning and exploration in low-rank mdps. *Journal of Machine Learning Research*, 25(6): 1–76.

Mouhoubi, Z. 2021. Perturbation and stability bounds for ergodic general state Markov chains with respect to various norms. *Le Matematiche*, 76(1): 243–276.

Nair, S.; Rajeswaran, A.; Kumar, V.; Finn, C.; and Gupta, A. 2023. R3m: A universal visual representation for robot manipulation. *Confernce on Robot Learning*.

Ortner, R.; and Ryabko, D. 2012. Online regret bounds for undiscounted continuous reinforcement learning. *Advances in Neural Information Processing Systems*, 25.

Osband, I.; and Van Roy, B. 2014. Model-based reinforcement learning and the eluder dimension. *Advances in Neural Information Processing Systems*, 27.

Ouyang, Y.; Gagrani, M.; Nayyar, A.; and Jain, R. 2017. Learning unknown Markov decision processes: A Thompson sampling approach. In *Advances in Neural Information Processing Systems*, 1333–1342.

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

Qian, J.; Fruit, R.; Pirotta, M.; and Lazaric, A. 2019. Exploration bonus for regret minimization in discrete and continuous average reward mdps. *Advances in Neural Information Processing Systems*, 32.

Rakhlin, A.; and Sridharan, K. 2014. Lecture notes for STAT928: Statistical Learning and Sequential Prediction. www.mit.edu/~rakhlin/courses/stat928/stat928_notes.pdf.

Sinclair, S. R.; Banerjee, S.; and Yu, C. L. 2023. Adaptive discretization in online reinforcement learning. *Operations Research*, 71(5): 1636–1652.

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

Strehl, A. L.; and Littman, M. L. 2008. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8): 1309–1331.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Tossou, A.; Basu, D.; and Dimitrakakis, C. 2019. Near-optimal optimistic reinforcement learning using empirical Bernstein inequalities. *arXiv preprint arXiv:1905.12425*.

Wei, C.-Y.; Jahromi, M. J.; Luo, H.; and Jain, R. 2021. Learning infinite-horizon average-reward MDPs with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, 3007–3015. PMLR.

Wu, R.; Sekhari, A.; Krishnamurthy, A.; and Sun, W. 2024. Computationally Efficient RL under Linear Bellman Completeness for Deterministic Dynamics. *CoRR*.

Yu, J. Y.; and Mannor, S. 2011. Unimodal Bandits. In *ICML*, 41–48.