Reinforcement Learning in Linear MDPs: Constant Regret and Representation Selection

Anonymous Author(s)
Affiliation
Address
email

Abstract

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

The ability of an agent to learn an informative mapping from complex observations to a succinct representation is one of the essential factors for the success of machine learning in fields such as computer vision, language modeling, and more broadly in deep learning [Bengio et al., 2013].

In supervised learning, it is well understood that a "good" representation is one that allows to accurately fit any target function of interest (e.g., correctly classify a set of objects in an image). We refer to such case, as a *realizable* representation.[1] 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, Roy and Dong, 2019, Lattimore et al., 2020, Hao et al., 2021]. For instance, a representation accurately approximating the optimal value function may not be good for actually learning it by, e.g., approximate value iteration. In fact, it may poorly approximate any of the intermediate value functions, leading to compounding errors. 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.[2]

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

[1] 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".

[2] As for "realizable" representation, we use "learnable" representation as a short-hand for a representation such that finding a near-optimal policy is a learnable problem.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

tabular Markov Decision Processes (MDPs) [e.g., Jaksch et al., 2010, Azar et al., 2012, 2017], low or zero inherent Bellman error [e.g., Jin et al., 2020, Zanette et al., 2020b,a, Jin et al., 2021] and linear mixture MDPs [e.g., Yang and Wang, 2019, Ayoub et al., 2020, Zhang et al., 2021]. 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. In this scenario, research has focused either on the problem of online representation selection for regret minimization [e.g., Ortner et al., 2014, 2019, Lee et al., 2021] or, only recently, on the sample complexity of online representation learning [e.g., 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. While it is well-known that properties such as dimensionality and scaling have an impact on the learning performance, Hao et al. [2020], Papini et al. [2021] proved that it is possible to achieve constant regret (i.e., not scaling with the number of learning steps) if a certain (necessary and sufficient) condition on the features associated with the optimal actions is satisfied. To the best of our knowledge, the impact of similar properties on RL algorithms is largely unexplored.

Contributions. In this paper, we investigate the concept of "good" representation in the context of regret minimization in finite-horizon problems with linear structure, under realizability and learnability assumptions. Concretely, we consider the settings of zero inherent Bellman error [Zanette et al., 2020b] and low-rank structure [e.g., Jin et al., 2020]. Similar to the bandit case [Hao et al., 2020, Papini et al., 2021], we study the impact of representations on the learning process. Our contributions are both fundamentals and algorithmic. **1)** We provide a necessary condition for a "good" representation to enable constant regret in any problem with linear reward parametrization. Notably, this result encompasses MDPs with zero inherent Bellman error, and linear mixture MDPs with linearly parametrized rewards. Intuitively, the condition generalizes a similar condition for linear contextual bandits and it requires that the features observed along trajectories generated by the optimal actions provide information on the whole feature space. **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 selection in low-rank MDPs. We prove that in low-rank MDPs, LSVI-LEADER suffers the regret of the best representation without knowing it in advance. Furthermore, LSVI-LEADER is able to combine representations and achieve constant regret even when none of the individual representations would, but the "good" representation condition is satisfied by a combination of them. This is indeed possible thanks to its ability to select a different representation for each stage, state, and action.

2 Preliminaries

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

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

where the expectation is over probability measures induced by the policy and the MDP over state-action sequences of length $H - h$. Under certain regularity conditions [e.g., Bertsekas and Shreve, 2004], there always exists an optimal policy $\pi^\star$ whose value functions are defined by $V_h^{\pi^\star}(s) := V_h^\star(s) = \sup_\pi V_h^\pi(s)$ and $Q_h^{\pi^\star}(s,a) := Q_h^\star(s,a) = \sup_\pi Q_h^\pi(s,a)$. The optimal Bellman equation (and Bellman operator L_h) at stage $h \in [H]$ is defined as:

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

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

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

Linear Representation. When the state space is large or continuous, value functions are often described through a parametric representation. A standard approach is to use linear representations of the state-action function $Q_h(s,a) = \phi_h(s,a)^\mathsf{T}\theta_h$, where $\phi_h : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$ is a time-inhomogeneous feature map and $\theta_h \in \mathbb{R}^d$ is an unknown parameter vector. This reduces the complexity of the problem to d, potentially paying for an approximation error.[3]

In this paper, we consider MDPs satisfying Bellman closure (i.e., zero Inherent Bellman Error) [Zanette et al., 2020b] or low-rank assumptions [e.g., Yang and Wang, 2019, Jin et al., 2020].

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

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

This definition implies that the optimal value function is realizable as $Q^\star_h \in \mathcal{Q}_h$. Furthermore, the function space $\mathcal{Q}$ is closed under the Bellman operator, i.e., for all $Q_{h+1} \in \mathcal{Q}_{h+1}$, $L_h Q_{h+1} \in \mathcal{Q}_h$. This enables learnability for value-iteration-based algorithms [Munos and Szepesvári, 2008]. In the context of regret minimization, Zanette et al. [2020b] proposed a model-free algorithm, called ELEANOR, that achieves sublinear regret under the Bellman closure assumption, but at the cost of computational intractability.[4] The design of a tractable algorithm for regret minimization under low IBE assumption is still an open question in the literature.

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

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

where $\mu_h : \mathcal{S} \to \mathbb{R}^d$*. Then, for any policy* $\pi \in \Pi$*,* $\exists \theta^\pi_h \in \Theta_h$ *such that* $Q^\pi_h(s,a) = \phi_h(s,a)^\mathsf{T}\theta^\pi_h$*. 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. The minimum positive gap provides a natural measure of the difficulty of an MDP.

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

$$\Delta_h(s,a) = V^\star_h(s) - Q^\star_h(s,a). \tag{1}$$

We assume the minimum positive gap $\Delta_{\min} = \min_{s,a,h}\{\Delta_h(s,a) | \Delta_h(s,a) > 0\}$ *is well defined.*

In Tab. 1, we summarize existing bounds in the two settings.

Another structural assumption that has gained popularity in the literature is the linear-mixture structure [Jia et al., 2020, Ayoub et al., 2020, Zhou et al., 2020]. Linear mixture MDPs admit a linear representation of the transition function of the form $p_h(s'|s,a) = \phi_h(s'|s,a)^\mathsf{T}\theta_h$. No structural requirement is made on the reward, which is typically assumed to be known. As a consequence, the

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

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

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

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

value function may not be linearly representable. However, the fact the reward is known and that it is possible to directly learn the parameters θ_h of the transition function allow to achieve sublinear regret (even logarithmic) through model-based algorithms. While in this paper we mostly focus on Asm. 1 and 2, in Sect. 3.1 we show that our condition is necessary for constant regret also for linear mixture MDPs with unknown linear reward.

3 Constant Regret for Linear MDPs

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

Assumption 4. *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:*

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

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

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

It is interesting to look into Asm. 4 from an alternative perspective. Denote by $0 \le \lambda_{h,1} \le \ldots \le \lambda_{h,d}$ the eigenvalues of the matrix $\Lambda_h := \mathbb{E}_{s\sim\rho_h^\star}\left[\phi_h^\star(s)\phi_h^\star(s)^\mathsf{T}\right]$ and by $\lambda_h^+ = \min\{\lambda_{h,i} > 0, i \in [d]\}$ the minimum positive eigenvalue. We notice that when the features are non-redundant (i.e., $\{\phi_h(s,a)\}$ spans $\mathbb{R}^d$) and the UNISOFT assumption holds, then $\lambda_h^+ = \lambda_{h,1} > 0$. As we will see, the minimum positive eigenvalue λ_h^+ plays a fundamental role in the constant regret bound, together with the minimum gap $\Delta_{\min}$.

3.1 UNISOFT is Necessary for Constant Regret

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

Theorem 5. *Let M be any MDP with finite states, arbitrary dynamics p, linear rewards (i.e., $r_h(s,a) = \phi_h(s,a)^\mathsf{T}\theta_h$) with Gaussian $\mathcal{N}(0,1)$ noise, unique optimal policy $\pi^\star$, and where 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, linear-mixture MDPs with unknown linear rewards, and MDPs with Bellman closure.

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

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

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

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. B) leverages techniques for deriving asymptotic lower bounds for linear contextual bandits [e.g., Lattimore and Szepesvari, 2017, Hao et al., 2020, Tirinzoni et al., 2020].

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

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

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

We now proceed by contradiction: suppose that A suffers constant expected regret on M even though the MDP does not satisfy the UNISOFT condition. Then, since A plays sub-optimal actions only a finite number of times, it is possible to show that, for each $h \in [H]$, there exists a positive constant $\lambda_M > 0$ such that $\mathbb{E}_M^{\mathsf{A}}[\Lambda_h^K] \preceq \Lambda_h^\star + \lambda_M I$, where $\Lambda_h^\star := K\sum_{s:\rho_h^\star(s)>0} \phi_h^\star(s)\phi_h^\star(s)^\mathsf{T}$. Furthermore, since UNISOFT does not hold, there exists a stage $h \in [H]$ and a sub-optimal policy π (i.e., with $\Delta(\pi) > 0$) such that the vector $\Psi_h^\pi - \Psi_h^\star$ does not belong to $\operatorname{span}\{\phi_h^\star(s) | \rho_h^\star(s) > 0\}$. Then, since such space is exactly the one spanned by all the eigenvectors of $\Lambda_h^\star$ associated with a non-zero eigenvalue, there exists a positive constant $\epsilon > 0$ (independent of K) such that $\|\Psi_h^\pi - \Psi_h^\star\|^2_{(\Lambda_h^\star + \lambda_M I)^{-1}} \geq \epsilon^2/\lambda_M$. Combining these steps with Lem. 7, we obtain

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

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

3.2 UNISOFT is Sufficient for Constant Regret

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

Theorem 8. *Consider an MDP and a representation $\{\phi_h\}_{h\in[H]}$ satisfying the Bellman closure (Asm. 1) and* UNISOFT *assumptions (Asm. 4). If $\Delta_{\min} > 0$ (Asm. 3), then with probability at least $1-3\delta$, ELEANOR[5] suffers a* constant *regret*

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

where $\overline{\tau} = H\overline{\kappa}$ and $\overline{\kappa} = \max_h\{\kappa_h\}$ is the last episode ELEANOR suffers a non-zero regret (see Eq. 5 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 9. *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. 8.

We can see that the regret bounds in Thm. 8 and Thm. 9 reflect the shape of the worst-case and logarithmic regret used in the derivation. The reason becomes clear when looking at the proof sketch.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The final bounds are then obtained by instantiating the specific values of β_k and $g(k)$ for the two algorithms we analyzed.

4 Representation Selection in Low-Rank MDPs

In Sec. 3, we have highlighted the benefits that a UNISOFT representation brings to optimistic algorithms in MDPs with Bellman closure and low rank structure. In this section, we take one step further and investigate the *representation selection* problem. Since ELEANOR is a computationally 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 $\Phi_j = \left\{\phi_h^{(j)}\right\}_{h\in[H]}$, 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 each stage $h \in [H]$ of episode $k \in [K]$, LSVI-LEADER solves N different regression problems to compute an optimistic value function for each representation. Then, the final estimate $\overline{Q}_h^k(s,a)$ is taken as the *minimum* across these different optimistic value functions. Notably, this implies that LSVI-LEADER implicitly *combines* representations, in the sense that the selected representations (i.e., those with tightest optimism) might vary for different stages. This is exploited in the following result, which shows that constant regret is achievable even if none of the given representations is globally UNISOFT.

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**

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

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

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

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

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

Theorem 10. *Given an MDP M and a set of representations $\{\Phi_j\}_{j\in[N]}$ satisfying the low-rank assumption (Asm. 2), let $\mathcal{Z}$ be the set of H^N representations obtained by combining those in $\{\Phi_j\}_{j\in[N]}$ across different stages.*[6] *Then, with probability at least $1-2\delta$, LSVI-LEADER suffers at most a regret*

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

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

This result shows that LSVI-LEADER adapts to the *best* representation automatically, i.e., without any prior knowledge about the properties of the representations. In particular, it shows a problem-dependent (or worst-case) bound when there is no UniSOFT representation, while it attains constant regret when a representation, potentially mixed through stages, is UniSOFT. This is similar to what was obtained by Papini et al. [2021] for linear contextual bandits. Indeed, LSVI-LEADER reduces to their algorithm in the case $H = 1$. While the cost of representation selection is only logarithmic in linear bandits, the cost becomes polynomial (i.e., $\sqrt{N}$) in 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].

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

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

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

The generality of this result allows us to easily derive, for any fixed representation $z \in \mathcal{Z}$, both the worst-case regret bound of Jin et al. [2020] and the problem-dependent one of He et al. [2020]. To see this, note that the regret decompositions in both of these two papers rely on an upper bound to $\overline{V}_h^k(s_h^k) - V_h^{\pi_k}(s_h^k)$ as a function of the *fixed* representation used by LSVI-UCB (see the proof of

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

Theorem 3.1 of Jin et al. [2020] and Lemma 6.2 of He et al. [2020]). Then, fix any $z \in \mathcal{Z}$ and call z_h its features at stage h. Note that $z_h \in \{\phi_h^{(j)}\}_{j\in[M]}$. Moreover, by definition of low-rank structure, since each Φ_j induces a low-rank MDP, their combination does too. Thus, z is learnable. Then, instantiating the concentration bound stated above for policy π^k, state s_h^k, action a_h^k, stage h, and by upper bounding the minimum with the representation selected in z_h, we get

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

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

4.1 Representation Selection under a Mixing Condition

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

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

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

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

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

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

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

5 Conclusions

We investigated the properties that make a representation efficient for online learning in MDPs with Bellman closure. We introduced UniSOFT, a necessary and sufficient condition to achieve a constant regret bound in this class of MDPs. We demonstrate that existing optimistic algorithms are able to adapt to the structure of the problem and achieve constant regret. Furthermore, we introduce an algorithm able to achieve constant regret by mixing representations across states, actions and stages in the case of low-rank MDPs.

An interesting direction raised by our paper is whether it is possible to leverage the UniSOFT structure for probably-efficient representation learning, rather than selection. Another direction can be to leverage these insights to drive the design of auxiliary losses for representation learning, for example in deep RL.

References

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

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

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

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

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

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

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

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

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

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

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

Botao Hao, Tor Lattimore, Csaba Szepesvári, and Mengdi Wang. Online sparse reinforcement learning. In *AISTATS*, volume 130 of *Proceedings of Machine Learning Research*, pages 316–324. PMLR, 2021.

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

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

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

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

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

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

Tor Lattimore, Csaba Szepesvári, and Gellért Weisz. Learning with good feature representations in bandits and in RL with a generative model. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 5662–5670. PMLR, 2020.

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

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

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

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

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

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

Benjamin Van Roy and Shi Dong. Comments on the du-kakade-wang-yang lower bounds. *CoRR*, abs/1911.07910, 2019.

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

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

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

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

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

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

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

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

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

Checklist

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