On Constant Regret for Low-Rank MDPs

Alexander Sturm[1] **Sebastian Tschiatschek**[2]

[1]Faculty of Computer Science, UniVie Doctoral School Computer Science, University of Vienna, Vienna, Austria
[2]Faculty of Computer Science, Research Network Data Science, University of Vienna, Vienna, Austria

Abstract

Although there exist instance-dependent regret bounds for linear Markov decision processes (MDPs) and low-rank bandits, extensions to low-rank MDPs remain unexplored. In this work, we close this gap and provide regret bounds for low-rank MDPs in an instance-dependent setting. Specifically, we introduce an algorithm, called UNISREP-UCB, which utilizes a constrained optimization objective to learn features with good spectral properties. Furthermore, we demonstrate that our algorithm enjoys constant regret if the minimal sub-optimality gap and the occupancy distribution of the optimal policy are well-defined and known. To the best of our knowledge, these are the first instance-dependent regret results for low-rank MDPs.

1 INTRODUCTION

The design of algorithms for RL problems involving large state spaces has been of great interest in recent years. As traditional tabular methods are intractable in this setting, algorithms that use function approximation to generalize across states have gained substantial attention. In particular, non-linear function approximation has demonstrated strong empirical successes [He et al., 2024, Zhang et al., 2022] with provably efficient algorithms emerging [Agarwal et al., 2020, Uehara et al., 2022, Modi et al., 2024].

Furthermore, in many RL applications, there is a common expectation that a good RL algorithm will eventually gain enough information to identify optimal behavior in finite time [Zhang et al., 2024]. In that regard, a key question is under which assumptions this expectation can be confirmed theoretically.

Recently, Jin et al. [2020] have shown that sample-efficient learning in large state-action spaces is possible in linear Markov decision processes (MDPs), where the transition operator $\mathcal{P}$ admits a low-rank decomposition $\mathcal{P}(s'|s,a) = \langle \phi(s,a), \mu(s') \rangle$ into (known) features ϕ and (unknown) signed measures μ. In this setting, Papini et al. [2021a] showed that features that fulfill a spectral property called UniSOFT (see Definition 3.2) are necessary and sufficient for constant instance-dependent regret, i.e., the regret does not scale with the number of iterations.

Similarly, in contextual linear bandits (CLB), where the reward function is linear in the features ϕ, Papini et al. [2021b] showed that a diversity condition called HLS [Hao et al., 2020], is necessary and sufficient for constant instance-dependent regret. Tirinzoni et al. [2022] were able to provide an algorithm that achieves constant instance-dependent regret for CLBs, even when the true features ϕ are unknown and must be learned over some (known) finite function class.

To the best of our knowledge, there exists neither an instance-dependent result nor an algorithm that achieves constant regret for low-rank MDPs [Agarwal et al., 2020]; that is, linear MDPs with unknown features ϕ.

In this work, we study low-rank MDPs and aim to close this gap by addressing the following research question.

Can we achieve constant instance-dependent regret in low-rank MDPs?

As we shall see, we can answer this question positively. In particular, we provide an instance-dependent analysis of our proposed algorithm UNISREP-UCB, which is an augmented version of the recently proposed REP-UCB algorithm [Uehara et al., 2022] that serves as the basis for many other works [Zhang et al., 2022a, Agarwal et al., 2023, Zhao et al., 2024] on low-rank MDPs. In our analysis, we leverage the insights of Cheng et al. [2023], who designed a UCB-style bonus term that serves as a trajectory-wise uncertainty measure. In particular, we show that the bonus term serves as an almost optimistic estimate of the average sub-optimality gaps. This allows us to perform an instance-

Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI 2025), PMLR 244:4044–4079.

dependent regret analysis, similar to Papini et al. [2021a], employing UniSOFT feature maps. More specifically, we contribute the following:

- We provide an algorithm called UNISREP-UCB (Algorithm 1) that, for T large enough, achieves $\tilde{O}(\sqrt{T})$ expected regret (Theorem 4.2) provided that the minimal sub-optimality gap (Definition 3.3) and the minimal optimal occupancy (Definition 3.4) are well-defined and we have access to an expressive enough function space (Assumption 4.1);
- We design a termination criterion that allows UNISREP-UCB to achieve constant regret (Theorem 4.3), provided that the minimal sub-optimality gap and the minimal optimal occupancy are known;
- We demonstrate that the existence of UniSOFT representations is fully characterized by the RL instance (Lemma 5.1). In particular, we show that in low-rank MDPs, feature space coverage is equivalent to state space coverage—a result which can be of interest on its own.

2 RELATED WORK

Linear MDPs Jin et al. [2020] proposed the first sample-efficient algorithm for linear MDPs without assuming access to a generative model or other restrictive assumptions on the transition operator. Their algorithm LSVI-UCB combines classical LSVI with UCB-style bonuses and achieves $\tilde{O}(\sqrt{T})$ worst-case regret. Later, He et al. [2021] provided the first instance-dependent regret analysis for linear MDPs, achieving a logarithmic $O(\Delta_{\text{min}}^{-1}\log(T))$ instance-dependent regret bound. Using features that satisfy a diversity condition, called UniSOFT (Definition 3.2), Papini et al. [2021a] showed that LSVI-UCB enjoys constant instance-dependent regret. In addition, they demonstrate that the UniSOFT property is necessary for constant expected regret, reinforcing the importance of good features. Using a similar diversity condition, in bilinear MDPs, Zhang et al. [2023] provided an algorithm that enjoys constant instance-dependent regret. However, both methods do not scale to large function classes or misspecified representations. Recently, Zhang et al. [2024] were able to provide an algorithm that achieves constant regret without prior assumption on the features. Remarkably, their result holds even if features have low point-wise misspecification w.r.t. the minimal sub-optimality gap.

Low-Rank MDPs In the much more challenging low-rank MDP setting, the seminal work of Agarwal et al. [2020] provided the first reward-free oracle-efficient algorithm called FLAMBE. They proposed learning representations using maximum likelihood estimation (MLE) and showed that their explore-then-commit style algorithm achieves polynomial sample complexity when provided with an MLE oracle. By interleaving representation learning, exploration, and exploitation, Uehara et al. [2022] provided an algorithm called REP-UCB that improves the sample complexity bound of FLAMBE in every relevant variable under the same MLE oracle assumptions. In particular, they employ an UCB-style bonus term, which provides optimism at the initial state distribution. Recently, Cheng et al. [2023] showed that this bonus term can also serve as a trajectory-wise uncertainty measure. They leverage this insight to design a value function that encourages exploration in the state-action space where the uncertainty in the model estimation error is large and subsequently, provide an improved sample complexity bound. Finally, Zhao et al. [2024] provided the first regret bound for low-rank MDPs, employing a double exploration strategy. However, to the best of our knowledge, in contrast to linear MDPs, there exists no instance-dependent regret bound for low-rank MDPs. Furthermore, under which conditions, constant regret is achievable is still an open problem.

Contextual Linear Bandits In contextual linear bandits (CLB), Papini et al. [2021b] showed that a diversity condition called HLS [Hao et al., 2020], similar to the UniSOFT property, is necessary and sufficient for constant instance-dependent regret. Relaxing the assumption of exact feature maps, Tirinzoni et al. [2022] provided an algorithm which achieves constant regret, introducing a constrained optimization objective which encourages the HLS property and enforces the representations to be exact.

3 PRELIMINARIES

Notation We denote the set of probability distributions on a measurable set A by $\Delta(A)$. Furthermore, let $\mathcal{U}(A)$ represent the uniform distribution over some finite set A and let $\text{Ber}(p)$ denote the Bernoulli distribution with success rate $p \in [0,1]$. Additionally, $[N] \coloneqq \{1, ..., N\}$ for any integer N. Finally, $\lesssim$ denotes inequalities up to absolute constants and $\tilde{O}(\cdot)$ hides absolute constants and poly-log terms.

We consider a finite-horizon episodic MDP described by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}^\star, r^\star, H, d_1)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the finite action space, $\mathcal{P}^\star = \{\mathcal{P}_h^\star\}_{h\in[H]}$ where $\mathcal{P}_h^\star \colon \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition operator (unknown) at time step $h \in [H]$, $r^\star = \{r_h^\star\}_{h\in[H]}$ where $r_h^\star \colon \mathcal{S} \times \mathcal{A} \to [0,1]$ is the deterministic reward function (known) at time step $h \in [H]$, $d_1 \in \Delta(\mathcal{S})$ the initial state distribution (known) and H is the episode length. We assume the reward function to be normalized, i.e., $\sum_{h=1}^H \sup_{s,a} r_h^\star(s,a) \leq 1$.

The agent interacts with MDP $\mathcal{M}$ in episodes. In particular, in each episode $t \in \mathbb{N}$, the agent starts in some initial state $s_1 \sim d_1$, for each time step $h \in [H]$ observes a state s_h, chooses some action $a_h \in \mathcal{A}$, receives a reward $r_h^\star(s_h, a_h)$ and transitions to a new state $s_{h+1} \sim \mathcal{P}_h^\star(\cdot|s_h, a_h)$. The interaction process in each episode ends at time step $H+1$.

By $\Pi = \{\pi = \{\pi_h\}_{h\in[H]} \mid \forall h \in [H] : \pi_h\colon \mathcal{S} \to \mathcal{A}\}$ we denote the policy space in which the elements are (deterministic[1]) decision rules that map states to actions for any time step h. We define the state value function $V^{\pi}_{\mathcal{P},r;h}(s) = \mathbb{E}[\sum_{i=h}^{H} r_i(s_i, a_i)|s_h = s, \mathcal{P}, \pi]$ to represent the expected total reward of policy π under $\mathcal{P}$ and r starting in state s at time step h. To simplify notation, we define the function $\mathcal{P}_h V^{\pi}_{\mathcal{P},r,h+1}(s,a) = \mathbb{E}_{s'\sim\mathcal{P}_h(\cdot|s,a)}[V^{\pi}_{\mathcal{P},r,h+1}(s')]$, where $\mathcal{P}_h$ should be viewed as an operator on functions $f\colon \mathcal{S} \to \mathbb{R}$ with $f \mapsto \mathcal{P}_h f$.

We define the Q-function as $Q^{\pi}_{\mathcal{P},r;h}(s,a) = r_h(s,a) + \mathcal{P}_h V^{\pi}_{\mathcal{P},r;h+1}(s,a)$ and let $V^{\pi;d_1}_{\mathcal{P},r;1} = \mathbb{E}_{s\sim d_1}[V^{\pi}_{\mathcal{P},r;1}(s)]$, given some initial state distribution d_1. The state-action occupancy distribution $d^{\pi}_{\mathcal{P};h}(s,a)$ denotes the probability of visiting state s at time step h and performing action a in model $\mathcal{P}$ with policy π. By abuse of notation, let $d^{\pi}_{\mathcal{P};h}(s) = \sum_{a\in\mathcal{A}} d^{\pi}_{\mathcal{P};h}(s,a)$ denote the state-occupancy distribution at time step h. We can sample a state s from $d^{\pi}_{\mathcal{P};h}$ by executing π for $h-1$ steps starting from state $s_1 \sim d_1$.

The agent's goal is to learn an optimal policy $\pi^\star \in \arg\max_{\pi\in\Pi} V^{\pi,d_1}_{\mathcal{P}^\star,r^\star,1}$, which maximizes the expected total reward under $\mathcal{P}^\star$, $r^\star$ and d_1. We evaluate the efficiency of an agent by the (expected) regret

$$\mathbb{E}[\mathcal{R}(T)] = \mathbb{E}[\sum_{t=1}^{T} V^{\pi^\star,d_1}_{\mathcal{P}^\star,r^\star,1} - V^{\pi_t,d_1}_{\mathcal{P}^\star,r^\star,1}], \qquad (1)$$

which measures the expected cumulative performance loss up to episode $T \in \mathbb{N}$. Note that the expectation in Equation 1 is taken w.r.t. any extra randomness induced by the algorithm.

Finally, we denote the sub-optimality gap of taking action a in state s at time step h as $\Delta_h(s,a) = V^{\pi^\star}_{\mathcal{P}^\star,r^\star;h}(s) - Q^{\pi^\star}_{\mathcal{P}^\star,r^\star;h}(s,a)$, which measures the loss in value of any sub-optimal action a.

3.1 STRUCTURAL ASSUMPTIONS

In this work, we are interested in MDPs with large, possibly infinite state spaces and hence require some form of structural assumptions such that efficient learning is possible. In particular, we assume that $\mathcal{P}^\star$ admits a low-rank decomposition.

Definition 3.1. *(Low-rank MDP [Agarwal et al., 2020]) An MDP $\mathcal{M}$ is* low-rank *or equivalently has* low-rank structure *with rank $d \in \mathbb{N}$ if for every $h \in [H]$ there exist two embedding functions $\phi^\star_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and $\mu^\star_h : \mathcal{S} \to \mathbb{R}^d$ such that*

$$\forall (s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}\colon \mathcal{P}^\star_h(s'|s,a) = \langle \phi^\star_h(s,a), \mu^\star_h(s')\rangle,$$

where, for normalization, $\|\phi^\star_h(s,a)\|_2 \leq 1$ and $\|\int_{\mathcal{S}} \mu^\star_h(s)g(s)ds\|_2 \leq \sqrt{d}\|g\|_\infty$, for any function $g\colon \mathcal{S} \to \mathbb{R}$, $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$.

As the embedding functions $\phi^\star_h$ and $\mu^\star_h$ are assumed to be unknown, we consider the *representation learning problem* of finding good representations for state-action pairs and states over (known) finite function spaces $\Phi = \Phi_1 \times ... \times \Phi_H$ and $\Psi = \Psi_1 \times ... \times \Psi_H$ where, for each $h \in [H]$, $\Phi_h \subseteq \{\phi_h\colon \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d\}$ and $\Psi_h \subseteq \{\mu_h\colon \mathcal{S} \to \mathbb{R}^d\}$. For notational brevity, we denote $\phi^\star = \{\phi^\star_h\}_{h\in[H]}$ and $\mu^\star = \{\mu^\star_h\}_{h\in[H]}$. To ensure tractability of this representation learning problem, we assume realizability of the function spaces [Agarwal et al., 2020, Uehara et al., 2022, Modi et al., 2024].

Assumption 3.1. *(Realizability) For all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, and any $(\phi_h, \mu_h) \in \Phi_h \times \Psi_h$, we have that $\|\phi_h(s,a)\|_2 \leq 1$, for any function $g\colon \mathcal{S} \to \mathbb{R}$, $\|\int_{\mathcal{S}} \mu_h(s)g(s)ds\|_2 \leq \sqrt{d}\|g\|_\infty$ and $\int_{\mathcal{S}}\langle\phi_h(s,a), \mu_h(s')\rangle ds' = 1$. Additionally, there exist (unknown) non-empty subsets $\Phi^\star \subseteq \Phi$ and $\Psi^\star \subseteq \Psi$ such that any $(\phi^\star, \mu^\star) \in \Phi^\star \times \Psi^\star$ fulfills the low-rank MDP Definition 3.1.*

Note that any tuple $(\phi,\mu) \in \Phi\times\Psi$ naturally induces a distribution over the state space in each time step and, in particular, a transition operator $\mathcal{P} \equiv \langle\phi,\mu\rangle$.

3.2 GOOD REPRESENTATIONS AND INSTANCE-DEPENDENT PROPERTIES

In favor of clarity, the main results are presented under the assumption of a unique optimal policy. In Section F we show how this assumption can be dropped. Let us denote $\Pi^\star$ as the set of all optimal (deterministic) policies.

Assumption 3.2. *(Unique optimal policy) There exists a unique optimal (deterministic) policy; that is, $|\Pi^\star| = 1$.*

We consider a feature mapping $\phi \in \Phi$ as *good* if it maps the set of state-action pairs reachable by the optimal policy to a set of vectors that span the whole feature space. In particular, good representations are non-redundant and UniSOFT.

Definition 3.2. *(UniSOFT Representation [Papini et al., 2021a]) A feature mapping $\phi \in \Phi$ is called UniSOFT (Universally Spanning Optimal FeaTures) if for all $h \in [H]$,*

$$\begin{aligned}&\text{span}\{\phi_h(s,a)|\forall(s,a) : \exists\pi \in \Pi : d^{\pi}_{\mathcal{P}^\star;h}(s,a) > 0\}\\ &\quad= \text{span}\{\phi_h(s,\pi^\star(s))|\forall s : d^{\pi^\star}_{\mathcal{P}^\star;h}(s) > 0\}\end{aligned}$$

holds. In particular, a UniSOFT feature mapping ϕ is non-redundant *if $\lambda^\star(\phi) > 0$ holds, where*

$$\lambda^\star(\phi) := \min_{h\in[H]} \lambda_{\min}(\mathbb{E}_{(s,a)\sim d^{\pi^\star}_{\mathcal{P}^\star,h}}[\phi_h(s,a)\phi_h(s,a)^T])$$

and $\lambda_{\min}(\cdot)$ returns the minimal eigenvalue.

[1] We require that our planning procedure outputs (w.l.o.g.) a deterministic policy to ensure that the representation learning oracle converges (see Appendix B).

Intuitively, non-redundant UniSOFT features allow an algorithm to efficiently explore the whole feature space by behaving optimally in the environment. How efficiently the feature space can be explored is dependent on $\lambda^\star(\cdot)$, which, as we will see, will play a major role in the regret bounds provided in the next chapter. Furthermore, we will say that a transition operator $\mathcal{P}$ *admits* a non-redundant UniSOFT representation, whenever there exists a representation $\langle \phi, \mu \rangle \equiv \mathcal{P}$ such that ϕ is UniSOFT and non-redundant.

We introduce two additional assumptions that will allow us to take advantage of good representations and perform an instance-dependent regret analysis. A very natural measure of hardness is the minimal sub-optimality gap, which captures the difficulty in detecting sub-optimal actions.

Assumption 3.3. *(Well-defined minimal sub-optimality gap) The quantity*

$$\Delta_{\text{min}} := \min_{s \in \mathcal{S}, a \in \mathcal{A}, h \in [H] : \Delta_h(s,a) > 0} \Delta_h(s,a)$$

is well-defined.

Finally, we assume that the minimal optimal occupancy exists. Intuitively, we ensure that when playing an optimal decision policy, we will eventually visit all states reachable by this policy.

Assumption 3.4. *(Well-defined minimal optimal occupancy) The quantity*

$$d^\star_{\text{min}} = \min_{s \in \mathcal{S}, a \in \mathcal{A}, h \in [H], \pi^\star \in \Pi^\star : d^{\pi^\star}_{\mathcal{P}^\star, h}(s,a) > 0} d^{\pi^\star}_{\mathcal{P}^\star, h}(s,a)$$

is well-defined.

Note that both assumptions are trivially satisfied whenever $\mathcal{S}$ and $\mathcal{A}$ are finite.

4 INSTANCE-DEPENDENT REGRET BOUNDS

This section provides an algorithm, called UNISREP-UCB (Upper Confidence Driven Universally Spanning Representation Learning, Algorithm 1), that achieves sub-linear expected regret under an additional simplifying assumption that guarantees the selection of good representations. Furthermore, we demonstrate that by introducing a carefully chosen termination criterion to UNISREP-UCB, resulting in the algorithm UNISREP-UCB + (Algorithm 1 with modifications shown in blue), we can identify optimal behavior with high probability whenever the minimal sub-optimality gap and the minimal optimal occupancy are known.

Algorithm 1 UNISREP-UCB (+)

```
Input: Function spaces $\{\Phi_h\}_{h=1}^H, \{\Psi_h\}_{h=1}^H$, Parameters $\lambda_t, \hat{\alpha}_t, \xi_t$ decreasing, $T$
Output: $\pi_t$
 1: Initialize: $\mathcal{D}_{0,h} = \emptyset$, $\mathcal{D}'_{0,h} = \emptyset$, $\pi_{0,h} \equiv \mathcal{U}(\mathcal{A})$, $\forall h \in [H]$
 2: for $t = 1, ..., T$ do
        // Interact with the MDP and collect transition data
 3:     $e_t \sim \text{Ber}(1 - \xi_{t-1})$
 4:     for $h = 1, ..., H$ do
 5:         $s_{h-1} \sim d^{\pi_{t-1}}_{P^\star; h-1}$
 6:         if $e_t = 1$ then
 7:             $a_{h-1} = \pi_{t-1,h-1}(s_{h-1})$, $s_h \sim P^\star_{h-1}$
 8:             $a_h = \pi_{t-1,h}(s_h)$, $s_{h+1} \sim P^\star_h$
 9:         else
10:             $a_{h-1} \sim \mathcal{U}(\mathcal{A})$, $s_h \sim P^\star_{h-1}$,
11:             $a_h \sim \mathcal{U}(\mathcal{A})$, $s_{h+1} \sim P^\star_h$
12:         end if
13:         $\mathcal{D}_{t,h-1} = \mathcal{D}_{t-1,h-1} \cup \{(s_{h-1}, a_{h-1})\}$
14:         $\mathcal{D}'_{t,h} = \mathcal{D}'_{t-1,h} \cup \{(s_h, a_h, s_{h+1})\}$
15:     end for
        // Learn representations & set bonus
16:     for $h = 1, ..., H$ do
17:         $\hat{\phi}_{t,h} = \arg\min_{\phi \in \Phi_h^{\text{MLE}}(\mathcal{D}'_{t,h})} \mathcal{L}^{\text{unisoft}}(\phi, \mathcal{D}_{t,h})$
18:         $\hat{\Sigma}_{t,h} = \sum_{(s,a) \in \mathcal{D}_{t,h}} \hat{\phi}_{t,h}(s,a) \hat{\phi}_{t,h}(s,a)^T$
19:             $+\lambda_t I$
20:         $\hat{b}_{t,h}(s,a) =$
21:             $\min\{\hat{\alpha}_t \sqrt{\hat{\phi}_{t,h}(s,a)^T \hat{\Sigma}^{-1}_{t,h} \hat{\phi}_{t,h}(s,a)}, 1\}$
22:         $\hat{\mathcal{P}}_{t,h}(s'|s,a) = \langle \hat{\phi}_{t,h}(s,a), \hat{\mu}_{t,h}(s') \rangle$
23:     end for
        // Update (deterministic) policy
24:     $\pi_t = \arg\max_{\pi \in \Pi} V^{\pi, d_1}_{\hat{\mathcal{P}}_t, \hat{b}_t + r^\star, 1}$
        // Check for optimality
25:     $\pi_t^b = \arg\max_{\pi \in \Pi} V^{\pi, d_1}_{\hat{\mathcal{P}}_t, \hat{b}_t, 1}$
26:     $c_t = 10H^2(V^{\pi_t^b, d_1}_{\hat{\mathcal{P}}_t, \hat{b}_t, 1} + \sqrt{\frac{|\mathcal{A}|}{\xi_t} \zeta_t})$
27:     if $c_t < \Delta_{\text{min}} d^\star_{\text{min}}$ then
28:         return $\pi_t$
29:     end if
30: end for
31: return $\pi_t$
```

4.1 ALGORITHM

On a high level, UNISREP-UCB is a finite-horizon adaptation of the REP-UCB algorithm proposed by Uehara et al. [2022]. However, unlike the REP-UCB algorithm, we employ a double exploration scheme to enable a regret bound, as proposed by Zhao et al. [2024] and augment the representation learning objective to encourage feature maps with good spectral properties.

Exploration (Lines 3-15) For each time step h, the algorithm samples the state-occupancy distribution $d^{\pi_{t-1}}_{\mathcal{P}^\star, h-1}$ and continues based on the result of a Bernoulli experiment with a success rate of $1-\xi_t$. If successful, the algorithm explores with the behavior policy π_{t-1}, and otherwise it explores by taking actions uniformly at random. This mechanism is key for enabling a regret bound, as otherwise the algorithm would explore uniformly at random in each episode and time step, preventing sub-linear regret. After time step $h+1$ the algorithm rolls-out to time step H according to π_{t-1}. Note that we only require the algorithm to interact with the environment in full trajectories due to a technicality when bounding the regret. Qualitatively, the algorithm does not change by resetting after $h+1$ time steps. Finally, we collect the transitions of the time steps $h-1$ and h in separate datasets.

Representation Learning (Lines 16-23) Similarly to Tirinzoni et al. [2022], we employ a constrained optimization objective (Line 17), to learn features that have good spectral properties and approximate the transition operator well enough. We define the following objective functions:

$$\mathcal{L}^{\text{MLE}}(\phi_h, \mu_h, \mathcal{D}) = \sum_{(s,a,s')\in\mathcal{D}} \log(\langle \phi_h(s,a), \mu_h(s') \rangle \quad (2)$$

$$\mathcal{L}^{\text{UniSOFT}}(\phi_h, \mathcal{D}) = -\lambda_{\min}\left(\sum_{(s,a)\in\mathcal{D}} \phi_h(s,a)\phi_h(s,a)^T\right) \quad (3)$$

Then, the set of representations that are the maximum likelihood solution of fitting the transition operator over some dataset $\mathcal{D}$, are defined as follows:

$$\Phi_h^{\text{MLE}}(\mathcal{D}) = \{\phi \in \Phi_h : \max_{\mu\in\Psi_h} \mathcal{L}^{\text{MLE}}(\phi, \mu, \mathcal{D}) = \max_{(\phi',\mu')\in\Phi_h\times\Psi_h} \mathcal{L}^{\text{MLE}}(\phi', \mu', \mathcal{D})\}$$

Similarly to previous work on low-rank MDPs [Agarwal et al., 2020, Uehara et al., 2022, Cheng et al., 2023], as a computational abstraction, we assume access to an optimization oracle.

Definition 4.1. *(Optimization Oracle) Consider the function class $\Phi\times\Psi$ and datasets $\mathcal{D}$ and $\mathcal{D}'$ consisting of (s,a) tuples and (s,a,s') triples, respectively. Then, the* optimization oracle *returns for any $h \in [H]$,*

$$\arg\min_{\phi\in\Phi_h^{\text{MLE}}(\mathcal{D}')} \mathcal{L}^{\text{UniSOFT}}(\phi, \mathcal{D}).$$

Note that although the oracle is computationally intractable, it can be reasonably well approximated in practice [Tirinzoni et al., 2022, Zhang et al., 2022]. After employing the oracle, we use the learned features to define an UCB-style bonus term and the estimated transition operator.

Planning (Line 24) We find an optimal (deterministic) policy for the bonus-augmented reward function in the estimated environment. Here, we assume access to a planning procedure that returns, for any given reward function r and transition operator $\mathcal{P} = \langle \phi, \mu \rangle$, an optimal (deterministic) policy $\arg\max_{\pi\in\Pi} V^{\pi,d_1}_{\mathcal{P},r,1}$. We note that planning in a known linear MDP can be performed efficiently, for example, with LSVI-UCB [Jin et al., 2020].

4.2 ANALYSIS

In the following lemma, we provide a baseline worst-case regret bound for UNISREP-UCB, which does not utilize UniSOFT features. We denote the regret incurred by algorithm 1 as $\tilde{\mathcal{R}}$, which differs from the regret incurred by behavior polices $\{\pi_t\}_{t=1}^T$ denoted as $\mathcal{R}$.

Lemma 4.1 (Expected Regret without UniSOFT)**.** *Let $\xi_t = t^{-1/4}$. Suppose Assumption 3.1 (realizability) holds. Then, for any $T \in \mathbb{N}$,* UNISREP-UCB *(Algorithm 1) satisfies*

$$\mathbb{E}[\tilde{\mathcal{R}}(T)] = \tilde{O}\left(H^3 d^2 |\mathcal{A}| T^{3/4}\right).$$

Our general strategy for improving the baseline regret given above is to show that there exists an episode after which UNISREP-UCB only selects good representations. Then, these good representations provide more efficient exploration, and we gain an improvement in learning efficiency. Hence, our regret bounds will only improve on the baseline regret result if we run the algorithm for long enough. Furthermore, establishing sub-linear regret without leveraging good representations is important for guaranteeing the selection of good representations at a later stage.

Nevertheless, to select good representations, we must ensure their existence. In that spirit, we introduce representations that approximately represent the ground-truth transition operator over the support of the occupancy distribution induced by the optimal policy.

Definition 4.2. *($\alpha^\star$-Approximate Representation) A representation $(\phi,\mu) \in \Phi\times\Psi$, with induced model $\mathcal{P}$, is* $\alpha^\star$-approximate *at level α if for all $h \in [H]$,*

$$\mathbb{E}_{(s,a)\sim d^{\pi^\star}_{\mathcal{P}^\star,h}}[\|\mathcal{P}_h(\cdot|s,a) - \mathcal{P}^\star_h(\cdot|s,a)\|_{\text{TV}}] \le \alpha.$$

Remark 4.1. *The set of $\alpha^\star$-approximate representations $\Phi_\alpha \times \Psi_\alpha \subseteq \Phi\times\Psi$ is non-empty for any $\alpha \ge 0$, whenever the realizability assumption 3.1 holds.*

Interestingly, we can show that the optimization oracle (Definition 4.1) converges uniformly over the occupancy distribution of the optimal policy (Lemma B.1), provided that the distribution is well-defined, that is, Assumption 3.4 (minimal optimal occupancy) holds. The following assumption exploits this convergence and ensures that we are guaranteed to find a good representation. In Section 5 we elaborate on how reasonable this assumption is.

Assumption 4.1. *($\alpha^\star$-Expressive Function Space) For all $\alpha^\star$-approximate representations $(\phi, \mu) \in \Phi_\alpha \times \Psi_\alpha$, there exists a representation $(\tilde{\phi}, \tilde{\mu}) \in \Phi \times \Psi$ that is non-redundant and UniSOFT, such that the induced models $\mathcal{P}$ and $\tilde{\mathcal{P}}$ agree on all $(s, a) \in \mathcal{S} \times \mathcal{A}$.*

We can show that the UniSOFT loss in Equation 3 eventually eliminates all redundant and all non-UniSOFT feature maps (Lemma B.4). Intuitively, if the exploration probabilities ξ_t are decreasing and the regret of the behavior policies is sub-linear, the collected transitions will eventually mostly be drawn from the optimal occupancy distribution. Then only good features minimize the UniSOFT loss, which are guaranteed to exist by the expressiveness assumption above.

Whenever the function space already consists of representations that have low model error on the optimal occupancy distribution, we can provide a purely gap-dependent regret bound.

Theorem 4.1 (Gap-dependent regret with UniSOFT)**.** *Let $\xi_t = t^{-1/3}$ and $\alpha = 1$. Suppose assumptions 3.1 (realizability), 3.3 (minimal sub-optimality gap), 4.1 ($\alpha^\star$-expressive function space) and 3.2 (unique optimal policy) hold. Then for any $T \in \mathbb{N}$, there exists a constant τ_{good}, such that* UniSREP-UCB *(Algorithm 1) satisfies the following:*

$$\mathbb{E}[\tilde{\mathcal{R}}(T)] = \tilde{O}(H^3 d^2 |\mathcal{A}| (\tau_{\text{good}} \wedge T)^{5/6} + \frac{1}{\lambda^\star_{\text{max}}} H^4 d |\mathcal{A}|^{1/2} T^{2/3})$$

where $\tau_{\text{good}} = \tilde{O}\left(\frac{H^{12} d^{12} |\mathcal{A}|^6}{(\Delta_{\text{min}} \lambda^\star_{\text{max}})^6}\right)$ *and* $\lambda^\star_{\text{max}} = \min_\alpha \max_{\phi \in \Phi_\alpha} \lambda^\star(\phi)$.

On a high level, τ_{good} captures the number of episodes UniSREP-UCB needs to eliminate all non-good representations. Hence, the theorem tells us that after some number of "warm-up" episodes τ_{good}, during which we incur expected regret according to the parameter-adjusted baseline result (Lemma 4.1), we gain an increase in learning efficiency provided by the properties of good representations. The duration of the warm-up and the gain in learning efficiency depend on the "goodness" of the available representations, captured by $\lambda^\star$. Notable is the worse dependence on the horizon.

If we additionally assume that the minimal optimal occupancy is well-defined (Assumption 3.4), we can show that the behavior policies are eventually optimal. In particular, we show that the bonus term serves as an almost optimistic estimate of expected sub-optimality gaps (Lemma C.3). Hence, if we are guaranteed to select good representations in each iteration (Lemma B.4), the bonus term decreases uniformly over the state-action space, leading to optimal behavior. However, since we bound sub-optimality gaps in expectation, we require $d^\star_{\text{min}}$ to be well-defined, in order to determine the optimality of any policy (Lemma D.1). We get the following improved result.

Theorem 4.2 (Expected regret with UniSOFT)**.** *Let $\alpha > 0$, $\gamma \in (2, 4]$ and $\xi_t = t^{-1/\gamma}$. Suppose assumptions 3.1 (realizability), 3.2 (unique optimal policy), 3.3 (minimal sub-optimality gap), 3.4 (minimal optimal occupancy) and 4.1 ($\alpha^\star$-expressive function space) hold. Then for any $T \in \mathbb{N}$, there exists a constant $\tau^\star$ such that* UniSREP-UCB *(Algorithm 1) satisfies*

$$\mathbb{E}[\tilde{\mathcal{R}}(T)] = \tilde{O}\left(H^3 d^2 |\mathcal{A}| (\tau^\star \wedge T)^{1/2 + 1/\gamma} + H T^{\frac{\gamma - 1}{\gamma}}\right),$$

where $\tau^\star = \tilde{O}\left(\left(\frac{H^2 d^2 |\mathcal{A}|}{\alpha \lambda^\star_{\text{max}} (\Delta_{\text{min}} d^\star_{\text{min}})^2}\right)^{\frac{2\gamma}{\gamma - 2}}\right)$.

In contrast to τ_{good}, $\tau^\star$ additionally captures the number of episodes UniSREP-UCB needs to fully explore the feature space and subsequently identify the optimal policy. However, our algorithm still explores uniformly with positive probability, preventing constant regret. We also incur dependence in α and $d^\star_{\text{min}}$, capturing the difficulty of selecting $\alpha^\star$-approximate representations.

Interestingly, if we assume that the quantities Δ_{min} and $d^\star_{\text{min}}$ are known[2], we can design a termination criterion, which stops the algorithm whenever the behavior policy is optimal. UniSREP-UCB + extends UniSREP-UCB by an evaluation phase (Lines 25-29) in which we measure the uncertainty in the learned model, through the value of the bonus term. If this uncertainty is below $\Delta_{\text{min}} d^\star_{\text{min}}$, we stop the algorithm and return the optimal policy with high probability.

Theorem 4.3 (Constant Regret)**.** *Let $\alpha > 0$, $\delta \in (0, 1)$ and $\xi_t = t^{-1/4}$. Suppose that the quantities Δ_{min} and $d^\star_{\text{min}}$ are known. Then, under the same assumptions as in Theorem 4.2, with probability at least $1 - 2\delta$,* UniSREP-UCB + *(Algorithm 1) satisfies the following:*

$$\tilde{\mathcal{R}}(T) \leq T \wedge \tau^\star,$$

where[3] $\tau^\star = \tilde{O}\left(\frac{H^8 d^8 |\mathcal{A}|^4}{(\alpha \lambda^\star_{\text{max}})^4 (\Delta_{\text{min}} d^\star_{\text{min}})^8}\right)$.

4.3 TECHNICAL CHALLENGES

The main technical challenge to providing instance-dependent regret lies in controlling the expected sub-optimality gaps. In Lemma C.5, we demonstrate that the expected gaps can be controlled w.r.t. the value of the bonus under policy π^b. Unfortunately, this is not the policy that interacts with the environment, and hence the elliptical potential lemma does not work here. Importantly, UniSOFT

[2] Extensions to lower bounds on Δ_{min} and $d^\star_{\text{min}}$ are straightforward.

[3] $\tilde{\mathcal{O}}$ hides a constant of order 2^{64}.

features uniformly decrease the confidence intervals, which allows us to proceed with our analysis. As such, the role of representation learning and, in particular, that of UniSOFT features is central for our instance-dependent bounds.

5 MORE ON GOOD REPRESENTATIONS

Note that within this section, we assume finiteness of the state space ($|\mathcal{S}| < \infty$) and that the transition operator has rank $\tilde{d}$ for all time steps, that is, rank$(\mathcal{P}_h^\star) = \tilde{d}$ for all $h \in [H]$. Furthermore, we denote by $\mathcal{X}_h^\star := \{(s, a) \in \mathcal{S} \times \mathcal{A} | d^{\pi^\star}_{\mathcal{P}^\star, h}(s, a) > 0\}$ the set of state-action pairs reachable by the optimal policy at time step $h \in [H]$. The following lemma provides a condition that is necessary and sufficient for the existence of non-redundant UniSOFT representations in low-rank MDPs.

Lemma 5.1 (Existence of good representations)**.** *Let $d \geq \tilde{d}$. Then, the following statements are equivalent:*

(1) span$\{\mathcal{P}_h^\star(\cdot|s, a) | (s, a) \in \mathcal{X}_h^\star\} = \mathbb{R}^{\tilde{d}}$ *and* $|\mathcal{X}_h^\star| \geq d$*,*

(2) *there exists a non-redundant UniSOFT representation* $\langle \tilde{\phi}_h, \tilde{\mu}_h \rangle_{\mathbb{R}^d} = \mathcal{P}_h^\star$*.*

Remark 5.1. *Note that the result is agnostic to the choice of policy. This implies that in low-rank MDPs feature space coverage is equivalent to state space coverage.*

Remark 5.2. *In section E, we provide a similar result for the existence of (possibly redundant) UniSOFT features. This implies, given that UniSOFT feature maps are necessary for constant expected regret in MDPs with linear rewards [Papini et al., 2021a], that to achieve constant expected regret in linear MDPs or low-rank MDPs with unknown rewards, the optimal policy must visit all states reachable by any policy with positive probability.*

Importantly, we see that the existence of good features ϕ is fully characterized by the ground-truth transition operator. That is, assuming the existence of non-redundant UniSOFT features, implicitly assumes that the optimal policy explores the whole reachable state space (Corollary E.1).

Nevertheless, if $\mathcal{P}^\star$ admits a non-redundant UniSOFT representation, good $\alpha^\star$-approximate representations are abundant. The following lemma supports the $\alpha^\star$-expressiveness assumption (Assumption 4.1).

Lemma 5.2. *Assume that Assumption 3.4 (minimal optimal occupancy) holds and that $\mathcal{P}^\star$ admits a non-redundant UniSOFT representation. Then, there exists an $\epsilon > 0$ such that for any $d \geq \tilde{d}$ the following holds: Let $\tilde{\alpha} < \alpha \leq \epsilon$ be arbitrary. There exist infinitely more $\alpha^\star$-approximate representations than $\tilde{\alpha}^\star$-approximate representations $\langle \phi, \mu \rangle_{\mathbb{R}^d} \equiv \hat{\mathcal{P}}$ that are UniSOFT and non-redundant.*

Remark 5.3. *On a high level, ϵ is upper bounded by the degree of linear independence between the (unknown) transition vectors of the optimal actions.*

6 DISCUSSION

6.1 COMPARISON WITH THE LITERATURE

In this subsection, we compare the constant regret result of Theorem 4.3 with related results from the literature. In Table 1 we provide an overview of algorithms achieving constant regret in different learning settings and compare their critical episodes; that is, the episode after which, with high probability, the respective algorithm does not incur additional regret.

LSVI-LEADER [Papini et al., 2021a] In the linear MDP setting, the LSVI-LEADER algorithm proposed by Papini et al. [2021a] assumes access to a set of realizable representations containing one UniSOFT representation, and that the unique optimal policy assumption 3.2 holds. However, their algorithm does not scale to large function spaces, as it learns a different representation for each state-action pair.

In comparison, UNISREP-UCB + can deal with large function spaces and misspecified representations. Additionally, we show how to generalize our regret bounds beyond the unique optimal policy assumption. However, we assume access to an optimization oracle, positive minimal optimal occupancy and known instance-dependent quantities.

In Table 1 we can see that, in contrast to LSVI-LEADER, the critical episode of UNISREP-UCB + depends on the size of the action space, which seems to be unavoidable in low-rank MDPs [Zhao et al., 2024]. We additionally incur a dependence on $d^\star_{\text{min}}$, which stems from bounding average sub-optimality gaps and on α as we must select representations with low model error. The overall smaller polynomial dependence for LSVI-LEADER follows from the overall tighter regret bound available for linear MDPs.

BanditSRL [Tirinzoni et al., 2022] In contextual linear bandits (CLB) the feature map ϕ must only linearly represent the reward function. Similarly to our work, BanditSRL learns a non-redundant representation with good spectral properties over a known finite function space. They do not rely on any oracle assumptions, as estimating the reward function can be done efficiently by minimizing the MSE. However, they rely on a restrictive misspecification assumption that allows them to eliminate all point-wise misspecified representations. In particular, they assume that the following quantity is well-defined:

$$\epsilon_{\text{min}} := \min_{\phi \in \Phi \setminus \Phi^\star} \min_{\theta : \|\theta\| \leq 1} \min_{\pi : \mathcal{S} \to \mathcal{A}} \mathbb{E}_{s \sim d_1}[(\langle \phi(s, \pi(s)), \theta \rangle - r^\star(s, \pi(s))^2] > 0.$$

Although estimating the reward function is conceptually different, we emphasize that our algorithm can deal with misspecified representations without making additional assumptions on the level of misspecification.

Table 1: Comparison of critical episodes; for ease of comparison, constants that refer to eigenvalue sizes are summarized with $\lambda^\star$.

Algorithm	Setting	Features ϕ	Critical Episode
LEADER [Papini et al., 2021b]	CLB	Known	$\tilde{O}((\frac{d}{\lambda^\star \Delta_{\text{min}}})^2)$
BanditSRL [Tirinzoni et al., 2022]	CLB	Unknown	$\tilde{O}(\frac{d^2}{(\lambda^\star \Delta_{\text{min}})^2 \epsilon_{\text{min}}})$
LSVI-LEADER [Papini et al., 2021a]	Linear MDP	Known	$\tilde{O}(\max\{\frac{d^3 H^4}{(\lambda^\star)^2}, \frac{d^2 H^4}{\Delta_{\text{min}}^2 (\lambda^\star)^3}\})$
UNISREP-UCB + (this work)	Low-rank MDP	Unknown	$\tilde{O}(\frac{H^{12} d^8 \lvert\mathcal{A}\rvert^4}{(\Delta_{\text{min}} d_{\text{min}}^\star)^8 (\alpha\lambda^\star)^4})$

Constant Regret with Misspecified Representations Interestingly, as far as we know, there exists no algorithm for linear MDPs that can identify optimal behavior, when features are only required to have small misspecification error on average. In fact, only very recently, Agarwal et al. [2023] provided the first sublinear regret result in this setting. On the other hand, Zhang et al. [2024] provided an algorithm that achieves constant instance-dependent regret for linear-MDPs with features that have low point-wise misspecification w.r.t. the minimal sub-optimality gap.

6.2 LIMITATIONS

Redundant Features Following a similar analysis as in Papini et al. [2021a], our regret bounds would also hold for redundant UniSOFT feature maps, provided that we are guaranteed to select them. In order to learn possibly redundant UniSOFT feature maps, Tirinzoni et al. [2022] provided the following loss function:

$$\min_{(s,a)\in\mathcal{D}} \phi(s,a)^T \left(\sum_{(s',a')\in\mathcal{D}} \phi(s',a')\phi(s',a')^T \right) \phi(s,a).$$

However, this loss function selects UniSOFT feature maps only if all state-action pairs are visited in finite time; otherwise, we cannot ensure that the features of optimal actions span the observable feature space.

Low-Rank Assumption The set of MDPs that admit a low-rank representation with small rank d w.r.t. $|\mathcal{S}|$ is inherently limited. In particular, Lee and Oh [2024] showed that the feature dimension is lower bounded by $\lfloor \frac{|\mathcal{S}|}{U} \rfloor$, where $U := \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} |\{s' \in \mathcal{S} : \mathcal{P}(s'|s,a) > 0\}|$ is the maximum number of directly reachable states. An immediate consequence is that, in deterministic environments, $d = |\mathcal{S}|$ holds. We refer to Section 4 in Lee and Oh [2024] for a more thorough discussion.

Minimal Optimal Occupancy In contrast to existing work on constant regret for linear MDPS [Papini et al., 2021a, Zhang et al., 2023, 2024], our bound has an additional dependence in $d_{\text{min}}^\star$. This dependence is caused by controlling expected sub-optimality gaps. A point-wise uncertainty quantification is generally not possible since the MLE objective is unbounded and we cannot use any standard uniform convergence techniques. Nevertheless, $d_{\text{min}}^\star \approx \lambda^\star$ is generally a reasonable approximation, where quantities similar to $\lambda^\star$ appear in many existing works (e.g., see Table 1) that leverage representations with good spectral properties.

The inherently undesirable trade-off between $d_{\text{min}}^\star$ and d is interesting to note here. We seek highly random transitions to hope for a small rank d, but deterministic transitions for a large value $d_{\text{min}}^\star$.

Computation Computationally, Algorithm 1 suffers from limitations similar to those of other existing works on low-rank MDPs. In particular, the optimization oracle cannot be efficiently solved accurately, as there is no practical mechanism to guarantee the normalization conditions for ϕ and μ [Zhang et al., 2022]. This, in particular, makes the constraint optimization objective in algorithm line 17 intractable. However, the MLE objective can be approximated with noise contrastive estimation (NCE) [Zhang et al., 2022], with the UniSOFT loss added as a regularization term.

7 CONCLUSION & FUTURE WORK

In this work, we studied low-rank MDPs characterized by the instance-dependent properties Δ_{min} (minimal sub-optimality gap) and $d_{\text{min}}^\star$ (minimal optimal occupancy). We proposed to extend the existing REP-UCB algorithm with a double exploration strategy and a constrained optimization objective, and showed that this novel algorithm can leverage good representations for more efficient exploration. Additionally, we demonstrated that our algorithm enjoys constant regret in low-rank MDPs and provided a condition that is sufficient and necessary for the existence of good representations.

An interesting direction for future work is the design of computationally efficient variants of our proposed algorithms and to test them on deep RL benchmarks. Furthermore, it would be interesting to understand whether UniSOFT features are necessary for instance-dependent regret.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24, 2011.

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in Neural Information Processing Systems*, 33:20095–20107, 2020.

Alekh Agarwal, Yuda Song, Wen Sun, Kaiwen Wang, Mengdi Wang, and Xuezhou Zhang. Provable benefits of representational transfer in reinforcement learning. In *The Conference on Learning Theory*, pages 2114–2187. PMLR, 2023.

Yuan Cheng, Ruiquan Huang, Yingbin Liang, and Jing Yang. Improved sample complexity for reward-free reinforcement learning under low-rank MDPs. In *International Conference on Learning Representations*, 2023.

Botao Hao, Tor Lattimore, and Csaba Szepesvari. Adaptive exploration in linear contextual bandit. In *International Conference on Artificial Intelligence and Statistics*, pages 3536–3545. PMLR, 2020.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 4171–4180. PMLR, 2021.

Qiang He, Tianyi Zhou, Meng Fang, and Setareh Maghsudi. Adaptive regularization of representation rank as an implicit constraint of bellman equation. In *International Conference on Learning Representations*, 2024.

Jiawei Huang, Li Zhao, Tao Qin, Wei Chen, Nan Jiang, and Tie-Yan Liu. Tiered reinforcement learning: Pessimism in the face of uncertainty and constant regret. *Advances in Neural Information Processing Systems*, 35:679–690, 2022.

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

Joongkyu Lee and Min-hwan Oh. Demystifying linear mdps and novel dynamics aggregation framework. In *International Conference on Learning Representations*, 2024.

Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free representation learning and exploration in low-rank mdps. *Journal of Machine Learning Research*, 25(6):1–76, 2024.

Matteo Papini, Andrea Tirinzoni, Aldo Pacchiano, Marcello Restelli, Alessandro Lazaric, and Matteo Pirotta. Reinforcement learning in linear mdps: Constant regret and representation selection. *Advances in Neural Information Processing Systems*, 34:16371–16383, 2021a.

Matteo Papini, Andrea Tirinzoni, Marcello Restelli, Alessandro Lazaric, and Matteo Pirotta. Leveraging good representations in linear contextual bandits. In *International Conference on Machine Learning*, pages 8371–8380. PMLR, 2021b.

Robert Piziak and Patrick L Odell. Full rank factorization of matrices. *Mathematics Magazine*, 72(3):193–201, 1999.

Andrea Tirinzoni, Matteo Papini, Ahmed Touati, Alessandro Lazaric, and Matteo Pirotta. Scalable representation learning in linear contextual bandits with constant regret guarantees. *Advances in Neural Information Processing Systems*, 35:2307–2319, 2022.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12:389–434, 2012.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline RL in low-rank MDPs. In *International Conference on Learning Representations*, 2022.

Tianjun Zhang, Tongzheng Ren, Mengjiao Yang, Joseph Gonzalez, Dale Schuurmans, and Bo Dai. Making linear mdps practical via contrastive representation learning. In *International Conference on Machine Learning*, pages 26447–26466. PMLR, 2022.

Weitong Zhang, Jiafan He, Dongruo Zhou, Q Gu, and A Zhang. Provably efficient representation selection in low-rank markov decision processes: from online to offline rl. In *Uncertainty in Artificial Intelligence*, pages 2488–2497. PMLR, 2023.

Weitong Zhang, Zhiyuan Fan, Jiafan He, and Quanquan Gu. Achieving constant regret in linear markov decision processes. In *Advances in Neural Information Processing Systems*, 2024.

Xuezhou Zhang, Yuda Song, Masatoshi Uehara, Mengdi Wang, Alekh Agarwal, and Wen Sun. Efficient reinforcement learning in block mdps: A model-free representation learning approach. In *International Conference on Machine Learning*, pages 26517–26547. PMLR, 2022a.

Canzhe Zhao, Ruofeng Yang, Baoxiang Wang, Xuezhou Zhang, and Shuai Li. Learning adversarial low-rank markov decision processes with unknown transition and full-information feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

On Constant Regret for Low-Rank MDPs (Supplementary Material)

Alexander Sturm[1] **Sebastian Tschiatschek**[2]

[1]Faculty of Computer Science, UniVie Doctoral School Computer Science, University of Vienna, Vienna, Austria
[2]Faculty of Computer Science, Research Network Data Science, University of Vienna, Vienna, Austria

Here we provide the omitted proofs of the main paper. In particular, Section A provides the proof for the baseline result in Theorem 4.1, in Section B we show how we guarantee the selection of good representations and in Section C and Section D we show how good representations can be leveraged to obtain an improved regret bound (Theorem 4.1) and constant regret (Theorem 4.3), respectively. Finally, in Section E we discuss the existence of good representations, in Section F we show how our results can be extended for multiple optimal policies and Section G provides auxiliary results.

We begin by introducing notation and good events. Let us denote

$$\tilde{\pi}_{t,h}(a|s) = \xi_{t-1} \cdot \frac{1}{|\mathcal{A}|} + (1 - \xi_{t-1}) \cdot \pi_{t-1,h}(a|s)$$

as the roll-out policy in episode t, which, with probability ξ_t, explores by taking an action uniformly at random and otherwise, selects an action according to the behavior policy $\pi_{t-1,h}$ from the previous episode. Importantly, we assume that the sequence $(\xi_t)_{t=1}^T$ is decreasing. Note that policy $\tilde{\pi}_{t,h}$ collects the transitions stored in the datasets of algorithm 1 and only interacts with the environment after sampling a state from $d^{\pi_{t-1}}_{\mathcal{P}^\star,h-1}$. Further, we denote the average roll-out policy as

$$\bar{\pi}_{t,h}(a|s) = \frac{1}{t} \sum_{i=0}^{t-1} \left(\xi_i \cdot \frac{1}{|\mathcal{A}|} + (1 - \xi_i) \cdot \pi_{i,h}(a|s) \right),$$

We define the mixture occupancy distributions

$$\rho_{t,h}(s) = \frac{1}{t} \sum_{i=0}^{t-1} d^{\pi_i}_{\mathcal{P}^\star,h}(s),$$

$$\gamma_{t,h}(s,a) = \frac{1}{t} \sum_{i=0}^{t-1} d^{\pi_i}_{\mathcal{P}^\star,h}(s,a),$$

$$\rho_{t,h}(s,a) = \rho_{t,h}(s)\bar{\pi}_{t,h}(a|s),$$

the next-state marginal distribution and next-state mixture occupancy distribution

$$\rho'_{t,h}(s') = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \rho_{t,h-1}(s,a)\mathcal{P}^\star_h(s'|s,a), \text{ and}$$

$$\rho'_{t,h}(s,a) = \rho'_{t,h}(s)\bar{\pi}_{t,h}(a|s),$$

respectively. Denote the total variation distance between the estimated model and the true model as

$$f_{t,h}(s,a) := \|\hat{\mathcal{P}}_{h,t}(\cdot|s,a) - \mathcal{P}^\star_h(\cdot|s,a)\|_{\text{TV}}.$$

Additionally, let

$$\Sigma_{\rho_t,\phi} = t\mathbb{E}_{(s,a)\sim\rho_t}[\phi(s,a)\phi(s,a)^T] + \lambda_t I,$$

where $\lambda_t = c_1 d \log(4tH|\Phi|/\delta)$, c_1 is a constant and $\rho_t \in \Delta(\mathcal{S} \times \mathcal{A})$ is an episode dependent distribution over the state-action space. Further we define the following two good events:

$$\begin{aligned}
&\mathcal{E}_1(\delta) = \{\forall t \in \mathbb{N}, h \in [H], s \in \mathcal{S}, a \in \mathcal{A} : \mathbb{E}_{(s,a)\sim\rho'_{t,h}}[f_{t,h}(s,a)^2] \leq \zeta_t\} \\
&\mathcal{E}_2(\delta) = \{\forall t \in \mathbb{N}, h \in [H], s \in \mathcal{S}, a \in \mathcal{A} : \\
&\qquad \frac{1}{5}\|\hat{\phi}_{t,h}(s,a)\|_{\Sigma^{-1}_{\rho_{t,h},\hat{\phi}_{t,h}}} \leq \|\hat{\phi}_{t,h}(s,a)\|_{\hat{\Sigma}^{-1}_{t,h}} \leq 3\|\hat{\phi}_{t,h}(s,a)\|_{\Sigma^{-1}_{\rho_{t,h},\hat{\phi}_{t,h}}}\},
\end{aligned}$$

where $\zeta_t = \frac{2\log(4t|\Phi||\Psi|H/\delta)}{t}$. Finally, let $\mathcal{E}(\delta) := \mathcal{E}_1(\delta/2) \cap \mathcal{E}_2(\delta/2)$. The good event $\mathcal{E}$ guarantees the convergence of the MLE oracle [Uehara et al., 2022] and the concentration of the bonus term.

Lemma .1. *Fix $\delta \in (0,1)$. Suppose Assumption 3.1 (realizability) holds and we run algorithm 1. Then, with probability at least $1-\delta$, the event $\mathcal{E}(\delta)$ occurs.*

Proof. By Lemma G.6, with probability at least $1-\delta/2$, event $\mathcal{E}_1(\delta/2)$ occurs. Furthermore, by Lemma 11 in Uehara et al. [2022], with probability at least $1-\delta/2$, event $\mathcal{E}_2(\delta/2)$ occurs. Taking an union bound concludes the proof. □

A SUB-LINEAR PSEUDO-REGRET WITHOUT GOOD REPRESENTATIONS

In this section, we show that the behavior policies of algorithm 1 achieve anytime sub-linear regret without exploiting the UniSOFT property. On a high level, this ensures that the algorithm plays optimal actions often enough, such that the MLE constrained oracle eventually selects UniSOFT features, which we leverage in subsequent sections to improve upon the baseline result. We note that the analysis in this section is purely based on known results and provided for completeness.

We start by providing two important results, first introduced by Uehara et al. [2022], which we will use to link the bonus of the learned features to the elliptical potential function of the true features. This allows us to track the progress of our algorithm through the standard elliptical potential lemma G.4.

Lemma A.1. *(One-step back inequality in the true model) Consider a set of functions $\{g_h\}_{h=1}^H$ that satisfies $g_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that $\|g_h\|_\infty \leq B$ for all $h \in [H]$. Then, for all $t \in \mathbb{N}$, $h > 1$ and any π,*

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim d^\pi_{\mathcal{P}^\star,h}}[g_h(s,a)] \\
&\qquad \leq \mathbb{E}_{(s,a)\sim d^\pi_{\mathcal{P}^\star,h-1}}[\|\phi^\star_{h-1}(s,a)\|_{\Sigma^{-1}_{\gamma_{t,h-1},\phi^\star_{h-1}}}]\sqrt{t\frac{|\mathcal{A}|}{\xi_t}\mathbb{E}_{(s,a)\sim\rho_{t,h}}[g_h(s,a)^2] + B^2\lambda_t d}
\end{aligned}$$

Proof. For $h = 2, ..., H$ we have,

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim d^\pi_{\mathcal{P}^\star,h}}[g_h(s,a)] \\
&= \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^\pi_{\mathcal{P}^\star,h-1}, s\sim\mathcal{P}^\star_{h-1}(\cdot|\tilde{s},\tilde{a}), a\sim\pi_h(\cdot|s)}[g_h(s,a)] \\
&= \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^\pi_{\mathcal{P}^\star,h-1}}[\langle\phi^\star_{h-1}(\tilde{s},\tilde{a}), \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mu^\star_{h-1}(s)\pi_h(a|s)g_h(s,a)\rangle] \\
&\overset{(i)}{\leq} \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^\pi_{\mathcal{P}^\star,h-1}}[\|\phi^\star_{h-1}(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\gamma_{t,h-1},\phi^\star_{h-1}}} \|\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mu^\star_{h-1}(s)\pi_h(a|s)g_h(s,a)\|_{\Sigma_{\gamma_{t,h-1},\phi^\star_{h-1}}}],
\end{aligned}$$

where (i) follows from the symmetry of the regularized covariance matrix and an application of the Cauchy-Schwarz inequality. Further we have for $h = 2, ..., H$,

$$
\begin{aligned}
&\|\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mu^\star_{h-1}(s)\pi_h(a|s)g_h(s,a)\|^2_{\Sigma_{\gamma_{t,h-1},\phi^\star_{h-1}}} \\
&\overset{(i)}{\leq} t\mathbb{E}_{(\tilde{s},\tilde{a})\sim\gamma_{t,h-1}}[(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \langle\phi^\star_{h-1}(\tilde{s},\tilde{a}), \mu^\star_{h-1}(s)\rangle\pi_h(a|s)g_h(s,a))^2] + B^2\lambda_t d \\
&= t\mathbb{E}_{(\tilde{s},\tilde{a})\sim\gamma_{t,h-1}}[\mathbb{E}_{s\sim\mathcal{P}^\star_{h-1}(\cdot|\tilde{s},\tilde{a}),a\sim\pi_h(\cdot|s)}[g_h(s,a)]^2] + B^2\lambda_t d \\
&\leq t\mathbb{E}_{s\sim\rho_{t,h},a\sim\pi_h(\cdot|s)}[g_h(s,a)^2] + B^2\lambda_t d \\
&\overset{(ii)}{\leq} t\max_{s,a}\frac{\rho_{t,h}(s)\pi_h(a|s)}{\rho_{t,h}(s)\bar{\pi}_{t,h}(a|s)}\mathbb{E}_{(s,a)\sim\rho_{t,h}}[g_h(s,a)^2] + B^2\lambda_t d \\
&\leq t\frac{1}{\frac{1}{t}\sum_{i=0}^{t-1}(\xi_i\cdot\frac{1}{|\mathcal{A}|})}\mathbb{E}_{(s,a)\sim\rho_{t,h}}[g_h(s,a)^2] + B^2\lambda_t d \\
&\overset{(iii)}{\leq} t\frac{|\mathcal{A}|}{\xi_t}\mathbb{E}_{(s,a)\sim\rho_{t,h}}[g_h(s,a)^2] + B^2\lambda_t d,
\end{aligned}
$$

where, (i) is by assumptions $\|g_h(s,a)\|_\infty \leq B$ and $\|\int_{\mathcal{S}}\mu^\star(s)h(s)p(s)\|_2 \leq \sqrt{d}$ for any $h : \mathcal{S} \to [0,1]$ (realizability, Assumption 3.1), (ii) is by importance sampling and (iii) follows from ξ_t being decreasing. □

Lemma A.2. *(One-step back inequality in the learned model) Consider a set of functions* $\{g_h\}_{h=1}^H$ *that satisfies* $g_h : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ *such that* $\|g_h\|_\infty \leq B$ *for all* $h\in[H]$. *Then, given that the event* $\mathcal{E}$ *occurs, for all* $t\in\mathbb{N}$, $h>1$ *and any* π,

$$
\begin{aligned}
&\mathbb{E}_{(s,a)\sim d^\pi_{\hat{\mathcal{P}}_{t,h}}}[g_h(s,a)] \\
&\quad\leq \mathbb{E}_{(s,a)\sim d^\pi_{\hat{\mathcal{P}}_{t,h-1}}}[\|\hat{\phi}_{t,h-1}(s,a)\|_{\Sigma^{-1}_{\rho_{t,h-1},\hat{\phi}_{t,h-1}}}]\sqrt{2t\frac{|\mathcal{A}|}{\xi_t}\mathbb{E}_{(s,a)\sim\rho'_{t,h}}[g_h(s,a)^2] + B^2\lambda_t d + 2t\frac{|\mathcal{A}|}{\xi_t}B^2\zeta_t}
\end{aligned}
$$

Proof. Let $t\in\mathbb{N}$ be arbitrary. For all $h = 2, ..., H$ we have,

$$
\begin{aligned}
&\mathbb{E}_{(s,a)\sim d^\pi_{\hat{\mathcal{P}}_{t,h}}}[g_h(s,a)] \\
&= \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^\pi_{\hat{\mathcal{P}}_{t,h-1}},s\sim\hat{\mathcal{P}}_{t,h-1}(\cdot|\tilde{s},\tilde{a}),a\sim\pi_h(\cdot|s)}[g_h(s,a)] \\
&= \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^\pi_{\hat{\mathcal{P}}_{t,h-1}}}[\langle\hat{\phi}_{t,h-1}(\tilde{s},\tilde{a}), \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\hat{\mu}_{t,h-1}(s)\pi_h(a|s)g_h(s,a)\rangle] \\
&\overset{(i)}{\leq} \mathbb{E}_{(\tilde{s},\tilde{a})\sim d^\pi_{\hat{\mathcal{P}}_{t,h-1}}}[\|\hat{\phi}_{t,h-1}(\tilde{s},\tilde{a})\|_{\Sigma^{-1}_{\rho_{t,h-1},\hat{\phi}_{t,h-1}}}\|\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\hat{\mu}_{t,h-1}(s)\pi_h(a|s)g_h(s,a)\|_{\Sigma_{\rho_{t,h-1},\hat{\phi}_{t,h-1}}}],
\end{aligned}
$$

where (i) follows from the symmetry of the covariance matrix and an application of the Cauchy-Schwarz inequality. Further we have for all $h = 2, ..., H$,

$$
\begin{aligned}
&\|\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\hat{\mu}_{t,h-1}(s)\pi_h(a|s)g_h(s,a)\|^2_{\Sigma_{\rho_{t,h-1},\hat{\phi}_{t,h-1}}} \\
&\overset{(i)}{\leq} t\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_{t,h-1}}[(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\langle\hat{\phi}_{t,h-1}(\tilde{s},\tilde{a}), \hat{\mu}_{t,h-1}(s)\rangle\pi_h(a|s)g_h(s,a))^2] + B^2\lambda_t d \\
&= t\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_{t,h-1}}[\mathbb{E}_{s\sim\hat{\mathcal{P}}_{t,h-1}(\cdot|\tilde{s},\tilde{a}),a\sim\pi_h(\cdot|s)}[g_h(s,a)]^2] + B^2\lambda_t d \\
&\overset{(ii)}{\leq} 2t\mathbb{E}_{(\tilde{s},\tilde{a})\sim\rho_{t,h-1}}[\mathbb{E}_{s\sim\mathcal{P}^\star_{h-1}(\cdot|\tilde{s},\tilde{a}),a\sim\pi_h(\cdot|s)}[g_h(s,a)]^2] + B^2\lambda_t d + 2tB^2\zeta_t \\
&\leq 2t\mathbb{E}_{s\sim\rho'_{t,h},a\sim\pi_h(\cdot|s)}[g_h(s,a)^2] + B^2\lambda_t d + 2t\frac{|\mathcal{A}|}{\xi_{t-1}}B^2\zeta_t \\
&\overset{(iii)}{\leq} 2t\frac{|\mathcal{A}|}{\xi_{t-1}}\mathbb{E}_{(s,a)\sim\rho'_{t,h}}[g_h(s,a)^2] + B^2\lambda_t d + 2t\frac{|\mathcal{A}|}{\xi_{t-1}}B^2\zeta_t,
\end{aligned}
$$

where, (i) is by assumptions $\|g_h(s,a)\|_\infty \leq B$ and $\| \int_{\mathcal{S}} \hat{\mu}(s)h(s)p(s)\|_2 \leq \sqrt{d}$ for any $h : \mathcal{S} \to [0,1]$ (realizability, Assumption 3.1), (ii) follows from $(a+b)^2 \leq 2a^2 + 2b^2$, importance sampling and the event $\mathcal{E}$ and (iii) is again by importance sampling. □

The following lemma exploits the one-step back inequalities to relate the bonus and the estimation error to elliptical potential functions. The formulation of the statement is inspired by Lemma 3 of Cheng et al. [2023].

Lemma A.3. *(Bonus relations) Given that the event $\mathcal{E}$ occurs, for all $t \in \mathbb{N}$, $h > 1$ and any π,*

$$\mathbb{E}_{(s,a)\sim d^\pi_{\hat{\mathcal{P}}_t,h}}[f_{t,h}(s,a)] \leq \alpha_t \mathbb{E}_{(s,a)\sim d^\pi_{\hat{\mathcal{P}}_t,h-1}}[\|\hat{\phi}_{t,h-1}(s,a)\|_{\Sigma^{-1}_{\rho_{t,h-1},\hat{\phi}_{t,h-1}}}],$$
$$\mathbb{E}_{(s,a)\sim d^\pi_{\mathcal{P}^\star,h}}[f_{t,h}(s,a)] \leq \alpha_t \mathbb{E}_{(s,a)\sim d^\pi_{\mathcal{P}^\star,h-1}}[\|\phi^\star_{h-1}(s,a)\|_{\Sigma^{-1}_{\rho_{t,h-1},\phi^\star_{h-1}}}],$$
$$\mathbb{E}_{(s,a)\sim d^\pi_{\mathcal{P}^\star,h}}[\hat{b}_{t,h}(s,a)] \leq \beta_t \mathbb{E}_{(s,a)\sim d^\pi_{\mathcal{P}^\star,h-1}}[\|\phi^\star_{h-1}(s,a)\|_{\Sigma^{-1}_{\gamma_{t,h-1},\phi^\star_{h-1}}}],$$

where $\alpha_t = \sqrt{4t\zeta_t \frac{|\mathcal{A}|}{\xi_t} + \lambda_t d}$ and $\beta_t = \sqrt{\frac{|\mathcal{A}|}{\xi_t} 40\alpha_t^2 d + \lambda_t d}$. In particular, for h=1,

$$\mathbb{E}_{s\sim d_1, a\sim\pi_1(\cdot|s)}[f_{t,1}(s,a)] \leq \sqrt{\frac{|\mathcal{A}|}{\xi_t}\zeta_t}, \qquad \mathbb{E}_{s\sim d_1, a\sim\pi_1(\cdot|s)}[\hat{b}_{t,1}(s,a)] \leq 15\alpha_t \sqrt{\frac{d|\mathcal{A}|}{t\xi_t}}.$$

Proof. Let $t \in \mathbb{N}$ be arbitrary. For all $h > 1$ we have,

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim d^\pi_{\hat{\mathcal{P}}_t,h}}[f_{t,h}(s,a)] \\
&\overset{(i)}{\leq} \mathbb{E}_{(s,a)\sim d^\pi_{\hat{\mathcal{P}}_t,h-1}}[\|\hat{\phi}_{t,h-1}(s,a)\|_{\Sigma^{-1}_{\rho_{t,h-1},\hat{\phi}_{t,h-1}}}]\sqrt{2t\frac{|\mathcal{A}|}{\xi_t}\mathbb{E}_{(s,a)\sim\rho'_{t,h}}[f_{t,h}(s,a)^2] + \lambda_t d + 2t\frac{|\mathcal{A}|}{\xi_t}\zeta_t} \\
&\overset{(ii)}{\leq} \alpha_t \mathbb{E}_{(s,a)\sim d^\pi_{\hat{\mathcal{P}}_t,h-1}}[\|\hat{\phi}_{t,h-1}(s,a)\|_{\Sigma^{-1}_{\rho_{t,h-1},\hat{\phi}_{t,h-1}}}],
\end{aligned}$$

where (i) is by Lemma A.2 and $\|f_{t,h}\|_\infty \leq 1$ and (ii) follows from the event $\mathcal{E}$. Similarly, for all $h > 1$,

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim d^\pi_{\mathcal{P}^\star_t,h}}[f_{t,h}(s,a)] \\
&\overset{(i)}{\leq} \mathbb{E}_{(s,a)\sim d^\pi_{\mathcal{P}^\star,h-1}}[\|\phi^\star_{h-1}(s,a)\|_{\Sigma^{-1}_{\rho_{t,h-1},\phi^\star_{h-1}}}]\sqrt{2t\frac{|\mathcal{A}|}{\xi_t}\mathbb{E}_{(s,a)\sim\rho'_{t,h}}[f_{t,h}(s,a)^2] + \lambda_t d} \\
&\overset{(ii)}{\leq} \alpha_t \mathbb{E}_{(s,a)\sim d^\pi_{\mathcal{P}^\star,h-1}}[\|\phi^\star_{h-1}(s,a)\|_{\Sigma^{-1}_{\rho_{t,h-1},\phi^\star_{h-1}}}],
\end{aligned}$$

where (i) is by Lemma A.1 and $\|f_{t,h}\|_\infty \leq 1$ and (ii) follows from the event $\mathcal{E}$. For $h = 1$ we have,

$$\mathbb{E}_{s\sim d_1, a\sim\pi_1(\cdot|s)}[f_{t,1}(s,a)] \overset{(i)}{\leq} \sqrt{\frac{|\mathcal{A}|}{\xi_t}\mathbb{E}_{(s,a)\sim\rho_{t,1}}[f_{t,1}(s,a)^2]} \leq \sqrt{\frac{|\mathcal{A}|}{\xi_t}\zeta_t},$$

where (i) is by importance sampling and Jensen's inequality. We can bound the bonus by,

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim d^\pi_{\mathcal{P}^\star_t,h}}[\hat{b}_{t,h}(s,a)] \\
&\leq \mathbb{E}_{(s,a)\sim d^\pi_{\mathcal{P}^\star,h-1}}[\|\phi^\star_{h-1}(s,a)\|_{\Sigma^{-1}_{\gamma_{t,h-1},\phi^\star_{h-1}}}]\sqrt{t\frac{|\mathcal{A}|}{\xi_t}\mathbb{E}_{(s,a)\sim\rho_{t,h}}[\hat{b}_{t,h}(s,a)^2] + \lambda_t d},
\end{aligned}$$

which follows from Lemma A.1 and $\|\hat{b}_{t,h}\|_\infty \leq 1$. Further,

$$\begin{aligned}
&t\mathbb{E}_{(s,a)\sim\rho_{t,h}}[\hat{b}_{t,h}(s,a)^2] \\
&\leq t\mathbb{E}_{(s,a)\sim\rho_{t,h}}[\hat{\alpha}_t^2\|\hat{\phi}_{t,h}(s,a)\|^2_{\hat{\Sigma}^{-1}_{t,h}}] \\
&\overset{(i)}{\leq} t\mathbb{E}_{(s,a)\sim\rho_{t,h}}[9\hat{\alpha}_t^2\|\hat{\phi}_{t,h}(s,a)\|^2_{\Sigma^{-1}_{\rho_{t,h},\hat{\phi}_{t,h}}}] \\
&= 9\hat{\alpha}_t^2 t\mathrm{Tr}\left(\mathbb{E}_{(s,a)\sim\rho_{t,h}}[\hat{\phi}_{t,h}(s,a)\hat{\phi}_{t,h}(s,a)^T](t\mathbb{E}_{(s,a)\sim\rho_{t,h}}[\hat{\phi}_{t,h}(s,a)\hat{\phi}_{t,h}(s,a)^T]+\lambda_t I)^{-1}\right) \\
&\overset{(ii)}{\leq} 9\hat{\alpha}_t^2 d,
\end{aligned}$$

where (i) follows from the event $\mathcal{E}$ and (ii) follows from Lemma G.3. Therefore,

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim d^\pi_{\mathcal{P}^\star_t,h}}[\hat{b}_{t,h}(s,a)] \\
&\leq \mathbb{E}_{(s,a)\sim d^\pi_{\mathcal{P}^\star,h-1}}[\|\phi^\star_{h-1}(s,a)\|_{\Sigma^{-1}_{\gamma_{t,h-1},\phi^\star_{h-1}}}]\sqrt{\frac{|\mathcal{A}|}{\xi_t}9\hat{\alpha}_t^2 d+\lambda_t d}.
\end{aligned}$$

Finally, for $h = 1$,

$$\begin{aligned}
\mathbb{E}_{s\sim d_1,a\sim\pi_1(\cdot|s)}[\hat{b}_{t,1}(s,a)] &\overset{(i)}{\leq} 3\hat{\alpha}_t\sqrt{\frac{|\mathcal{A}|}{\xi_t}\mathbb{E}_{(s,a)\sim\rho_{t,1}}[\|\hat{\phi}_{t,1}(s,a)\|^2_{\Sigma^{-1}_{\rho_{t,1},\hat{\phi}_{t,1}}}]} \\
&\overset{(ii)}{\leq} 15\alpha_t\sqrt{\frac{d|\mathcal{A}|}{t\xi_t}},
\end{aligned}$$

where (i) follows from the event $\mathcal{E}$, importance sampling and Jensen's inequality and (ii) follows from Lemma G.3. □

The next result shows that the optimal value w.r.t. the bonus-augmented reward function in the estimated environment provides an almost optimistic estimate of the true value achieved by any optimal policy.

Lemma A.4. *(Almost Optimism at the Initial Distribution) Given that the event $\mathcal{E}$ occurs, for all $t \in \mathbb{N}$,*

$$V^{\pi^\star,d_1}_{\mathcal{P}^\star,r^\star,1} - V^{\pi^\star,d_1}_{\hat{\mathcal{P}},r^\star+\hat{b}_t,1} \leq \sqrt{\frac{|\mathcal{A}|}{\xi_t}}\zeta_t$$

Proof. Let $t \in \mathbb{N}$ be arbitrary.

$$\begin{aligned}
&V^{\pi^\star,d_1}_{\mathcal{P}^\star,r^\star,1} - V^{\pi^\star,d_1}_{\hat{\mathcal{P}},r^\star+\hat{b}_t,1} \\
&\overset{(i)}{=} \sum_{h=1}^H \mathbb{E}_{(s,a)\sim d^{\pi^\star}_{\hat{\mathcal{P}}_t,h}}[(\mathcal{P}^\star_h - \hat{\mathcal{P}}_{t,h})V^{\pi^\star}_{\mathcal{P}^\star,r^\star,h+1}(s,a) - \hat{b}_{t,h}(s,a)] \\
&\overset{(ii)}{\leq} \sum_{h=1}^H \mathbb{E}_{(s,a)\sim d^{\pi^\star}_{\hat{\mathcal{P}}_t,h}}[f_{t,h}(s,a) - \min\{1, \frac{\hat{\alpha}_t}{5}\|\hat{\phi}_{t,h}(s,a)\|_{\Sigma^{-1}_{\rho_{t,h},\hat{\phi}_{t,h}}}\}] \\
&\overset{(iii)}{\leq} \sqrt{\frac{|\mathcal{A}|}{\xi_t}}\zeta_t + \sum_{h=1}^{H-1} \mathbb{E}_{(s,a)\sim d^{\pi^\star}_{\hat{\mathcal{P}}_t,h}}[\min\{1, \alpha_t\|\hat{\phi}_{t,h}(s,a)\|_{\Sigma^{-1}_{\rho_{t,h},\hat{\phi}_{t,h}}}\}] \\
&\qquad - \sum_{h=1}^H \mathbb{E}_{(s,a)\sim d^{\pi^\star}_{\hat{\mathcal{P}}_t,h}}[\min\{1, \alpha_t\|\hat{\phi}_{t,h}(s,a)\|_{\Sigma^{-1}_{\rho_{t,h},\hat{\phi}_{t,h}}}\}] \\
&\leq \sqrt{\frac{|\mathcal{A}|}{\xi_t}}\zeta_t,
\end{aligned}$$

where (i) follows from Lemma G.1, (ii) follows from the event $\mathcal{E}$ and $\|V^\pi_{\mathcal{P}^\star,r^\star}\|_\infty \leq 1$ and (iii) follows from Lemma A.3 and $\|f_{t,h}\|_\infty \leq 1$. □

We are now ready to show that algorithm 1 achieves sub-linear pseudo-regret; that is, the regret of the behavior polices is sub-linear. However, the actual regret of algorithm 1 might not be, as we explore uniformly at random in each episode with positive probability.

Lemma A.5. *(Sub-linear pseudo-regret without UniSOFT representations) Given that event $\mathcal{E}$ occurs for all $T \in \mathbb{N}$,*

$$\mathcal{R}(T) \lesssim H^2 d^2 |\mathcal{A}| \frac{\sqrt{T} \log^2(4TH|\Phi||\Psi|/\delta)}{\xi_T} \lesssim \tilde{O}(\frac{\sqrt{T}}{\xi_T}).$$

Proof. Let $T \in \mathbb{N}$ be arbitrary. Then, for all episodes $t \leq T$ we have,

$$\begin{aligned}
&V^{\pi^\star, d_1}_{\mathcal{P}^\star, r^\star, 1} - V^{\pi_t, d_1}_{\mathcal{P}^\star, r^\star, 1} \\
&= V^{\pi^\star, d_1}_{\hat{\mathcal{P}}_t, \hat{b}_t + r^\star, 1} - V^{\pi_t, d_1}_{\mathcal{P}^\star, r^\star, 1} + V^{\pi^\star, d_1}_{\mathcal{P}^\star, r^\star, 1} - V^{\pi^\star, d_1}_{\hat{\mathcal{P}}_t, \hat{b}_t + r^\star, 1} \\
&\overset{(i)}{\leq} V^{\pi_t, d_1}_{\hat{\mathcal{P}}_t, \hat{b}_t + r^\star, 1} - V^{\pi_t, d_1}_{\mathcal{P}^\star, r^\star, 1} + \sqrt{\frac{|\mathcal{A}|}{\xi_t} \zeta_t} \\
&\overset{(ii)}{=} \sum_{h=1}^{H} \mathbb{E}_{(s,a) \sim d^{\pi_t}_{\mathcal{P}^\star, h}} [\hat{b}_{t,h}(s,a) + (\hat{\mathcal{P}}_{t,h} - \mathcal{P}^\star_h) V^{\pi_t}_{\hat{\mathcal{P}}_t, r^\star + \hat{b}_t, h+1}(s,a)] + \sqrt{\frac{|\mathcal{A}|}{\xi_t} \zeta_t} \\
&\overset{(iii)}{\leq} 2H \sum_{h=1}^{H} \mathbb{E}_{(s,a) \sim d^{\pi_t}_{\mathcal{P}^\star, h}} [\hat{b}_{t,h}(s,a) + f_{t,h}(s,a)] + \sqrt{\frac{|\mathcal{A}|}{\xi_t} \zeta_t},
\end{aligned}$$

where (i) is by Lemma A.4, (ii) follows from Lemma G.1 and (iii) follows from $\|V^{\pi}_{\mathcal{P}, r^\star + \hat{b}}\|_\infty \leq 2H$. Then, by Lemma A.3,

$$\begin{aligned}
&\sum_{h=1}^{H} \mathbb{E}_{(s,a) \sim d^{\pi_t}_{\mathcal{P}^\star, h}} [\hat{b}_{t,h}(s,a) + f_{t,h}(s,a)] \\
&\leq \sqrt{\frac{\zeta_t |\mathcal{A}|}{\xi_t}} + 15\alpha_t \sqrt{\frac{d|\mathcal{A}|}{t\xi_t}} + \alpha_t \sum_{h=1}^{H-1} \mathbb{E}_{(s,a) \sim d^{\pi_t}_{\mathcal{P}^\star, h}} [\|\phi^\star_h(s,a)\|_{\Sigma^{-1}_{\rho_{t,h}, \phi^\star_h}}] \\
&\qquad + \beta_t \sum_{h=1}^{H-1} \mathbb{E}_{(s,a) \sim d^{\pi_t}_{\mathcal{P}^\star, h}} [\|\phi^\star_h(s,a)\|_{\Sigma^{-1}_{\gamma_{t,h}, \phi^\star_h}}].
\end{aligned}$$

Further, for all $h \in [H]$,

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}_{(s,a) \sim d^{\pi_t}_{\mathcal{P}^\star, h}} [\|\phi^\star_h(s,a)\|_{\Sigma^{-1}_{\gamma_{t,h}, \phi^\star_h}}] &\overset{(i)}{\leq} \sqrt{T \sum_{t=1}^{T} \mathbb{E}_{(s,a) \sim d^{\pi_t}_{\mathcal{P}^\star, h}} [\|\phi^\star_h(s,a)\|^2_{\Sigma^{-1}_{\gamma_{t,h}, \phi^\star_h}}]} \\
&= \sqrt{T \sum_{t=1}^{T} \text{tr}(\mathbb{E}_{(s,a) \sim d^{\pi_t}_{\mathcal{P}^\star, h}} [\phi^\star_h(s,a) \phi^\star_h(s,a)^T] \Sigma^{-1}_{\gamma_{t,h}, \phi^\star_h})} \\
&\overset{(ii)}{\leq} \sqrt{T d \log(1 + \frac{T}{d\lambda_1})}
\end{aligned}$$

where (i) follows from the Cauchy-Schwarz inequality and Jensen's inequality and (ii) follows from Lemma G.4 by noting that, $\Sigma^{-1}_{\gamma_{t,h}, \phi} - \lambda_t I = t\mathbb{E}_{\gamma_{t,h}}[\phi\phi^T] = \sum_{i=1}^{t} \mathbb{E}_{d^{\pi_i}_{\mathcal{P}^\star, h}}[\phi\phi^T]$ and that λ_t is increasing. Similarly, for all $h \in [H]$,

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}_{(s,a) \sim d^{\pi_t}_{\mathcal{P}^\star, h}} [\|\phi^\star_h(s,a)\|_{\Sigma^{-1}_{\rho_{t,h}, \phi^\star_h}}] &\overset{(i)}{\leq} \sqrt{T \frac{|\mathcal{A}|}{\xi_T} \sum_{t=1}^{T} \mathbb{E}_{s \sim d^{\pi_t}_{\mathcal{P}^\star, h}, a \sim \mathcal{U}(\mathcal{A})} [\|\phi^\star_h(s,a)\|^2_{\Sigma^{-1}_{\rho_{t,h}, \phi^\star_h}}]} \\
&\overset{(ii)}{\leq} \sqrt{T \frac{|\mathcal{A}|}{\xi_T} d \log(1 + \frac{T}{d\lambda_1})},
\end{aligned}$$

where (i) follows from the Cauchy-Schwarz inequality, Jensen's inequality, importance Sampling and ξ_t being decreasing and (ii) follows from Lemma G.4. Finally,

$$\begin{aligned}
&\sum_{t=1}^{T} V_{\mathcal{P}^\star,r^\star,1}^{\pi^\star,d_1} - V_{\mathcal{P}^\star,r^\star,1}^{\pi_t,d_1} \\
&\leq 8H\sqrt{\frac{\zeta_T T^2|\mathcal{A}|}{\xi_T}} + 30H\alpha_T\sqrt{\frac{Td|\mathcal{A}|}{\xi_T}} + 2H^2\alpha_T\sqrt{T\frac{|\mathcal{A}|}{\xi_T}d\log(1+\frac{T}{d\lambda_1})} \\
&\quad + 2H^2\beta_T\sqrt{Td\log(1+\frac{T}{d\lambda_1})} \\
&\lesssim H^2d^2|\mathcal{A}|\frac{\sqrt{T}\log^2(4HT|\Phi||\Psi|/\delta)}{\xi_T}
\end{aligned}$$

□

We proceed by providing an expected regret bound. Let $\mathbb{E}_\xi$ and $\mathbb{E}_\delta$ denote expectations w.r.t. the exploration probabilities and some good event $\mathcal{E}(\delta)$, respectively. Additionally, note that we sample from $d_{\mathcal{P}^\star,h}^{\pi_t}$ for each time step and hence produce H trajectories per episode. Then, the expected regret of algorithm 1 can be upper bounded as follows:

Lemma 4.1 (Expected Regret without UniSOFT). *Let $\xi_t = t^{-1/4}$. Suppose Assumption 3.1 (realizability) holds. Then, for any $T \in \mathbb{N}$,* UNISREP-UCB *(Algorithm 1) satisfies*

$$\mathbb{E}[\tilde{\mathcal{R}}(T)] = \tilde{O}\left(H^3d^2|\mathcal{A}|T^{3/4}\right).$$

Proof. Let T be given and fixed. Choose $\delta = T^{-1}$. Recall that Algorithm 1 explores for H time steps, for each $h \in [H]$ and episode t, by rolling into time step $h-1$ with policy π_{t-1}, taking actions according to $\tilde{\pi}_{t,h-1}$ and $\tilde{\pi}_{t,h}$ and finally, rolling out to time step H with policy π_{t-1}. Let us denote $\tilde{V}_{t,h}^{d_1}$ as the cumulative expected reward obtained by Algorithm 1 in episode t and time step h. Then,

$$\begin{aligned}
&\mathbb{E}_{\delta,\xi}[\tilde{\mathcal{R}}(T)] \\
&= \mathbb{E}_{\delta,\xi}[\sum_{t=1}^{T}\sum_{h=1}^{H}(V_{\mathcal{P}^\star,r^\star,1}^{\pi^\star,d_1} - \tilde{V}_{t,h}^{d_1})] \\
&\leq \mathbb{E}_{\delta,\xi}[\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{1}\{e_t=1\}\mathbb{1}\{\mathcal{E}(\delta)\}(V_{\mathcal{P}^\star,r^\star,1}^{\pi^\star,d_1} - \tilde{V}_{t,h}^{d_1})] + \mathbb{E}_{\delta,\xi}[\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{1}\{e_t=0\}(V_{\mathcal{P}^\star,r^\star,1}^{\pi^\star,d_1} - \tilde{V}_{t,h}^{d_1})] \\
&\quad + \mathbb{E}_{\delta,\xi}[\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{1}\{\mathcal{E}^c(\delta)\}(V_{\mathcal{P}^\star,r^\star,1}^{\pi^\star,d_1} - \tilde{V}_{t,h}^{d_1})] \\
&\overset{(i)}{\leq} \mathbb{E}_{\delta,\xi}[\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{1}\{e_t=1\}\mathbb{1}\{\mathcal{E}(\delta)\}(V_{\mathcal{P}^\star,r^\star,1}^{\pi^\star,d_1} - \tilde{V}_{t,h}^{d_1})] + \mathbb{E}_{\delta,\xi}[\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{1}\{e_t=0\} + \mathbb{1}\{\mathcal{E}^c(\delta)\}] \\
&\overset{(ii)}{\leq} \mathbb{E}_{\delta,\xi}[\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{1}\{e_t=1\}\mathbb{1}\{\mathcal{E}(\delta)\}(V_{\mathcal{P}^\star,r^\star,1}^{\pi^\star,d_1} - V_{\mathcal{P}^\star,r^\star,1}^{\pi_{t-1},d_1}] + H(T\delta + \sum_{t=1}^{T}\xi_t)\} \\
&\leq H\mathbb{E}_{\delta,\xi}[\sum_{t=1}^{T}\mathbb{1}\{\mathcal{E}(\delta)\}(V_{\mathcal{P}^\star,r^\star,1}^{\pi^\star,d_1} - V_{\mathcal{P}^\star,r^\star,1}^{\pi_{t-1},d_1}] + H(1 + \sum_{t=1}^{T}t^{-1/4})\} \\
&\overset{(iii)}{\leq} H\,\underbrace{\mathbb{E}_{\delta,\xi}[\sum_{t=1}^{T}\mathbb{1}\{\mathcal{E}(\delta)\}(V_{\mathcal{P}^\star,r^\star,1}^{\pi^\star,d_1} - V_{\mathcal{P}^\star,r^\star,1}^{\pi_t,d_1}]}_{(A)} + HT^{3/4} + 2H\}
\end{aligned}$$

where (i) follows from the optimality of $\pi^\star$ and $\|V_{\mathcal{P},r^\star}^{\pi}\|_\infty \leq 1$, (ii) follows from $\tilde{\pi}_t$ and π_{t-1} agreeing on the event $e_t = 1$ and Lemma .1 and (iii) follows from an index shift and $\|V_{\mathcal{P},r^\star}^{\pi}\|_\infty \leq 1$. Finally, we can leverage the pseudo-regret result of

Lemma A.5 to bound term (A),

$$\begin{aligned}(A) &= \mathbb{E}_{\delta,\xi}[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{E}(\delta)\}(V^{\pi^\star,d_1}_{\mathcal{P}^\star,r^\star,1} - V^{\pi_t,d_1}_{\mathcal{P}^\star,r^\star,1}] \\ &\lesssim H^2 d^2 |\mathcal{A}| \frac{\sqrt{T}\log^2(4TH|\Phi||\Psi|/\delta)}{\xi_T} \\ &\lesssim H^2 d^2 |\mathcal{A}| T^{3/4} \log^2(4TH|\Phi||\Psi|),\end{aligned}$$

and hence, conclude the proof. □

B SELECTING GOOD REPRESENTATIONS

In this section, we show how the expressiveness assumption 4.1 and the constrained optimization objective (Algorithm 1, Line 17) play together to guarantee the selection of good representations. The analysis builds on the sub-linear regret result for the behavior policies (Lemma A.5) provided in the previous section.

Selecting $\alpha^\star$-approximate representations We start by introducing an important result provided by Huang et al. [2022], which states that the average occupancy distribution induced by any sequence of deterministic policies that achieve low regret eventually provides a good approximation of the occupancy distribution of the optimal policy (assuming the optimal policy is unique).

Let us denote $\Pi^\star$ as the set of all optimal (deterministic) policies and $\Pi^\star_h(s)$ as the set of all optimal actions in state $s \in \mathcal{S}$ and time step $h \in [H]$. Then, we construct $\tilde{\pi}^\star_t := \{\tilde{\pi}^\star_{t,h}\}_{h\in[H]}$, where for each $h \in [H]$,

$$\tilde{\pi}^\star_{t,h}(s) = \begin{cases} \pi_{t,h}(s) & \text{if } \pi_{t,h}(s) \in \Pi^\star_h(s) \\ Select(\Pi^\star_h(s)) & \text{otherwise} \end{cases},$$

where $Select$ is a function which returns a fixed element of some set and π_t is the behavior policy of algorithm 1 at episode $t \in \mathbb{N}$. We define the mixture occupancy distribution of our constructed optimal policies $\tilde{\pi}^\star_t$ as

$$\tilde{\gamma}^\star_{t,h}(s,a) = \frac{1}{t}\sum_{i=0}^{t-1} d^{\tilde{\pi}^\star_i}_{\mathcal{P}^\star,h}(s,a).$$

Note that $\tilde{\gamma}^\star_{t,h} \equiv d^{\pi^\star}_{\mathcal{P}^\star,h}$ whenever there exists a unique optimal policy (Assumption 3.2).

Theorem B.1. *([Huang et al., 2022], Theorem 4.7) Suppose that we run algorithm 1. Then, for all $h \in [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$\sum_{i=1}^{t} d^{\pi_i}_{\mathcal{P}^\star,h}(s,a) \geq \sum_{i=1}^{t} d^{\tilde{\pi}^\star_i}_{\mathcal{P}^\star,h}(s,a) - \frac{1}{\Delta_{\text{min}}}\left(\sum_{i=1}^{t} V^{\tilde{\pi}^\star_i,d_1}_{\mathcal{P}^\star,r^\star,1} - V^{\pi_i,d_1}_{\mathcal{P}^\star,r^\star,1}\right).$$

Corollary B.1. *Suppose that we run algorithm 1 and assumption 3.3 (minimal sub-optimality gap) hold. Then, Theorem B.1 implies, for all $h \in [H]$, $t \in \mathbb{N}$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$\tilde{\gamma}^\star_{t,h}(s,a) \leq \gamma_{t,h}(s,a) + \frac{\mathcal{R}(t)}{t\Delta_{\text{min}}}.$$

We can leverage the above corollary to show that, whenever there exists a unique optimal policy, the MLE oracle converges uniformly on the optimal occupancy distribution, provided that the distribution is well defined for all states. Subsequently, for any given α, there must exist an episode after which algorithm 1 will only select representations that are $\alpha^\star$-approximate.

Lemma B.1. *(Selecting $\alpha^\star$-representations) Fix any $\alpha > 0$. Assume there exists an increasing sub-linear function g such that $\mathcal{R}(t) \leq g(t)$ for all $t \in \mathbb{N}$. Suppose we run algorithm 1 and assumptions 3.2 (unique optimal policy), 3.4 (minimal optimal occupancy) and 3.3 (minimal sub-optimality gap) hold. Then, given that the event $\mathcal{E}$ occurs, there exists an episode τ_α, such that for all $t \geq \tau_\alpha$ and $h \in [H]$, the learned feature maps $\hat{\phi}_{t,h}$ are $\alpha^\star$-approximate, where*

$$\tau_\alpha := \min\{t | t > \frac{1}{\alpha}\left(\frac{g(t)}{\Delta_{\text{min}} d^\star_{\text{min}}} + \frac{|\mathcal{A}|}{\xi_t}\sqrt{2t\log(4t|\Phi||\Psi|H/\delta)}\right)\}.$$

Proof. Let $t \in \mathbb{N}$ be arbitrary. Then, for all $h \in [H]$,

$$\begin{aligned}
\mathbb{E}_{(s,a)\sim d^{\pi^\star}_{\mathcal{P}^\star,h}}[f_{t,h}(s,a)] &= \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d^{\pi^\star}_{\mathcal{P}^\star,h}(s,a) f_{t,h}(s,a) \\
&\overset{(i)}{\leq} \sum_{(s,a):d^{\pi^\star}_{\mathcal{P}^\star,h}(s,a)>0} (\gamma_{t,h}(s,a) + \frac{\mathcal{R}(t)}{t\Delta_{\text{min}}}) f_{t,h}(s,a) \\
&\overset{(ii)}{\leq} \mathbb{E}_{(s,a)\sim\gamma_{t,h}}[f_{t,h}(s,a)] + \frac{g(t)}{t\Delta_{\text{min}}} \sum_{(s,a):d^{\pi^\star}_{\mathcal{P}^\star,h}(s,a)>0} \frac{d^{\pi^\star}_{\mathcal{P}^\star,h}(s,a)}{d^\star_{\text{min}}} \\
&\overset{(iii)}{\leq} \sqrt{\frac{|\mathcal{A}|^2}{\xi_t^2}\mathbb{E}_{(s,a)\sim\rho'_{t,h}}[f_{t,h}(s,a)^2]} + \frac{g(t)}{t\Delta_{\text{min}} d^\star_{\text{min}}} \\
&\overset{(iv)}{\leq} \frac{|\mathcal{A}|}{\xi_t}\sqrt{\zeta_t} + \frac{g(t)}{t\Delta_{\text{min}} d^\star_{\text{min}}},
\end{aligned}$$

where (i) is by Corollary B.1, (ii) follows from $\|f_{t,h}\|_\infty \leq 1$, (iii) is by importance sampling and Jensen's inequality and (iv) follows from the event $\mathcal{E}$. Since g is sub-linear, the above quantity decreases with t. Solving for t yields the result. □

Selecting non-redundant UniSOFT representations Although we can now be sure to select $\alpha^\star$-approximate representations, we still need to ensure that the UniSOFT loss in equation 3 will lead to Algorithm 1 actually selecting UniSOFT representations. Hence, we want to relate the eigenvalues of the expected covariance matrix of the optimal policy, which tells us if a feature map is UniSOFT, to the eigenvalues of the sample covariance matrix, which are captured by the UniSOFT loss in equation 3. We define the following good events:

$$\begin{aligned}
\mathcal{F}_1(\delta) &:= \{\forall t \in \mathbb{N}, h \in [H], \phi \in |\Phi| : \\
&\qquad \Sigma_{t,h} \succcurlyeq t\Sigma^\star_{t,h} + \lambda_t I - 2I\sum_{i=1}^{t}\xi_{i-1} - \Delta_{\text{min}}^{-1} g(t) I - 18I\sqrt{t\log(6tdH|\Phi|/\delta)}\} \\
\mathcal{F}_2(\delta) &:= \{\forall t \in \mathbb{N}, h \in [H], \phi \in |\Phi| : \\
&\qquad \Sigma_{t,h} \preccurlyeq t\Sigma^\star_{t,h} + \lambda_t I + 2I\sum_{i=1}^{t}\xi_{i-1} + \Delta_{\text{min}}^{-1} g(t) I + 18I\sqrt{t\log(6tdH|\Phi|/\delta)}\},
\end{aligned}$$

where $\Sigma^\star_{t,h} = \mathbb{E}_{(s,a)\sim\tilde{\gamma}^\star_{t,h}}[\phi(s,a)\phi(s,a)^T]$, $\Sigma_{t,h} = \sum_{(s,a)\in\mathcal{D}_{t,h}} \phi_h(s,a)\phi_h(s,a)^T$ and g is any increasing function such that $\mathcal{R}(t) \leq g(t)$ for all $t \in \mathbb{N}$. In addition, define $\mathcal{F}(\delta) := \mathcal{F}_1(\delta/2) \cap \mathcal{F}_2(\delta/2)$.

Lemma B.2. *(Eigenvalue bounds) Assume that there exists an increasing sub-linear function g such that $\mathcal{R}(t) \leq g(t)$ for all $t \in \mathbb{N}$. Assume that we run Algorithm 1 and that assumption 3.3 (minimal sub-optimality gap) holds. Then, with probability at least $1-\delta$, the event $\mathcal{F}(\delta)$ occurs.*

Proof. Recall that algorithm 1 produces for each time step $h \in [H]$, one trajectory τ_h, in any episode t. Furthermore, for each trajectory τ_h, we only use the transition at the time step h to construct the empirical covariance matrix $\hat{\Sigma}_{t,h}$.

Upper bound: Let $\tau^{(t,h)}$ denote the trajectory produced by rolling in with the behavior policy π_{t-1} and then taking action according to $\tilde{\pi}_{t,h}$ in episode $t \in \mathbb{N}$ for time step $h \in [H]$. Additionally, (s^τ_h, a^τ_h) denotes a state-action pair at time step $h \in [H]$ of trajectory τ. We define the set of trajectories of length $h \in [H]$ under which the (deterministic) behavior policy in some episode $t \in \mathbb{N}$ is optimal:

$$\Gamma^\star_{h,t} = \{\tau \in \Gamma_h : \pi_{t-1,i}(s^\tau_i) = \tilde{\pi}^\star_{t-1,i}(s^\tau_i) \text{ for } i = 1, ..., h\},$$

where Γ_h denotes the set of trajectories of length $h \in [H]$. The distribution over trajectories induced by any (deterministic) policy π is given by

$$\rho^\pi_h = d_1(s_1)\mathbb{1}[a_1 = \pi_1(s_1)]\mathcal{P}^\star_1(s_2|a_1, s_1)...\mathcal{P}^\star_{h-1}(s_h|a_{h-1}, s_{h-1})\mathbb{1}[a_h = \pi_h(s_h)].$$

Additionally, for any (deterministic) policy π, we denote

$$\rho_h^{\pi,\xi} = d_1(s_1)\mathbb{1}[a_1 = \pi_1(s_1)]\mathcal{P}_1^\star(s_2|a_1, s_1)...\mathcal{P}_{h-1}^\star(s_h|a_{h-1}, s_{h-1})\tilde{\pi}_{h,\xi}(a_h|s_h),$$

where $\tilde{\pi}_{h,\xi}(a_h|s_h) = \frac{\mathbb{1}[e=0]}{|\mathcal{A}|} + \mathbb{1}[e = 1]\mathbb{1}[a_h = \pi_h(s_h)]$ and $e \sim \text{Ber}(1 - \xi)$, as the trajectory distribution induced by algorithm 1. Finally, we denote $\tau_{1:h}^{(t,h)}$ as the trajectory $\tau^{(t,h)}$ cut off at time step $h \in [H]$. Then,

$$\begin{aligned}
\Sigma_{h,t} - \lambda_t I &= \sum_{i=1}^{t} \phi(s_h^{\tau^{(i,h+1)}}, a_h^{\tau^{(i,h+1)}})\phi(s_h^{\tau^{(i,h+1)}}, a_h^{\tau^{(i,h+1)}})^T \\
&\preccurlyeq \underbrace{\sum_{i=1}^{t} \mathbb{1}[e_i = 1]\mathbb{1}[\tau_{1:h}^{(i,h+1)} \in \Gamma_{h,i}^\star]\phi(s_h^{\tau^{(i,h+1)}}, a_h^{\tau^{(i,h+1)}})\phi(s_h^{\tau^{(i,h+1)}}, a_h^{\tau^{(i,h+1)}})^T}_{(A)} \\
&\quad + \underbrace{\sum_{i=1}^{t} \mathbb{1}[\tau_{1:h}^{(i,h+1)} \notin \Gamma_{h,i}^\star]\phi(s_h^{\tau^{(i,h+1)}}, a_h^{\tau^{(i,h+1)}})\phi(s_h^{\tau^{(i,h+1)}}, a_h^{\tau^{(i,h+1)}})^T}_{(B)} \\
&\quad + \underbrace{\sum_{i=1}^{t} \mathbb{1}[e_i = 0]\phi(s_h^{\tau^{(i,h+1)}}, a_h^{\tau^{(i,h+1)}})\phi(s_h^{\tau^{(i,h+1)}}, a_h^{\tau^{(i,h+1)}})^T}_{(C)}
\end{aligned}$$

Then, with probability of at least $1 - \delta/6$, for all $t \in \mathbb{N}$ and all $h \in [H]$ and $\phi \in \Phi$,

$$\begin{aligned}
(A) &= \sum_{i=1}^{t} \mathbb{1}[e_i = 1]\mathbb{1}[\tau_{1:h}^{(i,h+1)} \in \Gamma_{h,i}^\star]\phi(s_h^{\tau^{(i,h+1)}}, a_h^{\tau^{(i,h+1)}})\phi(s_h^{\tau^{(i,h+1)}}, a_h^{\tau^{(i,h+1)}})^T \\
&= \sum_{i=1}^{t} \mathbb{1}[e_i = 1]\mathbb{1}[\tau_{1:h}^{(i,h+1)} \in \Gamma_{h,i}^\star]\phi(s_h^{\tau^{(i,h+1)}}, \tilde{\pi}_{t-1,h}^\star(s_h^{\tau^{(i,h+1)}}))\phi(s_h^{\tau^{(i,h+1)}}, \tilde{\pi}_{t-1,h}^\star(s_h^{\tau^{(i,h+1)}}))^T \\
&= \sum_{i=1}^{t} \mathbb{E}_{\tau\sim\rho_h^{\pi_{i-1},\xi_{i-1}}}[\mathbb{1}[e = 1]\mathbb{1}[\tau \in \Gamma_{h,i}^\star]\phi(s_h^\tau, \tilde{\pi}_{t-1,h}^\star(s_h^\tau))\phi(s_h^\tau, \tilde{\pi}_{t-1,h}^\star(s_h^\tau))^T] \\
&\quad + \sum_{i=1}^{t} \mathbb{1}[e_i = 1]\mathbb{1}[\tau_{1:h}^{(i,h+1)} \in \Gamma_{h,i}^\star]\phi(s_h^{\tau^{(i,h+1)}}, \tilde{\pi}_{t-1,h}^\star(s_h^{\tau^{(i,h+1)}}))\phi(s_h^{\tau^{(i,h+1)}}, \tilde{\pi}_{t-1,h}^\star(s_h^{\tau^{(i,h+1)}}))^T \\
&\quad - \sum_{i=1}^{t} \mathbb{E}_{\tau\sim\rho_h^{\pi_{i-i},\xi_{i-1}}}[\mathbb{1}[e = 1]\mathbb{1}[\tau \in \Gamma_{h,i}^\star]\phi(s_h^\tau, \tilde{\pi}_{t-1,h}^\star(s_h^\tau))\phi(s_h^\tau, \tilde{\pi}_{t-1,h}^\star(s_h^\tau))^T] \\
&\overset{(i)}{\preccurlyeq} \underbrace{\sum_{i=1}^{t} \mathbb{E}_{\tau\sim\rho_h^{\pi_{i-1},\xi_{i-1}}}[\mathbb{1}[e = 1]\mathbb{1}[\tau \in \Gamma_{h,i}^\star]\phi(s_h^\tau, \tilde{\pi}_{t-1,h}^\star(s_h^\tau))\phi(s_h^\tau, \tilde{\pi}_{t-1,h}^\star(s_h^\tau))^T]}_{(A1)} + 8I\sqrt{t\log(6tdH|\Phi|/\delta)},
\end{aligned}$$

where (i) follows from $\|\phi_h\|_2 \leq 1$ and Proposition G.1 in combination with a union bound over all episodes $t \in \mathbb{N}$, time steps $h \in [H]$ and feature maps $\phi \in \Phi$. Further,

$$\begin{aligned}
(A1) &= \sum_{i=1}^{t} \mathbb{E}_{\tau\sim\rho_h^{\pi_{i-1},\xi_{i-1}}}[\mathbb{1}[e = 1]\mathbb{1}[\tau \in \Gamma_{h,i}^\star]\phi(s_h^\tau, \tilde{\pi}_{t-1,h}^\star(s_h^\tau))\phi(s_h^\tau, \tilde{\pi}_{t-1,h}^\star(s_h^\tau))^T] \\
&\overset{(i)}{=} \sum_{i=1}^{t} \mathbb{E}_{\tau\sim\rho_h^{\pi_{i-1}}}[\mathbb{1}[\tau \in \Gamma_{h,i}^\star]\phi(s_h^\tau, \tilde{\pi}_{t-1,h}^\star(s_h^\tau))\phi(s_h^\tau, \tilde{\pi}_{t-1,h}^\star(s_h^\tau))^T] \\
&\overset{(ii)}{\preccurlyeq} \sum_{i=1}^{t} \mathbb{E}_{\tau\sim\rho_h^{\tilde{\pi}_{i-1}^\star}}[\phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^T] \\
&\overset{(iii)}{=} t\mathbb{E}_{(s,a)\sim\tilde{\gamma}_{t,h}^\star}[\phi(s, a)\phi(s, a)^T],
\end{aligned}$$

where (i) follows from $\rho_h^{\pi,\xi}$ and ρ_h^{π} agreeing on the event $e = 1$ and (ii) follows from the occupancy distributions $d_{\mathcal{P}^\star,h}^{\tilde{\pi}_t^\star}$ and $d_{\mathcal{P}^\star,h}^{\pi_t}$ agreeing on $\Gamma_{h,t}^\star$ and for (iii) recall that $\tilde{\gamma}_{t,h}^\star(s,a) = \frac{1}{t}\sum_{i=o}^{t-1} d_{\mathcal{P}^\star,h}^{\tilde{\pi}_t^\star}(s,a)$. Similarly, with probability of at least $1 - \delta/6$, for all $t \in \mathbb{N}$ and all $h \in [H], \phi \in \Phi$,

$$\begin{aligned}
(B) &= \sum_{i=1}^{t} \mathbb{1}[\tau_{1:h}^{(i,h+1)} \notin \Gamma_{h,i}^\star]\phi(s_h^{\tau^{(i,h+1)}}, a_h^{\tau^{(i,h+1)}})\phi(s_h^{\tau^{(i,h+1)}}, a_h^{\tau^{(i,h+1)}})^T \\
&\overset{(i)}{\preccurlyeq} \sum_{i=1}^{t} \mathbb{E}_{\tau\sim\rho_h^{\pi_{i-1},\xi_{i-1}}}[\mathbb{1}[\tau \notin \Gamma_{h,i}^\star]\phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^T] + 8I\sqrt{t\log(6tdH|\Phi|/\delta)} \\
&\overset{(ii)}{\preccurlyeq} \underbrace{I\sum_{i=1}^{t} \mathbb{E}_{\tau\sim\rho_h^{\pi_{i-1},\xi_{i-1}}}[\mathbb{1}[\tau \notin \Gamma_{h,i}^\star]]}_{(B1)} + 8I\sqrt{t\log(6tdH|\Phi|/\delta)},
\end{aligned}$$

where (i) follows, similarly to before, from Proposition G.1 in combination with an union bound and (ii) is by $\|\phi_h\|_2 \leq 1$. Further,

$$\begin{aligned}
(B1) &= I\sum_{i=1}^{t} \mathbb{E}_{\tau\sim\rho_h^{\pi_{i-1},\xi_{i-1}}}[\mathbb{1}[\tau \notin \Gamma_{h,i}^\star]] \\
&= I\sum_{i=1}^{t} \mathbb{E}_{\tau\sim\rho_h^{\pi_{i-1},\xi_{i-1}}}[\mathbb{1}[e=1]\mathbb{1}[\tau \notin \Gamma_{h,i}^\star]] + \mathbb{E}_{\tau\sim\rho_h^{\pi_{i-1},\xi_{i-1}}}[\mathbb{1}[e=0]\mathbb{1}[\tau \notin \Gamma_{h,i}^\star]] \\
&\overset{(i)}{\preccurlyeq} I\sum_{i=1}^{t} \mathbb{E}_{\tau\sim\rho_h^{\pi_{i-1}}}[\mathbb{1}[\tau \notin \Gamma_{h,i}^\star]] + I\sum_{i=1}^{t} \mathbb{E}_{e\sim\text{Ber}(1-\xi_{i-1})}[\mathbb{1}[e=0]] \\
&\overset{(ii)}{\preccurlyeq} I\sum_{i=1}^{t}\sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim d_{\mathcal{P}^\star,h}^{\pi_{i-1}}}[\mathbb{1}[a \notin \Pi_h^\star(s)]] + I\sum_{i=1}^{t} \xi_{i-1} \\
&\preccurlyeq I\sum_{i=1}^{t}\sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim d_{\mathcal{P}^\star,h}^{\pi_{i-1}}}[\mathbb{1}[\Delta_h(s,a) \geq \Delta_{\text{min}}]] + I\sum_{i=1}^{t} \xi_{i-1} \\
&\preccurlyeq I\frac{1}{\Delta_{\text{min}}}\sum_{i=1}^{t}\sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim d_{\mathcal{P}^\star,h}^{\pi_{i-1}}}[\Delta_h(s,a)] + I\sum_{i=1}^{t} \xi_{i-1} \\
&\overset{(iii)}{=} \frac{\mathcal{R}(t)}{\Delta_{\text{min}}}I + I\sum_{i=1}^{t} \xi_{i-1},
\end{aligned}$$

where (i) follows from $\rho_h^{\pi,\xi}$ and ρ_h^{π} agreeing on the event $e = 1$, (ii) follows from the definition of $\tilde{\pi}_{t,h}^\star$ and (iii) follows from Lemma G.2. Finally, with probability at least $1 - \delta/6$, for all $t \in \mathbb{N}$ and $h \in [H]$,

$$\begin{aligned}
(C) &= \sum_{i=1}^{t} \mathbb{1}[e_i = 0]\phi_h(s_h^{\tau^{(i,h+1)}}, a_h^{\tau^{(i,h+1)}})\phi_h(s_h^{\tau^{(i,h+1)}}, a_h^{\tau^{(i,h+1)}})^T \\
&\overset{(i)}{\preccurlyeq} I\sum_{i=1}^{t} \mathbb{1}[e_i = 0] \\
&\overset{(ii)}{\preccurlyeq} I(\sum_{i=1}^{t} \mathbb{E}_{e\sim\text{Ber}(1-\xi_{i-1})}[\mathbb{1}[e=0]] + \sqrt{t\log(6tH/\delta)}) \\
&\preccurlyeq I\sum_{i=1}^{t} \xi_{i-1} + \sqrt{t\log(6tH/\delta)},
\end{aligned}$$

where (i) follows from $\|\phi_h\|_2 \leq 1$ and (ii) is by Hoeffding's inequality with a union bound over episodes and time steps.

Lower bound: The lower bound is easily derived by similar arguments. With probability at least $1 - \delta/2$, for all $t \in \mathbb{N}$, and all $\phi \in \Phi$, $h \in [H]$:

$$\begin{aligned}\Sigma_{h,t} - \lambda_t I &\succcurlyeq (A)\\ &\succcurlyeq (A1) - 8I\sqrt{t\log(6tdH|\Phi|/\delta)}\\ &\succcurlyeq t\mathbb{E}_{(s,a)\sim\tilde{\gamma}^\star_{t,h}}[\phi(s,a)\phi(s,a)^T] - (B) - (C) - 8I\sqrt{t\log(6tdH|\Phi|/\delta)}.\end{aligned}$$

We conclude the proof by performing an union bound over the results for the lower and upper bound. □

By the lower bound of the previous result, we immediately obtain the following:

Lemma B.3. *Consider a feature map $\phi \in \Phi$ that is non-redundant. Assume there exists an increasing sub-linear function g such that $\mathcal{R}(t) \leq g(t)$ for all $t \in \mathbb{N}$. Suppose assumptions 3.2 (unique optimal policy) and 3.3 (minimal sub-optimality gap) holds. Then, given that the event $\mathcal{F}$ occurs, there exists a constant τ_{inv} such that, for all $t \geq \tau_{\text{inv}}$, $h \in [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$\|\phi_h(s,a)\|_{\Sigma_{t,h}^{-1}} \leq (t\lambda_{\text{min}}(\Sigma^\star_{t,h}) + \lambda_t - 2\sum_{i=1}^{t}\xi_{i-1} - \Delta_{\text{min}}^{-1}g(t) - 18\sqrt{t\log(6tdH|\Phi|/\delta)})^{-1/2}.$$

Proof. Let τ_{inv} be large enough so that

$$t\lambda_{\text{min}}(\Sigma^\star_{t,h}) + \lambda_t > 2\sum_{i=1}^{t}\xi_{i-1} + \Delta_{\text{min}}^{-1}g(t) + 18\sqrt{t\log(6tdH|\Phi|/\delta)}$$

holds. Then, for all $t \geq \tau_{\text{inv}}$, $h \in [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$

$$\begin{aligned}\|\phi_h(s,a)\|_{\hat{\Sigma}_{t,h}^{-1}} &= (\phi_h(s,a)^T\hat{\Sigma}_{t,h}^{-1}\phi_h(s,a))^{1/2}\\ &\overset{(i)}{\leq} (\lambda_{\text{min}}(\hat{\Sigma}_{t,h}^{-1})\phi_h(s,a)^T\phi_h(s,a))^{1/2}\\ &\leq \lambda_{\text{min}}(\hat{\Sigma}_{t,h})^{-1/2},\end{aligned}$$

where (i) follows from the symmetry of the covariance matrix. We conclude the proof by substituting $\hat{\Sigma}_{t,h}$ with the lower bound provided by the event $\mathcal{F}_1$. □

Note that $\lambda_{\text{min}}(\Sigma^\star_{t,h}) > 0$ holds whenever there exists a unique optimal policy and the feature map is UniSOFT. The final lemma of this section shows that algorithm 1 is guaranteed to eventually select only good representations.

Lemma B.4. *(Selecting non-redundant UniSOFT representation) Fix any $\alpha > 0$. Assume that there exists an increasing sublinear function g such that $\mathcal{R}(t) \leq g(t)$ for all $t \in \mathbb{N}$. Suppose we run algorithm 1 and assumptions 3.2 (unique optimal policy), 4.1 (expressiveness) and 3.3 (minimal sub-optimality gap) hold. Additionally, if $\alpha < 1$, suppose that assumption 3.4 (minimal optimal occupancy) holds. Then, given that events $\mathcal{E}(\delta)$ and $\mathcal{F}(\delta)$ occur, there exists an episode $\tau_{\text{unisoft}} \geq \tau_\alpha$ such that for all subsequent episodes $t \geq \tau_{\text{unisoft}}$ and time steps $h \in [H]$, the learned feature maps $\hat{\phi}_{t,h}$ are UniSOFT and non-redundant, where*

$$\tau_{\text{unisoft}} := \min\{t | t > \left(\frac{2}{\lambda^\star_\alpha}(\Delta_{\text{min}}^{-1}\mathcal{R}(t) + 2\sum_{i=1}^{t}\xi_{i-1} + 18\sqrt{t\log(6tdH|\Phi|/\delta)}) \vee \tau_\alpha\right)\}.$$

Proof. Note that, by Lemma B.1, given that $\mathcal{E}$ occurs, there exists an episode τ_α such that for all $t \geq \tau_\alpha$ and $h \in [H]$, the learned features $\hat{\phi}_{t,h}$ are $\alpha^\star$-approximate.

Let $\Phi^{\text{unisoft}} \subseteq \Phi$ denote the set that contains only non-redundant UniSOFT feature mappings. By Lemma B.2, given that the event $\mathcal{F}$ occurs, for all $t \in \mathbb{N}$, $h \in [H]$, $\phi \in \Phi \setminus \Phi^{\text{unisoft}}$ and $\phi^{\text{unisoft}} \in \Phi^{\text{unisoft}}$,

$$\lambda_{\text{min}}(\Sigma_{t,h}(\phi^{\text{unisoft}}) - \lambda_t I) \geq t\lambda^\star(\phi) - 2\sum_{i=1}^{t}\xi_{i-1} - \Delta_{\text{min}}^{-1}g(t) - 18\sqrt{t\log(6tdH|\Phi|/\delta)},$$

$$\lambda_{\text{min}}(\Sigma_{t,h}(\phi) - \lambda_t I) \leq 2\sum_{i=1}^{t}\xi_{i-1} + \Delta_{\text{min}}^{-1}g(t) + 18\sqrt{t\log(6tdH|\Phi|/\delta)},$$

where $\Sigma_{t,h}(\phi) = \sum_{(s,a)\in\mathcal{D}_{t,h}} \phi_h(s,a)\phi_h(s,a)^T$. Let us denote $\Phi_\alpha \times \Psi_\alpha \subseteq \Phi \times \Psi$ as the set of $\alpha^\star$-approximate representations. Additionally, denote

$$\Phi_\alpha^{\text{unisoft}} \times \Psi_\alpha^{\text{unisoft}} = (\Phi_\alpha \times \Psi_\alpha) \cap \left(\Phi^{\text{unisoft}} \times \Psi\right),$$

as the set containing all $\alpha^\star$-approximate representations such that the feature map is non-redundant and UniSOFT, which is non-empty by assumption 4.1. A non-redundant UniSOFT representation ϕ^{unisoft} is selected in episode $t \geq \tau_\alpha$ if for all $\tilde{\alpha} \leq \alpha$,

$$\max_{\phi^{\text{unisoft}} \in \Phi_{\tilde{\alpha}}^{\text{unisoft}}} \lambda_{\text{min}}(\Sigma_{t,h}(\phi^{\text{unisoft}}) - \lambda_t I) > \max_{\phi \in \Phi_{\tilde{\alpha}} \setminus \Phi_{\tilde{\alpha}}^{\text{unisoft}}} \lambda_{\text{min}}(\Sigma_{t,h}(\phi) - \lambda_t I),$$

or equivalently,

$$t\lambda_\alpha^\star(\phi^{\text{unisoft}}) > 2(2\sum_{i=1}^{t} \xi_{i-1} + \Delta_{\text{min}}^{-1} g(t) + 18\sqrt{t\log(6tdH|\Phi|/\delta)}),$$

where $\lambda_\alpha^\star := \min_{\tilde{\alpha}\leq\alpha} \max_{\phi^{\text{unisoft}} \in \Phi_{\tilde{\alpha}}^{\text{unisoft}}} \lambda^\star(\phi^{\text{unisoft}})$. □

C IMPROVED PSEUDO-REGRET WITH GOOD REPRESENTATIONS

In this section, we show how we can use good representations to improve the pseudo-regret result A.5 provided in Section A. Subsequently, we can provide an improved expected regret result.

On a high level, we show that the bonus terms provide an almost optimistic estimate for the expected sub-optimality gaps incurred by the behavior policies of algorithm 1. We can then exploit the UniSOFT property of the good representations that we are guaranteed to select, as shown in the previous section, to show uniformly decreasing confidence intervals. Let us start by providing two results that are adapted from Cheng et al. [2023], which show that the bonus term can be used to provide a trajectory-wise uncertainty measure for the model estimation error over the occupancy distribution of the behavior policies.

Lemma C.1. *(Value difference of transition operators) For all $t \in \mathbb{N}$, any policy π, state $s \in \mathcal{S}$, time step $h \in [H]$ and set of reward function $\{r_h\}_{h=1}^H$ such that $r_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ and $\sum_{h=1}^H r_h \leq 1$,*

$$|V_{\mathcal{P}^\star,r,h}^\pi(s) - V_{\hat{\mathcal{P}}_t,r,h}^\pi(s)| \leq V_{\mathcal{P},f_t,h}^\pi(s),$$

where $\mathcal{P} \in \{\hat{\mathcal{P}}_t, \mathcal{P}^\star\}$.

Proof. We give a proof by induction. For $h = H+1$ and any $s \in \mathcal{S}$, we have $|V_{\mathcal{P}^\star,r,H+1}^\pi(s) - V_{\hat{\mathcal{P}}_t,r,H+1}^\pi(s)| = 0 = V_{\mathcal{P},f_t,H+1}^\pi$ for $\mathcal{P} \in \{\hat{\mathcal{P}}_t, \mathcal{P}^\star\}$. Suppose the induction hypothesis, $|V_{\mathcal{P}^\star,r,h+1}^\pi(s) - V_{\hat{\mathcal{P}}_t,r,h+1}^\pi(s)| \leq V_{\mathcal{P},f_t,h+1}^\pi(s)$ for $\mathcal{P} \in \{\hat{\mathcal{P}}_t, \mathcal{P}^\star\}$ and any $s \in \mathcal{S}$. Then, for any $h \in [H]$ and $s \in \mathcal{S}$,

$$\begin{aligned}
&|V_{\mathcal{P}^\star,r,h}^\pi(s) - V_{\hat{\mathcal{P}}_t,r,h}^\pi(s)| \\
&\leq \mathbb{E}_{a\sim\pi(\cdot|s)}[|Q_{\mathcal{P}^\star,r,h}^\pi(s,a) - Q_{\hat{\mathcal{P}}_t,r,h}^\pi(s,a)|] \\
&= \mathbb{E}_{a\sim\pi(\cdot|s)}[|\mathcal{P}_h^\star V_{\mathcal{P}^\star,r,h+1}^\pi(s,a) - \hat{\mathcal{P}}_{t,h} V_{\hat{\mathcal{P}}_t,r,h+1}^\pi(s,a)|] =: (A).
\end{aligned}$$

Then, the first claim ($\mathcal{P} = \hat{\mathcal{P}}_t$) follows from:

$$\begin{aligned}
(A) &= \mathbb{E}_{a\sim\pi(\cdot|s)}[|\hat{\mathcal{P}}_{t,h}(V_{\mathcal{P}^\star,r,h+1}^\pi - V_{\hat{\mathcal{P}}_t,r,h+1}^\pi)(s,a) + (\mathcal{P}_h^\star - \hat{\mathcal{P}}_{t,h})V_{\mathcal{P}^\star,r,h+1}^\pi(s,a)|] \\
&\overset{(i)}{\leq} \mathbb{E}_{a\sim\pi(\cdot|s)}[\hat{\mathcal{P}}_{t,h} V_{\hat{\mathcal{P}}_t,f_t,h+1}^\pi(s,a) + f_{t,h}(s,a)] \\
&= V_{\hat{\mathcal{P}}_t,f_t,h}^\pi(s),
\end{aligned}$$

where (i) follows from the induction hypothesis and $\|V_{\mathcal{P},r,h}^\pi\|_\infty \leq 1$. The second claim ($\mathcal{P} = \mathcal{P}^\star$) follows from:

$$\begin{aligned}
(A) &= \mathbb{E}_{a\sim\pi(\cdot|s)}[|\mathcal{P}_h^\star(V_{\mathcal{P}^\star,r,h+1}^\pi - V_{\hat{\mathcal{P}}_t,r,h+1}^\pi)(s,a) + (\mathcal{P}_h^\star - \hat{\mathcal{P}}_{t,h})V_{\hat{\mathcal{P}}_t,r,h+1}^\pi(s,a)|] \\
&\overset{(i)}{\leq} \mathbb{E}_{a\sim\pi(\cdot|s)}[\mathcal{P}_h^\star V_{\mathcal{P}^\star,f_t,h+1}^\pi(s,a) + f_{t,h}(s,a)] \\
&= V_{\mathcal{P}^\star,f_t,h}^\pi(s),
\end{aligned}$$

where (i) follows from the induction hypothesis and $\|V_{\mathcal{P},r,h}^\pi\|_\infty \leq 1$. □

Lemma C.2. *(Uncertainty bounded model estimation error) Given that the event $\mathcal{E}$ occurs, we have for all $t \in \mathbb{N}$ and any policy π,*

$$V^{\pi,d_1}_{\mathcal{P}^\star,f_t,1} \leq 2H\sqrt{\frac{|\mathcal{A}|}{\xi_t}}\zeta_t + 2HV^{\pi,d_1}_{\hat{\mathcal{P}}_t,\hat{b}_t,1}, \text{ and}$$

$$V^{\pi,d_1}_{\hat{\mathcal{P}}_t,f_t,1} \leq \sqrt{\frac{|\mathcal{A}|}{\xi_t}}\zeta_t + V^{\pi,d_1}_{\hat{\mathcal{P}}_t,\hat{b}_t,1}.$$

Proof. For all $h > 1$,

$$\begin{aligned}\mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{\mathcal{P}}_t;h}}[f_{t,h}(s,a)] &\overset{(i)}{\leq} \mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{\mathcal{P}}_t;h-1}}[\min\{1, \alpha_t\|\hat{\phi}_{t,h-1}(s,a)\|_{\Sigma^{-1}_{\rho_{t,h-1},\hat{\phi}_{t,h-1}}}\}] \\ &\overset{(ii)}{\leq} \mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{\mathcal{P}}_t;h-1}}[\hat{b}_{t,h-1}(s,a)],\end{aligned}$$

where (i) is by Lemma A.3 and $\|f_{t,h}\|_\infty \leq 1$ and (ii) follows from the event $\mathcal{E}$. Additionally, by Lemma A.3, we have,

$$\mathbb{E}_{(s,a)\sim d^{\pi}_{\hat{\mathcal{P}}_t;1}}[f_{t,1}(s,a)] \leq \sqrt{\frac{|\mathcal{A}|}{\xi_t}}\zeta_t,$$

which gives the second claim. Additionally,

$$\begin{aligned}V^{\pi,d_1}_{\mathcal{P}^\star,f_t,1} &\leq V^{\pi,d_1}_{\hat{\mathcal{P}}_t,f_t,1} + H|\frac{1}{H}V^{\pi,d_1}_{\mathcal{P}^\star,f_t,1} - \frac{1}{H}V^{\pi,d_1}_{\hat{\mathcal{P}}_t,f_t,1}| \\ &\overset{(i)}{\leq} V^{\pi,d_1}_{\hat{\mathcal{P}}_t,f_t,1} + HV^{\pi,d_1}_{\hat{\mathcal{P}}_t,f_t,1} \\ &\overset{(ii)}{\leq} 2H\sqrt{\frac{|\mathcal{A}|}{\xi_t}}\zeta_t + 2HV^{\pi,d_1}_{\hat{\mathcal{P}}_t,\hat{b}_t,1},\end{aligned}$$

where (i) is by Lemma C.1 and (ii) follows from the second claim. □

Next, we introduce an optimism result similar to that of Lemma A.4, which holds locally on the state-occupancy distribution of the behavior policies.

Lemma C.3. *(Almost Local Optimism) Given that the event $\mathcal{E}$ occurs, for all $t \in \mathbb{N}$ and $h \in [H]$,*

$$\mathbb{E}_{s\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[V^{\pi^\star}_{\mathcal{P}^\star,r^\star,h}(s) - V^{\pi^\star}_{\hat{\mathcal{P}}_t,r^\star+\hat{b}_t,h}(s)] \leq 2H\sqrt{\frac{|A|}{\xi_t}}\zeta_t + 2HV^{\pi^b_t,d_1}_{\hat{\mathcal{P}}_t,\hat{b}_t,1},$$

where $\pi^b_t = \arg\max_{\pi\in\Pi} V^{\pi,d_1}_{\hat{\mathcal{P}}_t,\hat{b}_t,1}$.

Proof. We have for all $h \in [H]$:

$$\begin{aligned}\mathbb{E}_{s\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[V^{\pi^\star}_{\mathcal{P}^\star,r^\star,h}(s) - V^{\pi^\star}_{\hat{\mathcal{P}}_t,r^\star+\hat{b}_t,h}(s)] &\leq \mathbb{E}_{s\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[V^{\pi^\star}_{\mathcal{P}^\star,r^\star,h}(s) - V^{\pi^\star}_{\hat{\mathcal{P}}_t,r^\star,h}(s)] \\ &\leq \mathbb{E}_{s\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[|V^{\pi^\star}_{\mathcal{P}^\star,r^\star,h}(s) - V^{\pi^\star}_{\hat{\mathcal{P}}_t,r^\star,h}(s)|] \\ &\overset{(i)}{\leq} \mathbb{E}_{s\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[V^{\pi^\star}_{\mathcal{P}^\star,f_t,h}(s)] =: (A),\end{aligned}$$

where (i) follows from Lemma C.1. Now, let $f^{(h:)}_{t,i}(s,a) = f_{t,i}(s,a)\mathbb{1}\{i \geq h\}$ and $\pi^{(h:)^\star}_{t,i}(a|s) = \pi_t(a|s)\mathbb{1}\{i < h\} + \pi^\star(a|s)\mathbb{1}\{i \geq h\}$ for any $h \in [H]$. Then,

$$(A) = V^{\pi^{(h:)^\star}_t,d_1}_{\mathcal{P}^\star,f^{(h:)}_t,1} \overset{(i)}{\leq} V^{\pi^{(h:)^\star}_t,d_1}_{\mathcal{P}^\star,f_t,1} \overset{(ii)}{\leq} 2H\sqrt{\frac{|A|}{\xi_t}}\zeta_t + 2HV^{\pi^{(h:)^\star}_t,d_1}_{\hat{\mathcal{P}}_t,\hat{b}_t,1},$$

where (i) follows from $f_{t,h} \geq 0$ being non-negative for all h and t and (ii) follows from Lemma C.2. Now, the claim follows by the definition of π^b_t. □

We continue by providing a local simulation lemma.

Lemma C.4. *For all $t \in \mathbb{N}$ and $h \in [H]$, we have*

$$\mathbb{E}_{s\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[V^{\pi_t}_{\hat{\mathcal{P}}_t,r^\star+b_{t,h},h}(s) - V^{\pi_t}_{\mathcal{P}^\star,r^\star,h}(s)] \leq 2HV^{\pi_t,d_1}_{\mathcal{P}^\star,\hat{b}_t+f_t,1}$$

Proof. We have,

$$\begin{aligned}
&\mathbb{E}_{s\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[V^{\pi_t}_{\hat{\mathcal{P}}_t,r^\star+\hat{b}_t,h}(s) - V^{\pi_t}_{\mathcal{P}^\star,r^\star,h}(s)]\\
&= \mathbb{E}_{(s,a)\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[Q^{\pi_t}_{\hat{\mathcal{P}}_t,r^\star+b_{t,h},h}(s,a) - Q^{\pi_t}_{\mathcal{P}^\star,r^\star,h}(s,a)]\\
&\leq \mathbb{E}_{(s,a)\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[\hat{b}_{h,t}(s,a)] + |\mathbb{E}_{(s,a)\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[(\hat{\mathcal{P}}_{t,h} - \mathcal{P}^\star_h)V^{\pi_t}_{\hat{\mathcal{P}}_t,r^\star+\hat{b}_t,h+1}(s,a)]|\\
&\qquad + \mathbb{E}_{(s,a)\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[\mathcal{P}^\star_h(V^{\pi_t}_{\hat{\mathcal{P}}_t,r^\star+\hat{b}_t,h+1} - V^{\pi_t}_{\mathcal{P}^\star,r^\star,h+1})(s,a)]\\
&\overset{(i)}{\leq} \mathbb{E}_{(s,a)\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[\hat{b}_{t,h}(s,a)] + 2H\mathbb{E}_{(s,a)\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[f_{t,h}(s,a)]\\
&\qquad + \mathbb{E}_{s\sim d^{\pi_t}_{\mathcal{P}^\star,h+1}}[V^{\pi_t}_{\hat{\mathcal{P}}_t,r^\star+\hat{b}_t,h+1}(s) - V^{\pi_t}_{\mathcal{P}^\star,r^\star,h+1}(s)],
\end{aligned}$$

where (i) follows from $\|V^\pi_{\mathcal{P},r^\star+\hat{b}_t}\|_\infty \leq 2H$. Unraveling the recursion gives the result. □

The previous four lemmata combined are enough to show that the bonus terms provide an almost optimistic estimate of the expected sub-optimality gaps incurred by the behavior policies of algorithm 1.

Lemma C.5. *(Sub-optimality gap to bonus) Given that the event $\mathcal{E}$ occurs, we have for all $t \in \mathbb{N}$ and $h \in [H]$,*

$$\mathbb{E}_{(s,a)\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[\Delta_h(s,a)] \leq 10H^2(\sqrt{\frac{|A|}{\xi_t}}\zeta_t + V^{\pi^b_t,d_1}_{\hat{\mathcal{P}}_t,\hat{b}_t,1}),$$

where $\pi^b_t = \arg\max_{\pi\in\Pi} V^{\pi,d_1}_{\hat{\mathcal{P}}_t,\hat{b}_t,1}$.

Proof. We have for all $h \in [H]$ and $t \in \mathbb{N}$:

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[\Delta_h(s,a)]\\
&\overset{(i)}{\leq} \mathbb{E}_{s\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[V^{\pi^\star}_{\mathcal{P}^\star,r^\star,h}(s) - V^{\pi_t}_{\mathcal{P}^\star,r^\star,h}(s)]\\
&\overset{(ii)}{\leq} \mathbb{E}_{s\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[V^{\pi_t}_{\hat{\mathcal{P}}_t,r^\star+\hat{b}_t,h}(s) - V^{\pi_t}_{\mathcal{P}^\star,r^\star,h}(s) + V^{\pi^\star}_{\mathcal{P}^\star,r^\star,h}(s) - V^{\pi^\star}_{\hat{\mathcal{P}}_t,r^\star+\hat{b}_t,h}(s)]\\
&\overset{(iii)}{\leq} 2HV^{\pi_t,d_1}_{\mathcal{P}^\star,\hat{b}_t,1} + 2HV^{\pi_t,d_1}_{\mathcal{P}^\star,f_t,1} + \mathbb{E}_{s\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[V^{\pi^\star}_{\mathcal{P}^\star,r^\star,h}(s) - V^{\pi^\star}_{\hat{\mathcal{P}}_t,r^\star+\hat{b}_t,h}(s)]\\
&\overset{(iv)}{\leq} 2H\underbrace{V^{\pi_t,d_1}_{\mathcal{P}^\star,\hat{b}_t,1}}_{=:(A)} + 2H\underbrace{V^{\pi_t,d_1}_{\mathcal{P}^\star,f_t,1}}_{=:(B)} + 2H\sqrt{\frac{|A|}{\xi_t}}\zeta_t + 2HV^{\pi^b_t,d_1}_{\hat{\mathcal{P}}_t,\hat{b}_t,1},
\end{aligned}$$

where (i) follows from the optimality of $\pi^\star$, (ii) by the optimality of π_t, (iii) follows from the local simulation Lemma C.4 and (iv) follows from the local optimism Lemma C.3. Further,

$$\begin{aligned}
(A) &= V^{\pi_t,d_1}_{\mathcal{P}^\star,\hat{b}_t,1}\\
&\leq V^{\pi_t,d_1}_{\hat{\mathcal{P}}_t,\hat{b}_t,1} + H|\frac{1}{H}V^{\pi_t,d_1}_{\mathcal{P}^\star,\hat{b}_t,1} - \frac{1}{H}V^{\pi_t,d_1}_{\hat{\mathcal{P}}_t,\hat{b}_t,1}|\\
&\overset{(i)}{\leq} V^{\pi_t,d_1}_{\hat{\mathcal{P}}_t,\hat{b}_t,1} + HV^{\pi_t,d_1}_{\hat{\mathcal{P}}_t,f_t,1}\\
&\overset{(ii)}{\leq} V^{\pi_t,d_1}_{\hat{\mathcal{P}}_t,\hat{b}_t,1} + H(\sqrt{\frac{|A|}{\xi_t}}\zeta_t + V^{\pi_t,d_1}_{\hat{\mathcal{P}}_t,\hat{b}_t,1})\\
&\overset{(iii)}{\leq} 2HV^{\pi^b_t,d_1}_{\hat{\mathcal{P}}_t,\hat{b}_t,1} + H\sqrt{\frac{|A|}{\xi_t}}\zeta_t,
\end{aligned}$$

where (i) follows from Lemma C.1, (ii) follows from Lemma C.2 and (iii) by the optimality of π_t^b. Similarly,

$$\begin{aligned}(B) &= V^{\pi_t,d_1}_{\mathcal{P}^\star,f_t,1}\\ &\overset{(i)}{\leq} 2H(\sqrt{\frac{|A|}{\xi_t}}\zeta_t + V^{\pi_t,d_1}_{\hat{\mathcal{P}}_t,\hat{b}_t,1})\\ &\overset{(ii)}{\leq} 2H(\sqrt{\frac{|A|}{\xi_t}}\zeta_t + V^{\pi_t^b,d_1}_{\hat{\mathcal{P}}_t,\hat{b}_t,1}),\end{aligned}$$

where (i) follows from Lemma C.2 and (ii) follows from the optimality of π_t^b. Finally, we get:

$$\mathbb{E}_{(s,a)\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[\Delta_h(s,a)] \leq 10H^2(\sqrt{\frac{|A|}{\xi_t}}\zeta_t + V^{\pi_t^b,d_1}_{\hat{\mathcal{P}}_t,\hat{b}_t,1})$$

□

We can now leverage the fact that we eventually select only good representations, which leads to the following improved pseudo-regret bound.

Lemma C.6. *(Sub-linear pseudo-regret with UniSOFT representations) Let $\xi_t = t^{-1/3}$ and $\alpha > 0$. Suppose assumptions 3.1 (realizability), 3.2 (unique optimal policy), 3.3 (minimal sub-optimality gap) and 4.1 ($\alpha^\star$-expressive function space) hold. Additionally, if $\alpha < 1$, suppose that assumption 3.4 (minimal optimal occupancy) holds. Then, given that events $\mathcal{E}(\delta)$ and $\mathcal{F}(\delta)$ occur, there exists a constant τ, such that for all $T \geq \tau$, the behavior policies $\{\pi_t\}_{t\geq 1}$ learned by algorithm 1, enjoy sub-linear regret:*

$$\mathcal{R}(T) \lesssim \frac{\sqrt{\tau}}{\xi_\tau} + \frac{1}{\lambda^\star_{\text{max}}}H^3 d|\mathcal{A}|^{1/2}T^{2/3}\log(4T|\Phi||\Psi|H/\delta) \lesssim \tilde{O}(T^{2/3})$$

Proof. Let $\tau := \{\tau_{\text{unisoft}} \vee \tau_{\text{inv}}\}$. Let $t \geq \tau$ be arbitrary. Then, since the event $\mathcal{E}$ occurs by assumption, by Lemma C.5, we can bound the expected sub-optimality gaps for all $h \in [H]$,

$$\mathbb{E}_{(s,a)\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[\Delta_h(s,a)] \leq 10H^2(\sqrt{\frac{|\mathcal{A}|}{\xi_t}}\zeta_t + V^{\pi_t^b,d_1}_{\hat{\mathcal{P}}_t,\hat{b}_t,1}) := (A).$$

Further, according to Lemma A.5, in the event $\mathcal{E}$, $\mathcal{R}(t) \leq g(t) = \tilde{O}(\sqrt{t}\xi_t^{-1})$ with $\hat{\alpha}_t = \tilde{O}(\xi_t^{-1/2})$. We note that if $\alpha = 1$, then all representations are $\alpha^\star$-approximate and hence we do not require assumption 3.4 (minimal optimal occupancy) to guarantee their selection in Lemma B.1. By Lemma B.4 and the events $\mathcal{F}$ and $\mathcal{E}$, for all $h \in [H]$, the learned feature maps $\hat{\phi}_{t,h}$ are non-redundant and UniSOFT. Then, by Lemma B.3 and the event $\mathcal{F}$,

$$\begin{aligned}V^{\pi_t^b,d_1}_{\hat{\mathcal{P}}_t,\hat{b}_t,1} &\leq \hat{\alpha}_t \sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim d^{\pi_t^b}_{\hat{\mathcal{P}}_t,h}}[\|\hat{\phi}_{t,h}(s,a)\|_{\hat{\Sigma}^{-1}_{t,h}}]\\ &\leq \frac{\hat{\alpha}_t H}{(\lambda^\star_{\text{max}}t + \lambda_t - \sum_{i=1}^{t}\xi_{i-1} - g(t)\Delta^{-1}_{\text{min}} - 18\sqrt{t\log(6tdH|\Phi|/\delta)})^{1/2}}\\ &\leq \tilde{O}(\frac{t^{1/6}}{t^{1/2}}) = \tilde{O}(t^{-1/3}).\end{aligned}$$

Since t was chosen arbitrarily, we get, for all $T \geq \tau$:

$$\begin{aligned}\mathcal{R}(T) &= \sum_{t=1}^{\tau}(V^{\pi^\star,d_1}_{\mathcal{P}^\star,r^\star,1} - V^{\pi_t,d_1}_{\mathcal{P}^\star,r^\star,1}) + \sum_{t=\tau}^{T}\sum_{h=1}^{H}\mathbb{E}_{(s,a)\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[\Delta(s,a)]\\ &\overset{(i)}{\lesssim} \tilde{O}(\frac{\sqrt{\tau}}{\xi_\tau}) + \frac{1}{\lambda^\star_{\text{max}}}H^3 d|\mathcal{A}|^{1/2}T^{2/3}\log(4T|\Phi||\Psi|H/\delta),\end{aligned}$$

where (i) follows from the pseudo-regret bound without UniSOFT representations given in Lemma A.5. □

Theorem C.1 (Instance-dependent regret with UniSOFT representations, Theorem 4.1)**.** *Let $\xi_t = t^{-1/3}$ and $\alpha \in (0,1]$. Suppose assumptions 3.1 (realizability), 3.3 (minimal sub-optimality gap), 4.1 ($\alpha^\star$-expressive function space) and 3.2 (unique optimal policy) hold. Additionally, if $\alpha < 1$, suppose that assumption 3.4 (minimal optimal occupancy) holds. Then for any $T \in \mathbb{N}$, UniSREP-UCB (Algorithm 1) satisfies the following:*

$$\mathbb{E}[\tilde{\mathcal{R}}(T)] = \tilde{O}\left(H^3 d^2 |\mathcal{A}| (\tau_{\text{good}} \wedge T)^{5/6} + \frac{1}{\lambda^\star_{\text{max}}} H^4 d |\mathcal{A}|^{1/2} T^{2/3}\right),$$

where

$$\begin{aligned}\tau_{\text{good}} &\lesssim \{\kappa_3^6 \cdot \log^{12}(\kappa_3 \cdot \kappa_2) \vee \kappa_1^6 \cdot \log^{12}(\kappa_1 \cdot \kappa_2)\} \\ &\lesssim \frac{H^{12} d^{12} |\mathcal{A}|^6}{(\Delta_{\text{min}}\{\alpha d^\star_{\text{min}} \wedge \lambda^\star_{\text{max}}\})^6} \cdot \log^{12}(TH^3 d^{3/2} |\mathcal{A}||\Phi||\Psi|),\end{aligned}$$

with $\kappa_1 = \frac{H^2 d^2 |\mathcal{A}|}{\alpha \Delta_{\text{min}} d^\star_{\text{min}}}$, $\kappa_2 = TH|\Phi||\Psi|$, $\kappa_3 = \frac{H^2 d^2 |\mathcal{A}|}{\lambda^\star_{\text{max}} \Delta_{\text{min}}}$ and $\lambda^\star_{\text{max}} = \min_{\tilde{\alpha} \le \alpha} \max_{\phi \in \Phi^{\text{unisoft}}_{\tilde{\alpha}}} \lambda^\star(\phi)$.

Proof. Let $\tau_{\text{good}} := \{\tau_{\text{unisoft}} \vee \tau_{\text{inv}}\}$ and $T \ge \tau_{\text{good}}$ be given and fixed. Choose $\delta = T^{-1}$. Recall that Algorithm 1 explores for H time steps, for each $h \in [H]$ and episode t, by rolling into time step $h-1$ with policy π_{t-1}, taking actions according to $\tilde{\pi}_{t,h-1}$ and $\tilde{\pi}_{t,h}$ and finally, rolling out to time step H with policy π_{t-1}. Let us denote $\tilde{V}^{d_1}_{t,h}$ as the cumulative expected reward obtained by Algorithm 1 in episode t and time step h. Then,

$$\begin{aligned}&\mathbb{E}_{\delta,\xi}[\tilde{\mathcal{R}}(T)] \\ &= \mathbb{E}_{\delta,\xi}[\sum_{t=1}^{T}\sum_{h=1}^{H}(V^{\pi^\star,d_1}_{\mathcal{P}^\star,r^\star,1} - \tilde{V}_{t,h})] \\ &\le \mathbb{E}_{\delta,\xi}[\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{1}\{e_t = 1\}\mathbb{1}\{\mathcal{E}(\delta)\}\mathbb{1}\{\mathcal{F}(\delta)\}(V^{\pi^\star,d_1}_{\mathcal{P}^\star,r^\star,1} - \tilde{V}_{t,h})] \\ &\quad + \mathbb{E}_{\delta,\xi}[\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{1}\{e_t = 0\}(V^{\pi^\star,d_1}_{\mathcal{P}^\star,r^\star,1} - \tilde{V}_{t,h})] + \mathbb{E}_{\delta,\xi}[\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{1}\{\mathcal{E}^c(\delta)\}(V^{\pi^\star,d_1}_{\mathcal{P}^\star,r^\star,1} - \tilde{V}_{t,h})] \\ &\quad + \mathbb{E}_{\delta,\xi}[\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{1}\{\mathcal{F}^c(\delta)\}(V^{\pi^\star,d_1}_{\mathcal{P}^\star,r^\star,1} - \tilde{V}_{t,h})]\end{aligned}$$

$$\begin{aligned}&\overset{(i)}{\le} \mathbb{E}_{\delta,\xi}[\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{1}\{e_t = 1\}\mathbb{1}\{\mathcal{E}(\delta)\}\mathbb{1}\{\mathcal{F}(\delta)\}(V^{\pi^\star,d_1}_{\mathcal{P}^\star,r^\star,1} - \tilde{V}_{t,h})] \\ &\quad + \mathbb{E}_{\delta,\xi}[\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{1}\{e_t = 0\} + \mathbb{1}\{\mathcal{E}^c(\delta)\} + \mathbb{1}\{\mathcal{F}^c(\delta)\}] \\ &\overset{(ii)}{\le} \mathbb{E}_{\delta,\xi}[\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{1}\{e_t = 1\}\mathbb{1}\{\mathcal{E}(\delta)\}\mathbb{1}\{\mathcal{F}(\delta)\}(V^{\pi^\star,d_1}_{\mathcal{P}^\star,r^\star,1} - V^{\pi_{t-1},d_1}_{\mathcal{P}^\star,r^\star,1}] + H(2T\delta + \sum_{t=1}^{T}\xi_t)\} \\ &\le H\mathbb{E}_{\delta,\xi}[\sum_{t=1}^{T}\mathbb{1}\{\mathcal{E}(\delta)\}\mathbb{1}\{\mathcal{F}(\delta)\}(V^{\pi^\star,d_1}_{\mathcal{P}^\star,r^\star,1} - V^{\pi_{t-1},d_1}_{\mathcal{P}^\star,r^\star,1}] + H(2 + \sum_{t=1}^{T}t^{-1/3})\} \\ &\overset{(iii)}{\le} H\,\underbrace{\mathbb{E}_{\delta,\xi}[\sum_{t=1}^{T}\mathbb{1}\{\mathcal{E}(\delta)\}\mathbb{1}\{\mathcal{F}(\delta)\}(V^{\pi^\star,d_1}_{\mathcal{P}^\star,r^\star,1} - V^{\pi_t,d_1}_{\mathcal{P}^\star,r^\star,1}]}_{(A)} + \frac{3}{2}HT^{2/3} + 3H\},\end{aligned}$$

where (i) follows from $\|V^\pi_{\mathcal{P},r^\star}\|_\infty \le 1$, (ii) follows from $\tilde{\pi}_t$ and π_{t-1} agreeing on the event $e_t = 1$, Lemma .1 and Lemma B.3 and (iii) follows from an index shift and $\|V^\pi_{\mathcal{P},r^\star}\|_\infty \le 1$. Now, we can leverage the pseudo-regret result of Lemma C.6

to bound term (A),

$$
\begin{aligned}
(A) &= \mathbb{E}_{\delta,\xi}[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{E}(\delta)\}\mathbb{1}\{\mathcal{F}(\delta)\}(V_{\mathcal{P}^\star,r^\star,1}^{\pi^\star,d_1} - V_{\mathcal{P}^\star,r^\star,1}^{\pi_t,d_1}] \\
&\lesssim \frac{\sqrt{\tau_{\text{good}}}}{\xi_{\tau_{\text{good}}}} + \frac{1}{\lambda^\star_{\text{max}}} H^3 d|\mathcal{A}|^{1/2} T^{2/3} \log(4T|\Phi||\Psi|H/\delta) \\
&\lesssim \tau_{\text{good}}^{5/6} + \tilde{O}(T^{2/3}).
\end{aligned}
$$

Substituting τ_{good} with the sufficient condition in Lemma D.3 with $\gamma = 3$ and using $T \gtrsim a \log^n(ab)$ as a sufficient condition for $T \geq a \log^n(bT)$, concludes the proof. □

D CONSTANT PSEUDO-REGRET WITH GOOD REPRESENTATIONS

In this section, we provide a constant pseudo-regret result that translates the uniform convergence of the confidence intervals to the expected sub-optimality gaps. We start by providing a sufficient condition that makes a deterministic policy optimal.

Lemma D.1. *Let π be any deterministic policy. Whenever,*

$$
\mathbb{E}_{(s,a)\sim d^\pi_{\mathcal{P}^\star,h}}[\Delta_h(s,a)] < d^\star_{\text{min}} \Delta_{\text{min}}
$$

holds for all $h \in [H]$ simultaneously, there exists an optimal policy $\tilde{\pi}^\star \in \Pi^\star$, such that, for all $h \in [H]$,

$$
d^{\tilde{\pi}^\star}_{\mathcal{P}^\star,h} \equiv d^\pi_{\mathcal{P}^\star,h}.
$$

Proof. We give a proof by induction. For $h = 1$ we have,

$$
\begin{aligned}
\mathbb{E}_{(s,a)\sim d^\pi_{\mathcal{P}^\star,1}}[\Delta_1(s,a)] &= \mathbb{E}_{s\sim d_1}[\Delta_1(s,\pi_1(s))] \\
&= \sum_{s\in\mathcal{S}} d_1(s)\Delta_1(s,\pi_1(s)) \\
&\geq d^\star_{\text{min}} \sum_{s:d_1(s)>0} \Delta_1(s,\pi_1(s))
\end{aligned}
$$

Hence, for all $s \in \mathcal{S}$ such that $d_1(s) > 0$,

$$
\Delta_1(s,\pi_1(s)) < \Delta_{\text{min}},
$$

and therefore, $\pi_1(s) \in \Pi^\star_1(s)$ for all $s \in \mathcal{S}$ such that $d_1(s) > 0$. Equivalently, there exits a policy $\tilde{\pi}^\star \in \Pi^\star$ such that,

$$
d^{\tilde{\pi}^\star}_{\mathcal{P}^\star,1} \equiv d^\pi_{\mathcal{P}^\star,1}.
$$

Suppose the induction hypothesis that for any time step $h \in [H]$ there exists an optimal policy $\tilde{\pi}^\star \in \Pi^\star$ such that, $d^{\tilde{\pi}^\star}_{\mathcal{P}^\star,h} \equiv d^\pi_{\mathcal{P}^\star,h}$ holds. Then, for an arbitrary $h \in [H]$,

$$
\begin{aligned}
\mathbb{E}_{(s,a)\sim d^\pi_{\mathcal{P}^\star,h+1}}[\Delta_{h+1}(s,a)] &\overset{(i)}{=} \mathbb{E}_{s\sim d^{\tilde{\pi}^\star}_{\mathcal{P}^\star,h+1}}[\Delta_{h+1}(s,\pi_{h+1}(s))] \\
&= \sum_{s\in\mathcal{S}} d^{\tilde{\pi}^\star}_{\mathcal{P}^\star,h+1}(s)\Delta_{h+1}(s,\pi_{h+1}(s)) \\
&\geq d^\star_{\text{min}} \sum_{s:d^{\pi^\star}_{\mathcal{P}^\star,h+1}(s)} \Delta_{h+1}(s,\pi_{h+1}(s)),
\end{aligned}
$$

where (i) follows from the induction hypothesis. Therefore, for all $s \in \mathcal{S}$ such that $d^{\tilde{\pi}^\star}_{\mathcal{P}^\star,h+1}(s) > 0$, we have $\pi_{h+1}(s) \in \Pi^\star_{h+1}(s)$. □

Lemma D.2. *(Constant pseudo-regret with UniSOFT representations) Let $\alpha \in (0,1]$, $\gamma \in (2,\infty)$ and $\xi_t = t^{-1/\gamma}$. Suppose assumptions 3.1 (realizability), 3.2 (unique optimal policy), 3.3 (minimal sub-optimality gap), 3.4 (minimal optimal occupancy) and 4.1 ($\alpha^\star$-expressive function space) hold. Then, given that events $\mathcal{E}(\delta)$ and $\mathcal{F}(\delta)$ occur, there exists a constant $\tau^\star$, after which the behavior policies $\{\pi_t\}_{t\geq 1}$ learned by algorithm 1, incur no additional regret and hence, for all $T \in \mathbb{N}$:*

$$\mathcal{R}(T) \lesssim \mathcal{R}(\tau^\star) = O(1)$$

Proof. Let t be arbitrary and large enough. Then, since the event $\mathcal{E}$ occurs by assumption, by Lemma C.5, we can bound the expected sub-optimality gaps for all $h \in [H]$,

$$\mathbb{E}_{(s,a)\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[\Delta_h(s,a)] \leq 10H^2(\sqrt{\frac{|\mathcal{A}|}{\xi_t}}\zeta_t + V^{\pi^b_t,d_1}_{\hat{\mathcal{P}}_t,\hat{b}_t,1}) := (A).$$

Further, according to Lemma A.5, in the event $\mathcal{E}$, $\mathcal{R}(t) \leq g(t) = \tilde{O}(\sqrt{t}\xi_t^{-1}) = O(t^{\frac{2+\gamma}{2\gamma}})$ with $\hat{\alpha}_t = \tilde{O}(\xi_t^{-1/2}) = \tilde{O}(t^{\frac{1}{2\gamma}})$. By Lemma B.4 and the events $\mathcal{F}$ and $\mathcal{E}$, for all $h \in [H]$, the learned feature maps $\hat{\phi}_{t,h}$ are non-redundant and UniSOFT. Then, by Lemma B.3, $\gamma > 2$ and the event $\mathcal{F}$,

$$\begin{aligned}
V^{\pi^b_t,d_1}_{\hat{\mathcal{P}}_t,\hat{b}_t,1} &\leq \hat{\alpha}_t \sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim d^{\pi^b_t}_{\hat{\mathcal{P}}_t,h}}[\|\hat{\phi}_{t,h}(s,a)\|_{\hat{\Sigma}^{-1}_{t,h}}] \\
&\leq \frac{\hat{\alpha}_t H}{(\lambda^\star_{\text{max}} t + \lambda_t - \sum_{i=1}^{t}\xi_{i-1} - g(t)\Delta^{-1}_{\text{min}} - 18\sqrt{t\log(6tdH|\Phi|/\delta)})^{1/2}} \\
&\leq \tilde{O}(\frac{t^{\frac{1}{2\gamma}}}{t^{1/2}}) = \tilde{O}(t^{-\frac{1}{2}(1-\frac{1}{\gamma})}) \xrightarrow[t\to\infty]{} 0,
\end{aligned}$$

Additionally, we have

$$\sqrt{\frac{|\mathcal{A}|}{\xi_t}}\zeta_t = \tilde{O}(t^{-\frac{1}{2}(1-\frac{1}{\gamma})}) \xrightarrow[t\to\infty]{} 0$$

Hence, there must exist an episode $\tau^\star$ such that

$$\mathbb{E}_{(s,a)\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[\Delta_h(s,a)] < \Delta_{\text{min}} d^\star_{\text{min}}$$

for all $t \geq \tau^\star$. Then by Lemma D.1, we get:

$$\begin{aligned}
\mathcal{R}(T) &\leq \sum_{t=1}^{\infty}\sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[\Delta(s,a)] \\
&\leq \sum_{t=1}^{\tau^\star}\sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[\Delta(s,a)] = \mathcal{R}(\tau^\star) = O(1).
\end{aligned}$$

□

Theorem 4.2 (Expected regret with UniSOFT)**.** *Let $\alpha > 0$, $\gamma \in (2,4]$ and $\xi_t = t^{-1/\gamma}$. Suppose assumptions 3.1 (realizability), 3.2 (unique optimal policy), 3.3 (minimal sub-optimality gap), 3.4 (minimal optimal occupancy) and 4.1 ($\alpha^\star$-expressive function space) hold. Then for any $T \in \mathbb{N}$, there exists a constant $\tau^\star$ such that* UNISREP-UCB *(Algorithm 1) satisfies*

$$\mathbb{E}[\tilde{\mathcal{R}}(T)] = \tilde{O}\left(H^3 d^2 |\mathcal{A}| (\tau^\star \wedge T)^{1/2+1/\gamma} + HT^{\frac{\gamma-1}{\gamma}}\right),$$

where $\tau^\star = \tilde{O}\left(\left(\frac{H^2 d^2 |\mathcal{A}|}{\alpha\lambda^\star_{\text{max}}(\Delta_{\text{min}} d^\star_{\text{min}})^2}\right)^{\frac{2\gamma}{\gamma-2}}\right)$.

Proof. Let T be given and fixed. Choose $\delta = \frac{1}{T}$. Then

$$\begin{aligned}
&\mathbb{E}_{\delta,\xi}[\tilde{\mathcal{R}}(T)] \\
&\overset{(i)}{\leq} H\mathbb{E}_{\delta,\xi}[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{E}(\delta)\}\mathbb{1}\{\mathcal{F}(\delta)\}(V^{\pi^\star,d_1}_{\mathcal{P}^\star,r^\star,1} - V^{\pi_{t-1},d_1}_{\mathcal{P}^\star,r^\star,1}] + H(2T\delta + \sum_{t=1}^{T} \xi_t)\} \\
&\overset{(ii)}{\lesssim} H(\tau^\star)^{1/2+1/\gamma} + H\sum_{t=1}^{T} t^{-1/\gamma} + 2H \\
&\lesssim H(\tau^\star)^{1/2+1/\gamma} + HT^{\frac{\gamma-1}{\gamma}} + 4H + 2H
\end{aligned}$$

where the details of (i) can be found in the proof of Theorem 4.1, (ii) follows from the constant pseudo-regret result of Lemma D.2. We conclude the proof by substituting $\tau^\star$ with the sufficient condition provided in Lemma D.3 and using $T \gtrsim a\log^n(ab)$ as a sufficient condition for $T \geq a\log^n(bT)$. □

Lemma D.3. *(Critical episodes) Let $\alpha \in (0, 1]$, $\gamma \in (2, 4]$ and $\xi_t = t^{-1/\gamma}$. Suppose assumptions 3.1 (realizability), 3.2 (unique optimal policy), 3.3 (minimal sub-optimality gap), 3.4 (minimal optimal occupancy) and 4.1 ($\alpha^\star$-expressive function space) hold. Suppose that we run algorithm 1. Then, given that events $\mathcal{E}(\delta)$ and $\mathcal{F}(\delta)$ occur:*

(1) *all non-$\alpha^\star$-approximate representations are eliminated after at most*

$$\tau_\alpha \lesssim \kappa_1^m \cdot \log^{2m}(\kappa_1 \cdot \kappa_2)$$

(2) *all redundant and non-UniSOFT representations are eliminated after at most*

$$\tau_{\text{good}} \lesssim \{\kappa_3^m \cdot \log^{2m}(\kappa_3 \cdot \kappa_2) \vee \tau_\alpha\}$$

(3) *the behavior policy π_t is optimal after at most*

$$\tau^\star \lesssim \{\kappa_4^{m'} \cdot \log^{m'}(\kappa_4 \cdot \kappa_2) \vee \tau_{\text{good}}\}$$

episodes, where $\kappa_1 = \frac{H^2 d^2 |\mathcal{A}|}{\alpha\Delta_{\min} d^\star_{\min}}$, $\kappa_2 = H|\Phi||\Psi|/\delta$, $\kappa_3 = \frac{H^2 d^2 |\mathcal{A}|}{\lambda^\star_{\max}\Delta_{\min}}$, $\kappa_4 = \frac{H^6 d^2 |\mathcal{A}|}{(\Delta_{\min} d^\star_{\min})^2 \lambda^\star_{\max}}$, $m = \frac{2\gamma}{\gamma-2}$ *and* $m' = \frac{\gamma}{\gamma-1}$.

Proof. By Lemma A.5, for all $t \in \mathbb{N}$,

$$\begin{aligned}
\mathcal{R}(t) &\leq c_3 H^2 d^2 |\mathcal{A}| \frac{\sqrt{t}\log^2(4tH|\Phi||\Psi|/\delta)}{\xi_t}, \\
\hat{\alpha}_t &= \sqrt{4t\zeta_t \frac{|\mathcal{A}|}{\xi_t} + \lambda_t d},
\end{aligned}$$

where c_3 is some universal constant. In the following, we will use $t \geq 3a\log(ab)$ as a sufficient condition for $t \geq a\log(bt)$ with reasonable values for a and b and $t > 0$. See Lemma 20 in Papini et al. [2021a] for details. In particular, by substituting t with $u = a^{\frac{1}{n}} t^{\frac{1}{mn}}$, we get that for any $n \geq 1$ and $m \geq 1$:

$$t > (mn)^n a^m (3\log(ab))^{mn} \Rightarrow t^{\frac{1}{m}} > a\log^n(bt). \tag{4}$$

We divide the analysis in four parts, where in each part we derive a sufficient condition for $\tau^\star$.
Part 1. $\tau^\star$ must satisfy the $\alpha^\star$-selection criteria in Lemma B.1.

$$\begin{aligned}
t &> \frac{1}{\alpha}(\frac{\mathcal{R}(t)}{\Delta_{\min} d^\star_{\min}} + \frac{|\mathcal{A}|}{\xi_t}\sqrt{2t\log(4tH|\Phi||\Psi|/\delta)}) \\
t &> \frac{1}{\alpha}(\frac{c_3 H^2 d^2 |\mathcal{A}| t^{(1/2+1/\gamma)} \log^2(4tH|\Phi||\Psi|)}{\Delta_{\min} d^\star_{\min}} + |\mathcal{A}| t^{(1/2+1/\gamma)}\sqrt{2\log(4t|\Phi||\Psi|H/\delta)}) \\
t &> t^{\frac{\gamma+2}{2\gamma}} \cdot c_3 2 \underbrace{\frac{H^2 d^2 |\mathcal{A}|}{\alpha\Delta_{\min} d^\star_{\min}}}_{\kappa_1} \cdot \log^2(t \cdot 4 \underbrace{H|\Phi||\Psi|/\delta}_{\kappa_2}) \\
t &\overset{(i)}{>} (2m)^2 (2c_3\kappa_1)^m 3^{2m} \log^{2m}(\kappa_1 \cdot 4\kappa_2) := \bar{\kappa}_1,
\end{aligned}$$

where (i) follows from the condition 4 with $m = \frac{2\gamma}{\gamma-2}$. We gain statement (1), by taking $\tau_\alpha = \bar{\kappa}_1$.

Part 2. $\tau^\star$ must satisfy the UniSOFT-selection criteria in Lemma B.4.

$$
\begin{aligned}
t &> \frac{2}{\lambda^\star_{\text{max}}}(\Delta_{\text{min}}^{-1}\mathcal{R}(t) + 2\sum_{i=1}^{t}\xi_{i-1} + 18\sqrt{t\log(6tdH|\Phi|/\delta)}) \\
t &> \frac{2}{\lambda^\star_{\text{max}}}(\frac{c_3H^2d^2|\mathcal{A}|t^{1/2+1/\gamma}\log^2(4tH|\Phi||\Psi|)}{\Delta_{\text{min}}} + 2\frac{\gamma}{\gamma-1}t^{1-1/\gamma} + 18\sqrt{t\log(6tdH|\Phi|/\delta)}) \\
t &\overset{(i)}{>} t^{\frac{2+\gamma}{2\gamma}} \cdot c_3 22\underbrace{\frac{H^2d^2|\mathcal{A}|}{\lambda^\star_{\text{max}}\Delta_{\text{min}}}}_{\kappa_3} \cdot \log^2(t\cdot 6\underbrace{dH|\Phi||\Psi|/\delta}_{\kappa_2}) \\
t &\overset{(ii)}{>} (2m)^2(22c_3\kappa_3)^m 3^{2m}\log^{2m}(\kappa_3\cdot 6\kappa_2) := \bar{\kappa}_2,
\end{aligned}
$$

where (i) follows from $\gamma \leq 4$ and (ii) follows from the condition 4 with $m = \frac{2\gamma}{\gamma-2}$.

Part 3. $\tau^\star$ must satisfy the invertibility condition from Lemma B.3.

$$
t > \frac{\mathcal{R}(t)\Delta_{\text{min}}^{-1} + 2\sum_{i=1}^{t}\xi_{i-1} + 18\sqrt{t\log(6tdH|\Phi|/\delta)}}{\lambda^\star_{\text{max}}},
$$

Note that the condition is fulfilled if $t \geq \bar{\kappa}_2$. By taking, $\tau_{\text{good}} := \max\{\bar{\kappa}_1, \bar{\kappa}_2\}$, we gain statement (2).

Part 4. First note we can upper bound,

$$
\begin{aligned}
\hat{\alpha}_t &= 5\sqrt{4t\zeta_t\frac{|\mathcal{A}|}{\xi_t} + \lambda_t d} \\
&= 5\sqrt{8\log(4|\Phi||\Psi|Ht/\delta)\frac{|\mathcal{A}|}{\xi_t} + c_1d^2\log(4tH|\Phi|/\delta)} \\
&\leq 5\sqrt{8c_1d^2|\mathcal{A}|t^{\frac{1}{\gamma}}\log(4|\Phi||\Psi|Ht/\delta)} \\
&\leq 5dt^{\frac{1}{2\gamma}}\sqrt{8|\mathcal{A}|c_1\log(4|\Phi||\Psi|Ht/\delta)}.
\end{aligned}
$$

For now we assume that $t \geq \bar{\kappa}_2$. Then,

$$
\begin{aligned}
\Delta_{\text{min}}d^\star_{\text{min}} &> 20H^2(\frac{\hat{\alpha}_tH}{(\lambda^\star_{\text{max}}t + \lambda_t - 2\sum_{i=1}^{t}\xi_{i-1} - \mathcal{R}(t)\Delta_{\text{min}}^{-1} - 18\sqrt{t\log(6tdH|\Phi|/\delta)})^{1/2}} + \sqrt{\frac{|A|}{\xi_t}\zeta_t}) \\
\Delta_{\text{min}}d^\star_{\text{min}} &\overset{(i)}{>} 20H^2(\frac{\hat{\alpha}_tH}{(\frac{3}{2}\lambda^\star_{\text{max}}t)^{1/2}} + \sqrt{2|A|t^{\frac{1}{\gamma}-1}\log(4t|\Phi||\Psi|H/\delta)}) \\
\Delta_{\text{min}}d^\star_{\text{min}} &> t^{-\frac{1}{2}(1-\frac{1}{\gamma})}\cdot 150\sqrt{c_1}\frac{H^3d|\mathcal{A}|^{1/2}}{(\lambda^\star_{\text{max}})^{1/2}} \cdot \sqrt{\log(t\cdot 4|\Phi||\Psi|H/\delta)},
\end{aligned}
$$

where (i) follows from $t \geq \bar{\kappa}_1$. After rearranging, we get:

$$
\begin{aligned}
t^{\frac{1}{2}(1-\frac{1}{\gamma})} &> 150\sqrt{c_1}\frac{H^3d|\mathcal{A}|^{1/2}}{\Delta_{\text{min}}d^\star_{\text{min}}(\lambda^\star_{\text{max}})^{1/2}} \cdot \log^{1/2}(t\cdot 4|\Phi||\Psi|H/\delta) \\
t^{(1-\frac{1}{\gamma})} &> 150^2c_1\underbrace{\frac{H^6d^2|\mathcal{A}|}{(\Delta_{\text{min}}d^\star_{\text{min}})^2\lambda^\star_{\text{max}}}}_{\kappa_4} \cdot \log(t\cdot 4\underbrace{|\Phi||\Psi|H/\delta}_{\kappa_2}) \\
t &\overset{(i)}{>} m450^{2m}(c_1\kappa_4)^m\log^m(\kappa_4\cdot 4\kappa_2) := \bar{\kappa}_3
\end{aligned}
$$

where (i) follows from condition 4 with $m = \frac{\gamma}{\gamma - 1}$. Finally, by taking

$$\tau^\star = \max\{\bar{\kappa}_1, \bar{\kappa}_2, \bar{\kappa}_3\}$$

we conclude. □

Theorem 4.3 (Constant Regret)**.** *Let $\alpha > 0$, $\delta \in (0,1)$ and $\xi_t = t^{-1/4}$. Suppose that the quantities Δ_{min} and $d^\star_{\text{min}}$ are known. Then, under the same assumptions as in Theorem 4.2, with probability at least $1 - 2\delta$,* UNISREP-UCB + *(Algorithm 1) satisfies the following:*

$$\tilde{\mathcal{R}}(T) \le T \wedge \tau^\star,$$

where[1] $\tau^\star = \tilde{O}\left(\frac{H^8 d^8 |\mathcal{A}|^4}{(\alpha\lambda^\star_{\text{max}})^4 (\Delta_{\text{min}} d^\star_{\text{min}})^8}\right)$.

Proof. We know, by the proof of Lemma C.6, that given that the events $\mathcal{E}$ and $\mathcal{F}$ hold, there exists an episode $\tau^\star$ such that, for all $t \ge \tau^\star$ and $h \in [H]$,

$$\begin{aligned} \mathbb{E}_{(s,a)\sim d^{\pi_t}_{\mathcal{P}^\star,h}}[\Delta_h(s,a)] &\le 10H^2\Big(\sqrt{\frac{|\mathcal{A}|}{\xi_t}}\zeta_t + V^{\pi^b_t, d_1}_{\mathcal{P}^\star, \hat{b}_t, 1}\Big) && (5) \\ &< \Delta_{\text{min}} d^\star_{\text{min}}. && (6) \end{aligned}$$

In particular, we know from Lemma D.1, that any deterministic policy satisfying the chain of inequalities above is optimal. Furthermore, the event $\mathcal{E}(\delta) \cap \mathcal{F}(\delta)$ holds with probability $1 - 2\delta$ by Lemma .1 and Lemma B.2. Hence, with probability at least $1 - 2\delta$, algorithm 1 returns an optimal policy after at most $\tau^\star$ episodes. □

E EXISTENCE OF GOOD REPRESENTATIONS

Note that within this section, we assume finiteness of the state space ($|\mathcal{S}| < \infty$) and that the transition operator has rank $\tilde{d}$ for all time steps, that is, $\text{rank}(\mathcal{P}^\star_h) = \tilde{d}$ for all $h \in [H]$. Recall that we denote $\mathcal{X}^\star_h := \{(s,a) \in \mathcal{S} \times \mathcal{A} | d^{\pi^\star}_{\mathcal{P}^\star,h}(s,a) > 0\}$ as the set of state-action pairs reachable by the optimal policy at time step $h \in [H]$. Similarly, we define $\mathcal{X}_h := \{(s,a) | \exists\pi : d^\pi_{\mathcal{P}^\star,h}(s,a) > 0\}$ as the set of state-action pairs reachable by any policy at time step $h \in [H]$. In the following, we provide the proofs for section 5.

Let us start by constructing a full rank factorization of $\mathcal{P}^\star_h$. Note that $\mathcal{P}^\star_h$ has rank $\tilde{d}$ by assumption and hence we can select $\tilde{d}$ columns of $\mathcal{P}^\star_h$ such that they form a basis for the column space of $\mathcal{P}^\star_h$. We collect them in a matrix $\Phi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times \tilde{d}}$, placing them in the same order as they appear in $\mathcal{P}^\star_h$. Now each column of $\mathcal{P}^\star_h$ can be expressed as a linear combination of the columns of Φ and we identify the row $\Phi_{sa,\cdot}$ with the feature $\phi(s,a)$. We denote $\Psi \in \mathbb{R}^{\tilde{d} \times |\mathcal{S}|}$ as the matrix uniquely determined by the coefficients in the linear combinations such that $\mathcal{P}^\star_h = \Phi\Psi$ and identify the column $\Psi_{\cdot,s}$ with $\mu(s)$. Then,

Lemma E.1. *Let $d \ge \tilde{d}$. Then, the following statements are equivalent:*

(1) span$\{\mathcal{P}^\star_h(\cdot|s,a) | (s,a) \in \mathcal{X}^\star_h\}$ = span$(\{\mathcal{P}^\star_h(\cdot|s,a) | (s,a) \in \mathcal{X}_h\})$

(2) *there exists a UniSOFT representation $\langle \tilde{\phi}_h, \tilde{\mu}_h \rangle_{\mathbb{R}^d} = \mathcal{P}^\star_h$.*

Proof. (1) $\Rightarrow$ (2). By construction, span$(\{\phi(s,a) | (s,a \in \mathcal{X}^\star_h)\}$ = span$(\{\phi(s,a) | (s,a) \in \mathcal{X}_h\})$. After extending Φ and Ψ with $d - \tilde{d}$ columns and rows of zero vectors, respectively, we see that $\Phi\Psi$ is a UniSOFT representation of $\mathcal{P}^\star_h$.

(2) $\Rightarrow$ (1). Let $\tilde{\Phi}\tilde{\Psi}$ be a UniSOFT representation. Then, we easily observe that,

$$\begin{aligned} \text{span}(\{\mathcal{P}^\star_h(\cdot|s,a) | (s,a) \in \mathcal{X}^\star_h\}) &= \text{span}(\{\tilde{\phi}(s,a)^T\tilde{\Psi} | (s,a) \in \mathcal{X}^\star_h\}) \\ &\overset{(i)}{=} \text{span}(\{\tilde{\phi}(s,a)^T\tilde{\Psi} | (s,a) \in \mathcal{X}_h\}) \\ &= \text{span}(\{\mathcal{P}^\star_h(\cdot|s,a) | (s,a) \in \mathcal{X}_h\}), \end{aligned}$$

where (i) follows from the UniSOFT property of $\tilde{\Phi}$. □

[1] $\tilde{O}$ hides a constant of order 2^{64}.

Lemma E.2 (Existence of good representations, Lemma 5.1)**.** *Let $d \geq \tilde{d}$. Then, the following statements are equivalent:*

(1) span$\{\mathcal{P}_h^\star(\cdot|s,a)|(s,a) \in \mathcal{X}_h^\star\} = \mathbb{R}^{\tilde{d}}$ *and* $|\mathcal{X}_h^\star| \geq d$*,*

(2) *there exists a non-redundant UniSOFT representation* $\langle \tilde{\phi}_h, \tilde{\mu}_h \rangle_{\mathbb{R}^d} = \mathcal{P}_h^\star$*,*

(3) *if* $d = \tilde{d}$*, any representation* $\langle \phi_h, \mu_h \rangle_{\mathbb{R}^d} = \mathcal{P}_h^\star$ *is UniSOFT.*

Proof. (1) $\Rightarrow$ (2). By construction, the rows of Φ corresponding to elements in $\mathcal{X}_h^\star$, i.e. the vectors $\{\phi(s,a)|(s,a) \in \mathcal{X}_h^\star\}$, form a basis of $\mathbb{R}^{\tilde{d}}$. As $|\mathcal{X}_h^\star| \geq d$ holds, we can extend Φ with $d - \tilde{d}$ columns of unit vectors, such that $\Phi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times d}$ and span$\{\phi(s,a)|(s,a) \in \mathcal{X}_h^\star\} = \mathbb{R}^d$. Hence, after appending $d - \tilde{d}$ rows of zero vectors to Ψ, we see that $\Phi\Psi$ is a non-redundant and UniSOFT representation of $\mathcal{P}_h^\star$.

(2) $\Rightarrow$ (1). First, note that $|\mathcal{X}_h^\star| \geq d$ must hold, in order to find d features ϕ that span the feature space $\mathbb{R}^d$. Second, note that rank$\{\mathcal{P}_h^\star(\cdot|s,a)|(s,a) \in \mathcal{X}_h^\star\} \leq \tilde{d}$ must hold, as otherwise rank$(\mathcal{P}_h^\star) > \tilde{d}$. We provide a proof by contradiction. Let $\tilde{\Phi}\tilde{\Psi}$ be any non-redundant UniSOFT representation. Suppose that rank$\{\mathcal{P}_h^\star(\cdot|s,a)|(s,a) \in \mathcal{X}_h^\star\} < \tilde{d}$ holds. Since $\tilde{\Phi}$ is UniSOFT and non-redundant by assumption, we have that rank$(\tilde{\Phi}\tilde{\Psi}) = $ rank$(\tilde{\Psi})$, which implies that rank$(\tilde{\Psi}) = \tilde{d}$ must be true, to match the rank of $\mathcal{P}_h^\star$. However, this further implies that

$$\tilde{d} \overset{(i)}{=} \text{rank}\{\tilde{\phi}(s,a)^T \tilde{\Psi}|(s,a) \in \mathcal{X}_h\} = \text{rank}\{\mathcal{P}_h^\star(\cdot|s,a)|(s,a) \in \mathcal{X}_h^\star\} \overset{(ii)}{<} \tilde{d}$$

holds, where (i) follows from $\tilde{\Phi}$ being UniSOFT and non-redundant and (ii) follows by assumption. This is, of course, absurd.

(2) $\Rightarrow$ (3). **(Case** $d = \tilde{d}$**)** Let $\mathcal{P}_h^\star = \Phi^\star\Psi^\star$ such that the representation is non-redundant and UniSOFT. By Theorem G.1 there exists an invertible matrix $R \in \mathbb{R}^{d \times d}$ such that $\bar{\Phi} = \Phi^\star R$ and $\bar{\Psi} = R^{-1}\Psi^\star$ for any other full rank factorization $\mathcal{P}_h^\star = \bar{\Phi}\bar{\Psi}$. Therefore, rows in $\Phi^\star$ that form a basis of $\mathbb{R}^d$ also form a basis of $\mathbb{R}^d$ in $\bar{\Phi}$.

(3) $\Rightarrow$ (1). The claim follows by the construction of Φ.

□

Corollary E.1. *Let $d \geq \tilde{d}$ and $\mathcal{Y}_h := \{s' \in \mathcal{S}|\exists(s,a) \in \mathcal{S} \times \mathcal{A} : \mathcal{P}_h^\star(s'|s,a) > 0\}$ be the set of states reachable by any other state (loops included). If there exists a state $s \in \mathcal{Y}_h$ s.t. $d^{\pi^\star}_{\mathcal{P}^\star,h+1}(s) = 0$, then there exists no factorization $P_h^\star = \Phi\Psi$ such that Φ is UniSOFT and non-redundant, where $\Phi \in \mathbb{R}^{|S||A| \times d}$ and $\Psi \in \mathbb{R}^{d \times |S|}$.*

Proof. First, note that

$$\text{rank}(\{\mathcal{P}_h^\star(\cdot|s,a)|(s,a) \in \mathcal{X}_h^\star\}) < \text{rank}(\{\mathcal{P}_h^\star(\cdot|s,a)|(s,a) \in \mathcal{S} \times \mathcal{A}\}) = \tilde{d}$$

must be true, since by assumption, there exists a state-action pair $(\tilde{s}, \tilde{a})$ such that $\mathcal{P}_h^\star(\cdot|\tilde{s},\tilde{a}) \notin$ span$(\{\mathcal{P}_h^\star(\cdot|s,a)|(s,a) \in \mathcal{X}_h\})$. Now, suppose that there exists a non-redundant UniSOFT representation. Then, by Lemma 5.1, we know that

$$\text{span}\{\mathcal{P}_h^\star(\cdot|s,a)|(s,a) \in \mathcal{X}_h\} = \mathbb{R}^{\tilde{d}}$$

must hold, which, however, contradicts the inequality derived above. □

Lemma E.3. *Suppose $(X, \|\cdot\|)$ is some normed space. Let $\{v_i\}_{i=1}^d$ be a set of linear independent vectors in X. Then, there exists some $\epsilon > 0$, such that any set of vectors $\{u_i\}_{i=1}^d$ in X with $\|v_i - u_i\| \leq \epsilon$ for all $i \in [d]$ is linear independent as well. In particular, $\epsilon < \min_{(\alpha_1,..,\alpha_d):\Sigma_i|\alpha_i|=1} \|\sum_{i=1}^d \alpha_i v_i\|/2$*

Proof. We provide a proof by contradiction. Let $S := \{(\alpha_1, ..., \alpha_d) \in \mathbb{R}^d|\sum_{i=1}^d |\alpha_i| = 1\}$. Suppose $\{u_i\}_{i=1}^d$ are linear dependent, that is, there exists some vector $(\alpha_1, ..., \alpha_d) \in \mathbb{R}^d$ such that

$$0 = \|\sum_{i=1}^d \alpha_i u_i\|.$$

In particular, w.l.o.g. we can assume that $(\alpha_1, ..., \alpha_d) \in S$. Let $\epsilon < \min_{(\alpha_1,..,\alpha_d):\Sigma_i|\alpha_i|=1} \| \sum_{i=1}^{d} \alpha_i v_i\|/2$ and positive. But then,

$$\begin{aligned} 0 = \| \sum_{i=1}^{d} \alpha_i u_i \| &= \| \sum_{i=1}^{d} \alpha_i v_i + \sum_{i=1}^{d} \alpha_i (u_i - v_i) \| \\ &\overset{(i)}{\geq} \| \sum_{i=1}^{d} \alpha_i v_i \| - \| \sum_{i=1}^{d} \alpha_i (u_i - v_i) \| \\ &\overset{(ii)}{>} 2\epsilon - \epsilon \sum_{i=1}^{d} |\alpha_i| = \epsilon, \end{aligned}$$

leads to a contradiction, where (i) follow from the reverse triangle inequality and (ii) follows from the Cauchy-Schwarz inequality. □

Lemma 5.2. *Assume that Assumption 3.4 (minimal optimal occupancy) holds and that $\mathcal{P}^\star$ admits a non-redundant UniSOFT representation. Then, there exists an $\epsilon > 0$ such that for any $d \geq \tilde{d}$ the following holds: Let $\tilde{\alpha} < \alpha \leq \epsilon$ be arbitrary. There exist infinitely more $\alpha^\star$-approximate representations than $\tilde{\alpha}^\star$-approximate representations $\langle \phi, \mu \rangle_{\mathbb{R}^d} \equiv \hat{\mathcal{P}}$ that are UniSOFT and non-redundant.*

Proof. Since $\mathcal{P}^\star$ is assumed to admit a non-redundant UniSOFT representation, by Lemma 5.1, there exist $\tilde{d}$ state-action pairs in $\mathcal{X}_h^\star$ such that their transition vectors in model $\mathcal{P}_h^\star$ span $\mathbb{R}^{\tilde{d}}$. Denote $\tilde{\mathcal{X}}_h^\star$ as the set that contains those $\tilde{d}$ state-action pairs. Let $\epsilon > 0$ arbitrary such that,

$$\epsilon < \min_{(\alpha_1,..,\alpha_d):\Sigma_i|\alpha_i|=1} \| \sum_{i=1}^{\tilde{d}} \alpha_i v_i \|_{\mathrm{TV}} \frac{d^\star_{\min}}{2},$$

where $\{v_i\}_{i=1}^{\tilde{d}} = \{\mathcal{P}_h^\star(\cdot|s,a) | (s,a) \in \tilde{\mathcal{X}}_h^\star\}$. Then, by continuity of norms and integrals, we can find an $\alpha^\star$-approximate representation with induced transition operator $\mathcal{P}$, such that for any $h \in [H]$ and $(s', a') \in \tilde{\mathcal{X}}_h^\star$,

$$\begin{aligned} \epsilon &= \mathbb{E}_{(s,a)\sim d^{\pi^\star}_{\mathcal{P}^\star,h}}[\|\mathcal{P}_h(\cdot|s,a) - \mathcal{P}_h^\star(\cdot|s,a)\|_{\mathrm{TV}}] \\ &= \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d^{\pi^\star}_{\mathcal{P}^\star,h}(s,a) \|\mathcal{P}_h(\cdot|s,a) - \mathcal{P}_h^\star(\cdot|s,a)\|_{\mathrm{TV}} \\ &\geq d^\star_{\min} \|\mathcal{P}_h(\cdot|s',a') - \mathcal{P}_h^\star(\cdot|s',a')\|_{\mathrm{TV}}. \end{aligned}$$

Then, by Lemma E.3, the vectors in $\{\mathcal{P}_h(\cdot|s,a) | (s,a) \in \tilde{\mathcal{X}}_h^\star\}$ are linear independent and, by Lemma E.2, there exists a non-redundant UniSOFT representation inducing $\mathcal{P}$. In particular, the existence of one good representation implies the existence of an infinite number of good representations. As ϵ was chosen arbitrarily, we conclude the proof.

□

F MULTIPLE OPTIMAL POLICIES

As noted in the preliminary section, we can extend our results to environments with multiple optimal policies as well. Recall that we denote $\Pi^\star$ as the set of all optimal (deterministic) policies. We say that a feature map ϕ is UniSOFT w.r.t. some policy π, if $\pi \in \Pi^\star$ and ϕ fulfills the UniSOFT property, as in definition 3.2, interchanging π and $\pi^\star$. In particular, a UniSOFT representation is non-redundant if $\lambda^\star(\phi) > 0$ where $\lambda^\star(\phi) := \min_{h\in[H],\pi^\star\in\Pi^\star} \lambda_{\min}(\mathbb{E}_{(s,a)\sim d^{\pi^\star}_{\mathcal{P}^\star,h}}[\phi_h(s,a)\phi_h(s,a)^T])$. We adjust the notion of $\alpha^\star$-approximate representations accordingly.

Definition F.1 ($(\sigma^\star, \alpha^\star)$-Approximate Representation)**.** *A representation $(\phi, \mu) \in \Phi \times \Psi$, with induced model $\mathcal{P}$, is $(\sigma^\star, \alpha)$-approximate at level α if for the finite sequence $\sigma^\star = (\pi_1^\star, \pi_2^\star, ..., \pi_t^\star)$ of optimal policies and for all $h \in [H]$,*

$$\mathbb{E}_{(s,a)\sim\gamma^\star_{t,h}}[\|\mathcal{P}_h(\cdot|s,a) - \mathcal{P}_h^\star(\cdot|s,a)\|_{\mathrm{TV}}] \leq \alpha,$$

where $\gamma^\star_{t,h}(s,a) = \frac{1}{t}\sum_{i=1}^{t} d^{\pi_i^\star}_{\mathcal{P}^\star,h}(s,a)$.

Assumption F.1 ($\alpha^\star$-Expressive Function Space). *Let $\sigma^\star$ be an arbitrary sequence of optimal policies of finite length. For all $(\sigma^\star, \alpha^\star)$-approximate representations $(\phi, \mu) \in \Phi \times \Psi$, there exists a non-redundant representation $(\tilde{\phi}, \tilde{\mu}) \in \Phi \times \Psi$ that is UniSOFT w.r.t. all $\pi^\star \in \sigma^\star$, such that the induced models $\mathcal{P}$ and $\tilde{\mathcal{P}}$ agree on all $(s, a) \in \mathcal{S} \times \mathcal{A}$, for which there exists a policy $\pi \in \Pi$, such that for any $h \in [H]$, we have $d^{\pi}_{\mathcal{P}^\star,h}(s, a) > 0$.*

Furthermore, recall that $\tilde{\pi}^\star_t := \{\tilde{\pi}^\star_{t,h}\}_{h\in[H]}$, where for each $h \in [H]$,

$$\tilde{\pi}^\star_{t,h}(s) = \begin{cases} \pi_{t,h}(s) & \text{if } \pi_{t,h}(s) \in \Pi^\star_h(s) \\ Select(\Pi^\star_h(s)) & \text{otherwise} \end{cases}.$$

We define $\tilde{\sigma}^\star_t := (\tilde{\pi}^\star_1, \tilde{\pi}^\star_2, ..., \tilde{\pi}^\star_t)$.

Compared to the unique optimal policy case, we must ensure the existence of feature maps that are UniSOFT w.r.t. all optimal policies, as we do not know in advance which distribution of optimal policies the algorithm converges to. In exchange for updating the expressiveness assumption 4.1 to the more restrictive assumption F.1, we can drop the unique optimal policy assumption. We note that allowing multiple optimal policies only worsens the sample complexity in the instance-dependent variables, which now depend on the 'worst' deterministic optimal policy.

The following two results ensure the selection of good representation. The remaining analysis can be performed analogously to the previous sections.

Lemma F.1. *(Selecting $(\tilde{\sigma}^\star_t, \alpha)$-representations) Fix any $\alpha > 0$. Assume there exists an increasing sub-linear function g such that $\mathcal{R}(t) \leq g(t)$ for all $t \in \mathbb{N}$. Suppose we run algorithm 1 and assumptions 3.4 (minimal optimal occupancy) and 3.3 (minimal sub-optimality gap) hold. Then, given that the event $\mathcal{E}$ occurs, there exists an episode τ_α such that for all episodes $t \geq \tau_\alpha$ and time steps $h \in [H]$, the learned feature maps $\hat{\phi}_{t,h}$ are $(\tilde{\sigma}^\star_t, \alpha)$-approximate, where*

$$\tau_\alpha := \min\{t | t > \frac{1}{\alpha}(\frac{\mathcal{R}(t)}{\Delta_{\text{min}} d^\star_{\text{min}}} + \frac{|\mathcal{A}|}{\xi_t}\sqrt{2t\log(4t|\Phi||\Psi|H/\delta)})\}.$$

Proof. Directly follows from Corollary B.1 and the proof of Lemma B.1. □

Lemma F.2. *(Selecting non-redundant UniSOFT representation) Fix any $\alpha > 0$. Assume there exists an increasing sub-linear function g such that $\mathcal{R}(t) \leq g(t)$ for all $t \in \mathbb{N}$. Suppose we run algorithm 1 and assumptions F.1 (expressiveness) and 3.3 (minimal sub-optimality gap) hold. Additionally, if $\alpha < 1$, suppose assumption 3.4 (minimal optimal occupancy) holds. Then, given that the events $\mathcal{E}(\delta)$ and $\mathcal{F}(\delta)$ occur, there exists an episode $\tau_{\text{unisoft}} \geq \tau_\alpha$ such that for all subsequent episodes $t \geq \tau_{\text{unisoft}}$ and time steps $h \in [H]$ the learned feature maps $\hat{\phi}_{t,h}$ are UniSOFT w.r.t. any optimal policy $\pi^\star \in \tilde{\sigma}^\star_t$, where*

$$\tau_{\text{unisoft}} := \min\{t | t > \left(\frac{2}{\lambda^\star_\alpha}(\Delta^{-1}_{\text{min}}\mathcal{R}(t) + 2\sum_{i=1}^{t}\xi_{i-1} + 18\sqrt{t\log(6dtH|\Phi|/\delta)}) \vee \tau_\alpha\right)\}.$$

Proof. Let $\Phi^{\text{unisoft}}_{\tilde{\sigma}^\star_t} \subseteq \Phi$ denote the set containing only non-redundant feature mappings that are UniSOFT w.r.t. at least one $\tilde{\pi}^\star \in \tilde{\sigma}^\star_t$. By Lemma B.2, with probability at least $1 - \delta$, for all $t \in \mathbb{N}$, $h \in [H]$, $\phi \in \Phi \setminus \Phi^{\text{unisoft}}_{\tilde{\sigma}^\star_t}$ and $\phi^{\text{unisoft}} \in \Phi^{\text{unisoft}}_{\tilde{\sigma}^\star_t}$,

$$\lambda_{\text{min}}(\Sigma_{t+1,h}(\phi^{\text{unisoft}}) - \lambda_t I) \geq t\lambda^\star(\phi^{\text{unisoft}}) - 2\sum_{i=1}^{t}\xi_{i-1} - \Delta^{-1}_{\text{min}}\mathcal{R}(t) - 18\sqrt{t\log(6dtH|\Phi|/\delta)},$$

$$\lambda_{\text{min}}(\Sigma_{t+1,h}(\phi) - \lambda_t I) \leq 2\sum_{i=1}^{t}\xi_{i-1} + \Delta^{-1}_{\text{min}}\mathcal{R}(t) + 18\sqrt{t\log(6dtH|\Phi|/\delta)},$$

where $\Sigma_{h,t+1}(\phi) = \sum_{(s,a)\in\mathcal{D}_{t,h}} \phi_h(s, a)\phi_h(s, a)^T$ and

$$\begin{aligned}\lambda^\star(\phi) &:= \min_{h\in[H],\pi^\star\in\Pi^\star} \lambda_{\text{min}}(\mathbb{E}_{(s,a)\sim d^{\pi^\star}_{\mathcal{P}^\star,h}}[\phi_h(s, a)\phi_h(s, a)^T]) \\ &\leq \min_{h\in[H]} \lambda_{\text{min}}(\mathbb{E}_{(s,a)\sim\tilde{\gamma}^\star_{t,h}}[\phi_h(s, a)\phi_h(s, a)^T]).\end{aligned}$$

Let us denote $\Phi_\alpha \times \Psi_\alpha \subseteq \Phi \times \Psi$ as the set of $(\tilde{\sigma}_t^\star, \alpha)$-approximate representations. Additionally, denote

$$\Phi_\alpha^{\text{unisoft}} \times \Psi_\alpha^{\text{unisoft}} = (\Phi_\alpha \times \Psi_\alpha) \cap \left(\Phi_{\tilde{\sigma}_t^\star}^{\text{unisoft}} \times \Psi\right),$$

as the set containing all $(\tilde{\sigma}_t^\star, \alpha)$-approximate representations such that the feature map is non-redundant and UniSOFT w.r.t. at least one $\pi \in \tilde{\sigma}_t^\star$, which is non-empty by Assumption F.1. A desired feature map is selected at episode $t \geq \tau_\alpha$ if for all $\tilde{\alpha} \leq \alpha$,

$$\max_{\phi^{\text{unisoft}} \in \Phi_{\tilde{\alpha}}^{\text{unisoft}}} \lambda_{\text{min}}(\Sigma_{t+1,h}(\phi^{\text{unisoft}}) - \lambda_t I) > \max_{\phi \in \Phi_{\tilde{\alpha}} \setminus \Phi_{\tilde{\alpha}}^{\text{unisoft}}} \lambda_{\text{min}}(\Sigma_{t+1,h}(\phi) - \lambda_t I),$$

or equivalently,

$$t\lambda_\alpha^\star(\phi^{\text{unisoft}}) > 2\left(\Delta_{\text{min}}^{-1}\mathcal{R}(t) + 2\sum_{i=1}^{t}\xi_i + 18\sqrt{t\log(6dtH|\Phi|/\delta)}\right),$$

where $\lambda_\alpha^\star := \min_{\tilde{\alpha}\leq\alpha} \max_{\phi^{\text{unisoft}} \in \Phi_{\tilde{\alpha}}^{\text{unisoft}}} \lambda^\star(\phi^{\text{unisoft}})$. □

G AUXILIARY RESULTS

Lemma G.1 (Simulation Lemma [Zhang et al., 2022a]). *Given two transition models $\mathcal{P}$ and $\mathcal{P}'$, we have:*

$$V_{\mathcal{P}',r+b,1}^{\pi,d_1} - V_{\mathcal{P},r,1}^{\pi,d_1} = \sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim d_{\mathcal{P}',h}^{\pi}}[b_h(s,a) + (\mathcal{P}_h' - \mathcal{P}_h)V_{\mathcal{P},r,h+1}^{\pi}(s,a)],$$

$$V_{\mathcal{P}',r+b,1}^{\pi,d_1} - V_{\mathcal{P},r,1}^{\pi,d_1} = \sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim d_{\mathcal{P},h}^{\pi}}[b_h(s,a) + (\mathcal{P}_h' - \mathcal{P}_h)V_{\mathcal{P}',r+b,h+1}^{\pi}(s,a)].$$

Lemma G.2 ([He et al., 2021]). *For any $h \in [H]$, $s \in \mathcal{S}$, and $\pi \in \Pi$:*

$$V_{\mathcal{P}^\star,r^\star,h}^{\pi^\star}(s) - V_{\mathcal{P}^\star,r^\star,h}^{\pi}(s) = \mathbb{E}[\sum_{h'=h}^{H} \Delta_{h'}(s_{h'}, a_{h'})|s_h = s, \pi, \mathcal{P}^\star],$$

Hence the regret after T episodes can be expressed as:

$$\begin{aligned}\mathcal{R}(T) &= \sum_{t=1}^{T} V_{\mathcal{P}^\star,r^\star,1}^{\pi^\star,d_1} - V_{\mathcal{P}^\star,r^\star,1}^{\pi_t,d_1} = \sum_{t=1}^{T} \mathbb{E}_{s\sim d_1}[\sum_{h=1}^{H} \Delta_h(s_h,a_h)|s_1 = s, \pi_t, \mathcal{P}^\star] \\ &= \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim d_{\mathcal{P}^\star,h}^{\pi_t}}[\Delta_h(s,a)]\end{aligned}$$

Proof.

$$\begin{aligned}&V_{\mathcal{P}^\star,r^\star,h}^{\pi^\star}(s) - V_{\mathcal{P}^\star,r^\star,h}^{\pi}(s) \\ &= \Delta_h(s,\pi_h(s)) + Q_{\mathcal{P}^\star,r^\star,h}^{\pi^\star}(s,\pi_h(s)) - V_{\mathcal{P}^\star,r^\star,h}^{\pi}(s) \\ &= \Delta_h(s,\pi_h(s)) + r_h^\star(s,\pi_h(s)) + \mathcal{P}_h^\star V_{\mathcal{P}^\star,r^\star,h+1}^{\pi^\star}(s,\pi_h(s)) - r_h^\star(s,\pi_h(s)) - \mathcal{P}_h^\star V_{\mathcal{P}^\star,r^\star,h+1}^{\pi}(s,\pi_h(s)) \\ &= \Delta_h(s,\pi_h(s)) + \mathcal{P}_h^\star(V_{h+1}^{\pi^\star} - V_{h+1}^{\pi})(s,\pi_h(s))\end{aligned}$$

Unravelling the recursion gives the result. □

Theorem G.1 ([Piziak and Odell, 1999]). *Every matrix $A \in \mathbb{C}^{n\times m}$ with* $\text{rank}(A) = r > 0$ *has infinitely many full rank factorizations. However, if $A = FG = \bar{F}\bar{G}$ are two full rank factorizations of A, then there exists an invertible matrix $R \in \mathbb{C}^{r\times r}$ such that $\bar{F} = FR$ and $\bar{G} = R^{-1}G$.*

Lemma G.3 (Lemma D.1. in Jin et al. [2020]). *Let $\Sigma_t = \lambda I + \sum_{i=1}^{t}\phi_i\phi_i^T$ where $\phi_i \in \mathbb{R}^d$ and $\lambda > 0$. Then,*

$$\sum_{i=1}^{t}\phi_i^T\Sigma_t^{-1}\phi_i = \text{Tr}(\Sigma_t^{-1}\sum_{i=1}^{t}\phi_i\phi_i^T) \leq d.$$

Lemma G.4 (Elliptical potential lemma [Abbasi-Yadkori et al., 2011]). *Consider a sequence of $d \times d$ positive semidefinite matrices $X_1, ..., X_T$ with $tr(X_t) \leq 1$ for all $t \in [T]$. Define $M_0 = \lambda_0 I$ and $M_t = M_{t-1} + X_t$. Then*

$$\sum_{t=1}^{T} tr(X_t M_{t-1}^{-1}) \leq 2d \log(1 + \frac{T}{d\lambda_0})$$

Proposition G.1 (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 $X_k^2 \preccurlyeq C_k^2$ almost surely. Then, with probability at least $1 - \delta$:*

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

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

Lemma G.5. *(Azuma's inequality) Let $(X_k)_{k=1}^t$ be a finite adapted sequence such that for all k, $\mathbb{E}_k[X_k] = 0$ and $|X_t| \leq a$ almost surely. Then, with probability at least $1 - \delta$:*

$$|\sum_{k=1}^{t} X_k| \leq a\sqrt{t \log(2/\delta)}$$

Lemma G.6 (MLE guarantee [Cheng et al., 2023]). *Fix $\delta \in (0, 1)$. Then, with probability $1 - \delta/2$,*

(1) *for all $h = 2, ..., H$ and $t \in \mathbb{N}$,*

$$\mathbb{E}_{(s,a)\sim\rho'_{t,h}(s,a)}[\|\hat{\mathcal{P}}_{h,t}(\cdot|s, a) - \mathcal{P}_h^\star(\cdot|s, a)\|_{\text{TV}}^2] \leq \zeta_t,$$

(2) *for $h = 1$ and all $t \in \mathbb{N}$,*

$$\mathbb{E}_{(s,a)\sim\rho_{t,h}(s,a)}[\|\hat{\mathcal{P}}_{h,t}(\cdot|s, a) - \mathcal{P}_h^\star(\cdot|s, a)\|_{\text{TV}}^2] \leq \zeta_t,$$

where $\zeta_t = \frac{2\log(4t|\Phi||\Psi|H/\delta)}{t}$.