Provably Adaptive Average Reward Reinforcement Learning for Metric Spaces

Anonymous Author
Anonymous Institution

Abstract

We study infinite-horizon average-reward reinforcement learning (RL) for Lipschitz MDPs and develop an algorithm `ZoRL` that discretizes the state-action space adaptively and zooms into "promising regions" of the state-action space. We show that its regret can be bounded as $\tilde{\mathcal{O}}\big(T^{1-d_{\text{eff.}}^{-1}}\big)$, 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. d_z is a problem-dependent quantity, which allows us to conclude that if an MDP is benign, then its regret will be small. We note that the existing notion of zooming dimension for average reward RL is defined in terms of policy coverings, and hence it can be huge when the policy class is rich even though the underlying MDP is simple, so that the regret upper bound is nearly $O(T)$. The zooming dimension proposed in the current work is bounded above by d, the dimension of the state-action space, and hence is *truly* adaptive, i.e., 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). The agent does not know the transition probabilities and reward function of the underlying MDP. Its goal is to choose actions sequentially so as to maximize the cumulative rewards.

The current work studies infinite-horizon average reward MDPs and derives RL algorithms for *continuous* state-action spaces that are endowed with a metric. Though discrete MDPs and linear MDPs have been studied in detail in RL literature, they are not suitable for real-world applications which mostly 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). 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 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 the *algorithm*. For example, 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, where near-optimality is relative to the optimal Q function for episodic RL. 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, it is shown in Kar and Singh (2024a) that zooming ideas and algorithms which were 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 on the regret with respect

Preliminary work. Under review by AISTATS 2025. Do not distribute.

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

to an ϵ suboptimal policy, where d_z^ϵ is the "ϵ-zooming dimension" and satisfies $d_z^\epsilon < d$. 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. Kar and Singh (2024b) rectify the issue to some extent by working directly in the policy space and showing zooming behavior in this space rather than the state-action space. Thus, their algorithm "activates" more number of policies from the near-optimal regions in the policy space. They obtain $d_{\text{eff.}} = 2d_{\mathcal{S}} + d_z^\Phi + 2$, where d_z^Φ measures the size of near-optimal policies in Φ. More specifically, if $\Phi^{(\beta)}$ denotes the set of $(\beta, 2\beta]$-suboptimal policies, then it upper bounds the number of plays of $\Phi^{(\beta)}$ in terms of its β-covering number. It defines d_z^Φ to be the logarithm of the β-covering number of $\Phi^{(\beta)} \cap \Phi$. However, if the problem, or the policy-set Φ is not structured, in the worst case the quantity d_z^Φ can be prohibitively large 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 infinite horizon average reward RL. `ZoRL` combines adaptive discretization with the principle of optimism and yields zooming behavior. We provide regret upper bound of `ZoRL` as a function of the zooming dimension d_z. We define d_z in terms of the suboptimality gap of the state action pairs (5). We show that the regret of `ZoRL` is upper bounded as $\tilde{\mathcal{O}}\left(T^{1-d_{\text{eff.}}^{-1}}\right)$, 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_z^Φ. The current work attains a small d_z 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. This allows us to bypass the need to work with policy coverings. Our proof hinges on the existence of certain "key cells" that are discussed next.

2. 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 a sufficient number of 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 suboptimality of policy and suboptimality gap of state-action pairs (5). This result plays a crucial role in proving the existence of the key cells.

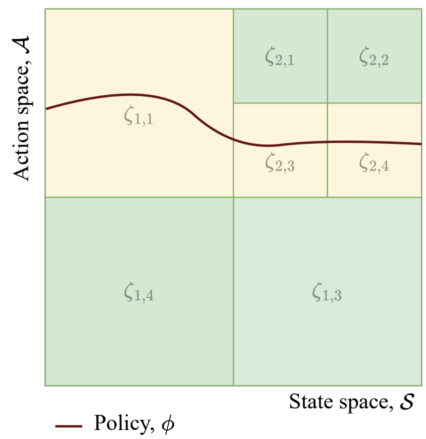


Figure 1: Key cell: The policy ϕ is played during the k-th episode. This diagram depicts the discretization grid at the beginning of the k-th episode. Then, one of the cells $\zeta_{1,1}$, $\zeta_{2,3}$ and $\zeta_{2,4}$ must be a key cell with a high probability (Lemma F.1). There must be a state s such that $(s, \phi(s))$ belongs to this cell, and s satisfies (27a) and (27b).

3. *Avoiding Over Exploration*: In order to prove the existence of the key cells, we have to overcome the problem of "over-exploration" that occurs in RL for continuous spaces. More specifically, the optimistic indices of policies might suffer from the problem of "excessive bonus terms," so we need to "control" the amount of optimism taken by `ZoRL`. We borrow the idea of bias-span constrained optimism that has been used in the literature on discrete MDPs to improve the regret of UCRL2, after suitably adapting it for our setup.

4. *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 attained by choosing the episode durations as a function of an estimate of the confidence diameter of the policy (25) that is played currently. We note that the popular approaches to choosing episode duration, such as doubling the episode length would fail to yield $d_{\text{eff.}} = 2d_{\mathcal{S}} + d_z + 3$.

5. Simulation experiments verify the gains of using adaptive discretization techniques as against popular fixed discretization-based algorithms.

1.2 Past Works

Lipschitz MDPs: Domingues et al. (2021) obtain $\tilde{\mathcal{O}}\left(H^3 K^{1-(2d+1)^{-1}}\right)$ regret by applying smoothing kernels. Cao and Krishnamurthy (2020) shows provable gains arising due to adaptive discretization and zooming and obtains $\tilde{\mathcal{O}}\left(H^{2.5+(2d_z+4)^{-1}}K^{1-(2d+1)^{-1}}\right)$ regret, where d_z is the zooming dimension defined specifically for episodic RL. Sinclair et al. (2023) proposes a model-based algorithm with adaptive discretization that has a regret upper bound of $\tilde{\mathcal{O}}\left(L_v H^{\frac{3}{2}} K^{1-(d_z+d_{\mathcal{S}})^{-1}}\right)$, where L_v is the Lipschitz constant for the value function. We note that as compared with the works on general function approximation, 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) 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 $\tilde{\mathcal{O}}(d\sqrt{DT})$, where D is the diameter (Wu et al., 2022). The current work develops algorithms 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. When the underlying continuous MDP has a transition kernel that is α-Hölder continuous and infinitely often smoothly differentiable, 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) 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. Moreover, the algorithm proposed in Kar and Singh (2024b) is computationally infeasible.

2 Problem Setup

Notation. The set of natural numbers is denoted by $\mathbb{N}$, the set of positive integers by $\mathbb{Z}_+$. 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., $\|\mu\|_{TV} := \sup\{|\mu(B)| : B \in \mathcal{F}\}$. We use $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r)$ to denote 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 $\mathcal{S}$ and $\mathcal{A}$ be endowed with Borel σ-algebra $\mathcal{B}_{\mathcal{S}}$ and $\mathcal{B}_{\mathcal{A}}$, respectively. 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. For $Z \subseteq \mathcal{S} \times \mathcal{A}$, $\text{diam}\,(Z) := \sup_{z_1,z_2\in Z} \rho(z_1, z_2)$. We denote the system state and action taken at time t by s_t, a_t respectively. The state s_t evolves as follows,

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

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,

$$\rho\left((s, a), (s', a')\right) \leq \rho_{\mathcal{S}}(s, s') + \rho_{\mathcal{A}}(a, a'), \tag{2}$$

for all $(s, a), (s', a') \in \mathcal{S} \times \mathcal{A}$.

Definition 2.1 (Stationary Deterministic Policy). *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 (Lai and Robbins, 1985) of a learning algorithm ψ until T is defined as,

$$\mathcal{R}(T;\psi) := T J^{\star}_{\mathcal{M}} - \mathbb{E}\left[\sum_{t=0}^{T-1} r(s_t, a_t)\right]. \tag{3}$$

The goal of this work is to design a learning algorithm with tight regret upper bound for Lipschitz MDPs. We now introduce the class of Lipschitz MDPs.

Assumption 2.2 (Lipschitz continuity). *(i) The reward function r is L_r-Lipschitz, i.e., $\forall\ s, s' \in \mathcal{S}, a, a' \in \mathcal{A}$,*

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

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

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

The following assumption ensures that the MDP is mixing and is typically required for average reward setup.

Assumption 2.3 (Ergodicity)**.** *There exists $\alpha \in (0,1)$ such that for each $s, s' \in \mathcal{S}$ and $a, a' \in \mathcal{A}$, we have*

$$\|p(s,a,\cdot) - p(s',a',\cdot)\|_{TV} \leq 2\alpha. \tag{4}$$

We note that even when $\mathcal{M}$ is known, (4) is the weakest known sufficient condition that ensures a computationally efficient way to obtain an optimal policy (Arapostathis et al., 1993). It is shown in Appendix A that if an MDP satisfies Assumptions 2.3, the controlled Markov chain (Hernández-Lerma, 2012) has a unique invariant distribution and is geometrically ergodic.

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 2.3, 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:

$$\begin{aligned} \text{gap}\,(s,a) :=& J^\star_{\mathcal{M}} + h_{\mathcal{M}}(s) - r(s,a) \\ &- \int_{\mathcal{S}} h_{\mathcal{M}}(s')\, p(s,a,s')\, ds'. \end{aligned} \tag{5}$$

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

$$d_z := \inf \left\{ d' > 0 \mid \mathcal{N}_{c_s\beta}\left(Z_\beta\right) \leq c_z \beta^{-d'},\ \forall \beta > 0 \right\}, \tag{6}$$

where $\mathcal{N}_{c_s\beta}\left(Z_\beta\right)$ denotes the $c_s\beta$-covering number of Z_β, c_s 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

Definition 3.1 (Cells)**.** *A cell is a dyadic cube with vertices from the set $\{2^{-\ell}(v_1, v_2, \ldots, v_d) : v_j \in \left[2^\ell\right], j = 1, 2, \ldots, d\}$ 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,*

$$\pi_{\mathcal{S}}(\zeta) : = \{ s \in \mathcal{S} \mid (s,a) \in \zeta \textit{ for some } a \in \mathcal{A} \}, \tag{7}$$

and its level is the same as that of ζ. For a cell/$\mathcal{S}$-cell ζ, we let $\ell(\zeta)$ denote its level. For a cell/$\mathcal{S}$-cell ζ, we let $q(\zeta)$ be 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 $q(\zeta) = z$. Denote the set of $\mathcal{S}$-cells of level ℓ by $\mathcal{Q}^{(\ell)}$.

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 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.*

The proposed algorithm `ZoRL` (2) 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 denote,*

$$N_{\max}(\zeta) := \frac{c_1 2^{d_{\mathcal{S}}+2} \log\left(\frac{T}{\delta}\right)}{\textit{diam}\,(\zeta)^{d_{\mathcal{S}}+2}}, \textit{ and,} \tag{8}$$

$$N_{\min}(\zeta) := \begin{cases} 1 & \textit{if } \zeta = \mathcal{S} \times \mathcal{A} \\ \frac{c_1 \log\left(\frac{T}{\delta}\right)}{\textit{diam}(\zeta)^{d_{\mathcal{S}}+2}}, & \textit{otherwise,} \end{cases} \tag{9}$$

where $c_1 > 0$ is a constant that is discussed in Lemma B.1, 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.,*

$$N_t(\zeta) := \sum_{i=0}^{t-1} \mathbb{1}_{\{(s_i, a_i) \in \zeta_i\}}, \tag{10}$$

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 $\mathcal{P}_t$.

We note that since the diameter of a child cell is equal to 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, i.e., $\mathcal{P}_t$, forms a partition of the state action space. Denote the collection of representative points of the active cells at time t by $Z_t := \{q(\zeta) : \zeta \in \mathcal{P}_t\}$. Denote $\mathcal{Q}_t = \mathcal{Q}^{(\ell_{\max,t})}$, where $\ell_{\max,t}$ is the level of the smallest cells in $\mathcal{P}_t$. Define the discrete state space at time t as $S_t := \{q(\zeta) : \zeta \in \mathcal{Q}_t\}$. Define the discretized transition kernel $\wp(z, s; Z, \mathcal{Q}_t)$ with support S_t as follows,

$$\wp(z, \cdot; Z, \mathcal{Q}_t) := p(z, q^{-1}(s)),\ \forall z \in Z, s \in S_t. \quad (11)$$

Estimating the Transition Kernel. We denote $\tilde{S}_t(z) := \{q(\xi) : \xi \in \mathcal{Q}^{(\ell_t(z))}\}$, where $\ell_t(z)$ denotes the level of the active cell that contains the state-action pair z at time t. We first construct an estimate $\hat{p}_t^{(d)}$ (12) of the discretized version of the true stochastic kernel. The support of $\hat{p}_t^{(d)}(z, \cdot)$ is $\tilde{S}_t(z)$, which is a finite set and depends upon diam (ζ), where ζ is the active cell which contains z. We choose the support adaptively as a function of diam (ζ) in order to bound regret in terms of d_z. The alternative is to set the number of support elements according to a covering of $\mathcal{S}$, but then the agent needs to know d_z. $\hat{p}_t^{(d)}(z, \cdot)$ is then extended to obtain a continuous kernel $\hat{p}_t$, which is then discretized w.r.t. the partition $\mathcal{Q}_t$ in order to again generate a discrete kernel $\hat{\wp}$. This construction of $\hat{\wp}$ from $\hat{p}_t^{(d)}$ ensures that the support of the discrete kernel at every point is the same (S_t), and hence this allows us to use `ScOpt`. Denote the total number of transitions from a cell ζ to a $\mathcal{S}$-cell ξ until t by $N_t(\zeta, \xi)$, i.e., $N_t(\zeta, \xi) := \sum_{i=1}^{t-1} \mathbb{1}_{\{(s_i, a_i, s_{i+1}) \in \zeta_i \times \xi\}}$. Define,

$$\hat{p}_t^{(d)}(z, s) := \frac{N_t\left(q^{-1}(z), q^{-1}(s)\right)}{1 \vee N_t\left(q^{-1}(z)\right)}, \quad (12)$$

$z \in Z_t, s \in \tilde{S}_t(z)$. It is a discretized estimate of the transition kernel with support $\tilde{S}_t(z)$[2]. The continuous extension of $\hat{p}_t^{(d)}$ on $\mathcal{S}$ is defined as,

$$\hat{p}_t(z, B) := \sum_{s \in \tilde{S}_t(z)} \frac{\lambda(B \cap q^{-1}(s))}{\lambda(q^{-1}(s))} \hat{p}_t^{(d)}(z, s), \quad (13)$$

where $z \in Z_t$, B is Borel measurable subset of $\mathcal{S}$, and $\lambda(\cdot)$ is the Lebesgue measure. Define the discretized version of $\hat{p}_t$ as follows:

$$\hat{\wp}_t(z, s; Z_t, \mathcal{Q}_t) := \hat{p}_t(z, q^{-1}(s)),\ \forall z \in Z_t, s \in S_t \quad (14)$$

Concentration Inequality. For a cell $\zeta \in \mathcal{P}_t$, the confidence radius associated with the estimate $\hat{\wp}_t(q(\zeta), \cdot; Z_t, \mathcal{Q}_t)$ is defined as follows,

$$\eta_t(\zeta) := \min\left\{2, (4-\alpha)\left(\frac{c_1 \log\left(\frac{T}{\delta}\right)}{N_t(\zeta)}\right)^{\frac{1}{d_{\mathcal{S}}+2}} + (3L_p + C_v)\mathrm{diam}\,(\zeta)\right\}, \quad (15)$$

where the constant $c_1 > 0$ is discussed in Lemma B.1, and C_v is as in Assumption 3.4. It follows from the property of the activation rule that for any cell ζ that is active at t, we must have $N_t(\zeta) \geq N_{\min}(\zeta)$, so that, $\eta_t(\zeta) \leq C_\eta \mathrm{diam}\,(\zeta)$, where $C_\eta := 4 - \alpha + 3L_p + C_v$. Let Z be a subset of $\mathcal{S} \times \mathcal{A}$ and $\mathcal{Q}$ be a partition of $\mathcal{S}$ that is composed of $\mathcal{S}$-cells. Let $\Theta(Z, \mathcal{Q})$ denote the set of all possible discrete transition kernels that are supported on $S = \{q(\xi) : \xi \in \mathcal{Q}\}$ and are defined on Z, i.e., $\Theta(Z, \mathcal{Q}) := \{\theta \mid \theta(z) : S \to [0,1], \sum_{s \in S} \theta(z, s) = 1, \theta(z, s) \geq 0$ for every $z \in Z\}$. Define

$$\mathcal{C}_t(Z) := \big\{\theta \in \Theta(Z, \mathcal{Q}_t) : \|\theta(z', \cdot) - \hat{\wp}_t(z, \cdot; Z_t, \mathcal{Q}_t)\|_1 \leq \eta_t(\zeta) \text{ for every } z \in Z_t, z' \in q^{-1}(z) \cap Z\big\}. \quad (16)$$

We make the following assumption on the true kernel p for deriving concentration results.

Assumption 3.4 (Bounded Radon-Nikodym derivative)**.** *The probability measures $\{p(s, a, \cdot)\}$ are absolutely-continuous w.r.t. the Lebesgue measure on $\mathcal{S}$, with density functions given by $\{f_{(s,a)}\}$. We assume that these densities satisfy*

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

where $\tilde{s} = (\tilde{s}(1), \tilde{s}(2), \cdots, \tilde{s}(d_{\mathcal{S}}))$ is the state vector, and C_v is known to the agent.

Assumption 3.4 ensures that the discretizations of $p(z, \cdot)$ with respect to the partitions $\mathcal{Q}^{(\ell_t(z))}$ and $\mathcal{Q}_t$ are at most $C_v \mathrm{diam}\left(q_t^{-1}(z)\right)$ distance apart (Lemma H.2). Lemma B.1 shows that under Assumption 2.2 and Assumption 3.4, the set $\cap_{t=0}^{T-1}\{\wp(\cdot, \cdot; \mathcal{S} \times \mathcal{A}, \mathcal{Q}_t) \in \mathcal{C}_t(\mathcal{S} \times \mathcal{A})\}$ has a high probability.

At the beginning of each episode, `ZoRL` constructs a set of MDPs denoted by $\mathcal{M}_t^+$ that have a) discrete transition kernels that are close to $\wp$, b) their reward function is the same, and equal to the sum of the true rewards at discrete points and a bonus term which compensates for the discretization error. It then solves an optimization problem (17) dubbed the Extended MDP that yields an *optimistic policy*, i.e., the optimal value of this problem exceeds the optimal average reward of the true MDP. It carefully ensures that this optimistic value is not too large, since otherwise it would lead to a high regret.

Extended MDP. We begin with some definitions. The set of all the relevant cells for $s \in \mathcal{S}$ at time t are defined as $Rel_t(s) := \{\zeta \in \mathcal{P}_t : \exists a \in \mathcal{A}$ such that $(s, a) \in \zeta\}$. These are those active cells whose $\mathcal{S}$-projection contain the state s. Thus, $Rel_t(s)$ can be seen as the set of those cells in the state-action space that are associated with state s currently. Recall that S_t is the discrete state space at time t. Define

$$A_t(s) := \{q(\pi_{\mathcal{A}}(\zeta)) \mid \zeta \in Rel_t(s)\}.$$

Thus, $A_t(s)$ 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 $A_t := \{A_t(s) :$

[2]Note that $\hat{p}_t^{(d)}(z, \cdot)$ and $\wp(z, \cdot; Z, \mathcal{Q}_t)$ have different support.

$s \in S_t\}$. Let Φ_t be the set of policies for such that for every $\phi \in \Phi_t$, $\phi(s) \in A_t(s),\ \forall s \in S_t$. Consider the following modified reward function defined for all $(s,a) \in S_t \times A_t$,

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

in which a bonus term proportional to the diameter of the cell has been included in order to compensate for the "discretization error." Let $\Phi_t(M,c) \subseteq \Phi_t$ denote the set of policies for which $\max_{\phi \in \Phi_t(M,c)} sp\left(h_M^{\phi}\right) \leq c$, where $c > 0$. Consider the following collection of MDPs $\mathcal{M}_t^+ = \{(S_t, A_t, \tilde{p}, \tilde{r}_t) : \tilde{p} \in \mathcal{C}_t(S_t \times A_t), \tilde{p}(s,a,s') \geq \frac{1-\alpha}{|S_t|T}\ \forall (s,a,s') \in S_t \times A_t \times S_t\}$. The following optimization problem is dubbed extended MDP:

$$\max_{M \in \mathcal{M}_t^+} \max_{\phi \in \Phi_t(M,c)} J_M(\phi), \tag{17}$$

where

$$c := \frac{1 + L_r}{(1-\alpha)(1-T^{-1})}. \tag{18}$$

Denote the optimal value of this problem as $J^\star_{\mathcal{M}_t^+,c}$. `ZoRL` uses the `ScOpt` (Fruit et al., 2018) algorithm in order to solve (17). This is discussed next.

Algorithm 1 `ScOpt`

Input Extended MDP $\mathcal{M}^+$, span bound c, contraction factor $\gamma \in (0,1)$, accuracy parameter $\epsilon > 0$, reference state $s_\star$.
Initialize $v_0 = \{0\}^{|S|}$, $n = 0$.
while True **do**
  $v_{n+1} = \Gamma_c v_n - \min_{s \in S} (\Gamma_c v_n)(s) e$
  if $sp\,(v_{n+1} - v_n) + \frac{2\gamma^n}{1-\gamma} sp\,(v_1) \leq \epsilon$ **then**
    break
  end if
  $n \leftarrow n+1$
end while
return $G_c v_n$

`ScOpt` (1) takes as an input a set of discrete MDPs $\mathcal{M}^+$, then solves the extended MDP (17), and returns a tuple (MDP, policy) that jointly optimize (17). Consider the set $\mathcal{M}^+ = \{(S, A, \tilde{p}, \tilde{r}) : \tilde{p} \in \mathcal{C}\}$. S is the discrete state space, and $A = \{A(s) : s \in S\}$ is the discrete action space, where $A(s)$ is the set of permissible actions at state s. $\mathcal{C}$ is a set of transition kernels from the points in $S \times A$ and the support of the transition distributions is S. $\tilde{r}$ is the reward function. Given the set $\mathcal{M}^+$ as above, define the following operators:

$$\mathcal{T}v(s) = \max_{\substack{a \in A(s) \\ \theta \in \mathcal{C}}} \Big\{\tilde{r}(s,a) + \sum_{s' \in S} \theta(s,a,s') v(s')\Big\}, \tag{19}$$

$$\text{and } \Gamma_c v(s) = \min\{\mathcal{T}v(s), \min_{s' \in S} \mathcal{T}v(s') + c\}, \tag{20}$$

where c is as in (18). We observe that Γ_c "truncates" the vector $\mathcal{T}v$ in case we have $sp\,(\mathcal{T}v) > c$. This truncation helps us in ensuring that the span of the iterates is upper bounded. This, in turn, allows us to upper bound the optimistic index of policies, and hence the regret. Γ_c is said to be feasible at $(v,s) \in \mathbb{R}^S \times S$ if there exists a (ϕ_v, θ_v) such that the following holds,

$$\Gamma_c v(s) = \sum_{a \in A(s)} \phi_v(a|s) \Big\{\tilde{r}(s,a) + \sum_{s' \in S} \theta_v(s,a,s') v(s')\Big\}, \tag{21}$$

where ϕ_v is a stationary randomized policy[3], and θ_v is a stochastic kernel in $\mathcal{C}$. We note that a maximizing (ϕ,θ) always exists in (19). However, since Γ_c truncates the vector $\mathcal{T}v$, it might be the case that there may not exist such (θ_v, ϕ_v) that satisfies (21), so that the feasibility of Γ_c is not automatically guaranteed. Given $v \in \mathbb{R}^S$, define the stationary randomized policy $G_c v$ as follows:

$$G_c v(s) := \tag{22}$$
$$\begin{cases} \phi_v(\cdot \mid s), \text{ if } (v,s) \text{ is feasible}, \\ \arg\min\limits_{\substack{a \in A(s), \\ \theta \in \mathcal{C}}} \{\tilde{r}(s,a) + \sum\limits_{s' \in S} \theta(s,a,s') v(s')\}, \text{ otherwise.} \end{cases}$$

We show in Lemma C.2 that with $\mathcal{M}^+$ set equal to $\mathcal{M}_t^+$, w.h.p. Γ_c is feasible at every $(v,s) \in \mathbb{R}^{S_t} \times S_t$ with $sp\,(v) \leq c$, and the corresponding policy $G_c v$ is deterministic. Also, note that Γ_c is a $(1-\gamma)$-contraction (Puterman, 2014, Theorem 6.6.6), where $\gamma := 1 - \frac{1-\alpha}{|S_t|T}$. So, w.h.p. Γ_c satisfies Assumption 6 and Assumption 9 of (Fruit et al., 2018), and hence, by Fruit et al. (2018, Theorem 10), we conclude that w.h.p. `ScOpt`$(M_t^+, c, \gamma, \epsilon, s_\star)$ returns a deterministic policy $\phi \in \Phi_t$ such that $J_{\mathcal{M}_t^+,c}(\phi) \geq J^\star_{\mathcal{M}_t^+,c} - \epsilon$, where

$$J_{\mathcal{M}_t^+,c}(\phi) := \sup_{M \in \mathcal{M}_t^+} J_M(\phi). \tag{23}$$

Let τ_k denote the time when the k-th episode starts. At time τ_k, `ZoRL` calls `ScOpt`$(M_{\tau_k}^+, c, \gamma, 1/T, s_\star)$, which then returns the policy $\tilde{\phi}_k$. It then extends $\tilde{\phi}_k$ on the continuous space $\mathcal{S}$, so as to yield a policy denoted by ϕ_k which is described 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.,

$$\phi_k(s) = \tilde{\phi}_k(q(\zeta)), \tag{24}$$

for every $s \in \xi$, for every $\xi \in \mathcal{Q}_{\tau_k}$.

Episode Duration. The diameter of a policy ϕ at time t is defined as

$$\text{diam}_t(\phi) := \int_{\mathcal{S}} \text{diam}\left(q_t^{-1}(s,\phi(s))\right) \mu_{\phi,p}^{(\infty)}(s) ds, \tag{25}$$

[3] $\phi_v(s \mid a) = \mathbb{P}(a_t = a \mid s_t = s)$ for any $t \in \mathbb{Z}_+$.

where $\mu^{(\infty)}_{\phi,p}$ denotes the stationary measure of the Markov process induced by ϕ when it is applied to p. The duration of the k-th episode is decided based on an estimate of $\mathrm{diam}_{\tau_k}(\phi)$. Since $\mu^{(\infty)}_{\phi,p}$ is unknown, we make the following assumption, which helps us in estimating $\mathrm{diam}_{\tau_k}(\phi_k)$.

Assumption 3.5. *There is a probability measure ν, and constants $0 < \kappa_1 \le \kappa_2$ such that for every policy $\phi \in \Phi_{SD}$, and for every $\zeta \in \mathcal{B}_{\mathcal{S}}$, we have, $\kappa_1 \nu(\zeta) \le \mu^{(\infty)}_{\phi,p}(\zeta) \le \kappa_2 \nu(\zeta)$.*

Remark. *Similar or more restrictive assumptions are often required in the average reward setup for continuous space MDPs. 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. (2023) 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 3.5.*

`ZoRL` assumes the knowledge of κ_1 and ν and constructs a lower bound of $\mathrm{diam}_{\tau_k}(\phi_k)$ as follows:

$$\underline{\mathrm{diam}}_{\tau_k}(\phi_k) := \kappa_1 \int_{\mathcal{S}} \mathrm{diam}\left(q^{-1}_{\tau_k}(s, \phi(s))\nu(s) ds\right). \quad (26)$$

H_k, the duration of the k-th episode, is chosen as $H_k = \frac{C_H \log(T/\delta)}{\underline{\mathrm{diam}}_{\tau_k}(\phi_k)^{2(d_{\mathcal{S}}+1)}}$ where C_H is a problem-dependent constant, discussed in Lemma F.2.

4 Regret Analysis

We let $\Delta(\phi) := J^\star_{\mathcal{M}} - J_{\mathcal{M}}(\phi)$ be the suboptimality of policy ϕ. The following result establishes a relation between the suboptimality gap of state-action pairs and the suboptimality of the policies. 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*

$$\Delta(\phi) = \int_{\mathcal{S}} gap\left(s, \phi(s)\right)\ \mu^{(\infty)}_{\phi,p}(s)\, ds.$$

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.

Algorithm 2 Zooming Algorithm for RL (`ZoRL`)

Input Horizon T, Constant L_r, L_p, α and C_v.
Initialize $h = 0$, $k = 0$, $H_0 = 0$, $\mathcal{P}_0 = \{\mathcal{S} \times \mathcal{A}\}$.
for $t = 0$ to $T-1$ **do**
  if $h \ge H_k$ **then**
    $k \leftarrow k+1$, $h \leftarrow 0$, $\tau_k = t$, $s_\star \in S_t$
    $\tilde{\phi}_k = \mathtt{ScOpt}(\mathcal{M}^+_{\tau_k}, c, \gamma, 1/T, s_\star)$.
    Obtain ϕ_k from $\tilde{\phi}_k$ according to (24).
    $H_k = \frac{C_H \log(T/\delta)}{\underline{\mathrm{diam}}_{\tau_k}(\phi_k)^{2(d_{\mathcal{S}}+1)}}$
  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**
    $\mathcal{P}_{t+1} = \mathcal{P}_t \cup Child(q_t^{-1}(s_t, a_t)) \setminus \{q_t^{-1}(s_t, a_t)\}$
  else
    $\mathcal{P}_{t+1} = \mathcal{P}_t$
  end if
end for

Theorem 4.2. *$\mathcal{R}(T; \mathtt{ZoRL})$ is upper bounded as $\mathcal{O}\left(T^{\frac{2d_{\mathcal{S}}+d_z+2}{2d_{\mathcal{S}}+d_z+3}}\right)$.*

Proof sketch:
Regret decomposition: We decompose the regret (3) in the following manner. Let $K(T)$ denote the total number of episodes during T timesteps. Then, for any learning algorithm ψ,

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

The term (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 stationary distribution. (a) and (b) are bounded separately below.

Bounding (a): Step 1: We show that the policy obtained by solving (17) is almost optimistic, i.e., w.h.p. $J^\star_{\mathcal{M}^+_t,c} \ge J^\star_{\mathcal{M}} - \mathcal{O}(1/T)$ (Corollary D.2). We also show that w.h.p., $J^\star_{\mathcal{M}^+_t,c} \le J^\star_{\mathcal{M}} + C_1 \mathrm{diam}_t(\phi_k) + \mathcal{O}(1/T)$, where $C_1 = 3L_r + \frac{2C_\eta}{1-\alpha}$ (Corollary D.4). As a consequence of the above two results, on a high probability set, a suboptimal policy ϕ satisfying $\Delta(\phi) > 1/T$ will never be played from episode k onwards if $\text{diam}_{\tau_k}(\phi) \leq \textit{const} \cdot \Delta(\phi)$. Note that the cumulative regret arising due to policies with $\Delta(\cdot)$ less than $1/T$ is at most a constant, so we can safely restrict analysis to regret arising from play of other policies.

Step 2: Combining Step 1 with Lemma 4.1, we show in Lemma F.1 that on a high probability set, for each k, there is a $s \in \mathcal{S}$ such that

$$\text{diam}\,(\zeta) \geq \frac{1}{3C_1} \max\{\text{gap}\,(s, \phi_k(s))\,, C_1 \text{diam}_{\tau_k}(\phi_k)\}, \tag{27a}$$

$$\mu^{(\infty)}_{\phi_k,p}(\pi_{\mathcal{S}}(\zeta)) \geq \left(\frac{\text{diam}_{\tau_k}(\phi_k)}{3}\right)^{d_{\mathcal{S}}+1}, \tag{27b}$$

where $\zeta = q^{-1}_{\tau_k}(s, \phi_k(s))$. We call the cell ζ a key cell of 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 $\mathcal{O}\left(\log\left(\frac{T}{\delta}\right)\text{diam}\,(\zeta)^{-2(d_{\mathcal{S}}+1)}\right)$ times during the k-th episode. This is done in Lemma F.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 F.3).

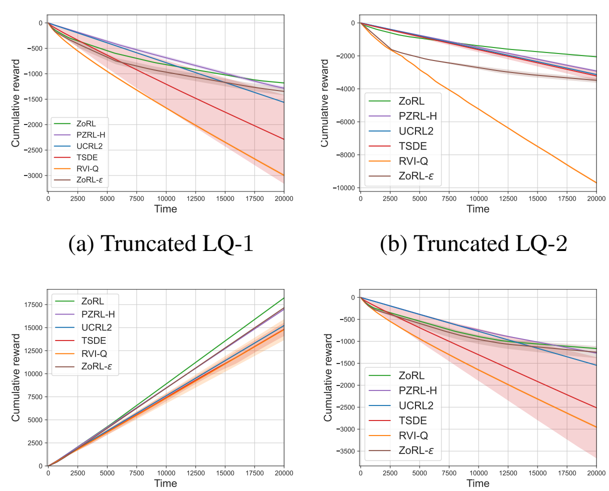
Step 5: The term (a) can be written as the sum of the regrets arising due to playing policies from the sets $\Phi_{2^{-i}}$, where $i = 1, 2, \ldots, \lceil\log\left(\frac{1}{\epsilon}\right)\rceil$, where $\epsilon = T^{-\frac{1}{2d_{\mathcal{S}}+d_z+3}}$. To bound the regret arising due to playing 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 $\lceil\log\left(\frac{1}{\epsilon}\right)\rceil$, and this gives us upper bound on (a) on a set with probability at least $1 - \delta$.

Bounding (b): Upper bound on the term (b) relies on the geometric ergodicity property (Meyn and Tweedie, 2012) of $\mathcal{M}$, that has been shown in Proposition A.1. Proposition F.4 shows that we must pay a constant penalty each time we change policy, which is $\mathcal{O}(K(T))$. 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).

The complement of the high probability set contributes δT to the regret. Taking $\delta = T^{-\frac{1}{2d_{\mathcal{S}}+d_z+3}}$, and summing the upper bounds on (a) and (b) along with δT, we obtain the desired regret bound. $\square$

5 Simulations

We compare the performance of ZoRL (Algorithm 2) 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) Linear Quadratic (LQ) control systems (Abbasi-Yadkori and Szepesvári, 2011) where the state evolves as $s_{t+1} = As_t + Ba_t + w_t$. A and B are matrices of appropriate dimensions. 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. (ii) Continuous RiverSwim, where the environment models an agent who is swimming in a river. (iii) Non-linear System where the state evolves as $s_{t+1} = Af(s_t) + Bg(a_t) + w_t$. A and B are matrices of appropriate dimensions and 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 (2024b), and also in Appendix G. We plot the cumulative rewards in Figure 2. ZoRL performs the best among all five algorithms on each of the environments.



(a) Truncated LQ-1 (b) Truncated LQ-2

(c) Continuous RiverSwim (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 supports the theoretical finding. ZoRL overperforms the popular fixed discretization-based algorithms as well as existing adaptive discretization-based algorithms.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. Advances in neural information processing systems, 24:2312–2320.

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

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

Auer, P. (2000). Using upper confidence bounds for online learning. In Proceedings 41st Annual Symposium on Foundations of Computer Science, pages 270–279. IEEE.

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

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

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

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

Domingues, O. D., Menard, P., Pirotta, M., Kaufmann, E., and Valko, M. (2021). Kernel-based reinforcement learning: A finite-time analysis. In Meila, M. and Zhang, T., editors, Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 2783–2792. PMLR.

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

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. (2018). Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In International Conference on Machine Learning, pages 1578–1586. PMLR.

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

He, J., Zhong, H., and Yang, Z. (2023). Sample-efficient learning of infinite-horizon average-reward mdps with general function approximation. In The Twelfth International Conference on Learning Representations.

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

Ibarz, J., Tan, J., Finn, C., Kalakrishnan, M., Pastor, P., and Levine, S. (2021). How to train your robot with deep reinforcement learning: lessons we have learned. The International Journal of Robotics Research, 40(4-5):698–721.

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

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In Conference on Learning Theory, pages 2137–2143. PMLR.

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

Kar, A. and Singh, R. (2024a). Adaptive discretization-based non-episodic reinforcement learning in metric spaces. arXiv preprint arXiv:2405.18793.

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

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

Kong, D. and Yang, L. (2022). Provably feedback-efficient reinforcement learning via active reward learning. Advances in Neural Information Processing Systems, 35:11063–11078.

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

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. Advances in applied mathematics, 6(1):4–22.

Maran, D., Metelli, A. M., Papini, M., and Restell, M. (2024a). No-regret reinforcement learning in smooth mdps. The Forty-first International Conference on Machine Learning.

Maran, D., Metelli, A. M., Papini, M., and Restelli, M. (2024b). Projection by convolution: Optimal sample complexity for reinforcement learning in continuous-space mdps. The 37th Annual Conference on Learning Theory.

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

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

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

Osband, I. and Van Roy, B. (2014). Model-based reinforcement learning and the eluder dimension. arXiv preprint arXiv:1406.1853.

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

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

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

Shah, D. and Xie, Q. (2018). Q-learning with nearest neighbors. Advances in Neural Information Processing Systems, 31.

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

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

Sodhi, P., Wu, F., Elenberg, E. R., Weinberger, K. Q., and McDonald, R. (2023). On the effectiveness of offline rl for dialogue response generation. In International Conference on Machine Learning, pages 32088–32104. PMLR.

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

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

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

Tossou, A., Basu, D., and Dimitrakakis, C. (2019). Near-optimal optimistic reinforcement learning using empirical bernstein inequalities. arXiv preprint arXiv:1905.12425.

Van Der Vaart, A. W., Wellner, J. A., van der Vaart, A. W., and Wellner, J. A. (1996). Weak convergence. Springer.

Wang, S., Blanchet, J., and Glynn, P. (2023). Optimal sample complexity for average reward markov decision processes. arXiv preprint arXiv:2310.08833.

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

Wu, Y., Zhou, D., and Gu, Q. (2022). 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.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes.**
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes.**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **Yes.**

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes.**
   (b) Complete proofs of all theoretical results. **Yes.**
   (c) Clear explanations of any assumptions. **Yes.**

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **Yes.**
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes.**
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Yes.**

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Yes.**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. **Not Applicable.**
   (b) The license information of the assets, if applicable. **Not Applicable.**
   (c) New assets either in the supplemental material or as a URL, if applicable. **Yes.**
   (d) Information about consent from data providers/curators. **Not Applicable.**
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not Applicable.**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. **Not Applicable.**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not Applicable.**
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not Applicable.**

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

Organization of the Appendix. Appendix A discusses some properties of MDPs that satisfy Assumption 2.3 and also proves Lemma 4.1. Appendix B derives concentration results for estimates of the discretized model. Appendix C derives some important properties of `ScOpt` when it is applied to solve the extended MDP (17). These properties hold true with a high probability. Appendix D proves that the extended MDP is nearly optimistic and not overly optimistic. Appendix E derives results that are used in Appendix F to produce a high probability lower bound on the number of visits to the key cells in each episode. We bound the regret in Theorem 4.2 Appendix F. Details of experiments and the associated environments are reported in Appendix G. Appendix H derives some key results that are used in the proof of Lemma B.1. Appendix I contains some known results that are used in this paper.

A General Results for MDPs

In this section, we discuss some important results on MDPs satisfying Assumption 2.3 and prove Lemma 4.1. Recall the AROE:

$$J + h(s) = \max_{a\in\mathcal{A}} \left\{ r(s,a) + \int_S h(s')\, p(s,a,s')\, ds' \right\}. \tag{28}$$

Under Assumption 2.3, there exists a solution to the AROE, and also, an optimal policy $\phi^\star \in \Phi_{SD}$ such that $\phi^\star(s)$ maximizes the r.h.s. of the AROE (Hernández-Lerma, 2012). In the first part of this section, we show a candidate solution to the AROE, prove the uniqueness of the solution under an additional condition, and prove a property of the solution.

Consider the controlled Markov process (CMP) $\{s_t\}$ described by the transition kernel p that evolves under the application of a stationary policy ϕ. The following result can be found in Arapostathis et al. (1993); Hernández-Lerma (2012). We provide the proof for the sake of completeness.

Proposition A.1. *Let the transition kernel $p : \mathcal{S}\times\mathcal{A}\times\mathcal{B}_\mathcal{S} \to [0,1]$ satisfies Assumption 2.3, i.e.,*

$$\|p(s,a,\cdot) - p(s',a',\cdot)\|_{TV} \le 2\alpha, \tag{29}$$

where $\alpha \in (0,1)$. Then, under the application of each stationary deterministic policy $\phi : \mathcal{S} \mapsto \mathcal{A}$, the controlled Markov process $\{s_t\}$ has a unique invariant distribution, denoted by $\mu^{(\infty)}_{\phi,p}$. Moreover $\{s_t\}$ is geometrically ergodic, i.e., for all $s \in \mathcal{S}$ the following holds,

$$\left\|\mu^{(t)}_{\phi,p,s} - \mu^{(\infty)}_{\phi,p}\right\|_{TV} \le 2\alpha^t,\ t\in\mathbb{N}, \tag{30}$$

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

Proof. Consider the CMP that is described by the transition kernel p and evolves under the application of policy ϕ, and consider two copies of that CMP with different initial state distributions, $\mu^{(0)}_1$ and $\mu^{(0)}_2$. Denote the distributions of s_t in the corresponding processes by $\mu^{(t)}_1$ and $\mu^{(t)}_2$, respectively. We show the following:

$$\left\|\mu^{(t)}_1 - \mu^{(t)}_2\right\|_{TV} \le \alpha^t \left\|\mu^{(0)}_1 - \mu^{(0)}_2\right\|_{TV},\ t\in\mathbb{N}. \tag{31}$$

We consider only deterministic policies. The proof for randomized policies goes in a similar line. Note that,

$$\begin{aligned}
\left\|\mu^{(1)}_1 - \mu^{(1)}_2\right\|_{TV} &= 2\sup_{A\subseteq\mathcal{S}} \left\{(\mu^{(1)}_1 - \mu^{(1)}_2)(A)\right\} \\
&= 2\sup_{A\subseteq\mathcal{S}} \left\{\int_\mathcal{S} p(s,\phi(s),A)(\mu^{(0)}_1 - \mu^{(0)}_2)(s)ds\right\} \\
&\le \sup_{\substack{A\subseteq\mathcal{S}\\ s,s'\in\mathcal{S}}} \{p(s,\phi(s),A) - p(s',\phi(s'),A)\} \left\|\mu^{(0)}_1 - \mu^{(0)}_2\right\|_{TV} \\
&\le \alpha \left\|\mu^{(0)}_1 - \mu^{(0)}_2\right\|_{TV},
\end{aligned}$$

where the first step follows from the definition of the total variation norm, while the third and the fourth step follow from Lemma I.6 and from Assumption 2.3, respectively. Applying the same argument recursively, we obtain (31). To see the existence of an invariant distribution, consider $\mu^{(0)}_2 = \mu^{(s)}_1$, $s\in\mathbb{N}$. Note that for all $n, s \in \mathbb{N}$,

$$\left\|\mu^{(t)}_1 - \mu^{(t+s)}_1\right\|_{TV} \le 2\alpha^t.$$

Hence, $\{\mu^{(t)}_1\}_t$ is a Cauchy sequence in the space of probability measures on $\mathcal{S}$ and attains a limit by the completeness of the space of probability measures. To see the uniqueness of the invariant distribution, assume that $\mu^{(\infty)}_i$, $i = 1, 2$, are two invariant distributions of the CMP. Let the initial distributions of the two processes be $\mu^{(\infty)}_1$ and $\mu^{(\infty)}_2$. Then, by the definition of invariant distribution, the distributions of the states in these two processes at time t must be $\mu_1^{(\infty)}$ and $\mu_2^{(\infty)}$, respectively, for every $t \in \mathbb{N}$. But, from (31), for all $t \in \mathbb{N}$,

$$\begin{aligned} \left\|\mu_1^{(t)} - \mu_2^{(t)}\right\|_{TV} &= \left\|\mu_1^{(\infty)} - \mu_2^{(\infty)}\right\|_{TV} \\ &\leq \alpha^t \left\|\mu_1^{(\infty)} - \mu_2^{(\infty)}\right\|. \end{aligned}$$

Taking limit $t \to \infty$, we have that $\mu_1^{(\infty)} = \mu_2^{(\infty)}$. Hence, the uniqueness of the invariant distribution of the CMP is established. To show the last part of the claim, we take $\mu_1^{(0)} = \mu_{\phi,p}^{(\infty)}$ and $\mu_2^{(0)} = \delta_s$. Then, (30) follows from (31). This completes the proof. □

Given a policy $\phi \in \Phi_{SD}$ and a reference state $s_\star$, consider the following set of equations:

$$\begin{aligned} &J + h(s) = r(s, \phi(s)) + \int_{\mathcal{S}} h(s')p(s, \phi(s), s')ds', \text{ and} \\ &h(s_\star) = 0. \end{aligned}$$

We show that the solution to the above equations is unique in the next lemma.

Lemma A.2. *Consider a policy $\phi \in \Phi_{SD}$ and a reference state $s_\star \in \mathcal{S}$. Then, $(J_{\mathcal{M}}(\phi), h_{\mathcal{M}}^{\phi})$ is the solution to the following equations,*

$$J + h(s) = r(s, \phi(s)) + \int_{\mathcal{S}} h(s')p(s, \phi(s), s')ds', \textit{ and} \tag{32a}$$

$$h(s_\star) = 0, \tag{32b}$$

where

$$\begin{aligned} J_{\mathcal{M}}(\phi) &= \liminf_{T\to\infty} \frac{1}{T} \mathbb{E}_\phi \left[\sum_{t=0}^{T-1} r(s_t, a_t)\right], \textit{ and} \\ h_{\mathcal{M}}^{\phi} &= \sum_{t=0}^{\infty} \int_{\mathcal{S}} \left(\mu_{\phi,p,s'}^{(t)} - \mu_{\phi,p,s_\star}^{(t)}\right) r(s, \phi(s))ds. \end{aligned}$$

Moreover, $sp\left(h_{\mathcal{M}}^{\phi}\right) \leq \frac{1}{1-\alpha}$.

Proof. Let (J_1, h_1) and (J_2, h_2) be two different solutions. Consider the function $h_1 - h_2$. Note that

$$(h_1 - h_2)(s) = -(J_1 - J_2) + \int_{\mathcal{S}} (h_1 - h_2)(s')p(s, \phi(s), s')ds'. \tag{33}$$

Then, it follows from Lemma I.6 that $sp\,(h_1 - h_2) \leq \alpha \; sp\,(h_1 - h_2)$. Since $\alpha < 1$, $sp\,(h_1 - h_2) = 0$. This is possible only if $h_1 = h_2 + c$ for some non-zero constant $c \in \mathbb{R}$. However, this cannot hold true because we have $h_1(s_\star) = h_2(s_\star) = 0$. Uniqueness of J follows trivially from here.

Now, see that $(J_1, h_1) \in \mathbb{R} \times \mathbb{R}^{\mathcal{S}}$ is a candidate solution of (32a) and (32b) where,

$$\begin{aligned} J_1 &= \int_{\mathcal{S}} \mu_{\phi,p}^{(\infty)}(s) r(s, \phi(s))ds, \text{ and,} \\ h_1(s') &= \sum_{t=0}^{\infty} \int_{\mathcal{S}} \left(\mu_{\phi,p,s'}^{(t)}(s) - \mu_{\phi,p,s_\star}^{(t)}(s)\right) r(s, \phi(s))ds. \end{aligned}$$

Next, we show that $sp\left(h_{\mathcal{M}}^{\phi}\right) \leq \frac{1}{1-\alpha}$.

$$\begin{aligned}
sp\left(h_{\mathcal{M}}^{\phi}\right) = &\ sp\left(\sum_{t=0}^{\infty}\int_{\mathcal{S}}\left(\mu_{\phi,p,\cdot}^{(t)} - \mu_{\phi,p,s_\star}^{(t)}\right)(s) r(s,\phi(s))ds\right) \\
\leq &\sum_{t=0}^{\infty} sp\left(\int_{\mathcal{S}}\left(\mu_{\phi,p,\cdot}^{(t)} - \mu_{\phi,p,s_\star}^{(t)}\right)(s) r(s,\phi(s))ds\right) \\
\leq &\sum_{t=0}^{\infty} \max_{s_1}\int_{\mathcal{S}}\left(\mu_{\phi,p,s_1}^{(t)} - \mu_{\phi,p,s_\star}^{(t)}\right)(s) r(s,\phi(s))ds \\
\leq &\ \frac{1}{2}\sum_{t=0}^{\infty}\max_{s_1,s_2}\left\|\mu_{\phi,p,s_1}^{(t)} - \mu_{\phi,p,s_2}^{(t)}\right\|_{TV} sp\left(r\right) \\
\leq &\ \frac{1}{2}\sum_{t=0}^{\infty}\alpha^t\max_{s_1,s_2}\left\|\mu_{\phi,p,s_1}^{(0)} - \mu_{\phi,p,s_2}^{(0)}\right\|_{TV} \\
= &\ \frac{1}{2}\sum_{t=0}^{\infty}2\alpha^t \\
= &\ \frac{1}{1-\alpha},
\end{aligned}$$

where the third inequality follows from Lemma I.6 and the fourth inequality follows from (31). It remains to prove $J_{\mathcal{M}}(\phi) = J_1$. To see this,

$$\begin{aligned}
J_{\mathcal{M}}(\phi) &= \liminf_{T\to\infty}\frac{1}{T}\mathbb{E}_{\phi}\left[\sum_{t=0}^{T-1} r(s_t,a_t)\right] \\
&= \liminf_{T\to\infty}\frac{1}{T}\left[\sum_{t=0}^{T-1}\int_{\mathcal{S}}\mu_{\phi,p,s_0}^{(t)}(s) r(s,\phi(s)ds)\right] \\
&= \liminf_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\int_{\mathcal{S}}\mu_{\phi,p}^{(\infty)}(s) r(s,\phi(s)ds) + \liminf_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\int_{\mathcal{S}}\left(\mu_{\phi,p,s_0}^{(t)} - \mu_{\phi,p}^{(\infty)}\right)(s) r(s,\phi(s)ds) \\
&= \int_{\mathcal{S}}\mu_{\phi,p}^{(\infty)}(s) r(s,\phi(s)ds),
\end{aligned}$$

where the last step follows from the fact that under Assumption 2.3, $\sum_{t=0}^{T-1}\int_{\mathcal{S}}\left(\mu_{\phi,p,s_0}^{(t)} - \mu_{\phi,p}^{(\infty)}\right)(s) r(s,\phi(s))ds$ is bounded. Hence, $(J_{\mathcal{M}}(\phi), h_{\mathcal{M}}^{\phi})$ is the unique solution to (32a) and (32b). $\square$

Corollary A.3. *Consider an arbitrarily chosen reference state $s_\star \in \mathcal{S}$. Then, $(J_{\mathcal{M}}^\star, h_{\mathcal{M}})$ is the solution to the following equations,*

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

where

$$\begin{aligned}
J_{\mathcal{M}}^\star &= \sup_{\phi}\lim_{T\to\infty}\frac{1}{T}\mathbb{E}_{\phi}\left[\sum_{t=0}^{T-1} r(s_t,a_t)\right], \\
h_{\mathcal{M}} &= \sum_{t=0}^{\infty}\int_{\mathcal{S}}\left(\mu_{\phi^\star,p,s'}^{(t)} - \mu_{\phi^\star,p,s_\star}^{(t)}\right)(s) r(s,\phi^\star(s))ds,
\end{aligned}$$

and $\phi^\star(s)$ satisfies,

$$\phi^\star(s) \in \arg\max_{a\in\mathcal{A}}\left\{r(s,a) + \int_{\mathcal{S}} h_{\mathcal{M}}^{\phi^\star}(s')p(s,a,s')ds'\right\}. \tag{34}$$

Moreover, $sp\left(h_{\mathcal{M}}\right) \leq \frac{1}{1-\alpha}$.

Proof. Proof follows from Lemma A.2 and from the fact that there exists an optimal policy $\phi^\star$ that satisfies the following equation:

$$r(s, \phi^\star(s)) + \int_{\mathcal{S}} h^{\phi}_{\mathcal{M}}(s') p(s, \phi^\star(s), s') ds' = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \int_{\mathcal{S}} h^{\phi^\star}_{\mathcal{M}}(s') p(s,a,s') ds' \right\},$$

which is proved in Hernández-Lerma (2012). □

Remark. *All of the above results hold true for discrete MDPs as well.*

The following result shows that under Assumption 2.3, any discretized stochastic kernel obtained from p is geometrically ergodic.

Proposition A.4. *Consider a state partition $\mathcal{Q}$ that is made of $\mathcal{S}$-cells and a set of state-action pairs Z. The discretized stochastic kernel $\wp(\cdot, \cdot; Z, \mathcal{Q})$ (11) satisfies,*

$$\left\| \wp(z, \cdot; Z, \mathcal{Q}) - \wp(z', \cdot; Z, \mathcal{Q}) \right\|_{TV} \leq 2\alpha, \ \forall z, z' \in Z.$$

Proof. Consider any two state-action pairs $z, z' \in Z$. Then,

$$\begin{aligned} \left\| \wp(z, \cdot; Z, \mathcal{Q}) - \wp(z', \cdot; Z, \mathcal{Q}) \right\|_{TV} &= 2 \sup_{B \subseteq \mathcal{Q}} \left\{ \sum_{\xi \in B} p(z, \xi) - p(z', \xi) \right\} \\ &\leq \sup_{B \in \mathcal{B}_{\mathcal{S}}} \left\{ p(z, B) - p(z', B) \right\} \\ &\leq 2\alpha, \end{aligned} \tag{35}$$

where the first inequality follows from the fact that every $\mathcal{S}$-cells are Borel measurable sets, and the last inequality follows from Assumption 2.3. □

A.1 Proof of Lemma 4.1

Let P be a kernel, i.e., $P(s)$ is a finite measure on $(\mathcal{S}, \mathcal{B}_{\mathcal{S}})$ for every $s \in \mathcal{S}$. Let h be a $\mathbb{R}$-valued function on $\mathcal{S}$. Then,

$$\langle P, h \rangle (s) := \int_{\mathcal{S}} h(s') P(s, s') ds'. \tag{36}$$

Also, let P^n denote the n-composition of P. Let I be the identity kernel, i.e., for every $h : \mathcal{S} \to \mathbb{R}$, $\langle I, h \rangle (s) = h(s)$.

Proof. From Lemma A.2 and Corollary A.3, we get that for every $s \in \mathcal{S}$,

$$\begin{aligned} J^\star_{\mathcal{M}} - J_{\mathcal{M}}(\phi) + h_{\mathcal{M}}(s) - h^{\phi}_{\mathcal{M}}(s) &= \max_{a \in \mathcal{A}} \left\{ r(s,a) + \int_{\mathcal{S}} p(s,a,s') h_{\mathcal{M}}(s') ds' \right\} - r(s, \phi(s)) \\ &\quad - \int_{\mathcal{S}} p(s, \phi(s), s') h_{\mathcal{M}}(s') ds' + \int_{\mathcal{S}} p(s, \phi(s), s') (h_{\mathcal{M}}(s') - h^{\phi}_{\mathcal{M}}(s')) ds' \\ &= \text{gap}\,(s, \phi(s)) + \int_{\mathcal{S}} p(s, \phi(s), s') (h_{\mathcal{M}}(s') - h^{\phi}_{\mathcal{M}}(s')) ds'. \end{aligned} \tag{37}$$

We can write this more compactly as follows,

$$(J^\star_{\mathcal{M}} - J_{\mathcal{M}}(\phi)) \cdot \mathbf{1} + h_{\mathcal{M}} - h^{\phi}_{\mathcal{M}} = \text{gap}^{\phi} + \left\langle p^{\phi}, (h_{\mathcal{M}} - h^{\phi}_{\mathcal{M}}) \right\rangle, \tag{38}$$

where $\mathbf{1} : \mathcal{S} \to \{1\}$ and $\text{gap}^{\phi} : \mathcal{S} \to \mathbb{R}$ be the function such that $\text{gap}^{\phi}(s) = \text{gap}\,(s, \phi(s))$ for every $s \in \mathcal{S}$. p^{ϕ} be the transition probability kernel induced by policy ϕ. i.e., $p^{\phi}(s, \cdot) = p(s, \phi(s), \cdot)$. From (38), we obtain

$$\begin{aligned}
(J^{\star}_{\mathcal{M}} - J_{\mathcal{M}}(\phi_{\mathcal{M}})) \cdot \mathbf{1} &= \text{gap}^{\phi} + \left\langle p^{\phi} - I, h_{\mathcal{M}} - h^{\phi}_{\mathcal{M}} \right\rangle \\
&= \text{gap}^{\phi} + \left\langle p^{\phi} - I, \text{gap}^{\phi} + \left\langle p^{\phi}, h_{\mathcal{M}} - h^{\phi}_{\mathcal{M}} \right\rangle - (J^{\star}_{\mathcal{M}} - J_{\mathcal{M}}(\phi)) \cdot \mathbf{1} \right\rangle \\
&= \left\langle p^{\phi}, \text{gap}^{\phi} \right\rangle + \left\langle \left(p^{\phi}\right)^2 - p^{\phi}, h_{\mathcal{M}} - h^{\phi}_{\mathcal{M}} \right\rangle \\
&= \left\langle p^{\phi}, \text{gap}^{\phi} \right\rangle + \left\langle \left(p^{\phi}\right)^2 - p^{\phi}, \text{gap}^{\phi} + \left\langle p^{\phi}, h_{\mathcal{M}} - h^{\phi}_{\mathcal{M}} \right\rangle - (J^{\star}_{\mathcal{M}} - J_{\mathcal{M}}(\phi)) \cdot \mathbf{1} \right\rangle \\
&= \left\langle \left(p^{\phi}\right)^2, \text{gap}^{\phi} \right\rangle + \left\langle \left(p^{\phi}\right)^3 - \left(p^{\phi}\right)^2, h_{\mathcal{M}} - h^{\phi}_{\mathcal{M}} \right\rangle.
\end{aligned}$$

Proceeding like this, we obtain

$$(J^{\star}_{\mathcal{M}} - J_{\mathcal{M}}(\phi)) \cdot \mathbf{1} = \left\langle \left(p^{\phi}\right)^t, \text{gap}^{\phi} \right\rangle + \left\langle \left(p^{\phi}\right)^{t+1} - \left(p^{\phi}\right)^t, h - h^{\phi} \right\rangle, \ \forall t \in \mathbb{N}. \tag{39}$$

We already know from Proposition A.1, that $\left(p^{\phi}\right)^t$ attains a limit geometrically fast with a rate of α. This implies that

$$\lim_{t \to \infty} \left\langle \left(p^{\phi}\right)^{t+1} - \left(p^{\phi}\right)^t, h - h^{\phi} \right\rangle = 0,$$

and also, by the same ergodicity property,

$$\lim_{t \to \infty} \left\langle \left(p^{\phi}\right)^t, \text{gap}^{\phi} \right\rangle = \int_{\mathcal{S}} \text{gap}^{\phi}(s) \mu^{(\infty)}_{\phi,p}(s) ds \cdot \mathbf{1}.$$

Hence, taking limit $t \to \infty$ in (39), we have that,

$$J^{\star}_{\mathcal{M}} - J_{\mathcal{M}}(\phi) = \int_{\mathcal{S}} \text{gap}^{\phi}(s) \mu^{(\infty)}_{\phi,p}(s) ds.$$

This concludes the proof. □

B 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 $Z \subseteq \mathcal{S} \times \mathcal{A}$, and $\mathcal{Q}$ be a partition of $\mathcal{S}$ that is made of $\mathcal{S}$-cells. Similar to (11), define the discrete transition kernel,

$$\wp(z, \cdot; Z, \mathcal{Q}) := p(z, q^{-1}(s)), \ \forall z \in Z, s \in \{q(\xi) : \xi \in \mathcal{Q}\}. \tag{40}$$

Denote the continuous extension of $\wp(z, \cdot; Z, \mathcal{Q})$ by $\bar{p}(z, \cdot; Z, \mathcal{Q})$, i.e.,

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

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

$$\mathcal{G}_1 := \cap_{t=0}^{T-1} \left\{ \|\wp(z', \cdot; \mathcal{S} \times \mathcal{A}, \mathcal{Q}_t) - \hat{\wp}_t(z, \cdot; Z_t, \mathcal{Q}_t)\|_1 \le \left(1 - \frac{1 - \alpha}{C_\eta}\right) \eta_t(\zeta) \text{ for every } z \in Z_t, z' \in q^{-1}(z) \right\}. \tag{41}$$

See that $\cap_{t=0}^{T-1} \{\wp(\cdot, \cdot; \mathcal{S} \times \mathcal{A}, \mathcal{Q}_t) \in \mathcal{C}_t(\mathcal{S} \times \mathcal{A})\} \subset \mathcal{G}_1$, where $\mathcal{C}_t(\mathcal{S} \times \mathcal{A})$ is as defined in (16). We show that $\mathcal{G}_1$ holds with a high probability.

Lemma B.1. *$\mathbb{P}(\mathcal{G}_1) \ge 1 - \frac{\delta}{2}$, where $\mathcal{G}_1$ is as in* (41).

Proof. Fix t, and consider a point $z \in Z_t$. Within this proof, we denote $q_t^{-1}(z)$ by ζ. Let ζ be a cell of level ℓ and note that ζ is an active cell at time t. Let z' be an arbitrary point in ζ. We want to get a high probability bound on $\|\hat{\wp}_t(z,\cdot;Z_t,\mathcal{Q}_t) - \wp(z',\cdot;\mathcal{S}\times\mathcal{A},\mathcal{Q}_t)\|_1$. We can write the following:

$$
\begin{aligned}
&\|\hat{\wp}_t(z,\cdot;\mathcal{Z}_t,\mathcal{Q}_t) - \wp(z',\cdot;\mathcal{S}\times\mathcal{A},\mathcal{Q}_t)\|_1\\
&= \|\hat{p}_t(z,\cdot) - \bar{p}(z',\cdot;\mathcal{S}\times\mathcal{A},\mathcal{Q}_t)\|_{TV}\\
&\le \left\|\hat{p}_t(z,\cdot) - \bar{p}(z',\cdot;\mathcal{S}\times\mathcal{A},\mathcal{Q}^{(\ell)})\right\|_{TV} + \left\|\bar{p}(z',\cdot;\mathcal{S}\times\mathcal{A},\mathcal{Q}^{(\ell)}) - \bar{p}(z',\cdot;\mathcal{S}\times\mathcal{A},\mathcal{Q}_t)\right\|_{TV}\\
&\le \left\|\hat{p}^{(d)}_t(z,\cdot) - \wp(z',\cdot;\mathcal{S}\times\mathcal{A},\mathcal{Q}^{(\ell)})\right\|_1 + \left\|\bar{p}(z',\cdot;\mathcal{S}\times\mathcal{A},\mathcal{Q}^{(\ell)}) - \bar{p}(z',\cdot;\mathcal{S}\times\mathcal{A},\mathcal{Q}_t)\right\|_{TV}.
\end{aligned}
\tag{42}
$$

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

$$
\left\|p(z',\cdot;\mathcal{S}\times\mathcal{A},\mathcal{Q}^{(\ell)}) - p(z',\cdot;\mathcal{S}\times\mathcal{A},\mathcal{Q}_t)\right\|_{TV} \le C_v \,\mathrm{diam}\,(\zeta)\,.
$$

Next, we will produce a high probability upperbound of the first term of r.h.s. of (42). We will denote $\wp(z',\cdot;\mathcal{S}\times\mathcal{A},\mathcal{Q}^{(\ell)})$ by $p^{(d)}_t(z',\cdot)$ in order to simplify the notation. Note that both $\hat{p}^{(d)}_t(z,\cdot)$ and $p^{(d)}_t(z',\cdot)$ has support $\tilde{S}_t(z)$ and $|\tilde{S}_t(z)| \le D^{d_\mathcal{S}}\mathrm{diam}\,(\zeta)^{-d_\mathcal{S}}$. Let $\tilde{S}^+_t(z)$ denote the collection of points of S_t such that for any $s \in \tilde{S}^+_t(z)$, we have $\hat{p}^{(d)}_t(z,s) - p^{(d)}_t(z',s) > 0$. So, we can write the following:

$$
\begin{aligned}
\mathbb{P}\left(\left\|\hat{p}^{(d)}_t(z,\cdot) - p^{(d)}_t(z',\cdot)\right\|_1 \ge \iota\right) &= \mathbb{P}\left(\max_{S'\subset\tilde{S}^+_t(z)} \sum_{s\in S'} \hat{p}^{(d)}_t(z,s) - p^{(d)}_t(z',s) \ge \frac{\iota}{2}\right)\\
&= \mathbb{P}\left(\cup_{S'\subset\tilde{S}^+_t(z)}\left\{\sum_{s\in S'} \hat{p}^{(d)}_t(z,s) - p^{(d)}_t(z',s) \ge \frac{\iota}{2}\right\}\right).
\end{aligned}
\tag{43}
$$

Note that if $S' \subset \tilde{S}^+_t(z)$, then $\tilde{S}_t(z)\setminus S' \not\subset \tilde{S}^+_t(z)$. Hence the number of subsets of $\tilde{S}^+_t(z)$ is at most $2^{|\tilde{S}_t(z)|-1}$. If $\mathbb{P}\left(\sum_{s\in S'} \hat{p}^{(d)}_t(z,s) - p^{(d)}_t(z',s) \ge \frac{\iota}{2}\right) \le b_\iota,\ \forall S' \subset \tilde{S}^+_t(z)$, then by an application of union bound on (43), we obtain that the following must hold,

$$
\mathbb{P}\left(\left\|\hat{p}^{(d)}_t(z,\cdot) - p^{(d)}_t(z',\cdot)\right\|_1 \ge \iota\right) \le 2^{|\tilde{S}_t(z)|-1} b_\iota. \tag{44}
$$

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

$$
v_i(z) := \mathbb{1}_{\{(s_i,a_i)\in\zeta_i\}}, \tag{45}
$$

$$
v_i(z,\xi) := \mathbb{1}_{\{(s_i,a_i,s_{i+1})\in\zeta_i\times\xi\}}, \tag{46}
$$

$$
w_i(z,\xi) := v_i(z,\xi) - p(s_i,a_i,\xi)v_i(z), \tag{47}
$$

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

$$
\begin{aligned}
\sum_{s\in S'} \hat{p}^{(d)}_t(z,s) - p^{(d)}_t(z',s) &= \frac{N_t\,(\zeta,\xi)}{N_t\,(\zeta)} - p(z',\xi)\\
&= \frac{N_t\,(\zeta,\xi) - p(z',\xi)N_t\,(\zeta)}{N_t\,(\zeta)}\\
&\le \frac{1}{N_t\,(\zeta)}\left(\sum_{i=0}^{t-1} w_i(z,\xi)\right) + \frac{L_p}{2N_t\,(\zeta)}\sum_{i=0}^{N_t(\zeta)} \mathrm{diam}\,(\zeta_{t_i})\\
&\le \frac{1}{N_t\,(\zeta)}\left(\sum_{i=0}^{t-1} w_i(z,\xi)\right) + 1.5L_p\,\mathrm{diam}\,(\zeta)\,,
\end{aligned}
\tag{48}
$$

where the last step follows from Lemma H.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, $|w_i(z,\zeta)| \le 1$. Hence from Lemma I.1 we have,

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

which when combined with (48) yields,

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

Upon using (44) in the above, and taking a union bound over all possible values of N, we obtain,

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

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

$$\mathbb{P}\left(\left\|\hat{p}_t^{(d)}(z,\cdot)-p_t^{(d)}(z',\cdot)\right\|_1\geq\operatorname{diam}(\zeta)^{-\frac{d_S}{2}}\sqrt{\frac{2\,D^{d_S}\log\left(\frac{2T}{\delta}\right)}{N_t(\zeta)}}+3L_p\operatorname{diam}(\zeta)\right)\leq\frac{\delta}{2}. \tag{49}$$

Let $\mathcal{N}_1:=2D^d\left(\frac{T}{c_1\log(T/\delta)}\right)^{\frac{d}{d_S+2}}$, 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 $1-\frac{\delta}{2}$, the following holds,

$$\left\|\hat{p}_t^{(d)}(z,\cdot)-p_t^{(d)}(z',\cdot)\right\|_1\leq 3\,\left(\frac{c_1\log\left(\frac{T}{\delta}\right)}{N_t(\zeta)}\right)^{\frac{1}{d_S+2}}+3L_p\operatorname{diam}(\zeta), \tag{50}$$

for every $z\in\zeta,\zeta\in\mathcal{P}_t$, and $t\in\{0,1,\ldots,T-1\}$, where c_1 be a constant such that $D^{d_S}\log\left(\frac{2T\mathcal{N}_1}{\delta}\right)=4.5c_1\log\left(\frac{T}{\delta}\right)$. The proof follows upon combining the upper bounds of the first and the second terms of (42). □

C Discussion on ScOpt

The next lemma shows that there exists an MDP in $\mathcal{M}_t^+$ (52) that is geometrically ergodic.

Lemma C.1. *Fix any $t\in\mathbb{N}$, and consider the MDP $\mathcal{M}_t^+$ (52). On the set $\mathcal{G}_1$, there exists at least one MDP $(S_t,A_t,\theta,\tilde{r}_t)\in\mathcal{M}_t^+$ such that*

$$\|\theta(s,a,\cdot)-\theta(s',a',\cdot)\|_1\leq 2\left(\alpha+\frac{1-\alpha}{T}\right),$$

for every $(s,a),(s',a')\in S_t\times A_t$.

Proof. From Lemma B.1, we have that on the set $\mathcal{G}_1$, for every $z\in Z_t$, $\|\wp(z',\cdot;S_t\times A_t,\mathcal{Q}_t)-\hat{\wp}_t(z,\cdot;Z_t,\mathcal{Q}_t)\|_1\leq(1-(1-\alpha)C_\eta^{-1})\eta_t(q^{-1}(z))$ where $z'\in S_t\times A_t$ belongs to $q^{-1}(z)$. Let $\theta_\wp$ denote the the projection of $\wp(\cdot,\cdot;S_t\times A_t,\mathcal{Q}_t)$ on $\{\theta\in\Theta(S_t\times A_t,\mathcal{Q}_t)\mid\theta(s,a,s')\geq\frac{1-\alpha}{T|S_t|},\ \forall(s,a,s')\in S_t\times A_t\times S_t\}$. See that $\|\theta_\wp(z',\cdot)-\wp(z',\cdot;S_t\times A_t,\mathcal{Q}_t)\|_1\leq\frac{1-\alpha}{T}<(1-\alpha)\operatorname{diam}\left(q^{-1}(z)\right)$. The last inequality holds because $q^{-1}(z)$ is active at time $t<T$. Hence, $\|\theta_\wp(z',\cdot)-\hat{\wp}_t(z',\cdot;S_t\times A_t,\mathcal{Q}_t)\|_1\leq\eta_t(q^{-1}(z))$, and $(S_t,A_t,\theta_\wp,\tilde{r}_t)\in\mathcal{M}_t^+$.

Let us fix $(s,a),(s',a')\in S_t\times A_t$. See that,

$$\begin{aligned}\|\theta_\wp(s,a,\cdot)-\theta_\wp(s',a',\cdot)\|_1&\leq\|\theta_\wp(s,a,\cdot)-\hat{\wp}_t(s,a,\cdot;S_t\times A_t,\mathcal{Q}_t)\|_1\\&\quad+\|\hat{\wp}_t(s,a,\cdot;S_t\times A_t,\mathcal{Q}_t)-\hat{\wp}_t(s',a',\cdot;S_t\times A_t,\mathcal{Q}_t)\|_1\\&\quad+\|\hat{\wp}_t(s',a',\cdot;S_t\times A_t,\mathcal{Q}_t)-\theta_\wp(s',a',\cdot)\|_1\\&\leq 2\left(\alpha+\frac{1-\alpha}{T}\right).\end{aligned}$$

This concludes the proof of the Lemma. □

Next, we prove two properties of the `ScOpt` algorithm when applied to solve the optimization problem (17).

Lemma C.2. *Consider the MDP $\mathcal{M}_t^+$, where $t \in \{0, 1, \dots, T-1\}$. Consider $(v, s) \in \mathbb{R}^{S_t} \times S_t$, such that $sp(v) \le c$, where c is as defined in* (18)*. Then, on the set $\mathcal{G}_1$, the operator Γ_c* (20) *corresponding to $\mathcal{M}_t^+$ is feasible at (v, s). Moreover, on the set $\mathcal{G}_1$, for every v with $sp(v) \le c$, the policy $G_c v$* (22) *is deterministic.*

Proof. Consider the Bellman operator corresponding to Let us fix $v \in \mathbb{R}^{S_t}$ such that $sp(v) \le c$. Let s_ℓ be the state corresponding to the minimum value of $\mathcal{T}v$, i.e.,

$$(\mathcal{T}v)(s_\ell) = \min_{s \in S_t} \max_{\substack{a \in A_t(s) \\ \theta \in \mathcal{C}}} \Big\{ \tilde{r}(s, a) + \sum_{s' \in S_t} \theta(s, a, s') v(s') \Big\}.$$

See that on the set $\mathcal{G}_1$,

$$\min_{s \in S_t} \max_{a \in A_t(s)} \Big\{ \tilde{r}(s, a) + \sum_{s' \in S_t} \theta_\wp(s, a, s') v(s') \Big\} \le (\mathcal{T}v)(s_\ell),$$

where $\theta_\wp(s, a, \cdot)$ is the projection of $\wp(s, a, \cdot; S_t \times A_t, \mathcal{Q}_t)$ on the set $\{\theta \in \Theta(S_t \times A_t, \mathcal{Q}_t) \mid \theta(s, a, s') \ge \frac{1-\alpha}{T|S_t|}, \forall (s, a, s') \in S_t \times A_t \times S_t\}$. Now, let us consider a state $\bar{s}$ for which

$$\max_{\substack{a \in A_t(\bar{s}) \\ \theta \in \mathcal{C}_t(S_t \times A_t)}} \Big\{ \tilde{r}(\bar{s}, a) + \sum_{s' \in S_t} \theta(\bar{s}, a, s') v(s') \Big\} \ge (\mathcal{T}v)(s_\ell) + c.$$

From Lemma C.1, we have that $\theta_\wp$ satisfies geometric ergodicity with ergodicity coefficient $\alpha + \frac{1-\alpha}{T}$. If $sp(v) \le c$, span of $\max_{a \in A(\cdot)} \Big\{ \tilde{r}(\cdot, a) + \sum_{s' \in S_t} \theta_\wp(\cdot, a, s') v(s') \Big\}$ is bounded by c where $c = \frac{1+L_r}{(1-\alpha)(1-T^{-1})}$. To see that,

$$\begin{aligned} sp\left(\max_{a \in A(\cdot)} \Big\{ \tilde{r}(\cdot, a) + \sum_{s' \in S_t} \theta_\wp(\cdot, a, s') v(s') \Big\} \right) &\le sp(\tilde{r}) + \frac{1}{2} \max_{(s,a),(s',a')} \|\theta_\wp(s, a) - \theta_\wp(s', a')\|_{TV}\ sp(v) \\ &\le c. \end{aligned}$$

Hence,

$$\begin{aligned} \max_{a \in A_t(\bar{s})} \Big\{ \tilde{r}(\bar{s}, a) + \sum_{s' \in S_t} \theta_\wp(\bar{s}, a, s') v(s') \Big\} &\le \min_{s \in S_t} \max_{a \in A_t(s)} \Big\{ \tilde{r}(s, a) + \sum_{s' \in S_t} \theta_\wp(s, a, s') v(s') \Big\}(s_\ell) + c \\ &\le (\mathcal{T}v)(s_\ell) + c. \end{aligned}$$

Hence, on the set $\mathcal{G}_1$, (v, s) is feasible for every $s \in S_t$. As there exists a stochastic kernel $\bar{\theta}$ for which

$$\Gamma_c v(s) = \max_{a \in A_t(s)} \Big\{ \tilde{r}(s, a) + \sum_{s' \in S_t} \bar{\theta}(s, a, s') v(s') \Big\},$$

and hence, $(G_c v)(s) = \arg\max_{a \in A_t(s)} \Big\{ \tilde{r}(s, a) + \sum_{s' \in S_t} \bar{\theta}(s, a, s') v(s') \Big\}$ is a deterministic policy. This concludes the proof of the lemma. □

D Properties of the Extended MDPs

Let us define the set of discrete MDPs at time t as follows,

$$\mathcal{M}_t^\dagger = \{(S_t, A_t, \tilde{p}, \tilde{r}_t) : \tilde{p} \in \mathcal{C}_t(S_t \times A_t)\},$$

where S_t and 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. $\tilde{p}$ belongs to a set of plausible discrete transition kernels. Also define the following extended MDP optimization problem:

$$\max_{M \in \mathcal{M}_t^\dagger} \max_{\phi \in \Phi_t(M, c)} J_M(\phi). \tag{51}$$

Solving (51) in a computationally efficient way is difficult. So, we slightly modify (51) to (17). Recall the set of discrete MDPs $\mathcal{M}_t^+$, defined in Section 3.

$$\mathcal{M}_t^+ = \{(S_t, A_t, \tilde{p}, \tilde{r}) : \tilde{p} \in \mathcal{C}_t(S_t \times A_t), \tilde{p}(s,a,s') \geq \frac{1-\alpha}{|S_t|T},\ \forall (s,a,s') \in S_t \times A_t \times S_t\}, \tag{52}$$

where $\tilde{p}$ belongs to a set of plausible discrete transition kernels which have the property that the outgoing transition probability from each (s,a) to each discrete state is at least $\frac{1-\alpha}{|S_t|T}$. The constraint $\tilde{p}(s,a,s') \geq \frac{1-\alpha}{|S_t|T}$ has been added in order that the corresponding operator Γ_c (20) becomes a contraction map. This helps us to conclude that `ScOpt` (1) converges but at the expense of an error of $\mathcal{O}(1/T)$ (Fruit et al., 2018, Theorem 11).

We denote the optimal value of (51) by $J^{\star,c}_{\mathcal{M}_t^\dagger}$. The operator Γ_c corresponding to the extended MDP (51) is not guaranteed to be a contraction map, and hence `ScOpt` (1) is not guaranteed to converge when applied to solve (51). We will now show that on the set $\mathcal{G}_1$, the solution to the extended MDP (51) is optimistic (Auer, 2000).

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

$$J^{\star,c}_{\mathcal{M}_t^\dagger} \geq J^\star_{\mathcal{M}} \textit{ for every } t \in \mathbb{N}, \tag{53}$$

where $J^{\star,c}_{\mathcal{M}_t^\dagger}$ is the optimal value of (17) *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 $s \in \mathcal{S}$,

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

We know that $\mathcal{M}$ is geometrically ergodic, and hence the following value iteration algorithm converges in the sense that $\lim_{n\to\infty} sp\,(V_{n+1} - V_n) = J^\star_{\mathcal{M}}$. 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 $h_{\mathcal{M}}$ is bounded (this is shown in Corollary A.3), it then follows that $\lim_{n\to\infty} \frac{1}{n} V_n(s) = J^\star_{\mathcal{M}}$.

Recall the operator Γ_c (20) for the MDP, $\mathcal{M}_t^\dagger$, i.e.,

$$\begin{aligned} \bar{V}_0(s) &= 0, \\ \bar{V}_{n+1}(s) &= (\Gamma_c \bar{V}_n)(s) \\ &= \min \Bigg\{ \max_{\substack{a \in A_t(s) \\ \theta \in \mathcal{C}_t(S_t \times A_t)}} \left\{ \tilde{r}(s,a) + \sum_{s' \in S} \theta(s,a,s') \bar{V}_n(s') \right\}, \\ &\qquad \min_{\tilde{s} \in S} \left\{ \max_{\substack{\tilde{a} \in A_t(s) \\ \theta \in \mathcal{C}_t(S_t \times A_t)}} \left\{ \tilde{r}(\tilde{s},\tilde{a}) + \sum_{s' \in S} \theta(\tilde{s},\tilde{a},s') \bar{V}_n(s') \right\} + c \right\} \Bigg\}, \end{aligned}$$

where $c = \frac{1+L_r}{1-\alpha}$. From Fruit et al. (2018, Theorem 21) and from Lemma C.2, it follows that on the set $\mathcal{G}_1$ we have $\lim_{n\to\infty} sp\left(\bar{V}_{n+1} - \bar{V}_n\right) = J^\star_{\mathcal{M}_t^\dagger,c}$ and $\lim_{n\to\infty} \frac{1}{n} \bar{V}_n(s) = J^\star_{\mathcal{M}_t^\dagger,c}$ for every $s \in S_t$.

We will prove that $\bar{V}_n(s) \geq V_n(s')$ for every $n \in \mathbb{N}$, $s \in S_t$ and $s' \in q^{-1}(s)$. We will show this via induction. The base cases (for $i = 0$) hold trivially. Next, assume that the following hold for all $i \in [n]$, where $n \in \mathbb{N}$,

$$\bar{V}_i(s) \geq V_i(s'),\ \forall s \in S_t,\ \forall s' \in q^{-1}(s). \tag{54}$$

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

$$\begin{aligned} r(s,a) + \int_{\mathcal{S}} p(s,a,s') V_n(s') ds' &\leq r(s,a) + \sum_{s' \in S_t} \wp(s,a,s'; S_t \times A_t, \mathcal{Q}_t) \bar{V}_n(s') \\ &\leq r(q(\zeta)) + L_r \mathrm{diam}\,(\zeta) + \sum_{s' \in S_t} \wp(s,a,s'; S_t \times A_t, \mathcal{Q}_t) \bar{V}_n(s') \end{aligned}$$

$$\leq \max_{\substack{\tilde{a}\in A_t(\tilde{s})\\ \theta\in\mathcal{C}_t(S_t\times A_t)}} \left\{\tilde{r}(\tilde{s},\tilde{a}) + \sum_{s'\in S_t} \theta(\tilde{s},\tilde{a},s')\bar{V}_n(s')\right\}, \tag{55}$$

where the first inequality follows from (54), the second inequality follows from Assumption 2.2(i), while the third inequality follows from the definition of the set $\mathcal{G}_1$. Note that $\wp(\cdot,\cdot;S_t\times A_t,\mathcal{Q}_t)$ satisfies geometric ergodicity property (4), and consequently we have,

$$\tilde{r}(\tilde{s},\tilde{a}) + \sum_{s'\in S_t} \wp(s,a,s';S_t\times A_t,\mathcal{Q}_t)\bar{V}_n(s') \leq \min_{\tilde{s}\in S} \max_{\substack{\tilde{a}\in A_t(s)\\ \theta\in\mathcal{C}_t(S_t\times A_t)}} \left\{\tilde{r}(\tilde{s},\tilde{a}) + \sum_{s'\in S} \theta(\tilde{s},\tilde{a},s')\bar{V}_n(s')\right\} + c. \tag{56}$$

Upon combining (55) and (56) we obtain,

$$r(s,a) + \int_{\mathcal{S}} p(s,a,s')V_n(s') \leq \bar{V}_{n+1}(\tilde{s}), \tag{57}$$

for every $\tilde{s}\in S_t$ and for every $(s,a)\in q^{-1}(\tilde{s})\times\mathcal{A}$. Taking maximum over all actions, we get $V_n(s)\leq \bar{V}_n(\tilde{s})$ for every $n\in\mathbb{N}$, $\tilde{s}\in S_t$ and $s\in q^{-1}(\tilde{s})$. The proof is then completed by dividing both sides of this inequality by n, and then taking limit $n\to\infty$. □

The following corollary shows that the extended MDP (17) that is described in Section 3, is "almost optimistic," i.e. its optimal value is an upper-bound on $J^{\star}_{\mathcal{M}}$ modulo an "error term" that arises since we had introduced the constraint that $\mathcal{M}^{+}_t$ contains MDPs with transition kernel $\tilde{p}$ such that $\tilde{p}(s,a,s')\geq \frac{1-\alpha}{|S_t|T}$ $\forall (s,a,s')$.

Corollary D.2. *On the set $\mathcal{G}_1$, we have,*

$$J^{\star,c}_{\mathcal{M}^{+}_t} \geq J^{\star}_{\mathcal{M}} - \frac{1+L_r}{|S_t|(T-1)}, \text{ for every } t\in\mathbb{N},$$

where $J^{\star,c}_{\mathcal{M}^{+}_t}$ is the optimal value of (17) *and $J^{\star}_{\mathcal{M}}$ is the optimal average reward of the MDP $\mathcal{M}$.*

Proof. From Theorem 11 of Fruit et al. (2018), we have $J^{\star,c}_{\mathcal{M}^{+}_t} \geq J^{\star,c}_{\mathcal{M}^{\dagger}_t} - \frac{1+L_r}{|S_t|(T-1)}$. The proof then follows from Lemma D.1. □

Fix time t. Consider a policy $\phi\in\Phi_{SD}$ such that for every $\xi\in\mathcal{Q}_t$, $\phi(s)=a$ for all $s\in\xi$ and $a\in A_t(q(\xi))$. Let us call such policies as permissible policies at time t. Consider the following iterations for evaluating the optimistic index of a permissible policy ϕ at time t.

$$\begin{aligned}
&\bar{V}^{\phi}_0(s) = 0,\\
&\bar{V}^{\phi}_{n+1}(s) = \min\Bigg\{ \max_{\theta\in\mathcal{C}_t(S_t\times A_t)} \left\{\tilde{r}(s,\phi(s)) + \sum_{s'\in S} \theta(s,\phi(s),s')\bar{V}_n(s')\right\},\\
&\qquad \min_{s\in S_t}\left\{ \max_{\theta\in\mathcal{C}_t(S_t\times A_t)} \left\{\tilde{r}(s,\phi(s)) + \sum_{s'\in S} \theta(s,\phi(s),s')\bar{V}_n(s')\right\}\right\} + c\Bigg\},\ \forall n\in\mathbb{N}.
\end{aligned} \tag{58}$$

Define the index of a policy as $J_{\mathcal{M}^{+}_t,c}(\phi) := \lim_{n\to\infty}\frac{1}{n}\bar{V}^{\phi}_n(s)$ for any $s\in S_t$. Also, consider the policy evaluation algorithm for the true MDP which as follows:

$$\begin{aligned}
&V^{\phi}_0(s) = 0,\\
&V^{\phi}_{n+1}(s) = r(s,\phi(s)) + \int_{\mathcal{S}} p(s,\phi(s),s')V_n(s')ds',\ \forall n\in\mathbb{N}.
\end{aligned} \tag{59}$$

Note that $\lim_{n\to\infty}\frac{1}{n}V^{\phi}_n(s) = J_{\mathcal{M}}(\phi)$ for any $s\in\mathcal{S}$. The next result establishes a relation between $J_{\mathcal{M}}(\phi)$ and $J_{\mathcal{M}^{+}_t,c}(\phi)$, which is the index of ϕ at time t.

Lemma D.3. *Consider time $t \in \mathbb{N}$ and a policy $\phi \in \Phi_{SD}$ that is permissible at time t. Then, we have that on the set $\mathcal{G}_1$,*

$$\max_{M \in \mathcal{M}_t^+} J_M(\phi) \leq J_{\mathcal{M}}(\phi) + C_1 \, diam_t(\phi), \tag{60}$$

where

$$C_1 := 3L_r + \frac{2(3 + 3L_p + C_v)}{1 - \alpha}. \tag{61}$$

Proof. Denote $q_t^{-1}(s, \phi(s))$ by $q_t^{-1}(s, \phi(s))$ within this proof. In order to prove this result, we will show that on the set $\mathcal{G}_1$, for every $n \in \mathbb{N}$, for every $s \in S_t$ and for every $s' \in q^{-1}(s)$, the following holds,

$$\bar{V}_n^{\phi}(s) \leq V_n^{\phi}(s') + C_1 \, \mathbb{E}_{p,\phi}\left[\sum_{i=0}^{n-1} \text{diam}\left(q_t^{-1}(s_i, \phi(s_i))\right) \middle| s_0 = s'\right], \tag{62}$$

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 $i \in [n]$, where $n \in \mathbb{N}$,

$$\bar{V}_i^{\phi}(s) \leq V_i^{\phi}(s') + C_1 \mathbb{E}_{p,\phi}\left[\sum_{j=0}^{i-1} \text{diam}\left(q_t^{-1}(s_j, \phi(s_j))\right) \middle| s_0 = s'\right], \tag{63}$$

for every $s \in S_t$ and for every $s' \in q^{-1}(s)$. Let us fix $s \in S_t$ and $s' \in q^{-1}(s)$ arbitrarily, then from (58) we obtain the following,

$$\begin{aligned}
\bar{V}_{n+1}^{\phi}(s) &= r(q(q_t^{-1}(s, \phi(s)))) + \max_{\theta \in \mathcal{C}_t(S_t \times A_t)} \sum_{s'' \in S_t} \theta(q(q_t^{-1}(s, \phi(s))), s'') \bar{V}_n^{\phi}(s'') + L_r \, \text{diam}\left(q_t^{-1}(s, \phi(s))\right) \\
&= r(q(q_t^{-1}(s, \phi(s)))) + \sum_{s'' \in S_t} \theta_n(q(q_t^{-1}(s, \phi(s))), s'') \bar{V}_n^{\phi}(s'') + L_r \, \text{diam}\left(q_t^{-1}(s, \phi(s))\right) \\
&\leq r(s', \phi(s')) + \sum_{s'' \in S_t} \wp(s', \phi(s'), s''; S_t \times A_t, \mathcal{Q}_t) \, \bar{V}_n^{\phi}(s'') + \eta_t(q_t^{-1}(s, \phi(s))) \, sp\left(\bar{V}_n^{\phi}\right) + 2L_r \, \text{diam}\left(q_t^{-1}(s, \phi(s))\right) \\
&\leq r(s', \phi(s')) + \int_{\mathcal{S}} p(s', \phi(s'), s'') V_n^{\phi}(s'') ds'' + C_1 \, \mathbb{E}_{p,\phi}\left[\sum_{i=1}^{n} \text{diam}\left(q_t^{-1}(s_i, \phi(s_i))\right) \middle| s_0 = s'\right] \\
&\quad + \left(\frac{C_\eta(1 + L_r)}{1 - \alpha} + 2L_r\right) \text{diam}\left(q_t^{-1}(s, \phi(s))\right) \\
&\leq r(s', \phi(s')) + \int_{\mathcal{S}} p(s', \phi(s'), s'') V_n^{\phi}(s'') ds'' + C_1 \, \mathbb{E}_{p,\phi}\left[\sum_{i=1}^{n} \text{diam}\left(q_t^{-1}(s_i, \phi(s_i))\right) \middle| s_0 = s'\right] \\
&\quad + \left(3L_r + \frac{2C_\eta}{1 - \alpha}\right) \text{diam}\left(q_t^{-1}(s, \phi(s))\right) \\
&= V_{n+1}^{\phi}(s) + C_1 \, \mathbb{E}_{p,\phi}\left[\sum_{i=0}^{n} \text{diam}\left(q_t^{-1}(s_i, \phi(s_i)))\right) \middle| s_0 = s\right],
\end{aligned}$$

where θ_n is a transition kernel belonging to the set $\mathcal{C}_t(S_t \times A_t)$ 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 I.6. The second inequality is obtained by invoking the induction hypothesis (63), and by using the upper bound on $sp\left(\bar{V}_i^{\phi}\right)$. This concludes the induction argument, and proves (62). The proof of the claim follows by dividing both side of (62) by n and taking limit $n \to \infty$. □

Corollary D.4. *Consider time $t \in \mathbb{N}$ and a policy $\phi \in \Phi_{SD}$ that is permissible at time t. Then, we have that on the set $\mathcal{G}_1$,*

$$\max_{M \in \mathcal{M}_t^{\dagger}} J_M(\phi) \leq J_{\mathcal{M}}(\phi) + C_1 \, diam_t(\phi) - \frac{1 + L_r}{|S_t|(T - 1)},$$

where

$$C_1 := 3L_r + \frac{2(3+3L_p+C_v)}{1-\alpha}. \tag{64}$$

Proof. Proof follows from Theorem 11 of Fruit et al. (2018), and from Lemma D.3. $\square$

E Guarantee on Number of Visits to Cells

Recall that $\mu^{(t)}_{\phi,p,s}$ 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 $\mu^{(\infty)}_{\phi,p}$ 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 $\mu^{(\infty)}_{\phi,p}(\xi) \geq (\epsilon/3)^{d_{\mathcal{S}}+1}$ for all stationary deterministic policies ϕ, where $\epsilon > 0$. Later we will choose an appropriate value for ϵ. From Proposition A.1 we get that for all $\phi \in \Phi_{SD}$ and for every initial state $s \in \mathcal{S}$ we have,

$$\mu^{(t)}_{\phi,p,s}(\xi) \geq \mu^{(\infty)}_{\phi,p}(\xi) - \alpha^t.$$

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

$$\mu^{(t)}_{\phi,p,s}(\xi) \geq \frac{1}{2}\mu^{(\infty)}_{\phi,p}(\xi),\ \forall t \geq t^\star(\epsilon), \tag{65}$$

where,

$$t^\star(\epsilon) := \left\lceil \frac{\log\left(2\left(\frac{3}{\epsilon}\right)^{d_{\mathcal{S}}+1}\right)}{\log\left(\frac{1}{\alpha}\right)} \right\rceil \tag{66}$$

$\lceil x \rceil$ denotes the nearest integer to $x \in \mathbb{R}$ such that $x \leq \lceil x \rceil$.

Lemma E.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 $1-\frac{\delta}{2}$, we have,*

$$n_k(\zeta) \geq \frac{H_k\ \mu^{(\infty)}_{\phi,p}(\xi)}{2t^\star(\epsilon)} - \sqrt{\frac{H_k}{t^\star(\epsilon)}\log\left(\frac{4T}{t^\star(\epsilon)\delta}\right)} - 1.$$

Proof. Denote $m := \lfloor H_k/t^\star(\epsilon) \rfloor$ and $t_i := \tau_k + i\, t^\star(\epsilon)$. Let $i^\star \in \mathbb{Z}_+$ 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

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

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 $\{f_i\}_i$ is a $\{0,1\}$-valued, $\{\mathcal{F}_{t_i}\}$-predictable stochastic process. Hence, we can use Corollary I.4 and obtain,

$$\mathbb{P}\left(\sum_{i=1}^{m+1}\mathbb{1}_{\left\{s_{t_i}\in\xi\right\}} \leq \sum_{i=1}^{m+1}\mathbb{E}\left[\mathbb{1}_{\left\{s_{t_i}\in\xi\right\}} \mid \mathcal{F}_{t_{i-1}}\right] - \sqrt{\frac{m+2}{2}\log\left(\frac{2(m+2)}{\delta}\right)}\right) \leq \frac{\delta}{2}. \tag{67}$$

From (65), (66) we have that

$$\mathbb{E}\left[\mathbb{1}_{\left\{s_{t_{i-1}}\in\xi\right\}} \mid \mathcal{F}_{t_{i-1}}\right] \geq \frac{1}{2}\mu^{(\infty)}_{\phi,p}(\xi). \tag{68}$$

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 (68) and $m+2 \leq 2m$ in (67), we obtain,

$$\mathbb{P}\left(\sum_{i=1}^{m} \mathbb{1}_{\{s_{t_i}\in\xi\}} \leq \frac{H_k\, \mu_{\phi,p}^{(\infty)}(\xi)}{2t^\star(\epsilon)} - \sqrt{\frac{H_k}{t^\star(\epsilon)}\log\left(\frac{4H_k}{t^\star(\epsilon)\delta}\right)} - 1\right) \leq \frac{\delta}{2}.$$

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

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

$$\mathcal{G}_{2,\epsilon} := \left\{ n_k(\zeta) \geq \frac{H_k\, \mu_{\phi,p}^{(\infty)}(\xi)}{2t^\star(\epsilon)} - \sqrt{\frac{H_k}{t^\star(\epsilon)}\log\left(\frac{8T^2D^d}{t^\star(\epsilon)\epsilon^d\delta}\right)} - 1,\ \forall (k,\xi,\zeta) \text{ that satisfies the above conditions.} \right\},$$

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

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 $(D/\epsilon)^d$. So, the total number of possible combinations of (k,ξ,ζ) that satisfies the given condition is at most $T(D/\epsilon)^d$. The proof then follows from Lemma E.1 by taking a union bound over all (k,ξ,ζ) and by the fact that $H_k \leq T$. $\square$

F Regret Analysis

Regret decomposition: Recall the decomposition of regret (3) of learning algorithm ψ w.r.t. Φ,

$$\begin{aligned}\mathcal{R}(T;\psi) &= \sum_{k=1}^{K(T)} \sum_{t=\tau_k}^{\tau_{k+1}-1} J^\star_{\mathcal{M}} - r(s_t,a_t)\\ &= \underbrace{\sum_{k=1}^{K(T)} H_k\Delta(\phi_k)}_{(a)} + \underbrace{\sum_{k=1}^{K(T)} H_k\, J_{\mathcal{M}}(\phi_k) - \sum_{t=\tau_k}^{\tau_{k+1}-1} r(s_t,\phi_k(s_t))}_{(b)}. \end{aligned} \tag{69}$$

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): The regret (a) 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 all ϵ-optimal policies. To bound the regret arising due to policies from $\Phi^{(2^{-i})}$, we count the number of timesteps 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 every episode, ZoRL solves (17) with accuracy parameter $1/T$. This "loss of accuracy" as compared to the case where ZoRL could solve (51) accurately at the beginning of every episode, leads to an additional term in the upper bound of (a). Since the difference between the two solutions is at most $\mathcal{O}(1/T)$ (Fruit et al., 2018, Theorem 11), this term for each episode can be upper bounded as $\mathcal{O}(1)$. We derive the upper bound of (a) considering ZoRL solves (51).

The regret arising due to playing policies from the set $\Phi^{(2^{-i})}$ is bounded as follows. Lemma F.1 shows 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 D. Lemma F.2 gives a lower bound of the number of plays of a key cell in any episode by ZoRL using Lemma F.1, Corollary E.2, and Lemma I.5. Next, Lemma F.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 F.1. *On the set $\mathcal{G}_1$, there exists $s \in \mathcal{S}$ such that diam $\left(q_{\tau_k}^{-1}(s, \phi_k(s))\right) \geq \frac{1}{3C_1} \max\left\{gap\left(s, \phi_k(s)\right), C_1 diam_{\tau_k}(\phi_k)\right\}$ and $\mu_{\phi_k,p}^{(\infty)}(\pi_{\mathcal{S}}(q_{\tau_k}^{-1}(s, \phi_k(s)))) \geq (diam_{\tau_k}(\phi_k)/3)^{d_{\mathcal{S}}+1}$. Such a $q_{\tau_k}^{-1}(s, \phi_k(s))$ is called a key cell for the k-th episode.*

Proof. As `ZoRL` plays policies that agree with the partition $\mathcal{P}_{\tau_k}$ in the k-th episode, we restrict our discussion to consider only such policies. Let us fix $k \in \mathbb{N}$ and a policy ϕ that agrees with $\mathcal{P}_{\tau_k}$.We will first show that if

$$\text{diam}_{\tau_k}(\phi) \leq \Delta(\phi)/C_1, \tag{70}$$

then ϕ will not be played from episode k onwards. From Lemma D.1 we have that on the set $\mathcal{G}_1$, $J^{\star}_{\mathcal{M}^{\dagger}_{\tau_k},c} = \max_{M \in \mathcal{M}^{\dagger}_{\tau_k}} J_M(\phi_k) \geq J^{\star}_{\mathcal{M}}$. Hence, if $\max_{M \in \mathcal{M}^{\dagger}_{\tau_k}} J_M(\phi) < J^{\star}_{\mathcal{M}}$, then the algorithm will not play ϕ. From Lemma D.3 we have that on the set $\mathcal{G}_1$, $\max_{M \in \mathcal{M}^{\dagger}_{\tau_k}} J_M(\phi) \leq J_{\mathcal{M}}(\phi) + C_1 \text{diam}_{\tau_k}(\phi)$. Thus, on $\mathcal{G}_1$, ϕ will never be played from the k-th episode onwards if

$$J_{\mathcal{M}}(\phi) + C_1 \,\text{diam}_{\tau_k}(\phi) \leq J^{\star}_{\mathcal{M}},$$

or, if $\text{diam}_{\tau_k}(\phi) \leq \Delta(\phi)/C_1$. In other words, on the set $\mathcal{G}_1$,

$$\text{diam}_{\tau_k}(\phi_k) > \Delta(\phi_k)/C_1. \tag{71}$$

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

$$\text{diam}\left(q_{\tau_k}^{-1}(s, \phi_k(s))\right) \leq \frac{1}{3C_1} \max\left\{\text{gap}\left(s, \phi_k(s)\right), C_1 \text{diam}_{\tau_k}(\phi_k)\right\}. \tag{72}$$

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

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

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 $\xi \in \mathcal{Q}^{(3)}$, $\frac{1}{3}\text{diam}_{\tau_k}(\phi_k) \leq \text{diam}\left(q_{\tau_k}^{-1}(q(\xi), \phi_k(q(\xi)))\right) \leq \frac{1}{3C_1} \min_{s \in \zeta}\{\text{gap}\left(s, \phi_k(s)\right)\}$. Then,

$$\begin{aligned}
\text{diam}_{\tau_k}(\phi_k) &= \int_{\mathcal{S}} \text{diam}\left(q_{\tau_k}^{-1}(s, \phi_k(s))\right) \mu_{\phi_k,p}^{(\infty)}(s)\, ds\\
&= \sum_{\xi \in \mathcal{Q}_{\tau_k}} \text{diam}\left(q_{\tau_k}^{-1}(q(\xi), \phi_k(q(\xi)))\right) \mu_{\phi_k,p}^{(\infty)}(\xi)\\
&= \sum_{\xi \in \mathcal{Q}^{(1)}} \text{diam}\left(q_{\tau_k}^{-1}(q(\xi), \phi_k(q(\xi)))\right) \mu_{\phi_k,p}^{(\infty)}(\xi) + \sum_{\xi \in \mathcal{Q}^{(2)}} \text{diam}\left(q_{\tau_k}^{-1}(q(\xi), \phi_k(q(\xi)))\right) \mu_{\phi_k,p}^{(\infty)}(\xi)\\
&\quad + \sum_{\xi \in \mathcal{Q}^{(3)}} \text{diam}\left(q_{\tau_k}^{-1}(q(\xi), \phi_k(q(\xi)))\right) \mu_{\phi_k,p}^{(\infty)}(\xi)\\
&\leq \frac{\text{diam}_{\tau_k}(\phi_k)}{3} + \frac{\text{diam}_{\tau_k}(\phi_k)}{3} + \frac{1}{3C_1} \int_{\mathcal{S}} \text{gap}\left(s, \phi_k(s)\right) \mu_{\phi_k,p}^{(\infty)}(s)\, ds\\
&= \frac{\text{diam}_{\tau_k}(\phi_k)}{3} + \frac{\text{diam}_{\tau_k}(\phi_k)}{3} + \frac{\Delta(\phi_k)}{3C_1}\\
&< \text{diam}_{\tau_k}(\phi_k),
\end{aligned}$$

However, this clearly cannot be true. Hence, our assumption (72) was wrong. This concludes the proof. ∎

Henceforth within this section we will set $\epsilon = T^{-\frac{1}{2d_{\mathcal{S}}+d_z+3}}$. Let $\tilde{\epsilon} = T^{-\frac{1}{d_{\mathcal{S}}+d+3}}$. Note that $\epsilon \geq \tilde{\epsilon}$ since $d_z \leq d$. Also, note that

$$t^\star(\epsilon) = \left\lceil \frac{\log\left(2\left(\frac{3}{\epsilon}\right)^{d_{\mathcal{S}}+1}\right)}{\log\left(\frac{1}{\alpha}\right)} \right\rceil \leq \left\lceil \frac{\log\left(2\left(\frac{3}{\epsilon}\right)^{d_{\mathcal{S}}+1}\right)}{\log\left(\frac{1}{\alpha}\right)} \right\rceil = t^\star(\tilde{\epsilon}).$$

Lemma F.2. *Consider the event $\mathcal{G}_1 \cap \mathcal{G}_{2,\epsilon}$, and fix episode k such that $\Delta(\phi_k) \geq C_1\epsilon$. Let ξ be a key cell in episode k, i.e., for some $s \in \xi$,*

$$\begin{aligned}
&\textit{diam}\left(q_{\tau_k}^{-1}(s, \phi_k(s))\right) > \frac{1}{3C_1}\max\left\{\textit{gap}\left(s, \phi_k(s)\right), C_1 \textit{diam}_{\tau_k}(\phi_k)\right\}, \textit{ and} \\
&\mu^{(\infty)}_{\phi_k,p}(\xi) \geq (\textit{diam}_{\tau_k}(\phi_k)/3)^{d_{\mathcal{S}}+1}.
\end{aligned}$$

Then, $q_{\tau_k}^{-1}(q(\xi), \phi_k(q(\xi)))$ will be visited at least $\frac{C_H \log\left(\frac{T}{\delta}\right)}{4\cdot 9^{d_{\mathcal{S}}+1}\cdot t^\star(\epsilon)}\textit{diam}\left(\zeta\right)^{-(d_{\mathcal{S}}+1)}$ times in the k-th episode, where,

$$C_H := 16\, t^\star(\tilde{\epsilon})\, 9^{d_{\mathcal{S}}+1} \frac{\log\left(\frac{8T^2D^d}{t^\star(\epsilon)\tilde{\epsilon}^d\delta}\right)+1}{\log(T/\delta)}.$$

Proof. We note that $\mu^{(\infty)}_{\phi_k,p}(\xi) \geq (\epsilon/3)^{d_{\mathcal{S}}+1}$. Recall that on the set $\mathcal{G}_1$, $\text{diam}_{\tau_k}(\phi_k) \geq \frac{\Delta(\phi_k)}{C_1}$ (71). Hence, $\text{diam}_{\tau_k}(\phi_k) > \epsilon$. So, using Corollary E.2, we can write,

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

where $\zeta = q_{\tau_k}^{-1}(q(\xi), \phi_k(q(\xi)))$. Next, we note that H_k can be lower bounded as follows,

$$\begin{aligned}
H_k &= \frac{C_H \log\left(T/\delta\right)}{\underline{\text{diam}}_{\tau_k}(\phi_k)^{2(d_{\mathcal{S}}+1)}} \\
&\geq 16\, t^\star(\tilde{\epsilon})\left(\frac{3}{\text{diam}_{\tau_k}(\phi_k)}\right)^{2(d_{\mathcal{S}}+1)}\left(\log\left(\frac{8T^2D^d}{t^\star(\epsilon)\tilde{\epsilon}^d\delta}\right)+1\right) \\
&\geq \frac{16t^\star(\epsilon)}{\mu^{(\infty)}_{\phi_k,p}(\xi)^2}\left(\log\left(\frac{8T^2D^d}{t^\star(\epsilon)\epsilon^d\delta}\right)+1\right),
\end{aligned} \tag{73}$$

where the first inequality uses the fact that $C_H\log(T/\delta) = 16\ t^\star(\tilde{\epsilon})\ 9^{d_{\mathcal{S}}+1}\left(\log\left(\frac{8T^2D^d}{t^\star(\epsilon)\tilde{\epsilon}^d\delta}\right)+1\right)$ and $\underline{\text{diam}}_{\tau_k}(\phi_k) \leq \text{diam}_{\tau_k}(\phi_k)$. The second inequality follows from the fact that $\mu^{(\infty)}_{\phi_k,p}(\xi) \geq (\text{diam}_{\tau_k}(\phi_k)/3)^{d_{\mathcal{S}}+1}$. An application of Lemma I.5 and (73) yields

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

or,

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

where the second inequality follows from the fact that $\underline{\text{diam}}_{\tau_k}(\phi_k) \leq \text{diam}_{\tau_k}(\phi_k) < 3\ \text{diam}\left(\zeta\right)$. This concludes the proof of the lemma. □

Lemma F.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. Recall that $Z_\beta \subseteq \mathcal{S} \times \mathcal{A}$ is the set of state-action pairs (s,a) such that $\text{gap}\,(s,a) \le \beta$. Let us call the smallest subset of $\mathcal{P}_t$ that covers Z_β the active covering of Z_β at time t. We note that if for all $j = 0, 1, \dots, i$, the active covering of $Z_{2^{-j}}$ at time τ_k has no cell ζ with $\text{diam}\,(\zeta) \ge \frac{D}{3C_1}\, 2^{-j}$ and $\mu^{(\infty)}_{\phi,p}(\xi) \ge (\Delta(\phi)/3)^{d_{\mathcal{S}}+1}$ 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 $\Phi^{(2^{-i})}$. Hence, `ZoRL`will not play a policy from $\Phi^{(2^{-i})}$.

Let $\mathcal{Y}_j$ denote the set of cells of diameter $\frac{D}{3C_1}\, 2^{-j}$ which has non-empty intersection with $Z_{2^{-j}}$. Note that all possible key cells for polices from the set $\Phi^{(2^{-i})}$ either belong to $\cup_{j=0}^{i}\mathcal{Y}_j$ or an ancestor of some cell that belongs to $\cup_{j=0}^{i}\mathcal{Y}_j$. The cardinality of $\mathcal{Y}_j$ is at most $c_z 2^{j d_z}$, where the scaling constant of the zooming dimension, $c_s = \frac{D}{3C_1}$ (6). Hence, the cardinality of $\{\zeta \mid \zeta \in \mathcal{Y}_j, \text{ or } \zeta \text{ is an ancestor of } \zeta' \in \mathcal{Y}_j\}$ is at most $\mathcal{O}(\log(j)\, 2^{j d_z})$. Finally, by summing these upper bounds $\mathcal{O}(\log(j)\, 2^{j d_z})$ from $j=0$ to i, the set of all possible key cells for polices from the set $\Phi^{(2^{-i})}$ belongs to a set with cardinality at most $c_2 \log(i)\, 2^{i d_z}$, where c_2 is a constant.

Every time some policy $\phi \in \Phi^{(2^{-i})}$ is chosen to play, there is at least one cell, say ζ, from $\cup_{j=0}^{i}\mathcal{Y}_j$ that is active, and $\mu^{(\infty)}_{\phi,p}(\xi) \ge (\Delta(\phi)/3)^{d_{\mathcal{S}}+1}$ where $\xi \in \mathcal{Q}_{\tau_k}, \xi \subseteq \pi_{\mathcal{S}}(\zeta)$. Every episode has at least one key cell. From Lemma F.2 and the activation rule for cells (3.3), we deduce that `ZoRL` requires at most $4 \cdot 6^{d_{\mathcal{S}}+1} \frac{c_1 t^\star(\epsilon)}{C_H} \text{diam}\,(\zeta)^{-1}$ episodes to split ζ. Let $\text{diam}\,(\zeta) = \frac{D}{3C_1}\, 2^{-j}$. Then ζ will be played as a key cell in at most $72 \cdot 6^{d_{\mathcal{S}}} \frac{c_1 C_1 t^\star(\epsilon)}{C_H D} 2^j$.

So, total number of episodes where policies from $\Phi^{(2^{-i})}$ is played, is no more than

$$\sum_{j=0}^{i} \left[72 \cdot 6^{d_{\mathcal{S}}} \frac{c_1 C_1 t^\star(\epsilon)}{C_H D} 2^j\right] \times \left[c_2 \log(i) 2^{i d_z}\right] \le 72 \cdot 6^{d_{\mathcal{S}}} \frac{c_1 c_2 C_1 t^\star(\epsilon)}{C_H D} \log(i) 2^{i(d_z+1)}. \tag{74}$$

Next, we will bound the duration of those episodes where policies from $\Phi^{(2^{-i})}$ are played. Recall that $\underline{\text{diam}}_{\tau_k}(\phi_k) = \kappa_1 \int_{\mathcal{S}} \text{diam}\left(q^{-1}_{\tau_k}(s,\phi(s))\nu(s)ds\right)$ (26). A simple calculation yields $\text{diam}_{\tau_k}(\phi_k) \le \frac{\kappa_2}{\kappa_1}\underline{\text{diam}}_{\tau_k}(\phi_k)$. Let `ZoRL` has played a policy from $\Phi^{(2^{-i})}$ in the k-th episode. Then the duration of the k-th episode can be bounded as follows:

$$\begin{aligned}
H_k &= \frac{C_H \log(T/\delta)}{\underline{\text{diam}}_{\tau_k}(\phi_k)}^{2(d_{\mathcal{S}}+1)} \\
&\le \frac{C_H \log(T/\delta)}{\left(\frac{\kappa_1}{\kappa_2}\text{diam}_{\tau_k}(\phi_k)\right)^{2(d_{\mathcal{S}}+1)}} \\
&\le C_H \log(T/\delta) \left(\frac{\kappa_2 C_1}{\kappa_1 \Delta(\phi_k)}\right)^{2(d_{\mathcal{S}}+1)} \\
&\le C_H \left(\frac{\kappa_2 C_1}{\kappa_1}\right)^{2(d_{\mathcal{S}}+1)} \log\left(\frac{T}{\delta}\right) 2^{i \cdot 2(d_{\mathcal{S}}+1)},
\end{aligned} \tag{75}$$

where the second and the last inequalities follow from the fact that $\text{diam}_{\tau_k}(\phi_k) > \Delta(\phi_k)/C_1 \ge 2^{-i}/C_1$.

Combining the above two results, i.e., (74) and (75), we have that the total number of time steps where policies from $\Phi^{(2^{-i})}$ is played is at most $C' 2^{i(2d_{\mathcal{S}}+d_z+3)}$ where

$$C' = 144 \cdot 6^{d_{\mathcal{S}}} \frac{c_1 c_2 C_1 t^\star(\epsilon)}{D} \left(\frac{\kappa_2 C_1}{\kappa_1}\right)^{2(d_{\mathcal{S}}+1)} \log\left(\frac{T}{\delta}\right).$$

This concludes the proof. $\square$

As has been discussed earlier, we will derive an upper bound on (a) by summing regret due to playing policies from the set

$\Phi^{(2^{-i})}$ for $i = 1, 2, \ldots, \lceil \log 1/\epsilon \rceil$, and ϵT, i.e.,

$$\begin{aligned}\sum_{k=1}^{K(T)} H_k \Delta(\phi_k) &\leq C' \sum_{i=1}^{i^\star} 2^{i(2d_{\mathcal{S}}+d_z+3)} \times 2^{-i+1} + T^{\frac{2d_{\mathcal{S}}+d_z+2}{2d_{\mathcal{S}}+d_z+3}} \\ &\leq 2C' \, 2^{i^\star(2d_{\mathcal{S}}+d_z+2)} + T^{\frac{2d_{\mathcal{S}}+d_z+2}{2d_{\mathcal{S}}+d_z+3}} \\ &\leq (2C'+1) \, T^{\frac{2d_{\mathcal{S}}+d_z+2}{2d_{\mathcal{S}}+d_z+3}}, \end{aligned} \tag{76}$$

where the second step follows from Lemma F.3, and by taking $\epsilon = T^{-\frac{1}{2d_{\mathcal{S}}+d_z+3}}$.

Bounding (b): We will now provide an upper bound on the term (b) of (69). This proof will rely on the geometric ergodicity property (Meyn and Tweedie, 2012) of the underlying MDP $\mathcal{M}$, that has been shown in Proposition A.1. Recall that for a stationary policy ϕ and kernel p, $\mu^{(t)}_{\phi,p,s}$ denotes the distribution of the Markov chain at time t induced by applying ϕ on p when the start state $s_0 = s$, and $\mu^{(\infty)}_{\phi,p}$ denotes the corresponding stationary distribution.

Proposition F.4. *On the set $\mathcal{G}_1 \cap \mathcal{G}_{2,\epsilon}$, the term (b) in* (69) *can be bounded as follows,*

$$\sum_{k=1}^{K(T)} \left[H_k \, J_{\mathcal{M}}(\phi_k) - \mathbb{E}\left[\sum_{t=\tau_k}^{\tau_{k+1}-1} r(s_t, \phi_k(s_t)) \right] \right] \leq \frac{2K(T)}{1-\alpha}, \tag{77}$$

where α is as in Assumption 2.3.

Proof. Due to the tower property of conditional expectation and due to the fact that expectation is bounded above by the supremum of the corresponding random variable, it suffices to derive the upper bound for

$$\sum_{k=1}^{K(T)} \left[H_k \, J_{\mathcal{M}}(\phi_k) - \mathbb{E}\left[\sum_{t=\tau_k}^{\tau_{k+1}-1} r(s_t, \phi_k(s_t)) \mid \mathcal{F}_{\tau_k} \right] \right].$$

We derive an upper bound on the absolute value of each term in the above summation. We have,

$$\begin{aligned}\left| H_k \, J_{\mathcal{M}}(\phi_k) - \mathbb{E}\left[\sum_{t=\tau_k}^{\tau_{k+1}-1} r(s_t, \phi_k(s_t)) \mid \mathcal{F}_{\tau_k} \right] \right| &\leq \sum_{h=0}^{H_k-1} \left| \int_{\mathcal{S}} \left(\mu^{(\infty)}_{\phi_k,p}(s) - \mu^{(h)}_{\phi_k,p,s_0}(s) \right) r(s, \phi_k(s)) \right| \\ &\leq 2 \sum_{h=0}^{H_k-1} \alpha^h \\ &\leq 2 \sum_{h=0}^{\infty} \alpha^h \\ &\leq \frac{2}{1-\alpha}, \end{aligned} \tag{78}$$

where the second inequality follows from Proposition A.1 and Lemma I.6. The proof then follows by summing (78) over $k = 1, 2, \ldots, K(T)$. □

Upon combining the upper bounds on all the terms of the regret decomposition, we obtain the following upper bound on the regret.

F.1 Proof of Theorem 4.2

Proof. We first derive an upper bound on 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 episode k is less than $T^{\frac{2d_{\mathcal{S}}+2}{2d_{\mathcal{S}}+d_z+3}}$, it means that all the corresponding key cells have diameters greater than $const \cdot T^{-\frac{1}{2d_{\mathcal{S}}+d_z+3}}$. Note 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 2^ℓ 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. Now, summing all the upper bounds of regret components from (76) and (77), taking $\delta = T^{-\frac{1}{2d_{\mathcal{S}}+d_z+3}}$ and replacing $K(T)$ by its upper bound, we get that,

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

Thus, we have the desired regret upper bound that holds with a high probability. □

G Simulation Experiments

We provide details of the simulation environments corresponding to the experiments we have conducted. We use the following environments.

1. `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

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

We have used the following two sets of system parameters:

(a) `Truncated LQ-1`:

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

$P = 0.4\ I_2$[4], $Q = 0.6\ I_2$ 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`.

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

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

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

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

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 $[-1, 1]^{d_{\mathcal{A}}}$.

Discussion on Hyperparameters. We note that several constants such as L_r, L_p (Assumption 2.2), α (Assumption 2.3), C_v (Assumption 3.4), κ_1, κ_2 (Assumption 3.5) are not known. Hence, certain quantities that are functions of the above parameters and are used by `ZoRL`, are incorporated as hyperparameters. Those are as follows:

1. C_a: Recall that `ZoRL` deactivates the active cell ζ if $N_t(\zeta) = \frac{c_1 2^{d_{\mathcal{S}}+2}\log\left(\frac{T}{\delta}\right)}{\text{diam}(\zeta)^{d_{\mathcal{S}}+2}}$ (8). Since evaluating c_1 requires knowledge of the parameters discussed above, we replace this condition with the following condition $N_t(\zeta) = \frac{C_a}{\text{diam}(\zeta)^{d_{\mathcal{S}}+2}}$.
2. L_r: We assume knowledge of an upper-bound on L_r, the Lipschitz constant (Assumption 2.2).
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 $\eta_t(\zeta) \leq C_\eta \text{diam}(\zeta)$. 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: `ZoRL` uses $c = \frac{1+L_r}{(1-\alpha)(1-T^{-1})}$ (18) as an upper bound of the span of the iterated of `ScOpt`.
5. C_H: C_H is the multiplicative constant associated with the episode duration, i.e., $H_k = C_H \underline{\text{diam}}_{\tau_k}(\phi_k)^{-2(d_{\mathcal{S}}+1)}$. We use $\frac{1}{|S_{\tau_k}|}\sum_{s \in S_{\tau_k}} q_{\tau_k}^{-1}(s, \phi_k(s))$ as an approximation for $\underline{\text{diam}}_{\tau_k}(\phi_k)$ (26).

The values of the above hyperparameters for all four experiments are taken as follows: $C_a = 10$, $L_r = 0.01$, $C_\eta = 10$, $c = 4$, and $C_H = 0.1$.

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.

Codebase. The anonymized codebase can be found here: `https://anonymous.4open.science/r/zorl-D367/`.

H 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 $3\,\text{diam}(\zeta)$. We use this in concentration inequality for the transition kernel estimate.

Lemma H.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*

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

Proof. By the activation rule (3.3), a cell ζ' can be played at most $N_{\max}(\zeta') - N_{\min}(\zeta') = \tilde{c}_1 2^{2\ell(\zeta')} + \frac{\tilde{c}_1}{3}\mathbb{1}_{\{\zeta'=\mathcal{S}\times\mathcal{A}\}}$ times while being active, where $\tilde{c}_1 = 3c_1 D^{-2}\log\left(\frac{T}{\epsilon\delta}\right)\epsilon^{-d_{\mathcal{S}}}$. We can write,

$$
\begin{aligned}
\frac{1}{N_t(\zeta)}\sum_{i=1}^{N_t(\zeta)} \operatorname{diam}(\zeta_{t_i}) &= \frac{1}{N_t(\zeta)}\sum_{i=1}^{N_{\min}(\zeta)} \operatorname{diam}(\zeta_{t_i}) + \frac{1}{N_t(\zeta)}\sum_{i=N_{\min}(\zeta)+1}^{N_t(\zeta)} \operatorname{diam}(\zeta_{t_i}) \\
&= \frac{\tilde{c}_1 D}{3N_t(\zeta)} + \frac{\tilde{c}_1 D}{N_t(\zeta)}\sum_{\ell=0}^{\ell(\zeta)-1} 2^{\ell} + \frac{N_t(\zeta)-N_{\min}(\zeta)-1}{N_t(\zeta)}\operatorname{diam}(\zeta) \\
&< \frac{\tilde{c}_1 D}{N_t(\zeta)}2^{\ell(\zeta)} + \frac{N_t(\zeta)-N_{\min}(\zeta)-1}{N_t(\zeta)}\operatorname{diam}(\zeta) \\
&= \frac{3N_{\min}(\zeta)}{N_t(\zeta)}\operatorname{diam}(\zeta) + \frac{N_t(\zeta)-N_{\min}(\zeta)-1}{N_t(\zeta)}\operatorname{diam}(\zeta) \\
&= \frac{(N_t(\zeta)+2N_{\min}(\zeta)-1)\operatorname{diam}(\zeta)}{N_t(\zeta)} \\
&\le 3\operatorname{diam}(\zeta),
\end{aligned}
$$

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

Next, we show that under Assumption 3.4, the total variation norm between $p(z,\cdot;Z,\mathcal{Q})$ and $p(z,\cdot;Z,\mathcal{Q}^{(\ell)})$ is bounded above by the discretization width of the partition $\mathcal{Q}^{(\ell)}$, where $\mathcal{Q}^{(\ell)}$ is a coarser partition of $\mathcal{S}$ than $\mathcal{Q}$. We use this result in Lemma B.1.

Lemma H.2. *Consider a partition $\mathcal{Q}$ of the state space. Let $\mathcal{Q}$ be finer than $\mathcal{Q}^{(\ell)}$ for some $\ell \in \mathbb{N}$. Let $Z \subseteq \mathcal{S}\times\mathcal{A}$. Then, under Assumption 3.4, we have that*

$$
\left\|\bar{p}(z,\cdot;Z,\mathcal{Q}) - \bar{p}(z,\cdot;Z,\mathcal{Q}^{(\ell)})\right\|_{TV} \le C_v D 2^{-\ell}
$$

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

Proof. Let us fix $z \in Z$, and $\xi \in \mathcal{Q}^{(\ell)}$ 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,

$$
\begin{aligned}
\sup_{B\subseteq\xi}\left|\bar{p}(z,B;Z,\mathcal{Q}^{(\ell)}) - p(z,B)\right| &\le \int_{\xi}(f-\bar{f})\mathbb{1}_{\{f\ge\bar{f}\}}d\lambda \\
&\le \int_{\xi}(\bar{f}+C_v D\epsilon)\mathbb{1}_{\{f\ge\bar{f}\}}d\lambda - \int_{\xi}\bar{f}\mathbb{1}_{\{f\ge\bar{f}\}}d\lambda \\
&\le C_v D\epsilon \times \epsilon^{d_{\mathcal{S}}},
\end{aligned}
$$

where $\epsilon = 2^{-\ell}$. Hence, by Assumption 3.4, we have that for every $z \in \mathcal{S}\times\mathcal{A}$ and for every $\xi \in \mathcal{Q}^{(\ell)}$,

$$
\begin{aligned}
\sup_{B\subseteq\xi}\left|\bar{p}(z,B;Z,\mathcal{Q}^{(\ell)}) - \bar{p}(z,B;Z,\mathcal{Q})\right| &\le \sup_{B\subseteq\xi}\left|\bar{p}(z,B;Z,\mathcal{Q}^{(\ell)}) - p(z,B)\right| \\
&\le C_v D\epsilon \times \epsilon^{d_{\mathcal{S}}}.
\end{aligned}
$$

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

$$
\begin{aligned}
\left\|\bar{p}(z,\cdot;Z,\mathcal{Q}^{(\ell)}) - \bar{p}(z,\cdot;Z,\mathcal{Q})\right\|_{TV} &\le \sum_{\xi\in\mathcal{Q}^{(\ell)}}\sup_{B\subseteq\xi}\left|\bar{p}(z,B;Z,\mathcal{Q}^{(\ell)}) - p(z,B)\right| \\
&\le C_v D\epsilon \times \epsilon^{d_{\mathcal{S}}} \times \epsilon^{-d_{\mathcal{S}}} \\
&\le C_v D\epsilon.
\end{aligned}
$$

Hence, we have proven the claim. □

I Useful Results

I.1 Concentration Inequalities

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

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i \ge \epsilon\right\} \le e^{-\frac{\epsilon^2}{2nc^2}} \tag{79}$$

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

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

$$\mathbb{P}\left(\sum_{i=1}^{n} |X_i - Np_i| \ge 2\sqrt{N}\epsilon\right) \le 2^n e^{-2\epsilon^2}. \tag{80}$$

Alternatively, for $\delta > 0$

$$\mathbb{P}\left(\sum_{i=1}^{n} \left|\frac{X_i}{N} - p_i\right| < \sqrt{\frac{2n}{N}\log\left(\frac{2}{\delta^{\frac{1}{n}}}\right)}\right) \ge 1 - \delta. \tag{81}$$

The following is essentially Theorem 1 of Abbasi-Yadkori et al. (2011).

Theorem I.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.,*

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

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 $t \ge 0$, 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 $t \ge 0$,

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

Corollary I.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 $k \ge 0$,

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

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

I.2 Other Useful Results

Lemma I.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 $x \geq 1.5\frac{a_1}{a_0^2}$ $f(x) \geq 0$. $\square$

Lemma I.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

$$\lambda(Z) = \lambda(Z_+) + \lambda(Z_-) = 0. \tag{82}$$

Also,

$$\lambda(Z_+) - \lambda(Z_-) = \left\|\mu_1 - \mu_2\right\|_{TV}. \tag{83}$$

Combining the above two, we get that

$$\lambda(Z_+) = \frac{1}{2}\left\|\mu_1 - \mu_2\right\|_{TV}. \tag{84}$$

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. $\square$