Reinforcement Learning in Linear MDPs: Constant Regret and Representation Selection

Matteo Papini [1] **Andrea Tirinzoni** [2] **Aldo Pacchiano** [3 4] **Marcello Restelli** [1] **Alessandro Lazaric** [3]
Matteo Pirotta [3]

Abstract

[1] We study the role of the representation in finite-horizon Markov Decision Processes (MDPs) with linear structure. We provide a necessary condition for achieving constant regret in any MDP with linear reward representation (even with known dynamics). This result encompasses the well-known scenario of low-rank MDPs and, more generally, zero inherent Bellman error. We demonstrate that this condition is not only necessary but also sufficient for these classes, by deriving a constant regret bound for two optimistic algorithms. As far as we know, this is the first constant regret result for MDPs. Finally, we study the problem of representation selection showing that our proposed algorithm achieves constant regret when one of the given representations is "good". Furthermore, our algorithm can combine representations and achieve constant regret also when none of the representations would.

1. Introduction

In supervised learning, it is well understood that a "good" representation is one that allows to accurately fit any target function of interest. We refer to such case, as a *realizable* representation.[2] In Reinforcement Learning (RL), realizability is a more subtle concept, 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 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., Du et al., 2020; Lattimore et al., 2020; Hao 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. We refer to such case as *learnable* representation.

Several works have focused on online learning and studied sufficient assumptions for learnable representations (e.g., Jaksch et al., 2010; Azar et al., 2012; 2017; Jin et al., 2020; 2021; Zanette et al., 2020b;a; Yang & Wang, 2019; Ayoub et al., 2020; Zhang et al., 2021). In all these settings, the representation enables efficient learning and it is provided as input to the algorithm. An alternative approach is to learn such representations (e.g., Ortner et al., 2014; 2019; Lee et al., 2021; Du et al., 2019; Agarwal et al., 2020; Modi et al., 2021). While this literature focuses on finding learnable representations, it does not study the impact of the representation on the learning process itself and its sample complexity or regret. On the other hand, Hao et al. (2020); Papini et al. (2021) have recently shown that certain learnable representations display non-trivial properties that enable much better performance in linear bandits. To the best of our knowledge, the impact of similar properties on RL algorithms is largely unexplored.

In this paper, we investigate the concept of "good" representation in the the settings of zero inherent Bellman error (Zanette et al., 2020b) and low-rank structure (e.g., Jin et al., 2020). **1)** We provide a necessary condition for a "good" representation to enable constant regret in any problem with linear reward parametrization. **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 "good" representation condition is satisfied. As a consequence, we show that good representations are not only necessary but also sufficient for constant regret in MDPs under Bellman closure (i.e., zero inherent Bellman error) or low-rank assumptions. **3)** We develop an algorithm, called LSVI-LEADER, for representation selec-

[1]Politecnico di Milano [2]INRIA Lille [3]Facebook AI Research [4]UC Berkeley. Correspondence to: Matteo Papini <matteo.papini@polimi.it>.

ICML 2021 Workshop on Reinforcement Learning Theory. Copyright 2021 by the author(s).

[1] Extended abstract. Full version available on arXiv.

[2] More formally, the realizability assumption states that the space of functions induced by the representation contains a function with zero risk for the target function at hand. As such, we use "realizable" representation as a short-hand for "a representation that defines a space satisfying the realizability condition". Similarly, we use "learnable" representation for a representation such that finding a near-optimal policy is a learnable problem.

tion in low-rank MDPs. We prove that in low-rank MDPs, LSVI-LEADER is able to combine representations to form a "good" one and achieve constant regret even when none of the individual representations would.

2. Preliminaries

We consider a 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 π as $Q_h^\pi(s,a) = r_h(s,a) + \mathbb{E}_\pi\Big[\sum_{i=h+1}^H r_i(s_i,a_i)\Big]$, and $V_h^\pi(s,a) = Q_h^\pi(s,\pi_h(s))$. The optimal Bellman equation (and Bellman operator L_h) at stage $h \in [H]$ is defined as:

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

Under certain regularity conditions (e.g., Bertsekas & Shreve, 2004), 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. 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 & 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,* $\forall h \in [H]$,

$$\sup_{Q_{h+1}\in\mathcal{Q}_{h+1}} \inf_{Q_h\in\mathcal{Q}_h} \|Q_h - L_h Q_{h+1}\|_\infty = 0.$$

This assumption enables learnability for value-iteration-based algorithms (Munos & Szepesvári, 2008). In the context of regret minimization, Zanette et al. (2020b) proposed an 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,* $\forall s, a, h, s'$,

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

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* $\max\{\|\mu_h(\mathcal{S})\|_2, \|\theta_h\|_2\} \le \sqrt{d}$ *and* $\|\phi_h(s,a)\|_2 \le 1$*, for any* s,a,h.

This assumption is *strictly* stronger than Bellman closure (Zanette et al., 2020b) and it implies the realizability of *any* value function. 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.

Assumption 3. *The suboptimality gap of taking action* a *in state* s *at stage* h *is defined as:* $\Delta_h(s,a) = V_h^\star(s) - Q_h^\star(s,a)$. *We assume the minimum positive gap* $\Delta_{\min} = \min_{s,a,h}\{\Delta_h(s,a) | \Delta_h(s,a) > 0\}$ *is well defined.*

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. *We assume the optimal policy* $\pi^\star$ *is unique, i.e.,* $|\operatorname{argmax}_a\{Q_h^\star(s,a)\}| = 1$ *for any* s,h. *A feature map is* UniSOFT *(Universally Spanning Optimal FeaTures) for an MDP if it is learnable (see Asm. 1-2), and for all* $h \in [H]$ *the following holds:*

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

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. 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.

[3] ELEANOR works under the weaker assumption of low IBE.

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.

It is interesting to look into Asm. 4 from an alternative perspective. Denote by $0 \leq \lambda_{h,1} \leq \ldots \leq \lambda_{h,d}$ 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.

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. The proof is reported in App. C.

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 the condition* UniSOFT *does not hold (Asm. 4). 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 (i.e., unstructured) 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 class of MDPs strictly generalizes low-rank MDPs (e.g., Jin et al., 2020), linear-mixture MDPs with unknown linear rewards (e.g., Yang & Wang, 2019), and MDPs with Bellman closure (e.g., Zanette et al., 2020b).

3.2. UniSOFT is Sufficient for Constant Regret

While the UniSOFT condition is necessary for achieving constant regret in a large class of MDPs, here, we show that ELEANOR and LSVI-UCB attain constant regret when the UniSOFT assumption holds. These results show that the UniSOFT condition is also sufficient in MDPs with low-rank and Bellman closure structure. Proofs in App. D.

Theorem 6. *Consider an MDP and a representation $\{\phi_h\}_{h\in[H]}$ satisfying the Bellman closure (Asm. 1) and* UniSOFT *assumptions (Asm. 4). If $\Delta_{\min} > 0$ (Asm. 3), then with probability at least $1 - 3\delta$, ELEANOR[4] 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 $\overline{\kappa} = \max_h\{\kappa_h\} < \infty$ is the last episode ELEANOR suffers a non-zero regret (see Eq. 4 in appendix for an implicit definition of $\overline{\kappa}$).

As expected, $\overline{\kappa}$ is independent of the number of episodes K, thus making the regret bound constant and depending only on "static" MDP and representation characteristics. Furthermore, the bound should be read as minimum between the constant regret and the minimax regret $O(\sqrt{K})$, which may be tighter for small K.

This bound leverages the minimax regret bound of ELEANOR. Unfortunately, whether ELEANOR can achieve problem-dependent logarithmic regret based on local gaps is an open question in the literature. The 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.

For LSVI-UCB, we derive a more refined constant-regret guarantee by leveraging the problem-dependent bound.

Theorem 7. *Consider an MDP and a representation $\{\phi_h\}_{h\in[H]}$ satisfying the low-rank (Asm. 2) and* UniSOFT *assumptions (Asm. 4). If $\Delta_{\min} > 0$ (Asm. 3), then with probability $1 - 3\delta$, LSVI-UCB suffers a* constant *regret*

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

where $\overline{\kappa}$ is defined as in Thm. 6.

4. Representation Selection

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 intractable algorithm, we focus on LSVI-UCB and low-rank MDPs (Asm. 2).

Given a set of N representations $\{\Phi_j\}_{j\in[N]}$ satisfying Asm. 2, where $\Phi_j = \{\phi_h^{(j)}\}_{h\in[H]}$, we show that it is possible to design a learning algorithm able to perform as well as the best representation, and thus achieve constant regret if a UniSOFT representation is present in this set. The algorithm, called LSVI-LEADER, is reported in Alg. 1.At

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

Algorithm 1 LSVI-LEADER

```
Input: Representations $\{\Phi_j\}_{j\in[M]}$, 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
        $\Lambda_h^k(j) = \lambda I + \sum_{i=1}^{k-1} \phi_h^{(j)}(s_h^i, a_h^i)\phi_h^{(j)}(s_h^i, a_h^i)^\mathsf{T} \;\forall\, j \in [M]$.
        $\boldsymbol{w}_h^k(j) = \Lambda_h^k(j)^{-1} \sum_{i=1}^{k-1} \phi_h^{(j)}(s_h^i, a_h^i) \left( r_h(s_h^i, a_h^i) + \max_{a\in\mathcal{A}} \overline{Q}_{h+1}^k(s_{h+1}^i, a) \right),\ \forall j \in [M]$
        $\overline{Q}_h^k(s,a) = \min\left\{ H, \min_{j\in[M]} \left( \phi_h^{(j)}(s,a)^\mathsf{T}\boldsymbol{w}_h^k(j) + \beta_k \left\| \phi_h^{(j)}(s,a) \right\|_{\Lambda_h^k(i)^{-1}} \right) \right\}$
    end for
    Collect a trajectory through policy $\pi_h^k(s_h^k) := \mathrm{argmax}_{a\in\mathcal{A}} \overline{Q}_h^k(s_h^k, a)$.
end for
```

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 states, actions, and stages. Proof are reported in App. E.

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 8. *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.*[5] *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. 7).*

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 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). 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).

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 9. *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 $\phi_h^{(j)}(s,a) \in \mathrm{span}\left\{\phi_h^{(j)}(s, \pi_h^\star(s)) | \rho_h^\star(s) > 0\right\}$.*

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 10. *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: $R(K) \le \min_{z\in\mathcal{Z}} \widetilde{R}(\widetilde{\kappa}, z, \{\beta_k\})$, where $\mathcal{Z}$, $\widetilde{R}$ and β_k are defined as in Thm. 8.*

Under the UNISOFT-mixing condition, LSVI-LEADER may not converge to selecting a single representation for

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

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. 8, 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

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *NIPS*, pp. 2312–2320, 2011.

Abbasi-Yadkori, Y., Bartlett, P. L., Bhatia, K., Lazic, N., Szepesvári, C., and Weisz, G. POLITEX: regret bounds for policy iteration using expert prediction. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3692–3702. PMLR, 2019.

Agarwal, A., Kakade, S. M., Krishnamurthy, A., and Sun, W. FLAMBE: structural complexity and representation learning of low rank mdps. In *NeurIPS*, 2020.

Ayoub, A., Jia, Z., Szepesvári, C., Wang, M., and Yang, L. Model-based reinforcement learning with value-targeted regression. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 463–474. PMLR, 2020.

Azar, M. G., Munos, R., and Kappen, B. On the sample complexity of reinforcement learning with a generative model. In *ICML*. icml.cc / Omnipress, 2012.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pp. 263–272. PMLR, 2017.

Bertsekas, D. P. and Shreve, S. *Stochastic optimal control: the discrete-time case*. 2004.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pp. 578–598. PMLR, 2021.

Du, S. S., Krishnamurthy, A., Jiang, N., Agarwal, A., Dudík, M., and Langford, J. Provably efficient RL with rich observations via latent state decoding. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1665–1674. PMLR, 2019.

Du, S. S., Kakade, S. M., Wang, R., and Yang, L. F. Is a good representation sufficient for sample efficient reinforcement learning? In *ICLR*. OpenReview.net, 2020.

Hao, B., Lattimore, T., and Szepesvári, C. Adaptive exploration in linear contextual bandit. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3536–3545. PMLR, 2020.

Hao, B., Lattimore, T., Szepesvári, C., and Wang, M. Online sparse reinforcement learning. In *AISTATS*, volume 130 of *Proceedings of Machine Learning Research*, pp. 316–324. PMLR, 2021.

He, J., Zhou, D., and Gu, Q. Logarithmic regret for reinforcement learning with linear function approximation. *CoRR*, abs/2011.11566, 2020.

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

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *COLT*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2137–2143. PMLR, 2020.

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

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

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

Lattimore, T., Szepesvári, C., and Weisz, G. Learning with good feature representations in bandits and in RL with a generative model. In *ICML*, volume 119 of *Proceedings* *of Machine Learning Research*, pp. 5662–5670. PMLR, 2020.

Lee, J. N., Pacchiano, A., Muthukumar, V., Kong, W., and Brunskill, E. Online model selection for reinforcement learning with function approximation. In *AISTATS*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3340–3348. PMLR, 2021.

Modi, A., Chen, J., Krishnamurthy, A., Jiang, N., and Agarwal, A. Model-free representation learning and exploration in low-rank mdps. *CoRR*, abs/2102.07035, 2021.

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

Ortner, R., Maillard, O., and Ryabko, D. Selecting near-optimal approximate state representations in reinforcement learning. In *ALT*, volume 8776 of *Lecture Notes in Computer Science*, pp. 140–154. Springer, 2014.

Ortner, R., Pirotta, M., Lazaric, A., Fruit, R., and Maillard, O. Regret bounds for learning state representations in reinforcement learning. In *NeurIPS*, pp. 12717–12727, 2019.

Papini, M., Tirinzoni, A., Restelli, M., Lazaric, A., and Pirotta, M. Leveraging good representations in linear contextual bandits. *CoRR*, abs/2104.03781, 2021.

Simchowitz, M. and Jamieson, K. G. Non-asymptotic gap-dependent regret bounds for tabular mdps. In *NeurIPS*, pp. 1151–1160, 2019.

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

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

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

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

Zanette, A., Brandfonbrener, D., Brunskill, E., Pirotta, M., and Lazaric, A. Frequentist regret bounds for randomized least-squares value iteration. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1954–1964. PMLR, 2020a.

Zanette, A., Lazaric, A., Kochenderfer, M. J., and Brunskill, E. Learning near optimal policies with low inherent bellman error. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10978–10989. PMLR, 2020b.

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

Content

A. Notation

Table 1. 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 (31) for ELEANOR and (41) 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
$\overline{V}_h^k(s)$	$=$	$\max_{a \in \mathcal{A}} \overline{Q}_h^k(s, a)$

B. Proof Sketches

We start providing brief proof sketches of the main results in the paper. Details will be provided in the following section.

B.1. 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 & Jamieson, 2019; Xu et al., 2021).

Definition 11 (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 such that $\mathbb{E}_M^{\mathsf{A}}\left[R(K)\right] \leq c_M K^{\alpha}$.*

For instance, LSVI-UCB and ELEANOR are $1/2$-consistent on the class of low-rank and Bellman-closure MDPs, respectively, where they enjoy worst-case $O(\sqrt{K})$ regret bounds.

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 & Szepesvari, 2017; Hao et al., 2020; Tirinzoni et al., 2020).

Lemma 12. *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 $\Delta(\pi) > 0$) such that the vector $\Psi_h^\pi - \Psi_h^\star$ does not belong to 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. 12, 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.

B.2. Combined proof sketch of Thm. 6 and Thm. 7.

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: $\pi_h^k(s) = \arg\max_{a\in\mathcal{A}}\{\overline{Q}_h^k(s,a)\}$.

(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. $\overline{Q}^k_h(s,a) + C_h\beta_k \|\phi_h(s,a)\|_{(\Lambda^k_h)^{-1}} \geq Q^\star_h(s,a)$.

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

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,

$$\Lambda^{k+1}_h \succeq k\Lambda^\star_h + \lambda I - \Delta^{-1}_{\min} g(k) I - \widetilde{O}(\sqrt{k}) I, \tag{1}$$

where $\Lambda^\star_h = \mathbb{E}_{s\sim\rho^\star_h}[\phi^\star_h(s)\phi^\star_h(s)^\mathsf{T}]$. 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),

$$\beta_k \|\phi_h(s,a)\|_{(\Lambda^k_h)^{-1}} \leq \beta_k \frac{k - \widetilde{O}(\sqrt{k})}{(k\lambda^+_h - \widetilde{O}(\sqrt{k}))^{3/2}} = \widetilde{O}(k^{-1/2}), \tag{2}$$

where λ^+_h is the minimum *nonzero* eigenvalue of $\Lambda^\star_h$. From (2), 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

$$V^\star_1(s^k_1) - V^{\pi^k}_1(s^k_1) = \mathbb{E}_{\pi^k}\left[\sum_{h=1}^{H}\Delta_h(s_h,a_h)|s_1 = s^k_1\right], \tag{3}$$

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

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

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 (2) to control the feature norms. Thus, there exists an episode κ_h independent of K satisfying

$$\Delta_h(s,\pi^k_h(s)) \leq \beta_{\kappa_h}\sum_{i=h}^{H}(2 + \mathbb{1}_{i=h>1}C_h)\frac{\kappa_h - 8\sqrt{\kappa_h\log(2d\kappa_h H/\delta)} - g(\kappa_h)}{(\kappa_h\lambda^+_i - 8\sqrt{\kappa_h\log(2d\kappa_h H/\delta)} - g(\kappa_h))^{3/2}} < \Delta_{\min}, \tag{4}$$

By definition of minimum gap, then $\Delta_h(s,\pi^k_h(s)) = 0$ for $k > \kappa_h$. Then, for $k > \overline{\kappa} = \max_h\{\kappa_h\}$, $V^\star_1(s^k_1) - V^{\pi^k}_1(s^k_1) = 0$. 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 (1). 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.

B.3. Proof sketch of Thm. 8.

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 $\pi \in \Pi, s \in \mathcal{S}, a \in \mathcal{A}, h \in [H], k \in [K]$,

$$\overline{Q}^k_h(s,a) - Q^\pi_h(s,a) \leq 2\beta_k \min_{j\in[N]}\left\|\phi^{(j)}_h(s,a)\right\|_{\Lambda^k_h(j)^{-1}} + \mathbb{E}_{s'\sim p_h(s,a)}\left[\overline{V}^k_{h+1}(s') - V^\pi_{h+1}(s')\right].$$

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. 7 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.

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 $\pi^\star$: $|\{a : Q_h^\star(s,a) = V_h^\star(s)\}| = 1$ and $\pi_h^\star(s) = \mathrm{argmax}_a\, Q_h^\star(s,a)$ 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 $\mathbb{P}_M^{\mathsf{A}}$ ($\mathbb{E}_M^{\mathsf{A}}$) to denote the probability (expectation) operator under MDP M and the chosen algorithm A.

C.1. Proof of Lemma 12

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,

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

where $\Lambda_h^K := \sum_{s,a} N_h^K(s,a)\phi(s,a)\phi(s,a)^T$ and $N_h^K(s,a) := \sum_{k=1}^K \mathbb{1}\left\{s_h^k = s, a_h^k = a\right\}$.

Suppose that, for sufficiently large K, the matrix $\mathbb{E}_M^{\mathsf{A}}[\Lambda_h^K]$ is invertible.[6] 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

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

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 $\epsilon \geq 0$,

$$\min_{\widetilde{M}\in\Gamma_h^\epsilon(\pi)} \mathrm{KL}(\mathbb{P}_M, \mathbb{P}_{\widetilde{M}}) = \frac{(\Delta(\pi)+\epsilon)^2}{2\|\Psi_h^\pi - \Psi_h^\star\|^2_{\mathbb{E}_M^{\mathsf{A}}[\Lambda_h^K]^{-1}}}. \tag{5}$$

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,

$$\mathrm{KL}(\mathbb{P}_M, \mathbb{P}_{\widetilde{M}}) \geq \log\frac{1}{\mathbb{P}_M(E_K) + \mathbb{P}_{\widetilde{M}}(E^c)} - \log 2. \tag{6}$$

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,

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

[6] (Lattimore & 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.

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 (6), 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 (5), 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 12 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),

$$\mathbb{E}_M^{\mathsf{A}}\left[\mathrm{R}(K)\right] \leq C_M. \tag{7}$$

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,

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

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 $\Psi_h^\pi \notin \mathrm{span}\left\{\phi_h^\star(s) | \rho_h^\star(s) > 0\right\}$.

Step 3. Concluding the proof. Combining Lemma 12 with Step 1 and Step 2, we have that, for some $h \in [H]$ and policy π such that $\Delta(\pi) > 0$ and $\Psi_h^\pi \notin \mathrm{span}\left\{\phi_h^\star(s) | \rho_h^\star(s) > 0\right\}$,

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

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 32, we have that there exists an $\epsilon > 0$ (independent of K) such that $\|\Psi_h^\pi - \Psi_h^\star\|_{(\Lambda_h^\star + \eta I)^{-1}} \ge \frac{\epsilon}{\sqrt{\eta}}$. Therefore, we get that

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

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

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

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 & 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 $h=2,\dots,H$, $k\geq 1$, $s\in\mathcal{S}$ 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 $K\geq 1$:*

$$R(K) \leq H\beta_K\sqrt{2dK\log(1+K/\lambda)} + 2H^2\sqrt{K\log(2HK/\delta)} = \widetilde{O}(\sqrt{K}). \tag{8}$$

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

$$\begin{aligned}
R(K) &= \sum_{k=1}^K V_1^\star(s_1^k) - V_1^{\pi^k}(s_1^k) \\
&\leq \sum_{k=1}^K \overline{V}_1^k(s_1^k) - V_1^{\pi^k}(s_1^k) && \text{(a)} \quad (9)\\
&\leq \underbrace{2\sum_{h=1}^H \beta_K \sum_{k=1}^K \left\|\phi_h(s_h^k,a_h^k)\right\|_{(\Lambda_h^k)^{-1}}}_{(A)} + \underbrace{\sum_{k=1}^K\sum_{h=1}^H \zeta_h^k}_{(B)}, && (10)
\end{aligned}$$

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

$$\zeta_h^k = \mathbb{E}_{s'\sim p_h(s_h^k,a_h^k)}[\overline{V}_{h+1}^k(s') - V_{h+1}^{\pi^k}(s')] - \overline{V}_{h+1}^k(s_{h+1}^k) + V_{h+1}^{\pi^k}(s_{h+1}^k), \tag{11}$$

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):

$$\begin{aligned}
(A) &= 2\beta_K \sum_{h=1}^H\sum_{k=1}^K \left\|\phi_h(s_h^k,a_h^k)\right\|_{(\Lambda_h^k)^{-1}} && (12)\\
2\beta_K\sum_{h=1}^H &\leq \sqrt{K\sum_{k=1}^K \left\|\phi_h(s_h^k,a_h^k)\right\|_{(\Lambda_h^k)^{-1}}^2} && (13)\\
&\leq H\beta_K\sqrt{2dK\log(1+K/\lambda)}. && (14)
\end{aligned}$$

Since ζ_h^k is a martingale difference sequence with $\zeta_h^k \leq 2H$, we can use Azuma's inequality (Prop. 25) to bound (B):

$$\sum_{k=1}^{K} \zeta_h^k \leq 2H\sqrt{K \log(2K/\delta_h)}, \tag{15}$$

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:

$$(B) = \sum_{h=1}^{H}\sum_{k=1}^{K} \zeta_h^k \leq 2H^2\sqrt{K \log(2HK/\delta)}. \tag{16}$$

□

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 $k \geq 1$:*

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

Proof.

$$\begin{aligned}
\Delta_h(s, \pi_h^k(s)) &= V_h^\star(s) - Q_h^\star(s_h^k, \pi_h^k(s)) && (17)\\
&\leq V_h^\star(s) - Q_h^{\pi^k}(s_h^k, \pi_h^k(s)) && (18)\\
&= V_h^\star(s) - V_h^{\pi^k}(s) && (19)\\
&= Q_h^\star(s, \pi_h^\star(s)) - V_h^{\pi^k}(s) && (20)\\
&\leq \overline{Q}_h^k(s, \pi_h^\star(s)) + \mathbb{1}\left\{h>1\right\} C_h \beta_k \left\|\phi_h(s,\pi_h^\star(s))\right\|_{(\Lambda_h^k)^{-1}} - V_h^{\pi^k}(s) && (21)\\
&\leq \overline{V}_h^k(s) + \mathbb{1}\left\{h>1\right\} C\sqrt{\gamma_{hk}} \left\|\phi_h(s,\pi_h^\star(s))\right\|_{\Lambda_{hk}^{-1}} - V_h^{\pi^k}(s) && (22)\\
&\leq 2\,\mathbb{E}_{\pi^k}\left[\sum_{i=h}^{H} \beta_k \left\|\phi_i(s_i,a_i)\right\|_{(\Lambda_i^k)^{-1}} \middle| s_h = s\right] && (23)\\
&\qquad + \mathbb{1}\left\{h>1\right\} C_h \beta_k \left\|\phi_h(s_h^k, \pi_h^\star(s_h^k))\right\|_{(\Lambda_h^k)^{-1}},
\end{aligned}$$

where (21) 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 $K \geq 1$:*

$$R(K) \leq g(\overline{\kappa}) = O(1), \tag{24}$$

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 31 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 $k \geq \widetilde{\kappa} = \max_{h\in[H]} \widetilde{\kappa}_h$:

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

$$+ \mathbb{1}\left\{h>1\right\} C_h \beta_k \left\|\phi^\star(s)\right\|_{(\Lambda_h^k)^{-1}} \tag{25}$$

$$\leq (2+\mathbb{1}\left\{h>1\right\} C_h)\beta_k \sum_{i=h}^{H} \frac{k+\lambda-g(k)-8\sqrt{k\log(2dHk/\delta)}}{(k\lambda_i^+ +\lambda - g(k) - 8\sqrt{k\log(2dHk/\delta)})^{3/2}}. \tag{26}$$

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 $k > \kappa_h$:

$$\Delta_h(s, \pi_h^k(s)) \leq \Delta_{\min}. \tag{27}$$

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. 27:

$$R(K) = \sum_{k=1}^{K} \mathbb{E}_{\pi^k}\left[\sum_{h=1}^{H} \Delta_h(s_h,a_h) \middle| s_1 = s_1^k\right] \tag{28}$$

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

$$= R(\overline{\kappa}) \leq g(\overline{\kappa}). \tag{30}$$

□

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 6.

Let:

$$\beta_k = H\sqrt{\frac{d}{2}\log(1+k/d) + d\log(1+4\sqrt{dk}) + \log\frac{2Hk^2}{\delta}} + 1, \tag{31}$$

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 $\boldsymbol{\theta}_1^\star, \dots, \boldsymbol{\theta}_H^\star$ such that (Lemma 6 from Zanette et al., 2020b):

$$Q_h^\star(s,a) = \phi_h(s,a)^\mathsf{T}\boldsymbol{\theta}_h^\star. \tag{32}$$

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

$$\left\|\boldsymbol{\theta}_h^\star - \widehat{\boldsymbol{\theta}}_h^k\right\|_{\Lambda_h^k} \leq \beta_k. \tag{33}$$

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

$$\left\|\boldsymbol{\theta}_h^\star - \overline{\boldsymbol{\theta}}_h^k\right\|_{\Lambda_h^k} \leq 2\beta_k. \tag{34}$$

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

$$\begin{aligned}
Q_h^\star(s_h,a_h) &= \phi_h(s_h,a_h)^\mathsf{T}\boldsymbol{\theta}_h^\star && (35)\\
&= \phi_h(s_h,a_h)^\mathsf{T}\overline{\boldsymbol{\theta}}_h^k + \phi_h(s_h,a_h)^\mathsf{T}(\boldsymbol{\theta}_h^\star - \overline{\boldsymbol{\theta}}_h^k) && (36)\\
&\leq \overline{Q}_h^k(s_h,a_h) + \|\phi(s_h,a_h)\|_{(\Lambda_h^k)^{-1}} \left\|\boldsymbol{\theta}_h^\star - \overline{\boldsymbol{\theta}}_h^k\right\|_{\Lambda_h^k} && (37)\\
&\leq \overline{Q}_h^k(s_h,a_h) + 2\beta_k \|\phi(s_h,a_h)\|_{(\Lambda_h^k)^{-1}}\,, && (38)
\end{aligned}$$

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:

$$\begin{aligned}
R(K) &\leq H^2\left(\sqrt{\frac{d}{2}\log(1+\overline{\kappa}/d) + d\log(1+4\sqrt{d\overline{\kappa}}) + \log(H\overline{\kappa}^2) + \log\frac{2}{\delta}} + H\right)\\
&\qquad \times \sqrt{2d\overline{\kappa}\log(1+\overline{\kappa}/\lambda)} + 2H^2\sqrt{\overline{\kappa}\log(2H\overline{\kappa}/\delta)} && (39)\\
&\lesssim H^{3/2}d\sqrt{\overline{\tau}\log\frac{\overline{\tau}}{\delta}}, && (40)
\end{aligned}$$

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 7.

Let:

$$\beta_k = c_\beta dH\sqrt{\log(2dHk/\delta)}, \tag{41}$$

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):

$$g(k) = 9HG(k)\log G(k) + \frac{16H^2}{3}\log\frac{\log\lceil Hk\rceil}{\delta} + 2, \tag{42}$$

where:

$$G(k) \propto \frac{d^3H^4\log(4dH^2k(k+1)\log(H/\Delta_{\min})/\delta)}{\Delta_{\min}}. \tag{43}$$

So:

$$R(K) \leq g(\overline{\kappa}) \simeq \frac{d^3H^5}{\Delta_{\min}}\log\left(dH^2\overline{\kappa}/\delta\right). \tag{44}$$

□

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.

E. Representation Selection: Proofs of Section 4

The main ingredient behind the proofs of Theorems 8 and 10 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 20. *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 $\|\boldsymbol{w}\| \le L$, $\beta \in [0,B]$ and the minimum eigenvalue of $\boldsymbol{\Lambda}_j$ satisfies $\lambda_{\min}(\boldsymbol{\Lambda}_j) \ge \lambda$. Assume $\|\boldsymbol{\phi}(s,a)\| \le 1$ for all (s,a) pairs and let $\mathcal{N}_\epsilon$ be the $\epsilon-$covering number of $\mathcal{V}$ with respect to the distance $\mathrm{dist}(V,V') = \sup_s |V(s) - V'(s)|$. Then,

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

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

$$V(\cdot) = \min\left(\min_{j\in[N]} \max_a \boldsymbol{w}_j^\top \phi_j(\cdot,a) + \sqrt{\phi_j(\cdot,a)^\top \boldsymbol{A}_j\phi_j(\cdot,a)}, H\right) \tag{45}$$

for $\|\boldsymbol{w}_j\| \le L$ and $\|\boldsymbol{A}_j\| \le B^2\lambda^{-1}$. For any two functions $V_1, V_2 \in \mathcal{V}$, let them take the form in Equation 45 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

$$\begin{aligned}
\mathrm{dist}(V_1,V_2) \le& \sup_{j,s,a}\Bigg|\left[\left(\boldsymbol{w}_j^{(1)}\right)^\top \phi_j(\cdot,a) + \sqrt{\phi_j(\cdot,a)^\top \boldsymbol{A}_j^{(1)}\phi_j(\cdot,a)}\right] - \\
&\left[\left(\boldsymbol{w}_j^{(2)}\right)^\top \phi_j(\cdot,a) + \sqrt{\phi_j(\cdot,a)^\top \boldsymbol{A}_j^{(2)}\phi_j(\cdot,a)}\right]\Bigg| \qquad (46)\\
\le& \sup_j\left(\sup_{\|\phi_j\|\le 1}\left|\left[\left(\boldsymbol{w}_j^{(1)}\right)^\top \phi_j + \sqrt{\phi_j^\top \boldsymbol{A}_j^{(1)}\phi_j}\right] - \left[\left(\boldsymbol{w}_j^{(2)}\right)^\top \phi_j + \sqrt{\phi_j^\top \boldsymbol{A}_j^{(2)}\phi_j}\right]\right|\right)\\
\le& \sup_j\left(\sup_{\|\phi_j\|\le 1}\left|\left(\boldsymbol{w}_j^{(1)} - \boldsymbol{w}_j^{(2)}\right)^\top \phi_j\right| + \sup_{\|\phi_j\|\le 1}\sqrt{\left|\phi_j^\top\left(\boldsymbol{A}_j^{(1)} - \boldsymbol{A}_j^{(2)}\right)\phi_j\right|}\right)\\
=& \sup_j \|\boldsymbol{w}_j^{(1)} - \boldsymbol{w}_j^{(2)}\| + \sqrt{\|\boldsymbol{A}_j^{(1)} - \boldsymbol{A}_j^{(2)}\|}\\
\le& \sup_j \|\boldsymbol{w}_j^{(1)} - \boldsymbol{w}_j^{(2)}\| + \sqrt{\|\boldsymbol{A}_j^{(1)} - \boldsymbol{A}_j^{(2)}\|_F} \qquad (47)
\end{aligned}$$

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

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

By Equation 47, 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 $\mathrm{dist}(V_1,V_2) \le \epsilon$. Hence it holds that $\mathcal{N}_\epsilon \le \left(|\mathcal{C}_j^{\boldsymbol{w}}||\mathcal{C}_j^{\boldsymbol{A}}|\right)^N$, 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 21 (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 $1-2\delta$:*

$$\left\|\sum_{i=1}^k \phi_h^{(j)}(s_h^i, a_h^i)\left(\overline{V}_{h+1}^k(s_{h+1}^i) - \mathbb{P}_h\overline{V}_{h+1}^k(s_h^i, a_h^i)\right)\right\|_{\Lambda_{h,k}^{-1}(j)} \leq CdH\sqrt{N\log(2N(c_\beta+1)dHk/\delta)}, \tag{48}$$

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 22 (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 $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$:

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

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,

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

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 21 (which holds with probability at least $1 - 2\delta$) we have,

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

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,

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

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

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

For some constant C'. The result follows.

□

Lemma 23. *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 $s\in\mathcal{S}$,*

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

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

Proof. Note that $\overline{V}^k_h(s) - V^{\pi^k}_h(s) = \overline{Q}^k_h(s,\pi^k_h(s)) - Q^{\pi^k}_h(s,\pi^k_h(s))$. Using Lemma 22, for any $j\in[N]$

$$\begin{aligned} Q^{\pi^k}_h(s,\pi^k_h(s)) \geq \langle\phi^{(j)}_h(s,\pi^k_h(s)), \boldsymbol{w}^k_h(j)\rangle - \\ \mathbb{E}_{s'\sim p_h(s,\pi^k_h(s))}[\overline{V}^k_{h+1}(s') - V^{\pi^k}_{h+1}(s')] - \beta_{h,k}\|\phi^{(j)}_h(s,\pi^k_h(s))\|_{\Lambda^{-1}_{h,k}(j)} \end{aligned}$$

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

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

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

$$\overline{V}^k_h(s) - V^{\pi^k}_h(s) \leq 2\beta_{h,k} \min_{j\in[N]} \|\phi^{(j)}_h(s,\pi^k_h(s))\|_{\Lambda^{-1}_{h,k}(j)} + \mathbb{E}_{s'\sim p_h(s,\pi^k_h(s))}[\overline{V}^k_{h+1}(s') - V^{\pi^k}_{h+1}(s')].$$

□

Finally we show this implies optimism holds,

Lemma 24. *[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 $\{\overline{Q}^k_h\}_{h\in[H]}$ functions of LSVI-LEADER satisfy,*

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

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 $\overline{Q}^k_H(s,a)\geq Q^*_H(s,a)$ since the value function at $H+1$ is zero and by Lemma 22 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 $j\in[N]$,

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

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

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

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

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

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 22 we have, for all $k \in [K]$ and all $j \in [N]$

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

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 8. Having proven Lemma 23 and that optimism holds for LSVI-LEADER (Lemma 24), 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. 7 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 10. The proof follows the template of Thm 7, but as shown in Lemma 23, the confidence sets of LSVI-LEADER scale with the minimum w.r.t. j of the feature norms. In place of Equation 2, and with the aid of Lemma 31 we see that since the collection of feature maps $\{\Phi_j\}_{j\in[M]}$ is UNISOFT-mixing for all reachable s, a:

$$\begin{aligned}\beta_k \min_{j\in[N]}\left\|\phi_h^{(j)}(s,a)\right\|_{\Lambda_{h,k}^{-1}(j)} \leq &\beta_k \frac{k+\lambda-g(k)-8\sqrt{k\log(2NdHk/\delta)}}{(k\lambda^+(h,s,a)+\lambda-g(k)-8\sqrt{k\log(2NdHk/\delta)})^{3/2}} \\ &= \widetilde{O}(k^{-1/2}),\end{aligned} \tag{49}$$

where $g(k) = \widetilde{O}(\sqrt{k})$ is the regret upper bound from Thm. 8,

$$\lambda^+(h,s,a) = \max_{j\in\mathcal{J}(h,s,a)} \lambda_{h,j}^+, \tag{50}$$

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 31 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. 9) guarantees $\mathcal{J}(h,s,a)$ is always nonempty.

By (49) 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 $k > \kappa_h$:

$$\begin{aligned}\Delta_h(s,\pi_h^k(s)) \leq 2\beta_k\, \mathbb{E}_{\pi^k}\left[\sum_{i=h}^{H} \frac{k+\lambda-g(k)-8\sqrt{k\log(2NdHk/\delta)}}{(k\lambda^+(i,s_i,a_i)+\lambda-g(k)-8\sqrt{k\log(2NdHk/\delta)})^{3/2}}\middle| s_h = s\right] \\ < \Delta_{\min}.\end{aligned} \tag{51}$$

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. 8 without the UNISOFT-mixing property.

F. Auxiliary Results

Proposition 25 (Azuma's inequality)**.** *Let $\{(Z_t, \mathcal{F}_t)\}_{t\in\mathbb{N}}$ be a martingale difference sequence such that $|Z_t| \le a$ almost surely for all $t \in \mathbb{N}$. Then, for all $\delta \in (0, 1)$,*

$$\mathbb{P}\left(\forall t \ge 1 : \left|\sum_{k=1}^{t} Z_k\right| \le a\sqrt{t\log(2t/\delta)}\right) \ge 1 - \delta. \tag{52}$$

Proposition 26 (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 $1 - \delta$:*

$$\lambda_{\max}\left(\sum_{k=1}^{t} X_k\right) \le \sqrt{8\sigma^2 \log(d/\delta)}, \tag{53}$$

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

Proposition 27 ((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:

$$\begin{aligned} V_h^\star(s) - V_h^\pi(s) &= Q_h^\star(s, \pi_h(s)) + \Delta_h(s, \pi_h(s)) - V_h^\pi(s) && (54)\\ &= r_h(s, \pi_h(s)) + \mathbb{E}_{s'\sim p_h(s,\pi_h(s))}[V_{h+1}^\star(s')] + \Delta_h(s, \pi_h(s)) - r_h(s, \pi_h(s)) \\ &\quad - \mathbb{E}_{s'\sim\mathbb{P}_h(s_h,\pi_h(s_h))}[V_{h+1}^\pi(s')] && (55)\\ &= \Delta_h(s_h, \pi_h(s_h)) + \mathbb{E}_{s'\sim\mathbb{P}_h(s_h,\pi_h(s_h))}[V_{h+1}^\star(s') - V_{h+1}^\pi(s')]. && (56) \end{aligned}$$

Unrolling the recursion up to H concludes the proof. □

Lemma 28. *Assume $R(k) \le g(k)$ for all $k \ge 1$ and Asm. 3 holds. Then, probability $1 - \delta$, for all h, k:*

$$\Lambda_h^{k+1} \succeq k\Lambda_h^\star + \lambda I - \Delta_{\min}^{-1} g(k) I - 8L^2 I\sqrt{k\log(2dkH/\delta)}. \tag{57}$$

Proof. Define a trajectory as a sequence of states and actions $\tau_h = (s_1, a_1, \dots, s_h, a_h)$. Let Γ_h denote the set of all trajectories of length h. The distribution over trajectories induced by a (deterministic) policy π is $p_h^\pi(\tau_h) = \mu(s_1)\mathbb{1}\{a_1 = \pi_1(s_1)\} p_1(s_2|s_1, a_1) \dots p_{h-1}(s_h|s_{h-1}, a_{h-1})\mathbb{1}\{a_h = \pi_h(s_h)\}$. 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:

$$E_h^k = \{\tau \in \Gamma_h \text{ s.t. } a_i = \pi_h^k(s_i) = \pi_h^\star(s_i) \text{ for } i = 1, \dots, h\}. \tag{58}$$

Then:

$$
\begin{aligned}
\Lambda_h^{k+1} - \lambda I &= \sum_{i=1}^{k} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\mathsf{T} \\
&\succeq \sum_{i=1}^{k} \mathbb{1}\left\{\tau_h^i \in E_h^i\right\} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\mathsf{T} \\
&= \sum_{i=1}^{k} \mathbb{1}\left\{\tau_h^i \in E_h^i\right\} \phi_h^\star(s_h^i)\phi_h^\star(s_h^i)^\mathsf{T} \qquad (59) \\
&= \underbrace{\sum_{i=1}^{k} \mathbb{E}_{\tau_h \sim p_h^i}\left[\mathbb{1}\left\{\tau_h \in E_h^i\right\} \phi_h^\star(s_h)\phi_h^\star(s_h)^\mathsf{T}\right]}_{(A)} \\
&\quad + \underbrace{\sum_{i=1}^{k} \Big(\mathbb{1}\left\{\tau_h^i \in E_h^i\right\} \phi_h^\star(s_h^i)\phi_h^\star(s_h^i)^\mathsf{T} - \mathbb{E}_{\tau_h \sim p_h^i}\left[\mathbb{1}\left\{\tau_h \in E_h^i\right\} \phi_h^\star(s_h)\phi_h^\star(s_h)^\mathsf{T}\right]\Big)}_{\text{(B)}},
\end{aligned}
$$

where (59) 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:

$$
(X_h^i)^2 \preceq \lambda_{\max}((X_h^i)^2) I \preceq \left\|X_h^i\right\|^2 I \preceq 4I. \qquad (60)
$$

Then by Proposition 26, with probability $1 - \delta_h^k$:

$$
\lambda_{\max}\left(\sum_{i=1}^{k} X_h^i\right) \le 4\sqrt{2k \log(d/\delta_h^k)}. \qquad (61)
$$

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:

$$
(B) = \sum_{i=1}^{k} X_h^i \preceq \lambda_{\max}\left(\sum_{i=1}^{k} X_h^i\right) I \preceq 8I\sqrt{k \log(2dHk/\delta)}. \qquad (62)
$$

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 $\tau_h \in E_h^k$:

$$
\begin{aligned}
p_h^k(\tau_h) &= \mu(s_1)\mathbb{1}\left\{a_1 = \pi_1^k(s_1)\right\} p_1(s_2|s_1, a_1) \dots p_{h-1}(s_h|s_{h-1}, a_{h-1})\mathbb{1}\left\{a_h = \pi_h^k(s_h)\right\} \\
&= \mu(s_1)\mathbb{1}\left\{a_1 = \pi_1^\star(s_1)\right\} p_1(s_2|s_1, a_1) \dots p_{h-1}(s_h|s_{h-1}, a_{h-1})\mathbb{1}\left\{a_h = \pi_h^\star(s_h)\right\} \qquad (63) \\
&= \mu(s_1)p_1(s_2|s_1, a_1) \dots p_{h-1}(s_h|s_{h-1}, a_{h-1}). \qquad (64)
\end{aligned}
$$

So:

$$
\begin{align}
(A) &= \sum_{i=1}^{k} \mathbb{E}_{\tau_h \sim p_h^i} [\mathbb{1}\left\{\tau_h \in E_h^i\right\} \phi_h^\star(s_h)\phi_h^\star(s_h)^\top] \notag \\
&= \sum_{i=1}^{k} \mathbb{E}_{\tau_h \sim p_h^\star} [\mathbb{1}\left\{\tau_h \in E_h^i\right\} \phi_h^\star(s_h)\phi_h^\star(s_h)^\top] \tag{65} \\
&= k\, \mathbb{E}_{\tau_h \sim p_h^\star} [\phi_h^\star(s_h)\phi_h^\star(s_h)^\top] - \sum_{i=1}^{k} \int_{\Gamma_h \setminus E_h^i} \phi_h^\star(s_h)\phi_h^\star(s_h)^\top p_h^\star(\mathrm{d}\tau_h) \tag{66} \\
&= k\, \mathbb{E}_{s \sim \rho_h^\star} [\phi_h^\star(s_h)\phi_h^\star(s_h)^\top] - \sum_{i=1}^{k} \int_{\Gamma_h \setminus E_h^i} \phi_h^\star(s_h)\phi_h^\star(s_h)^\top p_h^\star(\mathrm{d}\tau_h) \tag{67} \\
&\succeq k\, \mathbb{E}_{s \sim \rho_h^\star} [\phi_h^\star(s_h)\phi_h^\star(s_h)^\top] - I \sum_{i=1}^{k} \left(1 - \int_{E_h^i} p_h^\star(\mathrm{d}\tau_h)\right) \tag{68} \\
&= k\, \mathbb{E}_{s \sim \rho_h^\star} [\phi_h^\star(s_h)\phi_h^\star(s_h)^\top] - I \sum_{i=1}^{k} \left(1 - \int_{E_h^i} p_h^\star(\mathrm{d}\tau_h)\right) \tag{69} \\
&= k\, \mathbb{E}_{s \sim \rho_h^\star} [\phi_h^\star(s_h)\phi_h^\star(s_h)^\top] - I \underbrace{\sum_{i=1}^{k} \mathbb{E}_{\tau_h \sim p_h^i(\tau_h)} [\mathbb{1}\left\{\tau_h \notin E_h^i\right\}]}_{(C)}. \tag{70}
\end{align}
$$

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

$$
\begin{align}
(C) &= \sum_{i=1}^{k} \mathbb{E}_{\tau_h \sim p_h^i(\tau_h)} [\mathbb{1}\left\{\tau_h \notin E_h^i\right\}] \notag \\
&\le \sum_{i=1}^{k} \sum_{j=1}^{h} \mathbb{E}_{\pi^i} [\mathbb{1}\left\{a_j \neq \pi_j^\star(s_j)\right\}] \tag{71} \\
&\le \sum_{i=1}^{k} \sum_{j=1}^{h} \mathbb{E}_{\pi^i} [\mathbb{1}\left\{\Delta_j(s_j, a_j) \ge \Delta\right\}] \tag{72} \\
&\le \sum_{i=1}^{k} \sum_{j=1}^{h} \mathbb{E}_{\pi^i} \left[\frac{\Delta_j(s_j, a_j)}{\Delta_{\min}}\right] \tag{73} \\
&= \frac{1}{\Delta_{\min}} \sum_{i=1}^{k} \mathbb{E}_{\pi^i} \left[\sum_{j=1}^{h} \Delta_j(s_j, a_j)\right] \tag{74} \\
&\le \frac{1}{\Delta_{\min}} \sum_{i=1}^{k} \mathbb{E}_{\pi^i} \left[\sum_{h=1}^{H} \Delta_h(s_h, a_h)\right] \tag{75} \\
&= \frac{R(k)}{\Delta_{\min}} \le \frac{g(k)}{\Delta_{\min}}, \tag{76}
\end{align}
$$

where (71) is by definition of E_h^i, (72) is from the uniqueness of the optimal policy and Asm. 3, and (76) is from Proposition 27. □

Proposition 29 (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 $\lambda_1 \le \cdots \le \lambda_d$ 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 $i \in \mathcal{I}$:*

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

Proposition 30 (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 31. *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 $\phi_h(s,a) \in \operatorname{span}\{\phi^\star_h(s) | \rho^\star_h(s) > 0\}$,*

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

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 $\phi_h(s,a) \in \operatorname{span}\{\phi^\star_h(s)|\rho^\star_h(s) > 0\}$ and let $x = \phi_h(s,a)/\|\phi_h(s,a)\|$. From Lemma 28, with probability $1-\delta$, $\Lambda^k_h \succeq B^k_h$. So:

$$x^\mathsf{T}(\Lambda^k_h)^{-1}x \le x^\mathsf{T}(B^k_h)^{-1}x. \tag{77}$$

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:

$$k\lambda^+_h + \lambda - f(k), \tag{78}$$

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

$$k + \lambda - f(k). \tag{79}$$

From Proposition 29:

$$\begin{align} \|\phi_h(s,a)\|_{(\Lambda^k_h)^{-1}} &\le \sqrt{x^\mathsf{T}(\Lambda^k_h)^{-1}x} \tag{80}\\ &\le \sqrt{x^\mathsf{T}(B^k_h)^{-1}x} \tag{81}\\ &\le \frac{k+\lambda-f(k)}{k\lambda^+_h+\lambda-f(k)}\frac{1}{\sqrt{x^\mathsf{T}B^k_h x}}. \tag{82}\end{align}$$

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 30:

$$\begin{align} x^\mathsf{T}B^k_h x &= kx^\mathsf{T}\Lambda^\star_h x + \lambda - f(k) \tag{83}\\ &\ge k \min_{y\in\mathrm{Im}(\Lambda^\star_h),\|y\|=1} y^\mathsf{T}\Lambda^\star_h y + \lambda - f(k) \tag{84}\\ &= k\lambda^+_h + \lambda - f(k). \tag{85}\end{align}$$

Since $\beta_k = \widetilde{O}(1)$ and $f(k) = \widetilde{O}(\sqrt{k})$, from (82) and (85):

$$\beta_k \left\|\phi_h(s,a)\right\|_{(\Lambda_h^k)^{-1}} \le \beta_k \frac{k+\lambda-f(k)}{(k\lambda_h^+ + \lambda - f(k))^{3/2}} = \widetilde{O}(k^{-1/2}). \tag{86}$$

□

Lemma 32. *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 $v \notin \mathrm{span}\{\phi_j : j \in [n]\}$. Then, there exists a scalar $\epsilon > 0$ such that, for any $t \ge 0, \eta > 0$,*

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

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 $\mathrm{span}\{u_i : i \in [d]\} = \mathrm{span}\{\phi_j : j \in [n]\} \subset \mathbb{R}^d$. Then, Lemma 28 of (Papini et al., 2021) ensures that there exists a scalar $\epsilon > 0$ such that $|v^T u_i| \ge \epsilon$ 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

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

which concludes the proof. □

G. Example MDP

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

$$r_1(s,a) = 1 \qquad \text{for all } s \in \mathcal{S} \text{ and } a \in \mathcal{A}, \tag{87}$$

$$p_1(s_1|s_1,a_1) = 1, \qquad p_1(s_1|s_1,a_2) = \frac{1}{2}, \qquad p_1(s_1|s_2,a_1) = \frac{1}{2}, \qquad p_1(s_1|s_2,a_2) = \frac{3}{4}, \tag{88}$$

$$r_2(s_1,a_1) = 1, \qquad r_2(s_1,a_2) = \frac{7}{8}, \qquad r_2(s_2,a_1) = \frac{1}{2}, \qquad r_2(s_2,a_2) = \frac{5}{8}, \tag{89}$$

$\mu(s_1) = \mu(s_2) = 1/2$, 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:

$$\pi_1^\star(s_1) = a_1, \qquad \pi_1^\star(s_2) = a_2, \qquad \pi_2^\star(s_1) = a_1, \qquad \pi_2^\star(s_2) = a_2, \tag{90}$$

with the following values:

$$V_1^\star(s_1) = 2, \qquad V_1^\star(s_2) = \frac{61}{32}, \qquad V_2^\star(s_1) = 1, \qquad V_2^\star(s_2) = \frac{5}{8}. \tag{91}$$

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 $\Phi^{(1)}$:

$$\underline{\phi_1^{(1)}(s_1,a_1)} = \begin{bmatrix}1\\0\end{bmatrix} \qquad \phi_1^{(1)}(s_1,a_2) = \begin{bmatrix}1/2\\1/2\end{bmatrix} \qquad \phi_1^{(1)}(s_2,a_1) = \begin{bmatrix}1/2\\1/2\end{bmatrix} \qquad \underline{\phi_1^{(1)}(s_2,a_2)} = \begin{bmatrix}3/4\\1/4\end{bmatrix} \tag{92}$$

$$\underline{\phi_2^{(1)}(s_1,a_1)} = \begin{bmatrix}0\\1\end{bmatrix} \qquad \phi_2^{(1)}(s_1,a_2) = \begin{bmatrix}1/4\\3/4\end{bmatrix} \qquad \phi_2^{(1)}(s_2,a_1) = \begin{bmatrix}1\\0\end{bmatrix} \qquad \underline{\phi_2^{(1)}(s_2,a_2)} = \begin{bmatrix}3/4\\1/4\end{bmatrix}. \tag{93}$$

It is easy to check that the MDP is low-rank (Asm 2) and $\Phi^{(1)}$ is a realizable representation with $\boldsymbol{\theta}_1 = [1,1]^\mathsf{T}$, $\boldsymbol{\mu}_1(s_1) = [1,0]^\mathsf{T}$, $\boldsymbol{\mu}_1(s_2) = [0,1]^\mathsf{T}$, and $\boldsymbol{\theta}_2 = [1/2,1]^\mathsf{T}$. 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[7], so $\Phi^{(1)}$ is UNISOFT. The optimal covariance matrices are:

$$\Lambda_{1,\star}^{(1)} = \frac{1}{32}\begin{bmatrix}25 & 3\\ 3 & 1\end{bmatrix}, \qquad \Lambda_{2,\star}^{(1)} = \frac{1}{128}\begin{bmatrix}9 & 3\\ 3 & 113\end{bmatrix}. \tag{94}$$

Both are full rank, and their minimum eigenvalues are:

$$\lambda_{1,+}^{(1)} = \frac{13-3\sqrt{17}}{32} \simeq 0.02, \qquad \lambda_{2,+}^{(1)} = \frac{61-\sqrt{2713}}{128} \simeq 0.07. \tag{95}$$

As shown in Theorems 6 and 7, 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[8] of $\Phi^{(1)}$ to obtain an equivalent representation $\Phi^{(2)}$:

$$\underline{\phi_2^{(2)}(s_1,a_1)} = \begin{bmatrix}30/89\\ 74/89\end{bmatrix} \qquad \phi_2^{(2)}(s_1,a_2) = \begin{bmatrix}1/4\\ 3/4\end{bmatrix} \qquad \phi_2^{(2)}(s_2,a_1) = \begin{bmatrix}1\\ 0\end{bmatrix} \qquad \underline{\phi_2^{(2)}(s_2,a_2)} = \begin{bmatrix}75/356\\ 185/356\end{bmatrix}, \tag{96}$$

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.[9] 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:

$$\phi_{2,\star}^{(2)}(s_2) = \frac{5}{8}\phi_{2,\star}^{(2)}(s_1), \tag{97}$$

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.

[7] 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.

[8] Modifying the first stage would require a feature transformation that preserves realizability of both rewards and transitions. We are not aware of a general procedure to do so.

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