Provably Adaptive Average Reward Reinforcement Learning for Metric Spaces

Avik Kar[1] **Rahul Singh**[1]

[1] Department of Electrical Communication Engineering, Indian Institute of Science, Bengaluru

Abstract

We study infinite-horizon average-reward reinforcement learning (RL) for Lipschitz MDPs, a broad class that subsumes several important classes such as linear and RKHS MDPs, function approximation frameworks, and develop an adaptive algorithm ZoRL with regret bounded as $\mathcal{O}\big(T^{1-d_{\text{eff.}}^{-1}}\big)$, where $d_{\text{eff.}} = 2d_{\mathcal{S}} + d_z + 3$, $d_{\mathcal{S}}$ is the dimension of the state space and d_z is the zooming dimension. In contrast, algorithms with fixed discretization yield $d_{\text{eff.}} = 2(d_{\mathcal{S}} + d_{\mathcal{A}}) + 2$, $d_{\mathcal{A}}$ being the dimension of action space. ZoRL achieves this by discretizing the state-action space adaptively and zooming into "promising regions" of the state-action space. d_z, a problem-dependent quantity bounded by the state-action space's dimension, allows us to conclude that if an MDP is benign, then the regret of ZoRL will be small. The zooming dimension and ZoRL are truly adaptive, i.e., the current work shows how to capture adaptivity gains for infinite-horizon average-reward RL. ZoRL outperforms other state-of-the-art algorithms in experiments, thereby demonstrating the gains arising due to adaptivity.

1 INTRODUCTION

Reinforcement Learning (RL) [Sutton and Barto, 2018] is a popular model for systems involving real-time sequential decision-making and has applications in many fields such as robotics, natural language processing [Ibarz et al., 2021, Sodhi et al., 2023]. An agent interacts sequentially with an environment by applying actions and gathers rewards. The environment is modeled as a Markov decision process (MDP) [Puterman, 2014], its transition probabilities are not known to the agent. Its goal is to choose actions sequentially so as to maximize the cumulative rewards.

The current work develops an RL algorithm for infinite-horizon average reward Lipschitz MDPs on metric spaces. Popular frameworks such as tabular and linear MDPs that have been well-studied in detail in RL literature, are not suitable for real-world applications since these typically involve nonlinear systems that reside on continuous spaces [Kumar et al., 2021]. For continuous spaces, the learning regret could grow linearly with time horizon T unless the problem has some structure [Kleinberg et al., 2008]. Hence, we focus on Lipschitz MDPs, which is a very general class and subsumes several popular classes such as linear MDPs [Jin et al., 2020], RKHS MDPs [Chowdhury and Gopalan, 2019], linear mixture models, RKHS approximation, and the nonlinear function approximation framework [Osband and Van Roy, 2014, Kakade et al., 2020]. See Maran et al. [2024a,b] for more details.

Throughout, we use $d_{\mathcal{S}}, d_{\mathcal{A}}$ to denote the dimensions of the state-space and the action-space respectively, and $d := d_{\mathcal{S}} + d_{\mathcal{A}}$. In episodic RL for Lipschitz MDPs, the regret is known to scale as $\tilde{\mathcal{O}}\big(K^{1-d_{\text{eff.}}^{-1}}\big)$[1], where K is the number of episodes, while $d_{\text{eff.}}$ is the effective dimension associated with the *underlying MDP* and also importantly the *algorithm*. A naive algorithm that uses a fixed discretization has $d_{\text{eff.}} = d + 2$ [Song and Sun, 2019]. One can use problem structure to reduce $d_{\text{eff.}}$; prior works on episodic Lipschitz MDPs such as Sinclair et al. [2019], Cao and Krishnamurthy [2020] reduce effective dimension to $d_z + 2$, where the zooming dimension d_z measures the size of the near-optimal state-action pairs. These gains are achieved by performing an adaptive discretization of the state-action space and "zooming in" to only the promising regions of the state-action space by creating a finer grid around these as time progresses. However, Kar and Singh [2024a] show that zooming technique and algorithms developed for episodic MDPs are inappropriate for average reward RL tasks, in that $d_z \to d$ as $T \to \infty$, which is what one would have obtained via a naive fixed discretization scheme. Kar and Singh [2024a] derives an $\mathcal{O}(\epsilon^{2d_{\mathcal{S}}+d_z^{\epsilon}+1}\log T)$ upper-bound

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

Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI 2025), PMLR 244:1924–1964.

on the regret with respect to an ϵ suboptimal comparator policy class, where d_z^ϵ is the "ϵ-zooming dimension" and satisfies $d_z^\epsilon \leq d$. However, $d_z^\epsilon \to d$ in the limit $\epsilon \downarrow 0$, which shows that no adaptivity gains are achieved if the policy class contains optimal policy, i.e., one wants to attain optimal performance. In a later version of the same paper, Kar and Singh [2024b] rectifies this issue to some extent by competing against an optimal policy class. They work directly in the policy space, and show zooming behavior in this space rather than the state-action space, i.e., their algorithm "activates" more number of policies from the near-optimal regions in the policy space. They obtain $d_{\text{eff.}} = d_z^\Phi + 2$, where d_z^Φ measures the size of near-optimal policies in the set of policies Φ that can be chosen. d_z^Φ is the log-covering number of the set consisting of $(\beta, 2\beta]$-suboptimal policies in Φ. However, d_z^Φ can be prohibitively large if either the MDP or the policy-set Φ is not structured, since it involves coverings in function spaces [Guntuboyina and Sen, 2012]. The current work remedies this and upper-bounds the regret in terms of an alternative notion of zooming dimension, one that can be bounded by d in the worst case. Though the analysis of our algorithm is performed in the policy space, it relates the suboptimality of a policy with that of the associated state-action pairs, thereby deriving an upper-bound of the number of plays of suboptimal policies in terms of coverings of the state-action space.

1.1 CONTRIBUTIONS

We propose a computationally efficient algorithm ZoRL for Lipschitz MDPs in the infinite-horizon average reward RL setup. ZoRL combines adaptive discretization with the principle of optimism and yields zooming behavior. We provide a regret upper-bound of ZoRL as a function of the zooming dimension d_z, where d_z is defined in terms of the suboptimality gap of the state action pairs (2). We show that the regret of ZoRL is upper-bounded as $\tilde{\mathcal{O}}\big(T^{1-d_{\text{eff.}}^{-1}}\big)$, where $d_{\text{eff.}} = 2d_{\mathcal{S}} + d_z + 3$, and $d_z \leq d$. In order to attain a low $d_{\text{eff.}}$, we had to overcome several challenges. These are discussed in detail below.

1. *Bypassing Policy Covers*: As is discussed above, working with policy coverings could lead to a large $d_{\text{eff.}}$. Let $\Phi^{(\beta)}$ denote the set of all $(\beta, 2\beta]$-suboptimal policies. By establishing an upper-bound on the total number of plays of $\Phi^{(\beta)}$ in terms of the β-covering number of the set of all β-suboptimal state-action pairs, the current work attains a small $d_{\text{eff.}}$. Our proof hinges on the existence of certain "key cells." More specifically, we show that whenever ZoRL plays a suboptimal policy ϕ, there exists a ball in the state-action space that satisfies the following two properties: (i) it has not been visited sufficiently many times, and (ii) the stationary measure under ϕ assigns a large probability mass to it. Such a ball is called a "key cell" for that particular episode, see Fig. 1. Lemma 4.1 unveils a relation between the suboptimality of a policy, and the suboptimality gap of the state-action pairs through which this policy passes. This result plays a crucial role in proving the existence of key cells. We derive an upper-bound on the number of plays of a cell during which it is a key cell and policies from $\Phi^{(\beta)}$ are played; here β can be chosen from $(0, 1]$. This upper-bound helps us to express the regret in terms of a covering of a state-action space, which yields a bound that depends upon the zooming dimension (3).
2. *Adaptive Episode Durations*: In order to attain $d_{\text{eff.}} = 2d_{\mathcal{S}} + d_z + 3$, we have to ensure that with a high probability, the key cells are visited at least a certain number of times in each episode. This is achieved by choosing the episode durations as a function of the "proxy diameter" of the policy that is played currently. We note that the popular approaches for choosing episode duration, such as ending the episode upon doubling the number of visits to any cell, would fail to yield $d_{\text{eff.}} = 2d_{\mathcal{S}} + d_z + 3$.

We verify the gains of ZoRL over both popular fixed discretization-based algorithms and existing adaptive discretization-based algorithms through simulation experiments.

1.2 PAST WORKS

Lipschitz Bandits: The idea of zooming was first proposed in [Kleinberg et al., 2008] for Lipschitz multi-armed bandits. Bubeck et al. [2011] proposed a similar idea that uses a hierarchical partition of the arm space to perform adaptive discretization.

Lipschitz MDPs: Domingues et al. [2021] uses smoothing kernels in order to construct model estimates and obtain $\tilde{\mathcal{O}}\Big(H^3 K^{1-(2d+1)^{-1}}\Big)$ regret. Provable gains arising due to adaptive discretization and zooming is first demonstrated in [Cao and Krishnamurthy, 2020]. They obtain $\tilde{\mathcal{O}}\Big(H^{2.5+(2d_z+4)^{-1}} K^{1-(2d_z+1)^{-1}}\Big)$ regret, where d_z is the zooming dimension defined specifically for episodic RL. In another work, Sinclair et al. [2023] proposes a model-based algorithm with adaptive discretization and shows the regret to be upper-bounded as $\tilde{\mathcal{O}}\Big(L_v H^{\frac{3}{2}} K^{1-(d_z+d_{\mathcal{S}})^{-1}}\Big)$, where L_v is the Lipschitz constant for the value function. As compared the general function approximation-based works, regret bounds obtained in works on Lipschitz MDPs have a worse growth rate as a function of time horizon. However, this is expected since Lipschitz MDPs are a more general class of MDPs and have a regret lower-bound of $\Omega(K^{1-(d_z+2)^{-1}})$ [Sinclair et al., 2023].

Non-episodic RL: The minimax regret of state-of-the-art algorithms for finite MDPs [Jaksch et al., 2010, Tossou

et al., 2019] with S states and A actions is bounded as $\tilde{\mathcal{O}}(\sqrt{DSAT})$ where D is the diameter of the MDP. For finite MDPs in which the transition kernel is a mixture of d component transition kernels, regret is upper-bounded as $\tilde{\mathcal{O}}(d\sqrt{DT})$ [Wu et al., 2022]. The current work develops algorithm for continuous MDPs. Wei et al. [2021] analyzes continuous MDPs under the assumption that the relative value function is a linear function of the features, and obtains a $\tilde{\mathcal{O}}(\sqrt{T})$ regret. Another work, He et al. [2023] approximates the MDP, as well as the value function by using general function classes. They derive a regret upper-bound of $\tilde{\mathcal{O}}(poly(d_E, B)\sqrt{d_F T})$ regret, where B is the span of the relative value function, d_E, d_F are the eluder dimension and log-covering number of the function class, respectively. When the underlying continuous MDP has a α-Hölder continuous and infinitely often smoothly differentiable transition kernel, then Ortner and Ryabko [2012] shows how to obtain a $\tilde{\mathcal{O}}\left(T^{\frac{2d+\alpha}{2d+2\alpha}}\right)$ regret. To the best of our knowledge, only [Kar and Singh, 2024a,b][2] have studied adaptive discretization for average reward Lipschitz MDPs; however, they analyze regret with respect to a given class of policies. For [Kar and Singh, 2024a], when this class is "sufficiently rich" so that it contains an optimal policy, then their algorithm does not exhibit adaptivity gains, i.e., their zooming dimension reduces to d, which is what one would attain via a fixed discretization scheme. In Kar and Singh [2024b], the zooming dimension could be even larger than d if the policy class is complex.

2 PROBLEM SETUP

Notation. The set of natural numbers is denoted by $\mathbb{N}$. We denote the span of a $\mathbb{R}$-valued function $f \in \mathbb{R}^X$ by $sp\,(f)$, i.e., $sp\,(f) = \max_{x\in X} f(x) - \min_{x\in X} f(x)$. We abbreviate "with high probability" as "w.h.p." For a σ-algebra $\mathcal{F}$ and a measure $\mu : \mathcal{F} \to \mathbb{R}$, we let $\|\mu\|_{TV}$ denote its total variation norm [Folland, 2013], i.e., $\|\mu\|_{TV} := \sup\{|\mu(B)| : B \in \mathcal{F}\}$. $a \vee b$ denotes the maximum, and $a \wedge b$ denotes the minimum of two real numbers a and b. $\lceil a \rceil$ denotes the smallest integer that is greater than or equal to a. At certain places, we use a single variable (z) to denote state-action pairs.

Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r)$ be an MDP, where the state-space $\mathcal{S}$ and action-space $\mathcal{A}$ are compact sets of dimension $d_{\mathcal{S}}$ and $d_{\mathcal{A}}$, respectively. Let $\mathcal{S}$ be endowed with Borel σ-algebra $\mathcal{B}_{\mathcal{S}}$. To simplify exposition, we assume that $\mathcal{S} = [0,1]^{d_{\mathcal{S}}}$ and $\mathcal{A} = [0,1]^{d_{\mathcal{A}}}$ without loss of generality. We denote the system state and action taken at time t by s_t, a_t respectively. The state s_t evolves as follows,

$$\mathbb{P}\left(s_{t+1} \in B | s_t = s, a_t = a\right) = p(s, a, B), \text{ a.s.},$$
$$\forall (s, a, B) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}_{\mathcal{S}},\ t \in \{0\} \cup \mathbb{N},$$

where $p : \mathcal{S} \times \mathcal{A} \times \mathcal{B}_{\mathcal{S}} \to [0,1]$ is the transition kernel that is not known by the agent. The agent earns a reward $r(s_t, a_t)$ at time t, where the reward function $r : \mathcal{S} \times \mathcal{A} \to [0,1]$ is a measurable map. The goal of the agent is to maximize the infinite horizon average reward. 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,

$$\rho\left((s,a),(s',a')\right) \le \rho_{\mathcal{S}}(s,s') + \rho_{\mathcal{A}}(a,a'),$$

for all $(s,a),(s',a') \in \mathcal{S}\times\mathcal{A}$. For $\mathcal{Z} \subseteq \mathcal{S}\times\mathcal{A}$, diam $(\mathcal{Z}) := \sup_{z_1,z_2\in\mathcal{Z}} \rho(z_1,z_2)$. 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 of a policy ϕ when it acts on an MDP $\mathcal{M}$ is denoted by $J_{\mathcal{M}}(\phi)$, and is defined as,

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

where $\mathbb{E}_{\mathcal{M},\phi}$ denotes expectation taken under consideration that policy ϕ is used to take actions throughout on the MDP $\mathcal{M}$. The optimal average reward of the MDP $\mathcal{M}$ is defined as $J^\star_{\mathcal{M}} := \sup_{\phi\in\Phi_{SD}} J_{\mathcal{M}}(\phi)$. The regret [Lattimore and Szepesvári, 2020] of a learning algorithm ψ until T is defined as,

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

The goal of this work is to design a learning algorithm with tight regret upper bound for Lipschitz MDPs. An MDP is Lipschitz if it satisfies the assumption below.

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

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

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

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

The following assumption ensures that the underlying MDP is ergodic and is typically required for average reward setup [Ortner, 2020, Wei et al., 2021, Hao et al., 2021].

Assumption 2.2 (Uniform ergodicity)**.** *We assume that $\{s_t\}$, the controlled Markov process (CMP) induced by transition*

[2] Kar and Singh [2024b] is a later version of the same paper Kar and Singh [2024a].

kernel p under application of any stationary deterministic policy is uniformly ergodic [Douc et al., 2018], that is, for every $\phi \in \Phi_{SD}$, there exists a unique distribution $\mu^{(\infty)}_{\phi,p}$, two constants, $C \in (0,\infty)$ and $\alpha \in (0,1)$ such that

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

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

We note that even when $\mathcal{M}$ is known, (2.2) is the weakest known sufficient condition that ensures a computationally efficient way to obtain an optimal policy [Arapostathis et al., 1993]. Consider the Average Reward Optimality Equation (AROE) corresponding to the MDP $\mathcal{M}$, $J+h(s) = \max_{a\in\mathcal{A}} \left\{ r(s,a) + \int_S h(s')\, p(s,a,s')\, ds' \right\}$. It can be shown that under Assumption 2.2, there exists a function $h_{\mathcal{M}} : \mathcal{S} \to \mathbb{R}$ such that $(J^\star_{\mathcal{M}}, h_{\mathcal{M}})$ satisty the AROE [Hernández-Lerma, 2012] where $h_{\mathcal{M}}$ is the relative value function. Imposing an additional condition $h(s_\star) = 0$ results in unique solution to the AROE, where $s_\star$ is a designated state. Also, there exists a stationary deterministic policy $\phi^\star$ that is optimal, i.e., $J^\star_{\mathcal{M}} = J_{\mathcal{M}}(\phi^\star)$. Similarly, for a policy $\phi \in \Phi_{SD}$ there is a function $h^\phi_{\mathcal{M}} : \mathcal{S} \to \mathbb{R}$ such that $(J_{\mathcal{M}}(\phi), h^\phi_{\mathcal{M}})$ is the solution of $J + h(s) = r(s,\phi(s)) + \int_S h(s')\, p(s,\phi(s),s')\, ds'$. See Appendix A for more details on properties of average reward MDPs. The suboptimality gap [Burnetas and Katehakis, 1997] of a state-action pair is defined as follows:

$$\begin{aligned} \text{gap}\,(s,a) :=& J^\star_{\mathcal{M}} + h_{\mathcal{M}}(s) - r(s,a) \\ & - \int_{\mathcal{S}} h_{\mathcal{M}}(s')\, p(s,a,s')\, ds'. \end{aligned} \tag{2}$$

Zooming dimension. Let us denote the set of state-action pairs (s,a) such that $\text{gap}\,(s,a) \leq \beta$ by $\mathcal{Z}_\beta$. We define the zooming dimension as

$$d_z := \inf \left\{ d' > 0 \mid \mathcal{N}_{c_s\beta}\left(\mathcal{Z}_\beta\right) \leq c_z \beta^{-d'},\ \forall \beta > 0 \right\}, \tag{3}$$

where $\mathcal{N}_{c_s\beta}\left(\mathcal{Z}_\beta\right)$ denotes the $c_s\beta$-covering number [Cao and Krishnamurthy, 2020] of $\mathcal{Z}_\beta$, c_s (71) and c_z are problem-dependent constants. Note that d_z is logarithm of the covering number of a subset of $\mathcal{S}\times\mathcal{A}$, hence $d_z \leq d$.

3 ALGORITHM

The proposed algorithm, ZoRL discretizes the state-action space in a non-uniform grid adaptively, and the grid becomes finer as time progresses. In this section, first, we explain the adaptive discretization process.

Definition 3.1 (Cells)**.** *A cell is a dyadic cube with vertices from the set $\{2^{-\ell}(v_1, v_2, \ldots, v_d) : v_j \in \{0, 1, \ldots, 2^\ell\}, j = 1, 2, \ldots, d\}$ with sides of length $2^{-\ell}$, where $\ell \in \mathbb{N}$. The quantity ℓ is called the level of the cell. We also denote the collection of cells of level ℓ by $\mathcal{P}^{(\ell)}$. For a cell $\zeta \subseteq \mathcal{S}\times\mathcal{A}$, its $\mathcal{S}$-projection is called an $\mathcal{S}$-cell,*

$$\pi_{\mathcal{S}}(\zeta) := \{ s \in \mathcal{S} \mid (s,a) \in \zeta \text{ for some } a \in \mathcal{A} \}, \tag{4}$$

and its level is the same as that of ζ. Denote the set of $\mathcal{S}$-cells of level ℓ by $\mathcal{Q}^{(\ell)}$. For a cell/$\mathcal{S}$-cell ζ, we let $\ell(\zeta)$ denote its level, and let $q(\zeta)$ denote a point from ζ that is its unique representative point. q^{-1} maps a representative point to the cell/$\mathcal{S}$-cell that the point is representing, i.e., $q^{-1}(z) = \zeta$ such that $q(\zeta) = z$.[3]

Definition 3.2 (Partition tree)**.** *A partition tree of depth ℓ is a tree in which (i) Each node at a depth $m \leq \ell$ of the tree is a cell of level m. (ii) If ζ is a cell of level m, where $m < \ell$ then, a) all the cells of level $m+1$ that collectively generate a partition of ζ, are the child nodes of ζ. The corresponding cells are called child cells, and we use Child(ζ) to denote all the child cells of ζ. b) ζ is called the parent cell of these child nodes. The set of all ancestor nodes of cell ζ is called ancestors of ζ.*

ZoRL (3) maintains a set of "active cells." The following rule is used for activating and deactivating cells.

Definition 3.3 (Activation rule)**.** *For a cell ζ define,*

$$N_{\max}(\zeta) := \frac{c_a 2^{d_{\mathcal{S}}+2} \log\left(\frac{T}{\delta}\right)}{\textit{diam}\,(\zeta)^{d_{\mathcal{S}}+2}},\ \textit{and,} \tag{5}$$

$$N_{\min}(\zeta) := \begin{cases} 1 & \textit{if } \zeta = \mathcal{S}\times\mathcal{A} \\ \frac{c_a \log\left(\frac{T}{\delta}\right)}{\textit{diam}(\zeta)^{d_{\mathcal{S}}+2}}, & \textit{otherwise,} \end{cases} \tag{6}$$

where $c_a > 1$ is a constant that satisfies (92)*, and $\delta \in (0,1)$ is the confidence parameter. The number of visits to ζ is denoted $N_t(\zeta)$ and is defined as follows.*

1. *Any cell ζ is said to be active if $N_{\min}(\zeta) \leq N_t(\zeta) < N_{\max}(\zeta)$.*
2. *$N_t(\zeta)$ is defined for all cells as the number of times ζ or any of its ancestors has been visited while being active until time t, i.e.,*

$$N_t(\zeta) := \sum_{i=0}^{t-1} \mathbb{1}_{\{(s_i,a_i)\in\zeta_i\}}, \tag{7}$$

where ζ_i is the unique cell that is active at time i and satisfies $\zeta \subseteq \zeta_i$.

Denote the set of active cells at time t by $\mathcal{P}_t$.

[3] With a slight abuse of notation, we use the maps $\ell(\cdot)$, $q(\cdot)$ and $q^{-1}(\cdot)$ for both cells and $\mathcal{S}$-cells. Note that for cells and $\mathcal{S}$-cells, these maps have different domains and codomains.

We note that since the diameter of a child cell is half that of its parent, a parent cell is deactivated, and its child cells are activated simultaneously. Since a cell is partitioned by its child cells, the set of active cells at time t, $\mathcal{P}_t$ forms a partition of the state action space. ZoRL clusters all the state-action pairs into the active cells by utilizing the information gathered until t. Each point in an active cell (cluster) ζ looks similar for the purpose of generating optimal actions, and is hence represented via its unique representative point $q(\zeta)$. Denote the collection of representative points of the active cells at time t by $\mathcal{Z}_t := \{q(\zeta) : \zeta \in \mathcal{P}_t\}$. Let $\ell_{\max,t}$ be the level of the smallest cells in $\mathcal{P}_t$. At time t, ZoRL partitions the state-space into $\mathcal{S}$-cells of level $\ell_{\max,t}$. We denote this $\mathcal{S}$-cell partition by $\mathcal{Q}_t$, i.e., $\mathcal{Q}_t := \mathcal{Q}^{(\ell_{\max,t})}$, and the corresponding representative points by $\mathcal{S}_t$, i.e., $\mathcal{S}_t := \{q(\zeta) : \zeta \in \mathcal{Q}_t\}$. $\mathcal{S}_t$ can be thought of as the discretized state space at time t. ZoRL maintains estimates of the transition probability kernel that has support on $\mathcal{S}_t$.

Now, we introduce a generic notation for discretized transition kernels, which will be used often in this paper. Let $\tilde{\mathcal{S}}$ be a set of representative points of a partition of $\mathcal{S}$ consisting of only $\mathcal{S}$-cells. Then, for a continuous transition kernel $\tilde{p}$, and $\tilde{\mathcal{Z}} \subseteq \mathcal{S} \times \mathcal{A}$, we define $\wp_{\tilde{\mathcal{Z}} \to \tilde{\mathcal{S}}, \tilde{p}}(z, \cdot) : \tilde{\mathcal{Z}} \mapsto [0,1]^{\tilde{\mathcal{S}}}$ as follows,

$$\wp_{\tilde{\mathcal{Z}} \to \tilde{\mathcal{S}}, \tilde{p}}(z, s) := \tilde{p}(z, q^{-1}(s)),\ \forall z \in \tilde{\mathcal{Z}}, s \in \tilde{\mathcal{S}}. \quad (8)$$

The kernel $\wp_{\tilde{\mathcal{Z}} \to \tilde{\mathcal{S}}, \tilde{p}}$ can be viewed as a discretization of $\tilde{p}$.

Estimating the Transition Kernel. Let $N_t(\zeta, \xi)$ be the total number of transitions from a cell ζ, or from its active ancestors to a $\mathcal{S}$-cell ξ until t, i.e., $N_t(\zeta, \xi) := \sum_{i=1}^{t-1} \mathbb{1}_{\{(s_i, a_i, s_{i+1}) \in \zeta_i \times \xi\}}$. For any state-action pair z, we let $q_t^{-1}(z)$ denote the active cell that contains z. Denote $\tilde{\mathcal{S}}_t(z) := \{q(\xi) : \xi \in \mathcal{Q}^{(\ell(q_t^{-1}(z)))}\}$, which is the set of representative states of the $\mathcal{S}$-cells of level $\ell(q_t^{-1}(z))$. We first construct an estimate $\hat{p}_t^{(d)}$ (9) of the discretized version of the true stochastic kernel as follows,

$$\hat{p}_t^{(d)}(z, s) := \frac{N_t\left(q^{-1}(z), q^{-1}(s)\right)}{1 \vee N_t\left(q^{-1}(z)\right)}, \quad (9)$$

$z \in \mathcal{Z}_t, s \in \tilde{\mathcal{S}}_t(z)$. Note that the distribution $\hat{p}_t^{(d)}(z, \cdot)$ is supported on a finite set $\tilde{\mathcal{S}}_t(z)$, and the sets $\{\tilde{\mathcal{S}}_t(z)\}$ are adaptive. $\hat{p}_t^{(d)}(z, \cdot)$ is then extended to obtain a continuous kernel $\hat{p}_t$. $\hat{p}_t$ is defined as,

$$\hat{p}_t(z, B) := \sum_{s \in \tilde{\mathcal{S}}_t(z)} \frac{\lambda(B \cap q^{-1}(s))}{\lambda(q^{-1}(s))} \hat{p}_t^{(d)}(z, s), \quad (10)$$

where $z \in \mathcal{Z}_t$, $B \in \mathcal{B}_{\mathcal{S}}$, and $\lambda(\cdot)$ is the Lebesgue measure on $(\mathcal{S}, \mathcal{B}_{\mathcal{S}})$. To obtain a computationally feasible algorithm, we work with the discretization $\wp_{\mathcal{Z}_t \to \mathcal{S}_t, \hat{p}_t}$ of $\hat{p}_t$.

Note that the set $\tilde{S}_t(z)$ depends upon the diameter of the active cell containing z, so that the support of the discrete kernel $\hat{p}_t^{(d)}(z, \cdot)$ varies with z. The construction of $\wp_{\mathcal{Z}_t \to \mathcal{S}_t, \hat{p}_t}$ from $\hat{p}_t^{(d)}$ ensures that the support of the discrete kernel at every point is the same ($\mathcal{S}_t$). This allows us to use the EVI algorithm, which will be introduced later in this section.

Concentration Inequality. ZoRL constructs a confidence ball centered at $\wp_{\mathcal{Z}_t \to \mathcal{S}_t, \hat{p}_t}$ that contains discretized version of the true transition kernel, p w.h.p. For a cell $\zeta \in \mathcal{P}_t$, the confidence radius associated with the estimate $\wp_{\mathcal{Z}_t \to \mathcal{S}_t, \hat{p}_t}(q(\zeta), \cdot)$ is defined as follows,

$$\eta_t(\zeta) := \min \Bigg\{ 2, 3 \left(\frac{c_a \log\left(\frac{T}{\delta}\right)}{N_t(\zeta)} \right)^{\frac{1}{d_{\mathcal{S}}+2}} + (3L_p + C_p)\text{diam}\,(\zeta) \Bigg\}, \quad (11)$$

where C_p is an upper bound on the derivatives of the transition density functions, as described in Assumption 4.2, and the constant $c_a \geq 1$ satisfies (92). It turns out that the following value of c_a satisfies (92):

$$c_a = \frac{2d^{\frac{d_{\mathcal{S}}}{2}}}{9} \frac{\log\left(6d^{\frac{d}{2}}\right)}{\log\left(\frac{T}{\delta}\right)} + \frac{d}{d_{\mathcal{S}}+2} + 1. \quad (12)$$

Lemma F.1 shows that w.h.p.,

$$\|\wp_{\mathcal{S}\times\mathcal{A} \to \mathcal{S}_t, p}(z, \cdot) - \wp_{\mathcal{Z}_t \to \mathcal{S}_t, \hat{p}_t}(q(\zeta), \cdot)\|_{TV} \leq \eta_t(\zeta), \quad \forall z \in \zeta,$$

for every t and every $\zeta \in \mathcal{P}_t$. This leads to the definition of the confidence ball that ZoRL uses.

Now, we introduce the discrete state-action space that we will use in the definition of the confidence ball. The set of all the relevant cells for $s \in \mathcal{S}$ at time t are defined as $\mathit{Rel}_t(s) := \{\zeta \in \mathcal{P}_t \mid \exists a \in \mathcal{A} \text{ such that } (s, a) \in \zeta\}$. These are those active cells whose $\mathcal{S}$-projection contain the state s. Thus, $\mathit{Rel}_t(s)$ can be seen as the set of those cells in the state-action space that are associated with state s currently. Recall that $\mathcal{S}_t$ is the discrete state space at time t. Define

$$\mathcal{A}_t(s) := \cup_{\zeta \in \mathit{Rel}_t(s)} \{a \in \mathcal{A} \mid q(\zeta) = (s', a) \text{ for some } s' \in \mathcal{S}\}.$$

$\mathcal{A}_t(s)$ denotes the set of actions that are available to the agent that can be played by it currently in state s. The discrete action space at time t is given by $\mathcal{A}_t := \{\mathcal{A}_t(s) : s \in \mathcal{S}_t\}$. Let $\mathcal{S}_t \times \mathcal{A}_t := \{(s, a) \mid s \in \mathcal{S}_t, a \in \mathcal{A}_t(s)\}$. Define the confidence ball,

$$\mathcal{C}_t := \Big\{\theta : \mathcal{S}_t \times \mathcal{A}_t \mapsto [0,1]^{\mathcal{S}_t} \mid \sum_{s \in \mathcal{S}_t} \theta(z, s) = 1,\ \forall z \in \mathcal{S}_t \times \mathcal{A}_t, \|\theta(z', \cdot) - \wp_{\mathcal{Z}_t \to \mathcal{S}_t, \hat{\wp}_t}(\bar{z}, \cdot)\|_1 \leq \eta_t(q^{-1}(\bar{z})) \text{ for every } \bar{z} \in \mathcal{Z}_t, z' \in q^{-1}(\bar{z}) \cap \mathcal{S}_t \times \mathcal{A}_t\Big\}. \quad (13)$$

As a consequence of Lemma F.1, $\mathcal{C}_t$ contains $\wp_{\mathcal{S}_t\times\mathcal{A}_t\to\mathcal{S}_t,p}$ w.h.p. Denote the time when the k-th episode of ZoRL begins by τ_k. At the beginning of each episode k, ZoRL constructs a set of discrete MDPs $\mathcal{M}^+_{\tau_k}$ with transition kernel can be chosen from $\mathcal{C}_{\tau_k}$, and reward function is equal to the true rewards at the discrete points $\mathcal{S}_t \times \mathcal{A}_t$, plus a bonus term. Such a set of MDPs is called the "extended MDP" and it is commonly used to incorporate optimism in upper confidence bound-based RL algorithms [Jaksch et al., 2010]. The optimal average reward of the extended MDP exceeds the optimal average reward of the true MDP since $\mathcal{C}_t$ contains the true discretized transition kernel $\wp_{\mathcal{S}_t\times\mathcal{A}_t\to\mathcal{S}_t,p}$ w.h.p.; this yields an "optimistic push" which ensures "sufficient exploration." The confidence ball shrinks with the number of visits to different state-action pairs; this causes a reduction in the amount of optimism bonus. The extended MDP thus closely approximates the true MDP in the "important regions" (those necessary for recovering an optimal policy) of the state-action space as time progresses. Next, we discuss the extended MDP in detail, how to solve it, and its role in ZoRL.

Extended MDP. Consider the following modified reward function defined on $\mathcal{S}_t \times \mathcal{A}_t$,

$$\tilde{r}_t(s,a) = r(q(q_t^{-1}(s,a))) + L_r \text{diam}\left(q_t^{-1}(s,a)\right),$$

in which a bonus term proportional to the diameter of the active cell that contains (s,a) has been included in order to compensate for the "discretization error." Consider the following collection of MDPs $\mathcal{M}_t^+ := \{(\mathcal{S}_t, \mathcal{A}_t, \tilde{p}, \tilde{r}_t) : \tilde{p} \in \mathcal{C}_t\}$. One may view $\mathcal{M}_t^+$ as an MDP with the finite state space $\mathcal{S}_t$ and an extended action space, hence the name extended MDP. An element from the extended action space has two components: control input from $\mathcal{A}_t$, and a transition kernel from $\mathcal{C}_t$. Let Φ_t be the set of those policies ϕ that satisfy $\phi(s) \in \mathcal{A}_t(s),\ \forall s \in \mathcal{S}_t$. Denote the optimal average reward of $\mathcal{M}_t^+$ by $J^\star_{\mathcal{M}_t^+}$. ZoRL uses the EVI algorithm in order to obtain an optimal policy for the extended MDP at the beginning of every episode. This is discussed next.

Algorithm 1 Extended Value Iteration (EVI)

Input Extended MDP $\mathcal{M}^+$, accuracy parameter $\gamma > 0$.
Initialize $v_0 = \{0\}^{|S|}$, $n = 0$.
while True **do**
  $v_{n+1} = \mathcal{T} v_n$ (14)
  if $sp\left(v_{n+1} - v_n\right) \le \gamma$ **then**
    break
  end if
  $n \leftarrow n + 1$
end while
return Greedy Policy w.r.t. v_n

EVI (Algorithm 1) takes as input an extended MDP, and an error tolerance parameter $\gamma > 0$, and returns a policy whose average reward is γ-close to the optimal value of the extended MDP. A generic extended MDP $\mathcal{M}^+ = \{(S, A, \tilde{p}, \tilde{r}) : \tilde{p} \in \mathcal{C}\}$ has a discrete state space S, and discrete action space $A = \{A(s) : s \in S\}$ where $A(s)$ is the set of actions that are permissible in state s. $\mathcal{C}$ is a set of transition kernels which yield a distribution over S for each point in $S \times A$. $\tilde{r}$ is the reward function. Given the extended MDP $\mathcal{M}^+$, define the following operator $\mathcal{T} : \mathbb{R}^S \mapsto \mathbb{R}^S$,

$$\mathcal{T} v(s) = \max_{\substack{a\in A(s)\\ \theta\in\mathcal{C}}} \Big\{\tilde{r}(s,a) + \sum_{s'\in S} \theta(s,a,s') v(s')\Big\}. \quad (14)$$

See that $\mathcal{T}$ is the Bellman operator [Puterman, 2014] for the extended MDP, $\mathcal{M}^+$, where maximization of the value is done over the extended action space, $A(s) \times \mathcal{C}$. Recall that the Bellman operator for usual MDPs maximizes over the set of all actions. At time τ_k, ZoRL calls $\text{EVI}(M^+_{\tau_k}, 1/\sqrt{T})$. The EVI subroutine then applies the Bellman operator (14) for $M^+_{\tau_k}$ repetitively until stopping criterion is met and returns the policy $\tilde{\phi}_k \in \Phi_{\tau_k}$, which is $1/\sqrt{T}$-near optimal (Lemma G.1). ZoRL then extends $\tilde{\phi}_k$ on the entire continuous space $\mathcal{S}$ to obtain ϕ_k as follows: for every state in the $\mathcal{S}$-cell $\xi \in \mathcal{Q}_{\tau_k}$, ϕ_k plays $\tilde{\phi}_k(q(\xi))$, i.e.,

$$\phi_k(s) = \tilde{\phi}_k(q(\xi)), \forall s \in \xi,\ \xi \in \mathcal{Q}_{\tau_k}. \quad (15)$$

Algorithm 2 Extended Policy Evaluation (EPE)

Input Extended MDP $\mathcal{M}^+$, policy ϕ, accuracy parameter $\gamma > 0$, reference state $s_\star$.
Initialize $v_0 = \{0\}^{|S|}$, $n = 0$.
while True **do**
  $v_{n+1} = \max_{\theta\in\mathcal{C}} \Big\{\tilde{r}(s,\phi(s)) + \sum_{s'\in S} \theta(s,\phi(s),s') v_n(s')\Big\}$
  if $sp\left(v_{n+1} - v_n\right) \le \left(v_{n+1}(s_\star) - v_n(s_\star)\right)\gamma$ **then**
    break
  end if
  $n \leftarrow n + 1$
end while
return $v_{n+1}(s_\star) - v_n(s_\star)$

Episode Duration. ZoRL chooses the duration of the k-th episode as a function of the expected diameter of the states visited at stationarity of the chosen policy, ϕ_k. Define the extended MDP $\mathcal{M}_t^{d,+} = \{(\mathcal{S}_t, \mathcal{A}_t, \tilde{p}, d_t) : \tilde{p} \in \mathcal{C}_t\}$, where

$$d_t(s,a) := \text{diam}\left(q_t^{-1}(s,a)\right),\ \forall (s,a) \in \mathcal{S}_t \times \mathcal{A}_t.$$

Let $\tilde{\phi} \in \Phi_t$. We define the proxy diameter of $\tilde{\phi}$ at time t as the average reward of the policy $\tilde{\phi}$ evaluated on MDP $\mathcal{M}_t^{d,+}$ and denote it by $\widetilde{\text{diam}}_t(\tilde{\phi})$. To be precise, $\widetilde{\text{diam}}_t(\tilde{\phi})$ is the optimal value of $\mathcal{M}_t^{d,+}$ when the control input component of the extended action is chosen according to the policy $\tilde{\phi}$, and the transition kernel is chosen so as to maximize the average reward. Define the diameter of a policy $\phi \in \Phi_{SD}$ at time t as follows:

$$\text{diam}_t(\phi) := \int_{\mathcal{S}} \text{diam}\left(q_t^{-1}(s,\phi(s))\right) \mu^{(\infty)}_{\phi,p}(s) ds. \quad (16)$$

In Appendix C, we show that $\widetilde{\text{diam}}_{\tau_k}(\tilde{\phi}_k)$ is a tight upper-bound of $\text{diam}_{\tau_k}(\phi_k)$ for every k. The duration of the k-th episode, H_k is chosen as,

$$H_k = \frac{C_H \log(T/\delta)}{\widetilde{\text{diam}}_{\tau_k}(\tilde{\phi}_k)^{2(d_{\mathcal{S}}+1)}}, \tag{17}$$

where C_H, a problem-dependent quantity of $\mathcal{O}(\log(T))$, satisfies (68). This choice of episode duration ensures a reduction of the diameter of the chosen policy in every episode. ZoRL uses EPE (Algorithm 2) in order to compute $\widetilde{\text{diam}}_{\tau_k}(\tilde{\phi}_k)$. $\texttt{EPE}(\mathcal{M}_t^{d,+}, \tilde{\phi}, \gamma, s_\star)$ returns a value from $[(1+\gamma)^{-1}\widetilde{\text{diam}}_t(\tilde{\phi}), (1-\gamma)^{-1}\widetilde{\text{diam}}_t(\tilde{\phi})]$ (Corollary G.2) for any $\tilde{\phi} \in \Phi_t$ where γ is a parameter chosen by the agent.

Algorithm 3 Zooming Algorithm for RL (ZoRL)

Input Horizon T, upper-bounds on L_r, L_p, C_p, constants c_a, C_H and accuracy parameter $\gamma > 0$
Initialize $h = 0$, $k = 0$, $H_0 = 0$, $\mathcal{P}_0 = \{\mathcal{S} \times \mathcal{A}\}$
for $t = 0$ to $T-1$ **do**
  if $h \geq H_k$ **then**
    $k \leftarrow k+1$, $h \leftarrow 0$, $\tau_k = t$, $s_\star \in \mathcal{S}_t$
    Construct $\mathcal{M}_{\tau_k}^+$ and $\mathcal{M}_{\tau_k}^{d,+}$
    $\tilde{\phi}_k = \texttt{EVI}(\mathcal{M}_{\tau_k}^+, 1/\sqrt{T})$
    Obtain ϕ_k from $\tilde{\phi}_k$ according to (15)
    $d_k = \texttt{EPE}(\mathcal{M}_{\tau_k}^{d,+}, \tilde{\phi}_k, \gamma, s_\star)$
    $H_k = C_H \log(T/\delta)\, d_k^{-2(d_{\mathcal{S}}+1)}$
  end if
  $h \leftarrow h+1$
  Play $a_t = \phi_k(s_t)$, observe s_{t+1} and receive $r(s_t, a_t)$
  if $N_t(q_t^{-1}(s_t,a_t)) = N_{\max}(q_t^{-1}(s_t,a_t))$ **then**
    $\mathcal{P}_{t+1} = \mathcal{P}_t \cup \mathit{Child}(q_t^{-1}(s_t,a_t)) \setminus \{q_t^{-1}(s_t,a_t)\}$
  else
    $\mathcal{P}_{t+1} = \mathcal{P}_t$
  end if
end for

4 REGRET ANALYSIS

We let $\Delta(\phi) := J_{\mathcal{M}}^\star - J_{\mathcal{M}}(\phi)$ denote the suboptimality of policy ϕ. The following result establishes a relation between the suboptimality of a policy, and the suboptimality gap of the state-action pairs through which this policy passes, where suboptimality gap of state-action pair is defined in (2). Its proof is deferred to Appendix A.

Lemma 4.1. *Consider the MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r)$. For any policy $\phi \in \Phi_{SD}$, we have*

$$\Delta(\phi) = \int_{\mathcal{S}} \mathit{gap}(s, \phi(s))\ \mu_{\phi,p}^{(\infty)}(s)\, ds.$$

We make the following assumption on the true kernel p for deriving concentration bound for the estimate of the discretized transition kernel $\wp_{\mathcal{S}_t \times \mathcal{A}_t \to \mathcal{S}_t, p}$ (8).

Assumption 4.2 (Bounded Radon-Nikodym derivative)**.** *The probability measures $\{p(s,a,\cdot)\}$ are absolutely-continuous w.r.t. the Lebesgue measure on $(\mathcal{S}, \mathcal{B}_{\mathcal{S}})$, with density functions given by $\{f_{(s,a)}\}$. We assume that these densities satisfy*

$$\left\| \frac{\partial f_{(s,a)}(s^+)}{\partial s^+(i)} \right\|_\infty \leq C_p, \forall (s,a) \in \mathcal{S} \times \mathcal{A}, i = 1, 2, \ldots, d_{\mathcal{S}},$$

where the variable $s^+ = (s^+(1), s^+(2), \cdots, s^+(d_{\mathcal{S}}))$ represents the next state.

Assumption 4.2 ensures that the discretizations of $p(s,a,\cdot)$ with respect to the partitions $\mathcal{Q}^{(\ell(q_t^{-1}(s,a)))}$ and $\mathcal{Q}_t$ are at most C_p diam $\left(q_t^{-1}(s,a)\right)$ distance apart (Lemma I.2). Using this result, Lemma F.1 shows that under Assumption 2.1 and Assumption 4.2, $\cap_{t=0}^{T-1}\{\wp_{\mathcal{S}_t \times \mathcal{A}_t \to \mathcal{S}_t, p} \in \mathcal{C}_t\}$ occurs w.h.p. The following assumption allows us to derive an upper-bound on the span of the EVI iterates, which is essential to ensure that the algorithm is not overly optimistic.

Assumption 4.3 (Bound on Stationary Distributions)**.** *There is a constant $\kappa > 0$ such that for every policy $\phi \in \Phi_{SD}$, and for every $\zeta \in \mathcal{B}_{\mathcal{S}}$, we have, $\kappa \cdot \lambda(\zeta) \leq \mu_{\phi,p}^{(\infty)}(\zeta)$, where $\lambda(\cdot)$ denotes the Lebesgue measure on $(\mathcal{S}, \mathcal{B}_{\mathcal{S}})$.*

Remark (Regarding Assumptions)**.** *In the average reward setup for continuous space MDPs, assumptions similar to Assumption 4.3 or more restrictive assumptions are needed. For example, Ormoneit and Glynn [2002] assumes that the transition kernel of the underlying MDP has a strictly positive Radon-Nikodyn derivative in order to show that a proposed adaptive policy converges to an optimal policy. Wang et al. [2024] and Shah and Xie [2018] derive optimal sample complexity for average reward RL and for discounted reward RL, respectively, under an assumption that the m-step transition kernel is bounded below by a known measure. Kar and Singh [2024b] also make the same assumption as ours in order to derive the regret upper-bound of their adaptive discretization-based algorithm. Wei et al. [2021] bounds the regret for average reward RL algorithm when the relative value function is a linear function of a set of known feature maps. Their "uniformly excited features" assumption ensures that upon playing any policy, the confidence ball shrinks in each direction, which has a similar effect as Assumption 4.3.*

We now present our main result that provides an upper-bound on regret of ZoRL. We only provide a proof sketch here and delegate its detailed proof to the appendix.

Theorem 4.4. *Under Assumptions 2.1, 2.2, 4.2 and 4.3, with probability at least $1-\delta$, $\mathcal{R}(T;$ ZoRL$)$ is upper-bounded as $\mathcal{O}\left(T^{1-d_{\mathit{eff.}}^{-1}}\right)$ where $d_{\mathit{eff.}} = 2d_{\mathcal{S}} + d_z + 3$.*

Proof sketch. We decompose the regret (1) in the following manner. Let $K(T)$ denote the total number of episodes

during T timesteps. Then,

$$\mathcal{R}(T; \texttt{ZoRL}) = TJ^\star_{\mathcal{M}} - \sum_{k=1}^{K(T)} \sum_{t=\tau_k}^{\tau_{k+1}-1} r(s_t, a_t)$$
$$= \underbrace{\sum_{k=1}^{K(T)} H_k \left(J^\star_{\mathcal{M}} - J_{\mathcal{M}}(\phi_k)\right)}_{(a)}$$
$$+ \underbrace{\sum_{k=1}^{K(T)} \left(H_k \, J_{\mathcal{M}}(\phi_k) - \sum_{t=\tau_k}^{\tau_{k+1}-1} r(s_t, \phi_k(s_t)) \right)}_{(b)}.$$

(a) captures the regret arising due to playing a suboptimal policy ϕ_k during the k-th episode, while (b) captures the possible degradation in performance during the transient stage as compared with the average rewards of the chosen policies. (a) and (b) are bounded separately below.

Bounding (a): Step 1: In Lemma B.1, we show that the policy obtained by solving $\mathcal{M}_t^+$ is optimistic, i.e., w.h.p. $J^\star_{\mathcal{M}_t^+} \geq J^\star_{\mathcal{M}}$. Also, in Lemma B.3, we show that w.h.p., $J^\star_{\mathcal{M}_t^+} \leq J^\star_{\mathcal{M}} + C_{ub}\,\text{diam}_t(\phi_k)$, where C_{ub} is as defined in (53). As a consequence of the above two results, on a high probability set, a suboptimal policy ϕ will never be played from episode k onwards if $\text{diam}_{\tau_k}(\phi) \leq C_{ub}^{-1} \cdot \Delta(\phi)$. Note that the cumulative regret arising due to policies with $\Delta(\cdot)$ less than ϵ is at most ϵT. We choose ϵ optimally and restrict the analysis to regret arising from playing other policies.

Step 2: We combine Step 1 with Lemma 4.1 in Lemma E.1 and show that on a high probability set, in each episode k, there is a state $s \in \mathcal{S}$ such that

$$\text{diam}\,(\zeta) \geq \frac{1}{3C_{ub}} \max\{\text{gap}\,(s, \phi_k(s)), C_{ub}\text{diam}_{\tau_k}(\phi_k)\}, \quad (18a)$$
$$\mu^{(\infty)}_{\phi_k, p}(\pi_{\mathcal{S}}(\zeta)) \geq \left(\frac{\text{diam}_{\tau_k}(\phi_k)}{3}\right)^{d_{\mathcal{S}}+1}, \quad (18b)$$

where $\zeta = q^{-1}_{\tau_k}(s, \phi_k(s))$. This cell ζ is called a key cell in the k-th episode.

Step 3: Then we show that with a high probability, the key cells of the k-th episode are visited at least $\mathcal{O}\left(\log\left(\frac{T}{\delta}\right)\text{diam}\,(\zeta)^{-(d_{\mathcal{S}}+1)}\right)$ times during the k-th episode. This is done in Lemma E.2.

Step 4: We obtain a bound on the cardinality of the key cells associated with playing policies from the set $\Phi_{2^{-i}} = \{\phi \in \Phi_{SD} \mid \Delta(\phi) \in (2^{-i}, 2^{-i+1}]\}$ by showing that these cells are contained within a set of cells that has a cardinality at most $\mathcal{O}(2^{id_z})$. We then use this bound along with the lower-bound on the number of plays of the key cells, and conclude that the policies from $\Phi_{2^{-i}}$ are played for a maximum of $\mathcal{O}\left(\log\left(\frac{T}{\delta}\right)2^{i(2d_{\mathcal{S}}+d_z+3)}\right)$ time-steps (Lemma E.3).

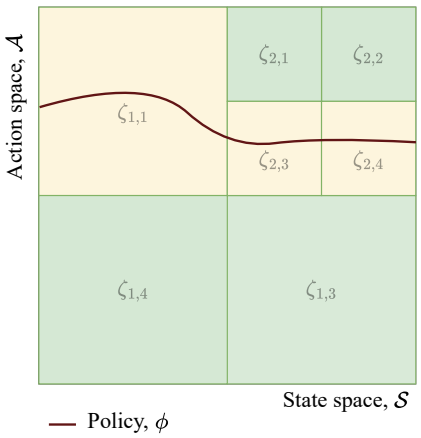


Figure 1: Key cell: The policy ϕ is played during the k-th episode. This diagram depicts the discretization grid at the beginning of the k-th episode. Then, one of the cells $\zeta_{1,1}$, $\zeta_{2,3}$ and $\zeta_{2,4}$ must be a key cell with a high probability (Lemma E.1). There must be a state s such that $(s, \phi(s))$ belongs to this cell, and s satisfies (18a) and (18b).

Step 5: The term (a) can be written as the sum of the regrets arising due to playing policies from the sets $\Phi_{2^{-i}}$, where $i = 1, 2, \ldots, \left\lceil \log\left(\frac{1}{\epsilon}\right)\right\rceil$, where $\epsilon = T^{-\frac{1}{2d_{\mathcal{S}}+d_z+3}}$. To bound the regret arising due to playing policies from $\Phi_{2^{-i}}$, we multiply $\mathcal{O}\left(\log\left(\frac{T}{\delta}\right)2^{i(2d_{\mathcal{S}}+d_z+3)}\right)$ by 2^{-i+1}. We then add these regret terms from $i = 1$ to $\left\lceil \log\left(\frac{1}{\epsilon}\right)\right\rceil$ and ϵT.

Step 6: Lastly, we add $\sqrt{T}$ to the final bound to compensate for the inaccuracy caused by `EVI` due to finite computational resources. This gives us the upper-bound on (a) w.h.p.

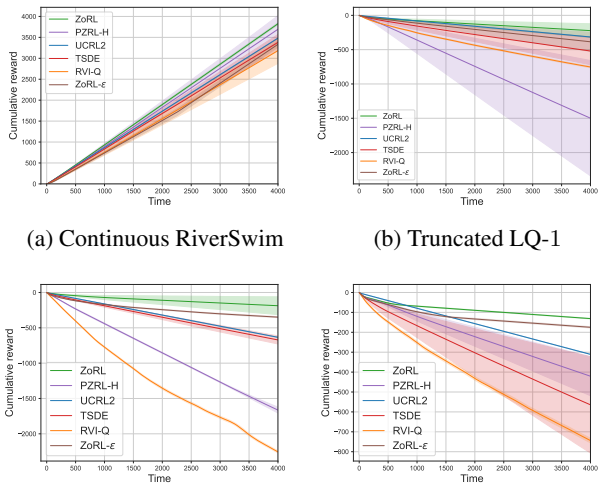
Bounding (b): upper-bound on the term (b) relies on the uniform ergodicity property (Assumption 2.2) of $\mathcal{M}$ and a trick that converts "Markovian noise" to "martingale noise" [Metivier and Priouret, 1984]. Proposition E.4 shows that on a high probability set, we must pay a constant penalty each time we change policy, which is $\mathcal{O}(K(T) + \sqrt{T})$. We show that the rule which decides when to start a new episode ensures that $K(T)$ is bounded above by $\mathcal{O}(T^{\frac{d_z+1}{2d_{\mathcal{S}}+d_z+3}})$, and so is the term (b).

Summing the upper-bounds on (a) and (b), we obtain the desired regret bound. □

5 SIMULATIONS

We compare the performance of `ZoRL` (Algorithm 3) with that of UCRL2 [Jaksch et al., 2010], TSDE [Ouyang et al., 2017], RVI-Q [Borkar and Meyn, 2000] which is a Q-learning algorithm for average-reward RL, ZoRL-ϵ [Kar and Singh, 2024a], and the heuristic algorithm PZRL-H [Kar and Singh, 2024b]. For competitor policies that are designed for finite state-action spaces, we apply

them on a uniform discretization of $\mathcal{S} \times \mathcal{A}$ performed at time $t = 0$. Simulation experiments are conducted on the following systems: (i) `Continuous RiverSwim`, where the environment models an agent who is swimming in a river. (ii) Linear Quadratic (LQ) control systems [Abbasi-Yadkori and Szepesvári, 2011] where the state evolves as $s_{t+1} = As_t + Ba_t + w_t$, and we truncate the state-action space in order to ensure that they are compact. Denote the two systems of dimension 2×2 and 2×4 as `Truncated LQ-1` and `Truncated LQ-2`, respectively. (iii) `Non-linear System` where the state evolves as $s_{t+1} = Af(s_t) + Bg(a_t) + w_t$, where f and g are non-linear functions. Similar to the truncated LQ systems, we truncate the state-action space. Details of the environments can be found in Kar and Singh [2024a], and also in Appendix H. We plot the cumulative rewards averaged over 50 runs in Figure 2. `ZoRL` performs the best among all six algorithms on each of the environments. Very recently, Kar and Singh [2024b] has replaced PZRL-H with two algorithms, PZRL-MB and PZRL-MF. In Appendix H we compare their performance with `ZoRL`.



(a) Continuous RiverSwim (b) Truncated LQ-1

(c) Truncated LQ-2 (d) Non-linear System

Figure 2: Cumulative Reward Plots.

6 CONCLUSION

We propose a computationally efficient algorithm for average reward RL for Lipschitz MDPs in continuous spaces, and show that it is truly adaptive, i.e. it achieves a regret of $\tilde{\mathcal{O}}\big(T^{1-d_{\text{eff.}}^{-1}}\big)$, where $d_{\text{eff.}} = 2d_{\mathcal{S}} + d_z + 3$. The zooming dimension d_z is a problem-dependent quantity, measures the size of near-optimal state-action pairs and is bounded above by d, the dimension of the state-action space. Simulation experiments support the theoretical findings. `ZoRL` overperforms the popular fixed discretization-based algorithms as well as adaptive discretization-based algorithms.

Acknowledgements

This work is partially supported by the SERB Grant SRG/2021/002308. The authors acknowledge the Prime Minister's Research Fellowship to Avik Kar.

References

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

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. Advances in neural information processing systems, 24:2312–2320, 2011.

Aristotle Arapostathis, Vivek S. Borkar, Emmanuel Fernández-Gaucherand, Mrinal K Ghosh, and Steven I Marcus. Discrete-time controlled Markov processes with average cost criterion: a survey. SIAM Journal on Control and Optimization, 31(2):282–344, 1993.

Vivek S Borkar and Sean P Meyn. The ODE method for convergence of stochastic approximation and reinforcement learning. SIAM Journal on Control and Optimization, 38 (2):447–469, 2000.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. Journal of Machine Learning Research, 12(5), 2011.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for Markov decision processes. Mathematics of Operations Research, 22(1):222–255, 1997.

Tongyi Cao and Akshay Krishnamurthy. Provably adaptive reinforcement learning in metric spaces. Advances in Neural Information Processing Systems, 33:9736–9744, 2020.

Sayak Ray Chowdhury and Aditya Gopalan. Online learning in kernelized Markov decision processes. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 3197–3205. PMLR, 2019.

Omar Darwiche Domingues, Pierre Menard, Matteo Pirotta, Emilie Kaufmann, and Michal Valko. Kernel-based reinforcement learning: A finite-time analysis. In Marina Meila and Tong Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 2783–2792. PMLR, 2021.

Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. Markov chains: Basic definitions. Springer, 2018.

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

Adityanand Guntuboyina and Bodhisattva Sen. L1 covering numbers for uniformly bounded convex functions. In Conference on Learning Theory, pages 12–1. JMLR Workshop and Conference Proceedings, 2012.

Botao Hao, Nevena Lazic, Yasin Abbasi-Yadkori, Pooria Joulani, and Csaba Szepesvári. Adaptive approximate policy iteration. In International Conference on Artificial Intelligence and Statistics, pages 523–531. PMLR, 2021.

Jianliang He, Han Zhong, and Zhuoran Yang. Sample-efficient learning of infinite-horizon average-reward MDPs with general function approximation. In The Twelfth International Conference on Learning Representations, 2023.

Onésimo Hernández-Lerma. Adaptive Markov control processes, volume 79. Springer Science & Business Media, 2012.

Onésimo Hernández-Lerma and Jean B Lasserre. Further topics on discrete-time Markov control processes, volume 42. Springer Science & Business Media, 2012.

Julian Ibarz, Jie Tan, Chelsea Finn, Mrinal Kalakrishnan, Peter Pastor, and Sergey Levine. How to train your robot with deep reinforcement learning: lessons we have learned. The International Journal of Robotics Research, 40(4-5):698–721, 2021.

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

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In Conference on Learning Theory, pages 2137–2143. PMLR, 2020.

Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. Advances in Neural Information Processing Systems, 33:15312–15325, 2020.

Avik Kar and Rahul Singh. Adaptive discretization-based non-episodic reinforcement learning in metric spaces. arXiv preprint arXiv:2405.18793v3, 2024a.

Avik Kar and Rahul Singh. Policy zooming: Adaptive discretization-based infinite-horizon average-reward reinforcement learning. arXiv preprint arXiv:2405.18793v1, 2024b.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In Proceedings of the fortieth annual ACM symposium on Theory of computing, pages 681–690, 2008.

Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. RMA: Rapid motor adaptation for legged robots. arXiv preprint arXiv:2107.04034, 2021.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. Cambridge University Press, 2020.

Davide Maran, Alberto Maria Metelli, Matteo Papini, and Marcello Restell. No-regret reinforcement learning in smooth MDPs. The Forty-first International Conference on Machine Learning, 2024a.

Davide Maran, Alberto Maria Metelli, Matteo Papini, and Marcello Restelli. Projection by convolution: Optimal sample complexity for reinforcement learning in continuous-space MDPs. The 37th Annual Conference on Learning Theory, 2024b.

Michel Metivier and Pierre Priouret. Applications of a Kushner and Clark lemma to general classes of stochastic algorithms. IEEE Transactions on Information Theory, 30(2):140–151, 1984.

Dirk Ormoneit and Peter Glynn. Kernel-based reinforcement learning in average-cost problems. IEEE Transactions on Automatic Control, 47(10):1624–1636, 2002.

Ronald Ortner. Regret bounds for reinforcement learning via Markov chain concentration. Journal of Artificial Intelligence Research, 67:115–128, 2020.

Ronald Ortner and Daniil Ryabko. Online regret bounds for undiscounted continuous reinforcement learning. Advances in Neural Information Processing Systems, 25, 2012.

Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. Advances in Neural Information Processing Systems, 27, 2014.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown Markov decision processes: A thompson sampling approach. In Advances in Neural Information Processing Systems, pages 1333–1342, 2017.

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

Maxim Raginsky, Igal Sason, et al. Concentration of measure inequalities in information theory, communications, and coding. Foundations and Trends® in Communications and Information Theory, 10(1-2):1–246, 2013.

Devavrat Shah and Qiaomin Xie. Q-learning with nearest neighbors. Advances in Neural Information Processing Systems, 31, 2018.

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

Sean R Sinclair, Siddhartha Banerjee, and Christina Lee Yu. Adaptive discretization in online reinforcement learning. Operations Research, 71(5):1636–1652, 2023.

Paloma Sodhi, Felix Wu, Ethan R Elenberg, Kilian Q Weinberger, and Ryan McDonald. On the effectiveness of offline RL for dialogue response generation. In International Conference on Machine Learning, pages 32088–32104. PMLR, 2023.

Zhao Song and Wen Sun. Efficient model-free reinforcement learning in metric spaces. arXiv preprint arXiv:1905.00475, 2019.

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

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

Aristide Tossou, Debabrota Basu, and Christos Dimitrakakis. Near-optimal optimistic reinforcement learning using empirical bernstein inequalities. arXiv preprint arXiv:1905.12425, 2019.

Aad W Van Der Vaart, Jon A Wellner, Aad W van der Vaart, and Jon A Wellner. Weak convergence. Springer, 1996.

Shengbo Wang, Jose Blanchet, and Peter Glynn. Optimal sample complexity for average reward Markov decision processes. In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=jOm5p3q7c7.

Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, and Rahul Jain. Learning infinite-horizon average-reward MDPs with linear function approximation. In International Conference on Artificial Intelligence and Statistics, pages 3007–3015. PMLR, 2021.

Yue Wu, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal regret for learning infinite-horizon average-reward MDPs with linear function approximation. In International Conference on Artificial Intelligence and Statistics, pages 3883–3913. PMLR, 2022.

Adaptive Discretization-based Non-Episodic Reinforcement Learning in Metric Spaces (Supplementary Material)

Avik Kar[1] **Rahul Singh**[1]

[1] Department of Electrical Communication Engineering, Indian Institute of Science, Bengaluru

Organization of the Appendix. Some properties of the MDPs that satisfy Assumption 2.2 are discussed in Appendix A. It also includes the proof of Lemma 4.1. Some important properties of extended MDPs can be found in Appendix B. We use these properties while analyzing the regret of ZoRL. Next, in Appendix C, we show certain properties of the proxy diameters of policies. Results obtained in Appendix B play a crucial role in deriving those properties. A high probability lower bound on the number of visits to the key cells in each episode is derived in Appendix D. In Appendix E, we derive the desired regret bound. Appendix F covers the concentration results for estimates of the discretized model. In Appendix G, we derive bounds on inaccuracy that EVI and EPE injects into ZoRL due to finite computation power. Details of the experiments, the associated environments and additional simulation results are reported in Appendix H. Appendix I derives some key results that are used in the proof of Lemma F.1. Appendix J contains some known results that are used in this paper.

A GENERAL RESULTS FOR MDPS

Consider an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r)$ and a policy $\phi \in \Phi_{SD}$ that maps states in $\mathcal{S}$ to actions in $\mathcal{A}$. We assume that the transition kernel p satisfies Assumption 2.2. Hence, there exists a unique invariant distribution $\mu_{\phi,p}^{(\infty)}$ for the controlled Markov process (CMP) induced by the transition kernel p under the application of policy ϕ. Under Assumption 2.2, there exists a solution to the following Poisson equation [Hernández-Lerma and Lasserre, 2012]:

$$J + h(s) = r(s, \phi(s)) + \int_{\mathcal{S}} h(s')p(s, \phi(s), ds'),\ \forall s \in \mathcal{S}. \tag{19}$$

Specifically, $(J_{\mathcal{M}}(\phi), h_{\mathcal{M}}^{\phi}) \in \mathbb{R} \times \mathbb{R}^{\mathcal{S}}$ satisfies (19), where

$$J_{\mathcal{M}}(\phi) = \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} r(s_t, \phi(s_t)) \mid s_0 = s\right] = \int_{\mathcal{S}} r(s, \phi(s))\mu_{\phi,p}^{(\infty)}(ds), \tag{20}$$

$$\text{and } h_{\mathcal{M}}^{\phi}(s') = \sum_{t=0}^{\infty} \int_{\mathcal{S}} r(s, \phi(s))(\mu_{\phi,p}^{(\infty)} - \mu_{\phi,p,s'}^{(t)})(ds),\ \forall s' \in \mathcal{S}. \tag{21}$$

Recall that $\mu_{\phi,p,s}^{(t)}$ denotes the distribution of s_t when initial state is $s_0 = s$, where $\{s_t\}_t$ is the CMP induced by the transition kernel p under the application of ϕ. $h_{\mathcal{M}}^{\phi}$ is called the relative value function of ϕ.

The following is popularly known as the average reward optimality equation (AROE),

$$J + h(s) = \max_{a\in\mathcal{A}} \left\{ r(s, a) + \int_{\mathcal{S}} h(s')p(s, a, s')ds' \right\}, \text{ and}$$
$$h(s_\star) = 0,$$

where $s_\star \in \mathcal{S}$ is a designated state. Hernández-Lerma [2012] shows that under Assumption 2.2, AROE has a solution. A

policy $\phi^\star$ is optimal if it satisfies the following,

$$\phi^\star(s) \in \arg\max_{a\in\mathcal{A}}\left\{ r(s,a) + \int_{\mathcal{S}} h_{\mathcal{M}}^{\phi^\star}(s')p(s,a,s')ds' \right\},\ \forall s \in \mathcal{S}. \tag{22}$$

In that case, $J_{\mathcal{M}}^\star = J_{\mathcal{M}}(\phi^\star)$ and $h_{\mathcal{M}} = h_{\mathcal{M}}^{\phi^\star}$ solve AROE.

Denote the t-stage transition kernel under the application of policy ϕ by $p_\phi^{(t)}$, i.e.,

$$p_\phi^{(t)}(s,B) = \mathbb{P}(s_{\tau+t} \in B \mid s_\tau = s, a_{t'} = \phi(s_{t'}), t' = \tau,\tau+1,\ldots,\tau+t-1),\ t\in\mathbb{N}, s\in\mathcal{S}, B\in\mathcal{B}_{\mathcal{S}}, \tau\in\mathbb{N}. \tag{23}$$

Our next result shows that when t is sufficiently large, then Assumption 2.2 is equivalent to saying that $p_\phi^{(t)}$ has the "contractive property," (24).

Lemma A.1. *Consider an MDP $\mathcal{M} = (\mathcal{S},\mathcal{A},p,r)$ such that p satisfies Assumption 2.2. Then, for every policy $\phi \in \Phi_{SD}$ we have,*

$$\left\| p_\phi^{(i)}(s,\cdot) - p_\phi^{(i)}(s',\cdot) \right\|_{TV} \le 2\alpha,\ \forall s,s'\in\mathcal{S}, i \ge m^\star, \tag{24}$$

where $p_\phi^{(i)}$ is the i-stage transition probability of the CMP induced by the transition kernel p under the application of policy ϕ as defined in (23)*, and*

$$m^\star := \left\lceil \log_{\frac{1}{\alpha}}(C) \right\rceil + 1. \tag{25}$$

Conversely, if

$$\left\| p_\phi^{(m)}(s,\cdot) - p_\phi^{(m)}(s',\cdot) \right\|_{TV} \le 2\alpha',\ \forall s,s'\in\mathcal{S},$$

for some $m\in\mathbb{N}$, then Assumption 2.2 holds with $C = \frac{2}{\alpha'}$ and $\alpha = {\alpha'}^{\frac{1}{m}}$.

Proof. We first note that $p_\phi^{(i)}(s,\cdot) = \mu_{\phi,p,s}^{(i)}$ for every $s\in\mathcal{S}$. Hence, for any $s,s'\in\mathcal{S}$,

$$\left\| p_\phi^{(i)}(s,\cdot) - p_\phi^{(i)}(s',\cdot) \right\|_{TV} \le \left\| \mu_{\phi,p,s}^{(i)} - \mu_{\phi,p}^{(\infty)} \right\|_{TV} + \left\| \mu_{\phi,p,s'}^{(i)} - \mu_{\phi,p}^{(\infty)} \right\|_{TV}.$$

Also, $C\alpha^i \le \alpha$ for $i \ge \log_{\frac{1}{\alpha}}(C) + 1$. Now, using Assumption 2.2, we have that when $i \ge m^\star$, then the following holds,

$$\begin{aligned}\left\| p_\phi^{(i)}(s,\cdot) - p_\phi^{(i)}(s',\cdot) \right\|_{TV} &\le \left\| \mu_{\phi,p,s}^{(i)} - \mu_{\phi,p}^{(\infty)} \right\|_{TV} + \left\| \mu_{\phi,p,s'}^{(i)} - \mu_{\phi,p}^{(\infty)} \right\|_{TV} \\ &\le 2\alpha.\end{aligned}$$

This concludes the proof of the first claim.

Now, we prove the second claim. Consider the CMP that is described by the transition kernel p and evolves under the application of the policy ϕ. Consider two copies of this CMP, where these copies differ in the distribution of the initial state. Denote these distributions by $\mu_1^{(0)}$ and $\mu_2^{(0)}$. Denote the distributions of s_i in the corresponding processes by $\mu_1^{(i)}$ and $\mu_2^{(i)}$, respectively. We show the following:

$$\left\| \mu_1^{(i)} - \mu_2^{(i)} \right\|_{TV} \le \tilde{C}\cdot\tilde{\alpha}^i \left\| \mu_1^{(0)} - \mu_2^{(0)} \right\|_{TV},\ \forall i\in\mathbb{N}, \tag{26}$$

where $\tilde{C} = \frac{1}{\alpha'}$ and $\tilde{\alpha} = {\alpha'}^{\frac{1}{m}}$. The claim then follows by letting $\mu_1^{(0)} = \delta_s$ and $\mu_2^{(0)} = \mu_{\phi,p}^{(\infty)}$. Note that,

$$\begin{aligned}\left\| \mu_1^{(m)} - \mu_2^{(m)} \right\|_{TV} &= 2\sup_{A\subseteq\mathcal{S}}\left\{ (\mu_1^{(m)} - \mu_2^{(m)})(A) \right\} \\ &= 2\sup_{A\subseteq\mathcal{S}}\left\{ \int_{\mathcal{S}} p_\phi^{(m)}(s,A)\, d(\mu_1^{(0)} - \mu_2^{(0)})(s) \right\} \\ &\le \sup_{\substack{A\subseteq\mathcal{S}\\ s,s'\in\mathcal{S}}}\left\{ p_\phi^{(m)}(s,A) - p_\phi^{(m)}(s',A) \right\} \left\| \mu_1^{(0)} - \mu_2^{(0)} \right\|_{TV} \\ &\le \alpha' \left\| \mu_1^{(0)} - \mu_2^{(0)} \right\|_{TV}.\end{aligned} \tag{27}$$

Also, note that for any $i \in \mathbb{N}$,

$$\begin{aligned}\left\|\mu_1^{(i)} - \mu_2^{(i)}\right\|_{TV} &= 2 \sup_{A \subseteq \mathcal{S}} \left\{(\mu_1^{(i)} - \mu_2^{(i)})(A)\right\} \\ &= 2 \sup_{A \subseteq \mathcal{S}} \left\{\int_{\mathcal{S}} p(s, \phi(s), A)\, d(\mu_1^{(i-1)} - \mu_2^{(i-1)})(s)\right\} \\ &\leq \sup_{\substack{A \subseteq \mathcal{S} \\ s, s' \in \mathcal{S}}} \{p(s, \phi(s), A) - p(s', \phi(s'), A)\} \left\|\mu_1^{(i-1)} - \mu_2^{(i-1)}\right\|_{TV} \\ &\leq \left\|\mu_1^{(i-1)} - \mu_2^{(i-1)}\right\|_{TV}, \end{aligned} \tag{28}$$

where the first step follows from the definition of the total variation norm, while the third step follows from Lemma J.6. Combining (27) and (28), we can write

$$\begin{aligned}\left\|\mu_1^{(i)} - \mu_2^{(i)}\right\|_{TV} &\leq \alpha'^{\left\lfloor \frac{i}{m} \right\rfloor} \left\|\mu_1^{(0)} - \mu_2^{(0)}\right\|_{TV} \\ &\leq \frac{1}{\alpha'} \left(\alpha'^{\frac{1}{m}}\right)^i \left\|\mu_1^{(0)} - \mu_2^{(0)}\right\|_{TV}, \ \forall i \in \mathbb{N}.\end{aligned}$$

This concludes the proof of the lemma. □

Consider two CMPs $\{s_{1,i}\}$ and $\{s_{2,i}\}$, both of which are induced by ϕ operating on the MDP $\mathcal{M}$ that has transition kernel p. Their initial state distributions are $\mu_1^{(0)}$ and $\mu_2^{(0)}$ respectively. Next, we derive an upper-bound on the cumulative sum of distances of the distributions of $s_{1,i}$ and $s_{2,i}$.

Lemma A.2. *Consider an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r)$ that satisfies Assumption 2.2, and a policy $\phi \in \Phi_{SD}$. Let $\{s_{1,i}\}$ and $\{s_{2,i}\}$ be two CMPs induced by ϕ when it is applied to $\mathcal{M}$. Let $\mu_1^{(i)}$ and $\mu_2^{(i)}$ denote the distributions of $s_{1,i}$ and $s_{2,i}$, respectively. Then,*

$$\sum_{i=0}^{\infty} \left\|\mu_1^{(i)} - \mu_2^{(i)}\right\|_{TV} \leq \frac{m^\star}{1-\alpha} \left\|\mu_1^{(0)} - \mu_2^{(0)}\right\|_{TV},$$

where $m^\star$ is as defined in (25)*.*

Proof. From Lemma A.1, we have that,

$$\left\|\mu_1^{(i)} - \mu_2^{(i)}\right\|_{TV} \leq \alpha \left\|\mu_1^{(0)} - \mu_2^{(0)}\right\|_{TV}, \text{ for } i \geq m^\star. \tag{29}$$

Also, for any $i \in \mathbb{N}$ we have,

$$\begin{aligned}\left\|\mu_1^{(i)} - \mu_2^{(i)}\right\|_{TV} &= 2 \sup_{A \subseteq \mathcal{S}} \left\{(\mu_1^{(i)} - \mu_2^{(i)})(A)\right\} \\ &= 2 \sup_{A \subseteq \mathcal{S}} \left\{\int_{\mathcal{S}} p(s, \phi(s), A)\, d(\mu_1^{(i-1)} - \mu_2^{(i-1)})(s)\right\} \\ &\leq \sup_{\substack{A \subseteq \mathcal{S} \\ s, s' \in \mathcal{S}}} \{p(s, \phi(s), A) - p(s', \phi(s'), A)\} \left\|\mu_1^{(i-1)} - \mu_2^{(i-1)}\right\|_{TV} \\ &\leq \left\|\mu_1^{(i-1)} - \mu_2^{(i-1)}\right\|_{TV},\end{aligned}$$

where the first step follows from the definition of the total variation norm, and the third step follows from Lemma J.6. Hence,

$$\left\|\mu_1^{(i)} - \mu_2^{(i)}\right\|_{TV} \leq \left\|\mu_1^{(0)} - \mu_2^{(0)}\right\|_{TV}, \ \forall i \in \mathbb{N} \tag{30}$$

Using (29) iteratively, and (30), we can write,

$$\sum_{t=0}^{\infty} \left\| \mu_1^{(i)} - \mu_2^{(i)} \right\|_{TV} = \sum_{m=0}^{m^\star - 1} \sum_{i=0}^{\infty} \left\| \mu_1^{(m+i\cdot m^\star)} - \mu_2^{(m+i\cdot m^\star)} \right\|_{TV} \leq \frac{m^\star}{1-\alpha} \left\| \mu_1^{(0)} - \mu_2^{(0)} \right\|_{TV}.$$

where $m^\star = \left\lceil \log_{\frac{1}{\alpha}} (C) \right\rceil + 1$. This concludes the proof. □

We now derive an upper-bound on the span of the relative value function $h^\phi_{\mathcal{M}}$ (21) associated with a policy $\phi \in \Phi_{SD}$.

Lemma A.3 (Bound on the span of relative value function)**.** *Consider an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r)$ such that p satisfies Assumption 2.2. For any policy $\phi \in \Phi_{SD}$, the span of the corresponding relative value function $h^\phi_{\mathcal{M}}$ (21) can be bounded as,*

$$sp\left(h^\phi_{\mathcal{M}}\right) \leq \frac{m^\star \, sp\,(r)}{1-\alpha}, \tag{31}$$

where $m^\star$ is as defined in (25)*.*

Proof. From the definition of $h^\phi_{\mathcal{M}}$ (21) we obtain,

$$\begin{aligned} sp\left(h^\phi_{\mathcal{M}}\right) &= sp\left(\sum_{t=0}^{\infty} \int_{\mathcal{S}} r(s,\phi(s)) \left(\mu^{(\infty)}_{\phi,p} - \mu^{(t)}_{\phi,p,\cdot}\right)(ds)\right) \\ &\leq \sum_{t=0}^{\infty} sp\left(\int_{\mathcal{S}} r(s,\phi(s)) \left(\mu^{(\infty)}_{\phi,p} - \mu^{(t)}_{\phi,p,\cdot}\right)(ds)\right) \\ &\leq \frac{1}{2} \sum_{t=0}^{\infty} \max_s \left\| \mu^{(\infty)}_{\phi,p} - \mu^{(t)}_{\phi,p,s} \right\|_{TV} sp\,(r)\,, \end{aligned} \tag{32}$$

where the first inequality follows since span is a seminorm [Puterman, 2014], while the second inequality follows from Lemma J.6. In Lemma A.2 we let $\mu_1^{(0)} = \mu^{(\infty)}_{\phi,p}$ and $\mu_2^{(0)} = \delta_s$, where δ_s is the Dirac measure on $(\mathcal{S}, \mathcal{B}_{\mathcal{S}})$ centered at s, and get the following,

$$\frac{1}{2} \sum_{t=0}^{\infty} \max_s \left\| \mu^{(\infty)}_{\phi,p} - \mu^{(t)}_{\phi,p,s} \right\|_{TV} sp\,(r) \leq \frac{m^\star \, sp\,(r)}{1-\alpha}.$$

This concludes the proof. □

Lemma A.4 (Bound on the span of policy evaluation iterates)**.** *Consider an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r)$ such that p satisfies Assumption 2.2, and consider the policy evaluation algorithm applied to obtain the average reward of a policy $\phi \in \Phi_{SD}$ on $\mathcal{M}$ i.e.,*

$$\begin{aligned} V_0^\phi(s) &= 0, \\ V_{i+1}^\phi(s) &= r(s,\phi(s)) + \int_{\mathcal{S}} p(s,\phi(s),s') V_i^\phi(s') ds', \; i = 1,2,\ldots. \end{aligned} \tag{33}$$

We have,

$$sp\left(V_i^\phi\right) \leq \frac{m^\star + 1}{1-\alpha}, \tag{34}$$

where $m^\star = \left\lceil \log_{\frac{1}{\alpha}} (C) \right\rceil + 1$.

Proof. Since Assumption 2.2 holds, Lemma A.1 gives us the following,

$$\left\| p_\phi^{(m^\star)}(s,\cdot) - p_\phi^{(m^\star)}(s',\cdot) \right\|_{TV} \leq 2\alpha,\ \forall s, s' \in \mathcal{S},$$

where $p_\phi^{(m)}$ (23) is the m-step transition kernel of the CMP induced by the transtion kernel p under the application of policy ϕ. Also, note that

$$V_{i+m^\star}^\phi(s) = \sum_{j=0}^{m^\star} \mathbb{E}\left[r(s_{i+j}, \phi(s_{i+j})) \mid s_i = s\right] + \int_{\mathcal{S}} p_\phi^{(m^\star)}(s,s') V_i^\phi(s') ds'.$$

Hence,

$$\begin{aligned} sp\left(V_{i+m^\star}^\phi\right) &\leq sp\left(\sum_{j=0}^{m^\star} \mathbb{E}\left[r(s_{i+j}, \phi(s_{i+j})) \mid s_i = s\right]\right) + sp\left(\int_{\mathcal{S}} p_\phi^{(m^\star)}(s,s') V_i^\phi(s') ds'\right) \\ &\leq m^\star + 1 + \frac{1}{2} sp\left(V_i^\phi\right) \left\| p_\phi^{(m^\star)}(s,\cdot) - p_\phi^{(m^\star)}(s',\cdot) \right\|_{TV} \\ &\leq m^\star + 1 + \alpha\, sp\left(V_i^\phi\right), \end{aligned}$$

where the second inequality follows from Lemma J.6. Using the above inequality, we have that for every $k \leq m^\star$,

$$\begin{aligned} sp\left(V_{i\cdot m^\star + k}^\phi\right) &\leq (m^\star + 1) \sum_{j=0}^{i-1} \alpha^j + \alpha^i\, sp\left(V_k^\phi\right) \\ &\leq (m^\star + 1) \sum_{j=0}^{i-1} \alpha^j + m^\star \alpha^i \\ &\leq \frac{m^\star + 1}{1-\alpha}. \end{aligned}$$

This concludes the proof. □

A.1 PROOF OF LEMMA 4.1

Proof. Using the definition of gap $(s, \phi(s))$ (2), we obtain that,

$$\begin{aligned} \int_{\mathcal{S}} \text{gap}\,(s,\phi(s))\ \mu_{\phi,p}^{(\infty)}(s)\, ds &= \int_{\mathcal{S}} \left(J_{\mathcal{M}}^\star + h_{\mathcal{M}}(s) - r(s,\phi(s)) - \int_{\mathcal{S}} h_{\mathcal{M}}(s')\, p(s,\phi(s),s')\, ds' \right) \mu_{\phi,p}^{(\infty)}(s)\, ds \\ &= J_{\mathcal{M}}^\star \int_{\mathcal{S}} \mu_{\phi,p}^{(\infty)}(s)\, ds + \int_{\mathcal{S}} h_{\mathcal{M}}(s) \mu_{\phi,p}^{(\infty)}(s)\, ds - \int_{\mathcal{S}} r(s,\phi(s)) \mu_{\phi,p}^{(\infty)}(s)\, ds \\ &\quad - \int_{\mathcal{S}} \left(\int_{\mathcal{S}} h_{\mathcal{M}}(s')\, p(s,\phi(s),s')\, ds' \right) \mu_{\phi,p}^{(\infty)}(s)\, ds \\ &= J_{\mathcal{M}}^\star + \int_{\mathcal{S}} h_{\mathcal{M}}(s) \mu_{\phi,p}^{(\infty)}(s)\, ds - J_{\mathcal{M}}(\phi) \\ &\quad - \int_{\mathcal{S}} h_{\mathcal{M}}(s') \left(\int_{\mathcal{S}} p(s,\phi(s),s') \mu_{\phi,p}^{(\infty)}(s)\, ds \right) ds' \\ &= J_{\mathcal{M}}^\star - J_{\mathcal{M}}(\phi) + \int_{\mathcal{S}} h_{\mathcal{M}}(s) \mu_{\phi,p}^{(\infty)}(s)\, ds - \int_{\mathcal{S}} h_{\mathcal{M}}(s) \mu_{\phi,p}^{(\infty)}(s)\, ds \\ &= \Delta(\phi), \end{aligned} \tag{35}$$

where the third equality follows from (20) and the fourth equality follows from the property of the stationary distribution. This concludes the proof. □

B PROPERTIES OF EXTENDED MDP

We present three results in this section. We begin by showing that extended MDPs constructed by ZoRL are optimistic, i.e., on the set $\mathcal{G}_1$ (82), the optimal average reward of the extended MDP $\mathcal{M}_t^+$ is greater than or equal to the optimal average reward of the true MDP for all $t \in \{0, 1, \ldots, T-1\}$. Next, we show that the span of the EPE iterates (42) for the extended MDP $\mathcal{M}_t^+$ and any $\phi \in \Phi_t$ are bounded for all $t \in \{0, 1, \ldots, T-1\}$. Lastly, we derive an upper-bound on the average reward of policy $\phi \in \Phi_t$ evaluated on MDP $\mathcal{M}_t^+$ for every $t \in \{0, 1, \ldots, T-1\}$.

Lemma B.1 (Optimism)**.** *On the set $\mathcal{G}_1$, we have,*

$$J^\star_{\mathcal{M}_t^+} \geq J^\star_{\mathcal{M}}, \text{ for every } t \in \{0, 1, \ldots, T-1\}, \tag{36}$$

where $J^\star_{\mathcal{M}_t^+}$ is the optimal average reward of the extended MDP $\mathcal{M}_t^+$, and $J^\star_{\mathcal{M}}$ is the optimal average reward of the MDP $\mathcal{M}$.

Proof. Consider the value iteration algorithm applied to the MDP $\mathcal{M}$. For every $s \in \mathcal{S}$,

$$\begin{aligned} V_0(s) &= 0, \\ V_{n+1}(s) &= \max_{a \in \mathcal{A}} \Big\{ r(s,a) + \int_{\mathcal{S}} p(s,a,s') V_n(s') ds' \Big\}, \ \forall n \in \mathbb{N}. \end{aligned} \tag{37}$$

We assumed that $\mathcal{M}$ is uniformly ergodic in Assumption 2.2, and hence the following value iteration algorithm converges, i.e., $\lim_{n\to\infty} sp\left(V_{n+1} - V_n\right) = J^\star_{\mathcal{M}}$. Also, it follows from [Hernández-Lerma, 2012] that $\lim_{n\to\infty} |V_n(s) - (nJ^\star_{\mathcal{M}} + h_{\mathcal{M}}(s))| = 0$ for every $s \in \mathcal{S}$. Since we have shown in Lemma A.3 that $h_{\mathcal{M}}$ is bounded, it then follows that

$$\lim_{n\to\infty} \frac{1}{n} V_n(s) = J^\star_{\mathcal{M}}, \ \forall s \in \mathcal{S}. \tag{38}$$

We will prove that $V_n(s') \leq v_n(s)$ for every $n \in \mathbb{N}$, $s \in \mathcal{S}_t$ and $s' \in q^{-1}(s)$. We prove this via induction. The base case, i.e. $n = 0$ is seen to hold trivially. Next, assume that the following hold for all $i \in [n]$, where $n \in \mathbb{N}$,

$$v_i(s) \geq V_i(s'), \ \forall s \in \mathcal{S}_t, \ \forall s' \in q^{-1}(s). \tag{39}$$

Consider a state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ and let $\tilde{s} \in \mathcal{S}_t$ such that $s \in q^{-1}(\tilde{s})$. Then,

$$\begin{aligned} r(s,a) + \int_{\mathcal{S}} p(s,a,s') V_n(s') ds' &\leq r(s,a) + \sum_{s' \in \mathcal{S}_t} \wp_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}_t, p}(s,a,s') v_n(s') \\ &\leq r(q(\zeta)) + L_r \text{diam}\,(\zeta) + \sum_{s' \in \mathcal{S}_t} \wp_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}_t, p}(s,a,s') v_n(s') \\ &\leq \max_{\substack{\tilde{a} \in A_t(\tilde{s}) \\ \theta \in \mathcal{C}_t}} \left\{ \tilde{r}_t(\tilde{s}, \tilde{a}) + \sum_{s' \in \mathcal{S}_t} \theta(\tilde{s}, \tilde{a}, s') v_n(s') \right\} \\ &= v_{n+1}(\tilde{s}), \end{aligned} \tag{40}$$

where the first inequality follows from (39), the second inequality follows from Assumption 2.1 (i), while the third inequality follows from the definition of the set $\mathcal{G}_1$. Since we have shown the above inequality for an arbitrary action a, we get,

$$\begin{aligned} V_{n+1}(s) &= \max_{a \in \mathcal{A}} \left\{ r(s,a) + \int_{\mathcal{S}} p(s,a,s') V_n(s') ds' \right\} \\ &\leq v_{n+1}(\tilde{s}). \end{aligned} \tag{41}$$

This completes the induction argument. The proof is then completed by dividing both sides of this inequality by n and then taking limit $n \to \infty$. □

Lemma B.2. *Let $t \in \{0, 1, \ldots, T-1\}$. Consider the extended MDP $\mathcal{M}_t^+$, a policy $\phi \in \Phi_t$ and the corresponding* EPE (2) *iterates:*

$$v_0^{\phi,t}(s) = 0,$$
$$v_{n+1}^{\phi,t}(s) = \max_{\theta \in \mathcal{C}_t} \left\{ \tilde{r}_t(s, \phi(s)) + \sum_{s' \in \mathcal{S}_t} \theta(s, \phi(s), s') v_n^{\phi,t}(s') \right\}, \ \forall s \in \mathcal{S}_t, n \in \mathbb{N}. \tag{42}$$

On the set $\mathcal{G}_1$, we have

$$sp\left(v_n^{\phi,t}\right) \leq C_v, \ \forall n \in \mathbb{N}, t \in \mathbb{N},$$

where,

$$C_v := \max \left\{ \frac{\overline{m}(\overline{m}+5)}{2} + \frac{3}{C\alpha^{\overline{m}+1}} + \frac{4\tilde{m}}{1-\alpha}, \frac{\left\lceil \log_{\left(\frac{1}{\alpha}\right)^{\tilde{m}-1}} \left(\frac{2}{\alpha}\right) + 1 \right\rceil}{1 - \alpha^{\tilde{m}-1}} \right\}, \tag{43}$$

$$\overline{m} := \left\lceil \log_{\frac{1}{\alpha}} \left(\frac{2C}{\kappa} \left(\frac{C_\eta \tilde{m} \sqrt{d}}{1-\alpha} \right)^{d_{\mathcal{S}}} \right) \right\rceil, \text{ and} \tag{44}$$

$$\tilde{m} := \left\lceil \log_{\frac{1}{\alpha}} \left(\frac{2C}{3\alpha - 1} \right) \right\rceil. \tag{45}$$

C and α are as in Assumption 2.2.

Proof. We first note that $v_n^{\phi,t}(s)$ is the optimal value of the expected reward for the extended MDP $\mathcal{M}_t^+$ that is accumulated during the first n steps when the process starts in state s. The first component of the extended action of the extended MDP is taken to be policy ϕ and doesn't need to be optimized, while the second component is the transition kernel that maximizes the r.h.s. of (42) in every step $i \in \{0, 1, \ldots, n-1\}$. We consider the following two cases separately.

Case 1: When,

$$\max_{s \in \mathcal{S}_t} \text{diam}\left(q_t^{-1}(s, \phi(s))\right) \geq \frac{1-\alpha}{2(3(1+L_p)+C_p)(\tilde{m}+1)}. \tag{46}$$

Let ζ be the cell with the largest diameter from the set $\{q_t^{-1}(s, \phi(s)) : \ s \in \mathcal{S}_t\}$. We first show that $\{s_i\}_{i=0}^{\infty}$, the CMP induced by the transition kernel p under the application of policy ϕ, hits $\pi_{\mathcal{S}}(\zeta)$ within

$$\frac{\overline{m}(\overline{m}+5)}{2} + \frac{3}{C\alpha^{\overline{m}+1}}$$

steps in expectation, where $\overline{m}$ is as defined in (44). From Assumption 2.2, Assumption 4.3 and (46), we have that for any $s' \in \mathcal{S}$,

$$\mu_{\phi,p,s'}^{(i)}(\pi_{\mathcal{S}}(\zeta)) \geq \frac{1}{2}\mu_{\phi,p}^{(\infty)}(\pi_{\mathcal{S}}(\zeta)), \text{ and } \mu_{\phi,p,s'}^{(i)}(\pi_{\mathcal{S}}(\zeta)) \leq \frac{3}{2}\mu_{\phi,p}^{(\infty)}(\pi_{\mathcal{S}}(\zeta)) \ \forall i \geq \overline{m}.$$

Now, consider another process $\{x_i\}_{i=0}^{\infty}$ that is independent across time; x_i assumes the value 1 with a probability $\mu_{\phi,p,s'}^{(i)}(\pi_{\mathcal{S}}(\zeta))$, and 0 with a probability $1 - \mu_{\phi,p,s'}^{(i)}(\pi_{\mathcal{S}}(\zeta))$. Define the following random variables $T_{\{1\}}^{(x)}$ and $T_{\pi_{\mathcal{S}}(\zeta),s'}^{(s)}$,

$$T_{\{1\}}^{(x)} := \inf\{i \geq 0 \mid x_i = 1\}, \text{ and}$$
$$T_{\pi_{\mathcal{S}}(\zeta),s'}^{(s)} := \inf\{i \geq 0 \mid s_i \in \pi_{\mathcal{S}}(\zeta), s_0 = s'\}.$$

We note that the distributions of $T_{\{1\}}^{(x)}$ and $T_{\pi_{\mathcal{S}}(\zeta),s'}^{(s)}$ are identical, so that $\mathbb{E}\left[T_{\{1\}}^{(x)}\right] = \mathbb{E}\left[T_{\pi_{\mathcal{S}}(\zeta),s'}^{(s)}\right]$. We derive an upper-bound

on $\mathbb{E}\left[T^{(x)}_{\{1\}}\right]$, and this would also serve as the upper-bound on $\mathbb{E}\left[T^{(s)}_{\pi_{\mathcal{S}}(\zeta),s'}\right]$. We have,

$$\begin{aligned}
\mathbb{E}\left[T^{(x)}_{\{1\}}\right] &= \sum_{i=0}^{\infty} i \cdot \mu^{(i)}_{\phi,p}(\pi_{\mathcal{S}}(\zeta)) \prod_{j=0}^{i-1}\left(1-\mu^{(j)}_{\phi,p,s}(\pi_{\mathcal{S}}(\zeta))\right) \\
&\leq \frac{\overline{m}(\overline{m}-1)}{2} + \sum_{i=\overline{m}}^{\infty} \frac{3i}{2}\mu^{(\infty)}_{\phi,p}(\pi_{\mathcal{S}}(\zeta)) \prod_{j=\overline{m}}^{i-1}\left(1-\frac{1}{2}\mu^{(\infty)}_{\phi,p}(\pi_{\mathcal{S}}(\zeta))\right) \\
&\leq \frac{\overline{m}(\overline{m}-1)}{2} + \frac{3}{2}\mu^{(\infty)}_{\phi,p}(\pi_{\mathcal{S}}(\zeta)) \sum_{i=0}^{\infty} i\left(1-\frac{1}{2}\mu^{(\infty)}_{\phi,p}(\pi_{\mathcal{S}}(\zeta))\right)^{i} + \frac{3\overline{m}}{2}\mu^{(\infty)}_{\phi,p}(\pi_{\mathcal{S}}(\zeta)) \sum_{i=0}^{\infty}\left(1-\frac{1}{2}\mu^{(\infty)}_{\phi,p}(\pi_{\mathcal{S}}(\zeta))\right)^{i} \\
&\leq \frac{\overline{m}(\overline{m}+5)}{2} + \frac{6}{\mu^{(\infty)}_{\phi,p}(\pi_{\mathcal{S}}(\zeta))}.
\end{aligned}$$

Furthermore, from Assumption 4.3, and since $\mathbb{E}\left[T^{(x)}_{\{1\}}\right] = \mathbb{E}\left[T^{(s)}_{\pi_{\mathcal{S}}(\zeta),s'}\right]$, we get,

$$\mathbb{E}\left[T^{(s)}_{\pi_{\mathcal{S}}(\zeta),s'}\right] \leq \frac{\overline{m}(\overline{m}+5)}{2} + \frac{6}{\kappa}\left(\frac{\sqrt{d}}{\text{diam}\,(\zeta)}\right)^{d_{\mathcal{S}}}.$$

From (46) we can write,

$$\begin{aligned}
\mathbb{E}\left[T^{(s)}_{\pi_{\mathcal{S}}(\zeta),s'}\right] &\leq \frac{\overline{m}(\overline{m}+5)}{2} + \frac{6}{\kappa}\left(\frac{(3(1+L_p)+C_p)\sqrt{d}(\tilde{m}+1)}{1-\alpha}\right)^{d_{\mathcal{S}}} \\
&\leq \frac{\overline{m}(\overline{m}+5)}{2} + \frac{3}{C\alpha^{\overline{m}+1}}.
\end{aligned}$$

Next, consider two states $\bar{s} \in \mathcal{S}_t$, and $\tilde{s} \in q^{-1}(\bar{s})$. We note that on the set $\mathcal{G}_1$, for the extended MDP $\mathcal{M}^+_t$ whenever the state is $\bar{s}$, there is an extended action such that the next state transition distribution is $p(\tilde{s}, \phi(\tilde{s}), \cdot)$. Hence, on the set $\mathcal{G}_1$, there is a sequence of extended actions such that starting from any state, in expectation, within $\frac{\overline{m}(\overline{m}+5)}{2} + \frac{3}{C\alpha^{\overline{m}+1}}$ steps the process hits $q(\pi_{\mathcal{S}}(\zeta))$ where $\pi_{\mathcal{S}}(\zeta)$ is the $\mathcal{S}$-projection of ζ, the largest cell in $\{q_t^{-1}(s,\phi(s)) :\ s \in \mathcal{S}_t\}$.

Now, consider the process $\{s_t\}$ associated with the extended MDP, in which the initial state is $s \in \mathcal{S}_t$. We claim that for any state s', there exists a sequence of extended actions where the first components of the extended actions are chosen by ϕ such that s' can be reached in $\frac{2}{(3(1+L_p)+C_p)\,\text{diam}\left(q_t^{-1}(s,\phi(s))\right)}$ steps in expectation. This is true because there is a transition kernel in $\mathcal{C}_t$ that assigns at least $\frac{3(1+L_p)+C_p}{2}\text{diam}\left(q_t^{-1}(s,\phi(s))\right)$ transition probability to s' when the current state is from s. To summarize, starting from any state using a sequence of actions the state process can reach $q(\zeta)$ in $\frac{\overline{m}(\overline{m}+5)}{2} + \frac{3}{C\alpha^{\overline{m}+1}}$ steps in expectation, and from $q(\zeta)$, again it can reach any other state using a sequence of actions in $\frac{2}{(3(1+L_p)+C_p)\text{diam}\left(q_t^{-1}(s,\phi(s))\right)}$. Therefore, there cannot be state s' such that

$$\max_{s\in\mathcal{S}_t} v_n^{\phi,t}(s) > v_n^{\phi,t}(s') + \frac{\overline{m}(\overline{m}+5)}{2} + \frac{3}{C\alpha^{\overline{m}+1}} + \frac{2}{(3(1+L_p)+C_p)\text{diam}\,(\zeta)}.$$

Now, from the lower-bound on diam (ζ) (46), we obtain that

$$sp\left(v_n^{\phi,t}\right) \leq \frac{\overline{m}(\overline{m}+5)}{2} + \frac{3}{C\alpha^{\overline{m}+1}} + \frac{4\tilde{m}}{1-\alpha}. \tag{47}$$

Case 2: In this case, we have that

$$\max\left\{\text{diam}\left(q_t^{-1}(s,\phi(s))\right) : s \in \mathcal{S}_t\right\} < \frac{1-\alpha}{2(3(1+L_p)+C_p)\,(\tilde{m}+1)}. \tag{48}$$

Let $\bar{\phi} \in \Phi_{SD}$ be the extension of policy $\phi \in \Phi_t$ such that

$$\bar{\phi}(s) = \phi(q(\pi_{\mathcal{S}}(\zeta))),\ \text{for ever } s \in \pi_{\mathcal{S}}(\zeta),\ \text{for every } \pi_{\mathcal{S}}(\zeta) \in \mathcal{Q}_t.$$

Claim: We claim that there is a sequence of extended actions for the extended MDP $\mathcal{M}_t^+$ such that the first components of the extended actions are governed by ϕ and on the set $\mathcal{G}_1$, the m-step state transition kernel prescribed by the sequence of extended actions is the same as the discretization of the m-step composition of true transition kernel induced under application of policy $\bar{\phi}$. Let the state process of the extended MDP be denoted by $\{\tilde{s}_i\}$ and let the state process of the extended MDP be denoted by $\{s_i\}$. Then, mathematically, our claim says that there exists a sequence of probability kernels $\{\tilde{p}_i \in \mathcal{C}_t : i \in \{1, 2, \ldots\}\}$ such that

$$\mathbb{P}(\tilde{s}_i = s' \mid \tilde{s}_0 = s, \tilde{p}, \phi) = \mathbb{P}(s_i \in q_t^{-1}(s') \mid s_0 = s, \bar{\phi}),\ \forall s, s' \in \mathcal{S}_t,$$

where $\mathbb{P}$ denotes the joint probability distribution of the processes $\{\tilde{s}_i\}$ and $\{s_i\}$, condition on $\tilde{p}$ and ϕ implies that the extended actions are governed by $\tilde{p}$ and ϕ. Similarly, condition on $\bar{\phi}$ implies that the actions are governed by $\bar{\phi}$. We show this using mathematical induction. The base cases follow from Lemma F.1. Let us assume that for every $s, s' \in \mathcal{S}_t$ and for every $j \in \{1, 2, \ldots i\}$,

$$\mathbb{P}(\tilde{s}_j = s' \mid \tilde{s}_0 = s, \tilde{p}, \phi) = \mathbb{P}(s_j \in q_t^{-1}(s') \mid s_0 = s, \bar{\phi}).$$

See that

$$\begin{aligned}\mathbb{P}(\tilde{s}_{i+1} = s' \mid \tilde{s}_0 = s, \tilde{p}, \phi) &= \sum_{\tilde{s} \in \mathcal{S}_t} \mathbb{P}(\tilde{s}_{i+1} = s' \mid \tilde{s}_i = \tilde{s}, \tilde{p}, \phi)\mathbb{P}(\tilde{s}_i = \tilde{s} \mid \tilde{s}_0 = s, \tilde{p}, \phi) \\ &= \sum_{\tilde{s} \in \mathcal{S}_t} \tilde{p}_{i+1}(\tilde{s}, \phi(\tilde{s}), s')\mathbb{P}(s_i = q_t^{-1}(\tilde{s}) \mid s_0 = s, \bar{\phi}).\end{aligned}$$

Here, we note that for every $s \in \mathcal{S} \times \mathcal{A}$, there is a kernel $\theta_s \in \mathcal{C}_t$ such that $\theta_s(q_t^{-1}(s, \phi(s)), s') = p(s, \bar{\phi}(s), q_t^{-1}(s'))$ for every $s' \in \mathcal{S}_t$. As the set $\mathcal{C}_t$ is convex, for any probability measure ν on $(\mathcal{S}, \mathcal{B}_\mathcal{S})$,

$$\int_\mathcal{S} \theta_s(\tilde{s}, \phi(\tilde{s}), s') d\nu(s) \in \mathcal{C}_t.$$

Taking ν to be a measure that satisfies $\nu(B) = \mathbb{P}(s_i \in B \mid s_i \in q_t^{-1}(\tilde{s}))$ for every $B \in \mathcal{B}_\mathcal{S}$, we get that

$$\int_\mathcal{S} \theta_s(\tilde{s}, \phi(\tilde{s}), s') d\nu(s) = \mathbb{P}(s_{i+1} \in q_t^{-1}(s') \mid s_i \in q_t^{-1}(\tilde{s})).$$

Taking $\tilde{p}_{i+1}(\tilde{s}, \phi(\tilde{s}), \cdot) = \int_\mathcal{S} \theta_s(\tilde{s}, \phi(\tilde{s}), \cdot) d\nu(s)$, we get that

$$\begin{aligned}\mathbb{P}(\tilde{s}_{i+1} = s' \mid \tilde{s}_0 = s, \tilde{p}, \phi) &= \sum_{\tilde{s} \in \mathcal{S}_t} \tilde{p}_{i+1}(\tilde{s}, \phi(\tilde{s}), s')\mathbb{P}(s_i = q_t^{-1}(\tilde{s}) \mid s_0 = s, \bar{\phi}) \\ &= \sum_{\tilde{s} \in \mathcal{S}_t} \mathbb{P}(s_{i+1} \in q_t^{-1}(s') \mid s_i \in q_t^{-1}(\tilde{s}))\mathbb{P}(s_i = q_t^{-1}(\tilde{s}) \mid s_0 = s, \bar{\phi}) \\ &= \mathbb{P}(s_{i+1} \in q_t^{-1}(s') \mid s_0 = s, \bar{\phi}).\end{aligned}$$

This completes the proof of our claim.

From (48), we have that for any $\theta \in \mathcal{C}_t$,

$$\max_{s \in \mathcal{S}_t} \|\theta(s, \phi(s), \cdot) - \tilde{p}_i(s, \phi(s), \cdot)\|_1 \leq \frac{1 - \alpha}{2\tilde{m}},\ \forall s \in \mathcal{S}_t, s' \in q_t^{-1}(s).$$

Define the discretization of the m-step transition kernel under the application of policy $\bar{\phi}$ as follows:

$$\wp_{t,\phi}^{(m)}(s, s') := p_\phi^{(m)}(s, q_t^{-1}(s')),\ \forall s \in \mathcal{S}, s' \in \mathcal{S}_t.$$

Let $\theta_\phi^{(m)}$ denote the m-step transition kernel of the CMP induced by θ under application of policy ϕ. From the previous claim and Lemma J.7, we have that

$$\left\| \wp_{t,\phi}^{(\tilde{m})}(s, \cdot) - \theta_\phi^{(\tilde{m})}(s, \cdot) \right\|_1 \leq \frac{1 - \alpha}{2}, \tag{49}$$

where $p_\phi^{(m)}$ is defined in (23). Also, observe that

$$\max_{s,s' \in \mathcal{S}_t} \left\| \wp_{t,\phi}^{(\tilde{m})}(s,\cdot) - \wp_{t,\phi}^{(\tilde{m})}(s',\cdot) \right\|_1 \leq \frac{3\alpha - 1}{2}. \tag{50}$$

Hence, combining (49) and (50), we have that for any $\theta \in \mathcal{C}_t$,

$$\begin{aligned}
\max_{s,s' \in \mathcal{S}_t} \left\| \theta_\phi^{(\tilde{m})}(s,\cdot) - \theta_\phi^{(\tilde{m})}(s',\cdot) \right\|_1 &\leq \max_{s,s' \in \mathcal{S}_t} \Big\{ \left\| \theta_\phi^{(\tilde{m})}(s,\cdot) - \wp_{t,\phi}^{(\tilde{m})}(s,\cdot) \right\|_1 + \left\| \wp_{t,\phi}^{(\tilde{m})}(s,\cdot) - \wp_{t,\phi}^{(\tilde{m})}(s',\cdot) \right\|_1 \\
&\quad + \left\| \wp_{t,\phi}^{(\tilde{m})}(s',\cdot) - \theta_\phi^{(\tilde{m})}(s',\cdot) \right\|_1 \Big\} \\
&\leq \frac{1-\alpha}{2} + 3\alpha - 1 + \frac{1-\alpha}{2} \\
&= 2\alpha.
\end{aligned}$$

Now, from Lemma A.1, we have that the Markov chain induced by the transition kernel θ under the application of policy ϕ is uniformly ergodic with constants $\frac{2}{\alpha}$ and $\alpha^{\tilde{m}^{-1}}$, i.e.,

$$\left\| \mu_{\phi,\theta,s}^{(i)} - \mu_{\phi,\theta}^{(\infty)} \right\|_1 \leq \frac{2}{\alpha} \cdot \left(\alpha^{\tilde{m}^{-1}} \right)^i, \ \forall i \in \mathbb{N}.$$

Hence, from Lemma A.4, we conclude that

$$sp\left(v_n^{\phi,t}\right) \leq \frac{\left\lceil \log_{\left(\frac{1}{\alpha}\right)^{\tilde{m}^{-1}}} \left(\frac{2}{\alpha}\right) \right\rceil + 1}{1 - \alpha^{\tilde{m}^{-1}}}. \tag{51}$$

Combining the upper-bounds from (47) and (51), we obtain the desired upper-bound. □

In the next lemma, we establish that the optimism injected by ZoRL is not huge.

Lemma B.3. *Consider time $t \in \mathbb{N}$ and a policy $\phi \in \Phi_t$. Let $\bar{\phi} \in \Phi_{SD}$ be the extension of ϕ as follows:*

$$\bar{\phi}(s) = \phi(q(\xi)), \text{ for every } s \in \xi, \text{ for every } \xi \in \mathcal{Q}_t.$$

Then, we have that on the set $\mathcal{G}_1$,

$$J_{\mathcal{M}_t^+}(\phi) \leq J_{\mathcal{M}}(\bar{\phi}) + C_{ub}\, diam_t(\bar{\phi}), \ \forall t \in \mathbb{N}, \phi \in \Phi_t, \tag{52}$$

where $J_{\mathcal{M}_t^+}(\phi)$ is the optimal value of $\mathcal{M}_t^+$ when the control input component of the extended action is chosen according to the policy ϕ, and the transition kernel is chosen so as to maximize the average reward, $diam_t(\bar{\phi})$ is as defined in (16)*, and*

$$C_{ub} := 2L_r + (3(1+L_p) + C_p)C_v. \tag{53}$$

L_r, L_p are as stated in Assumption 2.1, C_p is as stated in Assumption 4.3, and C_v is as defined in (43)*.*

Proof. Consider the iteration (42). From Corollary G.2 it follows that

$$\lim_{n\to\infty} \left(v_{n+1}^\phi(s) - v_n^\phi(s) \right) = J_{\mathcal{M}_t^+}(\phi), \ \text{ for every } s \in \mathcal{S}_t.$$

As the sequence of Cesaro means converges to the same limit, we can write

$$\lim_{n\to\infty} \frac{1}{n} v_n^\phi(s) = J_{\mathcal{M}_t^+}(\phi).$$

Similarly, from the policy evaluation iteration for the true MDP (33), we have that

$$\lim_{n\to\infty} \frac{1}{n} V_n^{\bar{\phi}}(s) = J_{\mathcal{M}}(\bar{\phi}).$$

In order to prove the lemma, we will show that on the set $\mathcal{G}_1$, for every $n \in \mathbb{N}$, for every $s \in \mathcal{S}_t$ and for every $s' \in q^{-1}(s)$, the following holds,

$$v_n^{\phi}(s) \leq V_n^{\bar{\phi}}(s') + C_{ub}\, \mathbb{E}_{p,\bar{\phi}} \left[\sum_{i=0}^{n-1} \text{diam}\left(q_t^{-1}(s_i, \bar{\phi}(s_i))\right) \middle| s_0 = s' \right], \tag{54}$$

where $\mathbb{E}_{p,\phi}$ denotes that the expectation is taken with respect to the measure induced by ϕ when it is applied to MDP with transition kernel p. We prove this using induction. The base case $(n = 0)$ is seen to hold trivially. Next, we assume that the following holds for $i \in \{0, 1, \ldots, n\}$, where $n \in \mathbb{N}$,

$$v_i^{\phi}(s) \leq V_i^{\bar{\phi}}(s') + C_{ub}\, \mathbb{E}_{p,\bar{\phi}} \left[\sum_{j=0}^{i-1} \text{diam}\left(q_t^{-1}(s_j, \bar{\phi}(s_j))\right) \middle| s_0 = s' \right], \tag{55}$$

for every $s \in \mathcal{S}_t$ and for every $s' \in q^{-1}(s)$. Let us fix $s \in \mathcal{S}_t$ and $s' \in q^{-1}(s)$ arbitrarily, then from (42) we obtain the following,

$$\begin{aligned}
v_{n+1}^{\phi}(s) &= r(q(q_t^{-1}(s, \phi(s)))) + \max_{\theta \in \mathcal{C}_t} \sum_{s'' \in \mathcal{S}_t} \theta(q(q_t^{-1}(s, \phi(s))), s'') v_n^{\phi}(s'') + L_r \,\text{diam}\left(q_t^{-1}(s, \phi(s))\right) \\
&= r(q(q_t^{-1}(s, \phi(s)))) + \sum_{s'' \in \mathcal{S}_t} \theta_n(q(q_t^{-1}(s, \phi(s))), s'') \bar{V}_n^{\phi}(s'') + L_r \,\text{diam}\left(q_t^{-1}(s, \phi(s))\right) \\
&\leq r(s', \phi(s')) + \sum_{s'' \in \mathcal{S}_t} \wp(s', \phi(s'), s''; \mathcal{S}_t \times A_t, \mathcal{Q}_t)\, v_n^{\phi}(s'') + \eta_t(q_t^{-1}(s, \phi(s)))\, sp\left(v_n^{\phi}\right) + 2L_r \,\text{diam}\left(q_t^{-1}(s, \phi(s))\right) \\
&\leq r(s', \phi(s')) + \int_{\mathcal{S}} p(s', \phi(s'), s'') V_n^{\phi}(s'') ds'' + C_{ub}\, \mathbb{E}_{p,\phi} \left[\sum_{i=1}^{n} \text{diam}\left(q_t^{-1}(s_i, \phi(s_i))\right) \middle| s_0 = s' \right] \\
&\quad + (2L_r + (3(1 + L_p) + C_p) C_v) \,\text{diam}\left(q_t^{-1}(s, \phi(s))\right) \\
&\leq r(s', \phi(s')) + \int_{\mathcal{S}} p(s', \phi(s'), s'') V_n^{\phi}(s'') ds'' + C_{ub}\, \mathbb{E}_{p,\phi} \left[\sum_{i=1}^{n} \text{diam}\left(q_t^{-1}(s_i, \phi(s_i))\right) \middle| s_0 = s' \right] \\
&\quad + (2L_r + (3(1 + L_p) + C_p) C_v) \,\text{diam}\left(q_t^{-1}(s, \phi(s))\right) \\
&= V_{n+1}^{\phi}(s) + C_{ub}\, \mathbb{E}_{p,\phi} \left[\sum_{i=0}^{n} \text{diam}\left(q_t^{-1}(s_i, \phi(s_i))\right) \middle| s_0 = s \right],
\end{aligned}$$

where θ_n is a transition kernel belonging to the set $\mathcal{C}_t$ that maximizes the expression in the r.h.s. of the first equality. The first inequality follows from Lipschitz continuity of the reward function, the definition of event $\mathcal{G}_1$ and from Lemma J.6. The second inequality is obtained by invoking the induction hypothesis (55), and by using the upper-bound on $sp\left(v_n^{\phi}\right)$ from Lemma B.2. This concludes the induction argument, and proves (54). The proof of the claim follows by dividing both side of (54) by n and taking limit $n \to \infty$. □

C PROPERTIES OF PROXY DIAMETER

In this section, we present three results as the corollaries of the results obtained in the previous section.

Corollary C.1. *Fix a time t. Let $\phi \in \Phi_t$ and $\bar{\phi} \in \Phi_{SD}$ be the unique extension of ϕ such that*

$$\bar{\phi}(s') = \phi(s), \text{ for every } s \in \mathcal{S}_t \text{ and } s' \in q^{-1}(s). \tag{56}$$

On the set $\mathcal{G}_1$, we have,

$$\widetilde{diam}_t(\phi) \geq diam_t(\phi),\ \forall t \in \{0, 1, \ldots, T-1\}, \phi \in \Phi_t. \tag{57}$$

where $\widetilde{diam}_t(\phi)$ is the average reward of policy ϕ evaluated on the extended MDP $\mathcal{M}_t^{d,+}$ and $diam_t(\bar{\phi}) = \int_{\mathcal{S}} q_t^{-1}(s, \phi(s)) \mu_{\phi,p}^{(\infty)}(s) ds$.

Proof. Define the MDP, $\mathcal{M}_t^d := (\mathcal{S}, \mathcal{A}, p, \tilde{d})$ where

$$\tilde{d}(s,a) = \text{diam}\left(q_t^{-1}(s,a)\right), \text{ for every } (s,a) \in \mathcal{S} \times \mathcal{A}.$$

As p satisfy Assumption 2.2,

$$J_{\mathcal{M}_t^d}(\bar{\phi}) = \text{diam}_t(\bar{\phi}), \text{ for every } \bar{\phi} \in \Phi_{SD}.$$

Note that the extended policy evaluation (42) and policy evaluation (33) algorithms are equivalent to extended value iteration (93) and value iteration (37) algorithms, respectively, except that the control inputs have to be chosen from singleton sets. Then the proof follows from Lemma B.1. □

Corollary C.2. *Let $t \in \{0, 1, \ldots, T-1\}$. Consider the extended MDP $\mathcal{M}_t^{d,+}$, a policy $\phi \in \Phi_t$ and the corresponding* `EPE` (2) *iterates:*

$$g_0^{\phi,t}(s) = 0,$$
$$g_{n+1}^{\phi,t}(s) = \max_{\theta \in \mathcal{C}_t} \left\{ d_t(s, \phi(s)) + \sum_{s' \in \mathcal{S}_t} \theta(s, \phi(s), s') g_n^{\phi,t}(s') \right\}, \ \forall s \in \mathcal{S}_t, n \in \mathbb{N}.$$

On the set $\mathcal{G}_1$, we have

$$sp\left(g_n^{\phi,t}\right) \leq C_v, \ \forall n \in \mathbb{N}, t \in \mathbb{N},$$

where, C_v, $\overline{m}$ and $\tilde{m}$ are defined in (43)*,* (44) *and* (45)*, respectively.*

Proof. Follows from Lemma B.2. □

Corollary C.3. *Consider time $t \in \mathbb{N}$ and a policy $\phi \in \Phi_t$. Let $\bar{\phi} \in \Phi_{SD}$ be the extension of ϕ as defined in* (56)*. Then, we have that on the set $\mathcal{G}_1$,*

$$\widetilde{diam}_t(\phi) \leq (C_{ub} + 1)\, diam_t(\bar{\phi}), \ \forall t \in \mathbb{N}, \phi \in \Phi_t,$$

where C_{ub} is as defined in (53)*.*

Proof. Noting that $J_{\mathcal{M}_t^d}(\bar{\phi}) = \text{diam}_t(\bar{\phi})$ and $J_{\mathcal{M}_t^{d,+}}(\phi) = \widetilde{\text{diam}}_t(\phi)$, the claim follows from Lemma B.3 and Corollary C.2. □

D GUARANTEE ON NUMBER OF VISITS TO CELLS

Recall that $\mu_{\phi,p,s}^{(t)}$ denotes the distribution of s_t when policy ϕ is applied to the MDP that has the transition kernel p and the initial state is s, and $\mu_{\phi,p}^{(\infty)}$ denotes the unique invariant distribution of the Markov chain induced by the policy ϕ on the MDP with transition kernel p. Consider an $\mathcal{S}$-cell ξ for which the diameter is greater than ϵ, and $\mu_{\phi,p}^{(\infty)}(\xi) \geq (\epsilon/3)^{d_{\mathcal{S}}+1}$ for all stationary deterministic policies ϕ, where $\epsilon > 0$. Later we will choose an appropriate value for ϵ. From Assumption 2.2 we get that for all $\phi \in \Phi_{SD}$ and for every initial state $s \in \mathcal{S}$ we have,

$$\mu_{\phi,p,s}^{(t)}(\xi) \geq \mu_{\phi,p}^{(\infty)}(\xi) - \frac{C}{2}\alpha^t.$$

Since $\mu_{\phi,p}^{(\infty)}(\xi) \geq (\epsilon/3)^{d_{\mathcal{S}}+1}$, we have

$$\mu_{\phi,p,s}^{(t)}(\xi) \geq \frac{1}{2}\mu_{\phi,p}^{(\infty)}(\xi), \ \forall t \geq t^\star(\epsilon), \tag{58}$$

where,

$$t^\star(\epsilon) := \left\lceil \log_{\frac{1}{\alpha}} \left(C \left(\frac{3}{\epsilon}\right)^{d_{\mathcal{S}}+1} \right) \right\rceil. \tag{59}$$

Lemma D.1. *Fix $k \in \mathbb{N}$ and consider a $\mathcal{S}$-cell $\xi \in \mathcal{Q}_{\tau_k}$ such that $\mu^{(\infty)}_{\phi,p}(\xi) \geq (\epsilon/3)^{d_{\mathcal{S}}+1}$. Let $\zeta \in \mathcal{P}_{\tau_k}$ denote the active cell that contains $\{(s,\phi_k(s))\}_{s\in\xi}$. Let $n_k(\zeta)$ be the number of visits to ζ in the k-th episode, and H_k be the duration of the k-th episode. Then, with a probability at least $1-\frac{\delta}{3}$, we have,*

$$n_k(\zeta) \geq \frac{H_k\, \mu^{(\infty)}_{\phi,p}(\xi)}{2t^\star(\epsilon)} - \sqrt{\frac{H_k}{t^\star(\epsilon)} \log\left(\frac{6T}{t^\star(\epsilon)\delta}\right)} - 1.$$

Proof. Denote $m := \lfloor H_k/t^\star(\epsilon) \rfloor$ and $t_i := \tau_k + i\; t^\star(\epsilon)$. Let $i^\star \in \{0\} \cup \mathbb{N}$ be such that $t_{i^\star} \leq T < t_{i^\star+1}$. Define the following martingale difference sequence $\{b_i\}_i$ w.r.t. the filtration $\{\mathcal{F}_{t_i}\}_i$,

$$b_i := \mathbb{1}_{\left\{s_{t_i}\in\xi\right\}} - \mathbb{E}\left[\mathbb{1}_{\left\{s_{t_i}\in\xi\right\}} \mid \mathcal{F}_{t_{i-1}}\right],\; i = 1,2,\ldots,i^\star.$$

Also, define

$$g_i := \mathbb{1}_{\{(i-1)t^\star(\epsilon)\leq H_k\}},\; i = 1,2,\ldots,i^\star,$$

and note that it is $\{\mathcal{F}_{t_i}\}_i$-predictable sequence. It can be shown that b_i's are conditionally $\frac{1}{2}$ sub-Gaussian, i.e., $\mathbb{E}[\exp(\beta\, b_i) \mid \mathcal{F}_{t_{i-1}}] \leq \exp(\beta^2/8)$ [Raginsky et al., 2013]. Also, note that $\{g_i\}_i$ is a $\{0,1\}$-valued, $\{\mathcal{F}_{t_i}\}$-predictable stochastic process. Hence, we can use Corollary J.4 and obtain,

$$\mathbb{P}\left(\sum_{i=1}^{m+1} \mathbb{1}_{\left\{s_{t_i}\in\xi\right\}} \leq \sum_{i=1}^{m+1} \mathbb{E}\left[\mathbb{1}_{\left\{s_{t_i}\in\xi\right\}} \mid \mathcal{F}_{t_{i-1}}\right] - \sqrt{\frac{m+2}{2}\log\left(\frac{3(m+2)}{\delta}\right)}\right) \leq \frac{\delta}{3}. \tag{60}$$

From (58), (59) we have that

$$\mathbb{E}\left[\mathbb{1}_{\left\{s_{t_{i-1}}\in\xi\right\}} \mid \mathcal{F}_{t_{i-1}}\right] \geq \frac{1}{2}\mu^{(\infty)}_{\phi,p}(\xi). \tag{61}$$

Also, observe that $m+1 > \frac{H_k}{t^\star(\epsilon)}$ and $m \leq \frac{H_k}{t^\star(\epsilon)}$. Since under ZoRL algorithm we have $H_k \geq 2t^\star(\epsilon)$, we get $m+2 \leq 2m$. Upon using (61) and $m+2 \leq 2m$ in (60), we obtain,

$$\mathbb{P}\left(\sum_{i=1}^{m} \mathbb{1}_{\left\{s_{t_i}\in\xi\right\}} \leq \frac{H_k\, \mu^{(\infty)}_{\phi,p}(\xi)}{2t^\star(\epsilon)} - \sqrt{\frac{H_k}{t^\star(\epsilon)}\log\left(\frac{6H_k}{t^\star(\epsilon)\delta}\right)} - 1\right) \leq \frac{\delta}{3}.$$

The claim then follows since $H_k \leq T$, and $\sum_{i=1}^{m} \mathbb{1}_{\left\{s_{t_i}\in\xi\right\}} \leq n_k(\zeta)$. □

Corollary D.2. *Fix an $\epsilon > 0$. Consider the triplet (k,ξ,ζ) such that $k \in \{0\}\cup\mathbb{N}$, $\xi \in \mathcal{Q}_{\tau_k}$, diam $(\xi) \geq \epsilon$, $\mu^{(\infty)}_{\phi,p}(\xi) \geq (\epsilon/3)^{d_{\mathcal{S}}+1}$, $\zeta \in \mathcal{P}_{\tau_k}$, and for every $s \in \xi$, $(s,\phi_k(s)) \in \zeta$. Define the event,*

$$\mathcal{G}_{2,\epsilon} := \left\{ n_k(\zeta) \geq \frac{H_k\, \mu^{(\infty)}_{\phi,p}(\xi)}{2t^\star(\epsilon)} - \sqrt{\frac{H_k}{t^\star(\epsilon)}\log\left(\frac{12T^2 d^{\frac{d}{2}}}{t^\star(\epsilon)\epsilon^d\delta}\right)} - 1,\; \forall (k,\xi,\zeta) \text{ that satisfies the above conditions.} \right\}, \tag{62}$$

where $t^\star(\epsilon) = \left\lceil \log_{\frac{1}{\alpha}}\left(C\left(\frac{3}{\epsilon}\right)^{d_{\mathcal{S}}+1}\right)\right\rceil$. We have, $\mathbb{P}(\mathcal{G}_{2,\epsilon}) \geq 1-\frac{\delta}{3}$.

Proof. Since k denotes the episode number, it can not exceed T. By definition of $\mathcal{P}_{\tau_k}$ and $\mathcal{Q}_{\tau_k}$, diam $(\zeta) \geq$ diam (ξ). Also, the number of cells that have a diameter greater than ϵ is less than $(\sqrt{d}/\epsilon)^d$. So, the total number of possible combinations of (k,ξ,ζ) that satisfies the given condition is at most $T(\sqrt{d}/\epsilon)^d$. The proof then follows from Lemma D.1 by taking a union bound over all (k,ξ,ζ) and by the fact that $H_k \leq T$. □

E REGRET ANALYSIS

Regret decomposition: Recall the regret (1) decomposition of ZoRL,

$$\begin{aligned}\mathcal{R}(T;\texttt{ZoRL}) &= TJ^\star_{\mathcal{M}} - \sum_{k=1}^{K(T)}\sum_{t=\tau_k}^{\tau_{k+1}-1} r(s_t,a_t)\\ &= \underbrace{\sum_{k=1}^{K(T)} H_k\left(J^\star_{\mathcal{M}} - J_{\mathcal{M}}(\phi_k)\right)}_{(a)} + \underbrace{\sum_{k=1}^{K(T)}\left(H_k\, J_{\mathcal{M}}(\phi_k) - \sum_{t=\tau_k}^{\tau_{k+1}-1} r(s_t,\phi_k(s_t))\right)}_{(b)}. \end{aligned} \tag{63}$$

The term (a) captures the regret arising due to the gap between the optimal value of the average reward and the average reward of the policies $\{\phi_k\}$ that are actually played in different episodes, while (b) captures the sub-optimality arising since the distribution of the induced Markov chain does not reach the stationary distribution in finite time. (a) and (b) are bounded separately.

Bounding (a): This term can be further decomposed into the sum of the regrets arising due to playing policies from the sets $\Phi^{(2^{-i})}$, for $i = 1, 2, \ldots, \lceil \log(1/\epsilon)\rceil$, and the regret arising from playing all ϵ-optimal policies. To bound the regret arising due to policies from $\Phi^{(2^{-i})}$, we count the number of timesteps in which policies from $\Phi^{(2^{-i})}$ are played, and then multiply it by 2^{-i+1}. We then add these regret terms from $i = 1$ to $\lceil \log(1/\epsilon)\rceil$. Note that the cumulative regret arising from playing the set of ϵ-optimal policies is upper-bounded by ϵT. Recall that at the beginning of the k-th episode, ZoRL solves $\mathcal{M}^+_{\tau_k}$ with the accuracy parameter set equal to $\frac{1}{\sqrt{T}}$. This "loss of accuracy" as compared to the case where ZoRL could have solved $\mathcal{M}^+_{\tau_k}$ accurately at the beginning of every episode, leads to an additional term in the upper-bound of (a). From Lemma G.1, the difference between the two solutions is at most $\frac{1}{\sqrt{T}}$ for each episode, hence this term can be upper-bounded as $\sqrt{T}$. Hence, we bound (a) by firstly considering that ZoRL solves $\mathcal{M}^+_{\tau_k}$ for the optimal policy (with complete accuracy), and then add $\sqrt{T}$ to obtain the upper-bound of term (a).

The regret arising due to playing policies from the set $\Phi^{(2^{-i})}$ is bounded as follows. Lemma E.1 proves the existence of a key cell in every episode on the set $\mathcal{G}_1$. Its proof relies crucially on Lemma 4.1 and on the properties of the index of policies that are derived in Section B. Lemma E.2 gives a lower-bound of the number of plays of a key cell in any episode by ZoRL using Lemma E.1, Corollary D.2, and Lemma J.5. Next, Lemma E.3 establishes an upper-bound on the number of timesteps when policies from $\Phi^{(2^{-i})}$ are played. This upper-bound multiplied by 2^{-i+1}, is the regret arising from playing policies from $\Phi^{(2^{-i})}$. Next, we derive an important property of the policy $\phi \in \Phi_{SD}$ that is played in the k-th episode. This is used to upper-bound the number of plays of sub optimal policies.

Lemma E.1. *Consider a sample path from the set $\mathcal{G}_1$ (82). For each $k = 1, 2, \ldots$, there exists at least one $s \in \mathcal{S}$ (where s could vary with k, and here we are suppressing dependence upon k) such that*

$$\begin{aligned}&\textit{diam}\left(q^{-1}_{\tau_k}(s,\phi_k(s))\right) \geq \frac{1}{3C_{ub}}\max\left\{\textit{gap}\left(s,\phi_k(s)\right), C_{ub}\,\textit{diam}_{\tau_k}(\phi_k)\right\},\\ &\textit{and } \mu^{(\infty)}_{\phi_k,p}(\pi_{\mathcal{S}}(q^{-1}_{\tau_k}(s,\phi_k(s)))) \geq (\textit{diam}_{\tau_k}(\phi_k)/3)^{d_{\mathcal{S}}+1}.\end{aligned}$$

Such a $q^{-1}_{\tau_k}(s,\phi_k(s))$ is called a key cell for the k-th episode.

Proof. Let us fix $k \in \mathbb{N}$ and a policy $\phi \in \Phi_{\tau_k}$. Let $\bar{\phi}$ be the unique continuous extension of ϕ as defined in (56). We will first show that if

$$\text{diam}_{\tau_k}(\bar{\phi}) \leq \Delta(\bar{\phi})/C_{ub}, \tag{64}$$

then $\bar{\phi}$ will not be played from episode k onwards. From Lemma B.1 we have that on the set $\mathcal{G}_1$, $J^\star_{\mathcal{M}^+_{\tau_k}} = J_{\mathcal{M}^+_{\tau_k}}(\tilde{\phi}_k) \geq J^\star_{\mathcal{M}}$. Hence, if $J_{\mathcal{M}^+_{\tau_k}}(\phi) < J^\star_{\mathcal{M}}$, then the algorithm will not play $\bar{\phi}$. From Lemma B.3 we have that on the set $\mathcal{G}_1$, $J_{\mathcal{M}^+_{\tau_k}}(\phi) \leq J_{\mathcal{M}}(\bar{\phi}) + C_{ub}\,\text{diam}_{\tau_k}(\bar{\phi})$. Thus, on $\mathcal{G}_1$, $\bar{\phi}$ will never be played from the k-th episode onwards if

$$J_{\mathcal{M}}(\bar{\phi}) + C_{ub}\,\text{diam}_{\tau_k}(\bar{\phi}) \leq J^\star_{\mathcal{M}},$$

or, if $\text{diam}_{\tau_k}(\bar{\phi}) \leq \Delta(\bar{\phi})/C_{ub}$. In other words, on the set $\mathcal{G}_1$,

$$\text{diam}_{\tau_k}(\phi_k) > \Delta(\phi_k)/C_{ub}. \tag{65}$$

We will prove the result by contradiction. Let us assume that for all $s \in \mathcal{S}$ that satisfy $\mu^{(\infty)}_{\phi_k,p}(\pi_{\mathcal{S}}(q^{-1}_{\tau_k}(s,\phi_k(s)))) \geq (\text{diam}_{\tau_k}(\phi_k)/3)^{d_{\mathcal{S}}+1}$, the following is true:

$$\text{diam}\left(q^{-1}_{\tau_k}(s,\phi_k(s))\right) \leq \frac{1}{3C_{ub}} \max\left\{\text{gap}\left(s,\phi_k(s)\right), C_{ub}\text{diam}_{\tau_k}(\phi_k)\right\}. \tag{66}$$

Define the following sets of $\mathcal{S}$-cells:

$$\begin{aligned}
\mathcal{Q}^{(1)} &:= \{\xi \in \mathcal{Q}_{\tau_k} \mid \mu^{(\infty)}_{\phi_k,p}(\xi) < (\text{diam}_{\tau_k}(\phi_k)/3)^{d_{\mathcal{S}}+1},\ \text{diam}\left(q^{-1}_{\tau_k}(q(\xi),\phi_k(q(\xi)))\right) \geq \text{diam}_{\tau_k}(\phi_k)/3\},\\
\mathcal{Q}^{(2)} &:= \{\xi \in \mathcal{Q}_{\tau_k} \mid \text{diam}\left(q^{-1}_{\tau_k}(q(\xi),\phi_k(q(\xi)))\right) < \text{diam}_{\tau_k}(\phi_k)/3\},\\
\mathcal{Q}^{(3)} &:= \{\xi \in \mathcal{Q}_{\tau_k} \mid \mu^{(\infty)}_{\phi_k,p}(\xi) \geq (\text{diam}_{\tau_k}(\phi_k)/3)^{d_{\mathcal{S}}+1},\ \text{diam}\left(q^{-1}_{\tau_k}(q(\xi),\phi_k(q(\xi)))\right) \geq \text{diam}_{\tau_k}(\phi_k)/3\}.
\end{aligned}$$

We observe that $\mathcal{Q}_{\tau_k}$ is partitioned by $\mathcal{Q}^{(1)}$, $\mathcal{Q}^{(3)}$ and $\mathcal{Q}^{(3)}$. Note that $\left|\mathcal{Q}^{(1)}\right| \leq (\text{diam}_{\tau_k}(\phi_k)/3)^{-d_{\mathcal{S}}}$. Also, note that by the necessary condition for ϕ_k to be played and by our assumption, for every $\xi \in \mathcal{Q}^{(3)}$, $\frac{1}{3}\text{diam}_{\tau_k}(\phi_k) \leq \text{diam}\left(q^{-1}_{\tau_k}(q(\xi),\phi_k(q(\xi)))\right) \leq \frac{1}{3C_{ub}} \min_{s\in\zeta}\{\text{gap}\left(s,\phi_k(s)\right)\}$. Then,

$$\begin{aligned}
\text{diam}_{\tau_k}(\phi_k) &= \int_{\mathcal{S}} \text{diam}\left(q^{-1}_{\tau_k}(s,\phi_k(s))\right) \mu^{(\infty)}_{\phi_k,p}(s)\, ds\\
&= \sum_{\xi\in\mathcal{Q}_{\tau_k}} \text{diam}\left(q^{-1}_{\tau_k}(q(\xi),\phi_k(q(\xi)))\right) \mu^{(\infty)}_{\phi_k,p}(\xi)\\
&= \sum_{\xi\in\mathcal{Q}^{(1)}} \text{diam}\left(q^{-1}_{\tau_k}(q(\xi),\phi_k(q(\xi)))\right) \mu^{(\infty)}_{\phi_k,p}(\xi) + \sum_{\xi\in\mathcal{Q}^{(2)}} \text{diam}\left(q^{-1}_{\tau_k}(q(\xi),\phi_k(q(\xi)))\right) \mu^{(\infty)}_{\phi_k,p}(\xi)\\
&\quad + \sum_{\xi\in\mathcal{Q}^{(3)}} \text{diam}\left(q^{-1}_{\tau_k}(q(\xi),\phi_k(q(\xi)))\right) \mu^{(\infty)}_{\phi_k,p}(\xi)\\
&\leq \frac{\text{diam}_{\tau_k}(\phi_k)}{3} + \frac{\text{diam}_{\tau_k}(\phi_k)}{3} + \frac{1}{3C_{ub}} \int_{\mathcal{S}} \text{gap}\left(s,\phi_k(s)\right) \mu^{(\infty)}_{\phi_k,p}(s)\, ds\\
&= \frac{\text{diam}_{\tau_k}(\phi_k)}{3} + \frac{\text{diam}_{\tau_k}(\phi_k)}{3} + \frac{\Delta(\phi_k)}{3\, C_{ub}}\\
&< \text{diam}_{\tau_k}(\phi_k),
\end{aligned}$$

which yields us a contradiction. Hence, we conclude that our assumption (66) was wrong. This concludes the proof. □

Define,

$$\epsilon(T) := T^{-\frac{1}{2d_{\mathcal{S}}+d_z+3}},\ \ \tilde{\epsilon}(T) := T^{-\frac{1}{2d_{\mathcal{S}}+d+3}} \tag{67}$$

Note that $\epsilon(T) \geq \tilde{\epsilon}(T)$ since $d_z \leq d$. Also, note that $t^\star(\epsilon(T)) \leq t^\star(\tilde{\epsilon}(T))$, where $t^\star(\cdot)$ is defined ins (59).

Choosing C_H: We choose the constant associated with the episode duration (17) of ZoRL as,

$$C_H \geq 16\, t^\star(\tilde{\epsilon}(T))\ \left(\frac{3(1+C_{ub})}{1-\gamma}\right)^{2(d_{\mathcal{S}}+1)} \frac{\log\left(\frac{12T^2 d^{\frac{d}{2}}}{t^\star(\epsilon(T))\tilde{\epsilon}(T)^d\delta}\right)+1}{\log(T/\delta)}. \tag{68}$$

Lemma E.2. *Pick a sample path from the set $\mathcal{G}_1 \cap \mathcal{G}_{2,\epsilon}$, where $\mathcal{G}_1$ and $\mathcal{G}_{2,\epsilon}$ are as in* (82) *and* (62)*, respectively. Let ζ be a key cell in episode k (such key cells have been shown to exist in Lemma E.1), i.e., for some $\xi \subseteq \pi_{\mathcal{S}}(\zeta)$ such that $\xi \in \mathcal{Q}_{\tau_k}$, and for some $s \in \xi$, the following holds,*

$$\begin{aligned}
&\textit{diam}\,(\zeta) > \frac{1}{3C_{ub}} \max\left\{\textit{gap}\left(s,\phi_k(s)\right), C_{ub}\ \textit{diam}_{\tau_k}(\phi_k)\right\},\ \textit{and},\\
&\mu^{(\infty)}_{\phi_k,p}(\xi) \geq (\textit{diam}_{\tau_k}(\phi_k)/3)^{d_{\mathcal{S}}+1}.
\end{aligned}$$

Then, if $\Delta(\phi_k) \geq \epsilon(T) C_{ub}$, then the number of visits to ζ during the k-th episode can be lower-bounded as follows,

$$n_k(\zeta) \geq \frac{4t^\star(\tilde{\epsilon}(T))}{t^\star(\epsilon(T))} \left(\log \left(\frac{12T^2 d^{\frac{d}{2}}}{t^\star(\epsilon(T))\tilde{\epsilon}(T)^d \delta} \right) + 1 \right) \textit{diam}\,(\zeta)^{-(d_{\mathcal{S}}+1)}. \tag{69}$$

Proof. Recall that on $\mathcal{G}_1$ we have $\text{diam}_{\tau_k}(\phi_k) \geq \frac{\Delta(\phi_k)}{C_{ub}}$ (65). Hence, $\text{diam}_{\tau_k}(\phi_k) > \epsilon(T)$ and $\mu^{(\infty)}_{\phi_k,p}(\xi) \geq (\epsilon(T)/3)^{d_{\mathcal{S}}+1}$. So, upon using Corollary D.2 we obtain,

$$n_k(\zeta) \geq \frac{H_k\, \mu^{(\infty)}_{\phi_k,p}(\xi)}{2t^\star(\epsilon(T))} - \sqrt{\frac{H_k}{t^\star(\epsilon(T))} \log \left(\frac{8T^2 d^{\frac{d}{2}}}{t^\star(\epsilon(T))\epsilon(T)^d \delta} \right)} - 1.$$

Next, we note that the duration of the k-th episode H_k can be lower-bounded as follows,

$$\begin{aligned} H_k &\geq \frac{C_H (1-\gamma)^{2(d_{\mathcal{S}}+1)} \log(T/\delta)}{\widetilde{\text{diam}}_{\tau_k}(\phi_k)^{2(d_{\mathcal{S}}+1)}} \\ &\geq \frac{C_H (1-\gamma)^{2(d_{\mathcal{S}}+1)} \log(T/\delta)}{(3(1+C_{ub}))^{2(d_{\mathcal{S}}+1)}} \left(\frac{3}{\text{diam}_{\tau_k}(\phi_k)} \right)^{2(d_{\mathcal{S}}+1)} \\ &\geq \frac{16 t^\star(\epsilon(T))}{\mu^{(\infty)}_{\phi_k,p}(\xi)^2} \left(\log \left(\frac{8T^2 d^{\frac{d}{2}}}{t^\star(\epsilon(T))\epsilon(T)^d \delta} \right) + 1 \right), \end{aligned} \tag{70}$$

where the first inequality follows from the lower-bound of H_k (96), the second inequality follows since from Corollary C.3 we have $\widetilde{\text{diam}}_{\tau_k}(\phi_k) \leq (1 + C_{ub})\text{diam}_{\tau_k}(\phi_k)$. The third inequality follows from the fact that $\mu^{(\infty)}_{\phi_k,p}(\xi) \geq (\text{diam}_{\tau_k}(\phi_k)/3)^{d_{\mathcal{S}}+1}$. Lemma J.5 when combined with (70) yields

$$\begin{aligned} n_k(\zeta) &\geq \frac{H_k\, \mu^{(\infty)}_{\phi_k,p}(\xi)}{2t^\star(\epsilon(T))} - \sqrt{\frac{H_k}{t^\star(\epsilon(T))} \log \left(\frac{8T^2 d^{\frac{d}{2}}}{t^\star(\epsilon(T))\epsilon(T)^d \delta} \right)} - 1 \\ &\geq \frac{H_k\, \mu^{(\infty)}_{\phi_k,p}(\xi)}{4t^\star(\epsilon(T))}, \end{aligned}$$

or,

$$\begin{aligned} n_k(\zeta) &\geq \frac{C_H (1-\gamma)^{2(d_{\mathcal{S}}+1)} \log\left(\frac{T}{\delta}\right)}{4\, t^\star(\epsilon(T))}\, \widetilde{\text{diam}}_{\tau_k}(\phi_k)^{-2(d_{\mathcal{S}}+1)} \times (\text{diam}_{\tau_k}(\phi_k)/3)^{d_{\mathcal{S}}+1} \\ &\geq \frac{C_H \log\left(\frac{T}{\delta}\right)}{4\, t^\star(\epsilon(T))\, (3(1+C_{ub})^2)^{d_{\mathcal{S}}+1}}\, \text{diam}_{\tau_k}(\phi_k)^{-(d_{\mathcal{S}}+1)} \\ &\geq \frac{C_H \log\left(\frac{T}{\delta}\right)}{4\, t^\star(\epsilon(T))\, (3(1+C_{ub}))^{2(d_{\mathcal{S}}+1)}}\, \text{diam}\,(\zeta)^{-(d_{\mathcal{S}}+1)} \\ &\geq \frac{4t^\star(\tilde{\epsilon}(T))}{t^\star(\epsilon(T))} \left(\log \left(\frac{12T^2 d^{\frac{d}{2}}}{t^\star(\epsilon(T))\tilde{\epsilon}(T)^d \delta} \right) + 1 \right) \text{diam}\,(\zeta)^{-(d_{\mathcal{S}}+1)}, \end{aligned}$$

where the first inequality follows from the lower-bound of H_k (96) and from the fact that $\mu^{(\infty)}_{\phi_k,p}(\xi) \geq (\text{diam}_{\tau_k}(\phi_k)/3)^{d_{\mathcal{S}}+1}$. The second and the third inequality follow from the fact that $\widetilde{\text{diam}}_{\tau_k}(\phi_k) \leq (1 + C_{ub})\text{diam}_{\tau_k}(\phi_k)$, and $\text{diam}_{\tau_k}(\phi_k) < 3\,\text{diam}\,(\zeta)$, respectively. The fourth inequality follows from (68). This concludes the proof. □

Lemma E.3. *Consider the set of policies $\Phi^{(2^{-i})} = \{\phi \in \Phi_{SD} \mid \Delta(\phi) \in (2^{-i}, 2^{-i+1}]\}$, where $i \in \mathbb{N}$. On the set $\mathcal{G}_1$, `ZoRL` can play policies from the set $\Phi^{(2^{-i})}$ for a maximum of $\mathcal{O}(\log\left(\frac{T}{\delta}\right) 2^{i(2d_{\mathcal{S}}+d_z+3)})$ time steps.*

Proof. We prove this lemma in the following three steps: First, we derive the number of episodes in which a cell can serve as a key cell while policies from $\Phi^{(2^{-i})},\ i \in \mathbb{N}$ are being played. Secondly, we derive an upper-bound on the episode duration when policies from $\Phi^{(2^{-i})}$ are played. Thirdly, we multiply upper-bounds on the number of episodes with the upper-bound

on the duration of the episodes and then sum it over all possible key cells corresponding to policies in $\Phi^{(2^{-i})}$, and this yields the desired upperbound on cumulative plays from $\Phi^{(2^{-i})}$.

Before proceeding with proving these three properties, we begin with some preliminary results. Recall that for $\beta > 0$, the set $\mathcal{Z}_\beta \subseteq \mathcal{S} \times \mathcal{A}$ consists of those state-action pairs (s, a) for which gap $(s, a) \leq \beta$. Let us denote the smallest subset of $\mathcal{P}_t$ that covers $\mathcal{Z}_\beta$, as the active covering of $\mathcal{Z}_\beta$ at time t. From Lemma E.1, we obtain that if for all $j = 0, 1, \ldots, i$, the active covering of $\mathcal{Z}_{2^{-j}}$ at time τ_k does not contain a cell ζ that satisfies the following conditions,

1. $\text{diam}\,(\zeta) \geq \frac{\sqrt{d}}{3C_{ub}}\, 2^{-j}$, and
2. $\mu^{(\infty)}_{\phi,p}(\xi) \geq (\Delta(\phi)/3C_{ub})^{d_{\mathcal{S}}+1}$ for all ξ which satisfy $\xi \in \mathcal{Q}_{\tau_k}$ and $\xi \subseteq \pi_{\mathcal{S}}(\zeta)$,

then there is no cell that qualifies to be a key cell for a policy from the set $\Phi^{(2^{-i})}$. Thus, under the above condition, ZoRL will not play a policy from $\Phi^{(2^{-i})}$ k-th episode onwards. Let $\mathcal{Y}_j$ be the covering of $\mathcal{Z}_{2^{-j}}$ by cells of diameter $\frac{\sqrt{d}}{3C_{ub}}\, 2^{-j}$. We make the following observation: If every cell in $\mathcal{Y}_j$ for $j = 1, 2, \ldots, i$ is split, then no cell in the active covers of $\mathcal{Z}_{2^{-j}}$ for $j = 1, 2, \ldots, i$ can serve as the key cell while playing policies from $\Phi^{(2^{-i})}$. This is a sufficient condition for any policy from $\Phi^{(2^{-i})}$ to be not played by ZoRL.

Step 1: First, we bound the number of episodes when a cell $\zeta \in \mathcal{Y}_i$ or any of its ancestors has served as a key cell. From the cell activation rule (3.3), we have that ζ would be split when the number of visits to ζ exceeds $c_a 2^{d_{\mathcal{S}}+2} \log\left(\frac{T}{\delta}\right) \text{diam}\,(\zeta)^{-(d_{\mathcal{S}}+2)}$. In Lemma E.3, we derived the lower-bound on the number of visits to a key cell. Invoking that lower-bound, we obtain that ζ can be played in at most

$$\frac{c_a t^\star(\epsilon(T)) 2^{d_{\mathcal{S}}+2} \log\left(\frac{T}{\delta}\right)}{4t^\star(\tilde{\epsilon}(T)) \left(\log\left(\frac{12T^2 d^{\frac{d}{2}}}{t^\star(\epsilon(T))\tilde{\epsilon}(T)^d \delta}\right) + 1\right)} \text{diam}\,(\zeta)^{-1}$$

episode as a key cell when the corresponding episode plays a policy from $\Phi^{(2^{-i})}$. Replacing $\text{diam}\,(\zeta)$ with $\frac{\sqrt{d}}{3C_{ub}}\, 2^{-j}$, we obtain that ζ can be played in at most

$$\frac{3c_a t^\star(\epsilon(T)) C_{ub} 2^{d_{\mathcal{S}}+2} \log\left(\frac{T}{\delta}\right)}{4t^\star(\tilde{\epsilon}(T))\sqrt{d} \left(\log\left(\frac{12T^2 d^{\frac{d}{2}}}{t^\star(\epsilon(T))\tilde{\epsilon}(T)^d \delta}\right) + 1\right)}\, 2^j$$

episode as a key cell when the corresponding episode plays a policy from $\Phi^{(2^{-i})}$.

Step 2: Now, we produce an upper-bound on the length of the episodes while playing policies from $\Phi^{(2^{-i})}$. See that

$$\begin{aligned} H_k &\leq \frac{C_H (1+\gamma)^{2(d_{\mathcal{S}}+1)} \log\left(\frac{T}{\delta}\right)}{\widetilde{\text{diam}}_{\tau_k}(\phi_k)^{2(d_{\mathcal{S}}+1)}} \\ &\leq \frac{C_H (1+\gamma)^{2(d_{\mathcal{S}}+1)} \log\left(\frac{T}{\delta}\right)}{\text{diam}_{\tau_k}(\phi_k)^{2(d_{\mathcal{S}}+1)}} \\ &\leq \frac{C_H ((1+\gamma)C_{ub})^{2(d_{\mathcal{S}}+1)} \log\left(\frac{T}{\delta}\right)}{2^{-i2(d_{\mathcal{S}}+1)}}, \end{aligned}$$

where the first inequality follows from the upper-bound on H_k (96), the second inequality follows from Corollary C.1, and the third inequality follows from the definition of $\Phi^{(2^{-i})}$.

Step 3: First, we note that the cardinality of $\mathcal{Y}_j$ is at most $c_z 2^{j d_z}$ for every $j \in \mathbb{N}$, where the scaling constant of the zooming dimension,

$$c_s := \frac{\sqrt{d}}{3C_{ub}}. \tag{71}$$

This follows from the definition of the zooming dimension (3). Multiplying the bounds from step 1 and step 2, we obtain an upper-bound on the number of plays of a cell $\zeta \in \mathcal{Y}_j$ as a key cell while playing policies from $\Phi^{(2^{-i})}$. Summing this

upper-bound for all cells in $\mathcal{Y}_j$ and then summing those terms over $j = 1, 2, \ldots, i$, we obtain that the total number of time steps in which policies from $\Phi^{(2^{-i})}$ is played, can be bounded above by

$$\begin{aligned}
&\sum_{j=1}^{i} \sum_{\zeta \in \mathcal{Y}_j} \left(\frac{3c_a t^\star(\epsilon(T)) C_{ub} 2^{d_\mathcal{S}+2} \log\left(\frac{T}{\delta}\right)}{4t^\star(\tilde{\epsilon}(T))\sqrt{d}\left(\log\left(\frac{12T^2 d^{\frac{d}{2}}}{t^\star(\epsilon(T))\tilde{\epsilon}(T)^d \delta}\right)+1\right)} 2^j \right) \times \left(\frac{C_H((1+\gamma)C_{ub})^{2(d_\mathcal{S}+1)} \log\left(\frac{T}{\delta}\right)}{2^{-i2(d_\mathcal{S}+1)}} \right) \\
&= \frac{3c_a c_z t^\star(\epsilon(T)) C_H C_{ub}^{2d_\mathcal{S}+3} (1+\gamma)^{2(d_\mathcal{S}+1)} 2^{d_\mathcal{S}+2} \left(\log\left(\frac{T}{\delta}\right)\right)^2}{4t^\star(\tilde{\epsilon}(T))\sqrt{d}\left(\log\left(\frac{12T^2 d^{\frac{d}{2}}}{t^\star(\epsilon(T))\tilde{\epsilon}(T)^d \delta}\right)+1\right)} 2^{i2(d_\mathcal{S}+1)} \sum_{j=0}^{i} 2^{j(d_z+1)} \\
&\leq \frac{3c_a c_z t^\star(\epsilon(T)) C_H C_{ub}^{2d_\mathcal{S}+3} (1+\gamma)^{2(d_\mathcal{S}+1)} 2^{d_\mathcal{S}+1} \left(\log\left(\frac{T}{\delta}\right)\right)^2}{t^\star(\tilde{\epsilon}(T))\sqrt{d}\left(\log\left(\frac{12T^2 d^{\frac{d}{2}}}{t^\star(\epsilon(T))\tilde{\epsilon}(T)^d \delta}\right)+1\right)} 2^{i(2d_\mathcal{S}+d_z+3)}.
\end{aligned}$$

This concludes the proof. □

Let us denote

$$C' := \frac{3c_a c_z t^\star(\epsilon(T)) C_H C_{ub}^{2d_\mathcal{S}+3} (1+\gamma)^{2(d_\mathcal{S}+1)} 2^{d_\mathcal{S}+1} \left(\log\left(\frac{T}{\delta}\right)\right)^2}{t^\star(\tilde{\epsilon}(T))\sqrt{d}\left(\log\left(\frac{12T^2 d^{\frac{d}{2}}}{t^\star(\epsilon(T))\tilde{\epsilon}(T)^d \delta}\right)+1\right)}. \tag{72}$$

As has been discussed earlier at the beginning of this section, we derive an upper-bound on (a) of (63) by summing the three terms: the regret due to playing policies from the set $\Phi^{(2^{-i})},\ i = 1, 2, \ldots, \lceil \log 1/\epsilon(T) \rceil$, the regret due to playing other policies, and the suboptimality that arises due to the inaccuracy in the solution of the extended MDPs at the beginning of every episode, which can be bounded by $\sqrt{T}$. The first term is bounded using the bound obtained on the number of plays of policies from $\Phi^{(i)}$ in Lemma E.3. The regret arising from playing policies that are not in $\cup_{i=1}^{\lceil \log 1/\epsilon \rceil} \Phi^{(2^{-i})}$ is at most $\epsilon(T)T$. Hence,

$$\begin{aligned}
\sum_{k=1}^{K(T)} H_k(J^\star_\mathcal{M} - J_\mathcal{M}(\phi_k) &\leq C' \sum_{i=1}^{i^\star} 2^{i(2d_\mathcal{S}+d_z+3)} \times 2^{-i+1} + \epsilon(T)T + \sqrt{T} \\
&\leq 2C'\, 2^{i^\star(2d_\mathcal{S}+d_z+2)} + T^{\frac{2d_\mathcal{S}+d_z+2}{2d_\mathcal{S}+d_z+3}} + \sqrt{T} \\
&\leq (2C'+1)\, T^{\frac{2d_\mathcal{S}+d_z+2}{2d_\mathcal{S}+d_z+3}} + \sqrt{T},
\end{aligned} \tag{73}$$

where the second step follows from Lemma E.3.

Bounding (b): We now provide an upper-bound on the term (b) of (63). This proof relies on the uniform ergodicity property (Assumption 2.2) of the underlying MDP $\mathcal{M}$ and a trick that converts Markovian noise to martingale noise using the Poisson equation (19) [Metivier and Priouret, 1984].

Proposition E.4. *Define*

$$\mathcal{G}_3 := \{\omega :\ (75) \textit{ holds } \}, \tag{74}$$

$$\sum_{k=1}^{K(T)} \sum_{t=\tau_k}^{\tau_{k+1}-1} J_\mathcal{M}(\phi_k) - r(s_t, \phi_k(s_t)) \leq \frac{m^\star}{1-\alpha}\sqrt{\frac{T}{2}\log\left(\frac{3}{\delta}\right)} + \frac{m^\star}{1-\alpha}(1+K(T)), \tag{75}$$

where $K(T)$ denotes the total number of episodes until time T, and $m^\star = \left\lceil \log_{\frac{1}{\alpha}}(C) \right\rceil + 1$. Then, we have,

$$\mathbb{P}(\mathcal{G}_3) \geq 1 - \frac{\delta}{3},\ \delta \in (0,1). \tag{76}$$

Proof. Let us denote the episode index at time t by $k(t)$. We begin by converting the Markovian noise to a martingale difference sequence, i.e.,

$$\begin{aligned}
&\sum_{t=0}^{T-1} J_{\mathcal{M}}(\phi_{k(t)}) - r(s_t, \phi_{k(t)}(s_t)) \\
&= \sum_{t=0}^{T-1} \int_{\mathcal{S}} h_{\mathcal{M}}^{\phi_{k(t)}}(s) p(s_t, \phi_{k(t)}(s_t), ds) - h_{\mathcal{M}}^{\phi_{k(t)}}(s_t) \\
&= \sum_{t=1}^{T-1} \int_{\mathcal{S}} h_{\mathcal{M}}^{\phi_{k(t)}}(s) p(s_{t-1}, \phi_{k(t-1)}(s_{t-1}), ds) - h_{\mathcal{M}}^{\phi_{k(t)}}(s_t) \\
&\quad + \sum_{t=1}^{T-1} \int_{\mathcal{S}} h_{\mathcal{M}}^{\phi_{k(t)}}(s) p(s_t, \phi_{k(t)}(s_t), ds) - \int_{\mathcal{S}} h_{\mathcal{M}}^{\phi_{k(t)}}(s) p(s_{t-1}, \phi_{k(t-1)}(s_{t-1}), ds) \\
&\quad + \int_{\mathcal{S}} h_{\mathcal{M}}^{\phi_1}(s) p(s_0, \phi_1(s_0), ds) - h_{\mathcal{M}}^{\phi_1}(s_0) \\
&= \sum_{t=1}^{T-1} \int_{\mathcal{S}} h_{\mathcal{M}}^{\phi_{k(t)}}(s) p(s_{t-1}, \phi_{k(t-1)}(s_{t-1}), ds) - h_{\mathcal{M}}^{\phi_{k(t)}}(s_t) \\
&\quad + \sum_{t=1}^{T-1} \int_{\mathcal{S}} \left(h_{\mathcal{M}}^{\phi_{k(t)}}(s) - h_{\mathcal{M}}^{\phi_{k(t-1)}}(s) \right) p(s_{t-1}, \phi_{k(t-1)}(s_{t-1}), ds) \\
&\quad + \int_{\mathcal{S}} h_{\mathcal{M}}^{\phi_{k(T-1)}}(s) p(s_{T-1}, \phi_{k(T-1)}(s_{T-1}), ds) - h_{\mathcal{M}}^{\phi_1}(s_0). \qquad (77)
\end{aligned}$$

Now consider the first summation term in the r.h.s. of (77). Denote $m_t = \int_{\mathcal{S}} h_{\mathcal{M}}^{\phi_{k(t)}}(s) p(s_{t-1}, \phi_{k(t-1)}(s_{t-1}), ds) - h_{\mathcal{M}}^{\phi_{k(t)}}(s_t)$. Noting that ϕ_k is $\mathcal{F}_{\tau_k - 1}$-measurable, we obtain the following:

$$\begin{aligned}
\mathbb{E}\left[m_t \mid \mathcal{F}_{t-1} \right] &= \mathbb{E}\left[\int_{\mathcal{S}} h_{\mathcal{M}}^{\phi_{k(t)}}(s) p(s_{t-1}, \phi_{k(t-1)}(s_{t-1}), ds) - h_{\mathcal{M}}^{\phi_{k(t)}}(s_t) \mid \mathcal{F}_{t-1} \right] \\
&= \int_{\mathcal{S}} h_{\mathcal{M}}^{\phi_{k(t)}}(s) p(s_{t-1}, \phi_{k(t-1)}(s_{t-1}), ds) - \int_{\mathcal{S}} h_{\mathcal{M}}^{\phi_{k(t)}}(s) p(s_{t-1}, \phi_{k(t-1)}(s_{t-1}), ds) \\
&= 0.
\end{aligned}$$

Hence, $\{m_t\}$ is a martingale difference sequence. Also, from the bound on the span of $h_{\mathcal{M}}^{\phi}$ that was derived in Lemma A.3, we have that $m_t \in \left[-\frac{m^\star}{1-\alpha}, \frac{m^\star}{1-\alpha} \right]$. An application of Azuma-Hoeffding inequality (Lemma J.1), yields the following: for each $\delta \in (0,1)$, with probability at least $1 - \frac{\delta}{3}$ we have,

$$\sum_{t=1}^{T-1} \int_{\mathcal{S}} h_{\mathcal{M}}^{\phi_{k(t)}}(s) p(s_{t-1}, \phi_{k(t-1)}(s_{t-1}), ds) - h_{\mathcal{M}}^{\phi_{k(t)}}(s_t) \leq \frac{m^\star}{1-\alpha} \sqrt{\frac{T}{2} \log\left(\frac{3}{\delta} \right)}. \qquad (78)$$

Now, consider the second summation term in the r.h.s. of (77). The t-th element in this summation can assume a non-zero value only when a new episode starts at time t. Hence, upon using Lemma A.3, we conclude that this summation can be upper-bounded as

$$\sum_{t=1}^{T-1} \int_{\mathcal{S}} \left(h_{\mathcal{M}}^{\phi_{k(t)}}(s) - h_{\mathcal{M}}^{\phi_{k(t-1)}}(s) \right) p(s_{t-1}, \phi_{k(t-1)}(s_{t-1}), ds) \leq \frac{m^\star}{1-\alpha} K(T), \qquad (79)$$

where $K(T)$ denotes the number of episodes that have been started until time T by the learning algorithm. Again by using Lemma A.3, the third term can be bounded as,

$$\int_{\mathcal{S}} h_{\mathcal{M}}^{\phi_{k(T-1)}}(s) p(s_{T-1}, \phi_{k(T-1)}(s_{T-1}), ds) - h_{\mathcal{M}}^{\phi_1}(s_0) \leq \frac{m^\star}{1-\alpha}. \qquad (80)$$

Putting all the individual bounds from (78), (79) and (80) together, we have that for any $\delta \in (0,1)$ with probability at least $1-\delta$,

$$\sum_{t=1}^{T-1} J_{\mathcal{M}}(\phi_{k(t)}) - r(s_t, \phi_{k(t)}(s_t)) \leq \frac{m^\star}{1-\alpha}\sqrt{\frac{T}{2}\log\left(\frac{3}{\delta}\right)} + \frac{m^\star}{1-\alpha}(1+K(T)). \tag{81}$$

This concludes the proof. □

Upon combining the upper-bounds on all the terms of the regret decomposition, we obtain the upper-bound on the regret. This is done in the next section.

E.1 PROOF OF THEOREM 4.4

Proof. We first derive an upper-bound on $K(T)$, which is the total number of episodes. The number of episodes of length greater than $T^{\frac{2d_{\mathcal{S}}+2}{2d_{\mathcal{S}}+d_z+3}}$ is trivially bounded above by $T^{\frac{d_z+1}{2d_{\mathcal{S}}+d_z+3}}$. Now let us bound the number of episodes of length less than $T^{\frac{2d_{\mathcal{S}}+2}{2d_{\mathcal{S}}+d_z+3}}$. If the length of the k-th episode is less than $T^{\frac{2d_{\mathcal{S}}+2}{2d_{\mathcal{S}}+d_z+3}}$, then from the rule of setting episode duration (17), we have

$$\frac{C_H \log\left(\frac{T}{\delta}\right)}{\widetilde{\text{diam}}_{\tau_k}(\phi_k)^{2(d_{\mathcal{S}}+1)}} \leq T^{\frac{2d_{\mathcal{S}}+2}{2d_{\mathcal{S}}+d_z+3}},$$

or

$$\widetilde{\text{diam}}_{\tau_k}(\phi_k) \geq \left(C_H \log\left(\frac{T}{\delta}\right)\right)^{\frac{1}{2(d_{\mathcal{S}}+1)}} T^{-\frac{1}{2d_{\mathcal{S}}+d_z+3}}.$$

From Corollary C.1 and Corollary C.3, we obtain that

$$\frac{1}{3(C_{ub}+1)}\widetilde{\text{diam}}_{\tau_k}(\phi_k) \leq \frac{1}{3}\text{diam}_{\tau_k}(\phi_k).$$

Also, from the condition of a cell ζ to be a key cell in the k-th episode, we have that

$$\text{diam}\,(\zeta) \geq \frac{1}{3}\text{diam}_{\tau_k}(\phi_k).$$

Combining the above three relations, we obtain that if the length of the k-th episode is less than $T^{\frac{2d_{\mathcal{S}}+2}{2d_{\mathcal{S}}+d_z+3}}$, then the diameter of the corresponding key cell is greater than

$$\frac{\left(C_H \log\left(\frac{T}{\delta}\right)\right)^{\frac{1}{2(d_{\mathcal{S}}+1)}}}{3(C_{ub}+1)} T^{-\frac{1}{2d_{\mathcal{S}}+d_z+3}}.$$

From the definition of the zooming dimension (3), it follows that there can at most be $\mathcal{O}\left(T^{\frac{d_z}{2d_{\mathcal{S}}+d_z+3}}\right)$ such key cells activated by ZoRL, and each key cell of level ℓ becomes deactivated when it has been played in $\mathcal{O}(2^\ell)$ episodes. Hence there can be at most $\mathcal{O}\left(T^{\frac{d_z+1}{2d_{\mathcal{S}}+d_z+3}}\right)$ episodes of length less than $T^{\frac{2d_{\mathcal{S}}+2}{2d_{\mathcal{S}}+d_z+3}}$. Hence,

$$K(T) \leq C_K T^{\frac{d_z+1}{2d_{\mathcal{S}}+d_z+3}},$$

where C_K is a constant.

We now add all the upper-bounds of various regret components from (73) and (81), and use the upper-bound on $K(T)$ derived above. This yields,

$$\begin{aligned}
\mathcal{R}(T; \texttt{ZoRL}) &\leq (2C'+1)\, T^{\frac{2d_{\mathcal{S}}+d_z+2}{2d_{\mathcal{S}}+d_z+3}} + \sqrt{T} + \frac{m^\star}{1-\alpha}\sqrt{\frac{T}{2}\log\left(\frac{3}{\delta}\right)} + \frac{m^\star}{1-\alpha}(1+K(T)) \\
&\leq (2C'+1)\, T^{\frac{2d_{\mathcal{S}}+d_z+2}{2d_{\mathcal{S}}+d_z+3}} + \left(1 + \frac{m^\star}{1-\alpha}\sqrt{\frac{1}{2}\log\left(\frac{3}{\delta}\right)}\right)\sqrt{T} + \frac{m^\star}{1-\alpha}\left(1 + C_K T^{\frac{d_z+1}{2d_{\mathcal{S}}+d_z+3}}\right) \\
&= \tilde{\mathcal{O}}\left(T^{\frac{2d_{\mathcal{S}}+d_z+2}{2d_{\mathcal{S}}+d_z+3}}\right).
\end{aligned}$$

Note that $\mathbb{P}(\mathcal{G}_1 \cap \mathcal{G}_{2,\epsilon} \cap \mathcal{G}_3) \geq 1-\delta$. Thus, we have the desired regret upper-bound with probability at least $1-\delta$. $\square$

F CONCENTRATION INEQUALITY

In this section, we will show that the discretized MDP kernel belongs to a confidence ball around its estimate. First, let us introduce some notations. Let $\tilde{\mathcal{Z}} \subseteq \mathcal{S} \times \mathcal{A}$, and $\tilde{\mathcal{Q}}$ be a partition of $\mathcal{S}$ that is made of $\mathcal{S}$-cells. Let $\tilde{\mathcal{S}}$ be the set of representative points of the $\mathcal{S}$-cells in $\tilde{\mathcal{Q}}$. Recall the discretization of p given $\tilde{\mathcal{Z}}$ and $\tilde{\mathcal{S}}$, $\wp_{\tilde{\mathcal{Z}}\to\tilde{\mathcal{S}},p}$ (8). Denote the continuous extension of $\wp_{\tilde{\mathcal{Z}}\to\tilde{\mathcal{S}},p}$ by $\bar{\wp}_{\tilde{\mathcal{Z}}\to\tilde{\mathcal{S}},p}$, i.e.,

$$\bar{\wp}_{\tilde{\mathcal{Z}}\to\tilde{\mathcal{S}},p}(z,B) := \sum_{\xi\in\mathcal{Q}} \frac{\lambda(B\cap\xi)}{\lambda(\xi)} \wp_{\tilde{\mathcal{Z}}\to\tilde{\mathcal{S}},p}(z,q(\xi)),$$

for every $z \in \mathcal{Z}, B \in \mathcal{B}_{\mathcal{S}}$. Define the set,

$$\mathcal{G}_1 := \cap_{t=0}^{T-1} \left\{ \left\| \wp_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}_t,p}(z',\cdot) - \wp_{\mathcal{Z}_t\to\mathcal{S}_t,\hat{p}_t}(z,\cdot) \right\|_1 \right\} \leq \eta_t(\zeta) \text{ for every } z \in \mathcal{Z}_t, z' \in q^{-1}(z). \tag{82}$$

We show that $\mathcal{G}_1$ holds with a high probability.

Lemma F.1. $\mathbb{P}(\mathcal{G}_1) \geq 1-\frac{\delta}{3}$, *where* $\mathcal{G}_1$ *is as in* (82).

Proof. Fix t, and consider a point $z \in \mathcal{Z}_t$. Within this proof, we denote $q_t^{-1}(z)$ by ζ. Let ζ be of level ℓ, and note that ζ is active at time t. Let z' be an arbitrary point in ζ. We want to get a high probability bound on $\left\|\wp_{\mathcal{Z}_t\to\mathcal{S}_t,\hat{p}_t}(z,\cdot) - \wp_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}_t,p}(z,\cdot)\right\|_1$. We have,

$$\begin{aligned}
&\left\|\wp_{\mathcal{Z}_t\to\mathcal{Q}_t,\hat{p}_t}(z,\cdot) - \wp_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}_t,p}(z,\cdot)\right\|_1 \\
&= \left\|\hat{p}_t(z,\cdot) - \bar{\wp}_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}_t,p}(z',\cdot)\right\|_{TV} \\
&\leq \left\|\hat{p}_t(z,\cdot) - \bar{\wp}_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}^{(\ell)},p}(z',\cdot)\right\|_{TV} + \left\|\bar{\wp}_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}^{(\ell)},p}(z',\cdot) - \bar{\wp}_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}_t,p}(z',\cdot)\right\|_{TV} \\
&\leq \left\|\hat{p}_t^{(d)}(z,\cdot) - \wp_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}^{(\ell)},p}(z',\cdot)\right\|_1 + \left\|\bar{\wp}_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}^{(\ell)},p}(z',\cdot) - \bar{\wp}_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}_t,p}(z',\cdot)\right\|_{TV}.
\end{aligned} \tag{83}$$

By definition, $\mathcal{Q}_t$ is a finer partition of $\mathcal{S}$ than $\mathcal{Q}^{(\ell)}$. Hence, from Lemma I.2, we have that

$$\left\|\bar{\wp}_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}^{(\ell)},p}(z',\cdot) - \bar{\wp}_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}_t,p}(z',\cdot)\right\|_{TV} \leq C_p \operatorname{diam}(\zeta).$$

Next, we will provide a high probability upperbound on the first term of r.h.s. of (83). We will denote $\wp_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}^{(\ell)},p}(z',\cdot)$ by $p_t^{(d)}(z',\cdot)$ in order to simplify the notation. Note that both $\hat{p}_t^{(d)}(z,\cdot)$ and $p_t^{(d)}(z',\cdot)$ have the support $\tilde{\mathcal{S}}_t(z)$, where $|\tilde{\mathcal{S}}_t(z)| \leq d^{\frac{d_{\mathcal{S}}}{2}} \operatorname{diam}(\zeta)^{-d_{\mathcal{S}}}$. Let $\tilde{\mathcal{S}}_t^+(z)$ denote the collection of those points in $\mathcal{S}_t$ such that for any $s \in \tilde{\mathcal{S}}_t^+(z)$, we have $\hat{p}_t^{(d)}(z,s) - p_t^{(d)}(z',s) > 0$. So, we can write the following:

$$\begin{aligned}
\mathbb{P}\left(\left\|\hat{p}_t^{(d)}(z,\cdot) - p_t^{(d)}(z',\cdot)\right\|_1 \geq \iota\right) &= \mathbb{P}\left(\max_{\mathcal{S}'\subset\tilde{\mathcal{S}}_t^+(z)} \sum_{s\in\mathcal{S}'} \hat{p}_t^{(d)}(z,s) - p_t^{(d)}(z',s) \geq \frac{\iota}{2}\right) \\
&= \mathbb{P}\left(\cup_{\mathcal{S}'\subset\tilde{\mathcal{S}}_t^+(z)} \left\{\sum_{s\in\mathcal{S}'} \hat{p}_t^{(d)}(z,s) - p_t^{(d)}(z',s) \geq \frac{\iota}{2}\right\}\right).
\end{aligned} \tag{84}$$

Note that if $\mathcal{S}' \subset \tilde{\mathcal{S}}_t^+(z)$, then $\tilde{\mathcal{S}}_t(z) \setminus \mathcal{S}' \not\subset \tilde{\mathcal{S}}_t^+(z)$. Hence the number of subsets of $\tilde{\mathcal{S}}_t^+(z)$ is at most $2^{|\tilde{\mathcal{S}}_t(z)|-1}$. If $\mathbb{P}\left(\sum_{s\in\mathcal{S}'} \hat{p}_t^{(d)}(z,s) - p_t^{(d)}(z',s) \geq \frac{\iota}{2}\right) \leq b_\iota,\ \forall \mathcal{S}' \subset \tilde{\mathcal{S}}_t^+(z)$, then by an application of union bound in (84), we obtain that the following must hold,

$$\mathbb{P}\left(\left\|\hat{p}_t^{(d)}(z,\cdot) - p_t^{(d)}(z',\cdot)\right\|_1 \geq \iota\right) \leq 2^{|\tilde{\mathcal{S}}_t(z)|-1} b_\iota. \tag{85}$$

Consider a fixed $\xi \subseteq \mathcal{S}$. Define the following random processes,

$$v_i(z) := \mathbb{1}_{\{(s_i,a_i)\in\zeta_i\}}, \tag{86}$$

$$v_i(z,\xi) := \mathbb{1}_{\{(s_i,a_i,s_{i+1})\in\zeta_i\times\xi\}}, \tag{87}$$

$$w_i(z,\xi) := v_i(z,\xi) - p(s_i,a_i,\xi) v_i(z), \tag{88}$$

where $i = 0, 1, \ldots, T-1$. Let $\mathcal{S}' \subset S_t^+$ and $\xi = \cup_{s \in \mathcal{S}'} q^{-1}(s)$. Then we have,

$$\begin{aligned}
\sum_{s \in \mathcal{S}'} \hat{p}_t^{(d)}(z, s) - p_t^{(d)}(z', s) &= \frac{N_t(\zeta, \xi)}{N_t(\zeta)} - p(z', \xi) \\
&= \frac{N_t(\zeta, \xi) - p(z', \xi) N_t(\zeta)}{N_t(\zeta)} \\
&\leq \frac{1}{N_t(\zeta)} \left(\sum_{i=0}^{t-1} w_i(z, \xi) \right) + \frac{L_p}{2 N_t(\zeta)} \sum_{i=0}^{N_t(\zeta)} \operatorname{diam}(\zeta_{t_i}) \\
&\leq \frac{1}{N_t(\zeta)} \left(\sum_{i=0}^{t-1} w_i(z, \xi) \right) + 1.5 L_p \operatorname{diam}(\zeta), \qquad (89)
\end{aligned}$$

where the last step follows from Lemma I.1. Note that $\{w_i(z, \zeta)\}_{i \in [T-1]}$ is martingale difference sequence w.r.t. $\{\mathcal{F}_i\}_{i \in [T-1]}$. Moreover, $|w_i(z, \zeta)| \leq 1$. Hence from Lemma J.1 we have,

$$\mathbb{P}\left(\left\{ \frac{\sum_{i=0}^{t-1} w_i(z, \xi)}{N_t(\zeta)} \geq \sqrt{\frac{2}{N_t(\zeta)} \log\left(\frac{3}{\delta}\right)}, N_t(\zeta) = N \right\} \right) \leq \frac{\delta}{3}.$$

Upon combining this with (89) we get,

$$\mathbb{P}\left(\left\{ \sum_{s \in \mathcal{S}'} \hat{p}_t^{(d)}(z, s) - p_t^{(d)}(z', s) \geq \sqrt{\frac{2}{N_t(\zeta)} \log\left(\frac{3}{\delta}\right)} + 1.5 L_p \operatorname{diam}(\zeta), N_t(\zeta) = N \right\} \right) \leq \frac{\delta}{3}.$$

Upon using (85) in the above, and taking a union bound over all possible values of N, we obtain,

$$\mathbb{P}\left(\left\{ \left\| \hat{p}_t^{(d)}(z, \cdot) - p_t^{(d)}(z', \cdot) \right\|_1 \geq \sqrt{\frac{2|\tilde{\mathcal{S}}_t(z)|}{N_t(\zeta)} \log\left(\frac{3T}{\delta}\right)} + 3 L_p \operatorname{diam}(\zeta), N_t(\zeta) = N \right\} \right) \leq \frac{\delta}{3}.$$

Note that we do not have to take a union over all possible values of $\tilde{\mathcal{S}}_t(z)$ because of the one-to-one correspondence between $N_t(\zeta)$ and $\tilde{\mathcal{S}}_t(z)$. Replacing $|\tilde{\mathcal{S}}_t(z)|$ by its upper-bound $d^{\frac{d_{\mathcal{S}}}{2}} \operatorname{diam}(\zeta)^{-d_{\mathcal{S}}}$, we have,

$$\mathbb{P}\left(\left\| \hat{p}_t^{(d)}(z, \cdot) - p_t^{(d)}(z', \cdot) \right\|_1 \geq \operatorname{diam}(\zeta)^{-\frac{d_{\mathcal{S}}}{2}} \sqrt{\frac{2\, d^{\frac{d_{\mathcal{S}}}{2}} \log\left(\frac{3T}{\delta}\right)}{N_t(\zeta)}} + 3 L_p \operatorname{diam}(\zeta) \right) \leq \frac{\delta}{3}. \qquad (90)$$

Let $\mathcal{N}_1 := 2 d^{\frac{d}{2}} \left(\frac{T}{c_a \log(T/\delta)} \right)^{\frac{d}{d_{\mathcal{S}}+2}}$, which is the number of cells the ZoRL can activate under all sample paths. Upon taking union bound over all the cells that could possibly be activated in all possible sample paths at some t and using the fact that $N_t(\zeta) \geq N_{\min}(\zeta)$, the above inequality yields that with a probability at least $1 - \frac{\delta}{3}$, the following holds,

$$\left\| \hat{p}_t^{(d)}(z, \cdot) - p_t^{(d)}(z', \cdot) \right\|_1 \leq 3 \left(\frac{c_a \log\left(\frac{T}{\delta}\right)}{N_t(\zeta)} \right)^{\frac{1}{d_{\mathcal{S}}+2}} + 3 L_p \operatorname{diam}(\zeta), \qquad (91)$$

for every $z \in \zeta$, $\zeta \in \mathcal{P}_t$, and $t \in \{0, 1, \ldots, T-1\}$, where c_a is a constant that satisfies

$$d^{\frac{d_{\mathcal{S}}}{2}} \log\left(\frac{3T\mathcal{N}_1}{\delta}\right) \leq 4.5 c_a \log\left(\frac{T}{\delta}\right). \qquad (92)$$

After some algebraic manipulation, we obtain that it suffices to have,

$$c_a = \frac{2 d^{\frac{d_{\mathcal{S}}}{2}}}{9} \frac{\log\left(6 d^{\frac{d}{2}}\right)}{\log\left(\frac{T}{\delta}\right)} + \frac{d}{d_{\mathcal{S}}+2} + 1.$$

The proof follows upon combining the upper-bounds of the first and the second terms of (83). □

Remark. *See that* $\cap_{t=0}^{T-1} \{\wp_{\mathcal{S}_t \times \mathcal{A}_t \to \mathcal{S}_t, p}(\cdot, \cdot) \in \mathcal{C}_t\} \subseteq \mathcal{G}_1$*, where* $\mathcal{C}_t$ *is as defined in* (13)*. Hence,*

$$\mathbb{P}\left(\cap_{t=0}^{T-1} \{\wp_{\mathcal{S}_t \times \mathcal{A}_t \to \mathcal{S}_t, p} \in \mathcal{C}_t\} \right) \geq 1 - \frac{\delta}{3}.$$

G PROPERTIES OF EXTENDED VALUE ITERATION (EVI) AND EXTENDED POLICY EVALUATION (EPE)

We recall the definition of Extended MDP at time t that was discussed in Section 3,

$$\mathcal{M}_t^+ = \{(\mathcal{S}_t, \mathcal{A}_t, \tilde{p}, \tilde{r}_t) : \tilde{p} \in \mathcal{C}_t\},$$

where $\mathcal{S}_t$ and $\mathcal{A}_t$ are the discretized state and action space respectively, at time t, while $\tilde{r}$ is the discretized reward function with an additional bonus term. $\mathcal{C}_t$ is a set of plausible discrete transition kernels. Note that `ZoRL` calls the `EVI` subroutine (Algorithm 1) with a parameter γ which specifies the desired accuracy; upon calling `EVI` with accuracy parameter γ, it returns a policy that is γ-optimal for the extended MDP. We begin with introducing some notation. For $\phi \in \Phi_t$, $J_{\mathcal{M}_t^+}(\phi)$ denotes the value of the policy ϕ evaluated on the extended MDP $\mathcal{M}_t^+$. To be precise, this is the optimal average reward when the control action for the extended MDP is chosen according to the policy ϕ, and the kernel is chosen so as to maximize the average reward. The next result is similar in spirit to Jaksch et al. [2010, Theorem 7].

Lemma G.1. *Fix a time $t \in \mathbb{N}$. Consider the extended MDP $\mathcal{M}_t^+$ and the corresponding `EVI` iterates:*

$$\begin{aligned} v_0(s) &= 0, \\ v_{n+1}(s) &= \max_{\substack{a \in \mathcal{A}_t(s) \\ \theta \in \mathcal{C}_t}} \left\{ \tilde{r}_t(s,a) + \sum_{s' \in \mathcal{S}_t} \theta(s,a,s') v_n(s') \right\}, \ \forall s \in \mathcal{S}_t, n \in \mathbb{N}. \end{aligned} \tag{93}$$

Then,

$$\lim_{n \to \infty} (v_{n+1}(s) - v_n(s)) = J^\star_{\mathcal{M}_t^+}.$$

Moreover, whenever $sp\,(v_{n+1} - v_n) \leq \gamma$, the policy that chooses greedy actions which are optimal w.r.t. v_n, is γ-optimal.

Proof. Consider the n-th step of the `EVI` iteration, and let the action $a_n(s)$ and the kernel θ_n maximize the r.h.s. of (93), i.e.,

$$(a_n(s), \theta_n) \in \operatorname*{arg\,max}_{\substack{a \in \mathcal{A}_t(s) \\ \theta \in \mathcal{C}_t}} \left\{ \tilde{r}_t(s,a) + \sum_{s' \in \mathcal{S}_t} \theta(s,a,s') v_n(s') \right\}, \text{ for every } s \in \mathcal{S}_t.$$

Let $s^\star \in \arg\max_{s \in \mathcal{S}_t} v_n(s)$. Then, $\theta_i(s, \cdot)$ has to be chosen from the set $\mathcal{C}_t$ in such a manner that one assigns the maximum possible probability to a state in $s^\star$. Thus, we must have $\theta_i(s, s^\star) \geq \min\left\{1, \frac{1}{2}\eta_t(q_t^{-1}(s, a_n(s)))\right\}$, where $q_t^{-1}(s, a_n(s))$ is the active cell at time t that contains $(s, a_n(s))$. Since $\eta_t(q_t^{-1}(s, a_n(s))) > 0$ for all $s \in \mathcal{S}_t$, it follows that $\theta_i(s^\star, s^\star) > 0$. It is evident that the associated Markov chain is aperiodic. The proof then follows from Puterman [2014, Theorem 9.4.4]. The second claim follows from Puterman [2014, Theorem 8.5.6]. □

The next result follows from the previous result. It proves the convergence of the `EPE` algorithm (2), also derives the gap between the true value of a policy and that returned by the `EPE`.

Corollary G.2. *Fix a time $t \in \mathbb{N}$. Recall the extended MDP $\mathcal{M}_t^{d,+} = \{(\mathcal{S}_t, \mathcal{A}_t, \tilde{p}, d_t) : \tilde{p} \in \mathcal{C}_t\}$, where*

$$d_t(s,a) = \textit{diam}\left(q_t^{-1}(s,a)\right), \ \forall (s,a) \in \mathcal{S}_t \times \mathcal{A}_t,$$

policy $\phi \in \Phi_t$ and the corresponding `EPE` iterates:

$$\begin{aligned} g_0^\phi(s) &= 0, \\ g_{n+1}^\phi(s) &= \max_{\theta \in \mathcal{C}_t} \left\{ d_t(s, \phi(s)) + \sum_{s' \in \mathcal{S}_t} \theta(s, \phi(s), s') g_n^\phi(s') \right\}, \ \forall s \in \mathcal{S}_t, n \in \mathbb{N}. \end{aligned} \tag{94}$$

Then

$$\lim_{n \to \infty} \left(g_{n+1}^\phi(s) - g_n^\phi(s) \right) = \widetilde{diam}_t(\phi).$$

Moreover, when $sp\left(g_{n+1}^{\phi} - g_n^{\phi}\right) \leq \gamma(g_{n+1}^{\phi}(s_\star) - g_n^{\phi}(s_\star))$*, i.e. the stopping criteria is met, then* $(g_{n+1}^{\phi}(s_\star) - g_n^{\phi}(s_\star))$ *satisfies the following:*

$$\frac{\widetilde{diam}_t(\phi)}{1+\gamma} \leq (g_{n+1}^{\phi}(s_\star) - g_n^{\phi}(s_\star)) \leq \frac{\widetilde{diam}_t(\phi)}{1-\gamma}. \tag{95}$$

Proof. Similar to the proof of Lemma G.1, one can show that the transition kernels which maximize the r.h.s. in every iteration of EPE (94) are aperiodic. The convergence of EPE then follows from Puterman [2014, Theorem 9.4.4]. From Puterman [2014, Theorem 8.5.6], it follows that

$$\left|(g_{n+1}^{\phi}(s_\star) - g_n^{\phi}(s_\star)) - \widetilde{\text{diam}}_t(\phi)\right| \leq \gamma\,(g_{n+1}^{\phi}(s_\star) - g_n^{\phi}(s_\star)),$$

or

$$g_{n+1}^{\phi}(s_\star) - g_n^{\phi}(s_\star) \leq \frac{\widetilde{\text{diam}}_t(\phi)}{1-\gamma}, \text{ and, } g_{n+1}^{\phi}(s_\star) - g_n^{\phi}(s_\star) \geq \frac{\widetilde{\text{diam}}_t(\phi)}{1+\gamma}.$$

This concludes the proof. □

Remark (Upper and lower-bounds of episode duration)**.** *Let* $d_k = EPE(\mathcal{M}_{\tau_k}^{d,+}, \tilde{\phi}_k, \gamma, s_\star)$ *be the value of the policy* $\tilde{\phi}_k$ *evaluated on* $\mathcal{M}_{\tau_k}^{d,+}$*. From Corollary G.2 we have,*

$$\frac{\widetilde{diam}_{\tau_k}(\tilde{\phi}_k)}{1+\gamma} \leq d_k \leq \frac{\widetilde{diam}_{\tau_k}(\tilde{\phi}_k)}{1-\gamma}.$$

As $H_k = \frac{C_H \log\left(\frac{T}{\delta}\right)}{d_k^{2(d_{\mathcal{S}}+1)}}$*, we conclude that*

$$\frac{C_H(1-\gamma)^{2(d_{\mathcal{S}}+1)}\log\left(\frac{T}{\delta}\right)}{\widetilde{diam}_{\tau_k}(\tilde{\phi}_k)^{2(d_{\mathcal{S}}+1)}} \leq H_k \leq \frac{C_H(1+\gamma)^{2(d_{\mathcal{S}}+1)}\log\left(\frac{T}{\delta}\right)}{\widetilde{diam}_{\tau_k}(\tilde{\phi}_k)^{2(d_{\mathcal{S}}+1)}}. \tag{96}$$

H SIMULATION EXPERIMENTS

We perform simulations on the following environments.

1. `Continuous RiverSwim`: This environment models an agent who is swimming in a river [Strehl and Littman, 2008]. Though the original MDP is discrete, we use a continuous version of it. The state denotes the location of the agent in the river in a single dimension, and the action captures the movement of the agent. The state and action spaces are $[0, 6]$ and $[0, 1]$, respectively. The state of the system evolves as follows:

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

where $\{w_t\}$ is a 0-mean i.i.d. Gaussian random sequence. The reward function is given by

$$r(s, a) = 0.005(((s-6)/6)^4 + ((a-1)/2)^4) + 0.5((s/6)^4 + ((a+1)/2)^4).$$

2. `Truncated LQ System`: The state of an LQ [Abbasi-Yadkori and Szepesvári, 2011] system evolves as follows:

$$s_{t+1} = As_t + Ba_t + w_t,$$

where A, B are matrices of appropriate dimensions, and w_t is i.i.d. Gaussian noise. The reward at time t is $-s_t^\top P s_t - a_t^\top Q a_t$. We clip the state vector since our framework allows only compact state-action spaces. More specifically, we ensure that the state value for each coordinate lies within the interval $[c_\ell, c_u]$, and restrict the action space to be $[-1, 1]^{d_{\mathcal{A}}}$. Hence, the i-th coordinate of the state process evolves as

$$s_{t+1}(i) = \max\{\min\{(As_t + Ba_t + w_t)(i), c_u\}, c_\ell\}.$$

We have used the following two sets of system parameters:

(a) `Truncated LQ-1`:

$$A = \begin{bmatrix} -0.2 & -0.07 \\ 0.6 & 0.07 \end{bmatrix}, \quad B = \begin{bmatrix} 0.07 & 0.09 \\ -0.03 & -0.1 \end{bmatrix},$$

$P = 0.4\ I_2$[1], $Q = 0.6\ I_2$ and mean and standard deviation of w_t are 0 and 0.05, respectively. We consider $c_u = -c_\ell = 4$.

(b) `Truncated LQ-2`:

$$A = \begin{bmatrix} -0.2 & -0.07 \\ 0.6 & 0.07 \end{bmatrix}, \quad B = \begin{bmatrix} 0.1 & -0.01 & 0.12 & 0.08 \\ 0.02 & -0.1 & 0.3 & 0.001 \end{bmatrix}.$$

Values of P, Q, c_u, c_ℓ and mean and standard deviation of w_t are the same as `Truncated LQ-1`.

3. `Non-linear System`: We consider a non-linear system [Kakade et al., 2020] where the state evolves as

$$s_{t+1}(i) = \max\{\min\{(Af(s_t) + Bg(a_t) + w_t)(i), c_u\}, c_\ell\},$$

where f and g are non-linear functions, A, B are matrices of appropriate dimensions, and w_t is noise sequence. This system can be viewed as a generalization of the LQ control system in which the dynamics are linear in the feature vectors corresponding to state-action values. The feature maps $f(\cdot), g(\cdot)$ can be non-linear functions. The reward function is a function of the state and the actions. We have set the values for the matrices A, B, P, Q, c_u and c_ℓ to be the same as that of `Truncated LQ-1`. We set

$$f(s)(i) = 0.5s(i) + 0.5s(i)^2, \text{ for } i \in \{1, 2\}, \text{ and } g(a) = a^2,$$

where $v(i)$ denotes the i-th element of vector v. Similar to the `LQ system`, we consider the action space to be $[-1, 1]^{d_{\mathcal{A}}}$.

H.1 CHOOSING HYPERPARAMETERS

Since L_r (Assumption 2.1), c_a (92), C_η, C_H (68) may not be known, we instead provide their estimates/appropriate upper-bounds to `ZoRL` in lieu of these parameters. Our theoretical upper-bounds on regret continue to hold, we simply replace these parameters with the chosen upper-bounds. In addition to these `ZoRL` we pass δ and γ as hyperparameters to `ZoRL`. A brief description of these quantities are as follows:

1. L_r: We assume the knowledge of an upper-bound on L_r, the Lipschitz constant for the reward function (Assumption 2.1).
2. c_a: `ZoRL` activates a cell ζ if $N_t(\zeta) \geq \frac{c_a \log\left(\frac{T}{\delta}\right)}{\text{diam}(\zeta)^{d_{\mathcal{S}}+2}}$ (6), and deactivates ζ if $N_t(\zeta) \geq \frac{c_a 2^{d_{\mathcal{S}}+2} \log\left(\frac{T}{\delta}\right)}{\text{diam}(\zeta)^{d_{\mathcal{S}}+2}}$ (5).
3. C_η: Recall from Section 3 that if ζ is an active cell at time t, then its confidence radius $\eta_t(\zeta)$ satisfies $\eta_t(\zeta) \leq C_\eta \text{ diam}(\zeta)$, where $C_\eta = 3(1 + L_p) + C_p$. In order to avoid computing $\eta_t(\zeta)$, we use $C_\eta \text{ diam}(\zeta)$ as a substitute for $\eta_t(\zeta)$, and choose C_η as a hyperparameter for ZoRL
4. C_H: C_H is the multiplicative constant associated with the episode duration that satisfies (68).
5. δ: $\delta \in (0, 1)$ is the probability parameter.
6. γ: $\gamma > 0$ is the accuracy parameter for `EPE` subroutine that is used by `ZoRL` in order to compute the proxy diameter of the chosen policy in an episode.

The values of the following three hyperparameters are kept unchanged across four experiments: $L_r = 0.001$, $\delta = 0.1$ and $\gamma = 0.05$. Values of the rest of the parameters are reported in Table 1.

H.2 COMPARISON WITH PZRL-MF AND PZRL-MB

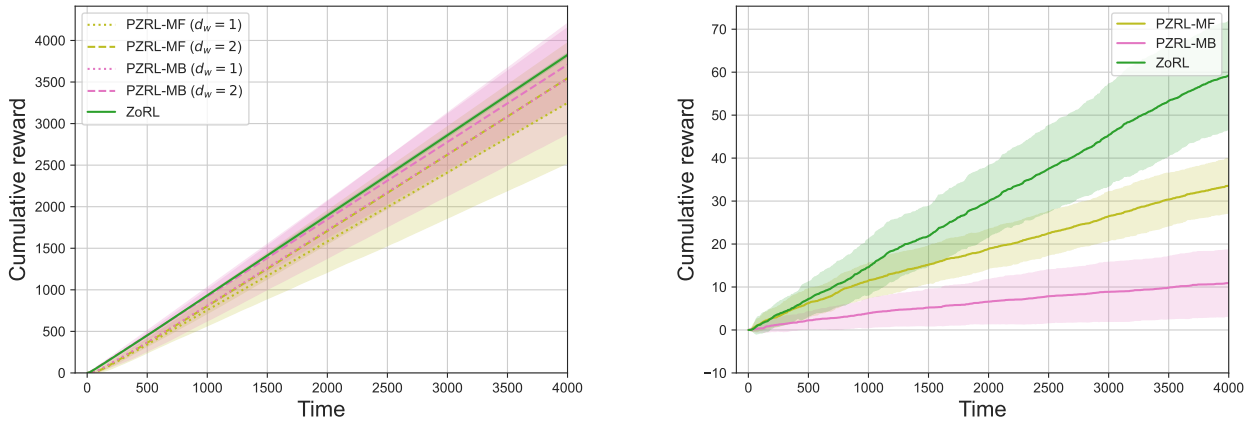
Kar and Singh [2024b] allows the agent to play policies from a parametric class. The latest version of Kar and Singh [2024b] proposes two new algorithms PZRL-MF and PZRL-MB[2]. Due to paucity of time, we could not compare PZRL-MF and PZRL-MB with `ZoRL` on the four environments discussed in Section 5. However, here we compare PZRL-MB

[1] I_n denotes identity matrix of size $n \times n$.

[2] These replace the PZRL-H algorithm, that has been proposed in an earlier version of the same paper.

Experiments	C_a	C_η	C_H
Truncated LQ-1	0.2	1	0.1
Truncated LQ-2	0.1	1	0.001
Continuous RiverSwim	0.1	1	0.001
Non-linear System	1	5	0.1

Table 1: ZoRL hyper-parameters.



(a) Continuous RiverSwim

(b) Continuous GridWorld

Figure 3: Comparison with PZRL-MF and PZRL-MB.

and PZRL-MF with `ZoRL` on the following two environments: (i) `Continuous RiverSwim` and (ii) `Continuous GridWorld`.

`Continuous RiverSwim`: This environment has been discussed above. We use the following two policy parameterization schemes for PZRL-MF and PZRL-MB.

1. $\phi(s; w) = w \cdot s$, $w \in [-1, 1]$.
2. $\phi(s; w) = w(1) + w(2)s^2$, $w = (w(1), w(2)) \in [-1, 1]^2$.

`Continuous GridWorld`: In GridWorld environment [Sutton and Barto, 2018], the agent moves around a compact space, and the space contains a designated reward-yielding region such that the agent earns a reward of 1 whenever it stays inside the reward-yielding region, and earns no reward otherwise. We design a continuous version of the same environment; the reward-yielding region is taken to be a circle of radius 0.1 units whose center is $[0.8, 0.8]$. The state space is $[0, 1]^2$ and the action space is $[0, 2\pi]$. The state of the system evolves as follows:

$$y_{t+1} = s_t + \beta \begin{bmatrix} \cos a_t \\ \sin a_t \end{bmatrix} + w_t, \text{ and}$$
$$s_t(i) = (0 \vee y_t(i)) \wedge 1, \text{ for } i = 1, 2, \ \forall t \in \{0\} \cup \mathbb{N},$$

where w_t is a zero-mean i.i.d. Gaussian noise, and $\beta > 0$ is the step-size. The standard deviation of w_t is set to 0.1, and we use a step size $\beta = 0.2$. For this environment, we parametrize the policies as follows: $\phi(s; w) = w(0) + s(0)w(1) + s(1)w(2)$, where $w \in [0, 1]^3$ and $s \in [0, 1]^2$.

We plot the cumulative rewards incurred by PZRL-MF, PZRL-MB, and `ZoRL`, averaged over 50 runs for both the systems in Figure 3.

Computing resources. We have conducted experiments on a 11-th Gen Intel Core-i7, 2.5GHz CPU processor with 16GB RAM using Python-3 and PyTorch library.

I AUXILIARY RESULTS

In this section, we derive some useful properties of the algorithm that are used in the proof of regret upper-bound. The first lemma shows that for any active cell ζ at time t, the quantity $\frac{1}{N_t(\zeta)}\sum_{i=1}^{N_t(\zeta)} \text{diam}\,(\zeta_{t_i})$ is bounded above by 3 diam (ζ). We use this in concentration inequality for the transition kernel estimate.

Lemma I.1. *For all $t \in [T-1]$ and $\zeta \in \mathcal{P}_t$, let t_i denote the time instance when ζ or any of its ancestor was visited by* ZoRL *for the i-th time. Then*

$$\frac{1}{N_t(\zeta)} \sum_{i=1}^{N_t(\zeta)} \textit{diam}\,(\zeta_{t_i}) \leq 3\ \textit{diam}\,(\zeta)\,.$$

Proof. By the activation rule (3.3), a cell ζ' can be played at most $N_{\max}(\zeta') - N_{\min}(\zeta') = \tilde{c}_a 2^{2\ell(\zeta')} + \frac{\tilde{c}_a}{3}\mathbb{1}_{\{\zeta'=\mathcal{S}\times\mathcal{A}\}}$ times while being active, where $\tilde{c}_a = 3c_a d^{-1}\log\left(\frac{T}{\epsilon\delta}\right)\epsilon^{-d_{\mathcal{S}}}$. We can write,

$$\begin{aligned}
\frac{1}{N_t(\zeta)} \sum_{i=1}^{N_t(\zeta)} \text{diam}\,(\zeta_{t_i}) &= \frac{1}{N_t(\zeta)} \sum_{i=1}^{N_{\min}(\zeta)} \text{diam}\,(\zeta_{t_i}) + \frac{1}{N_t(\zeta)} \sum_{i=N_{\min}(\zeta)+1}^{N_t(\zeta)} \text{diam}\,(\zeta_{t_i}) \\
&= \frac{\tilde{c}_a\sqrt{d}}{3N_t(\zeta)} + \frac{\tilde{c}_a\sqrt{d}}{N_t(\zeta)} \sum_{\ell=0}^{\ell(\zeta)-1} 2^\ell + \frac{N_t(\zeta) - N_{\min}(\zeta) - 1}{N_t(\zeta)} \text{diam}\,(\zeta) \\
&< \frac{\tilde{c}_a\sqrt{d}}{N_t(\zeta)} 2^{\ell(\zeta)} + \frac{N_t(\zeta) - N_{\min}(\zeta) - 1}{N_t(\zeta)} \text{diam}\,(\zeta) \\
&= \frac{3N_{\min}(\zeta)}{N_t(\zeta)} \text{diam}\,(\zeta) + \frac{N_t(\zeta) - N_{\min}(\zeta) - 1}{N_t(\zeta)} \text{diam}\,(\zeta) \\
&= \frac{(N_t(\zeta) + 2N_{\min}(\zeta) - 1)\ \text{diam}\,(\zeta)}{N_t(\zeta)} \\
&\leq 3\ \text{diam}\,(\zeta)\,,
\end{aligned}$$

where the last step is due to the fact that $N_{\min}(\zeta) \leq N_t(\zeta)$. □

Next, we show that under Assumption 4.2, the total variation norm between $\bar{\wp}_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}_t,p}(z,\cdot)$ and $\bar{\wp}_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}^{(\ell)},p}(z,\cdot)$ is bounded above by the discretization width of the partition $\mathcal{Q}^{(\ell)}$. We use this result in Lemma F.1.

Lemma I.2. *Let us fix any state-action pair z and time t. Let $\ell = \ell(q_t^{-1}(z))$. Recall distributions $\bar{\wp}_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}_t,p}(z,\cdot)$ and $\bar{\wp}_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}^{(\ell)},p}(z,\cdot)$ from Lemma F.1. Under Assumption 4.2, we have that*

$$\left\|\bar{\wp}_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}^{(\ell)},p}(z,\cdot) - \bar{\wp}_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}_t,p}(z,\cdot)\right\|_{TV} \leq C_p\sqrt{d}\,2^{-\ell}$$

for every $z \in \mathcal{S}\times\mathcal{A}$.

Proof. Recall that $\mathcal{S}_t$ is the set of representative points of $\mathcal{Q}_t$ and that $\mathcal{Q}^{(\ell)}$ is a coarser partition of $\mathcal{S}$ than $\mathcal{Q}_t$. Let us fix $\xi \in \mathcal{Q}^{(\ell)}$, and let us denote the Radon-Nikodym derivative of the distribution $p(z,\cdot)$ by f. Let $\bar{f} = p(z,\xi)/\lambda(\xi)$. We have,

$$\begin{aligned}
\sup_{B\subseteq\xi} \left|\bar{\wp}_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}^{(\ell)},p}(z,B) - p(z,B)\right| &\leq \int_\xi (f-\bar{f})\mathbb{1}_{\{f\geq\bar{f}\}} d\lambda \\
&\leq \int_\xi (\bar{f} + C_p\sqrt{d}\epsilon)\mathbb{1}_{\{f\geq\bar{f}\}} d\lambda - \int_\xi \bar{f}\mathbb{1}_{\{f\geq\bar{f}\}} d\lambda \\
&\leq C_p\sqrt{d}\epsilon \times \epsilon^{d_{\mathcal{S}}},
\end{aligned}$$

where $\epsilon = 2^{-\ell}$. Hence, by Assumption 4.2, we have that for every $z \in \mathcal{S}\times\mathcal{A}$ and for every $\xi \in \mathcal{Q}^{(\ell)}$,

$$\begin{aligned}
\sup_{B\subseteq\xi} \left|\bar{\wp}_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}^{(\ell)},p}(z,B) - \bar{\wp}_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}_t,p}(z,B)\right| &\leq \sup_{B\subseteq\xi} \left|\bar{\wp}_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}^{(\ell)},p}(z,B) - p(z,B)\right| \\
&\leq C_p\sqrt{d}\epsilon \times \epsilon^{d_{\mathcal{S}}}.
\end{aligned}$$

As $\mathcal{Q}^{(\ell)}$ is coarser than $\mathcal{Q}$, it follows that

$$\begin{aligned}\left\|\bar{\wp}_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}^{(\ell)},p}(z,\cdot)-\bar{\wp}_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}_t,p}(z,\cdot)\right\|_{TV} &\leq \sum_{\xi\in\mathcal{Q}^{(\ell)}}\sup_{B\subseteq\xi}\left|\bar{\wp}_{\mathcal{S}\times\mathcal{A}\to\mathcal{S}^{(\ell)},p}(z,B)-p(z,B)\right| \\ &\leq C_p\sqrt{d}\epsilon\times\epsilon^{d_{\mathcal{S}}}\times\epsilon^{-d_{\mathcal{S}}} \\ &\leq C_p\sqrt{d}\epsilon.\end{aligned}$$

Hence, we have proven the claim. □

J USEFUL RESULTS

J.1 CONCENTRATION INEQUALITIES

Lemma J.1 (Azuma-Hoeffding inequality)**.** *Let $X_1, X_2, \ldots$ be a martingale difference sequence with $|X_i| \leq c\forall i$. Then for all $\epsilon > 0$ and $n \in \mathbb{N}$,*

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i \geq \epsilon\right\} \leq e^{-\frac{\epsilon^2}{2nc^2}} \tag{97}$$

The following inequality is Proposition $A.6.6$ of Van Der Vaart et al. [1996].

Lemma J.2 (Bretagnolle-Huber-Carol inequality)**.** *If the random vector $(X_1, X_2, \ldots, X_n)$ is multinomially distributed with parameters N and $(p_1, p_2, \ldots, p_n)$, then for $\epsilon > 0$*

$$\mathbb{P}\left(\sum_{i=1}^{n}|X_i - Np_i| \geq 2\sqrt{N}\epsilon\right) \leq 2^n e^{-2\epsilon^2}. \tag{98}$$

Alternatively, for $\delta > 0$

$$\mathbb{P}\left(\sum_{i=1}^{n}\left|\frac{X_i}{N} - p_i\right| < \sqrt{\frac{2n}{N}\log\left(\frac{2}{\delta^{\frac{1}{n}}}\right)}\right) \geq 1-\delta. \tag{99}$$

The following is essentially Theorem 1 of Abbasi-Yadkori et al. [2011].

Theorem J.3 (Self-Normalized Tail Inequality for Vector-Valued Martingales)**.** *Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\eta_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that η_t is $\mathcal{F}_t$ measurable and η_t is conditionally R sub-Gaussian for some $R > 0$, i.e.,*

$$\mathbb{E}\left[\exp(\lambda\eta_t)|\mathcal{F}_{t-1}\right] \leq \exp\left(\lambda^2R^2/2\right), \forall\lambda \in \mathbb{R}.$$

Let $\{X_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$ valued stochastic process such that X_t is $\mathcal{F}_{t-1}$ measurable. Assume that V is a $d\times d$ positive definite matrix. For any $t \geq 0$, define

$$\bar{V}_t := V + \sum_{s=1}^{t} X_s X_s^{\top},$$

and

$$S_t := \sum_{s=1}^{t}\eta_s X_s.$$

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

$$\|S_t\|^2_{\bar{V}_t^{-1}} \leq 2R^2\log\left(\frac{\det(\bar{V}_t)^{1/2}\det(V)^{-1/2}}{\delta}\right).$$

Corollary J.4 (Self-Normalized Tail Inequality for Martingales). *Let $\{\mathcal{F}_i\}_{i=0}^{\infty}$ be a filtration. Let $\{\eta_i\}_{i=1}^{\infty}$ be a $\{\mathcal{F}_i\}_{i=0}^{\infty}$ measurable stochastic process and η_t is conditionaly R sub-Gaussian for some $R > 0$. Let $\{X_i\}_{i=1}^{\infty}$ be a $\{0,1\}$-valued $\mathcal{F}_{i-1}$ measurable stochastic process.*

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

$$\left|\sum_{i=1}^{k} \eta_i X_i\right| \leq R\sqrt{2\left(1+\sum_{i=1}^{k} X_i\right)\log\left(\frac{1+\sum_{i=1}^{k} X_i}{\delta}\right)}.$$

Proof. Taking $V = 1$, we have that $\bar{V}_t = 1 + \sum_{s=1}^{t} X_s$. The claim follows from Theorem J.3. □

J.2 OTHER USEFUL RESULTS

Lemma J.5. *Consider the following function $f(x)$ such that $0 < a_0 \leq \frac{a_1}{4}$,*

$$f(x) = a_0 x - \sqrt{a_1 x} - 1.$$

Then for all $x \geq 1.5\frac{a_1}{a_0^2}$, $f(x) \geq 0$.

Proof. See that $f(x) \geq 0$ for all $x \geq \left(\frac{\sqrt{a_1}+\sqrt{a_1+4a_0}}{2a_0}\right)^2$. Since $a_1 \leq 4a_0$, we have that for all $x \geq 1.5\frac{a_1}{a_0^2}$ $f(x) \geq 0$. □

Lemma J.6. *Let μ_1 and μ_2 be two probability measures on Z and let v be an $\mathbb{R}$-valued bounded function on Z. Then, the following holds.*

$$\left|\int_Z (\mu_1 - \mu_2)(z)v(z)dz\right| \leq \frac{1}{2}\left\|\mu_1 - \mu_2\right\|_{TV} \; sp\left(v\right).$$

Proof. Denote $\lambda(\cdot) := \mu_1(\cdot) - \mu_2(\cdot)$. Now let $Z_+, Z_- \subset Z$ be such that $\lambda(B) \geq 0$ for every $B \subseteq Z_+$ and $\lambda(B) < 0$ for every $B \subseteq Z_-$. We have that

$$\lambda(Z) = \lambda(Z_+) + \lambda(Z_-) = 0. \tag{100}$$

Also,

$$\lambda(Z_+) - \lambda(Z_-) = \left\|\mu_1 - \mu_2\right\|_{TV}. \tag{101}$$

Combining the above two, we get that

$$\lambda(Z_+) = \frac{1}{2}\left\|\mu_1 - \mu_2\right\|_{TV}. \tag{102}$$

Now,

$$\begin{aligned}
\left|\int_Z \lambda(z)v(z)dz\right| &= \left|\int_{Z_+} \lambda(z)v(z)dz + \int_{Z_-} \lambda(z)v(z)dz\right| \\
&\leq \left|\lambda(Z_+)\sup_{z\in Z} v(z) + \lambda(Z_-)\inf_{z\in Z} v(z)\right| \\
&= \left|\lambda(Z_+)\sup_{z\in Z} v(z) - \lambda(Z_+)\inf_{z\in Z} v(z) + \lambda(Z_+)\inf_{z\in Z} v(z) + \lambda(Z_-)\inf_{z\in Z} v(z)\right| \\
&= \lambda(Z_+)\left(\sup_{z\in Z} v(z) - \inf_{z\in Z} v(z)\right) \\
&= \frac{1}{2}\left\|\mu_1 - \mu_2\right\|_{TV} \; sp\left(v\right).
\end{aligned}$$

Hence, we have proven the lemma. □

Lemma J.7. *Let θ_1 and θ_2 be two transition probability kernels of two Markov chains with common state space $\mathcal{S}$. Let $\max_{s\in\mathcal{S}} \|\theta_1(s,\cdot) - \theta_2(s,\cdot)\|_{TV} \leq c$. Then,*

$$\left\|\theta_1^{(m)}(s,\cdot) - \theta_2^{(m)}(s,\cdot)\right\|_{TV} \leq m \cdot c,\ \forall m \in \mathbb{N}.$$

where $\theta_i^{(m)}$ is the m-step transition kernel of the Markov chain with one-step transition kernel θ_i for $i = 1, 2$.

Proof. We shall prove this using mathematical induction. The base case is given. Let us assume that,

$$\left\|\theta_1^{(i)}(s,\cdot) - \theta_2^{(i)}(s,\cdot)\right\|_{TV} \leq i \cdot c,\ \forall i = 1, 2, \ldots, m-1.$$

See that

$$\begin{aligned}
\left\|\theta_1^{(m)}(s,\cdot) - \theta_2^{(m)}(s,\cdot)\right\|_{TV} &= \Bigg\| \int_{\mathcal{S}} \theta_1^{(m-1)}(s,s')\theta_1(s',\cdot)ds' - \int_{\mathcal{S}} \theta_2^{(m-1)}(s,s')\theta_1(s',\cdot)ds' \\
&\qquad + \int_{\mathcal{S}} \theta_2^{(m-1)}(s,s')\theta_1(s',\cdot)ds' - \int_{\mathcal{S}} \theta_2^{(m-1)}(s,s')\theta_2(s',\cdot)ds' \Bigg\|_{TV} \\
&\leq 2 \sup_{A\in\mathcal{B}_{\mathcal{S}}} \int_{\mathcal{S}} \left(\theta_1^{(m-1)}(s,s') - \theta_2^{(m-1)}(s,s')\right) \theta_1(s',A)ds' \\
&\quad + 2 \sup_{A\in\mathcal{B}_{\mathcal{S}}} \int_{\mathcal{S}} \theta_2^{(m-1)}(s,s') \left(\theta_1(s',A) - \theta_2(s',A)\right) ds' \\
&\leq \left\|\theta_1^{(m-1)}(s,\cdot) - \theta_2^{(m-1)}(s,\cdot)\right\|_{TV} \sup_{A\in\mathcal{B}_{\mathcal{S}}} sp\left(\theta_1(\cdot,A)\right) \\
&\quad + \int_{\mathcal{S}} \theta_2^{(m-1)}(s,s') \left\|\theta_1(s',\cdot) - \theta_2(s',\cdot)\right\|_{TV} ds' \\
&\leq \left\|\theta_1^{(m-1)}(s,\cdot) - \theta_2^{(m-1)}(s,\cdot)\right\|_{TV} + \max_{s'\in\mathcal{S}} \|\theta_1(s',\cdot) - \theta_2(s',\cdot)\|_{TV},
\end{aligned}$$

where the first inequality follows from triangle inequality and from the definition of total variation distance, the second inequality follows from Lemma J.6 and by taking the supremum inside integration. This concludes the proof of the lemma. □