Reinforcement Learning in Linear MDPs: Constant Regret and Representation Selection

Matteo Papini
Universitat Pompeu Fabra
matteo.papini@upf.edu

Andrea Tirinzoni
INRIA Lille
andrea.tirinzoni@inria.fr

Aldo Pacchiano*
Microsoft Research
apacchiano@microsoft.com

Marcello Restilli
Politecnico di Milano
marcello.restelli@polimi.it

Alessandro Lazaric
Facebook AI Research
lazaric@fb.com

Matteo Pirotta
Facebook AI Research
pirotta@fb.com

Abstract

We study the role of the representation of state-action value functions in regret minimization in finite-horizon Markov Decision Processes (MDPs) with linear structure. We first derive a necessary condition on the representation, called universally spanning optimal features (UniSOFT), to achieve constant regret in any MDP with linear reward function. This result encompasses the well-known settings of low-rank MDPs and, more generally, zero inherent Bellman error (also known as the Bellman closure assumption). We then demonstrate that this condition is also sufficient for these classes of problems by deriving a constant regret bound for two optimistic algorithms (LSVI-UCB and ELEANOR). Finally, we propose an algorithm for representation selection and we prove that it achieves constant regret when one of the given representations, or a suitable combination of them, satisfies the UniSOFT condition.

1 Introduction

The ability of an agent to learn an informative mapping from complex observations to a succinct representation is one of the essential factors for the success of machine learning in fields such as computer vision, language modeling, and more broadly in deep learning [Bengio et al., 2013]. In supervised learning, it is well understood that a "good" representation is one that allows to accurately fit any target function of interest (e.g., correctly classify a set of objects in an image). In Reinforcement Learning (RL), this concept is more subtle, as it can be applied to different aspects of the problem, such as the optimal value function or the optimal policy. Furthermore, recent works have shown that realizability (e.g., being able to represent the optimal value function) is not a sufficient condition for solving an RL problem, as the sample complexity using realizable representations is exponential in the worst case [e.g., Weisz et al., 2021]. As such, a desirable property of a "good" representation in RL is to enable learning a near-optimal policy with a polynomial sample complexity (or similarly sublinear regret bound).

Several works have focused on online learning — considering sample complexity or regret minimization — and identified sufficient assumptions for efficient learning. Standard examples are

*Work done while at Facebook AI Research.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

tabular Markov Decision Processes (MDPs) [e.g., Jaksch et al., 2010, Azar et al., 2012, 2017], low or zero inherent Bellman error [e.g., Jin et al., 2020, Zanette et al., 2020b,a, Jin et al., 2021] and linear mixture MDPs [e.g., Yang and Wang, 2019, Ayoub et al., 2020, Zhang et al., 2021]. While, in these settings, the representation is provided as input to the algorithm, an alternative scenario is to learn such representations. In this case, research has focused either on the problem of online representation selection for regret minimization [e.g., Ortner et al., 2014, 2019, Lee et al., 2021] or, more recently, on the sample complexity of online representation learning [e.g., Du et al., 2019, Agarwal et al., 2020, Modi et al., 2021]. Refer to App. A for more details. While this literature has focused on finding a representation enabling learning a near-optimal policy with sublinear regret or polynomial sample complexity, there may be several of such "good" representations with significantly different learning performance and existing approaches are not guaranteed to find the most efficient one. Intuitively, we would like to find representations that require the minimum level of exploration to solve the task. For example, representations that would allow the algorithm to stop exploring after a finite time and play only optimal actions forever (i.e., achieving constant regret), if they exist. This aspect of representation learning was recently studied by Hao et al. [2020], Papini et al. [2021] in contextual linear bandits, where they showed that certain representations display non-trivial properties that enable much better learning performance. While it is well-known that properties such as dimensionality and norm of the features have an impact on the learning performance, Hao et al. [2020], Papini et al. [2021] proved that it is possible to achieve constant regret (i.e., not scaling with the number of learning steps) if a certain (necessary and sufficient) condition on the features associated with the optimal actions is satisfied. To the best of our knowledge, the impact of similar properties on RL algorithms and how to find such representations is largely unexplored.

Contributions. In this paper, we investigate the concept of "good" representations in the context of regret minimization in finite-horizon MDPs with linear structure. In particular, we consider the settings of zero inherent Bellman error (also referred to as Bellman closure) [Zanette et al., 2020b] and low-rank structure [e.g., Jin et al., 2020]. Similarly to the bandit case [Hao et al., 2020, Papini et al., 2021], we study the impact of representations on the learning process. Our contributions are both fundamental and algorithmic. **1)** We provide a necessary condition (called UniSOFT) for a representation to enable constant regret in any problem with linear reward parametrization. Notably, this result encompasses MDPs with zero inherent Bellman error, and linear mixture MDPs with linearly parametrized rewards. Intuitively, the condition generalizes a similar condition for linear contextual bandits and it requires that the features observed along trajectories generated by the optimal actions provide information on the whole feature space (see Asm. 4). **2)** We provide the first *constant regret bound* for MDPs for both ELEANOR [Zanette et al., 2020b] and LSVI-UCB [Jin et al., 2020] when the UniSOFT condition is satisfied. As a consequence, we show that good representations are not only necessary but also sufficient for constant regret in MDPs with zero inherent Bellman error or low-rank assumptions. **3)** We develop an algorithm, called LSVI-LEADER, for representation selection in low-rank MDPs. We prove that in low-rank MDPs, LSVI-LEADER suffers the regret of the best representation without knowing it in advance. Furthermore, LSVI-LEADER achieves constant regret even when only a suitable combination of the representations satisfies the UniSOFT condition despite none of them being "good". This is indeed possible thanks to its ability to select a different representation for each stage, state, and action.

2 Preliminaries

We consider a time-inhomogeneous finite-horizon Markov decision process (MDP) $M = \big(\mathcal{S}, \mathcal{A}, H, \{r_h\}_{h=1}^H, \{p_h\}_{h=1}^H, \mu\big)$ where $\mathcal{S}$ is the state space and $\mathcal{A}$ is the action space, H is the length of the episode, $\{r_h\}$ and $\{p_h\}$ are reward functions and state-transition probability measures, and μ is the initial state distribution. We denote by $r_h(s,a)$ the expected reward of a pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ at stage h. We assume that $\mathcal{S}$ is a measurable space with a possibly infinite number of elements and $\mathcal{A}$ is a finite set. A policy $\pi = (\pi_1, \dots, \pi_H) \in \Pi$ is a sequence of decision rules $\pi_h : \mathcal{S} \to \mathcal{A}$. For every $h \in [H] := \{1, \dots, H\}$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, we define the value functions of a policy π as

$$Q_h^\pi(s,a) = r_h(s,a) + \mathbb{E}_\pi \left[\sum_{i=h+1}^{H} r_i(s_i, a_i) \right], \qquad V_h^\pi(s,a) = Q_h^\pi(s, \pi_h(s)),$$

where the expectation is over probability measures induced by the policy and the MDP over state-action sequences of length $H - h$. Under certain regularity conditions [e.g., Bertsekas and Shreve,

2004], there always exists an optimal policy $\pi^\star$ whose value functions are defined by $V_h^{\pi^\star}(s) := V_h^\star(s) = \sup_\pi V_h^\pi(s)$ and $Q_h^{\pi^\star}(s,a) := Q_h^\star(s,a) = \sup_\pi Q_h^\pi(s,a)$. The optimal Bellman equation (and Bellman operator L_h) at stage $h \in [H]$ is defined as:

$$Q_h^\star(s,a) := L_h Q_{h+1}^\star(s,a) = r_h(s,a) + \mathbb{E}_{s' \sim p_h(s,a)} \left[\max_{a'} Q_{h+1}^\star(s',a') \right].$$

The value iteration algorithm (a.k.a. backward induction) computes $Q^\star$ or Q^π by applying the Bellman equations starting from stage H down to 1, with $V_{H+1}^\pi(s) = 0$ by definition for any s and π. The optimal policy is simply the greedy policy w.r.t. $Q^\star$: $\pi_h^\star(s) = \operatorname{argmax}_{a \in \mathcal{A}} Q_h^\star(s,a)$.

In online learning, the agent interacts with an *unknown* MDP in a sequence of K episodes. At each episode k, the agent observes an initial state s_1^k, it selects a policy π_k, it collects the samples observed along a trajectory obtained by executing π_k, it updates the policy, and reiterates over the next episode. We evaluate the performance of a learning agent through the regret: $R(K) := \sum_{k=1}^K V_1^\star(s_1^k) - V_1^{\pi_k}(s_1^k)$.

Linear Representation. When the state space is large or continuous, value functions are often described through a parametric representation. A standard approach is to use linear representations of the state-action function $Q_h(s,a) = \phi_h(s,a)^\mathsf{T}\theta_h$, where $\phi_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is a time-inhomogeneous feature map and $\theta_h \in \mathbb{R}^d$ is an unknown parameter vector.[2] In this paper, we consider MDPs satisfying Bellman closure (i.e., zero Inherent Bellman Error) [Zanette et al., 2020b] or low-rank assumptions [e.g., Yang and Wang, 2019, Jin et al., 2020].

Assumption 1 (Bellman Closure)**.** *Define the set of bounded value function* $\mathcal{Q}_h = \{Q_h | \theta_h \in \Theta_h : Q_h(s,a) = \phi_h(s,a)^\mathsf{T}\theta_h, \forall (s,a)\}$ *and the associated parameter space* $\Theta_h = \{\theta_h \in \mathbb{R}^d : |\phi_h(s,a)^\mathsf{T}\theta_h| \le D\}$*. An MDP has zero Inherent Bellman Error (IBE) if*

$$\forall h \in [H], \qquad \sup_{Q_{h+1} \in \mathcal{Q}_{h+1}} \inf_{Q_h \in \mathcal{Q}_h} \|Q_h - L_h Q_{h+1}\|_\infty = 0.$$

This definition implies that the optimal value function is realizable as $Q_h^\star \in \mathcal{Q}_h$. Furthermore, the function space $\mathcal{Q}$ is closed under the Bellman operator, i.e., for all $Q_{h+1} \in \mathcal{Q}_{h+1}$, $L_h Q_{h+1} \in \mathcal{Q}_h$. Under this assumption, value-iteration-based algorithms are guaranteed to converge to the optimal policy in the limit of samples and iterations [Munos and Szepesvári, 2008]. In the context of regret minimization, Zanette et al. [2020b] proposed a model-free algorithm, called ELEANOR, that achieves sublinear regret under the Bellman closure assumption, but at the cost of computational intractability.[3] The design of a tractable algorithm for regret minimization under low IBE assumption is still an open question in the literature.

Assumption 2 (Low-Rank MDP)**.** *Let* $\Theta_h = \mathbb{R}^d$*, then an MDP has low-rank structure if*

$$\forall s, a, h, s', \quad r_h(s,a) = \phi_h(s,a)^\mathsf{T}\theta_h, \quad p_h(s'|s,a) = \phi_h(s,a)^\mathsf{T}\mu_h(s')$$

where $\mu_h : \mathcal{S} \to \mathbb{R}^d$*. Then, for any policy* $\pi \in \Pi$*,* $\exists \theta_h^\pi \in \Theta_h$ *such that* $Q_h^\pi(s,a) = \phi_h(s,a)^\mathsf{T}\theta_h^\pi$*. We assume* $\|\theta_h\|_2 \le \sqrt{d}$*,* $\|\int_{s'} \mu_h(s')v(s')\mathrm{d}s'\|_2 \le \sqrt{d}\|v\|_\infty$ *and* $\|\phi_h(s,a)\|_2 \le 1$*, for any* s, a, h*, and function* $v : \mathcal{S} \to \mathbb{R}$*.*

This assumption is *strictly* stronger than Bellman closure [Zanette et al., 2020b] and it implies the value function of *any* policy is linear in the features. Furthermore, under Asm. 2 sublinear regret is achievable using, e.g., LSVI-UCB [Jin et al., 2020], a tractable algorithm for low-rank MDPs. He et al. [2020] have recently established a problem-dependent logarithmic regret bound for LSVI-UCB under a strictly-positive minimum gap. The minimum positive gap provides a natural measure of the difficulty of an MDP.

Assumption 3. *The suboptimality gap of taking action* a *in state* s *at stage* h *is defined as:*

$$\Delta_h(s,a) = V_h^\star(s) - Q_h^\star(s,a). \tag{1}$$

We assume the minimum positive gap $\Delta_{\min} = \min_{s,a,h}\{\Delta_h(s,a) | \Delta_h(s,a) > 0\}$ *is well defined and that the optimal action is unique, i.e.,* $|\operatorname{argmax}_a\{Q_h^\star(s,a)\}| = 1$*, for any* $s \in \mathcal{S}$*,* $h \in [H]$*.*

[2] It is possible to extend the setting to different feature dimensions $\{d_h\}_{h \in [H]}$.

[3] ELEANOR works under the weaker assumption of low IBE. Jin et al. [2021] considered the more general case of low Bellman Eluder dimension. Their algorithm reduces to ELEANOR in the case of low IBE.

Algorithm (setting)	Minimax	Problem-Dependent Logarithmic	Constant with UNISOFT *(this work)*
ELEANOR *(Bellman Closure)*	$\widetilde{O}(\sqrt{d^2H^3T})$ [Zanette et al., 2020b]	N/A	$\frac{d^2H^4}{\Delta_{\min}\lambda_+^{3/2}}\log^{1/2}\left(\frac{d^2H^5}{\delta\Delta_{\min}^2\lambda_+^3}\right)$ (Thm. 8)
LSVI-UCB *(low-rank MDPs)*	$\widetilde{O}(\sqrt{d^3H^3T})$ [Jin et al., 2020]	$O(\frac{d^3H^5}{\Delta_{\min}}\log^2(T))$ [He et al., 2020]	$\frac{d^3H^5}{\Delta_{\min}}\log\left(\frac{d^4H^6}{\delta\Delta_{\min}^2\lambda_+^3}\right)$ (Thm. 9)
Lower Bound	$\Omega(\sqrt{d^2H^2T})$ [Zhou et al., 2020, Remark 5.8]	$\Omega(\frac{dH}{\Delta_{\min}})$ [He et al., 2020]	N/A

Table 1: Regret comparisons of ELEANOR and LSVI-UCB. For ELEANOR, we consider the special case of Bellman closure.

In Tab. 1, we summarize existing bounds in the two settings. Another structural assumption that has gained popularity in the literature is the linear-mixture structure [Jia et al., 2020, Ayoub et al., 2020, Zhou et al., 2020], where the transition function admits a form $p_h(s'|s,a) = \phi_h(s'|s,a)^\mathsf{T}\theta_h$. No structural requirement is made on the reward, which is typically assumed to be known. As a consequence, the value function may not be linearly representable. However, the fact the reward is known and that it is possible to directly learn the parameters θ_h of the transition function allow to achieve sublinear regret (even logarithmic) through model-based algorithms. While in this paper we mostly focus on Asm. 1 and 2, in Sect. 3.1 we show that our condition is necessary for constant regret also for linear-mixture MDPs with unknown linear reward.

3 Constant Regret for Linear MDPs

In this section, we introduce UNISOFT, a necessary condition for constant regret in any MDP with linear rewards. We show that this condition is also sufficient in MDPs with Bellman closure.

Assumption 4. *A feature map is* UNISOFT *(Universally Spanning Optimal FeaTures) for an MDP if it satisfies Asm. 1 or 2, and for all $h \in [H]$ the following holds:*

$$\operatorname{span}\Big\{\phi_h(s,a) \mid \forall(s,a),\ \exists\pi\in\Pi : \rho_h^\pi(s,a) > 0\Big\} = \operatorname{span}\Big\{\phi_h^\star(s) \mid \forall s,\ \rho_h^\star(s) > 0\Big\},$$

where $\rho_h^\pi(s) = \mathbb{E}[\mathbb{1}\{s_h = s\}|M,\pi]$ is the occupancy measure of a policy π, $\rho_h^\pi(s,a) = \rho_h(s)\mathbb{1}\{\pi_h(s) = a\}$, $\rho_h^\star(s) := \rho_h^{\pi^\star}(s)$, and $\phi_h^\star(s) := \phi_h(s,\pi_h^\star(s))$.

Intuitively, features that are observed by only playing optimal actions must provide information on the whole space of reachable features at each stage h. We notice that Asm. 4 reduces to the HLS property for contextual bandits considered by Hao et al. [2020], Papini et al. [2021]. The key difference is that, in RL, the reachability of a state plays a fundamental role. For example, features of states that are not reachable by any policy are irrelevant, while features of optimal actions in states that are not reachable by the optimal policy (i.e., $\phi_h^\star(s)$ in a state with $\rho_h^\star(s) = 0$) do not contribute to the span of optimal features since they can only be reached by acting sub-optimally. In RL, a related structural assumption to Asm. 4 is the "uniformly excited feature" assumption used by Abbasi-Yadkori et al. [2019, Asm. A4] for average reward problems. Their assumption is strictly stronger than ours since it requires that all policies generate an occupancy measure under which the features span all directions uniformly well. Such an assumption can be related to the ergodicity assumption for tabular MDPs, which is known to be restrictive. Another related quantity is the "explorability" coefficient introduced by Zanette et al. [2020c]. This term represents how explorative (in the feature space) are the optimal policies of the tasks compatible with the MDP, i.e., considering any possible parameter $\theta_h \in \Theta_h$. This coefficient is important in reward-free exploration where the objective is to learn a near optimal policy for any task, which is revealed only once learning has completed. In our setting, we focus only on the properties of the optimal policy for the single task we aim to solve.

It is interesting to look into Asm. 4 from an alternative perspective. Denote by $0 \le \lambda_{h,1} \le \ldots \le \lambda_{h,d}$ the eigenvalues of the matrix $\Lambda_h := \mathbb{E}_{s\sim\rho_h^\star}\left[\phi_h^\star(s)\phi_h^\star(s)^\mathsf{T}\right]$ and by $\lambda_h^+ = \min\{\lambda_{h,i} > 0, i \in [d]\}$ the minimum positive eigenvalue. We notice that when the features are non-redundant (i.e., $\{\phi_h(s,a)\}$

spans $\mathbb{R}^d$) and the UNISOFT assumption holds, then $\lambda_h^+ = \lambda_{h,1} > 0$. As we will see, the minimum positive eigenvalue λ_h^+ plays a fundamental role in the constant regret bound, together with the minimum gap $\Delta_{\min}$. We provide examples of UNISOFT and Non-UNISOFT representations in App. G, as well as their impact on the learning process.

3.1 UNISOFT is Necessary for Constant Regret

The following theorem shows that the UNISOFT condition is necessary to achieve constant regret in a large class of MDPs.

Theorem 5. *Let M be any MDP with finite states, arbitrary dynamics p, linear rewards (i.e., $r_h(s,a) = \phi_h(s,a)^\mathsf{T}\theta_h$) with Gaussian $\mathcal{N}(0,1)$ noise, unique optimal policy $\pi^\star$, and where condition* UNISOFT *(Asm. 4) is not satisfied. Let $\mathcal{M}$ be the set of MDPs with same dynamics as M but different reward parameters $\{\theta_h\}_{h\in[H]}$. Then, there exists no algorithm that suffers sub-linear regret in all MDPs in $\mathcal{M}$ while suffering constant regret in M.*

Thm. 5 states that in MDPs with linear reward, the UNISOFT condition is *necessary* to achieve constant regret for any "provably efficient" algorithm. Notably, this result does not put any restriction on the transition model, which can be arbitrary and known. This means that as soon as the reward is linear and unknown to the learning agent, the UNISOFT condition is necessary to attain constant regret. This result applies to low-rank MDPs, linear-mixture MDPs with unknown linear rewards, and MDPs with Bellman closure (Bellman closure implies linear rewards, see Prop. 2 by Zanette et al. [2020b]).

Proof sketch of Theorem 5. The key intuition behind the proof is that an algorithm achieving a constant regret must select sub-optimal actions only a finite number of times. Nonetheless, in order to learn the optimal policy, all features associated with suboptimal actions should be explored enough. Since UNISOFT does not hold, this cannot happen by executing the optimal policy alone and requires selecting suboptimal policies for long enough, thus preventing constant regret.

More formally, we call an algorithm "provably efficient" if it suffers sub-linear regret on the given class of MDPs $\mathcal{M}$. Formally, we use the following definition, which is standard to prove problem-dependent lower bounds [e.g., Simchowitz and Jamieson, 2019, Xu et al., 2021].

Definition 6 (α-consistency)**.** *Let $\alpha \in (0,1)$, then an algorithm A is α-consistent on a class of MDPs $\mathcal{M}$ if, for each $M \in \mathcal{M}$ and $K \geq 1$, there exists a constant c_M (independent from K) such that $\mathbb{E}_M^{\mathsf{A}}[R(K)] \leq c_M K^\alpha$.*[4]

The following lemma is the key result for proving Thm. 5 and it might be of independent interest. It shows that any consistent algorithm must explore sufficiently all relevant directions in the feature space to discriminate any sub-optimal policy from the optimal one. The proof (reported in App. C) leverages techniques for deriving asymptotic lower bounds for linear contextual bandits [e.g., Lattimore and Szepesvari, 2017, Hao et al., 2020, Tirinzoni et al., 2020].

Lemma 7. *Let $M, \mathcal{M}$ be as in Thm. 5 and A be any α-consistent algorithm on $\mathcal{M}$. For any $\pi \in \Pi$, denote by $\Psi_h^\pi := \sum_{s,a} \rho_h^\pi(s,a)\phi_h(s,a)$ its expected features at stage h and $\Delta(\pi) := V_1^\star - V_1^\pi$ its sub-optimality gap. Then, for any $\pi \in \Pi$ with $\Delta(\pi) > 0$ and $h \in [H]$,*

$$\limsup_{K\to\infty} \log(K)\|\Psi_h^\pi - \Psi_h^\star\|^2_{\mathbb{E}_M^{\mathsf{A}}[\Lambda_h^K]^{-1}} \leq \frac{\Delta(\pi)^2}{2(1-\alpha)},$$

where $\Psi_h^\star := \Psi_h^{\pi^\star}$ and $\Lambda_h^K := \sum_{k=1}^K \phi_h(s_h^k, a_h^k)\phi_h(s_h^k, a_h^k)^\mathsf{T}$.

We now proceed by contradiction: suppose that A suffers constant expected regret on M even though the MDP does not satisfy the UNISOFT condition. Then, since A plays sub-optimal actions only a finite number of times, it is possible to show that, for each $h \in [H]$, there exists a positive constant $\lambda_M > 0$ such that $\mathbb{E}_M^{\mathsf{A}}[\Lambda_h^K] \preceq \Lambda_h^\star + \lambda_M I$, where $\Lambda_h^\star := K\sum_{s:\rho_h^\star(s)>0} \phi_h^\star(s)\phi_h^\star(s)^\mathsf{T}$. Furthermore, since UNISOFT does not hold, there exists a stage $h \in [H]$ and a sub-optimal policy π (i.e., with $\Delta(\pi) > 0$)

[4] In practice, all existing "provably-efficient" algorithms we are interested in are included in this class and c_M is polynomial in all problem-dependent quantities (e.g., d, H). For instance, LSVI-UCB and ELEANOR are $1/2$-consistent on the class of low-rank and Bellman-closure MDPs, where they enjoy worst-case $\widetilde{O}(\sqrt{K})$ regret bounds (with c_M being $O(\sqrt{d^3H^4})$ and $O(\sqrt{d^2H^4})$, respectively).

such that the vector $\Psi_h^\pi - \Psi_h^\star$ does not belong to $\text{span}\{\phi_h^\star(s)|\rho_h^\star(s) > 0\}$. Then, since such space is exactly the one spanned by all the eigenvectors of $\Lambda_h^\star$ associated with a non-zero eigenvalue, there exists a positive constant $\epsilon > 0$ (independent of K) such that $\|\Psi_h^\pi - \Psi_h^\star\|^2_{(\Lambda_h^\star + \lambda_M I)^{-1}} \geq \epsilon^2/\lambda_M$. Combining these steps with Lem. 7, we obtain

$$\frac{\Delta(\pi)^2}{2(1-\alpha)} \geq \limsup_{K\to\infty} \log(K)\|\Psi_h^\pi - \Psi_h^\star\|^2_{(\Lambda_h^\star+\eta I)^{-1}} \geq \frac{\epsilon^2}{\lambda_M} \limsup_{K\to\infty} \log(K),$$

which is clearly a contradiction. Therefore, A cannot suffer constant regret in M while suffering sub-linear regret in all other MDPs in $\mathcal{M}$, and our claim follows.

3.2 UniSOFT is Sufficient for Constant Regret

While the UniSOFT condition is necessary for achieving constant regret in a large class of MDPs, in the following, we prove that ELEANOR and LSVI-UCB attain constant regret when the UniSOFT assumption holds, thus implying that it is a sufficient condition in MDPs with low-rank and Bellman closure structure.

Theorem 8. *Consider an MDP and a representation $\{\phi_h\}_{h\in[H]}$ satisfying the Bellman closure (Asm. 1) and* UniSOFT *assumptions (Asm. 4). Under Asm. 3, with probability at least $1-3\delta$, ELEANOR*[5] *suffers a* constant *regret*

$$R(K) \lesssim H^{3/2} d \sqrt{\overline{\tau} \log \frac{\overline{\tau}}{\delta}},$$

where $\overline{\tau} = H\overline{\kappa}$ and $\overline{\kappa}$ is the last episode ELEANOR suffers a non-zero regret. Furthermore, $\overline{\kappa} \lesssim \max\left\{\frac{d^2H^4}{\lambda_+^2}, \frac{dH^4}{\Delta_{\min}^2\lambda_+^3}\right\}$[6]*, where $\lambda_+ := \min_h\{\lambda_h^+\} > 0$.*

Alternatively, we can prove the following result for LSVI-UCB.

Theorem 9. *Consider an MDP and a representation $\{\phi_h\}_{h\in[H]}$ satisfying the low-rank (Asm. 2) and* UniSOFT *assumptions (Asm. 4). Under Asm. 3, with probability $1-3\delta$, LSVI-UCB suffers a* constant *regret*

$$R(K) \lesssim \frac{d^3H^5}{\Delta_{\min}} \log\left(dH^2\overline{\kappa}/\delta\right),$$

where $\overline{\kappa}$ is the last episode LSVI-UCB suffers a non-zero regret and is upper-bounded as $\overline{\kappa} \lesssim \max\left\{\frac{d^3H^4}{\lambda_+^2}, \frac{d^2H^4}{\Delta_{\min}^2\lambda_+^3}\right\}$, where $\lambda_+ := \min_h\{\lambda_h^+\} > 0$.

In both cases, $\overline{\kappa}$ is polynomial in all the problem-dependent terms and independent of the number of episodes K (see Lem. 21 and 20). As a result, ELEANOR and LSVI-UCB achieves a constant regret that only depends on "static" MDP and representation characteristics, thus indicating that after a finite time the agent only executes the optimal policy. Notice also that the bounds should be read as minimum between the constant regret and the minimax regret $O(\sqrt{K})$, which may be tighter for small K.The main difference between the two previous bounds is that for ELEANOR we build on the anytime minimax regret bound, while for LSVI-UCB, we derive a more refined constant-regret guarantee by building on its problem-dependent bound of He et al. [2020]. Unfortunately, limiting factor for applying the analysis in [He et al., 2020] seems to be the fact that ELEANOR is not optimistic at each stage h but rather only at the first stage. As such, whether ELEANOR can achieve a problem-dependent logarithmic regret based on local gaps that can be leverage to improve our analysis is an open question in the literature.

Combined proof sketch of Thm. 8 and Thm. 9. We provide a general proof sketch that can be instantiated to both ELEANOR and LSVI-UCB. The purpose is to illustrate what properties an algorithm must have to exploit good representations, and how this leads to constant regret. Consider a learnable feature map $\{\phi_h\}_{h\in[H]}$ and an algorithm with the following properties:

(a) Greedy w.r.t. a Q-function estimate: $\pi_h^k(s) = \arg\max_{a\in\mathcal{A}}\{\overline{Q}_h^k(s,a)\}$.

[5] ELEANOR and LSVI-UCB are defined up to a regularization parameter λ that we set to $\lambda = 1$.

[6] Here $\lesssim$ hides logarithmic terms in λ_+, H, and d, but not in K.

(b) Global optimism: $\overline{V}_1^k(s) \geq V_1^\star(s)$ where, for all $h \geq 1$, we set $\overline{V}_h^k(s) = \max_{a\in\mathcal{A}}\{\overline{Q}_h^k(s,a)\}$.

(c) Almost local optimism: $\forall h > 1, \exists C_h \geq 0$ s.t. $\overline{Q}_h^k(s,a) + C_h\beta_k \left\|\phi_h(s,a)\right\|_{(\Lambda_h^k)^{-1}} \geq Q_h^\star(s,a)$.

(d) Confidence set: let $\Lambda_h^k = \sum_{i=1}^{k-1} \phi_h(s_h^i, a_h^i)\phi_h(s_h^i, a_h^i)^\mathsf{T} + \lambda I$ and $\beta_k \in \mathbb{R}_+$ be logarithmic in k, then $\overline{V}_h^k(s_h^k) - V_h^{\pi_k}(s_h^k) \leq 2\beta_k \left\|\phi_h(s_h^k, a_h^k)\right\|_{(\Lambda_h^k)^{-1}} + \mathbb{E}_{s'\sim p_h(s_h^k,a_h^k)}\left[\overline{V}_{h+1}^k(s') - V_{h+1}^{\pi_k}(s')\right]$.

These properties are verified by ELEANOR [Zanette et al., 2020b, App. C] and LSVI-UCB [Jin et al., 2020, Lem. B.4, B.5]. Note that for LSVI-UCB condition (c) is trivially verified since the algorithm is optimistic at each stage ($C_h = 0$). On the other hand, ELEANOR is only guaranteed to be optimistic at the first stage, and (c) is thus important ($C_h = 2$). First, we use existing techniques to establish an *any-time* regret bound, either worst-case or problem-dependent. We call this $g(k)$ and prove that $R(k) \leq g(k) \leq \widetilde{O}(\sqrt{k})$ for any k with probability $1 - 2\delta$.

Next, we show that, under Asm. 4, the eigenvalues of the design matrix grow almost linearly, making the confidence intervals decrease at a $1/\sqrt{k}$ rate. From some algebra and a martingale argument,

$$\Lambda_h^{k+1} \succeq k\Lambda_h^\star + \lambda I - \Delta_{\min}^{-1} g(k) I - \widetilde{O}(\sqrt{k}) I, \tag{2}$$

where $\Lambda_h^\star = \mathbb{E}_{s\sim\rho_h^\star}[\phi_h^\star(s)\phi_h^\star(s)^\mathsf{T}]$. The UNISOFT property ensures that the linear term is nonzero in relevant directions, while the regret bound of the algorithm makes the penalty term sublinear. Then, we show that, for any *reachable* (s,a),

$$\beta_k \left\|\phi_h(s,a)\right\|_{(\Lambda_h^k)^{-1}} \leq \beta_k \frac{k - \widetilde{O}(\sqrt{k})}{(k\lambda_h^+ - \widetilde{O}(\sqrt{k}))^{3/2}} = \widetilde{O}(k^{-1/2}), \tag{3}$$

where λ_h^+ is the minimum *nonzero* eigenvalue of $\Lambda_h^\star$. From (3), we can see that λ_h^+ plays a fundamental role in the rate of decrease. Finally, we show that, under the gap assumption, these uniformly-decreasing confidence intervals allow learning the optimal policy in a finite time. From the Bellman equations, we have that

$$V_1^\star(s_1^k) - V_1^{\pi^k}(s_1^k) = \mathbb{E}_{\pi^k}\left[\sum_{h=1}^H \Delta_h(s_h, a_h) | s_1 = s_1^k\right], \tag{4}$$

while from (a)-(d), for any reachable state,

$$\Delta_h(s, \pi_h^k(s)) \leq 2\,\mathbb{E}_{\pi^k}\left[\sum_{i=h}^H \beta_k \left\|\phi_i(s_i,a_i)\right\|_{(\Lambda_i^k)^{-1}} | s_h = s\right] + \mathbb{1}_{h>1} C_h \beta_k \left\|\phi_h^\star(s)\right\|_{(\Lambda_h^k)^{-1}}.$$

The second term (with $\mathbb{1}_{h>1}$) accounts for the almost-optimism of ELEANOR, while it is zero in LSVI-UCB due to the stage-wise optimism. Then, for every $h \in [H]$, we can use (3) to control the feature norms. Thus, there exists an episode κ_h independent of K satisfying

$$\Delta_h(s, \pi_h^k(s)) \leq \beta_{\kappa_h} \sum_{i=h}^H (2 + \mathbb{1}_{i=h>1} C_h) \frac{\kappa_h - 8\sqrt{\kappa_h \log(2d\kappa_h H/\delta)} - g(\kappa_h)}{(\kappa_h\lambda_i^+ - 8\sqrt{\kappa_h \log(2d\kappa_h H/\delta)} - g(\kappa_h))^{3/2}} < \Delta_{\min}, \tag{5}$$

By definition of minimum gap, then $\Delta_h(s, \pi_h^k(s)) = 0$ for $k > \kappa_h$. Then, for $k > \overline{\kappa} = \max_h\{\kappa_h\}$, $V_1^\star(s_1^k) - V_1^{\pi^k}(s_1^k) = 0$. But this means the algorithm only accumulates regret up to $\overline{\kappa}$, that is, $R(K) = R(\overline{\kappa}) \leq g(\overline{\kappa}) = O(1)$ for all $K > \overline{\kappa}$. This holds with probability $1 - 3\delta$, also taking into account the martingale argument from (2). Note that $\{\kappa_h\}$ are by definition monotone for LSVI-UCB. The final bounds are then obtained by instantiating the specific values of β_k and $g(k)$ for the two algorithms we analyzed.

4 Representation Selection in Low-Rank MDPs

In Sec. 3, we have highlighted the benefits that a UNISOFT representation brings to optimistic algorithms in MDPs with Bellman closure and low rank structure. In this section, we take one step further and investigate the *representation selection* problem. Since ELEANOR is a computationally

Algorithm 1: LSVI-LEADER

Input: Representations $\{\Phi_j\}_{j\in[N]}$, confidence values $\{\beta_k\}_{k\in[K]}$

for $k = 1, \dots, K$ **do**

  Receive the initial state s_1^k

  for $h = H, \dots, 1$ **do**

    $\Lambda_h^k(j) = \lambda I + \sum_{i=1}^{k-1} \phi_h^{(j)}(s_h^i, a_h^i)\phi_h^{(j)}(s_h^i, a_h^i)^\top \; \forall\, j \in [N]$.

    $\boldsymbol{w}_h^k(j) = \Lambda_h^k(j)^{-1} \sum_{i=1}^{k-1} \phi_h^{(j)}(s_h^i, a_h^i) \left(r_h(s_h^i, a_h^i) + \max_{a\in\mathcal{A}} \overline{Q}_{h+1}^k(s_{h+1}^i, a) \right), \; \forall j \in [N]$

    $\overline{Q}_h^k(s,a) = \min \left\{ H, \min_{j\in[N]} \left(\phi_h^{(j)}(s,a)^\top \boldsymbol{w}_h^k(j) + \beta_k \left\| \phi_h^{(j)}(s,a) \right\|_{\Lambda_h^k(j)^{-1}} \right) \right\}$

  for $h = 1, \dots, H$ **do**

    Execute action $a_h^k = \pi_h^k(s_h^k) := \operatorname{argmax}_{a\in\mathcal{A}} \overline{Q}_h^k(s_h^k, a)$.

intractable algorithm, we build on LSVI-UCB and low-rank MDPs (Asm. 2) and we introduce LSVI-LEADER (Alg. 1), an algorithm that adaptively selects representations in a given set.

Given a set of N representations $\{\Phi_j\}_{j\in[N]}$ satisfying Asm. 2, where $\Phi_j = \left\{\phi_h^{(j)}\right\}_{h\in[H]}$, at each stage $h \in [H]$ of episode $k \in [K]$, LSVI-LEADER solves N different regression problems to compute an optimistic value function for each representation. Then, the final estimate $\overline{Q}_h^k(s,a)$ is taken as the *minimum* across these different optimistic value functions. Notably, this implies that LSVI-LEADER implicitly *combines* representations, in the sense that the selected representations (i.e., those with tightest optimism) might vary for different stages. This is exploited in the following result, which shows that constant regret is achievable even if none of the given representations is globally UNISOFT.

Theorem 10. *Given an MDP M and a set of representations $\{\Phi_j\}_{j\in[N]}$ satisfying the low-rank assumption (Asm. 2), let $\mathcal{Z}$ be the set of H^N representations obtained by combining those in $\{\Phi_j\}_{j\in[N]}$ across different stages.*[7] *Then, with probability at least $1-2\delta$, LSVI-LEADER suffers at most a regret*

$$R(K) \le \min_{z\in\mathcal{Z}} \widetilde{R}(K, z, \{\beta_k\}),$$

where $\widetilde{R}(K, z, \beta_k)$ is either the worst-case regret bound of LSVI-UCB [Jin et al., 2020] or the problem-dependent one [He et al., 2020] when the algorithm is executed with representation z and confidence values $\beta_k \propto dH\sqrt{N\log(2dNHk/\delta)}$. Moreover, if $\mathcal{Z}$ contains a UNISOFT *representation $z^\star$, then LSVI-LEADER achieves constant regret with problem-dependent values of $z^\star$ (see Thm. 9).*

This result shows that LSVI-LEADER adapts to the *best* representation automatically, i.e., without any prior knowledge about the properties of the representations. In particular, it shows a problem-dependent (or worst-case) bound when there is no UNISOFT representation, while it attains constant regret when a representation, potentially mixed through stages, is UNISOFT. This is similar to what was obtained by Papini et al. [2021] for linear contextual bandits. Indeed, LSVI-LEADER reduces to their algorithm in the case $H = 1$. While the cost of representation selection is only logarithmic in linear bandits, the cost becomes polynomial (i.e., $\sqrt{N}$ in the worst-case bound and N in the problem-dependent one) in RL. This is due to the structure induced by the Bellman equation, which requires a cover argument over H^N functions (more details in the proof sketch). Note that for $H = 1$, the analysis can be refined to obtain a $\log(N)$ dependence, due to the lack of propagation through stages, and recover the result in [Papini et al., 2021]. We refer the read to App. G for a numerical validation.

Proof sketch of Thm. 10. The proof relies on the following important result, which extends Lem. B.4 of Jin et al. [2020] and shows that the deviation between the optimistic value function computed by LSVI-LEADER and the true one scales with the *minimum* confidence interval across the different representations. Formally, with probability $1-2\delta$, for any $\pi \in \Pi, s \in \mathcal{S}, a \in \mathcal{A}, h \in$

[7] Note that any combination of features in Φ_j is learnable, since each representation is learnable in the low-rank MDP sense.

$[H], k \in [K]$,

$$\overline{Q}_h^k(s,a) - Q_h^\pi(s,a) \leq 2\beta_k \min_{j\in[N]} \left\|\phi_h^{(j)}(s,a)\right\|_{\Lambda_h^k(j)^{-1}} + \mathbb{E}_{s'\sim p_h(s,a)}\left[\overline{V}_{h+1}^k(s') - V_{h+1}^\pi(s')\right].$$

As in [Jin et al., 2020], the derivation of this result combines the well-known self-normalized martingale bound in [Abbasi-Yadkori et al., 2011] with a covering argument over the space of possible optimistic value functions. In our setting, the structure of such function space requires us to build N different covers, one for each different representation. This, in turn, requires the confidence values β_k to be inflated by an extra factor $\sqrt{N}$ w.r.t. learning with a single representation.

The generality of this result allows us to easily derive, for any fixed representation $z \in \mathcal{Z}$, both the worst-case regret bound of Jin et al. [2020] and the problem-dependent one of He et al. [2020]. To see this, note that the regret decompositions in both of these two papers rely on an upper bound to $\overline{V}_h^k(s_h^k) - V_h^{\pi_k}(s_h^k)$ as a function of the *fixed* representation used by LSVI-UCB (see the proof of Theorem 3.1 of Jin et al. [2020] and Lemma 6.2 of He et al. [2020]). Then, fix any $z \in \mathcal{Z}$ and call z_h its features at stage h. Note that $z_h \in \{\phi_h^{(j)}\}_{j\in[M]}$. Moreover, by definition of low-rank structure, since each Φ_j induces a low-rank MDP, their combination does too. Thus, z is learnable. Then, instantiating the concentration bound stated above for policy π^k, state s_h^k, action a_h^k, stage h, and by upper bounding the minimum with the representation selected in z_h, we get

$$\overline{V}_h^k(s_h^k) - V_h^{\pi_k}(s_h^k) \leq 2\beta_k \left\|z_h(s_h^k, a_h^k)\right\|_{\Lambda_h^k(j)^{-1}} + \mathbb{E}_{s'\sim p_h(s_h^k,a_h^k)}\left[\overline{V}_{h+1}^k(s') - V_{h+1}^{\pi_k}(s')\right].$$

From here, one can carry out exactly the same proofs of Jin et al. [2020] and He et al. [2020], thus obtaining the same regret bound that LSVI-UCB enjoys when executed with the fixed representation $z \in \mathcal{Z}$ and confidence values $\{\beta_k\}_{k\in[K]}$. Hence, we conclude that the regret of LSVI-LEADER is upper bounded by the minimum of these regret bounds for all representations $z \in \mathcal{Z}$, thus proving the first result. To obtain the second result, simply notice that, if $z^\star \in \mathcal{Z}$ is UNISOFT, then we can use the refined analysis for LSVI-UCB of Thm. 9 to show that $\widetilde{R}(K, z^\star, \{\beta_k\})$ is upper bounded by a constant independent of K, hence proving constant regret for LSVI-LEADER.

4.1 Representation Selection under a Mixing Condition

We show that the LSVI-LEADER algorithm not only is able to select the best representation among a set of viable representations, and to combine representations for the different stages, but also to stitch representations together *across states and actions*. With this in mind we introduce the notion of a mixed ensemble of representations.

Definition 11. *Consider an MDP M and a set of representations $\{\Phi_j\}_{j\in[N]}$ satisfying the low-rank assumption (Asm. 2). The collection of feature maps $\{\Phi_j\}_{j\in[M]}$ is* UNISOFT*-mixing if for all $s, a \in \mathcal{S}\times\mathcal{A}$ and $h \in [H]$, there exists j such that $\phi_h^{(j)}(s,a) \in \operatorname{span}\left\{\phi_h^{(j)}(s,\pi_h^\star(s)) | \rho_h^\star(s) > 0\right\}$.*

We show that when presented with a UNISOFT-mixing family of representations, LSVI-LEADER is able to successfully combine these and obtain a regret guarantee that may be better than what is achievable by running LSVI-UCB using any of these representations in isolation.

Theorem 12. *Consider an MDP M and a set of representations $\{\Phi_j\}_{j\in[N]}$ satisfying the low-rank (Asm. 2) and* UNISOFT*-mixing assumptions. If $\Delta_{\min} > 0$ (Asm. 3), then with probability at least $1 - 3\delta$, there exist a constant $\widetilde{\kappa} = \max_h\{\kappa_h\}$ independent from K such that the regret of LSVI-LEADER after K episodes is at most:*

$$R(K) \leq \min_{z\in\mathcal{Z}} \widetilde{R}\big(\widetilde{\kappa}, z, \{\beta_k\}\big),$$

where $\mathcal{Z}$, $\widetilde{R}$ and β_k are defined as in Thm. 10.

Under the UNISOFT-mixing condition, LSVI-LEADER may not converge to selecting a single representation for each stage h but rather to mixing multiple representations. In fact, it may select a different representation in different regions of the state-action space. This is the main difference w.r.t. Thm. 10, where constant regret is shown when there exists a representation $z^\star$ that is UNISOFT, and the value κ_h depends on the minimum positive eigenvalue of $z_h^\star$. In the case of UNISOFT-mixing, κ_h depends on properties of a combination of representations at stage h. We provide a characterization of κ_h in the full proof in App. E.

5 Conclusions

We investigated the properties that make a representation efficient for online learning in MDPs with Bellman closure. We introduced UNISOFT, a necessary and sufficient condition to achieve a constant regret bound in this class of MDPs. We demonstrate that existing optimistic algorithms are able to adapt to the structure of the problem and achieve constant regret. Furthermore, we introduce an algorithm able to achieve constant regret by mixing representations across states, actions and stages in the case of low-rank MDPs. An interesting direction raised by our paper is whether it is possible to leverage the UNISOFT structure for probably-efficient representation learning, rather than selection. Another direction can be to leverage these insights to drive the design of auxiliary losses for representation learning, for example in deep RL.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *NIPS*, pages 2312–2320, 2011.

Yasin Abbasi-Yadkori, Peter L. Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvári, and Gellért Weisz. POLITEX: regret bounds for policy iteration using expert prediction. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 3692–3702. PMLR, 2019.

Alekh Agarwal, Sham M. Kakade, Akshay Krishnamurthy, and Wen Sun. FLAMBE: structural complexity and representation learning of low rank mdps. In *NeurIPS*, 2020.

Alex Ayoub, Zeyu Jia, Csaba Szepesvári, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 463–474. PMLR, 2020.

Mohammad Gheshlaghi Azar, Rémi Munos, and Bert Kappen. On the sample complexity of reinforcement learning with a generative model. In *ICML*. icml.cc / Omnipress, 2012.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 263–272. PMLR, 2017.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, August 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2013.50. URL `https://doi.org/10.1109/TPAMI.2013.50`.

Dimitir P Bertsekas and Steven Shreve. *Stochastic optimal control: the discrete-time case*. 2004.

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pages 578–598. PMLR, 2021.

Simon S. Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudík, and John Langford. Provably efficient RL with rich observations via latent state decoding. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 1665–1674. PMLR, 2019.

Botao Hao, Tor Lattimore, and Csaba Szepesvári. Adaptive exploration in linear contextual bandit. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pages 3536–3545. PMLR, 2020.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. *CoRR*, abs/2011.11566, 2020.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, 2010.

Zeyu Jia, Lin Yang, Csaba Szepesvári, and Mengdi Wang. Model-based reinforcement learning with value-targeted regression. In *L4DC*, volume 120 of *Proceedings of Machine Learning Research*, pages 666–686. PMLR, 2020.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *COLT*, volume 125 of *Proceedings of Machine Learning Research*, pages 2137–2143. PMLR, 2020.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. *CoRR*, abs/2102.00815, 2021.

Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737. PMLR, 2017.

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

Jonathan N. Lee, Aldo Pacchiano, Vidya Muthukumar, Weihao Kong, and Emma Brunskill. Online model selection for reinforcement learning with function approximation. In *AISTATS*, volume 130 of *Proceedings of Machine Learning Research*, pages 3340–3348. PMLR, 2021.

Rui Lu, Gao Huang, and Simon S. Du. On the power of multitask representation learning in linear MDP. *CoRR*, abs/2106.08053, 2021.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1928–1937. JMLR.org, 2016.

Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free representation learning and exploration in low-rank mdps. *CoRR*, abs/2102.07035, 2021.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *J. Mach. Learn. Res.*, 9:815–857, 2008.

Ronald Ortner, Odalric-Ambrym Maillard, and Daniil Ryabko. Selecting near-optimal approximate state representations in reinforcement learning. In *ALT*, volume 8776 of *Lecture Notes in Computer Science*, pages 140–154. Springer, 2014.

Ronald Ortner, Matteo Pirotta, Alessandro Lazaric, Ronan Fruit, and Odalric-Ambrym Maillard. Regret bounds for learning state representations in reinforcement learning. In *NeurIPS*, pages 12717–12727, 2019.

Matteo Papini, Andrea Tirinzoni, Marcello Restelli, Alessandro Lazaric, and Matteo Pirotta. Leveraging good representations in linear contextual bandits. *CoRR*, abs/2104.03781, 2021.

Antonin Raffin. Rl baselines3 zoo. `https://github.com/DLR-RM/rl-baselines3-zoo`, 2020.

Antonin Raffin, Ashley Hill, Maximilian Ernestus, Adam Gleave, Anssi Kanervisto, and Noah Dormann. Stable baselines3. `https://github.com/DLR-RM/stable-baselines3`, 2019.

Max Simchowitz and Kevin G. Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. In *NeurIPS*, pages 1151–1160, 2019.

Andrea Tirinzoni, Matteo Pirotta, Marcello Restelli, and Alessandro Lazaric. An asymptotically optimal primal-dual incremental algorithm for contextual linear bandits. *Advances in Neural Information Processing Systems*, 33, 2020.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4): 389–434, 2012.

Gellért Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In *ALT*, volume 132 of *Proceedings of Machine Learning Research*, pages 1237–1264. PMLR, 2021.

Haike Xu, Tengyu Ma, and Simon S. Du. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. *CoRR*, abs/2102.04692, 2021.

Lin F. Yang and Mengdi Wang. Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. *CoRR*, abs/1905.10389, 2019.

Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pages 1954–1964. PMLR, 2020a.

Andrea Zanette, Alessandro Lazaric, Mykel J. Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 10978–10989. PMLR, 2020b.

Andrea Zanette, Alessandro Lazaric, Mykel J. Kochenderfer, and Emma Brunskill. Provably efficient reward-agnostic navigation with linear value iteration. In *NeurIPS*, 2020c.

Zihan Zhang, Jiaqi Yang, Xiangyang Ji, and Simon S. Du. Variance-aware confidence set: Variance-dependent bound for linear bandits and horizon-free bound for linear mixture MDP. *CoRR*, abs/2101.12745, 2021.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvári. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. *CoRR*, abs/2012.08507, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See conclusions (Sec. 5).
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Due to its theoretical nature, we see no ethical or immediate societal consequence of our work
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] We perform simple numerical simulations to support our theoretical findings. The provided pseudocode and experimental protocol is sufficient to reproduce our results.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] See answer 3.(a)
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Related Work

The representation selection problem has been originally studied in the context of tabular MDPs. Given a set of representation mapping histories to (sequences of actions, observations, and rewards) to a finite set of states, the goal of the learning agent is to solve the MDP under an appropriate representation. The standard assumption is that at least one representation induces an MDP. Several papers have investigated this online learning problem and provided algorithms based on the optimism principle [e.g., Ortner et al., 2014, 2019]. The settings and the representation learning objective are different from ours. In particular, this line of research aims at finding any representation that is good for learning but the methods are not guaranteed to find the most efficient.

Recently, a few papers have focused on representation learning with theoretical guarantees. Du et al. [2019] considered the representation learning problem in block MDPs with rich observations, where the objective is to learn the compact latent representation. Representation learning in low-rank MDPs was recently studied in [Agarwal et al., 2020, Modi et al., 2021, Lu et al., 2021]. We believe that these papers are orthogonal to our work for several reasons. We start considering the setting in [Agarwal et al., 2020, Modi et al., 2021]. First, they operate in the reward-free setting where the objective is to learn a representation of the low-rank MDP that can be used to efficiently learn an optimal policy once a reward is given. For us, a reward is given from the start and learning/selecting a good representation in the meantime is just a way to suffer less regret. Second, our representation selection objective is different (and arguably more challenging) than the one considered by Agarwal et al. [2020], Modi et al. [2021]. They aim at finding a representation with low mean square error, i.e. any realizable representation of the low-rank MDP. On the other hand, we wish to find a UNISOFT representation among a set of realizable representations, which makes the representation learning problem harder. In App. F, Papini et al. [2021] showed that reducing the MSE is not enough for this purpose. It is shown that this only allows the algorithm to end up with a set of realizable representations, but after that, a different algorithmic scheme, whose primary objective is reducing regret (like LSVI-LEADER), is needed to find the UNISOFT one. Therefore, even if the approach in these papers could be extended to the regret minimization setting, there would be no guarantee that running that algorithm would recover a UNISOFT representation as in our case. Finally, it is unclear how to transform their sample complexity into a regret bound. In particular, it is not just a matter of translating a sample complexity bound into a regret bound: both the interaction protocol and the algorithmic schemes are different w.r.t. our work. Even if we directly translated the sample complexity bounds of these papers into regret bounds, we note that, while it is true that they could scale as $\log(|\Phi|)$, they also contain several dependencies which are orders of magnitude worse than in our work. For instance, the sample complexity provided in [Modi et al., 2021, Thm. 2] scales as A^{13} (A is the number of actions). This is also an unreasonable dependence in any case of practical interest we can think of. Finally, Lu et al. [2021] studied the effect of representation learning on the sample complexity in multi-task settings, which is quite different from the single-task regret minimization problem considered in this paper.

B Notation

Table 2: Notation.

$\mathcal{S}$		state space
$\mathcal{A}$		action space
H		episode length
r_h		reward function at stage h
p_h		transition function at stage h
μ		initial-state distribution
K		number of episodes
T	$=$	HK, total number of interactions
π_h		policy for stage h
Π		policy space
Q_h^π		state-action value function of policy π at stage h
V_h^π	$=$	$Q_h^\pi(s, \pi_h(s))$
$\pi_h^\star$		optimal policy for stage h
$Q_h^\star$	$=$	$Q_h^{\pi^\star}$, optimal value function at stage h
$V_h^\star(s)$	$=$	$\max_{a\in\mathcal{A}} Q_h^\star(s,a)$
L_h		Bellman's optimality operator for stage h
π_h^k		policy played by the algorithm at stage h of episode k
ϕ_h		feature map for stage h
$R(K)$		regret suffered in the first K episodes
d		feature dimension
D	$=$	H, value function upper bound
$\mathcal{Q}_h$		set of linear bounded value functions for stage h
Θ_h		set of parameters of linear bounded value functions for stage h
$\Delta_h(s,a)$	$=$	$V_h^\star(s) - Q_h^\star(s,a)$, suboptimality gap
$\Delta_{\min}$		minimum positive gap (see Asm. 3)
$\phi_h^\star(s)$	$=$	$\phi_h(s, \pi_h^\star(s))$, optimal features for state s at stage h
ρ_h^π		occupancy measure of policy π at stage h (see Asm. 4)
$\Lambda_h^\star$	$=$	$\mathbb{E}_{s\sim\rho_h^\star}[\phi_h^\star(s)\phi_h^\star(s)]$, optimal covariance matrix
λ_h^+		minimum nonzero eigenvalue of $\Lambda_h^\star$
δ		failure probability
$\overline{\kappa}$		last episode at which nonzero regret is paid (see proof of Thm. 19)
$\overline{\tau}$	$=$	$H\overline{\kappa}$
β_k		confidence radius, see (32) for ELEANOR and (42) for LSVI-UCB
λ	$=$	1, regularization parameter
Λ_h^k	$=$	$\lambda I + \sum_{i=1}^{k-1} \phi_h(s_h^i, a_h^i)\phi_h(s_h^i, a_h^i)^\mathsf{T}$, design matrix
$\overline{Q}_h^k$		optimistic value function for stage h at episode k
$\overline{V}_h^k(s)$	$=$	$\max_{a\in\mathcal{A}} \overline{Q}_h^k(s,a)$

C UNISOFT is Necessary: Proofs of Section 3.1

We illustrate all the detailed proofs needed for showing that the UNISOFT condition is necessary to achieve constant regret (Thm. 5). For the sake of completeness, we restate here all the assumptions on the MDP M under consideration.

Assumptions on MDP M.

- $\mathcal{S}$ and $\mathcal{A}$ finite, $H \geq 1$ arbitrary;
- Linear rewards: $r_h(s,a) = \langle \theta_h, \phi(s,a) \rangle$ with $\mathcal{N}(0,1)$ noise;
- Arbitrary transition probabilities $\{p_h\}_{h\in[H]}$ and initial-state distribution μ;
- Unique optimal policy $\pi^\star$: $|\{a : Q^\star_h(s,a) = V^\star_h(s)\}| = 1$ and $\pi^\star_h(s) = \operatorname{argmax}_a Q^\star_h(s,a)$ for all s, h;
- UNISOFT condition (Asm. 4 does not hold).

Moreover, recall that we define $\mathcal{M}$ as any set of MDPs that contains (but it can be larger than) all the MDPs which are equivalent to M in all components except for the reward parameters $\{\theta_h\}_{h\in[H]}$, which can be arbitrary vectors in $\mathbb{R}^d$. Formally,

$$\mathcal{M} \supseteq \left\{ \widetilde{M} = \left(\mathcal{S}, \mathcal{A}, H, \{\widetilde{r}_h\}_{h=1}^H, \{p_h\}_{h=1}^H, \mu\right) \mid \forall h \in [H], \exists \widetilde{\theta}_h \in \mathbb{R}^d : \widetilde{r}_h(s,a) = \langle \widetilde{\theta}_h, \phi(s,a) \rangle \right\}.$$

Intuitively, $\mathcal{M}$ contains at least all the MDPs that could be faced by an agent that knows the linear-reward structure of the problem but that does not know the true parameters $\{\theta_h\}_{h\in[H]}$. Obviously, if the agent knows all the components of M except for the reward parameters, the set $\mathcal{M}$ can be taken exactly as the set on the righthand side above (which would contain all and only the realizable MDPs). On the other hand, in the more general case where the agent does not know the dynamics as well, set $\mathcal{M}$ can be enlarged by including all the realizable MDPs with different transition probabilities (e.g., those with low-rank or low-IBE structure, or even the whole set of unstructured dynamics). Our proof that UNISOFT is necessary for constant regret holds for an agent that only knows that the true MDP M belongs to this general set $\mathcal{M}$ and thus encompasses all the relevant settings mentioned in Sec. 3.1.

In the following proofs we shall write $\mathbb{P}^{\mathsf{A}}_M$ ($\mathbb{E}^{\mathsf{A}}_M$) to denote the probability (expectation) operator under MDP M and the chosen algorithm A.

C.1 Proof of Lemma 7

Let M be our true MDP and $\widetilde{M} \in \mathcal{M}$ be any other MDP which is equivalent to M in all components except for the reward parameters, which are given by $\{\widetilde{\theta}_h\}_{h\in[H]}$. We start by a standard decomposition of the expected log-likelihood ratio between the observations generated in the two MDPs. Fix $K \geq 1$ and let $\mathrm{KL}(\mathbb{P}_M, \mathbb{P}_{\widetilde{M}})$ denote the KL-divergence between the distributions of the observations collected by algorithm A over K episodes. Using, e.g., Lemma 5 of Domingues et al. [2021] together with the closed-form of the KL divergence between Gaussian distributions,

$$\mathrm{KL}(\mathbb{P}_M, \mathbb{P}_{\widetilde{M}}) = \sum_{s,a} \sum_{h\in[H]} \mathbb{E}^{\mathsf{A}}_M[N^K_h(s,a)] \frac{(\langle \phi(s,a), \theta_h - \widetilde{\theta}_h \rangle)^2}{2} = \frac{1}{2} \sum_{h\in[H]} \|\theta_h - \widetilde{\theta}_h\|^2_{\mathbb{E}^{\mathsf{A}}_M[\Lambda^K_h]},$$

where $\Lambda^K_h := \sum_{s,a} N^K_h(s,a)\phi(s,a)\phi(s,a)^T$ and $N^K_h(s,a) := \sum_{k=1}^K \mathbb{1}\left\{s^k_h = s, a^k_h = a\right\}$.

Suppose that, for sufficiently large K, the matrix $\mathbb{E}^{\mathsf{A}}_M[\Lambda^K_h]$ is invertible.[8] We now proceed as follows. For a fixed $h \in [H]$ and sub-optimal policy $\pi \in \Pi$ (i.e., with $\Delta(\pi) > 0$), we seek the hardest MDP $\widetilde{M}$ to discriminate from M (i.e., that minimizes $\mathrm{KL}(\mathbb{P}_M, \mathbb{P}_{\widetilde{M}})$) where policy π is strictly better (in terms of expected return) than $\pi^\star$ and where we change only the parameter θ_h w.r.t. M. Formally, we minimize

$$\text{minimize}_{\widetilde{\theta}_h \in \mathbb{R}^d} \|\theta_h - \widetilde{\theta}_h\|^2_{\mathbb{E}^{\mathsf{A}}_M[\Lambda^K_h]}$$

[8] Lattimore and Szepesvari [2017] proved that this is indeed true for consistent algorithms. Otherwise, one could simply make the matrix positive-definite by adding λI for some arbitrary $\lambda > 0$ and the derivation still holds.

subject to the constraint $\widetilde{V}_1^{\pi} \geq \widetilde{V}_1^{\pi^\star} + \epsilon$. First note that the expected return of policy π can be equivalently written as

$$V_1^{\pi} = \sum_{s,a} \sum_{h \in [H]} \rho_h^{\pi}(s,a) r_h(s,a) = \sum_{h \in [H]} \langle \theta_h, \sum_{s,a} \rho_h^{\pi}(s,a) \phi(s,a) \rangle = \sum_{h \in [H]} \langle \theta_h, \Psi_h^{\pi} \rangle.$$

Moreover, since M and $\widetilde{M}$ have same transition probabilities, $\Psi_h^{\pi} = \widetilde{\Psi}_h^{\pi}$ for each π, h. Thus, $\widetilde{V}_1^{\pi} = \sum_{h \in [H]} \langle \widetilde{\theta}_h, \Psi_h^{\pi} \rangle$ and the constraint can be rewritten in the more convenient form

$$\sum_{h \in [H]} \langle \widetilde{\theta}_h, \Psi_h^{\pi} \rangle \geq \sum_{h \in [H]} \langle \widetilde{\theta}_h, \Psi_h^{\star} \rangle + \epsilon.$$

Using Lemma 13, the optimization problem has a closed-form expression. Therefore, let $\Gamma_h^{\epsilon}(\pi) \subseteq \mathcal{M}$ be the set of MDPs over which we are optimizing, that is, with (1) same transition probabilities as $\mathcal{M}$, (2) same reward parameters as $\mathcal{M}$ at all stages except h, and (3) $\widetilde{V}_1^{\pi} \geq \widetilde{V}_1^{\pi^\star} + \epsilon$. Using Lemma 13 together with the rewritings above, for any $\pi \in \Pi$, $h \in [H]$ and $\epsilon \geq 0$,

$$\min_{\widetilde{M} \in \Gamma_h^{\epsilon}(\pi)} \mathrm{KL}(\mathbb{P}_M, \mathbb{P}_{\widetilde{M}}) = \frac{(\Delta(\pi) + \epsilon)^2}{2 \|\Psi_h^{\pi} - \Psi_h^{\star}\|^2_{\mathbb{E}_M^{\mathsf{A}}[\Lambda_h^K]^{-1}}}. \tag{6}$$

We now show that $\mathrm{KL}(\mathbb{P}_M, \mathbb{P}_{\widetilde{M}})$ is lower bounded by a quantity that increases logarithmically in K for any $\widetilde{M} \in \Gamma_h^{\epsilon}(\pi)$ with $\epsilon > 0$. Let $E_K := \{\sum_{\pi \in \Pi^\star} N_K(\pi) < f(K)\}$, where $N_K(\pi) := \sum_{k=1}^{K} \mathbb{1}\left\{\pi^k = \pi\right\}$, $\Pi^\star$ is the set of all deterministic policies with maximal expected return in M, and $f(K)$ will be specified later. Using Lemma 14,

$$\mathrm{KL}(\mathbb{P}_M, \mathbb{P}_{\widetilde{M}}) \geq \log \frac{1}{\mathbb{P}_M(E_K) + \mathbb{P}_{\widetilde{M}}(E^c)} - \log 2. \tag{7}$$

Now note that, under the assumption that A is α-consistent,

$$c_M K^{\alpha} \geq \mathbb{E}_M^{\mathsf{A}}\left[R(K)\right] = \sum_{\pi \in \Pi} \mathbb{E}_M^{\mathsf{A}}\left[N_K(\pi)\right] \Delta(\pi) \geq \Delta \sum_{\pi \notin \Pi^\star} \mathbb{E}_M^{\mathsf{A}}\left[N_K(\pi)\right].$$

Here, with some abuse of notation, Δ is the minimum policy gap. Therefore,

$$\mathbb{P}_M(E_K) = \mathbb{P}_M\left(K - \sum_{\pi \notin \Pi^\star} N_K(\pi) < f(K)\right) \leq \frac{\sum_{\pi \notin \Pi^\star} \mathbb{E}_M^{\mathsf{A}}\left[N_K(\pi)\right]}{K - f(K)} \leq \frac{K^{\alpha} c_M / \Delta}{K - f(K)},$$

where the first inequality is Markov's inequality. Note that, since $\Psi_h^{\pi} = \Psi_h^{\star}$ for all optimal policies $\pi \in \Pi^\star$ and since the transition probablities of M and $\widetilde{M}$ are the same, $\widetilde{V}_1^{\pi} = \widetilde{V}_1^{\pi^\star}$ for all $\pi \in \Pi^\star$. Hence, all optimal policies for M have a gap of at least ϵ in $\widetilde{M}$. This implies that

$$c_{\widetilde{M}} K^{\alpha} \geq \mathbb{E}_{\widetilde{M}}^{\mathsf{A}}\left[R(K)\right] \geq \epsilon \mathbb{E}_{\widetilde{M}}^{\mathsf{A}}\left[\sum_{\pi \in \Pi^\star} N_K(\pi)\right].$$

Therefore,

$$\mathbb{P}_{\widetilde{M}}(E_K^c) = \mathbb{P}_{\widetilde{\mathcal{M}}}\left(\sum_{\pi \in \Pi^\star} N_K(\pi) \geq f(K)\right) \leq \frac{\mathbb{E}_{\widetilde{M}}^{\mathsf{A}}\left[\sum_{\pi \in \Pi^\star} N_K(\pi)\right]}{f(K)} \leq \frac{K^{\alpha} c_{\widetilde{M}} / \epsilon}{f(K)}.$$

If we set $f(K) = K/2$ and plug the two bounds above into (7), we obtain

$$\mathrm{KL}(\mathbb{P}_M, \mathbb{P}_{\widetilde{M}}) \geq \log \frac{K^{1-\alpha}}{2c_M/\Delta + 2c_{\widetilde{M}}/\epsilon} - \log 2.$$

Finally, for any $\widetilde{M} \in \Gamma_h^{\epsilon}(\pi)$ with $\epsilon > 0$,

$$\liminf_{K \to \infty} \frac{\mathrm{KL}(\mathbb{P}_M, \mathbb{P}_{\widetilde{M}})}{\log(K)} \geq 1 - \alpha.$$

This holds for any $\epsilon > 0$. Hence, in combination with (6), we proved that, for any sub-optimal policy π and stage h,

$$\liminf_{K \to \infty} \frac{1}{\log(K)} \frac{\Delta(\pi)^2}{2 \|\Psi_h^{\pi} - \Psi_h^{\star}\|^2_{\mathbb{E}_M^{\mathsf{A}}[\Lambda_h^K]^{-1}}} \geq 1 - \alpha.$$

Rearranging concludes the proof.

C.2 Proof of Theorem 5

We now use Lemma 7 to prove that the UNISOFT condition is necessary for constant regret. We proceed in different steps.

Step 1. Controlling the design matrix. Suppose that the algorithm suffers constant regret on instance M. This means that, for some constant C_M (different from the c_M used in the definition of α-consistence),

$$\mathbb{E}_M^{\mathsf{A}}\left[\mathrm{R}(K)\right] \leq C_M. \tag{8}$$

Since $\mathbb{E}_M^{\mathsf{A}}\left[\mathrm{R}(K)\right] = \sum_h \sum_{s,a} \mathbb{E}_M^{\mathsf{A}}\left[N_h^K(s,a)\right] \Delta_h(s,a)$, we have that $\sum_h \sum_{s,a\neq\pi_h^\star(s)} \mathbb{E}_M^{\mathsf{A}}\left[N_h^K(s,a)\right] \leq C_M/\Delta_{\min}$, where $\Delta_{\min}$ is the minimum value-function gap. Therefore, the expected design matrix at each $h \in [H]$ satifies

$$\begin{aligned}
\mathbb{E}_M^{\mathsf{A}}[\Lambda_h^K] &= \sum_{s,a} \mathbb{E}_M^{\mathsf{A}}[N_h^K(s,a)]\phi(s,a)\phi(s,a)^T \\
&= \sum_s \mathbb{E}_M^{\mathsf{A}}[N_h^K(s,\phi_h^\star(s))]\phi_h^\star(s)\phi_h^\star(s)^T + \sum_{s,a\neq\pi_h^\star(s)} \mathbb{E}_M^{\mathsf{A}}[N_h^K(s,a)]\phi(s,a)\phi(s,a)^T \\
&\preceq \sum_s \mathbb{E}_M^{\mathsf{A}}[N_h^K(s)]\phi_h^\star(s)\phi_h^\star(s)^T + L^2\frac{C_M}{\Delta_{\min}} I \\
&\preceq K \sum_{s:\rho_h^\star(s)>0} \phi_h^\star(s)\phi_h^\star(s)^T + \sum_{s:\rho_h^\star(s)=0} \mathbb{E}_M^{\mathsf{A}}[N_h^K(s)]\phi_h^\star(s)\phi_h^\star(s)^T + L^2\frac{C_M}{\Delta_{\min}} I.
\end{aligned}$$

We now bound the expected number of times the algorithm visit states which are not visited by an optimal policy. Take any s such that $\rho_h^\star(s) = 0$. Since any optimal policy has the same state distribution $\rho_h^\star$, the event $s_h^k = s$ implies that $\pi^k \notin \Pi^\star$. Therefore,

$$\mathbb{E}_M^{\mathsf{A}}[N_h^K(s)] = \mathbb{E}_M^{\mathsf{A}}[\sum_{k=1}^K \mathbb{1}\left\{s_h^k = s\right\}] \leq \mathbb{E}_M^{\mathsf{A}}[\sum_{k=1}^K \mathbb{1}\left\{\pi^k \notin \Pi^\star\right\}] = \mathbb{E}_M^{\mathsf{A}}[\sum_{\pi\notin\Pi^\star} N_K(\pi)].$$

Moreover, since the algorithm suffers constant regret,

$$\Delta\mathbb{E}_M^{\mathsf{A}}[\sum_{\pi\notin\Pi^\star} N_K(\pi)] \leq \mathbb{E}_M^{\mathsf{A}}\left[\mathrm{R}(K)\right] \leq C_M.$$

Therefore, we conclude that

$$\mathbb{E}_M^{\mathsf{A}}[\Lambda_h^K] \preceq K \sum_{s:\rho_h^\star(s)>0} \phi_h^\star(s)\phi_h^\star(s)^T + L^2\left(\frac{C_M}{\Delta_{\min}} + S_h\frac{C_M}{\Delta}\right) I,$$

where $S_h := S - |\mathrm{supp}(\rho_h^\star))|$.

Step 2. Controlling the feature expectations. We now show that, since UNISOFT does not hold, there exists a sub-optimal policy π such that Ψ_h^π is not in the span of the optimal features. By directly using the definition of UNISOFT (Asm. 4), we have that there must exist a state-action pair s, a which is reachable at time h (i.e., $\exists\pi \in \Pi : \rho_h^\pi(s,a) > 0$) such that $\phi(s,a) \notin \mathrm{span}\left\{\phi_h^\star(s)|\rho_h^\star(s) > 0\right\}$. Clearly, we have only two cases:

1. $\rho_h^\star(s) > 0$ and $a \neq \pi_h^\star(s)$;
2. $\rho_h^\star(s) = 0$ and a is arbitrary (even an optimal action).

For Case 1, simply take a policy π that is equivalent to $\pi^\star$ everywhere except that $\pi_h(s) = a$. Clearly, the policy is sub-optimal, in the sense that $\Delta(\pi) = V_1^\star - V_1^\pi > 0$. Moreover, it is easy to check that $\Psi_h^\pi - \Psi_h^\star = \rho_h^\star(s)(\phi(s,a) - \phi_h^\star(s))$. Therefore, $\Psi_h^\pi \notin \mathrm{span}\left\{\phi_h^\star(s)|\rho_h^\star(s) > 0\right\}$.

For Case 2, choose π in such a way that $\rho_h^\pi(s) > 0$ (we know that one such policy exists due to the reachability of s). This only requires selecting the actions of π for all stages $h' < h$. For all stages

$h' > h$, set π equal to $\pi^\star$ except for $\pi_h(s) = a$. Note that, even if a is optimal at time h, π is strictly sub-optimal (i.e., $\Delta(\pi) > 0$) since no optimal policy can achieve the condition $\rho_h^\pi(s) > 0$ by the uniqueness of the optimal state distribution. Moreover,

$$\begin{aligned}\Psi_h^\pi - \Psi_h^\star &= \sum_{s',a'} \rho_h^\pi(s',a')\phi(s',a') - \sum_{s'} \rho_h^\star(s')\phi_h^\star(s') \\ &= \rho_h^\pi(s)\phi(s,a) - \underbrace{\rho_h^\star(s)}_{=0}\phi_h^\star(s) + \sum_{s' \neq s} (\rho_h^\pi(s',a') - \rho_h^\star(s'))\phi_h^\star(s').\end{aligned}$$

Thus, we still conclude $\Psi_h^\pi \notin \operatorname{span}\{\phi_h^\star(s) | \rho_h^\star(s) > 0\}$.

Step 3. Concluding the proof. Combining Lemma 7 with Step 1 and Step 2, we have that, for some $h \in [H]$ and policy π such that $\Delta(\pi) > 0$ and $\Psi_h^\pi \notin \operatorname{span}\{\phi_h^\star(s) | \rho_h^\star(s) > 0\}$,

$$\limsup_{K\to\infty} \log(K) \|\Psi_h^\pi - \Psi_h^\star\|^2_{(\Lambda_h^\star + \eta I)^{-1}} \leq \frac{\Delta(\pi)^2}{2(1-\alpha)},$$

where $\Lambda_h^\star := K \sum_{s:\rho_h^\star(s)>0} \phi_h^\star(s)\phi_h^\star(s)^T$ and $\eta := L^2\left(\frac{C_M}{\Delta_{\min}} + S_h \frac{C_M}{\Delta}\right) > 0$. Using Lemma 34, we have that there exists an $\epsilon > 0$ (independent of K) such that $\|\Psi_h^\pi - \Psi_h^\star\|_{(\Lambda_h^\star+\eta I)^{-1}} \geq \frac{\epsilon}{\sqrt{\eta}}$. Therefore, we get that

$$\limsup_{K\to\infty} \log(K) \leq \frac{\eta \Delta(\pi)^2}{2\epsilon^2(1-\alpha)},$$

which clearly does not hold since the left-hand side grows with K while the right-hand side is constant. Therefore, we have a contradiction, and the algorithm A cannot achieve constant regret on this non-UNISOFT instance while being consistent on all other instances in $\mathcal{M}$. Our claim that UNISOFT is necessary follows.

C.3 Auxiliary Results

Lemma 13. *Let $A \in \mathbb{R}^{d\times d}$ be any positive semi-definite invertible matrix. For $\pi \in \Pi$, $h \in [H]$, and $\epsilon \geq 0$, consider the following optimization problem:*

$$\begin{aligned}\min_{\theta\in\mathbb{R}^d} \quad & \|\theta - \theta_h\|_A^2 \\ \text{subject to} \quad & \sum_{l\in[H], l\neq h} \langle \theta_l, \Psi_l^\pi - \Psi_l^\star\rangle + \langle \theta, \Psi_h^\pi - \Psi_h^\star\rangle \geq \epsilon\end{aligned}$$

Then, for $\overline{\theta}$ a minimizer we have

$$\|\overline{\theta} - \theta_h\|_A^2 = \frac{(\Delta(\pi)+\epsilon)^2}{\|\Psi_h^\pi - \Psi_h^\star\|^2_{A^{-1}}}.$$

Proof. To simplify notation, let us define $b := \sum_{l\in[H], l\neq h} \langle \theta_l, \Psi_l^\pi - \Psi_l^\star\rangle$. The corresponding Lagrange dual problem is

$$\max_{\lambda\geq 0} \min_{\theta\in\mathbb{R}^d} \left\{\|\theta - \theta_h\|_A^2 - \lambda\left(\langle \theta, \Psi_h^\pi - \Psi_h^\star\rangle + b - \epsilon\right)\right\}.$$

Let $f(\theta, \lambda)$ denote the resulting objective function. Taking the gradient w.r.t. θ,

$$\nabla_\theta f(\theta,\lambda) = 2A(\theta - \theta_h) - \lambda(\Psi_h^\pi - \Psi_h^\star),$$

and equating it to zero, we obtain

$$\theta = \theta_h + \frac{\lambda}{2} A^{-1}(\Psi_h^\pi - \Psi_h^\star).$$

Plugging this back to the original objective we get

$$
\begin{aligned}
f(\lambda) &= \frac{\lambda^2}{4}\|A^{-1}(\Psi_h^\pi - \Psi_h^\star)\|_A^2 - \lambda\left(\langle\theta_h, \Psi_h^\pi - \Psi_h^\star\rangle + \frac{\lambda}{2}\|\Psi_h^\pi - \Psi_h^\star\|_{A^{-1}}^2 + b - \epsilon\right) \\
&= -\frac{\lambda^2}{4}\|\Psi_h^\pi - \Psi_h^\star\|_{A^{-1}}^2 - \lambda\left(\langle\theta_h, \Psi_h^\pi - \Psi_h^\star\rangle + \sum_{l\in[H], l\neq h}\langle\theta_l, \Psi_l^\pi - \Psi_l^{\pi^\star}\rangle - \epsilon\right) \\
&= -\frac{\lambda^2}{4}\|\Psi_h^\pi - \Psi_h^\star\|_{A^{-1}}^2 + \lambda\left(\Delta(\pi) + \epsilon\right).
\end{aligned}
$$

Differentiating with respect to λ and equating to zero we obtain

$$
\lambda = \frac{2\left(\Delta(\pi) + \epsilon\right)}{\|\Psi_h^\pi - \Psi_h^\star\|_{A^{-1}}^2}.
$$

Therefore, plugging this back into the objective value

$$
\|\overline{\theta} - \theta_h\|_A^2 = \frac{\left(\Delta(\pi) + \epsilon\right)^2}{\|\Psi_h^\pi - \Psi_h^\star\|_{A^{-1}}^2}.
$$

□

Lemma 14 (Bretagnolle–Huber inequality, see, e.g., Thm. 14.2 of Lattimore and Szepesvári [2020])**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on the same measurable space $(\Omega, \mathcal{F})$ and let $E \in \mathcal{F}$ be an arbitrary event. Then,*

$$
\mathbb{P}(E) + \mathbb{Q}(E^c) \geq \frac{1}{2}e^{-\mathrm{KL}(\mathbb{P},\mathbb{Q})}.
$$

D UNISOFT is Sufficient: Proofs of Section 3.2

We first prove that UNISOFT is sufficient for a whole class of algorithms, as done in the proof sketch of Section 3.2. We will then instantiate this result to ELEANOR and LSVI-UCB.

Consider the following assumptions.

Assumption 15. *Consider a feature map $\{\phi_h\}_{h\in[H]}$ and a Q-function estimate $\overline{Q}_h^k$. There is an event $G(\delta)$ that holds with probability at least $1-\delta$ under which:*

(a) Global optimism: $\overline{V}_1^k(s) \geq V_1^\star(s)$ where $\overline{V}_h^k(s) = \max_{a\in\mathcal{A}}\{\overline{Q}_h^k(s,a)\}$,

(b) Confidence set: let $\Lambda_h^k = \sum_{i=1}^{k-1}\phi_h(s_h^i, a_h^i)\phi_h(s_h^i, a_h^i)^\mathsf{T} + \lambda I$ and $\beta_k \in \mathbb{R}_+$ be increasing and logarithmic in k, then $\overline{V}_h^k(s_h^k) - V_h^{\pi^k}(s_h^k) \leq 2\beta_k\left\|\phi_h(s_h^k, a_h^k)\right\|_{(\Lambda_h^k)^{-1}} + \mathbb{E}_{s'\sim p_h(s_h^k, a_h^k)}\left[\overline{V}_{h+1}^k(s') - V_{h+1}^{\pi^k}(s')\right]$,

simultaneously for all $h \in [H]$, $k \geq 1$ and $s \in \mathcal{S}$, where $\delta \in (0,1)$ is a parameter of the algorithm.

Assumption 16. *The algorithm satisfies Assumption 15, and additionally there exist a set of constants $(C_h)_{h\in[H]}$ such that, under the event $G(\delta)$:*

(c) (Almost) local optimism: $\overline{Q}_h^k(s,a) + C_h\beta_k\left\|\phi_h(s,a)\right\|_{(\Lambda_h^k)^{-1}} \geq Q_h^\star(s,a)$,

for all $h = 2, \ldots, H$, $k \geq 1$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

Assumption 16 characterizes the class of algorithms for which we are going to prove a constant bound on the regret under UNISOFT. However, we first study the regret under the weaker Assumption 15, following the proof pattern from [Jin et al., 2020].

Lemma 17. *Under Assumption 15, assuming event $G(\delta)$ holds, there exists a $\widetilde{O}(\sqrt{K})$ function g such that, with probability $1-\delta$, for all $K \geq 1$:*

$$
R(K) \leq H\beta_K\sqrt{2dK\log(1+K/\lambda)} + 2H^2\sqrt{K\log(2HK/\delta)} = \widetilde{O}(\sqrt{K}). \tag{9}
$$

Proof. Under event $G(\delta)$:

$$
\begin{aligned}
R(K) &= \sum_{k=1}^{K} V_1^\star(s_1^k) - V_1^{\pi^k}(s_1^k) \\
&\leq \sum_{k=1}^{K} \overline{V}_1^k(s_1^k) - V_1^{\pi^k}(s_1^k) && \text{(a)} && (10) \\
&\leq \underbrace{2\sum_{h=1}^{H} \beta_K \sum_{k=1}^{K} \left\|\phi_h(s_h^k, a_h^k)\right\|_{(\Lambda_h^k)^{-1}}}_{(A)} + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} \zeta_h^k}_{(B)}, && && (11)
\end{aligned}
$$

where the last inequality is from recursive application of (b) and the fact that β_k is increasing, and:

$$
\zeta_h^k = \mathbb{E}_{s' \sim p_h(s_h^k, a_h^k)}[\overline{V}_{h+1}^k(s') - V_{h+1}^{\pi^k}(s')] - \overline{V}_{h+1}^k(s_{h+1}^k) + V_{h+1}^{\pi^k}(s_{h+1}^k), \tag{12}
$$

where expectations are conditioned on the history up to the beginning of episode k. We bound (A) using the Elliptical Potential Lemma [e.g., Abbasi-Yadkori et al., 2011]:

$$
\begin{aligned}
(A) &= 2\beta_K \sum_{h=1}^{H}\sum_{k=1}^{K} \left\|\phi_h(s_h^k, a_h^k)\right\|_{(\Lambda_h^k)^{-1}} && (13) \\
2\beta_K \sum_{h=1}^{H} &\leq \sqrt{K \sum_{k=1}^{K} \left\|\phi_h(s_h^k, a_h^k)\right\|^2_{(\Lambda_h^k)^{-1}}} && (14) \\
&\leq H\beta_K \sqrt{2dK\log(1+K/\lambda)}. && (15)
\end{aligned}
$$

Since ζ_h^k is a martingale difference sequence with $\zeta_h^k \leq 2H$, we can use Azuma's inequality (Prop. 27) to bound (B):

$$
\sum_{k=1}^{K} \zeta_h^k \leq 2H\sqrt{K\log(2K/\delta_h)}, \tag{16}
$$

with probability $1-\delta_h$ for all $K \geq 1$. To make it hold with probability $1-\delta$ for all $h \in [H]$, we set $\delta_h = \delta/H$. Finally:

$$
(B) = \sum_{h=1}^{H}\sum_{k=1}^{K} \zeta_h^k \leq 2H^2\sqrt{K\log(2HK/\delta)}. \tag{17}
$$

□

The stronger Assumption 16 is needed to upper-bound the gaps.

Lemma 18. *Under Assumption 16, assuming event $G(\delta)$ holds, for all $s \in \mathcal{S}$, $h \in [H]$ and $k \geq 1$:*

$$
\Delta_h(s, \pi_h^k(s)) \leq 2\,\mathbb{E}_{\pi^k}\left[\sum_{i=h}^{H} \beta_k \left\|\phi_i(s_i, a_i)\right\|_{(\Lambda_i^k)^{-1}} \middle| s_h = s\right] + \mathbb{1}\{h > 1\}\, C_h \beta_k \left\|\phi^\star(s)\right\|_{(\Lambda_h^k)^{-1}}.
$$

Proof.

$$\Delta_h(s,\pi_h^k(s)) = V_h^\star(s) - Q_h^\star(s_h^k,\pi_h^k(s)) \tag{18}$$
$$\leq V_h^\star(s) - Q_h^{\pi^k}(s_h^k,\pi_h^k(s)) \tag{19}$$
$$= V_h^\star(s) - V_h^{\pi^k}(s) \tag{20}$$
$$= Q_h^\star(s,\pi_h^\star(s)) - V_h^{\pi^k}(s) \tag{21}$$
$$\leq \overline{Q}_h^k(s,\pi_h^\star(s)) + \mathbb{1}\{h>1\}\, C_h\beta_k \left\|\phi_h(s,\pi_h^\star(s))\right\|_{(\Lambda_h^k)^{-1}} - V_h^{\pi^k}(s) \tag{22}$$
$$\leq \overline{V}_h^k(s) + \mathbb{1}\{h>1\}\, C\sqrt{\gamma_{hk}}\left\|\phi_h(s,\pi_h^\star(s))\right\|_{\Lambda_{hk}^{-1}} - V_h^{\pi^k}(s) \tag{23}$$
$$\leq 2\,\mathbb{E}_{\pi^k}\left[\sum_{i=h}^H \beta_k \left\|\phi_i(s_i,a_i)\right\|_{(\Lambda_i^k)^{-1}} \middle| s_h = s\right] \tag{24}$$
$$+ \mathbb{1}\{h>1\}\, C_h\beta_k \left\|\phi_h(s_h^k,\pi_h^\star(s_h^k))\right\|_{(\Lambda_h^k)^{-1}},$$

where (22) uses (a) for $h=1$ and (c) for $h>1$, while the last inequality is from recursive application of (b). □

Now we can prove our main result on constant regret:

Theorem 19. *Any algorithm satisfying Assumption 16 enjoys constant regret if the representation has the* UNISOFT *property (Asm. 4) and Assumption 3 on the minimum gap holds. In general, let $g:\mathbb{N}\to\mathbb{R}_+$ be any increasing $\widetilde{O}(\sqrt{K})$ function such that, with probability $1-2\delta$ for all $K\geq 1$, $R(K)\leq g(K)$. Then, under Assumptions 3, 4, 16, with probability $1-3\delta$ for all $K\geq 1$:*

$$R(K) \leq g(\overline{\kappa}) = O(1), \tag{25}$$

where $\overline{\kappa}$ is a constant independent of K.

Proof. First notice that a valid regret upper bound $g(K)$ always exists due to Lemma 17. Moreover, due to Asm. 4, for all $h\in[H]$ and $k\geq 1$, we have $\phi_h(s,\pi_h^k(s))\in \operatorname{span}\{\phi_h^\star(s)|\rho_h^\star(s)>0\}$ for all $s\in\mathcal{S}$ such that $\rho_h^{\pi^k}(s)>0$. Hence, with probability $1-2\delta$, the requirements of Lemma 33 are satisfied and we can apply it to the gap upper bound from Lemma 18. So, with probability $1-3\delta$, for all $s\in\mathcal{S}$, $h\in[H]$ and $k\geq\widetilde{\kappa}=\max_{h\in[H]}\widetilde{\kappa}_h$:

$$\Delta_h(s,\pi_h^k(s)) \leq 2\,\mathbb{E}_{\pi^k}\left[\sum_{i=h}^H \beta_k \left\|\phi_i(s_i,a_i)\right\|_{(\Lambda_i^k)^{-1}} \middle| s_h = s\right]$$
$$+ \mathbb{1}\{h>1\}\, C_h\beta_k \left\|\phi^\star(s)\right\|_{(\Lambda_h^k)^{-1}} \tag{26}$$
$$\leq (2+\mathbb{1}\{h>1\}\, C_h)\beta_k \sum_{i=h}^H \frac{k+\lambda-g(k)-8\sqrt{k\log(2dHk/\delta)}}{(k\lambda_i^+ + \lambda - g(k) - 8\sqrt{k\log(2dHk/\delta)})^{3/2}}. \tag{27}$$

Assume for now that $k\geq\widetilde{\kappa}$. From the previous inequality, since $g(k)=\widetilde{O}(\sqrt{k})$ and $\beta_k=\widetilde{O}(1)$, there exists a κ_h independent of K such that, for $k>\kappa_h$:

$$\Delta_h(s,\pi_h^k(s)) \leq \Delta_{\min}. \tag{28}$$

Under Asm. 3, this implies $\Delta_h(s,\pi_h^k(s))=0$. Let $\overline{\kappa}=\max\{\widetilde{\kappa},\max_h\{\kappa_h\}\}$. For $k>\overline{\kappa}$, all the gaps are zero. Finally, by Prop. 29:

$$R(K) = \sum_{k=1}^K \mathbb{E}_{\pi^k}\left[\sum_{h=1}^H \Delta_h(s_h,a_h) \middle| s_1 = s_1^k\right] \tag{29}$$
$$= \sum_{k=1}^{\overline{\kappa}} \mathbb{E}_{\pi^k}\left[\sum_{h=1}^H \Delta_h(s_h,a_h) \middle| s_1 = s_1^k\right] + \sum_{k=\overline{\kappa}+1}^K \mathbb{E}_{\pi^k}\left[\sum_{h=1}^H \underbrace{\Delta_h(s_h,a_h)}_{=0} \middle| s_1 = s_1^k\right] \tag{30}$$
$$= R(\overline{\kappa}) \leq g(\overline{\kappa}). \tag{31}$$

□

Finally, we instantiate the general result of 19 to ELEANOR on MDPs with Bellman closure and LSVI-UCB on low-rank MDPs, by showing that they satisfy Assumption 16.

Proof of Theorem 8.

Let:

$$\beta_k = H\sqrt{\frac{d}{2}\log(1+k/d) + d\log(1+4\sqrt{dk}) + \log\frac{2Hk^2}{\delta}} + 1, \tag{32}$$

and define event $G(\delta)$ as in Lemma 2 from [Zanette et al., 2020b]. We have (a) by Lemma 7 from [Zanette et al., 2020b], while (b) can be extracted from the proof of Theorem 1 from [Zanette et al., 2020b]. To prove (c), we use the fact that the MDP satisfies Bellman closure, hence there exist $\theta_1^\star, \dots, \theta_H^\star$ such that [Lemma 6 from Zanette et al., 2020b]:

$$Q_h^\star(s,a) = \phi_h(s,a)^\mathsf{T}\theta_h^\star. \tag{33}$$

By Lemma 7 from [Zanette et al., 2020b], $\theta_1^\star, \dots, \theta_H^\star$ is a feasible solution for $\overline{\theta}_1, \dots, \overline{\theta}_H$ in ELEANOR's program [Definition 2 from Zanette et al., 2020b]. Due to the program's constraints:

$$\left\|\theta_h^\star - \widehat{\theta}_h^k\right\|_{\Lambda_h^k} \le \beta_k. \tag{34}$$

Let $\overline{\theta}_1^k, \dots, \overline{\theta}_H^k$ be the values that are actually selected by ELEANOR's program. Since they are subject to the same constraints, by the triangular inequality:

$$\left\|\theta_h^\star - \overline{\theta}_h^k\right\|_{\Lambda_h^k} \le 2\beta_k. \tag{35}$$

Finally, since $\overline{Q}_h^k(s,a) = \phi_h(s,a)^\mathsf{T}\overline{\theta}_h^k$:

$$\begin{aligned}
Q_h^\star(s_h,a_h) &= \phi_h(s_h,a_h)^\mathsf{T}\theta_h^\star && (36)\\
&= \phi_h(s_h,a_h)^\mathsf{T}\overline{\theta}_h^k + \phi_h(s_h,a_h)^\mathsf{T}(\theta_h^\star - \overline{\theta}_h^k) && (37)\\
&\le \overline{Q}_h^k(s_h,a_h) + \|\phi(s_h,a_h)\|_{(\Lambda_h^k)^{-1}} \left\|\theta_h^\star - \overline{\theta}_h^k\right\|_{\Lambda_h^k} && (38)\\
&\le \overline{Q}_h^k(s_h,a_h) + 2\beta_k \|\phi(s_h,a_h)\|_{(\Lambda_h^k)^{-1}}, && (39)
\end{aligned}$$

so (c) holds with $C_h = 2$. So Asm. 16 holds and we can invoke Theorem 19 with the upper bound g from Lemma 17 and the β_k given above to obtain:

$$\begin{aligned}
R(K) &\le H^2\left(\sqrt{\frac{d}{2}\log(1+\overline{\kappa}/d) + d\log(1+4\sqrt{d\overline{\kappa}}) + \log(H\overline{\kappa}^2) + \log\frac{2}{\delta}} + H\right)\\
&\quad\times\sqrt{2d\overline{\kappa}\log(1+\overline{\kappa}/\lambda)} + 2H^2\sqrt{\overline{\kappa}\log(2H\overline{\kappa}/\delta)} && (40)\\
&\lesssim H^{3/2}d\sqrt{\overline{\tau}\log\frac{\overline{\tau}}{\delta}}, && (41)
\end{aligned}$$

where $\overline{\tau} = H\overline{\kappa}$. □

Remark 1. We have slightly modified the ELEANOR algorithm to obtain any-time regret bounds. In particular, we have replaced the fixed $\delta' = \delta/(2T)$ term in the original β_k (see the proof of Lemma 2 in [Zanette et al., 2020b]) with the adaptive $\delta/(2Hk^2)$. This still makes event $G(\delta)$ hold with probability $1-\delta$, but without knowledge of the horizon K. This only affects logarithmic terms. Also notice that we have considered the case of zero inherent Bellman error ($\mathcal{I} = 0$), which corresponds to Bellman closure, and we have taken $[0, H]$, not $[0, 1]$, as the range of the value function (see the comment following Theorem 1 in [Zanette et al., 2020b]).

For LSVI-UCB, we can instantiate Theorem 19 with the problem-dependent logarithmic lower bound by He et al. [2020] in place of the worst-case upper bound from Lemma 17.

Proof of Theorem 9.

Let:

$$\beta_k = c_\beta dH\sqrt{\log(2dHk/\delta)}, \tag{42}$$

where c_β is a constant defined in Lemma C.3 from [Jin et al., 2020], and define event $G(\delta)$ as in Lemma B.3 from [Jin et al., 2020]. Then since the MDP is low-rank, by Lemma B.5 from [Jin et al., 2020] we have both (a) and (c) with $C_h = 0$. We get (b) by Lemma B.4 from Jin et al. [2020]. So Asm. 16 holds and, under Asm 3, we can instantiate Theorem 19 with the logarithmic regret bound from Theorem 4.4 by He et al. [2020]:

$$g(k) = 9HG(k)\log G(k) + \frac{16H^2}{3}\log\frac{\log\lceil Hk\rceil}{\delta} + 2, \tag{43}$$

where:

$$G(k) \propto \frac{d^3H^4\log(4dH^2k(k+1)\log(H/\Delta_{\min})/\delta)}{\Delta_{\min}}. \tag{44}$$

So:

$$R(K) \leq g(\overline{\kappa}) \simeq \frac{d^3H^5}{\Delta_{\min}}\log\left(dH^2\overline{\kappa}/\delta\right). \tag{45}$$

□

Remark 2. We have slightly modified the LSVI-UCB algorithm to obtain any-time regret bounds. In particular, we have replaced the fixed $\iota = \log(2dT/\delta)$ term in the original β_k (see Theorem 3.1 from [Jin et al., 2020]) with the adaptive $\log(4dHk^2/\delta)$. This still makes event $G(\delta)$ hold with probability $1-\delta$, but without knowledge of the horizon K. We have also re-written the logarithmic regret bound by He et al. [2020] (Theorem 4.4) to hold with probability $1-2\delta$. These changes only affect logarithmic terms.

Lemma 20. *The critical time $\overline{\kappa}$ from Theorem 9 for LSVI-UCB is upper bounded as:*

$$\overline{\kappa} \leq \max\left\{\frac{48c_1^2H^4d^3}{\lambda_+^2}\log\left(\frac{32c_1^2H^5d^4}{\lambda_+^2\delta}\right), \frac{432c_2^2H^4d^2}{\Delta_{\min}^2\lambda_+^3}\log\left(\frac{288d^3H^5c_2^2}{\Delta_{\min}^2\lambda_+^3\delta}\right)\right\} \tag{46}$$

where $\lambda_+ = \min_{h\in[H]}\{\lambda_h^+\}$ and c_1, c_2 are universal constants.

Proof. For LSVI-UCB we have (see the proof of Theorem 9):

$$g(k) \leq c_1H^2d^{3/2}\sqrt{k\log(2dHk/\delta)}, \tag{47}$$

$$\beta_k = c_2dH\sqrt{\log(2dHk/\delta)}, \tag{48}$$

for some universal constants c_1, c_2. We assume $\lambda = 1$ and $c_1 \geq 8$.

We will use the fact that a sufficient condition for $k \geq a\log(bk)$ is $k \geq 3a\log(ab)$ for $k > 0$ and reasonable values of the constants a, b. See App. C.6 from Papini et al. [2021] for details. This immediately implies that a sufficient condition for $k \geq a\sqrt{k\log(bk)}$ is:

$$k \geq 3a^2\log(a^2b) \tag{49}$$

We divide the rest of the proof in three parts:

Part 1. First, κ must satisfy the invertibility conditions from Lemma 33. To make matrix $B_h^k = k\Lambda_h^\star + \lambda I - g(k) + 8\sqrt{k\log(2dHk/\delta)}$ invertible for each h, we first require the positive eigenvalues of $\Lambda_h^\star$ to map into positive eigenvalues of B_h^k. A sufficient condition for this is:

$$k\lambda_+ > 1 + g(k) + 8\sqrt{k\log(2dHk/\delta)} \tag{50}$$

$$k \geq \frac{c_1H^2d^{3/2}+8}{\lambda_+}\sqrt{k\log(2dHk/\delta)} \tag{51}$$

$$k \geq \frac{2c_1H^2d^{3/2}}{\lambda_+}\sqrt{k\log(2dHk/\delta)} \tag{52}$$

$$k \geq \frac{12c_1^2H^4d^3}{\lambda_+^2}\log\left(\frac{8c_1^2H^5d^4}{\lambda_+^2\delta}\right) \triangleq \overline{\kappa}_1, \tag{53}$$

where the latter is from (49). We also need the zero eigenvalues of $\Lambda_h^\star$ to map into negative eigenvalues of B_h^k. However, this just requires $\lambda - g(k) + 8\sqrt{k\log(2dHk/\delta)} < 0$ which is already true for $k = 1$ given $\lambda = 1$.

Part 2. We require $\overline{\kappa}$ to satisfy the following, which will make the analysis of Part 3 easier:

$$g(k) + 8\sqrt{k\log(2dHk/\delta)} \leq \frac{k\lambda_+}{2}. \tag{54}$$

□

After rearranging, we can proceed precisely as in Part 1, only with different numerical constants, obtaining:

$$k \geq \frac{48c_1^2H^4d^3}{\lambda_+^2}\log\left(\frac{32c_1^2H^5d^4}{\lambda_+^2\delta}\right) \triangleq \overline{\kappa}_2. \tag{55}$$

Part 3. Assume for now that $k \geq \overline{\kappa}_2$. Since $\overline{\kappa}_2 \geq \overline{\kappa}_1$, the invertibility conditions from Lemma 33 are satisfied and, by the proof of Theorem 19, regret is zero for all time k such that:

$$(2 + \mathbb{1}\{h > 1\}C_h)\beta_k \sum_{i=h}^{H} \frac{k + \lambda - g(k) - 8\sqrt{k\log(2dHk/\delta)}}{(k\lambda_i^+ + \lambda - g(k) - 8\sqrt{k\log(2dHk/\delta)})^{3/2}} \leq \Delta_{\min}, \tag{56}$$

for all h. Using the definition of λ^+, $\lambda = 1$ and $C_h = 0$ for LSVI-UCB, a sufficient condition is:

$$2H\beta_k \frac{k + 1 - g(k) - 8\sqrt{k\log(2dHk/\delta)}}{(k\lambda_+ + 1 - g(k) - 8\sqrt{k\log(2dHk/\delta)})^{3/2}} \leq \Delta_{\min}, \tag{57}$$

$$2H\beta_k \frac{2k}{(k\lambda_+ - g(k) - 8\sqrt{k\log(2dHk/\delta)})^{3/2}} \leq \Delta_{\min}. \tag{58}$$

Since $k \geq \overline{\kappa}_2$, by (54), we just need:

$$2H\beta_k \frac{2k}{\left(\frac{1}{2}k\lambda_+\right)^{3/2}} \leq \Delta_{\min}. \tag{59}$$

Rearranging and using the definition of β_k:

$$\sqrt{k} \geq \frac{12c_2H^2d}{\Delta_{\min}\lambda_+^{3/2}}\sqrt{\log(2dHk/\delta)} \tag{60}$$

$$k \geq \frac{12c_2H^2d}{\Delta_{\min}\lambda_+^{3/2}}\sqrt{k\log(2dHk/\delta)}, \tag{61}$$

and again from (49):

$$k \geq \frac{432c_2^2H^4d^2}{\Delta_{\min}^2\lambda_+^3}\log\left(\frac{288c_2^2d^3H^5}{\Delta_{\min}^2\lambda_+^2\delta}\right) \triangleq \overline{\kappa}_3. \tag{62}$$

The proof is concluded by taking $\overline{\kappa} = \max\{\overline{\kappa}_2, \overline{\kappa}_3\}$.

Lemma 21. *The critical time $\overline{\kappa}$ from Theorem 8 for ELEANOR is upper bounded as:*

$$\overline{\kappa} \leq \max\left\{\frac{48c_1^2H^4d^2}{\lambda_+^2}\log\left(\frac{32c_1^2H^5d^3}{\lambda_+^2\delta}\right), \frac{432c_2^2H^4d}{\Delta_{\min}^2\lambda_+^3}\log\left(\frac{288d^2H^5c_2^2}{\Delta_{\min}^2\lambda_+^3\delta}\right)\right\} \tag{63}$$

where $\lambda_+ = \min_{h\in[H]}\{\lambda_h^+\}$ and c_1, c_2 are universal constants.

Proof. The proof is the same as for Lemma 20, except that for ELEANOR we have (see the proof of Theorem 8):

$$g(k) \leq c_1H^2d\sqrt{k\log(2dHk/\delta)} \tag{64}$$

$$\beta_k \leq c_2H\sqrt{d\log(2dHk/\delta)}, \tag{65}$$

where c_1, c_2 are universal constants. The three critical times are then:

$$\overline{\kappa}_1 = \frac{12c_1^2 H^4 d^2}{\lambda_+^2} \log\left(\frac{8c_1^2 H^5 d^3}{\lambda_+^2 \delta}\right) \tag{66}$$

$$\overline{\kappa}_2 = \frac{48c_1^2 H^4 d^2}{\lambda_+^2} \log\left(\frac{32c_1^2 H^5 d^3}{\lambda_+^2 \delta}\right) \geq \overline{\kappa}_1 \tag{67}$$

$$\overline{\kappa}_3 = \frac{432c_2^2 H^4 d}{\Delta_{\min}^2 \lambda_+^3} \log\left(\frac{288c_2^2 d^2 H^5}{\Delta_{\min}^2 \lambda_+^2 \delta}\right), \tag{68}$$

and we can take $\overline{\kappa} = \max\{\overline{\kappa}_2, \overline{\kappa}_3\}$. □

E Representation Selection: Proofs of Section 4

The main ingredient behind the proofs of Theorems 10 and 12 In order to show a regret guarantee for the LSVI-LEADER algorithm, we start by showing a version of Lemma B.4 in [Jin et al., 2020] that takes into account the presence of multiple representations.

First we need the corresponding version of Lemma D.6 in [Jin et al., 2020].

Lemma 22. *Given an MDP M and a set of representations $\{\Phi_j\}_{j\in[N]}$ satisfying the low-rank assumption (Asm. 2). Let $\mathcal{V}$ denote a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with the following parametric form,*

$$V(\cdot) = \min\left(\min_{j\in[N]} \max_a \boldsymbol{w}_j^\top \phi_j(\cdot, a) + \beta\sqrt{\phi_j(\cdot, a)^\top \boldsymbol{\Lambda}_j^{-1} \phi_j(\cdot, a)}, H\right)$$

where the parameters $\{\boldsymbol{w}_j, \boldsymbol{\Lambda}_j\}_{j=1}^N, \beta$ satisfy $\|\boldsymbol{w}\| \leq L$, $\beta \in [0, B]$ and the minimum eigenvalue of $\boldsymbol{\Lambda}_j$ satisfies $\lambda_{\min}(\boldsymbol{\Lambda}_j) \geq \lambda$. Assume $\|\boldsymbol{\phi}(s, a)\| \leq 1$ for all (s, a) pairs and let $\mathcal{N}_\epsilon$ be the $\epsilon-$covering number of $\mathcal{V}$ with respect to the distance $\operatorname{dist}(V, V') = \sup_s |V(s) - V'(s)|$. Then,

$$\log \mathcal{N}_\epsilon \leq N\left(d\log(1 + 4L/\epsilon) + d^2 \log\left(1 + 8d^{1/2}B^2/(\lambda\epsilon^2)\right)\right)$$

Proof. Let's reparametrize the function class $\mathcal{V}$ by $\mathbf{A}_j = \beta^2 \boldsymbol{\Lambda}_j^{-1}$, so we have,

$$V(\cdot) = \min\left(\min_{j\in[N]} \max_a \boldsymbol{w}_j^\top \phi_j(\cdot, a) + \sqrt{\phi_j(\cdot, a)^\top \boldsymbol{A}_j \phi_j(\cdot, a)}, H\right) \tag{69}$$

for $\|\boldsymbol{w}_j\| \leq L$ and $\|\boldsymbol{A}_j\| \leq B^2\lambda^{-1}$. For any two functions $V_1, V_2 \in \mathcal{V}$, let them take the form in Equation 69 with parameters $(\{\boldsymbol{w}_j^{(1)}, \boldsymbol{A}_j^{(1)}\}_{j=1}^N$ and $(\{\boldsymbol{w}_j^{(2)}, \boldsymbol{A}_j^{(2)}\}_{j=1}^N$. Then since $\min_j$, $\min(\cdot, H)$ and $\max_a$ are contraction maps, we have

$$\begin{aligned}
\operatorname{dist}(V_1, V_2) &\leq \sup_{j,s,a} \Bigg| \left[\left(\boldsymbol{w}_j^{(1)}\right)^\top \phi_j(\cdot, a) + \sqrt{\phi_j(\cdot, a)^\top \boldsymbol{A}_j^{(1)} \phi_j(\cdot, a)}\right] - \\
&\quad \left[\left(\boldsymbol{w}_j^{(2)}\right)^\top \phi_j(\cdot, a) + \sqrt{\phi_j(\cdot, a)^\top \boldsymbol{A}_j^{(2)} \phi_j(\cdot, a)}\right] \Bigg| \qquad (70) \\
&\leq \sup_j \left(\sup_{\|\phi_j\|\leq 1} \left|\left[\left(\boldsymbol{w}_j^{(1)}\right)^\top \phi_j + \sqrt{\phi_j^\top \boldsymbol{A}_j^{(1)} \phi_j}\right] - \left[\left(\boldsymbol{w}_j^{(2)}\right)^\top \phi_j + \sqrt{\phi_j^\top \boldsymbol{A}_j^{(2)} \phi_j}\right]\right|\right) \\
&\leq \sup_j \left(\sup_{\|\phi_j\|\leq 1} \left|\left(\boldsymbol{w}_j^{(1)} - \boldsymbol{w}_j^{(2)}\right)^\top \phi_j\right| + \sup_{\|\phi_j\|\leq 1} \sqrt{\left|\phi_j^\top \left(\boldsymbol{A}_j^{(1)} - \boldsymbol{A}_j^{(2)}\right) \phi_j\right|}\right) \\
&= \sup_j \|\boldsymbol{w}_j^{(1)} - \boldsymbol{w}_j^{(2)}\| + \sqrt{\|\boldsymbol{A}_j^{(1)} - \boldsymbol{A}_j^{(2)}\|} \\
&\leq \sup_j \|\boldsymbol{w}_j^{(1)} - \boldsymbol{w}_j^{(2)}\| + \sqrt{\|\boldsymbol{A}_j^{(1)} - \boldsymbol{A}_j^{(2)}\|_F} \qquad (71)
\end{aligned}$$

For matrices $\|\cdot\|$ and $\|\cdot\|_F$ denote the matrix operator norm and the frobenius norm respectively.

Let $\mathcal{C}_j^{\boldsymbol{w}}$ be an $\epsilon/2$ cover of $\{\boldsymbol{w}_j \in \mathbb{R}^d | \|\boldsymbol{w}_j\| \leq L\}$ with respect to the 2-norm and let $\mathcal{C}_j^{\boldsymbol{A}}$ be an $\epsilon^2/4-$cover of $\{\boldsymbol{A} \in \mathbb{R}^{d\times d} | \|\boldsymbol{A}\|_F \leq d^{1/2}B^2\lambda^{-1}\}$ with respect to the Frobenius norm. By Lemma D.5. in [Jin et al., 2020] we know that,

$$|\mathcal{C}_j^{\boldsymbol{w}}| \leq (1+4L/\epsilon)^d, \qquad |\mathcal{C}_j^{\boldsymbol{A}}| \leq \left(1+8d^{1/2}B^2/(\lambda\epsilon^2)\right)^{d^2}$$

By Equation 71, for any $V_1 \in \mathcal{V}$ there exists points $\{\boldsymbol{w}_j^{(2)}\}_{j=1}^N$ and $\{\boldsymbol{A}_j^{(2)}\}_{j=1}^N$ such that V_2 parametrized by $(\{\boldsymbol{w}_j^{(2)}\}_{j=1}^N, \boldsymbol{A}_j^{(2)}\}_{j=1}^N)$ satisfies $\text{dist}(V_1, V_2) \leq \epsilon$. Hence it holds that $\mathcal{N}_\epsilon \leq \left(|\mathcal{C}_j^{\boldsymbol{w}}||\mathcal{C}_j^{\boldsymbol{A}}|\right)^N$, which gives:

$$\log \mathcal{N}_\epsilon \leq N\left(d\log(1+4L/\epsilon) + d^2\log\left(1+8d^{1/2}B^2/(\lambda\epsilon^2)\right)\right).$$

□

Lemma 23 (Multi-representation version of Lemma B.3 in [Jin et al., 2020])**.** *Given an MDP M and a set of representations $\{\Phi_j\}_{j\in[N]}$ satisfying the low-rank assumption (Asm. 2). For all $k \in \mathbb{N}, h \in [H]$, with probability $1-2\delta$:*

$$\left\|\sum_{i=1}^k \phi_h^{(j)}(s_h^i, a_h^i)\left(\overline{V}_{h+1}^k(s_{h+1}^i) - \mathbb{P}_h\overline{V}_{h+1}^k(s_h^i, a_h^i)\right)\right\|_{\Lambda_{h,k}^{-1}(j)} \leq CdH\sqrt{N\log(2N(c_\beta+1)dHk/\delta)}, \tag{72}$$

for all $j \in [N]$ and for some constant C independent of c_β.

Proof. This result follows from a simple use of an anytime version of Lemma D.4 from [Jin et al., 2020] with $\epsilon = dH/k$ and $\delta' = \frac{\delta}{2N}$ and $\lambda = 1$. Let $j \in [N]$ be one of the representations.

$$\begin{aligned}
&\left\|\sum_{i=1}^k \phi_h^{(j)}(s_h^i, a_h^i)\left(\overline{V}_{h+1}^k(s_{h+1}^i) - \mathbb{P}_h\overline{V}_{h+1}^k(s_h^i, a_h^i)\right)\right\|^2_{\Lambda_{h,k}^{-1}(j)} \\
&\leq 4H^2\left[\frac{d}{2}\log\left(\frac{k+\lambda}{\lambda}\right) + 2\log\frac{\pi k}{\sqrt{6}} + \log\frac{2}{\delta} + dN\log\left(1+\frac{8k^{3/2}}{\sqrt{\lambda d}}\right) + \right. \\
&\quad \left. d^2N\log\left(1+\frac{8\sqrt{d}c_\beta^2k^2\log(2dHk/\delta)}{\lambda}\right)\right] + \frac{8d^2H^2}{\lambda} \\
&= \mathcal{O}(d^2NH^2\log(2N(c_\beta+1)dHk/\delta))
\end{aligned}$$

A simple union bound over all representations in $\{\Phi_j\}_{j\in[N]}$ yields the desired result.

□

We have now the necessary ingredients to prove an equivalent version to Lemma B.4 from [Jin et al., 2020] for the case of multiple representations.

Lemma 24 (Equivalent to Lemma B.4 in [Jin et al., 2020])**.** *Given an MDP M and a set of representations $\{\Phi_j\}_{j\in[N]}$ satisfying the low-rank assumption (Asm. 2). With probability at least $1-2\delta$, for any policy π, any episode $k \in \mathbb{N}$, stage $h \in [H]$, state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$,*

$$\left|\langle\phi_h^{(j)}(s,a), \boldsymbol{w}_h^k(j)\rangle - Q_h^\pi(s,a) - \mathbb{P}_h\left(\overline{V}_{h+1}^k - V_{h+1}^\pi\right)(s,a)\right| \leq \beta_k\left\|\phi^{(j)}(s,a)\right\|_{\Lambda_{h,k}(j)^{-1}}$$

where $\beta_k = C'dH\sqrt{N\log(2N(c_\beta+1)dHk/\delta)}$. For some absolute constant C'.

Proof. We know that for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$:

$$Q_h^\pi(s,a) = \langle \phi_h^{(j)}(s,a), \boldsymbol{w}_h^\pi(j) \rangle = \left(r_h + \mathbb{P}_h V_{h+1}^\pi\right)(s,a) \quad \forall j \in [N],$$

This gives

$$\begin{aligned}
\boldsymbol{w}_h^k(j) - \boldsymbol{w}_h^\pi(j) &= \Lambda_{h,k}^{-1}(j) \sum_{i=1}^{k-1} \phi_h^{(j)}(s_h^i, a_h^i) \left(r_h(s_h^i, a_h^i) + \max_{a \in \mathcal{A}} \overline{Q}_{h+1}^{k-1}(s_{h+1}^i, a) \right) - \boldsymbol{w}_h^\pi \\
&= \Lambda_{h,k}(j)^{-1} \left\{ -\lambda \boldsymbol{w}_h^\pi + \sum_{i=1}^{k-1} \phi_h^{(j)}(s_h^i, a_h^i) \left(\overline{V}_{h+1}^k(s_{h+1}^i) - \mathbb{P}_h V_{h+1}^\pi(s_h^i, a_h^i) \right) \right\} \\
&= \underbrace{-\lambda \Lambda_{h,k}^{-1}(j) \boldsymbol{w}_h^\pi(j)}_{\boldsymbol{q}_1} + \underbrace{\Lambda_{h,k}^{-1}(j) \sum_{i=1}^{k-1} \phi_h^{(j)}(s_h^i, a_h^i) \left(\overline{V}_{h+1}^k(s_{h+1}^i) - \mathbb{P}_h \overline{V}_{h+1}^k(s_h^i, a_h^i) \right)}_{\boldsymbol{q}_2} + \\
&\quad \underbrace{\Lambda_{h,k}^{-1}(j) \left(\sum_{i=1}^{k-1} \phi_h^{(j)}(s_h^i, a_h^i) \mathbb{P}_h \left(\overline{V}_{h+1}^k - V_{h+1}^\pi \right)(s_h^i, a_h^i) \right)}_{\boldsymbol{q}_3}
\end{aligned}$$

Now we bound the terms on the right hand side. For the first term,

$$\left| \langle \phi_h^{(j)}(s,a), \boldsymbol{q}_1 \rangle \right| = \left| \lambda \langle \phi_h^{(j)}(s,a), \Lambda_{h,k}^{-1}(j) \boldsymbol{w}_h^\pi \rangle \right| \leq \sqrt{\lambda} \|\boldsymbol{w}_h^\pi\| \left\| \phi_h^{(j)}(s,a) \right\|_{\Lambda_{h,k}^{-1}(j)} \overset{(i)}{\leq} 2H\sqrt{d\lambda} \left\| \phi_h^{(j)}(s,a) \right\|_{\Lambda_{h,k}^{-1}(j)}$$

Inequality (i) above holds because of Lemma B.1 of [Jin et al., 2020]. For the second term $\boldsymbol{q}_2$, given the event defined in Lemma 23 (which holds with probability at least $1 - 2\delta$) we have,

$$\left| \langle \phi_h^{(j)}(s,a), \boldsymbol{q}_2 \rangle \right| \leq CdH \sqrt{N \log(2N(c_\beta + 1)dHk/\delta)} \left\| \phi_h^{(j)}(s,a) \right\|_{\Lambda_{h,k}^{-1}(j)}$$

For the third term,

$$\begin{aligned}
&\langle \phi_h^{(j)}(s,a), \boldsymbol{q}_3 \rangle \\
&= \left\langle \phi_h^{(j)}(s,a), \left(\Lambda_{h,k}^{-1}(j) \right) \sum_{i=1}^{k-1} \phi_h^{(j)}(s_h^i, a_h^i) \mathbb{P}_h \left(\overline{V}_{h+1}^k - V_{h+1}^\pi \right)(s_h^i, a_h^i) \right\rangle \\
&= \left\langle \phi_h^{(j)}(s,a), \left(\Lambda_{h,k}^{-1}(j) \right) \sum_{i=1}^{k-1} \phi_h^{(j)}(s_h^i, a_h^i) \phi_j^\top(s_h^i, a_h^i) \int \left(\overline{V}_{h+1}^k - V_{h+1}^\pi \right)(s'_{h+1}) d\boldsymbol{\mu}_h^j(s'_{h+1} | s_h^i, a_h^i) \right\rangle \\
&= \underbrace{\left\langle \phi_h^{(j)}(s,a), \int \left(\overline{V}_{h+1}^k - V_{h+1}^\pi \right)(s'_{h+1}) d\boldsymbol{\mu}_h^j(s'_{h+1} | s_h^i, a_h^i) \right\rangle}_{p_1} - \\
&\quad \underbrace{\lambda \left\langle \phi_h^{(j)}(s,a), \Lambda_{h,k}^{-1}(j) \int \left(\overline{V}_{h+1}^k - V_{h+1}^\pi \right)(s'_{h+1}) d\boldsymbol{\mu}_h^j(s'_{h+1} | s_h^i, a_h^i) \right\rangle}_{p_2}
\end{aligned}$$

And therefore,

$$p_1 = \mathbb{P}_h \left(\overline{V}_{h+1}^k - V_{h+1}^\pi \right)(s,a), \qquad |p_2| \leq 2H\sqrt{d\lambda} \left\| \phi_h^{(j)}(s,a) \right\|_{\Lambda_{h,k}^{-1}(j)}$$

Finally since $\langle \phi_h^{(j)}(s,a), \boldsymbol{w}_h^k(j)\rangle - Q_h^\pi(s,a) = \langle \phi_h^{(j)}(s,a), \boldsymbol{w}_h^k - \boldsymbol{w}_h^\pi\rangle = \langle \phi_h^{(j)}(s,a), \boldsymbol{q}_1 + \boldsymbol{q}_2 + \boldsymbol{q}_3\rangle$, we have

$$\begin{aligned}
&\left|\langle \phi_h^{(j)}(s,a), \boldsymbol{w}_h^k(j)\rangle - Q_h^\pi(s,a) - \mathbb{P}_h\left(\overline{V}_{h+1}^k - V_{h+1}^\pi\right)(s,a)\right| \\
&\leq \left(CdH\sqrt{N\log(2N(c_\beta+1)dHk/\delta)} + 4H\sqrt{d\lambda}\right)\left\|\phi_h^{(j)}(s,a)\right\|_{\Lambda_{h,k}^{-1}(j)} \\
&\leq C'dH\sqrt{N\log(2N(c_\beta+1)dHk/\delta)}\left\|\phi_h^{(j)}(s,a)\right\|_{\Lambda_{h,k}^{-1}(j)}
\end{aligned}$$

For some constant C'. The result follows.

□

Lemma 25. *Given an MDP M and a set of representations $\{\Phi_j\}_{j\in[N]}$ satisfying the low-rank assumption (Asm. 2). With probability at least $1-2\delta$, for any episode $k \in \mathbb{N}$, stage $h \in [H]$, and state $s \in \mathcal{S}$,*

$$\overline{V}_h^k(s) - V_h^{\pi^k}(s) \leq 2\beta_k \min_{j\in[N]} \|\phi_h^{(j)}(s,\pi_h^k(s))\|_{\Lambda_{h,k}^{-1}(j)} + \mathbb{E}_{s'\sim p_h(s,\pi_h^k(s))}\left[\overline{V}_{h+1}^k(s') - V_{h+1}^{\pi^k}(s')\right].$$

Where $\beta_k = C'dH\sqrt{N\log(2N(c_\beta+1)dHk/\delta)}$.

Proof. Note that $\overline{V}_h^k(s) - V_h^{\pi^k}(s) = \overline{Q}_h^k(s,\pi_h^k(s)) - Q_h^{\pi^k}(s,\pi_h^k(s))$. Using Lemma 24, for any $j \in [N]$

$$\begin{aligned}
Q_h^{\pi^k}(s,\pi_h^k(s)) \geq \langle \phi_h^{(j)}(s,\pi_h^k(s)), \boldsymbol{w}_h^k(j)\rangle - \\
\mathbb{E}_{s'\sim p_h(s,\pi_h^k(s))}[\overline{V}_{h+1}^k(s') - V_{h+1}^{\pi^k}(s')] - \beta_{h,k}\|\phi_h^{(j)}(s,\pi_h^k(s))\|_{\Lambda_{h,k}^{-1}(j)}
\end{aligned}$$

And therefore for all $j \in [N]$,

$$\begin{aligned}
\langle \phi_h^{(j)}(s,\pi_h^k(s)), \boldsymbol{w}_h^k(j)\rangle + \beta_{h,k}\|\phi_h^{(j)}(s,\pi_h^k(s))\|_{\Lambda_{h,k}^{-1}(j)} - V_h^{\pi^k}(s) \leq \\
2\beta_{h,k}\|\phi_h^{(j)}(s,\pi_h^k(s))\|_{\Lambda_{h,k}^{-1}(j)} + \mathbb{E}_{s'\sim p_h(s,\pi_h^k(s))}[\overline{V}_{h+1}^k(s') - V_{h+1}^{\pi^k}(s')]
\end{aligned}$$

Taking the minimum over $j \in [N]$ (and H) on the LHS yields the result,

$$\overline{V}_h^k(s) - V_h^{\pi^k}(s) \leq 2\beta_{h,k} \min_{j\in[N]} \|\phi_h^{(j)}(s,\pi_h^k(s))\|_{\Lambda_{h,k}^{-1}(j)} + \mathbb{E}_{s'\sim p_h(s,\pi_h^k(s))}[\overline{V}_{h+1}^k(s') - V_{h+1}^{\pi^k}(s')].$$

□

Finally we show this implies optimism holds,

Lemma 26. *[Optimism. Equivalent version of Lemma B.5 in [Jin et al., 2020]] With probability $1-\delta$ and for all $s,a \in \mathcal{S}\times\mathcal{A}$, $k \in \mathbb{N}$ and $h \in [H]$, the $\{\overline{Q}_h^k\}_{h\in[H]}$ functions of LSVI-LEADER satisfy,*

$$\overline{Q}_h^k(s,a) \geq Q_h^*(s,a).$$

Proof. The same proof as in Lemma B.5 in [Jin et al., 2020] works just simply modifying it to have a minimum over $j \in [N]$ in the necessary places. We reproduce the argument here for completeness. The proof of the Lemma proceeds by induction.

First, we prove the base case, at the last step H. The statement holds because $\overline{Q}_H^k(s,a) \geq Q_H^*(s,a)$ since the value function at $H+1$ is zero and by Lemma 24 we have that with probability at least $1-2\delta$ for all $k \in \mathbb{N}$, $s \in \mathcal{S}, a \in \mathcal{A}$ and any $j \in [N]$,

$$\left|\langle \phi_h^{(j)}(s,a), \boldsymbol{w}_H^k(j)\rangle - Q_H^{\pi_*}(s,a)\right| \leq C'dH\sqrt{N\log(2N(c_\beta+1)dHk/\delta)}\left\|\phi_h^{(j)}(s,a)\right\|_{\Lambda_{H,k}^{-1}(j)}$$

Therefore for all $j \in [N]$, with probability at least $1 - 2\delta$,

$$\langle \phi_h^{(j)}(s,a), \boldsymbol{w}_H^k(j)\rangle + C' dH \sqrt{N \log(2N(c_\beta+1)dHk/\delta)} \left\|\phi_h^{(j)}(s,a)\right\|_{\Lambda_{H,k}^{-1}(j)} \geq Q_H^{\pi_*}(s,a)$$

Since $H \geq Q_H^{\pi_*}(s,a)$ by definition, we conclude that taking the mimimum over $j \in [N]$ (and H), and using the fact that

$$\overline{Q}_h^k(s,a) = \min\left(\min_{j\in[N]} \langle \phi_h^{(j)}(s,a), \boldsymbol{w}_H^k(j)\rangle + C' dH \sqrt{N \log(2N(c_\beta+1)dHk/\delta)} \left\|\phi_h^{(j)}(s,a)\right\|_{\Lambda_{H,k}^{-1}(j)}, H\right)$$

We conclude that,

$$\overline{Q}_H^k(s,a) \geq Q_H^{\pi_*}(s,a).$$

Now, suppose the statement holds true at step $h+1$ and consider step h. Again by Lemma 24 we have, for all $k \in [K]$ and all $j \in [N]$

$$\begin{aligned}&\left|\langle \phi_h^{(j)}(s,a), \boldsymbol{w}_h^k(j)\rangle - Q_h^{\pi_*}(s,a) - \mathbb{P}_h\left(\overline{V}_{h+1}^k - V_{h+1}^{\pi_*}\right)(s,a)\right| \\ &\leq C' dH \sqrt{N \log(2N(c_\beta+1)dHk/\delta)} \left\|\phi_j(s,a)\right\|_{\Lambda_{h,k}^{-1}(j)}\end{aligned}$$

By the induction assumption that $\mathbb{P}_h\left(\overline{V}_{h+1}^k - V_{h+1}^{\pi_*}\right)(s,a) \geq 0$, we have for all $j \in [N]$:

$$Q_h^{\pi_*}(s,a) \leq \min\left(\langle \phi_h^{(j)}(s,a), \boldsymbol{w}_h^k(j)\rangle + C' dH \sqrt{N \log(2N(c_\beta+1)dHk/\delta)} \left\|\phi_h^{(j)}(s,a)\right\|_{\Lambda_{h,k}^{-1}(j)}, H\right)$$

The result follows by taking a minimum over $j \in [N]$. □

Finishing the proof of Theorem 10. Having proven Lemma 25 and that optimism holds for LSVI-LEADER (Lemma 26), we conclude that an equivalent version of Assumption 16 holds. The same logic of the proofs of Lemmas 17, 18 and Theorem 19 apply in this case. Hence, we conclude that the regret of LSVI-LEADER is upper bounded by the minimum of these regret bounds for all representations $z \in \mathcal{Z}$, thus proving the first result. To obtain the second result, simply notice that, if $z^\star \in \mathcal{Z}$ is UNISOFT, then we can use the refined analysis for LSVI-UCB of Thm. 9 to show that $\widetilde{R}(K, z^\star, \{\beta_k\})$ is upper bounded by a constant independent of K, hence proving constant regret for LSVI-LEADER.

Proof of Theorem 12. The proof follows the template of Thm 9, but as shown in Lemma 25, the confidence sets of LSVI-LEADER scale with the minimum w.r.t. j of the feature norms. In place of Equation 3, and with the aid of Lemma 33 we see that since the collection of feature maps $\{\Phi_j\}_{j\in[M]}$ is UNISOFT-mixing for all reachable s, a:

$$\begin{aligned}\beta_k \min_{j\in[N]} \left\|\phi_h^{(j)}(s,a)\right\|_{\Lambda_{h,k}^{-1}(j)} \leq& \beta_k \frac{k+\lambda-g(k)-8\sqrt{k\log(2NdHk/\delta)}}{(k\lambda^+(h,s,a)+\lambda-g(k)-8\sqrt{k\log(2NdHk/\delta)})^{3/2}} \\ =& \widetilde{O}(k^{-1/2}),\end{aligned} \tag{73}$$

where $g(k) = \widetilde{O}(\sqrt{k})$ is the regret upper bound from Thm. 10,

$$\lambda^+(h,s,a) = \max_{j\in\mathcal{J}(h,s,a)} \lambda_{h,j}^+, \tag{74}$$

and $\mathcal{J}(h,s,a) \subseteq [N]$ is such that $j \in \mathcal{J}(h,s,a)$ if $\phi_h^{(j)}(s,a) \in \operatorname{span}\left\{\phi_h^{(j)}(s,\pi_h^*(s)) | \rho_h^\star(s) > 0\right\}$. To see this, notice that we can instantiate Lemma 33 with any representation $j \in [N]$ such that $\phi_h^{(j)}(s,a)$ belongs to the span of optimal features. So we use the representation with the largest eigenvalue $\lambda_{h,j}^+$. The UNISOFT-mixing property (Def. 11) guarantees $\mathcal{J}(h,s,a)$ is always nonempty.

By (73) and Lemma 18 (where $C_h = 0$ thanks to local optimism), for each $h \in [H]$ there exists an episode κ_h independent of K such that, for all reachable s and $k > \kappa_h$:

$$\begin{aligned}\Delta_h(s,\pi_h^k(s)) \leq& 2\beta_k \mathbb{E}_{\pi^k}\left[\sum_{i=h}^H \frac{k+\lambda-g(k)-8\sqrt{k\log(2NdHk/\delta)}}{(k\lambda^+(i,s_i,a_i)+\lambda-g(k)-8\sqrt{k\log(2NdHk/\delta)})^{3/2}} \middle| s_h = s\right] \\ <& \Delta_{\min}.\end{aligned} \tag{75}$$

So after $\widetilde{\kappa} = \max_h\{\kappa_h\}$ episodes, LSVI-UCB suffers zero regret. Finally, the regret up to $\widetilde{\kappa}$ cannot be worse than that obtained in Thm. 10 without the UNISOFT-mixing property.

F Auxiliary Results

Proposition 27 (Azuma's inequality)**.** *Let $\{(Z_t, \mathcal{F}_t)\}_{t\in\mathbb{N}}$ be a martingale difference sequence such that $|Z_t| \leq a$ almost surely for all $t \in \mathbb{N}$. Then, for all $\delta \in (0,1)$,*

$$\mathbb{P}\left(\forall t \geq 1 : \left|\sum_{k=1}^{t} Z_k\right| \leq a\sqrt{t\log(2t/\delta)}\right) \geq 1-\delta. \tag{76}$$

Proposition 28 (Matrix Azuma, Tropp, 2012)**.** *Let $\{X_k\}_{k=1}^t$ be a finite adapted sequence of symmetric matrices of dimension d, and $\{C_k\}_{k=1}^t$ a sequence of symmetric matrices such that for all k, $\mathbb{E}_k[X_k] = 0$ and $X_k^2 \preceq C_k^2$ almost surely. Then, with probability at least $1-\delta$:*

$$\lambda_{\max}\left(\sum_{k=1}^{t} X_k\right) \leq \sqrt{8\sigma^2\log(d/\delta)}, \tag{77}$$

where $\sigma^2 = \left\|\sum_{k=1}^t C_k^2\right\|$.

Proposition 29 (He et al. [2020])**.** *For any $h \in [H]$, $s \in \mathcal{S}$, and $\pi \in \Pi$:*

$$V_h^\star(s) - V_h^\pi(s) = \mathbb{E}_\pi\left[\sum_{i=h}^{H} \Delta_i(s_i, a_i)\middle| s_h = s\right],$$

Hence the regret after K episodes can be expressed as:

$$R(K) = \sum_{k=1}^{K} V_1^\star(s_1^k) - V_1^{\pi^k}(s_1^k) = \sum_{k=1}^{K} \mathbb{E}_{\pi^k}\left[\sum_{h=1}^{H} \Delta_h(s_h, a_h)\middle| s_1 = s_1^k\right].$$

Proof. By definition of Δ_h:

$$\begin{aligned}
V_h^\star(s) - V_h^\pi(s) &= Q_h^\star(s, \pi_h(s)) + \Delta_h(s, \pi_h(s)) - V_h^\pi(s) && (78)\\
&= r_h(s, \pi_h(s)) + \mathbb{E}_{s'\sim p_h(s,\pi_h(s))}[V_{h+1}^\star(s')] + \Delta_h(s, \pi_h(s)) - r_h(s, \pi_h(s))\\
&\quad - \mathbb{E}_{s'\sim \mathbb{P}_h(s_h,\pi_h(s_h))}[V_{h+1}^\pi(s')] && (79)\\
&= \Delta_h(s_h, \pi_h(s_h)) + \mathbb{E}_{s'\sim \mathbb{P}_h(s_h,\pi_h(s_h))}[V_{h+1}^\star(s') - V_{h+1}^\pi(s')]. && (80)
\end{aligned}$$

Unrolling the recursion up to H concludes the proof. □

Lemma 30. *Assume $R(k) \leq g(k)$ for all $k \geq 1$ and Asm. 3 holds. Then, probability $1-\delta$, for all h, k:*

$$\Lambda_h^{k+1} \succeq k\Lambda_h^\star + \lambda I - \Delta_{\min}^{-1} g(k) I - 8L^2 I\sqrt{k\log(2dkH/\delta)}. \tag{81}$$

Proof. Define a trajectory as a sequence of states and actions $\tau_h = (s_1, a_1, \ldots, s_h, a_h)$. Let Γ_h denote the set of all trajectories of length h. The distribution over trajectories induced by a (deterministic) policy π is $p_h^\pi(\tau_h) = \mu(s_1)\mathbb{1}\{a_1 = \pi_1(s_1)\}\, p_1(s_2|s_1, a_1)\ldots p_{h-1}(s_h|s_{h-1}, a_{h-1})\mathbb{1}\{a_h = \pi_h(s_h)\}$. We abbreviate as $p_h^\star$ the distribution induced by the optimal policy $\pi^\star$ and as p_h^k the one induced by π^k, the algorithm's policy at episode k. Let us define the following event:

$$E_h^k = \{\tau \in \Gamma_h \text{ s.t. } a_i = \pi_h^k(s_i) = \pi_h^\star(s_i) \text{ for } i = 1, \ldots, h\}. \tag{82}$$

Then:

$$
\begin{aligned}
\Lambda_h^{k+1} - \lambda I &= \sum_{i=1}^{k} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\mathsf{T} \\
&\succeq \sum_{i=1}^{k} \mathbb{1}\left\{\tau_h^i \in E_h^i\right\} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\mathsf{T} \\
&= \sum_{i=1}^{k} \mathbb{1}\left\{\tau_h^i \in E_h^i\right\} \phi_h^\star(s_h^i)\phi_h^\star(s_h^i)^\mathsf{T} \qquad (83) \\
&= \underbrace{\sum_{i=1}^{k} \mathbb{E}_{\tau_h \sim p_h^i}\left[\mathbb{1}\left\{\tau_h \in E_h^i\right\} \phi_h^\star(s_h)\phi_h^\star(s_h)^\mathsf{T}\right]}_{(A)} \\
&\quad + \underbrace{\sum_{i=1}^{k} \left(\mathbb{1}\left\{\tau_h^i \in E_h^i\right\} \phi_h^\star(s_h^i)\phi_h^\star(s_h^i)^\mathsf{T} - \mathbb{E}_{\tau_h \sim p_h^i}\left[\mathbb{1}\left\{\tau_h \in E_h^i\right\} \phi_h^\star(s_h)\phi_h^\star(s_h)^\mathsf{T}\right]\right)}_{\text{(B)}},
\end{aligned}
$$

where (83) is by definition of E_h^i and expectations are conditioned on history up to the beginning of the i-th episode. We first bound (B) with a matrix version of Azuma's inequality. Let:

$$
X_h^i = \mathbb{1}\left\{\tau_h^i \in E_h^i\right\} \phi_h^\star(s_h^i)\phi_h^\star(s_h^i)^\mathsf{T} - \mathbb{E}_{\tau_h \sim p_h^i}\left[\mathbb{1}\left\{\tau_h \in E_h^i\right\} \phi_h^\star(s_h)\phi_h^\star(s_h)^\mathsf{T}\right].
$$

Clearly $\mathbb{E}[X_h^i] = 0$. Moreover, since X_h^i is symmetric:

$$
(X_h^i)^2 \preceq \lambda_{\max}((X_h^i)^2) I \preceq \left\|X_h^i\right\|^2 I \preceq 4I. \qquad (84)
$$

Then by Proposition 28, with probability $1 - \delta_h^k$:

$$
\lambda_{\max}\left(\sum_{i=1}^{k} X_h^i\right) \leq 4\sqrt{2k\log(d/\delta_h^k)}. \qquad (85)
$$

Setting $\delta_h^k = \delta/(2Hk^2)$ we can perform a union bound over episodes and stages to obtain, with probability $1-\delta$, for all h, k:

$$
(B) = \sum_{i=1}^{k} X_h^i \preceq \lambda_{\max}\left(\sum_{i=1}^{k} X_h^i\right) I \preceq 8I\sqrt{k\log(2dHk/\delta)}. \qquad (86)
$$

Now we focus on the (A) term. First, observe that the probability measures p_h^k and $p_h^\star$ agree on E_h^k. Indeed, if $\tau_h \in E_h^k$:

$$
\begin{aligned}
p_h^k(\tau_h) &= \mu(s_1)\mathbb{1}\left\{a_1 = \pi_1^k(s_1)\right\} p_1(s_2|s_1,a_1)\dots p_{h-1}(s_h|s_{h-1},a_{h-1})\mathbb{1}\left\{a_h = \pi_h^k(s_h)\right\} \\
&= \mu(s_1)\mathbb{1}\left\{a_1 = \pi_1^\star(s_1)\right\} p_1(s_2|s_1,a_1)\dots p_{h-1}(s_h|s_{h-1},a_{h-1})\mathbb{1}\left\{a_h = \pi_h^\star(s_h)\right\} \qquad (87) \\
&= \mu(s_1)p_1(s_2|s_1,a_1)\dots p_{h-1}(s_h|s_{h-1},a_{h-1}). \qquad (88)
\end{aligned}
$$

So:

$$
\begin{aligned}
(A) &= \sum_{i=1}^{k} \mathbb{E}_{\tau_h \sim p_h^i}[\mathbb{1}\left\{\tau_h \in E_h^i\right\} \phi_h^\star(s_h)\phi_h^\star(s_h)^\mathsf{T}] \\
&= \sum_{i=1}^{k} \mathbb{E}_{\tau_h \sim p_h^\star}[\mathbb{1}\left\{\tau_h \in E_h^i\right\} \phi_h^\star(s_h)\phi_h^\star(s_h)^\mathsf{T}] && (89)\\
&= k\,\mathbb{E}_{\tau_h \sim p_h^\star}[\phi_h^\star(s_h)\phi_h^\star(s_h)^\mathsf{T}] - \sum_{i=1}^{k} \int_{\Gamma_h \setminus E_h^i} \phi_h^\star(s_h)\phi_h^\star(s_h)^\mathsf{T} p_h^\star(\mathrm{d}\tau_h) && (90)\\
&= k\,\mathbb{E}_{s \sim \rho_h^\star}[\phi_h^\star(s_h)\phi_h^\star(s_h)^\mathsf{T}] - \sum_{i=1}^{k} \int_{\Gamma_h \setminus E_h^i} \phi_h^\star(s_h)\phi_h^\star(s_h)^\mathsf{T} p_h^\star(\mathrm{d}\tau_h) && (91)\\
&\succeq k\,\mathbb{E}_{s \sim \rho_h^\star}[\phi_h^\star(s_h)\phi_h^\star(s_h)^\mathsf{T}] - I \sum_{i=1}^{k} \left(1 - \int_{E_h^i} p_h^\star(\mathrm{d}\tau_h)\right) && (92)\\
&= k\,\mathbb{E}_{s \sim \rho_h^\star}[\phi_h^\star(s_h)\phi_h^\star(s_h)^\mathsf{T}] - I \sum_{i=1}^{k} \left(1 - \int_{E_h^i} p_h^\star(\mathrm{d}\tau_h)\right) && (93)\\
&= k\,\mathbb{E}_{s \sim \rho_h^\star}[\phi_h^\star(s_h)\phi_h^\star(s_h)^\mathsf{T}] - I \underbrace{\sum_{i=1}^{k} \mathbb{E}_{\tau_h \sim p_h^i(\tau_h)}[\mathbb{1}\left\{\tau_h \notin E_h^i\right\}]}_{(C)}. && (94)
\end{aligned}
$$

Finally, under Asm. 3 and the regret upper bound:

$$
\begin{aligned}
(C) &= \sum_{i=1}^{k} \mathbb{E}_{\tau_h \sim p_h^i(\tau_h)}[\mathbb{1}\left\{\tau_h \notin E_h^i\right\}] \\
&\leq \sum_{i=1}^{k} \sum_{j=1}^{h} \mathbb{E}_{\pi^i}[\mathbb{1}\left\{a_j \neq \pi_j^\star(s_j)\right\}] && (95)\\
&\leq \sum_{i=1}^{k} \sum_{j=1}^{h} \mathbb{E}_{\pi^i}[\mathbb{1}\left\{\Delta_j(s_j, a_j) \geq \Delta\right\}] && (96)\\
&\leq \sum_{i=1}^{k} \sum_{j=1}^{h} \mathbb{E}_{\pi^i}\left[\frac{\Delta_j(s_j, a_j)}{\Delta_{\min}}\right] && (97)\\
&= \frac{1}{\Delta_{\min}} \sum_{i=1}^{k} \mathbb{E}_{\pi^i}\left[\sum_{j=1}^{h} \Delta_j(s_j, a_j)\right] && (98)\\
&\leq \frac{1}{\Delta_{\min}} \sum_{i=1}^{k} \mathbb{E}_{\pi^i}\left[\sum_{h=1}^{H} \Delta_h(s_h, a_h)\right] && (99)\\
&= \frac{R(k)}{\Delta_{\min}} \leq \frac{g(k)}{\Delta_{\min}}, && (100)
\end{aligned}
$$

where (95) is by definition of E_h^i, (96) is from the uniqueness of the optimal policy and Asm. 3, and (100) is from Proposition 29. □

Proposition 31 (Lemma 29 from [Papini et al., 2021]). *Let $\boldsymbol{v} \in \mathbb{R}^d$ with $\|\boldsymbol{v}\| = 1$ and $A \in \mathbb{R}^{d \times d}$ symmetric invertible with non-zero eigenvalues $\lambda_1 \leq \cdots \leq \lambda_d$ and corresponding orthonormal eigenvectors $u_1, \ldots, u_d$. Let $\mathcal{I} \subseteq [d]$ be any index set. If $\boldsymbol{v} \in \operatorname{span}\{u_i\}_{i \in \mathcal{I}}$ and $\lambda_i > 0$ for all $i \in \mathcal{I}$:*

$$
\boldsymbol{v}^\mathsf{T} A^{-1} \boldsymbol{v} \leq \frac{(\max_{i \in \mathcal{I}} \lambda_i + \min_{i \in \mathcal{I}} \lambda_i)^2}{4 \max_{i \in \mathcal{I}} \lambda_i \min_{i \in \mathcal{I}} \lambda_i} \frac{1}{\boldsymbol{v}^\mathsf{T} A \boldsymbol{v}}.
$$

Proposition 32 (e.g., Lemma 30 from [Papini et al., 2021]). *The smallest nonzero eigenvalue of a symmetric p.s.d. matrix $A \in \mathbb{R}^{d\times d}$ is:*

$$\lambda^+_{\min}(A) = \min_{\substack{\boldsymbol{v}\in \mathrm{Im}(A)\\ \|\boldsymbol{v}\|=1}} \boldsymbol{v}^\mathsf{T} A \boldsymbol{v},$$

where $\mathrm{Im}(A)$ denotes the column space of A.

Lemma 33. *Consider a d-dimensional representation $(\phi_h)_{h\in[H]}$. Assume there exists an increasing $\widetilde{O}(\sqrt{k})$ function g such that $R(k) \le g(k)$ for all $k \ge 1$, Asm. 3 holds, and $\beta_k = \widetilde{O}(1)$. Then with probability $1-\delta$, for all h, there exists a constant $\widetilde{\kappa}_h$ such that, for every $k \ge \widetilde{\kappa}_h$ and all s, a having $\phi_h(s,a) \in \mathrm{span}\{\phi^\star_h(s) | \rho^\star_h(s) > 0\}$,*

$$\beta_k \|\phi_h(s,a)\|_{(\Lambda^k_h)^{-1}} \le \beta_k \frac{k+\lambda-g(k)-8\sqrt{k\log(2dHk/\delta)}}{(k\lambda^+_h+\lambda-g(k)-8\sqrt{k\log(2dHk/\delta)})^{3/2}} = \widetilde{O}(k^{-1/2}),$$

where λ^+_h is the minimum nonzero eigenvalue of $\Lambda^\star_h$.

Proof. We follow the proof scheme of Lemma 19 from [Papini et al., 2021]. Let $f(k) = g(k) + 8\sqrt{k\log(2dHk/\delta)} = \widetilde{O}(\sqrt{k})$. Notice that $f(k)$ is positive.

Fix h and let $B^k_h = k\Lambda^\star_h + \lambda I - f(k)I$. First, notice that B^k_h is an affine transformation of $\Lambda^\star_h$. As such, B^k_h has the same orthonormal eigenvectors as $\Lambda^\star_h$, and we can define a mapping between the eigenvalues of the two matrices. Next, notice that B^k_h is always invertible for sufficiently large k. Indeed, zero eigenvalues of $\Lambda^\star_h$ are mapped to negative eigenvalues of B^k_h for sufficiently large k — and since $f(k)$ is increasing and sublinear, positive eigenvalues of $\Lambda^\star_h$ are mapped to positive eigenvalues of B^k_h for sufficiently large k. We call $\widetilde{\kappa}_h$ the smallest k such as both conditions hold. For the rest of the proof assume $k \ge \kappa_h$. We have shown that B^k_h is invertible and all and only the nonzero eigenvalues of $\Lambda^\star_h$ are mapped into positive eigenvalues of B^k_h, with the same orthonormal eigenvectors.

Now fix (s,a) such that $\phi_h(s,a) \in \mathrm{span}\{\phi^\star_h(s)|\rho^\star_h(s) > 0\}$ and let $x = \phi_h(s,a)/\|\phi_h(s,a)\|$. From Lemma 30, with probability $1-\delta$, $\Lambda^k_h \succeq B^k_h$. So:

$$x^\mathsf{T}(\Lambda^k_h)^{-1}x \le x^\mathsf{T}(B^k_h)^{-1}x. \tag{101}$$

By hypothesis x belongs to the column space $\mathrm{Im}(\Lambda^\star_h)$, so it belongs to the span of $\widetilde{d} \le d$ orthonormal eigenvectors of $\Lambda^\star_h$. From the properties of B^k_h stated above, x belongs to the span of $\widetilde{d}$ orthonormal eigenvectors of B^k_h corresponding to positive eigenvalues. The smallest such eigenvalue is:

$$k\lambda^+_h + \lambda - f(k), \tag{102}$$

where λ^+_h is the smallest nonzero eigenvalue of $M^\star_h$. Moreover, all the eigenvalues are upper bounded by:

$$k + \lambda - f(k). \tag{103}$$

From Proposition 31:

$$\|\phi_h(s,a)\|_{(\Lambda^k_h)^{-1}} \le \sqrt{x^\mathsf{T}(\Lambda^k_h)^{-1}x} \tag{104}$$

$$\le \sqrt{x^\mathsf{T}(B^k_h)^{-1}x} \tag{105}$$

$$\le \frac{k+\lambda-f(k)}{k\lambda^+_h+\lambda-f(k)} \frac{1}{\sqrt{x^\mathsf{T}B^k_h x}}. \tag{106}$$

Again from the properties of B^k_h, x is orthogonal to all the orthonormal eigenvector of B^k_h that correspond to zero eigenvalues of $\Lambda^\star_h$. Hence by Proposition 32:

$$x^\mathsf{T}B^k_h x = kx^\mathsf{T}\Lambda^\star_h x + \lambda - f(k) \tag{107}$$

$$\ge k \min_{y\in\mathrm{Im}(\Lambda^\star_h),\|y\|=1} y^\mathsf{T}\Lambda^\star_h y + \lambda - f(k) \tag{108}$$

$$= k\lambda^+_h + \lambda - f(k). \tag{109}$$

Since $\beta_k = \widetilde{O}(1)$ and $f(k) = \widetilde{O}(\sqrt{k})$, from (106) and (109):

$$\beta_k \left\|\phi_h(s,a)\right\|_{(\Lambda_h^k)^{-1}} \leq \beta_k \frac{k+\lambda-f(k)}{(k\lambda_h^+ + \lambda - f(k))^{3/2}} = \widetilde{O}(k^{-1/2}). \tag{110}$$

□

Lemma 34. *Let $\{\phi_j\}_{j\in[n]}$ be a set of n vectors in $\mathbb{R}^d$ and $v \in \mathbb{R}^d$ be such that $v \notin \operatorname{span}\{\phi_j : j \in [n]\}$. Then, there exists a scalar $\epsilon > 0$ such that, for any $t \geq 0, \eta > 0$,*

$$\|v\|_{(t\sum_{j\in[n]}\phi_j\phi_j^T+\eta I)^{-1}} \geq \frac{\epsilon}{\sqrt{\eta}}.$$

Proof. Let $\{\lambda_i, u_i\}_{i\in[d]}$ denote the eigenvalues/eigenvectors of the matrix $\sum_{j\in[n]}\phi_j\phi_j^T$. Note that $\operatorname{span}\{u_i : i \in [d]\} = \operatorname{span}\{\phi_j : j \in [n]\} \subset \mathbb{R}^d$. Then, Lemma 28 of Papini et al. [2021] ensures that there exists a scalar $\epsilon > 0$ such that $|v^T u_i| \geq \epsilon$ for at least one eigenvector u_i associated with a zero eigenvalue. Noting that the eigenvectors of $(t\sum_{j\in[n]}\phi_j\phi_j^T + \eta I)^{-1}$ are the same as the those of $\sum_{j\in[n]}\phi_j\phi_j^T$, we have that

$$\|v\|^2_{(t\sum_{j\in[n]}\phi_j\phi_j^T+\eta I)^{-1}} = \sum_{j\in[d]} \frac{(v^T u_j)^2}{\eta+\lambda_j} \geq \frac{(v^T u_i)^2}{\eta} \geq \frac{\epsilon^2}{\eta},$$

which concludes the proof. □

G Examples and Numerical Validations

Consider the following two-stage MDP ($H = 2$) with states $\mathcal{S} = \{s_1, s_2\}$ and actions $\mathcal{A} = \{a_1, a_2\}$:

$$r_1(s,a) = 1 \qquad \text{for all } s \in \mathcal{S} \text{ and } a \in \mathcal{A}, \tag{111}$$

$$p_1(s_1|s_1,a_1) = 1, \quad p_1(s_1|s_1,a_2) = \frac{1}{2}, \quad p_1(s_1|s_2,a_1) = \frac{1}{2}, \quad p_1(s_1|s_2,a_2) = \frac{3}{4}, \tag{112}$$

$$r_2(s_1,a_1) = 1, \quad r_2(s_1,a_2) = \frac{7}{8}, \quad r_2(s_2,a_1) = \frac{1}{2}, \quad r_2(s_2,a_2) = \frac{5}{8}, \tag{113}$$

$\mu(s_1) = \mu(s_2) = 1/2$, and of course $p(s_2|s,a) = 1 - p(s_1|s,a)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Backward induction shows that the (unique) optimal policy is:

$$\pi_1^\star(s_1) = a_1, \quad \pi_1^\star(s_2) = a_2, \quad \pi_2^\star(s_1) = a_1, \quad \pi_2^\star(s_2) = a_2, \tag{114}$$

with the following values:

$$V_1^\star(s_1) = 2, \quad V_1^\star(s_2) = \frac{61}{32}, \quad V_2^\star(s_1) = 1, \quad V_2^\star(s_2) = \frac{5}{8}. \tag{115}$$

Notice also that all states and actions are reachable, i.e. $\rho_h(s,a) > 0$ for all $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $h \in [H]$.

UNISOFT representation. Consider the following 2-dimensional representation $\Phi^{(1)}$:

$$\underline{\phi_1^{(1)}(s_1,a_1)} = \begin{bmatrix}1\\0\end{bmatrix} \quad \phi_1^{(1)}(s_1,a_2) = \begin{bmatrix}1/2\\1/2\end{bmatrix} \quad \phi_1^{(1)}(s_2,a_1) = \begin{bmatrix}1/2\\1/2\end{bmatrix} \quad \underline{\phi_1^{(1)}(s_2,a_2)} = \begin{bmatrix}3/4\\1/4\end{bmatrix} \tag{116}$$

$$\underline{\phi_2^{(1)}(s_1,a_1)} = \begin{bmatrix}0\\1\end{bmatrix} \quad \phi_2^{(1)}(s_1,a_2) = \begin{bmatrix}1/4\\3/4\end{bmatrix} \quad \phi_2^{(1)}(s_2,a_1) = \begin{bmatrix}1\\0\end{bmatrix} \quad \underline{\phi_2^{(1)}(s_2,a_2)} = \begin{bmatrix}3/4\\1/4\end{bmatrix}. \tag{117}$$

It is easy to check that the MDP is low-rank (Asm 2) and $\Phi^{(1)}$ is a realizable representation with $\theta_1 = [1,1]^\mathsf{T}$, $\boldsymbol{\mu}_1(s_1) = [1,0]^\mathsf{T}$, $\boldsymbol{\mu}_1(s_2) = [0,1]^\mathsf{T}$, and $\theta_2 = [1/2,1]^\mathsf{T}$. This is an example of low-rank MDP with *simplex feature space* (see Example 2.2 in [Jin et al., 2020]). We have underlined optimal

features. It is easy to see that optimal features span $\mathbb{R}^2$ at both stages[9], so $\Phi^{(1)}$ is UNISOFT. The optimal covariance matrices are:

$$\Lambda^{(1)}_{1,\star} = \frac{1}{32}\begin{bmatrix} 25 & 3 \\ 3 & 1 \end{bmatrix}, \qquad \Lambda^{(1)}_{2,\star} = \frac{1}{128}\begin{bmatrix} 9 & 3 \\ 3 & 113 \end{bmatrix}. \tag{118}$$

Both are full rank, and their minimum eigenvalues are:

$$\lambda^{(1)}_{1,+} = \frac{13 - 3\sqrt{17}}{32} \simeq 0.02, \qquad \lambda^{(1)}_{2,+} = \frac{61 - \sqrt{2713}}{128} \simeq 0.07. \tag{119}$$

As shown in Theorems 8 and 9, both LSVI-UCB and ELEANOR will only suffer constant regret on this problem.

Non-UNISOFT representation. We apply the procedure described in the proof of Lemma 7 from [Papini et al., 2021] to the second stage of $\Phi^{(1)}$ to obtain an equivalent representation $\Phi^{(2)}$:

$$\underline{\phi^{(2)}_2(s_1, a_1)} = \begin{bmatrix} 30/89 \\ 74/89 \end{bmatrix} \quad \phi^{(2)}_2(s_1, a_2) = \begin{bmatrix} 1/4 \\ 3/4 \end{bmatrix} \quad \phi^{(2)}_2(s_2, a_1) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \underline{\phi^{(2)}_2(s_2, a_2)} = \begin{bmatrix} 75/356 \\ 185/356 \end{bmatrix}, \tag{120}$$

while the feature map for $h = 1$ is the same. It is easy to check that this is still a realizable representation for our MDP with the same parameters.[10] Although the UNISOFT property holds for $h = 1$, it no longer does for $h = 2$. Indeed, we have the following linear dependence between optimal features:

$$\phi^{(2)}_{2,\star}(s_2) = \frac{5}{8}\phi^{(2)}_{2,\star}(s_1), \tag{121}$$

so optimal features only span $\mathbb{R}^1$. However, suboptimal features still span $\mathbb{R}^2$, e.g., by taking action a_2 in s_1 and a_1 in s_2 (recall that all state-action pairs are reachable). Due to Theorem 5, neither LSVI-UCB nor ELEANOR will achieve constant regret on this problem.

Alternative Non-UNISOFT representation. It is also easy to build a representation that is non-UNISOFT by changing the representation at the first stage. For example, let h be any stage (e.g., $h = 1$ in our example) for which we want to transform a UNISOFT representation (in our case $\phi^{(1)}$) into a non-UNISOFT one. We can define a new representation $\phi^{(3)}$ as follows

$$s \in \mathcal{S}, \;\; \phi^{(3)}_h(s, a^\star) = \begin{bmatrix} 0_d \\ \phi^{(1)}_h(s, a^\star_s) \end{bmatrix} \qquad \forall a \neq a^\star_s, \;\; \phi^{(3)}_h(s, a) = \begin{bmatrix} \phi^{(1)}_h(s, a) \\ 0_d \end{bmatrix} \tag{122}$$

$$s' \in \mathcal{S}, \;\; \mu^{(3)}_h(s') = \begin{bmatrix} \mu^{(1)}_h(s') \\ \mu^{(1)}_h(s') \end{bmatrix} \qquad \theta^{(3)}_h = \begin{bmatrix} \theta^{(1)}_h \\ \theta^{(1)}_h \end{bmatrix} \tag{123}$$

Since all states are reachable, it is easy to verify that $\lambda_{\min}\Big(\mathbb{E}_{s\sim\rho^\star_h}\Big[\phi^{(3)}_h(s, a^\star_s)^\intercal \phi^{(3)}_h(s, a^\star_s)\Big]\Big) = 0$ and that

$$\operatorname{span}\Big\{\phi_h(s, a) \mid \forall (s, a), \; \exists \pi \in \Pi : \rho^\pi_h(s, a) > 0\Big\} \neq \operatorname{span}\Big\{\phi^\star_h(s) \mid \forall s, \; \rho^\star_h(s) > 0\Big\}.$$

Then, the representation is not UNISOFT at stage h.

G.1 Numerical Validations

We provide a numerical validation of the behavior of the algorithms with and without a UNISOFT representation. We consider the following representations: $\phi^{(1)}$, $\phi^{(2)}$, $\phi^{(3)}$ which is obtained by applying the transformation in Eq. 122-123 to $\phi^{(2)}$ at stage $h = 1$, and $\phi^{(4)}$ which is obtained by

[9] It may appear counterintuitive that simplex features, which live on a one-dimensional manifold, can span $\mathbb{R}^2$. However, notice that the simplex is not a *linear* subspace of the Euclidean space (it does not include the origin). Indeed, we could describe the example MDP with less parameters, but we would loose the linear structure.

[10] However, notice that some of the new features do not belong to the simplex.

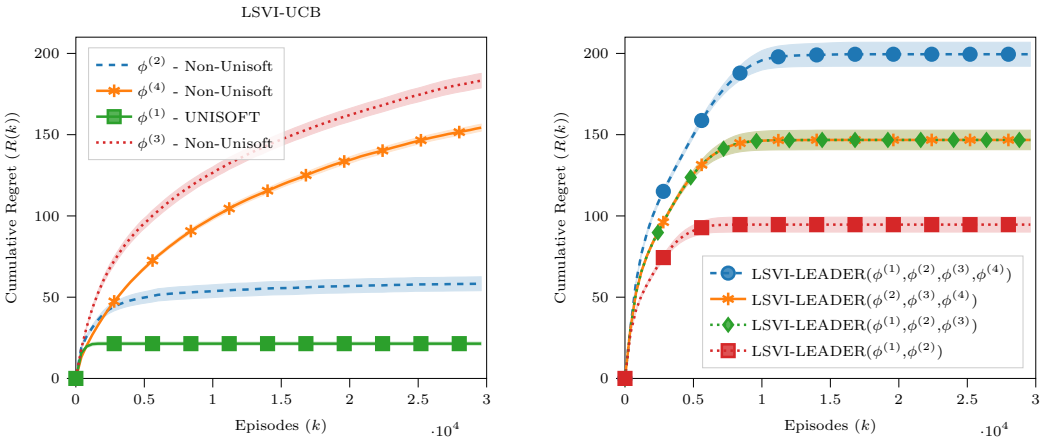

Figure 1: Cumulative regret of LSVI-UCB and LSVI-LEADER with different representations. The performance of LSVI-LEADER with $\{\phi^{(1)}, \phi^{(2)}, \phi^{(4)}\}$ is the same of the one with $\{\phi^{(1)}, \phi^{(2)}, \phi^{(3)}\}$ and $\{\phi^{(2)}, \phi^{(3)}, \phi^{(4)}\}$.

applying the transformation in Eq. 122-123 to $\phi^{(1)}$ at stage $h = 1$. Note that we have $d_1 = 4$ and $d_2 = 2$ for $\phi^{(3)}$ and $\phi^{(4)}$. Furthermore, $\lambda^{(3)}_{1,1} = \lambda^{(4)}_{1,1} = 0$, while $\lambda^{(3)}_{2,1} = 0$ and $\lambda^{(4)}_{2,1} > 0$, which means that $\phi^{(4)}$ is "locally" UNISOFT at stage $h = 2$. The reward is stochastic and drawn from a Bernoulli distribution: $r_{h,t} \sim \text{Ber}(r_h(s_t, a_t))$. We tested both LSVI-UCB on each individual representation and LSVI-LEADER with different combinations of the representations. We consider $\beta_{h,k} = c_\beta d_h H \sqrt{\log(d_h K)}$ and $\beta_{h,k} = c_\beta d_h H \sqrt{N \log(N d_h K)}$ for LSVI-UCB and LSVI-LEADER, respectively. We set $c_\beta = 0.2$ and $K = 30000$. The regret is shown in Fig. 1, averaged over the same 100 seeds.

As expected from the theoretical analysis, LSVI-UCB with UNISOFT representation suffers constant regret since, after the initial exploration phase, it only selects optimal actions. On the other hand, when the representation is Non-UNISOFT, LSVI-UCB suffers a non-constant regret that grows over episodes. LSVI-LEADER is able to exploit the structure of the UNISOFT representation and it achieves constant regret as well in all the configurations containing a UNISOFT representation. The higher regret is due to a longer exploration phase that is a consequence of the enlarged confidence intervals; this is also in line with the theoretical analysis. It is interesting to notice that LSVI-LEADER performs equally good with all the combinations of representations of dimension three (i.e., $\{\phi^{(1)}, \phi^{(2)}, \phi^{(4)}\}$, $\{\phi^{(1)}, \phi^{(2)}, \phi^{(3)}\}$ and $\{\phi^{(2)}, \phi^{(3)}, \phi^{(4)}\}$). LSVI-LEADER is indeed able to mix representations and achieve constant regret even when none of the individual representation would. In the case of $\{\phi^{(2)}, \phi^{(3)}, \phi^{(4)}\}$, LSVI-LEADER is able to mix $\phi^{(2)}$ and $\phi^{(4)}$, that are UNISOFT in stage $h = 1$ and $h = 2$, respectively.

UNISOFT in DeepRL. We wanted also to verify the existence of UNISOFT representations in DeepRL. We trained A2C [Mnih et al., 2016] on different domains and evaluated whether the recovered representation (i.e., last layer of the neural network used to approximate $V^\star$) satisfies the UNISOFT assumptions. Standard benchmark problems are not finite-horizon, we thus considered the following "strong" UNISOFT condition $\lambda_{\min}\left(\mathbb{E}_{s\sim\rho^\star}[\phi^\star(s)\phi^\star(s)^\mathsf{T}]\right) > 0$, which was evaluated by simulating multiple trajectories:

$$\Lambda^\pi_m = \frac{1}{m}\sum_{i=1}^{m}\sum_{t=1}^{T_i}\phi(s_t, a_t)\phi(s_t, a_t)^\mathsf{T} \tag{124}$$

where $a_t = \pi(s_t)$. We use a deterministic version of the policy recovered by A2C for evaluation. We trained A2C using the implementation provided by stable-baselines3 [Raffin et al., 2019]. We use the default parameters (provided by rl-baselines3-zoo [Raffin, 2020]) and tested different network architectures. Since we did not optimize the parameter, we reported only the domains where we obtained good results with at least one network architecture (highlighted in the table). Since A2C

Domain	mean reward	std reward	eval timesteps	eval episodes (m)	$\mathrm{rank}(\Lambda_m^\pi)$	$\lambda_{\min}(\Lambda_m^\pi)$	UNISOFT
Acrobot-v1	-84.5	20.7	149923	1753	16	0.02	✔
AntBulletEnv-v0	2303.9	68.3	150000	150	15	0	
BipedalWalker-v3	2.2	1.6	148800	93	10	0	
CartPole-v1	500.0	0.0	150000	300	1	0	
HopperBulletEnv-v0	836.3	536.2	149982	372	16	0	
MountainCar-v0	-124.9	31.4	149979	1201	16	0.01	✔
MountainCarContinuous-v0	91.6	0.2	149966	1736	5	0	
Pendulum-v0	-173.5	107.0	150000	750	16	0	

Table 3: Results for A2C policy network of dimension $[64, 16]$ and value network of dimension $[64, 16]$. These dimensions represent the size of the hidden layers (with tanh activation function). We highlighted the environments where A2C achieved good performance.

Domain	mean reward	std reward	eval timesteps	eval episodes (m)	$\mathrm{rank}(\Lambda_m^\pi)$	$\lambda_{\min}(\Lambda_m^\pi)$	UNISOFT
Acrobot-v1	-84.9	29.4	149987	1747	32	0.0018	✔
AntBulletEnv-v0	2109.9	46.1	150000	150	32	0.0010	✔
BipedalWalker-v3	267.3	53.3	149278	201	24	0	
CartPole-v1	500.0	0.0	150000	300	1	0	
HopperBulletEnv-v0	1461.6	707.1	149123	205	32	0.0001	✔
MountainCar-v0	-116.5	28.0	149999	1288	32	0.0001	✔
MountainCarContinuous-v0	91.5	0.2	149975	1742	10	0	
Pendulum-v0	-236.5	187.7	150000	750	26	0	

Table 4: Results for A2C policy network of dimension $[64, 32]$ and value network of dimension $[64, 32]$. These dimensions represent the size of the hidden layers (with tanh activation function). We highlighted the environments where A2C achieved good performance.

estimates directly $V^\star$, we used the features of the last layer as features of the optimal policy (i.e., $\phi^\star(s)$) to test for the "strong" UNISOFT condition.

Tables 3–5 show that in several domains the learnt representation is UNISOFT, although the minimum eigenvalue is small. As expected, the number of "strong" UNISOFT representations decreases as the size of the last layer increases. This initial experiment shows that UNISOFT representations are not uncommon in practice but also leave open the possibility of designing algorithms that explicitly try to force the UNISOFT while learning. We believe this is an interesting direction for future work.

Domain	mean reward	std reward	eval timesteps	eval episodes (m)	$\mathrm{rank}(\Lambda_m^\pi)$	$\lambda_{\min}(\Lambda_m^\pi)$	UNISOFT
Acrobot-v1	-83.3	17.1	149970	1778	64	0.0003	✔
AntBulletEnv-v0	1912.7	106.0	150000	150	64	0.0008	✔
BipedalWalker-v3	276.1	25.8	149707	198	28	0	
CartPole-v1	500.0	0.0	150000	300	2	0	
HopperBulletEnv-v0	14.0	0.8	149997	26620	59	0	
MountainCar-v0	-107.3	20.1	149944	1397	44	0	
MountainCarContinuous-v0	92.4	0.1	149984	1948	10	0	
Pendulum-v0	-153.3	92.9	150000	750	32	0	

Table 5: Results for A2C policy network of dimension $[64, 64]$ and value network of dimension $[64, 64]$. These dimensions represent the size of the hidden layers (with tanh activation function). We highlighted the environments where A2C achieved good performance.