universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

On Constant Regret for Low-Rank MDPs

verfasst von | submitted by

Alexander Sturm

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 645
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Data Science
Betreut von | Supervisor:	Assoz. Prof. Dipl.-Ing. Dr.techn. Sebastian Tschiatschek BSc

Acknowledgements

I would like to thank my girlfriend Manuela Wieser for her emotional support throughout my studies. I would also like to thank Professor Sebastian Tschiatschek for his guidance and support.

Abstract

Although there exist instance-dependent regret results for linear Markov decision processes (MDPs) and low-rank Bandits, extensions to low-rank MDPs remain unexplored. In this master thesis, we close this gap and provide expected 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 feature maps with good spectral properties. We show that for any low-rank MDP with positive minimal sub-optimality gap, UniSREP-UCB achieves expected regret $\tilde{O}(H^4 d^{1/2}|\mathcal{A}|T^{2/3})$, after some warm-up episodes. Furthermore, we demonstrate that optimal policy identification is possible, as long as the minimal sub-optimality gap and the occupancy distributions of optimal policies are well-defined and known. To the best of our knowledge, these are the first instance-dependent regret results for low-rank MDPs.

Kurzfassung

Obwohl bereits problemabhängige Regret-Bounds für lineare Markov Decision Processes (MDPs) und Low-Rank Bandits existieren, bleiben Erweiterungen auf Low-Rank MDPs unerforscht. In dieser Masterarbeit schließen wir diese Lücke und liefern Expected-Regret-Bounds für Low-Rank MDPs in einem problemabhängigen Kontext. Konkret stellen wir einen Algorithmus namens UniSREP-UCB vor, der ein beschränktes Optimierungsziel nutzt, um Representationen mit guten spektralen Eigenschaften zu lernen. Wir zeigen, dass für jeden Low-Rank MDP mit einem positiven minimalen Sub-Optimalitygap, UniSREP-UCB nach einigen Aufwärmepisoden einen Expected-Regret von $\tilde{O}(H^4 d^{1/2}|\mathcal{A}|T^{2/3})$ erreicht. Darüber hinaus zeigen wir, dass eine Identifikation der optimalen Policy möglich ist, solange der minimale Sub-Optimalitygap und die Occupancy-Distributions der optimalen Policies wohldefiniert und bekannt sind. Nach bestem Wissen sind dies die ersten problemabhängigen Regret-Bounds für Low-Rank MDPs.

Contents

Acknowledgements i

Abstract iii

Kurzfassung v

List of Tables ix

List of Algorithms xi

1. Introduction 1
- 1.1.1. Research Question & Summary of Contributions 2
- 1.1.2. Related Work 3
- 1.1.3. Thesis Structure 4

2. Preliminaries 7

3. Instance-Dependent Regret Bounds & Optimal Policy Identification for Low-Rank MDPs 13
- 3.1. Algorithm 13
- 3.2. Instance-Dependent Regret Bounds 15
- 3.3. Further Results 19

4. Discussion 23
- 4.1. Comparison with the Literature 23
- 4.2. Limitations 25

5. Conclusions & Future Work 27

Bibliography 29

A. Appendix 37
- A.1. Sub-Linear Pseudo-Regret without UniSOFT Representations 38
- A.2. Selecting Non-Redundant UniSOFT Representations 46
- A.3. Improved Pseudo-Regret with UniSOFT Representations 53
- A.4. Constant Pseudo-Regret with UniSOFT Representations 60
- A.5. Existence of UniSOFT Representations 66
- A.6. Multiple Optimal Policies 68
- A.7. Auxiliary Results 69

List of Tables

4.1. Critical episodes for algorithms achieving constant regret. 23

1. Notation 33
2. Notation 34
3. Notation 35
4. Notation 36

List of Algorithms

1. Episodic RL Framework 8

2. UniSREP-UCB (Upper Confidence Bound driven Universally Spanning Representation Learning) 21
3. UniSREP-UCB+ 22

1. Introduction

Reinforcement learning (RL) is a framework for describing sequential decision-making problems through the interaction of an agent with an unknown environment. It is often assumed that the interaction between agent and environment is episodic, i.e., the agent interacts with the environment in episodes, where each episode terminates in finite time, upon which the agent returns to a (possibly random) starting state. Additionally, RL often assumes that the environment is described by an episodic Markov Decision Process (MDP), which consists of a state space, an action space, a transition operator, describing the transition probabilities between states after performing some action, a reward function, providing feedback on actions and a horizon, determining the length of an episode. The goal is to learn an optimal decision policy from information gained by exploring the environment.

We can measure the performance of a learning algorithm by the regret; that is, the cumulative difference in expected total rewards between the behavior policies employed by the learning algorithm and an optimal decision policy. Generally, regret bounds hold for all environments of a given class; that is, a set of environments sharing a specific property (e.g. all environments with a gaussian transition kernel) and hence are 'worst-case' in nature. However, whenever additional information about an environment is available, we would like to perform a more refined regret analysis. We say that the regret is instance/problem-dependent whenever it depends on properties of the environment, usually characterizing the hardness of the RL instance/problem. Generally, a RL algorithm is efficient in learning an optimal decision policy, whenever the regret grows sub-linearly in the number of learning steps.

In many applications of RL, there is a common expectation that a good RL algorithm will eventually gain enough knowledge on the environment, such that it will identify an optimal decision policy in finite time [ZFHG24]. Therefore, a very interesting question for any RL problem is, under what assumptions this expectation can be confirmed theoretically. In particular, under which conditions does there exist an algorithm that identifies the optimal policy in finite time. In that regard, we say an algorithm enjoys constant regret, whenever the regret does not scale with the number of learning steps.

Nevertheless, when the environment is described by a high-, possibly infinite-dimensional state space, efficient learning without some form of representation learning, i.e., finding a meaningful and easy to process representation of the state space, is generally not possible [OVR16, SJ19]. In RL however, representation learning is particularly challenging when considering the exploitation versus exploration dilemma [AKKS20]. Intuitively, one re-

quires a high quality dataset collected from the environment to learn good representations, but conversely, also requires good representations to efficiently explore the environment in the first place. Naturally, good representations not only describe the environment reasonably well but additionally allow for more efficient exploration.

Recently, [JYWJ20] has shown that sample efficient learning in a high-dimensional state-action space is possible in linear 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, [PTP$^+$21] showed that, features holding a spectral property called UniSOFT (see Definition 2.1.2), are necessary and sufficient for constant instance-dependent regret. Additionally, they provide a representation selection algorithm that is able to achieve constant instance-dependent regret when provided with a known set of exact feature maps (these are feature maps that exactly represent the transition operator) containing one that is UniSOFT.

Similarly, in contextual linear bandits (CLB), where the reward function is linear in the features ϕ, [PTR$^+$21] showed that a diversity condition called HLS [HLS20], is necessary and sufficient for constant instance-dependent regret. In contrast to linear MDPs, [TPT$^+$22] were able to provide an algorithm that achieves constant instance-dependent regret for CLBs, even when the true features ϕ are unknown and have to 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 identifies the optimal policy for low-rank MDPs; that is, linear MDPs with unknown features ϕ. Nevertheless, many existing works (e.g. [AKKS20, UZS22, ZSU$^+$2a, MCK$^+$24]) have shown that, given access to an optimization oracle, sample efficient learning in low-rank MDPs is possible. In that regard, we call an algorithm oracle-efficient whenever sample efficient learning is made possible, by providing access to an optimization oracle; i.e. a known function that returns a solution to an optimization problem. Finally, [ZYW$^+$24] achieved the first sub-linear regret guarantee for low-rank MDPs.

1.1.1. Research Question & Summary of Contributions

In this thesis, we study low-rank MDPs and aim to close the gap on instance-dependent regret results. In particular, we address the following important open research question:

Is it possible to define an oracle-efficient RL algorithm that enjoys constant instance-dependent regret in low-rank MDPs?

As we will see, we can answer the aforementioned question positively. In particular, we provide an instance-dependent analysis of a slightly augmented version of the recently proposed REP-UCB algorithm [UZS22], which serves as a basic framework for many other works [ZSU$^+$2a, ASS$^+$23, ZYW$^+$24] on low-rank MDPs. In our analysis, we leverage

the insights of [CHYL23], which designed an UCB-style bonus term that serves as a trajetory-wise uncertainty measure for the approximation of the transition operator. This allows as to perform an instance-dependent regret analysis, similar to [PTP$^+$21], by employing UniSOFT feature maps and a double exploration strategy introduced by [ZYW$^+$24]. More specifically:

- We provide an algorithm that achieves $\tilde{O}(T^{2/3})$ expected regret (Theorem 3.2.1) whenever the minimal sub-optimality gap exists and we have access to an expressive enough function space (Assumption 3.2.1);
- We show that optimal policy identification is possible for low-rank MDPs (Theorem 3.2.3), provided that the minimal sub-optimality gap and the minimal optimal occupancy (Definition 2.1.4) are known;
- We demonstrate that the existence of exact UniSOFT representations is fully characterized by the RL instance whenever the rank is minimal (Section 3.3).

1.1.2. Related Work

We provide a non-exhaustive summary of related work. For a more thorough discussion on how this thesis aligns with the literature see Chapter 4.

Linear MDPs

There exists a large body of literature providing regret bounds for linear MDPs, which assume that the transition operator admits a low-rank decomposition into known features and unknown signed measures. In this setting, [JYWJ20] proposed the first sample efficient algorithm without assuming access to a generative model or other restrictive assumptions on the transition operator. Their model-free algorithm LSVI-UCB combines classical LSVI with UCB-style bonuses and achieves $\tilde{O}(\sqrt{T})$ worst-case regret. Later, [HZG21] provided the first instance-dependent regret analysis for linear MDPs and provided a logarithmic $O(\Delta_{\min}^{-1} \log(T))$ instance-dependent regret bound, given some minimal sub-optimality gap $\Delta_{\min}$. By leveraging features that fulfill a diversity condition, called called UniSOFT (see Def. 2.1.2), [PTP$^+$21] showed that LSVI-UCB enjoys constant instance-dependent regret. In particular, they relate the per-iteration regret to confidence intervals, which decrease uniformly when the features hold the UniSOFT property. Further, they demonstrate that the UniSOFT property is necessary for constant expected regret, reinforcing the importance of good features. Nevertheless, [ZFHG24] were able to provide an algorithm that achieves constant regret without prior assumption on the features. Remarkably, features are not required to be accurate for the constant regret result to hold, as long as they have low point-wise misspecification w.r.t. the minimal sub-optimality gap. Their algorithm employs a phased elimination scheme which eliminates actions based on their value under decreasing levels of uncertainty.

Low-rank MDPs

In the much more challenging low-rank MDP setting, where the features are unknown as well, the seminal work of [AKKS20] provided the first reward-free oracle-efficient algorithm called FLAMBE. They proposed to learn representations using maximum likelihood estimation (MLE) and showed that their explore-then-commit style algorithm achieves polynomial sample complexity, when provided with a MLE oracle. By interleaving representation learning, exploration and exploitation together, [UZS22] provided an algorithm called REP-UCB which 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 almost optimism at the initial state distribution and track the progress of the algorithm through the potential function of the unknown true features. Recently, [CHYL23] showed that the bonus term of REP-UCB also serves 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, [ZYW$^+$24] provided the first regret bound for low-rank MDPs, by employing a double exploration strategy. However, as far as we known, in contrast to linear MDPs, there exists no instance-dependent regret result for low-rank MDPs. Furthermore, whether we can identify an optimal policy, such as in linear MDPs, is still an open problem.

Contextual Linear Bandits

In contextual linear bandits (CLB), the equivalent to linear MDPs with horizon one, [PTR$^+$21] showed that a diversity condition called HLS [HLS20], similar to the UniSOFT property for linear MDPs, is necessary and sufficient for constant instance-dependent regret. Relaxing the assumption of exact feature maps, [TPT$^+$22] provided an algorithm which achieves constant regret, by introducing a constrained optimization objective which encourages the HLS property and enforces representations to be exact.

1.1.3. Thesis Structure

The thesis is structured as follows:

Chapter 2 - Preliminaries
Introduces the premise of this thesis.

Chapter 3 - Instance-dependent regret bounds & optimal policy identification for low-rank MDPs
States the main results of this thesis.

Chapter 4 - Discussion
Compares the results of this thesis with the literature and discusses limitations.

Chapter 5 - Conclusions & Future Work
Summarizes the findings of this thesis and discusses possible directions for future work.

Chapter A - Appendix
Provides the proofs for the results stated in Chapter 3.

2. Preliminaries

We start with some general notations. We let $\Delta(A)$ denote the set of probability distributions over a set A. Further, let $\mathcal{U}(A)$ represent the uniform distribution over some set A and let $\text{Ber}(p)$ denote the Bernoulli distribution with success rate $p \in [0,1]$. Additionally, $[N] := \{1, ..., N\}$ for any integer N. For some index set $\mathcal{I}$ and any set of functions $\{f_i : X \to Y\}_{i \in \mathcal{I}}$ we denote $\prod_{i \in \mathcal{I}} f_i$ as the Cartesian product $X^{\mathcal{I}} \to Y^{\mathcal{I}}$ with $(f_1 \times f_2 \times ... \times f_{\mathcal{I}})(x_1, x_2, ..., x_{\mathcal{I}}) = (f_1(x_1), f_2(x_2), ..., f_{\mathcal{I}}(x_{\mathcal{I}}))$. Finally, $\lesssim$ denotes inequalities up to absolute constants, $O(\cdot)$ hides absolute constants, and $\tilde{O}(\cdot)$ hides absolute constants as well as poly-log terms.

We consider a finite-horizon episodic Markov Decision Process (MDP) described by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}^\star, r^\star, H, d_1)$, where $\mathcal{S}$ is the finite state space, $\mathcal{A}$ is the finite action space, $\mathcal{P}^\star = \prod_{h \in [H]} \mathcal{P}_h^\star$ where $\mathcal{P}_h^\star : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition operator (unknown) at time step $h \in [H]$, $r^\star = \prod_{h \in [H]} r_h^\star$ where $r_h^\star : \mathcal{S} \times \mathcal{A} \to [0,1]$ 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 that the reward function is normalized; that is, $\sum_{h=1}^{H} \sup_{a,s} r_h(s,a)^\star \leq 1$.

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

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

Additionally, given some initial state distribution d_1 let $V_{\mathcal{P},r;1}^{\pi,d_1} = \mathbb{E}_{s \sim d_1}[V_{\mathcal{P},r;1}^\pi(s)]$ denote the expected total reward of policy π under $\mathcal{P}$, r and d_1. Further, let us define the Q-function as $Q_{\mathcal{P},r;h}^\pi(s,a) = r_h(s,a) + \mathcal{P}_h V_{\mathcal{P},r;h+1}^\pi(s,a)$ which represents the expected total reward of performing action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ at time step $h \in [H]$ and then following policy π under $\mathcal{P}$ and r.

Another important quantity is the state-action occupancy distribution $d^{\pi}_{\mathcal{P};h}(s,a)$, which denotes the probability of visiting state $s \in \mathcal{S}$ at time step $h \in [H]$ and performing action $a \in \mathcal{A}$ under model $\mathcal{P}$ and 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 $h \in [H]$. We can sample a state $s \in \mathcal{S}$ from $d^{\pi}_{\mathcal{P};h}$ by executing π for $h-1$ steps starting from state $s_1 \sim d_1$.

Algorithm 1 Episodic RL Framework

Input: MDP $\mathcal{M}$, Learning Algorithm Alg, Total number of episodes T
Output: π_T

Initialize: $\mathcal{D} = \emptyset$, $\pi_0 = \mathcal{U}(\mathcal{A})$
for $t = 1, ..., T$ **do**
  $\tau = \mathcal{M}(\pi_{t-1})$ ▷ Agent interacts with environment and collects data τ
  $\mathcal{D}_t = \mathcal{D}_{t-1} \cup \{\tau\}$ ▷ Store data
  $\pi_t = \text{Alg}(\mathcal{D}_t)$ ▷ update policy
end for
return π_T

The goal of the agent 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, by interacting with the environment as described in Algorithm 1. In each episode t, the agents first explores the environment $\mathcal{M}$ with the policy π_{t-1} and collects data $\tau = \mathcal{M}(\pi_{t-1})$ (e.g. a trajectory $(s_1, a_1, s_2, ..., a_H, s_H)$). Then, the agent updates its exploration strategy $\pi_t = \text{ALG}(\mathcal{D}_t)$ according to some learning algorithm ALG by considering the newly gained experience $\mathcal{D}_t = \mathcal{D}_{t-1} \cup \{\tau\}$. Finally, after T episodes, the algorithm terminates and returns the policy π_T.

Additionally we want the agent to learn an optimal policy efficiently, where we evaluate the efficiency of an agent by the (expected) regret

$$\mathbb{E}[\mathcal{R}(T)] = \mathbb{E}[\sum_{t=1}^{T} V^{\pi^{\star},d_1}_{\mathcal{P}^{\star},r^{\star},1} - V^{\pi_t,d_1}_{\mathcal{P}^{\star},r^{\star},1}], \tag{2.1}$$

which measures the expected cumulative performance loss up to episode $T \in \mathbb{N}$. More specifically, $\mathcal{R}(T)$ denotes the cumulative difference in expected total rewards up to episode T by following the decision rules of the behavior policy π_t instead of $\pi^{\star}$. Note that the expectation in equation 2.1 is taken w.r.t. any extra randomness induced by the algorithm.

Finally, we denote the sub-optimality gap of taking action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ at time step $h \in [H]$ 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.

Structural Assumptions

In this thesis 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 [UZS22, AKKS20, PTP$^+$21].

Definition 2.1.1. *(Low-rank MDP, [AKKS20]) 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 $\phi_h^\star : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and $\mu_h^\star : \mathcal{S} \to \mathbb{R}^d$ such that*

$$\forall (s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} : \mathcal{P}_h^\star(s'|s, a) = \langle \phi_h^\star(s, a), \mu_h^\star(s') \rangle,$$

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

Remark 2.1.1. *We can extend the definition such that the rank d is time step dependent. The results in the following chapters would then hold with d replaced by $\max_{h\in[H]} d_h$, where d_h denotes the rank at time step h.*

Remark 2.1.2. *Ignoring the normalization conditions of Def. 2.1.1, the existence of one low-rank representation, implies the existence of an infinite amount of low-rank representations (see Appendix A.5).*

Note that, in order to efficiently learn in the environment $\mathcal{M}$, we need to have at least a good approximation of the transition operator $\mathcal{P}^\star$. Hence, as the embedding functions $\phi_h^\star$ and $\mu_h^\star$ are 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 $h \in [H]$, $\Phi_h \subseteq \{\phi_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d\}$ and $\Psi_h \subseteq \{\mu_h : \mathcal{S} \to \mathbb{R}^d\}$. For notational brevity, we denote $\phi^\star = \prod_{h\in[H]} \phi_h^\star$ and $\mu^\star = \prod_{h\in[H]} \mu_h^\star$. Note that, as mentioned in the introduction, without some tool to estimate the transition operator, efficient learning is generally not possible.

To make this representation learning problem tractable, we need to ensure that we are able to select a low-rank decomposition that exactly represents $\mathcal{P}^\star$, i.e., $\phi^\star \in \Phi$ and $\mu^\star \in \Psi$, regularity of the function classes and that all function pairs induce a distribution over the state space, i.e. for $(\phi, \mu) \in \Phi \times \Psi$ we have that $\langle \phi, \mu \rangle \in (\Delta\mathcal{S})^H$ [AKKS20, UZS22, PTP$^+$21].

Assumption 2.1.1. *(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 $g : \mathcal{S} \to \mathbb{R}$, $\| \int_{\mathcal{S}} \mu_h(s) g(s) ds\|_2 \leq \sqrt{d}\|g\|_\infty$ 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 definition 2.1.1.*

Note that any tuple $(\phi, \mu) \in \Phi \times \Psi$ naturally induces a distribution over the state-space in each coordinate and in particular, a transition operator $\mathcal{P} \equiv \langle \phi, \mu \rangle$. In the following we will introduce good representations and instance-dependent quantities that measure learnability of an environment.

Good representations and instance-dependent properties

In favor of clarity, the main results of this thesis are formulated under the assumption of an unique optimal policy. But, as we will see, we can extend them to hold for multiple optimal policies as well. Let us denote $\Pi^\star$ as set of all optimal (deterministic) policies.

Assumption 2.1.2. *(Unique optimal policy) There exists an 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 2.1.2. *(UniSOFT Representation, [PTP+21]) A feature mapping* $\phi \in \Phi$ *is called UniSOFT (Universally Spanning Optimal FeaTures) if for all* $\pi^\star \in \Pi^\star$ *and* $h \in [H]$,

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

holds. In particular, a UniSOFT feature mapping ϕ *is* non-redundant *if it spans* $\mathbb{R}^d$ *in each coordinate or equivalently* $\lambda^\star(\phi) > 0$ *holds, where*

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

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

Intuitively, non-redundant UniSOFT features allow an algorithm to efficiently explore the whole feature space, by only deciding optimally w.r.t. the reward signal. 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 leverage 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 2.1.3. *(Well-defined minimal sub-optimality gap) The quantity*

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

is well defined.

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

Assumption 2.1.4. *(Well-defined minimal optimal occupancy) The quantity*

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

is well defined.

Note that both assumptions trivially hold whenever $\mathcal{S}$ and $\mathcal{A}$ are finite, which we assumed in the beginning. However, introducing the assumptions above allows us to easily extend most results to infinite state spaces.

3. Instance-Dependent Regret Bounds & Optimal Policy Identification for Low-Rank MDPs

This chapter provides an algorithm, called UniSREP-UCB (Algorithm 2), that achieves sub-linear expected regret under an additional simplifying assumption that guarantees the selection of non-redundant UniSOFT feature maps. Furthermore, we demonstrate that by introducing a carefully chosen termination criteria to UniSREP-UCB, resulting in algorithm UniSREP-UCB+ (Algorithm 3), we can identify the optimal policy with high probability whenever the minimal sub-optimality gap and the minimal optimal occupancy are known. Finally, we will see that the existence of exact UniSOFT features is fully determined by the environment, whenever d is minimal.

3.1. Algorithm

In this section, we describe the proposed UniSREP-UCB algorithm. On a high level, the algorithm is a finite horizon adaption of the REP-UCB algorithm proposed in [UZS22]. However, different from the REP-UCB algorithm, we employ a double exploration scheme, as proposed in [ZYW+24] and augment the representation learning objective to encourage feature maps with good spectral properties. In each episode t, the algorithm runs through three phases which we explain in more detail below.

Exploration (Lines 4-13)

In the exploration phase, the algorithm carefully collects samples from the environment which later will be used for learning representations of the environment. In particular, for each time step h, the algorithm samples from the state-occupancy distribution $d^{\pi_{t-1}}_{\mathcal{P}^\star,h-1}$ (rolls-in at time step $h-1$) and continues for the next two time steps (rolls-out to time step $h+1$) based on the outcome of a Bernoulli experiment with success rate $1-\xi_t$. If successful, the algorithm explores with the behavior policy π_{t-1} of the last episode, and otherwise, 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. In particular, taking actions uniformly at random with positive probability is necessary, for employing a crucial representation learning guarantee (Lemma A.7.6). 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 to provide a regret bound. Qualitatively, the algorithm does not change by resetting after $h+1$ time steps, provided the environment allows such an action. Finally, after interacting for H steps, the environment is reset, we collect the transitions of time steps $h-1$ and h in separate datasets and proceed to roll-in at time step $h+1$. The necessity of two separate datasets will be made clear in the next section.

Representation learning (Lines 16-21)

In the representation learning phase, we learn representations using the samples collected in the exploration phase, define a UCB-style bonus term and set the estimated transition operator which will be used for planning. Similar to [TPT+22], we employ a constrained optimization objective (Line 18), to learn features that have good spectral properties and approximate the transition operator well enough. In particular, the representation learning objective forces representations to be the maximum likelihood solution to fitting the unknown transition operator, while the feature maps additionally are optimized to span the feature space. Let us define the objective functions that enforce the desired properties over some dataset $\mathcal{D}$:

$$\mathcal{L}^{\text{likelihood}}(\phi_h, \mu_h, \mathcal{D}) = \sum_{(s,a,s')\in\mathcal{D}} \log(\langle \phi_h(s,a), \mu_h(s')\rangle \tag{3.1}$$

$$\mathcal{L}^{\text{unisoft}}(\phi_h, \mathcal{D}) = -\lambda_{\min}\left(\sum_{(s,a)\in\mathcal{D}} \phi_h(s,a)\phi_h(s,a)^T\right) \tag{3.2}$$

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:

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

Similar to previous works on low-rank MDPs [AKKS20, UZS22, CHYL23], as a computational abstraction, we assume access to an optimization oracle.

Definition 3.1.1. *(Optimization Oracle) Consider the function class $\Phi \times \Psi$ and a dataset $\mathcal{D}$ consisting of (s,a,s') triples, the optimization oracle returns, for any $h \in [H]$,*

$$\operatorname*{argmin}_{\phi\in\Phi_h^{\text{MLE}}(\mathcal{D})} \mathcal{L}^{\text{unisoft}}(\phi, \mathcal{D}).$$

Note that in practice, the above oracle can be approximated reasonably well [TPT+22, ZRY+22]. In algorithm 2, we leverage the samples of both datasets collected during exploration to obtain an approximation guarantee for the transition operator on transitions collected while rolling-out (Lemma A.7.6). We then use the learned features to define an UCB-style bonus term (Line 20) which intuitively represents the uncertainty in the direction of some feature vector. In particular, [CHYL23] found that the bonus term

serves as a trajectory-wise uncertainty measure, in the sense of Lemma A.3.2. Note that the bonus term employs samples collected only from one dataset, which allows us to leverage the elliptical potential Lemma A.7.4 to show decreasing confidence intervals. Finally, we define the estimated transition operator in Line 21.

Planning (Line 25)

In the last phase, we find an optimal policy for the bonus augmented reward function in the estimated environment obtained during the representation learning phase. Note that this step does not require any interactions with the true environment. However, 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 $\operatorname{argmax}_{\pi \in \Pi} V^{\pi, d_1}_{\mathcal{P}, r, 1}$. We note, that planning in a known linear MDP can be done efficiently by, for example, LSVI-UCB [JYWJ20].

In the following lemma we provide a baseline worst-case regret bound for algorithm 2, which only requires the realizability assumption to hold. We denote the regret incurred by algorithm 2 as $\tilde{\mathcal{R}}$, which differs from the regret incurred by the behavior polices $\{\pi_t\}_{t=1}^T$ denoted as $\mathcal{R}$.

Lemma 3.1.1 (Regret bound without UniSOFT representations)**.** *Let $\xi_t = t^{-1/4}$. Suppose Assumption 2.1.1 (realizability) holds. Then, for any $T \in \mathbb{N}$, algorithm 2 satisfies:*

$$\mathbb{E}[\tilde{\mathcal{R}}(T)] \lesssim H^3 d^{3/2} |\mathcal{A}| T^{3/4} \log^2(TH|\Phi||\Psi|) = \tilde{O}(H^3 d^{3/2} |\mathcal{A}| T^{3/4})$$

Proof. The proof is given in Appendix A.1. □

We see that the regret bound does not depend on the size of the state space, which lets us generalize to infinite state spaces. Additionally, the size of the function space does only appear in a logarithmic term which would allow us to extend the result to hold for infinite function classes with bounded statistical complexity [AKKS20]. Importantly, the regret bound is worst case and holds without leveraging good features.

In the following sections we will see how the result above can be improved for environments with minimal sub-optimality gap (Assumption 2.1.3). In particular, we will see how good representations can be leveraged to improve the learning efficiency.

3.2. Instance-Dependent Regret Bounds

Our general strategy for improving the baseline regret given in the previous section, is to show that there exists an episode after which algorithm 2 only selects good representations. Then, these good representations provide more efficient exploration and we gain an improvement in learning efficiency. Hence, the results provided in this chapter will only improve upon the baseline regret result if we run the algorithm for long enough.

We start by introducing representations that approximately represent the ground truth transition operator over the support of the occupancy distribution of the optimal policy.

Definition 3.2.1. *($\alpha^\star$-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 $h \in [H]$,*

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

Remark 3.2.1. *Note that $\alpha^\star$-approximate representations are closely related to the notion of ϵ-accurate system identification [CHYL23]. Additionally, note that the set of $\alpha^\star$-approximate representations $\Phi_\alpha \times \Psi_\alpha \subseteq \Phi \times \Psi$ is non-empty for any $\alpha \geq 0$, whenever the realizability assumption 2.1.1 holds.*

Interestingly, we can show that the optimization oracle (Definition 3.1.1) will converge uniformly over the occupancy distribution of the optimal policy (Lemma A.2.1), provided that said distribution is well defined for all state-action pairs, i.e., Assumption 2.1.4 (minimal optimal occupancy) holds. The following assumption exploits this convergence and ensures that we are guaranteed to find a good representation.

Assumption 3.2.1. *(α-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 $(s,a) \in \mathcal{S} \times \mathcal{A}$, for which there exists a policy $\pi \in \Pi$, such that for any $h \in [H]$, we have $d^\pi_{\mathcal{P}^\star,h}(s,a) > 0$.*

With this assumption at hand, we can show that the UniSOFT loss in equation 3.2 eventually eliminates all redundant and all non-UniSOFT feature maps (Lemma A.2.4). 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 non-redundant UniSOFT features minimize the UniSOFT loss, which are guaranteed to exist by the expressiveness assumption above.

Theorem 3.2.1 (Instance-dependent regret with UniSOFT representations)**.** *Let $\xi_t = t^{-1/3}$ and $\alpha \in (0, 1]$. Suppose assumptions 2.1.1 (realizability), 2.1.3 (minimal suboptimality gap), 3.2.1 (α-expressive function space) and 2.1.2 (unique optimal policy) hold. Additionally, if $\alpha < 1$, suppose assumption 2.1.4 (minimal optimal occupancy) holds. Then, for any $T \in \mathbb{N}$ large enough, algorithm 2 satisfies:*

$$\begin{aligned}\mathbb{E}[\tilde{\mathcal{R}}(T)] &\lesssim H\tau^{5/6} + \frac{1}{\lambda^\star_{\max}} H^4 d^{1/2} |\mathcal{A}|^{1/2} T^{2/3} \log(TH|\Phi||\Psi|) \\ &\lesssim \tilde{O}(\frac{1}{\lambda^\star_{\max}} H^4 \sqrt{d|\mathcal{A}|} T^{2/3}),\end{aligned}$$

where

$$\begin{aligned}\tau &\lesssim \{\kappa_3^m \cdot \log^{2m}(\kappa_3 \cdot \kappa_2) \vee \kappa_1^m \cdot \log^{2m}(\kappa_1 \cdot \kappa_2)\} \\ &\lesssim \frac{H^{12} d^9 |\mathcal{A}|^6}{\Delta_{\min}\{\alpha d^\star_{\min} \wedge \lambda^\star_{\max}\}} \cdot \log^{12}(TH^3 d^{3/2} |\mathcal{A}||\Phi||\Psi|),\end{aligned}$$

with $\kappa_1 = \frac{H^2 d^{3/2}|\mathcal{A}|}{\alpha \Delta_{\min} d^\star_{\min}}$, $\kappa_2 = TH|\Phi||\Psi|$, $\kappa_3 = \frac{H^2 d^{3/2}|\mathcal{A}|}{\lambda^\star_{\max}\Delta_{\min}}$ *and* $\lambda^\star_{\max} = \min_{\tilde{\alpha}\leq\alpha} \max_{\phi\in\Phi^{\text{unisoft}}_{\tilde{\alpha}}} \lambda^\star(\phi)$. *In particular,* T *must be large enough such that* $T \geq \tilde{O}(\tau)$ *holds.*

Proof. The proof is provided in Appendix A.3. □

Remark 3.2.2. *If* $\alpha = 1$*, then all representations are* $\alpha^\star$*-approximate and we don't require assumption 2.1.4 (minimal optimal occupancy) to ensure the selection thereof. In particular,* τ *would loose its dependence in* α *and* $d^\star_{\min}$*.*

Remark 3.2.3. *The actual regret should be read as the minimum of the baseline result in Lemma 3.1.1 and the bound given in the Theorem above.*

Compared to the baseline result in Lemma 3.1.1, we see that, except for the horizon, the dependence in each relevant variable has improved. However, we gain an additional dependence in $\lambda^\star_{\max}$, which intuitively captures how efficiently we can explore the feature space by playing optimally. Furthermore, we require additional assumptions that generally are hard to check in practice. Nevertheless, whenever the function space is expressive enough, such that we can choose $\alpha = 1$, we only require the minimal optimality gap to exist. The unique optimal policy assumption can be dropped, in exchange for a slightly stronger assumption on the function space, as we will see in the next section (Section 3.3).

On a high level, τ captures the number of episodes algorithm 2 needs to eliminate all non-UniSOFT representations. Hence, the theorem tells us, that after interacting with the environment for some number of "warm-up" episodes τ, during which we incur expected regret according to the parameter adjusted baseline result of Lemma 3.1.1, 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 and in particular, on how sharp the available baseline regret result is.

In order to eliminate all non-UniSOFT-representations, the algorithm must first eliminate all non-$\alpha^\star$-approximate representations, as otherwise, there is no guarantee that any UniSOFT representation fulfills the MLE constraint. This is captured by τ's dependence in α and $d^\star_{\min}$. Then, the algorithm must collect enough samples from the optimal occupancy distribution, such that the UniSOFT loss in equation 3.2 leads to the selection of UniSOFT feature maps. Here, τ incurs its dependence in $\lambda^\star_{\max}$.

Whenever the minimal optimal occupancy is well defined, we get an even stronger result. In particular, we then can show that the behavior policies of algorithm 2 will eventually be optimal and the only regret we incur is due to the exploration schedule ξ_t.

Theorem 3.2.2 (Instance-dependent sub-linear expected regret with UniSOFT representations and constant pseudo-regret)**.** *Let* $\alpha > 0$*,* $\gamma \in (2, 4]$ *and* $\xi_t = t^{-1/\gamma}$*. Suppose assumptions 2.1.1 (realizability), 2.1.2 (unique optimal policy), 2.1.3 (minimal sub-optimality*

gap), 2.1.4 (minimal optimal occupancy) and 3.2.1 (α-expressive function space) hold. Then for any $T \in \mathbb{N}$ large enough, algorithm 2 satisfies:

$$\mathbb{E}_{\delta,\xi}[\tilde{\mathcal{R}}(T)] \lesssim H(\tau^\star)^{1/2+1/\gamma} + HT^{\frac{\gamma-1}{\gamma}} \lesssim \tilde{O}(HT^{\frac{\gamma-1}{\gamma}}),$$

where

$$\begin{aligned}\tau^\star &\lesssim \{\kappa_3^m \cdot \log^{2m}(\kappa_3 \cdot \kappa_2) \vee \kappa_1^m \cdot \log^{2m}(\kappa_1 \cdot \kappa_2) \vee \kappa_4^{m'} \cdot \log^{m'}(\kappa_4 \cdot \kappa_2)\} \\ &\lesssim \left(\frac{H^3 d^2 |\mathcal{A}|}{\alpha \lambda^\star_{\max}(\Delta_{\min} d^\star_{\min})^2}\right)^{\frac{2\gamma}{\gamma-2}} \cdot \left(\log(\frac{TH^4 d^2 |\mathcal{A}|\Phi||\Psi|}{\alpha \lambda^\star_{\max}(\Delta_{\min} d^\star_{\min})^2})\right)^{\frac{4\gamma}{\gamma-2}},\end{aligned}$$

with $\kappa_1 = \frac{H^2 d^{3/2}|\mathcal{A}|}{\alpha \Delta_{\min} d^\star_{\min}}$, $\kappa_2 = TH|\Phi||\Psi|$, $\kappa_3 = \frac{H^2 d^{3/2}|\mathcal{A}|}{\lambda^\star_{\max}\Delta_{\min}}$, $\kappa_4 = \frac{H^3 d^2 |\mathcal{A}|}{(\Delta_{\min} d^\star_{\min})^2 \lambda^\star_{\max}}$, $m = \frac{2\gamma}{\gamma-2}$ *and* $m' = \frac{\gamma}{\gamma-1}$. *In particular, T must be large enough such that $T \geq \tilde{O}(\tau^\star)$ holds.*

Proof. The proof is given in Appendix A.4. □

Notably, compared to the previous theorem, we can now choose $\gamma < 3$ to get an improved dependence in T. However, decreasing γ worsens the baseline regret result and we require more warm-up episodes. The term $\tau^\star$ now additionally captures the number of episodes we require to fully explore the feature space with UniSOFT representations, and subsequently identify the optimal policy. Any further improvements in expected regret of algorithm 2 are limited by the baseline regret result.

However, if we assume the quantities $\Delta_{\min}$ and $d^\star_{\min}$ to be known, we can design a termination criteria, which stops the algorithm whenever the behavior policy is optimal. Algorithm 3 extends algorithm 2 by an evaluation phase, in which we measures the uncertainty in estimating the transition operator by the maximal value achieved in the learned model, where the bonus term serves as the reward function. If this uncertainty is below some threshold we stop the algorithm and return the optimal policy with high probability.

Theorem 3.2.3 (Optimal policy identification)**.** *Let $\alpha > 0$, $\gamma \in (2, 4]$ and $\xi_t = t^{-1/\gamma}$. suppose the quantities $\Delta_{\min}$ and $d^\star_{\min}$ are known. Let T be large enough such that $T \geq \tau^\star$ holds. Then, under the same assumptions as in Theorem 3.2.2, with probability at least $1-2T^{-1}$, algorithm 3 will return the optimal policy after at most T episodes. In particular, if $\gamma = 4$,*

$$T \geq \tilde{O}\left(\frac{H^{12} d^8 |\mathcal{A}|^4}{(\alpha \lambda^\star_{\max})^4 (\Delta_{\min} d^\star_{\min})^8}\right).$$

Proof. The proof is given in Appendix A.4. □

Remark 3.2.4. *We can equivalently say that algorithm 3 enjoys constant high-probability regret, by employing the identified optimal policy in the environment.*

3.3. Further Results

In this section we provide sufficient conditions for the existence of UniSOFT representations and show that we are able to drop the restrictive unique optimal policy assumption.

Existence of good representations

We start by characterizing the situation in which d coincides with the rank of the transition operator $\mathcal{P}^\star$; that is, $\text{rank}(\mathcal{P}_h^\star) = d$ for all $h \in [H]$. The following Lemma provides a necessary and sufficient condition for the existence of exact UniSOFT representations. In particular, the occupancy distribution of the optimal policy must be non-zero for at least d state-action pairs at every time step h.

Lemma 3.3.1 (Existence of exact UniSOFT representations with minimal dimension)**.** *Assume that* $\text{rank}(\mathcal{P}_h^\star) = \tilde{d}$ *for each* $h \in [H]$*. Let* $\mathcal{X}_h := \{(s,a) \in \mathcal{S} \times \mathcal{A} | d^{\pi^\star}_{\mathcal{P}^\star,h}(s,a) > 0\}$ *be the set of state-action pairs reachable by the optimal policy at time step* $h \in [H]$*. Then, the following statements are equivalent:*

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

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

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

Proof. The proof is given in the Appendix A.5. □

Remark 3.3.1. *We immediately gain* (1) *as a sufficient condition for values of* $d \geq \text{rank}(\mathcal{P}^\star)$.

Importantly, this is also a negative result for exact representations with minimal dimension. In particular, whenever we choose d to be minimal, we might not be able to learn representations that are UniSOFT and exact. As slightly misspecified representations can still hold the UniSOFT property, we find that good representations are not necessarily exact.

Since the linear independence property of a set of vectors is robust to small perturbations (Lemma A.5.1), the above existence result holds for all models that approximate the true transition model on the support of the occupancy distribution of the optimal policy well enough.

Lemma 3.3.2. *Assume that* $\text{rank}(\mathcal{P}_h^\star) = \tilde{d}$ *for each* $h \in [H]$ *and that assumption 2.1.4 (minimal optimal occupancy) holds. Further assume that* $\mathcal{P}^\star$ *admits an UniSOFT representation. Then, there exists an* $\epsilon > 0$ *such that all* $\alpha^\star$*-approximate representations* $\langle \phi, \mu \rangle_{\mathbb{R}^{\tilde{d}}} \equiv \hat{\mathcal{P}}$ *with* $\alpha \leq \epsilon$ *are UniSOFT.*

Proof. The proof is given in the Appendix A.5. □

Remark 3.3.2. *On a high level, ϵ is upper bounded by the degree of linear independence between the (unknown) transition vectors corresponding to optimal actions, which can generally be arbitrarily small.*

The Lemma tells us, that whenever we learn representations with minimal dimension, they are guaranteed to eventually hold the UniSOFT property, provided that the environment fulfills the necessary condition in Lemma 3.3.1. In the context of the previous section, we could drop the α-expressiveness assumption, solely optimize the MLE loss and we would still end up selecting good representations.

Multiple optimal policies

As noted in the preliminary chapter, we can extend our results to hold for 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. We adjust the notion of $\alpha^\star$-approximate representations accordingly.

Definition 3.3.1 ($(\sigma^\star, \alpha)$-Approximate Representation)**.** *A representation $(\phi, \mu) \in \Phi \times \Psi$, with induced model $\mathcal{P}$, is $(\sigma^\star, \alpha)$-approximate if for the finite sequence $\sigma^\star = (\pi_1^\star, \pi_2^\star, ..., \pi_t^\star)$ of optimal policies and for all $h \in [H]$,*

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

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 3.3.1 (α-Expressive Function Space)**.** *Let $\sigma^\star$ be an arbitrary sequence of optimal polices of finite length. For all $(\sigma^\star, \alpha)$-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 $d^{\pi}_{\mathcal{P}^\star,h}(s,a) > 0$.*

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, to which distribution of optimal polices the algorithm converges. In exchange for updating the expressiveness assumption 3.2.1 to the more restrictive assumption 3.3.1, 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.

Algorithm 2 UniSREP-UCB (Upper Confidence Bound driven Universally Spanning Representation Learning)

Input: Function spaces $\{\Phi_h\}_{h=1}^H, \{\Psi_h\}_{h=1}^H$, Parameters $\lambda_t, \hat{\alpha}_t, \xi_t$ deacreasing, T
Output: π_t

Initialize: $\mathcal{D}_{0,h} = \emptyset$, $\mathcal{D}'_{0,h} = \emptyset$, $\pi_{0,h} \equiv \mathcal{U}(\mathcal{A})$, $\forall h \in [H]$
for $t = 1, ..., T$ **do**
  // Interact with the MDP and collect transition data
  $e_t \sim \mathrm{Ber}(1 - \xi_t)$
  for $h = 1, ..., H$ **do**
    $s \sim d^{\pi_{t-1}}_{P^\star;h-1}$
    if $e_t = 1$ **then**
      $a = \pi_{t-1,h-1}(s)$, $s' \sim P^\star_{h-1}(\cdot|a, s)$, $a' = \pi_{t-1,h}(s')$, $s'' \sim P^\star_h(\cdot|a', s')$
    else
      $a \sim \mathcal{U}(\mathcal{A})$, $s' \sim P^\star_{h-1}(\cdot|a, s)$, $a' \sim \mathcal{U}(\mathcal{A})$, $s'' \sim P^\star_h(\cdot|a', s')$
    end if
    $\mathcal{D}_{t,h} = \mathcal{D}_{t-1,h} \cup \{(s, a, s')\}$, and $\mathcal{D}'_{t,h} = \mathcal{D}'_{t-1,h} \cup \{(s', a', s'')\}$
  end for

  // Learn representations
  for $h = 1, ..., H$ **do**
    ▷ Learn UniSOFT representations via MLE constraint
    $\hat{\phi}_{t,h} = \arg\min_{\phi \in \Phi_h^{\mathrm{MLE}}(\mathcal{D}_{t,h} \cup \mathcal{D}'_{t,h})} \mathcal{L}^{\mathrm{unisoft}}(\phi, \mathcal{D}_{t,h} \cup \mathcal{D}'_{t,h})$
    $\hat{\Sigma}_{t,h} = \sum_{(s,a) \in \mathcal{D}_{t,h}} \hat{\phi}_{t,h}(s, a) \hat{\phi}_{t,h}(s, a)^T + \lambda_t I$ ▷ Update covariance matrix
    $\hat{b}_{t,h}(s, a) = \min\{\hat{\alpha}_t \sqrt{\hat{\phi}_{t,h}(s, a)^T \hat{\Sigma}^{-1}_{t,h} \hat{\phi}_{t,h}(s, a)}, 1\}$ ▷ Set bonus
    $\hat{\mathcal{P}}_{t,h}(s'|s, a) = \langle \hat{\phi}_{t,h}(s, a), \hat{\mu}_{t,h}(s') \rangle$ ▷ Update transition operator
  end for

  // Update (deterministic) policy
  $\pi_t = \mathrm{argmax}_{\pi \in \Pi} V^{\pi, d_1}_{\hat{\mathcal{P}}_t, \hat{b}_t + r^\star, 1}$
end for
return π_t

Algorithm 3 UniSREP-UCB+

Input: Function spaces $\{\Phi_h\}_{h=1}^H, \{\Psi_h\}_{h=1}^H$, Parameters $\lambda_t, \hat{\alpha}_t, \xi_t$ deacreasing, T
Output: π_t

Initialize: $\mathcal{D}_{0,h} = \emptyset$, $\mathcal{D}'_{0,h} = \emptyset$, $\pi_{0,h} \equiv \mathcal{U}(\mathcal{A})$, $\forall h \in [H]$
for $t = 1, ..., T$ **do**
  // Interact with the MDP and collect transition data
  for $h = 1, ..., H$ **do**
    $s \sim d^{\pi_{t-1}}_{P^\star;h-1}$
    if $\mathrm{Ber}(1-\xi_t)$ **then**
      $a = \pi_{t-1,h-1}(s)$, $s' \sim P^\star_{h-1}(\cdot|a,s)$, $a' = \pi_{t-1,h}(s')$, $s'' \sim P^\star_h(\cdot|a',s')$
    else
      $a \sim \mathcal{U}(\mathcal{A})$, $s' \sim P^\star_{h-1}(\cdot|a,s)$, $a' \sim \mathcal{U}(\mathcal{A})$, $s'' \sim P^\star_h(\cdot|a',s')$
    end if
    $\mathcal{D}_{t,h} = \mathcal{D}_{t-1,h} \cup \{(s,a,s')\}$, and $\mathcal{D}'_{t,h} = \mathcal{D}'_{t-1,h} \cup \{(s',a',s'')\}$
  end for

  // Learn representations
  for $h = 1, ..., H$ **do**
    ▷ Learn UniSOFT representations via MLE constraint
    $\hat{\phi}_{t,h} = \arg\min_{\phi \in \Phi_h^{\mathrm{MLE}}(\mathcal{D}_{t,h} \cup \mathcal{D}'_{t,h})} \mathcal{L}^{\mathrm{unisoft}}(\phi, \mathcal{D}_{t,h} \cup \mathcal{D}'_{t,h})$
    $\hat{\Sigma}_{t,h} = \sum_{(s,a)\in\mathcal{D}_{t,h}} \hat{\phi}_{t,h}(s,a)\hat{\phi}_{t,h}(s,a)^T + \lambda_t I$ ▷ Update covariance matrix
    $\hat{b}_{t,h}(s,a) = \min\{\hat{\alpha}_t \sqrt{\hat{\phi}_{t,h}(s,a)^T \hat{\Sigma}^{-1}_{t,h} \hat{\phi}_{t,h}(s,a)}, 1\}$ ▷ Set bonus
    $\hat{\mathcal{P}}_{t,h}(s'|s,a) = \langle \hat{\phi}_{t,h}(s,a), \hat{\mu}_{t,h}(s') \rangle$ ▷ Update transition operator
  end for

  // Update (deterministic) policy
  $\pi_t = \mathrm{argmax}_{\pi\in\Pi} V^{\pi,d_1}_{\hat{\mathcal{P}}_t, \hat{b}_t + r^\star, 1}$

  // Learn policy for optimality condition
  $\pi^b_t = \mathrm{argmax}_{\pi\in\Pi} V^{\pi,d_1}_{\hat{\mathcal{P}}_t, \hat{b}_t, 1}$

  // Evaluate policy
  $c_t = 10H^2 (V^{\pi^b_t,d_1}_{\hat{\mathcal{P}}_t, \hat{b}_t, 1} + \sqrt{\frac{|\mathcal{A}|}{\xi_t}\zeta_t})$

  // Check for optimality
  if $c_t < \Delta_{\min} d^\star_{\min}$ **then**
    return π_t
  end if
end for
return π_t

4. Discussion

4.1. Comparison with the Literature

Here we compare the optimal policy identification (constant regret) result of Theorem 3.2.3 with related results from the literature. In Table 4.1 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 incurs no additional regret.

LSVI-LEADER [PTP$^+$21]

The LSVI-LEADER algorithm proposed by [PTP$^+$21] achieves constant instance-dependent regret, when provided with a set of exact representations containing one exact UniSOFT representation and the unique optimal policy assumption 2.1.2 holds. Their algorithm does not scale to large function classes as it learns a different representation for each state-action pair, which becomes computationally intractable when the function space is large [TPT$^+$22]. Algorithm 3 does not suffer form this issue, as we learn only one representation per episode. Additionally, we extend our results beyond the unique optimal policy assumption as shown in Appendix A.6. However, we need to assume access to an optimization oracle and that certain instance-dependent quantities are known, which might be unreasonable in practice.

In Table 4.1 we can see that, in contrast to LSVI-LEADER, the critical episode of Algorithm 3 depends on the size of the action space, which seems to be unavoidable in low-rank MDPs [ZYW$^+$24]. We additionally incur a dependence on $d^\star_{\min}$, which stems from bounding average sub-optimality gaps. The overall smaller polynomial dependence

Table 4.1.: Critical episodes for algorithms achieving constant regret.

Algorithm	Setting	Critical Episode
LEADER [PTR$^+$21]	CLB	$\tilde{O}((\frac{d}{\lambda^\star \Delta_{\min}})^2)$
BanditSRL [TPT$^+$22]	CLB	$\tilde{O}(\frac{d^2}{(\lambda^\star \Delta_{\min})^2 \epsilon_{\min}})$
LSVI-LEADER [PTP$^+$21]	Linear MDP	$\tilde{O}(\max\{\frac{d^3H^4}{(\lambda^\star)^2}, \frac{d^2H^4}{\Delta^2_{\min}(\lambda^\star)^3}\})$
UniSREP (Our)	Low-rank MDP	$\tilde{O}(\frac{H^{12}d^8\lvert\mathcal{A}\rvert^4}{(\Delta_{\min} d^\star_{\min})^8(\alpha\lambda^\star)^4})$

for LSVI-LEADER follows from the overall tighter regret bound available for linear MDPs compared to low-rank MDPs.

LEADER [PTR+21]

In contextual linear bandits (CLB), the LEADER algorithm enjoys constant regret, when provided with a set of exact representations containing one exact HLS representation. A representation ϕ is HLS if

$$\lambda_{\min}\left(\mathbb{E}_{s\sim p}[\phi(s,a)\phi(s,a)^T]\right) > 0,$$

where $p \in \Delta(\mathcal{S})$ is some distribution over the state space. Note that in CLBs the feature map ϕ must only linearly represent the reward function.

The LEADER algorithm suffers from similar issues as its counterpart for linear MDPs (LSVI-LEADER). However, for CLBs, [TPT+22] were able to provide an algorithm that achieves constant regret without assuming access to a set of exact representations.

BanditSRL [TPT+22]

The BanditSRL algorithm achieves a constant instance-dependent regret result for CLBs when provided with a finite function class. However, [TPT+22] 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:

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

Note that this assumption is about representing the reward function, and hence is conceptually different to representing a transition model. However, we want to emphasize that by leveraging the MLE oracle we can deal with misspecified representations without making additional assumptions on the level of misspecification. In particular, we only need one realizable representation to exist and can identify the optimal policy with representations that have low misspecification error on average. Additionally, by leveraging the MLE oracle, eliminating all point-wise misspecified representations is generally not possible in the first place, since the MLE objective is unbounded and we can not use any standard uniform convergence techniques [AKKS20].

Table 4.1 shows that the critical episode of BanditSRL also depends on an additional instance-dependent property when compared to the critical episode achieved by LEADER. This stems from the fact that the LEADER algorithm assumes that all representations are realizable, whereas BanditSRL first needs to eliminate all non-realizable ones, which depends on the misspecification level $\epsilon_{\min}$ introduced above.

Constant regret with misspecified representations

Interestingly, as far as we know, there exists no algorithm for linear MDPs, that can identify the optimal policy, when features are only required to have small misspecification error on average. In fact, only very recently, [ASS+23] provided the first sublinear regret result in this setting. On the other hand, [ZFHG24] provided an algorithm that achieves constant instance-dependent regret for linear-MDPs for features that have low point-wise misspecification w.r.t. the minimal sub-optimality gap.

4.2. Limitations

Infinite Function classes

We note that all results are easily extended to hold for infinite function classes with bounded statistical complexity (e.g. Rademacher complexity [BM02]) as we leveraged the finiteness only through uniform convergence arguments [AKKS20].

Infinite Action Space

We can extend the results to hold for an infinite action space if the transition operator satisfies a Hölder smoothness condition w.r.t. actions and the reward function is also Hölder smooth [OBK24]. The complexity bounds would then depend on the order of smoothness instead of the size of the action space.

Infinite State Space

All results easily generalize to infinite state spaces, by interchanging all sums over the state space with integrals. The existence results however, require more careful arguments as $\mathcal{P}$ is now a matrix with infinitely many rows and columns. We leave such an extension for future work.

Redundant Features

Throughout this thesis we have introduced assumptions such that we are guaranteed to select non-redundant features. We note that, by following a similar analysis as in [PTP+21], our results would also hold for UniSOFT feature maps that are redundant, provided that we are guaranteed to select them. In order to learn UniSOFT feature maps that are redundant, [TPT+22] provided the following loss function

$$\mathcal{L}^{\text{unisoft}}_{\text{weak}}(\phi, \mu, \mathcal{D}) = \min_{(s,a)\in\mathcal{D}} \phi(s,a)^T \left(\sum_{(s',a')\in\mathcal{D}} \phi(s',a')\phi(s',a')^T \right) \phi(s,a).$$

However, this loss function only leads to selecting UniSOFT feature maps, if we are guaranteed to visit any state-action pair in finite time. Hence, we would require a reachability assumption, which guarantees that we will eventually sample all state-action

pairs reachable by any policy. Otherwise, we can not be sure that the embeddings of optimal actions actually span the observable feature space. We leave such an extension for future work.

Unknown reward function

We can additionally estimate an unknown reward function, as long as the reward function is linear in the feature maps. We then would require an additional UCB-style bonus term that accounts for the estimation uncertainty, similar to [TPT+22]. Note that any representation that accurately represents the transition operator in $\mathbb{R}^d$ can easily be extended to additionally represent the reward function in $\mathbb{R}^{d+1}$. We leave such an extension for future work.

Low-rank assumption

Assuming low-rank structure offers a significant advantage in statistical complexity, as it allows us to provide learning guarantees that scale with the feature dimension instead of the size of the state space, which can generally be uncountable large. However, as noted in [LO24], the set of MDPs that satisfy a low-rank representation with small rank d w.r.t. $|\mathcal{S}|$ is inherently limited. In particular, [LO24] 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 [LO24] for a more thorough discussion.

Computation

Computationally, algorithm 2 suffers from similar limitations as most other existing works for low-rank MDPs. In particular, the optimization oracle can not be accurately solved efficiently, as there exists no practical mechanism for guaranteeing the normalization conditions for ϕ and μ [ZRY+22]. This, in particular, makes the constraint optimization objective in algorithm line 18 intractable. However, as shown in [ZRY+22], we can approximate the MLE objective with noise contrastive estimation (NCE) and similar to [TPT+22], can add the UniSOFT loss as a regularization term, as a reasonable proxy. Additionally, by Lemma 15 in [AKKS20], we can leverage the LSVI-UCB algorithm [JYWJ20] to find, with probability at least $1-\delta$, a policy that is ϵ-optimal, after sampling $\tilde{O}(\frac{d^3H^6\log(2/\delta)}{\epsilon^4})$ transitions.

5. Conclusions & Future Work

In this thesis we studied low-rank MDPs characterized by the instance-dependent properties $\Delta_{\min}$ (minimal sub-optimality gap) and $d^{\star}_{\min}$ (minimal optimal occupancy). We showed that the existing algorithm REP-UCB, augmented with a double exploration strategy and a constrained optimization objective, can leverage good representations for more efficient exploration. Additionally, we demonstrated that optimal policy identification is possible for low-rank MDPs and provided sufficient conditions for the existence of good representations.

An interesting direction for future work, would be to design computationally efficient variants of our proposed algorithms and test them on deep RL benchmarks. Additionally, our results permit several extensions to more general settings as noted in the section on limitations. Finally, whether constant regret is possible for low-rank MDPs without knowledge on the minimal sub-optimality gap or the minimal optimal occupancy is an important open problem worth investigating.

Bibliography

[AKKS20] 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.

[ASS+23] 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.

[AYPS11] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing systems*, 24, 2011.

[BM02] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[CHYL23] Yuan Cheng, Ruiquan Huang, Jing Yang, and Yingbin Liang. Improved sample complexity for reward-free reinforcement learning under low-rank mdps. *arXiv preprint arXiv:2303.10859*, 2023.

[HLS20] 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.

[HZG21] 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.

[HZQ+22] 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.

[JYWJ20] 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.

[LO24] Joongkyu Lee and Min-hwan Oh. Demystifying linear mdps and novel dynamics aggregation framework. In *The Twelfth International Conference on Learning Representations*, 2024.

[MCK+24] 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.

[OBK24] Miruna Oprescu, Andrew Bennett, and Nathan Kallus. Low-rank mdps with continuous action spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 4069–4077. PMLR, 2024.

[OVR16] Ian Osband and Benjamin Van Roy. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*, 2016.

[PO99] Robert Piziak and Patrick L Odell. Full rank factorization of matrices. *Mathematics magazine*, 72(3):193–201, 1999.

[PTP+21] 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, 2021.

[PTR+21] 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, 2021.

[SJ19] Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32, 2019.

[TPT+22] 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.

[Tro12] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12:389–434, 2012.

[UZS22] 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.

[ZFHG24] Weitong Zhang, Zhiyuan Fan, Jiafan He, and Quanquan Gu. Settling constant regrets in linear markov decision processes. *arXiv preprint arXiv:2404.10745*, 2024.

[ZRY+22] 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.

[ZSU+2a] 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.

[ZYW+24] 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.

$\mathcal{S}$	Finite state space
$\mathcal{A}$	Finite action space
H	Horizon
$r_h^\star : \mathcal{S} \times \mathcal{A} \to [0,1]$	Reward function at time step h
$r^\star = \prod_{h\in[H]} r_h^\star$	Reward function
$\mathcal{P}_h^\star : \mathcal{S} \times \mathcal{A} \to \Delta\mathcal{S}$	Transition operator at time step h
$\mathcal{P}^\star = \prod_{h\in[H]} \mathcal{P}_h^\star$	True transition operator
d_1	Initial state distribution
$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}^\star, r^\star, H, d_1)$	MDP
T	Number of total episodes
d	Rank of $\mathcal{M}$
$\pi_h : \mathcal{S} \to \mathcal{A}$	Deterministic policy at time step h
$\pi = \prod_{h=1}^H \pi_h$	Deterministic policy
Π	Deterministic policy space
$\Pi^\star$	Subspace of optimal policies
$\Pi_h^\star(s)$	Optimal actions at state s and time step h
$V_{\mathcal{P},r;h}^\pi(s) = \mathbb{E}[\sum_{i=h}^H r_i(s_i,a_i) \vert s_h = s, \mathcal{P}, \pi]$	State value function
$V_{\mathcal{P},r;1}^{\pi,d_1} = \mathbb{E}_{s\sim d_1}[V_{\mathcal{P},r;1}^\pi(s)]$	Expected value function at the initial state distribution
$\mathcal{P}_h V_{\mathcal{P},r,h+1}^\pi(s,a) = \mathbb{E}_{s'\sim\mathcal{P}_h(\cdot\vert s,a)}[V_{\mathcal{P},r,h+1}^\pi(s')]$	Operator view
$Q_{\mathcal{P},r;h}^\pi(s,a) = r_h(s,a) + \mathcal{P}_h V_{\mathcal{P},r;h+1}^\pi(s,a)$	Q-function
$\pi^\star \in \arg\max_{\pi\in\Pi} V_{\mathcal{P}^\star,r^\star,1}^{\pi,d_1}$	Optimal policy
$\mathcal{R}(T) = \sum_{t=1}^T V_{\mathcal{P}^\star,r^\star,1}^{\pi^\star,d_1} - V_{\mathcal{P}^\star,r^\star,1}^{\pi_t,d_1}$	Regret of behavior policies
$\Delta_h(s,a) = V_{\mathcal{P}^\star,r^\star;h}^{\pi^\star}(s) - Q_{\mathcal{P}^\star,r^\star;h}^{\pi^\star}(s,a)$	Sub-optimality gap

Table 1.: Notation

$\phi_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$	Feature map at time step h
$\phi = \prod_{h\in[H]} \phi_h$	Feature map
$\mu_h : \mathcal{S} \to \mathbb{R}^d$	Signed measure at time step h
$\mu = \prod_{h\in[H]} \mu_h$	Signed measure
$\Phi \times \Psi$	Function space
$\Phi^\star \times \Psi^\star$	Subspace of exact representations
$\Phi_\alpha \times \Psi_\alpha$	Subspace of $\alpha^\star$-approx. representations
$\Phi^{\text{unisoft}} \times \Psi$	Subspace of UniSOFT feature maps
$\phi^\star, \mu^\star$	Exact embedding functions
$\hat{\phi}_t, \hat{\mu}_t$	Estimated embedding functions
$\hat{\mathcal{P}}_t \equiv \langle \hat{\phi}_t, \hat{\mu}_t \rangle$	Induced transition operator
$d^\pi_{\mathcal{P},h}(s,a)$	Occupancy distribution of
$d^\pi_{\mathcal{P},h}(s) = \sum_{a\in\mathcal{A}} d^\pi_{\mathcal{P},h}(s,a)$	policy π in model $\mathcal{P}$
$\lambda^\star(\phi) := \min_{h\in[H],\pi^\star\in\Pi^\star} \lambda_{\min}(\mathbb{E}_{(s,a)\sim d^{\pi^\star}_{\mathcal{P}^\star,h}}[\phi_h(s,a)\phi_h(s,a)^T])$	Minimal eigenvalue w.r.t
$\lambda^\star_{\max} = \min_{\tilde{\alpha}\le\alpha} \max_{\phi\in\Phi^{\text{unisoft}}_{\tilde{\alpha}}} \lambda^\star(\phi)$	the optimal occupancy
$\Delta_{\min} := \min_{s\in\mathcal{S},a\in\mathcal{A},h\in[H]:\Delta_h(s,a)>0} \Delta_h(s,a)$	Minimal sub-optimality gap
$d^\star_{\min} = \min_{s\in\mathcal{S},a\in\mathcal{A},h\in[H],\pi^\star\in\Pi^\star:d^{\pi^\star}_{\mathcal{P}^\star,h}(s,a)>0} d^{\pi^\star}_{\mathcal{P}^\star,h}(s,a)$	Minimal optimal occupancy
$f_{t,h}(s,a) := \|\hat{\mathcal{P}}_{h,t}(\cdot\|s,a) - \mathcal{P}^\star_h(\cdot\|s,a)\|_{\text{TV}}$	Total variation distance between estimated and true model
ξ_t	Exploration probability
$\zeta_t = \frac{2\log(4t^2\|\Phi\|\|\Psi\|H/\delta)}{t}$	MLE concentration rate

Table 2.: Notation

$\lambda_t = c_1 d \log(4t^2 H \lvert\Phi\rvert/\delta)$	Parameter
$\alpha_t = \sqrt{4t\zeta_t \frac{\lvert\mathcal{A}\rvert}{\xi_t} + \lambda_t d}$	
$\hat{\alpha}_t = 5\alpha_t$	
$\beta_t = \sqrt{\frac{\lvert\mathcal{A}\rvert}{\xi_t} 40\alpha_t^2 d + \lambda_t d}$	
$\mathcal{D}_{t,h}$	Dataset collected up until episode t
$\Sigma_{\rho_t,\phi} = t\mathbb{E}_{(s,a)\sim\rho_t}[\phi(s,a)\phi(s,a)^T] + \lambda_t I$	Expected covariance matrix
$\hat{\Sigma}_{t,h} = \sum_{(s,a)\in\mathcal{D}_{t,h}} \hat{\phi}_{t,h}(s,a)\hat{\phi}_{t,h}(s,a)^T + \lambda_t I$	Empirical covariance matrix
$\hat{b}_{t,h}(s,a) = \min\{\hat{\alpha}_t \sqrt{\hat{\phi}_{t,h}(s,a)^T \hat{\Sigma}_{t,h}^{-1} \hat{\phi}_{t,h}(s,a)}, 1\}$	Bonus
$\pi_t^b = \arg\max_{\pi\in\Pi} V^{\pi,d_1}_{\hat{\mathcal{P}}_t,\hat{b}_t,1}$	Policy for optimality condition
$\pi_t \in \operatorname{argmax}_{\pi\in\Pi} V^{\pi,d_1}_{\hat{\mathcal{P}}_t,\hat{b}_t+r^\star,1}$	Behavior policy
$\tilde{\pi}_{t,h}(a\lvert s) = \xi_t \cdot \frac{1}{\lvert\mathcal{A}\rvert} + (1-\xi_t)\cdot \pi_{t,h}(a\lvert s)$	Exploration policy
$\tilde{\mathcal{R}}$	Regret incurred by the exploration policy
$\bar{\pi}_{t,h}(a\lvert s) = \frac{1}{t}\sum_{i=1}^t \tilde{\pi}_{t,h}(a\lvert s)$	Average roll-out policy
$\rho_{t,h}(s) = \frac{1}{t}\sum_{i=1}^t d^{\pi_i}_{\mathcal{P}^\star,h}(s)$	Mixture distributions
$\gamma_{t,h}(s,a) = \frac{1}{t}\sum_{i=1}^t d^{\pi_i}_{\mathcal{P}^\star,h}(s,a)$	
$\rho_{t,h}(s,a) = \rho_{t,h}(s)\bar{\pi}_{t,h}(a\lvert s)$	
$\rho'_{t,h}(s') = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \rho_{t,h-1}(s,a)\mathcal{P}_h^\star(s'\lvert s,a)$	
$\rho'_{t,h}(s,a) = \rho'_{t,h}(s)\bar{\pi}_{t,h}(a\lvert s)$	

Table 3.: Notation

$\mathcal{E}_1 = \{\mathbb{E}_{(s,a)\sim\rho_{t,h}}[f_{t,h}(s,a)^2] \le \zeta_t \text{ and } \mathbb{E}_{(s,a)\sim\rho'_{t,h}}[f_{t,h}(s,a)^2] \le \zeta_t\}$	MLE guarantee
$\mathcal{E}_2 = \{\frac{1}{5}\|\hat{\phi}_{t,h}(s,a)\|_{\Sigma^{-1}_{\rho_{t,h},\hat{\phi}_{t,h}}} \le \|\hat{\phi}_{t,h}(s,a)\|_{\hat{\Sigma}^{-1}_{t,h}} \le 3\|\hat{\phi}_{t,h}(s,a)\|_{\Sigma^{-1}_{\rho_{t,h},\hat{\phi}_{t,h}}}\}$	Bonus Concentration
$\mathcal{F}_1 = \{\Sigma_{t+1,h} \succcurlyeq t\Sigma^\star_{t,h} + \lambda_t I - 2I\sum_{i=1}^t \xi_i - \Delta_{\min}^{-1} g(t) I - 18I\sqrt{t\log(6tdH\lvert\Phi\rvert/\delta)}\}$	Eigenvalue bounds
$\mathcal{F}_2 = \{\Sigma_{t+1,h} \preccurlyeq t\Sigma^\star_{t,h} + \lambda_t I + 2I\sum_{i=1}^t \xi_i + \Delta_{\min}^{-1} g(t) I + 18I\sqrt{t\log(6tdH\lvert\Phi\rvert/\delta)}\}$	
$\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$	Good events
$\mathcal{F} = \mathcal{F}_1 \cap \mathcal{F}_2$	
$\tilde{\pi}^\star_{t,h}(s) = \begin{cases} \pi_{t,h}(s) & \text{if } \pi_{t,h}(s) \in \Pi^\star_h(s) \\ Select(\Pi^\star_h(s)) & \text{otherwise} \end{cases}$	Constructed optimal policy
$\tilde{\gamma}^\star_{t,h}(s,a) = \frac{1}{t}\sum_{i=1}^t d^{\tilde{\pi}^\star_i}_{\mathcal{P}^\star,h}(s,a)$	Optimal mixture distribution

Table 4.: Notation

A. Appendix

In this chapter we will provide the omitted proofs of chapter 3. In particular, Section A.1 provides the proof for the baseline result in Theorem 3.1.1, in Section A.2 we show how we guarantee the selection of good representations and in Section A.3 and Section A.4 we show how good representations can be leveraged to obtain an improved regret bound (Theorem 3.2.1) and constant regret (Theorem 3.2.3), respectively. Finally, in Section A.5 we discuss the existence of good representations, in Section A.6 we show how our results can be extended for multiple optimal policies and Section A.7 provides auxiliary results.

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

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

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 2 and only interacts with the environment after sampling a state from $d_{\mathcal{P}^\star,h-1}^{\pi_{t-1}}$. 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

$$\begin{aligned}
\rho_{t,h}(s) &= \frac{1}{t} \sum_{i=0}^{t-1} d_{\mathcal{P}^\star,h}^{\pi_i}(s), \\
\gamma_{t,h}(s,a) &= \frac{1}{t} \sum_{i=0}^{t-1} d_{\mathcal{P}^\star,h}^{\pi_i}(s,a), \\
\rho_{t,h}(s,a) &= \rho_{t,h}(s) \bar{\pi}_{t,h}(a|s),
\end{aligned}$$

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

$$\begin{aligned}
\rho'_{t,h}(s') &= \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho_{t,h-1}(s,a) \mathcal{P}_h^\star(s'|s,a), \text{ and} \\
\rho'_{t,h}(s,a) &= \rho'_{t,h}(s) \bar{\pi}_{t,h}(a|s),
\end{aligned}$$

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

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

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 $\lambda_t = c_1 d \log(4t^2 H|\Phi|/\delta)$, 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:

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

where $\zeta_t = \frac{2\log(4t^2|\Phi||\Psi|H/\delta)}{t}$. 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 [UZS22] and the concentration of the bonus term.

Lemma A.0.1. *Fix $\delta \in (0,1)$. Suppose Assumption 2.1.1 (realizability) holds and we run algorithm 2. Then, with probability at least $1-\delta$, the event $\mathcal{E}(\delta)$ occurs.*

Proof. By Lemma A.7.6, with probability at least $1-\delta/2$, event $\mathcal{E}_1(\delta/2)$ occurs, as

$$\frac{1}{2}\mathbb{E}_{(s,a)\sim\rho(s,a)}[f(s,a)^2] \le \mathbb{E}_{(s,a)\sim(\frac{1}{2}\rho(s,a)+\frac{1}{2}\rho'(s,a))}[f(s,a)^2],$$

holds by the non-negativity of f^2. Further, by Lemma 11 in [UZS22], with probability at least $1-\delta/2$, event $\mathcal{E}_2(\delta/2)$ occurs. Taking an union bound concludes. □

A.1. Sub-Linear Pseudo-Regret without UniSOFT Representations

In this section we show that the behavior policies of algorithm 2 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 chapter is purely based on known results, but provided for completeness.

We start by providing two important results, first introduced by [UZS22], 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 A.7.4.

Lemma A.1.1. *(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,*

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

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 2.1.1), (ii) is by importance sampling and (iii) follows from ξ_t being decreasing. □

Lemma A.1.2. *(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,*

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

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 $h=2,...,H$,

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

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 2.1.1), (ii) follows from $(a+b)^2\leq 2a^2+2b^2$ and the event $\mathcal{E}$ and (iii) is 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 [CHYL23].

Lemma A.1.3. *(Bonus relations) Given that the event $\mathcal{E}$ occurs, for all $t \in \mathbb{N}$, $h > 1$ and any π,*

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

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,

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

where (i) is by Lemma A.1.2 and $\|f_{t,h}\|_\infty \leq 1$ and (ii) follows from the event $\mathcal{E}$. Similarly, for all $h > 1$,

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

where (i) is by Lemma A.1.1 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.1.1 and $\|\hat{b}_{t,h}\|_\infty \leq 1$. Further,

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

where (i) follows from the event $\mathcal{E}$ and (ii) follows from Lemma A.7.3. Therefore,

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

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 A.7.3. □

The next result shows that the value function for the estimated environment with the bonus augmented reward function provides an almost optimistic estimate of the true value achieved by any optimal policy.

Lemma A.1.4. *(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 A.7.1, (ii) follows from the event $\mathcal{E}$ and $\|V^{\pi}_{\mathcal{P}^\star,r^\star}\|_\infty \leq 1$ and (iii) follows from Lemma A.1.3 and $\|f_{t,h}\|_\infty \leq 1$. □

We are now ready to show that algorithm 2 achieves sub-linear pseudo-regret; that is, the regret of the behavior polices is sub-linear. The actual regret of algorithm 2 might not be, as we explore uniformly at random in each episode, with positive probability.

Lemma A.1.5. *(Sub-linear pseudo-regret without UniSOFT representations) Given that the event $\mathcal{E}$ occurs, for all $T \in \mathbb{N}$,*

$$
\mathcal{R}(T) \lesssim H^2 d^{3/2} |\mathcal{A}| \frac{\sqrt{T}\log^2(2TH|\Phi||\Psi|/\delta)}{\xi_T} \lesssim \tilde{O}(\frac{\sqrt{T}}{\xi_T})
$$

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.1.4, (ii) follows from Lemma A.7.1 and (iii) follows from $\|V^{\pi}_{\mathcal{P},r^\star+\hat{b}}\|_\infty \leq 2H$. Then, by Lemma A.1.3,

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

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

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

where (i) follows from the Cauchy-Schwarz inequality and Jensen's inequality and (ii) follows from Lemma A.7.4 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. Finally,

$$\begin{aligned}
&\sum_{t=1}^{T} V^{\pi^\star,d_1}_{\mathcal{P}^\star,r^\star,1} - V^{\pi_t,d_1}_{\mathcal{P}^\star,r^\star,1} \\
&\leq 2H\left(4\sqrt{\frac{\zeta_T T^2|\mathcal{A}|}{\xi_T}} + 30\alpha_T\sqrt{\frac{Td|\mathcal{A}|}{\xi_T}} + H(\alpha_T+\beta_T)\sqrt{Td\log(1+\frac{T}{d\lambda_1})}\right) \\
&\lesssim H^2 d^{3/2}|\mathcal{A}|\frac{\sqrt{T}\log^2(2HT|\Phi||\Psi|/\delta)}{\xi_T}
\end{aligned}$$

□

We can now proceed to provide an expected regret bound, by leveraging the result above. 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 $d^{\pi_t}_{\mathcal{P}^\star,h}$ for each time step and hence produce H trajectories per episode. Then, the expected regret of algorithm 2 can be upper bounded as follows:

Lemma 3.1.1 (Regret bound without UniSOFT representations)**.** *Let $\xi_t = t^{-1/4}$. Suppose Assumption 2.1.1 (realizability) holds. Then, for any $T \in \mathbb{N}$, algorithm 2 satisfies:*

$$\mathbb{E}[\tilde{\mathcal{R}}(T)] \lesssim H^3 d^{3/2}|\mathcal{A}|T^{3/4}\log^2(TH|\Phi||\Psi|) = \tilde{O}(H^3 d^{3/2}|\mathcal{A}|T^{3/4})$$

Proof. Let T be given and fixed. Choose $\delta = T^{-1}$. Recall that Algorithm 2 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 2 in episode t and time step h. Then,

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

where (i) follows from the optimality of $\pi^\star$ and $\|V^{\pi}_{\mathcal{P},r^\star}\|_\infty \leq 1$, (ii) follows from $\tilde{\pi}_t$ and π_{t-1} agreeing on the event $e_t = 1$ and Lemma A.0.1 and (iii) follows from an index shift and $\|V^{\pi}_{\mathcal{P},r^\star}\|_\infty \leq 1$. Finally, we can leverage the pseudo-regret result of Lemma A.1.5 to bound term (A),

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

and hence, conclude the proof.

□

A.2. Selecting Non-Redundant UniSOFT Representations

In this section, we show how the $\alpha^\star$-expressiveness assumption 3.2.1 and the constrained optimization objective (Algorithm 2, Line 18) play together, to guarantee that eventually, algorithm 2 will select only good representations. The analysis builds on the sub-linear regret result for the behaviour policies (Lemma A.1.5) provided in the previous chapter.

Selecting $\alpha^\star$-approximate representations

We start by introducing an important result provided by [HZQ+22], 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 set of all optimal (deterministic) policies and $\Pi_h^\star(s)$ as the set of all optimal actions in state $s \in \mathcal{S}$ and time step $h \in [H]$. Then, we construct $\tilde{\pi}_t^\star := \prod_{h\in[H]} \tilde{\pi}_{t,h}^\star$, where for each $h \in [H]$,

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

where *Select* is a function which returns a fixed element of some set and π_t is the behavior policy of algorithm 2 at episode $t \in \mathbb{N}$. We define the mixture occupancy distribution of our constructed optimal policies $\tilde{\pi}_t^\star$ as

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

Note that $\tilde{\gamma}_{t,h}^\star \equiv d_{\mathcal{P}^\star,h}^{\pi^\star}$ whenever there exists an unique optimal policy (Assumption 2.1.2).

Theorem A.2.1. *([HZQ+22], Theorem 4.7) Suppose we run algorithm 2. Then, for all $h \in [H]$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$,*

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

Corollary A.2.1. *Suppose we run algorithm 2 and assumption 2.1.3 (minimal sub-optimality gap) hold. Then, Theorem A.2.1 implies, for all $h \in [H]$, $t \in \mathbb{N}$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$,*

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

We can leverage the above corollary to show that, whenever there exists an unique optimal policy, the MLE oracle converges uniformly on the optimal occupancy distribution, provided that said distribution is well defined for all states. Subsequently, for any given α, there must exist an episode after which algorithm 2 will only select representations that are $\alpha^\star$-approximate.

Lemma A.2.1. *(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 2 and assumptions 2.1.2 (unique optimal policy), 2.1.4 (minimal optimal occupancy) and 2.1.3 (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*

$$\tau_\alpha := \min\{t | t > \frac{1}{\alpha}\left(\frac{g(t)}{\Delta_{\min} d^\star_{\min}} + \sqrt{\frac{t|\mathcal{A}|}{\xi_t}\log(6t^2|\Phi||\Psi|H/\delta)}\right)\}.$$

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

$$\begin{aligned}
\mathbb{E}_{(s,a)\sim d^{\pi^\star}_{\mathcal{P}^\star,h}}[f_{t,h}(s,a)] &= \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d^{\pi^\star}_{\mathcal{P}^\star,h}(s,a) f_{t,h}(s,a) \\
&\overset{(i)}{\leq} \sum_{(s,a):d^{\pi^\star}_{\mathcal{P}^\star,h}(s,a)>0} (\gamma_{t,h}(s,a) + \frac{\mathcal{R}(t)}{t\Delta_{\min}}) f_{t,h}(s,a) \\
&\overset{(ii)}{\leq} \mathbb{E}_{(s,a)\sim\gamma_{t,h}}[f_{t,h}(s,a)] + \frac{g(t)}{t\Delta_{\min}} \sum_{(s,a):d^{\pi^\star}_{\mathcal{P}^\star,h}(s,a)>0} \frac{d^{\pi^\star}_{\mathcal{P}^\star,h}(s,a)}{d^\star_{\min}} \\
&\overset{(iii)}{\leq} \sqrt{\frac{|\mathcal{A}|}{\xi_t}\mathbb{E}_{(s,a)\sim\rho_{t,h}}[f_{t,h}(s,a)^2]} + \frac{g(t)}{t\Delta_{\min} d^\star_{\min}} \\
&\overset{(iv)}{\leq} \sqrt{\frac{|\mathcal{A}|}{\xi_t}\zeta_t} + \frac{g(t)}{t\Delta_{\min} d^\star_{\min}},
\end{aligned}$$

where (i) is by Corollary A.2.1, (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 UniSOFT representations

Although we now can be sure to select $\alpha^\star$ approximate representations, we still need to ensure that the UniSOFT loss in equation 3.2 will lead to Algorithm 2 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 3.2. We define the following good events:

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

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

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 $t \in \mathbb{N}$. Further, define $\mathcal{F}(\delta) := \mathcal{F}_1(\delta/2) \cap \mathcal{F}_2(\delta/2)$.

Lemma A.2.2. *(Eigenvalue bounds) Assume 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 2 and that assumption 2.1.3 (minimal sub-optimality gap) holds. Then, with probability at least $1-\delta$, the event $\mathcal{F}(\delta)$ occurs.*

Proof. Recall that algorithm 2 produces for each time step $h \in [H]$, one trajectory τ_h, in any episode t. Further, for each trajectory τ_h, we only employ the transition at time step h for constructing 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) behaviour 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

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

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 2. Finally, we denote $\tau^{(t,h)}_{1:h}$ as the trajectory $\tau^{(t,h)}$ cut

off at time step $h \in [H]$. Then,

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

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

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

where (i) follows from $\|\phi_h\|_2 \leq 1$ and Proposition A.7.1 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,

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

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 $d^{\tilde{\pi}^\star_t}_{\mathcal{P}^\star,h}$ and $d^{\pi_t}_{\mathcal{P}^\star,h}$ agreeing on $\Gamma^\star_{h,t}$ and for (iii) recall that $\tilde{\gamma}^\star_{t,h}(s,a) = \frac{1}{t} \sum_{i=o}^{t-1} d^{\tilde{\pi}^\star_t}_{\mathcal{P}^\star,h}(s,a)$. Similarly, with probability of at least $1 - \delta/6$, for all $t \in \mathbb{N}$ and all $h \in [H], \phi \in \Phi$,

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

where (i) follows, similarly to before, from Proposition A.7.1 in combination with an

union bound and (ii) is by $\|\phi_h\|_2 \leq 1$. Further,

$$\begin{aligned}
(B1) &= I\sum_{i=1}^{t}\mathbb{E}_{\tau\sim\rho_h^{\pi_{i-1},\xi_i}}[\mathbb{1}[\tau\notin\Gamma^\star_{h,i}]] \\
&= I\sum_{i=1}^{t}\mathbb{E}_{\tau\sim\rho_h^{\pi_{i-1},\xi_i}}[\mathbb{1}[e=1]\mathbb{1}[\tau\notin\Gamma^\star_{h,i}]] + \mathbb{E}_{\tau\sim\rho_h^{\pi_{i-1},\xi_i}}[\mathbb{1}[e=0]\mathbb{1}[\tau\notin\Gamma^\star_{h,i}]] \\
&\overset{(i)}{\preccurlyeq} I\sum_{i=1}^{t}\mathbb{E}_{\tau\sim\rho_h^{\pi_{i-1}}}[\mathbb{1}[\tau\notin\Gamma^\star_{h,i}]] + I\sum_{i=1}^{t}\mathbb{E}_{e\sim\mathrm{Ber}(1-\xi_i)}[\mathbb{1}[e=0]] \\
&\overset{(ii)}{\preccurlyeq} I\sum_{i=1}^{t}\sum_{h=1}^{H}\mathbb{E}_{(s,a)\sim d^{\pi_{i-1}}_{\mathcal{P}^\star,h}}[\mathbb{1}[a\notin\Pi^\star_h(s)]] + I\sum_{i=1}^{t}\xi_i \\
&\preccurlyeq I\sum_{i=1}^{t}\sum_{h=1}^{H}\mathbb{E}_{(s,a)\sim d^{\pi_{i-1}}_{\mathcal{P}^\star,h}}[\mathbb{1}[\Delta_h(s,a)\geq\Delta_{\min}]] + I\sum_{i=1}^{t}\xi_i \\
&\preccurlyeq I\frac{1}{\Delta_{\min}}\sum_{i=1}^{t}\sum_{h=1}^{H}\mathbb{E}_{(s,a)\sim d^{\pi_{i-1}}_{\mathcal{P}^\star,h}}[\Delta_h(s,a)] + I\sum_{i=1}^{t}\xi_t \\
&\overset{(iii)}{=} \frac{\mathcal{R}(t)}{\Delta_{\min}}I + I\sum_{i=1}^{t}\xi_i,
\end{aligned}$$

where (i) follows from $\rho_h^{\pi,\xi}$ and ρ_h^{π} agreeing on the event $e=1$, (ii) follows from the definition of $\tilde{\pi}^\star_{t,h}$ and (iii) follows from Lemma A.7.2. Finally, with probability at least $1-\delta/6$, for all $t\in\mathbb{N}$ and $h\in[H]$,

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

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, which will be crucial to show uniformly decreasing confidence intervals:

Lemma A.2.3. *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 2.1.2 (unique optimal policy) and 2.1.3 (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 $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

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

Proof. Let τ_{inv} be large enough such that,

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

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

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

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

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

Lemma A.2.4. *(Selecting non-redundant UniSOFT representation) Fix any $\alpha > 0$. Assume 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 2 and assumptions 2.1.2 (unique optimal policy), 3.2.1 (expressiveness) and 2.1.3 (minimal sub-optimality gap) hold. Additionally, if $\alpha < 1$,*

suppose that assumption 2.1.4 (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 episode $t \geq \tau_{\text{unisoft}}$ and time steps $h \in [H]$, the learned feature maps $\hat{\phi}_{t,h}$ are UniSOFT and non-redundant, where

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

Proof. Note that, by Lemma A.2.1, 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 containing only non-redundant UniSOFT feature mappings. By Lemma A.2.2, given that the event $\mathcal{F}$ occurs, for all $t \in \mathbb{N}$, $h \in [H]$, $\phi \in \Phi \setminus \Phi^{\text{unisoft}}$ and $\phi^{\text{unisoft}} \in \Phi^{\text{unisoft}}$,

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

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

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

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

as the set containing all $\tilde{\alpha}^\star$-approximate representations such that the feature map is non-redundant and UniSOFT, which is non-empty by Assumption 3.2.1. A non-redundant UniSOFT representation ϕ^{unisoft} is selected at episode $t \geq \tau_\alpha$ if for all $\tilde{\alpha} \leq \alpha$,

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

or equivalently,

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

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

A.3. Improved Pseudo-Regret with UniSOFT Representations

In this section, we show how we can leverage good representations to improve the pseudo-regret result A.1.5 provided in the section A.1. Subsequently, we can provide an improved

expected regret result.

On a high level, we show that the bonus terms can be used to provide an almost optimistic estimate for the expected sub-optimality gaps incurred by the behaviour polices of algorithm 2. We can then exploit the UniSOFT property, of the good representations we are guaranteed to select as show in the previous section, to show uniformly decreasing confidence intervals. Let us start by providing two results that are adapted from [CHYL23] which show that the bonus term can be employed to provide a trajectory-wise uncertainty measure for the model estimation error over the occupancy distribution of the behavior policies.

Lemma A.3.1. *(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 $\sum_{h=1}^H r_h \leq 1$,*

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

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 $|V^{\pi}_{\mathcal{P}^\star,r,H+1}(s) - V^{\pi}_{\hat{\mathcal{P}}_t,r,H+1}(s)| = 0 = V^{\pi}_{\mathcal{P},f_t,H+1}$ for $\mathcal{P} \in \{\hat{\mathcal{P}}_t, \mathcal{P}^\star\}$. Suppose the induction hypothesis, $|V^{\pi}_{\mathcal{P}^\star,r,h+1}(s) - V^{\pi}_{\hat{\mathcal{P}}_t,r,h+1}(s)| \leq V^{\pi}_{\mathcal{P},f_t,h+1}(s)$ for $\mathcal{P} \in \{\hat{\mathcal{P}}_t, \mathcal{P}^\star\}$ and any $s \in \mathcal{S}$. Then, for any $h \in [H]$ and $s \in \mathcal{S}$,

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

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

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

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

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

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

Lemma A.3.2. *(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.1.3 and $\|f_{t,h}\|_\infty \leq 1$ and (ii) follows from the event $\mathcal{E}$. Additionally, by Lemma A.1.3, 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 A.3.1 and (ii) follows from the second claim. □

Next we introduce an optimism result similar to that of Lemma A.1.4, which holds locally on the state-occupancy distribution of the behavior policies.

Lemma A.3.3. *(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 A.3.1. 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 A.3.2. Now, the claim follows by the definition of π^b_t. □

We continue by providing a local simulation lemma.

Lemma A.3.4. *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 from the behavior policies of algorithm 2.

Lemma A.3.5. *(Sub-optimality gap to bonus) Given that the event $\mathcal{E}$ occurs, we have for all $t \in \mathbb{N}$ and $h \in [H]$,*

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

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 A.3.4 and (iv) follows from the local optimism Lemma A.3.3. 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 A.3.1, (ii) follows from Lemma A.3.2 and (iii) by the optimality of π^b_t. Similarly,

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

where (i) follows from Lemma A.3.2 and (ii) follows from the optimality of π^b_t. Finally, we get:

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

□

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

Lemma A.3.6. *(Sub-linear pseudo-regret with UniSOFT representations) Let $\xi_t = t^{-1/3}$ and $\alpha > 0$. Suppose assumptions 2.1.1 (realizability), 2.1.2 (unique optimal policy), 2.1.3 (minimal sub-optimality gap) and 3.2.1 (α-expressive function space) hold. Additionally, if $\alpha < 1$, suppose that assumption 2.1.4 (minimal optimal occupancy) holds. Then, given that the events $\mathcal{E}(\delta)$ and $\mathcal{F}(\delta)$ occur there exists a constant τ, such that for all $T \geq \tau$, the behaviour policies $\{\pi_t\}_{t\geq 1}$ learned by algorithm 2, enjoy sub-linear regret:*

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

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 A.3.5, we can bound the expected sub-optimality gaps for all $h \in [H]$,

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

Further, by Lemma A.1.5, under 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 2.1.4 (minimal optimal occupancy) to guarantee their selection in Lemma A.2.1. By Lemma A.2.4 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 A.2.3 and the event $\mathcal{F}$,

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

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

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

where (i) follows from the pseudo-regret bound without UniSOFT representations given in Lemma A.1.5. □

Remark A.3.1. *The choice for $\xi_t = t^{-1/3}$ is somewhat optimal when considering the expected regret, as, in expectation, a faster exploration rate worsens the pseudo regret result, but reduces the number of episodes in which we have to explore uniformly at random.*

Theorem 3.2.1 (Instance-dependent regret with UniSOFT representations)**.** *Let $\xi_t = t^{-1/3}$ and $\alpha \in (0, 1]$. Suppose assumptions 2.1.1 (realizability), 2.1.3 (minimal sub-optimality gap), 3.2.1 (α-expressive function space) and 2.1.2 (unique optimal policy) hold. Additionally, if $\alpha < 1$, suppose assumption 2.1.4 (minimal optimal occupancy) holds. Then, for any $T \in \mathbb{N}$ large enough, algorithm 2 satisfies:*

$$\begin{aligned}\mathbb{E}[\tilde{\mathcal{R}}(T)] &\lesssim H\tau^{5/6} + \frac{1}{\lambda^\star_{\max}} H^4 d^{1/2} |\mathcal{A}|^{1/2} T^{2/3} \log(TH|\Phi||\Psi|) \\ &\lesssim \tilde{O}(\frac{1}{\lambda^\star_{\max}} H^4 \sqrt{d|\mathcal{A}|} T^{2/3}),\end{aligned}$$

where

$$\begin{aligned}\tau &\lesssim \{\kappa_3^m \cdot \log^{2m}(\kappa_3 \cdot \kappa_2) \vee \kappa_1^m \cdot \log^{2m}(\kappa_1 \cdot \kappa_2)\} \\ &\lesssim \frac{H^{12} d^9 |\mathcal{A}|^6}{\Delta_{\min}\{\alpha d^\star_{\min} \wedge \lambda^\star_{\max}\}} \cdot \log^{12}(TH^3 d^{3/2} |\mathcal{A}||\Phi||\Psi|),\end{aligned}$$

with $\kappa_1 = \frac{H^2 d^{3/2}|\mathcal{A}|}{\alpha \Delta_{\min} d^\star_{\min}}$, $\kappa_2 = TH|\Phi||\Psi|$, $\kappa_3 = \frac{H^2 d^{3/2}|\mathcal{A}|}{\lambda^\star_{\max}\Delta_{\min}}$ and $\lambda^\star_{\max} = \min_{\tilde{\alpha}\le\alpha} \max_{\phi \in \Phi^{\text{unisoft}}_{\tilde{\alpha}}} \lambda^\star(\phi)$. In particular, T must be large enough such that $T \ge \tilde{O}(\tau)$ holds.

Proof. Let $\tau := \{\tau_{\text{unisoft}} \vee \tau_{\text{inv}}\}$ and $T \ge \tau$ be given and fixed. Choose $\delta = T^{-1}$. Recall that Algorithm 2 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 2 in episode t and time step h. Then,

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

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

where (i) follows from $\|V^{\pi}_{\mathcal{P},r^\star}\|_\infty \leq 1$, (ii) follows from $\tilde{\pi}_t$ and π_{t-1} agreeing on the event $e_t = 1$, Lemma A.0.1 and Lemma A.2.3 and (iii) follows from an index shift and $\|V^{\pi}_{\mathcal{P},r^\star}\|_\infty \leq 1$. Now, we can leverage the pseudo-regret result of Lemma A.3.6 to bound term (A),

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

Substituting τ with the sufficient condition from Lemma A.4.3 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. □

A.4. Constant Pseudo-Regret with UniSOFT Representations

In this section we provide a further improved 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 renders a deterministic policy to be optimal.

Lemma A.4.1. *Let π be any deterministic policy. Whenever,*

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

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,

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

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

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

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 a 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 $h \in [H]$,

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

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 A.4.2. *(Constant pseudo-regret with UniSOFT representations) Let $\alpha \in (0,1]$, $\gamma \in (2,\infty)$ and $\xi_t = t^{-1/\gamma}$. Suppose assumptions 2.1.1 (realizability), 2.1.2 (unique optimal policy), 2.1.3 (minimal sub-optimality gap), 2.1.4 (minimal optimal occupancy) and 3.2.1 (α-expressive function space) hold. Then, given that the events $\{\mathcal{E}_t\}_{t\geq 1}$ and $\{\mathcal{F}^{(t)}(\delta_t)\}_{t\geq 1}$ occur, there exists a constant $\tau^\star$, after which, the behaviour policies $\{\pi_t\}_{t\geq 1}$ learned by algorithm 2, 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 $\mathcal{E}^{(t)}$ occurs by assumption, by Lemma A.3.5, we can bound the expected sub-optimality gaps for all $h \in [H]$,

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

Further, by Lemma A.1.5, under the event $\mathcal{E}_t$, $\mathcal{R}(t) \leq g(t) = \tilde{O}(\sqrt{t}\xi_t^{-1}) = O(t^{\frac{2+\gamma}{2\gamma}})$ with $\hat{\alpha}_t = \tilde{O}(\xi_t^{-1/2}) = \tilde{O}(t^{\frac{1}{2\gamma}})$. By Lemma A.2.4 and the events $\mathcal{F}^{(t)}$ and $\mathcal{E}^{(t)}$, for all $h \in [H]$, the learned feature maps $\hat{\phi}_{t,h}$ are non-redundant and UniSOFT. Then, by Lemma A.2.3, $\gamma > 2$ and the event $\mathcal{F}^{(t)}$,

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

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

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

for all $t \geq \tau^\star$. Then by Lemma A.4.1, 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 3.2.2 (Instance-dependent sub-linear expected regret with UniSOFT representations and constant pseudo-regret)**.** *Let $\alpha > 0$, $\gamma \in (2, 4]$ and $\xi_t = t^{-1/\gamma}$. Suppose assumptions 2.1.1 (realizability), 2.1.2 (unique optimal policy), 2.1.3 (minimal sub-optimality gap), 2.1.4 (minimal optimal occupancy) and 3.2.1 (α-expressive function space) hold. Then for any $T \in \mathbb{N}$ large enough, algorithm 2 satisfies:*

$$
\mathbb{E}_{\delta,\xi}[\tilde{\mathcal{R}}(T)] \lesssim H(\tau^\star)^{1/2+1/\gamma} + HT^{\frac{\gamma-1}{\gamma}} \lesssim \tilde{O}(HT^{\frac{\gamma-1}{\gamma}}),
$$

where

$$
\begin{aligned}
\tau^\star &\lesssim \{\kappa_3^m \cdot \log^{2m}(\kappa_3 \cdot \kappa_2) \vee \kappa_1^m \cdot \log^{2m}(\kappa_1 \cdot \kappa_2) \vee \kappa_4^{m'} \cdot \log^{m'}(\kappa_4 \cdot \kappa_2)\} \\
&\lesssim \left(\frac{H^3 d^2 |\mathcal{A}|}{\alpha \lambda^\star_{\max}(\Delta_{\min} d^\star_{\min})^2}\right)^{\frac{2\gamma}{\gamma-2}} \cdot \left(\log(\frac{TH^4 d^2 |\mathcal{A}|\Phi||\Psi|}{\alpha\lambda^\star_{\max}(\Delta_{\min} d^\star_{\min})^2})\right)^{\frac{4\gamma}{\gamma-2}},
\end{aligned}
$$

with $\kappa_1 = \frac{H^2 d^{3/2}|\mathcal{A}|}{\alpha \Delta_{\min} d^\star_{\min}}$, $\kappa_2 = TH|\Phi||\Psi|$, $\kappa_3 = \frac{H^2 d^{3/2}|\mathcal{A}|}{\lambda^\star_{\max}\Delta_{\min}}$, $\kappa_4 = \frac{H^3 d^2 |\mathcal{A}|}{(\Delta_{\min} d^\star_{\min})^2 \lambda^\star_{\max}}$, $m = \frac{2\gamma}{\gamma-2}$ and $m' = \frac{\gamma}{\gamma-1}$. In particular, T must be large enough such that $T \geq \tilde{O}(\tau^\star)$ holds.

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 3.2.1, (ii) follows from the constant pseudo-regret result of Lemma A.4.2. We conclude the proof by substituting $\tau^\star$ with the sufficient condition provided in Lemma A.4.3 and using $T \gtrsim a\log^n(ab)$ as a sufficient condition for $T \geq a\log^n(bT)$. □

Lemma A.4.3. *(Critical episodes) Let $\alpha \in (0,1]$, $\gamma \in (2,4]$ and $\xi_t = t^{-1/\gamma}$. Suppose assumptions 2.1.1 (realizability), 2.1.2 (unique optimal policy), 2.1.3 (minimal sub-optimality gap), 2.1.4 (minimal optimal occupancy) and 3.2.1 (α-expressive function space) hold. Supppose we run algorithm 2. Then, given that the events $\mathcal{E}(\delta)$ and $\mathcal{F}(\delta)$ occur:*

(1) *all non-UniSOFT representations are eliminated after at most*

$$\tau \lesssim \{\kappa_3^m \cdot \log^{2m}(\kappa_3 \cdot \kappa_2) \vee \kappa_1^m \cdot \log^{2m}(\kappa_1 \cdot \kappa_2)\}$$

(2) *the behavior policies $\{\pi_t\}_{t\geq 1}$ are optimal after at most*

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

episodes, where $\kappa_1 = \frac{H^2 d^{3/2}|\mathcal{A}|}{\alpha\Delta_{\min} d^\star_{\min}}$, $\kappa_2 = H|\Phi||\Psi|/\delta$, $\kappa_3 = \frac{H^2 d^{3/2}|\mathcal{A}|}{\lambda^\star_{\max}\Delta_{\min}}$, $\kappa_4 = \frac{H^3 d^2|\mathcal{A}|}{(\Delta_{\min} d^\star_{\min})^2\lambda^\star_{\max}}$, $m = \frac{2\gamma}{\gamma-2}$ and $m' = \frac{\gamma}{\gamma-1}$.

Proof. By Lemma A.1.5, for all $t \in \mathbb{N}$,

$$\begin{aligned}
\mathcal{R}(t) &\leq c_3 H^2 d^{3/2}|\mathcal{A}|\frac{\sqrt{t}\log^2(4tH|\Phi||\Psi|/\delta)}{\xi_t}, \\
\hat{\alpha}_t &= \sqrt{4t\zeta_t\frac{|\mathcal{A}|}{\xi_t} + \lambda_t d},
\end{aligned}$$

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 [PTP+21] 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 $m \geq 1$:

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

A. Appendix

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 A.2.1.

$$
\begin{aligned}
t &> \frac{1}{\alpha}\Big(\frac{\mathcal{R}(t)}{\Delta_{\min} d^\star_{\min}} + \sqrt{\frac{t|\mathcal{A}|}{\xi_t}\log(6t^2 H|\Phi||\Psi|/\delta)}\Big)\\
t &> \frac{1}{\alpha}\Big(\frac{c_3 H^2 d^{3/2}|\mathcal{A}|t^{(1/2+1/\gamma)}\log^2(4t^3 H|\Phi||\Psi|)}{\Delta_{\min} d^\star_{\min}} + \sqrt{2t^{1+1/\gamma}|\mathcal{A}|\log(4t|\Phi||\Psi|H/\delta)}\Big)\\
t &> t^{\frac{\gamma+2}{2\gamma}}\cdot c_3 6\underbrace{\frac{H^2 d^{3/2}|\mathcal{A}|}{\alpha\Delta_{\min} d^\star_{\min}}}_{\kappa_1}\cdot\log^2(t\cdot 4\underbrace{H|\Phi||\Psi|/\delta)}_{\kappa_2}\\
t &\overset{(i)}{>} (2m)^2 6^m c_3^m \kappa_1^m \log^{2m}(\kappa_1\cdot 4\kappa_2) := \bar{\kappa}_1,
\end{aligned}
$$

where (i) follows from the condition A.1 with $m = \frac{2\gamma}{\gamma-2}$.

Part 2. $\tau^\star$ must satisfy the UniSOFT-selection criteria in Lemma A.2.4.

$$
\begin{aligned}
t &> \frac{2}{\lambda^\star_{\max}}\Big(\Delta_{\min}^{-1}\mathcal{R}(t) + \sum_{i=1}^{t}\xi_i + 18\sqrt{t\log(6tdH|\Phi|/\delta)}\Big)\\
t &> \frac{2}{\lambda^\star_{\max}}\Big(\frac{c_3 H^2 d^{3/2}|\mathcal{A}|t^{1/2+1/\gamma}\log^2(4t^2 H|\Phi||\Psi|)}{\Delta_{\min}} + \frac{\gamma}{\gamma-1}t^{1-1/\gamma} + 18\sqrt{t\log(6t|\Phi|H/\delta)}\Big)\\
t &\overset{(i)}{>} t^{\frac{2+\gamma}{2\gamma}}\cdot c_3 2^7\underbrace{\frac{H^2 d^{3/2}|\mathcal{A}|}{\lambda^\star_{\max}\Delta_{\min}}}_{\kappa_3}\cdot\log^2(t\cdot 6\underbrace{H|\Phi||\Psi|/\delta)}_{\kappa_2}\\
t &\overset{(ii)}{>} (2m)^2 2^{14m} c_3^m \kappa_3^m \log^{2m}(\kappa_3\cdot 8\kappa_2) := \bar{\kappa}_2,
\end{aligned}
$$

where (i) follows from $\gamma \le 4$ and (ii) follows from the condition A.1 with $m = \frac{2\gamma}{\gamma-2}$.

Part 3. $\tau^\star$ must satisfy the invertibility condition from Lemma A.2.3.

$$
t > \frac{\lambda_t + \mathcal{R}(t)\Delta_{\min}^{-1} + \sum_{i=1}^{t}\xi_i + 8\sqrt{t\log(6tdH|\Phi|/\delta)}}{\lambda^\star_{\max}},
$$

Note that the condition is fulfilled if $t \ge \bar{\kappa}_2$. By taking, $\tau := \max\{\bar{\kappa}_1, \bar{\kappa}_2\}$, we gain statement (1).

Part 4. First note we can upper bound,

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

For now we assume that $t \geq \bar{\kappa}_1$. Then,

$$\begin{aligned}\Delta_{\min}d^\star_{\min} &> 20H^2\Big(\frac{\hat{\alpha}_t H}{(\lambda^\star_{\max}t+\lambda_t-\sum_{i=1}^t\xi_i-\mathcal{R}(t)\Delta^{-1}_{\min}-18\sqrt{t\log(6tdH|\Phi|/\delta)})^{1/2}}+\sqrt{\frac{|A|}{\xi_t}\zeta_t}\Big)\\ \Delta_{\min}d^\star_{\min} &> 20H^2\Big(\frac{\hat{\alpha}_t H}{(\lambda^\star_{\max}t+\lambda_t-\frac{\gamma}{\gamma-1}t^{1-\frac{1}{\gamma}}-\mathcal{R}(t)\Delta^{-1}_{\min}-18\sqrt{t\log(6tdH|\Phi|/\delta)})^{1/2}}+\sqrt{\frac{|A|}{\xi_t}\zeta_t}\Big)\\ \Delta_{\min}d^\star_{\min} &\overset{(i)}{>} 20H^2\Big(\frac{\hat{\alpha}_t H}{(\frac{1}{2}\lambda^\star_{\max}t)^{1/2}}+\sqrt{\frac{2|A|t^{\frac{1}{\gamma}}\log(4t|\Phi||\Psi|H/\delta)}{t}}\Big)\\ \Delta_{\min}d^\star_{\min} &> t^{-\frac{1}{2}(1-\frac{1}{\gamma})}\cdot 2^{12}\sqrt{c_1}\frac{H^3d|\mathcal{A}|^{1/2}}{(\lambda^\star_{\max})^{1/2}}\cdot\sqrt{\log(t\cdot 4|\Phi||\Psi|H/\delta)},\end{aligned}$$

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

$$\begin{aligned}t^{\frac{1}{2}(1-\frac{1}{\gamma})} &> 2^{12}\sqrt{c_1}\frac{H^3d|\mathcal{A}|^{1/2}}{\Delta_{\min}d^\star_{\min}(\lambda^\star_{\max})^{1/2}}\cdot\log^{1/2}(t\cdot 4|\Phi||\Psi|H/\delta)\\ t^{(1-\frac{1}{\gamma})} &> 2^{24}c_1\underbrace{\frac{H^3d^2|\mathcal{A}|}{(\Delta_{\min}d^\star_{\min})^2\lambda^\star_{\max}}}_{\kappa_4}\cdot\log(t\cdot 4\underbrace{|\Phi||\Psi|H/\delta)}_{\kappa_2}\\ t &\overset{(i)}{>} m2^{24m}c_1^m\kappa_4^m\log^m(\kappa_4\cdot 4\kappa_2) := \bar{\kappa}_3\end{aligned}$$

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

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

we conclude. □

Theorem 3.2.3 (Optimal policy identification)**.** *Let $\alpha > 0$, $\gamma \in (2, 4]$ and $\xi_t = t^{-1/\gamma}$. suppose the quantities $\Delta_{\min}$ and $d^\star_{\min}$ are known. Let T be large enough such that $T \geq \tau^\star$ holds. Then, under the same assumptions as in Theorem 3.2.2, with probability at least*

$1-2T^{-1}$, *algorithm 3 will return the optimal policy after at most* T *episodes. In particular, if* $\gamma = 4$,

$$T \geq \tilde{O}\left(\frac{H^{12}d^8|\mathcal{A}|^4}{(\alpha\lambda^\star_{\max})^4(\Delta_{\min}d^\star_{\min})^8}\right).$$

Proof. We know, by the proof of Lemma A.3.6, that given that the events $\mathcal{E}$ and $\mathcal{F}$ hold, there exists an episode $\tau^\star$ such that, for all $t \geq \tau^\star$ and $h \in [H]$,

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

$$< \Delta_{\min}d^\star_{\min}. \tag{A.3}$$

In particular, we know from Lemma A.4.1, 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 A.0.1 and Lemma A.2.2. Hence, with probability at least $1-2\delta$, algorithm 3 returns an optimal policy after at most $\tau^\star$ episodes. The requirement on T follows from the complexity bound of Theorem 3.2.2 and the choice $\delta = T^{-1}$. □

A.5. Existence of UniSOFT Representations

Here we restate and prove the existence results from section 3.3.

Theorem A.5.1. *([PO99]) Every matrix* $A \in \mathbb{C}^{n\times m}$ *with* $\operatorname{rank}(A) = r > 0$ *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 3.3.1 (Existence of exact UniSOFT representations with minimal dimension)**.** *Assume that* $\operatorname{rank}(\mathcal{P}^\star_h) = \tilde{d}$ *for each* $h \in [H]$. *Let* $\mathcal{X}_h := \{(s,a) \in \mathcal{S}\times\mathcal{A} | d^{\pi^\star}_{\mathcal{P}^\star,h}(s,a) > 0\}$ *be the set of state-action pairs reachable by the optimal policy at time step* $h \in [H]$. *Then, the following statements are equivalent:*

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

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

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

Proof. Let us start by constructing a full rank factorization of $\mathcal{P}^\star_h$. Note that $\mathcal{P}^\star_h$ has rank d by assumption and hence we can select 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 $\Phi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times d}$, 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. We denote $\Psi \in \mathbb{R}^{d\times|\mathcal{S}|}$ as the matrix uniquely determined by the coefficients in the linear combinations such that $\mathcal{P}^\star_h = \Phi\Psi$. Then,

(1) $\Rightarrow$ (2). By construction, the rows of Φ corresponding to elements in $\mathcal{X}_h$ form a basis of $\mathbb{R}^d$ and hence $\Phi\Psi$ is a non-redundant and UniSOFT representation of $\mathcal{P}^\star_h$.

(2) $\Rightarrow$ (3). Let $\mathcal{P}^\star_h = \Phi^\star\Psi^\star$ such that the representation is non-redundant and UniSOFT. By Theorem A.5.1 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 $\mathcal{P}^\star_h = \bar{\Phi}\bar{\Psi}$. 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.

□

Lemma A.5.1. *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 $i = 1, ..., d$ 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. 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, then there exists some tuple $(\alpha_1, ..., \alpha_d) \in \mathbb{R}^d$ such that

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

In particular, w.l.o.g. we can assume that $(\alpha_1, ..., \alpha_d) \in S$. But then,

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

leads to a contradiction, where (i) follow from the reverse triangle inequality and (ii) by the choice of ϵ. □

Remark A.5.1. *The statement is generally not true for a set of linear dependent vectors.*

Lemma 3.3.2. *Assume that $\text{rank}(\mathcal{P}^\star_h) = \tilde{d}$ for each $h \in [H]$ and that assumption 2.1.4 (minimal optimal occupancy) holds. Further assume that $\mathcal{P}^\star$ admits an UniSOFT representation. Then, there exists an $\epsilon > 0$ such that all $\alpha^\star$-approximate representations $\langle \phi, \mu \rangle_{\mathbb{R}^{\tilde{d}}} \equiv \hat{\mathcal{P}}$ with $\alpha \leq \epsilon$ are UniSOFT.*

Proof. Let $\alpha \leq \epsilon$ be arbitrary and $\mathcal{X}_h := \{(s, a) \in \mathcal{S} \times \mathcal{A} | d^{\pi^\star}_{\mathcal{P}^\star, h}(s, a) > 0\}$ be the set of state-action pairs reachable by the optimal policy. Then, since $\mathcal{P}^\star$ is assumed to admit an

UniSOFT representation, by Lemma 3.3.1, there exist at least d state-action pairs in $\mathcal{X}_h$ such that their transition vectors in model $\mathcal{P}_h^\star$ span $\mathbb{R}^d$. Denote $\tilde{\mathcal{X}}_h$ as the set containing those d state-action pairs. Further, for any $\alpha^\star$-approximate representation with induced transition operator $\mathcal{P}$, any $h \in [H]$ and $(s', a') \in \mathcal{X}_h$,

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

Then, by Lemma A.5.1, if ϵ is small enough, the vectors in $\{\mathcal{P}_h(\cdot|s,a)|(s,a) \in \tilde{\mathcal{X}}_h\}$ are linear independent and hence, by Lemma 3.3.1, all representations inducing $\mathcal{P}$ are UniSOFT. In particular,

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

where $\{v_i\}_{i=1}^d = \{\mathcal{P}_h^\star(\cdot|s,a)|(s,a) \in \tilde{\mathcal{X}}_h\}$. □

A.6. Multiple Optimal Policies

We provide two results, that ensure the selection of good representations under multiple optimal policies.

Lemma A.6.1. *(Selecting $(\tilde{\sigma}_t^\star, \alpha)$-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 2 and assumptions 2.1.4 (minimal optimal occupancy) and 2.1.3 (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 $(\tilde{\sigma}_t^\star, \alpha)$-approximate, where*

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

Proof. Directly follows from Corollary A.2.1 and the proof of Lemma A.2.1. □

Lemma A.6.2. *(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 2 and assumptions 3.3.1 (expressiveness) and 2.1.3 (minimal sub-optimality gap) hold. Additionally, if $\alpha < 1$, suppose assumption 2.1.4 (minimal optimal occupancy) holds. Then, given that the events $\mathcal{E}(\delta)$ and $\mathcal{F}(\delta)$ occur, there exists an episode $\tau_{\mathrm{unisoft}} \geq \tau_\alpha$ such that for all subsequent episodes $t \geq \tau_{\mathrm{unisoft}}$ and time steps*

$h \in [H]$ the learned feature maps $\hat{\phi}_{t,h}$ are UniSOFT w.r.t. any optimal policy $\pi^\star \in \tilde{\sigma}_t^\star$, where

$$\tau_{\text{unisoft}} := \min\{t | t > \left(\frac{2}{\lambda_\alpha^\star}(\Delta_{\min}^{-1}\mathcal{R}(t) + 18\sqrt{t\log(6dtH|\Phi|/\delta)}) \wedge \tau_\alpha \right)\}.$$

Proof. Let $\Phi_{\tilde{\sigma}_t^\star}^{\text{unisoft}} \subseteq \Phi$ denote the set containing only non-redundant feature mappings that are UniSOFT w.r.t. at least one $\tilde{\pi}^\star \in \tilde{\sigma}_t^\star$. By Lemma A.2.2, with probability at least $1-\delta$, for all $t \in \mathbb{N}$, $h \in [H]$, $\phi \in \Phi \setminus \Phi_{\tilde{\sigma}_t^\star}^{\text{unisoft}}$ and $\phi^{\text{unisoft}} \in \Phi_{\tilde{\sigma}_t^\star}^{\text{unisoft}}$,

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

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

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

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

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

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

as the set containing all $(\tilde{\sigma}_t^\star, \tilde{\alpha})$-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 3.3.1. A desired feature map is selected at episode $t \geq \tau_\alpha$ if for any $\tilde{\alpha} \leq \alpha$,

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

or equivalently,

$$t\lambda_\alpha^\star(\phi^{\text{unisoft}}) > 2\Delta_{\min}^{-1}\mathcal{R}(t) + 2\sum_{i=1}^{t}\xi_i + 32\sqrt{t\log(6dtH|\Phi|/\delta)},$$

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

A.7. Auxiliary Results

Lemma A.7.1. *([ZSU+ 2a], Simulation Lemma) Given two transition models $\mathcal{P}$ and $\mathcal{P}'$, we have:*

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

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

Lemma A.7.2. *([HZG21]) For any $h \in [H]$, $s \in \mathcal{S}$, and $\pi \in \Pi$:*

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

Hence the regret after T episodes can be expressed as:

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

Proof.

$$\begin{aligned}&V^{\pi^\star}_{\mathcal{P}^\star,r^\star,h}(s) - V^{\pi}_{\mathcal{P}^\star,r^\star,h}(s) \\ &= \Delta_h(s,\pi_h(s)) + Q^{\pi^\star}_{\mathcal{P}^\star,r^\star,h}(s,\pi_h(s)) - V^{\pi}_{\mathcal{P}^\star,r^\star,h}(s) \\ &= \Delta_h(s,\pi_h(s)) + r^\star_h(s,\pi_h(s)) + \mathcal{P}^\star_h V^{\pi^\star}_{\mathcal{P}^\star,r^\star,h+1}(s,\pi_h(s)) - r^\star_h(s,\pi_h(s)) - \mathcal{P}^\star_h V^{\pi}_{\mathcal{P}^\star,r^\star,h+1}(s,\pi_h(s)) \\ &= \Delta_h(s,\pi_h(s)) + \mathcal{P}^\star_h(V^{\pi^\star}_{h+1} - V^{\pi}_{h+1})(s,\pi_h(s))\end{aligned}$$

Unravelling the recursion gives the result. □

Lemma A.7.3. *([JYWJ20], Lemma D.1.) 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,*

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

Lemma A.7.4. *(Elliptical potential lemma, [AYPS11]) Consider a sequence of $d \times d$ positive semidefinite matrices $X_1, ..., X_T$ with $tr(X_t) \le 1$ for all $t \in [T]$. Define $M_0 = \lambda_0 I$ and $M_t = M_{t-1} + X_t$. Then*

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

Proposition A.7.1. *(Matrix Azuma, [Tro12]) 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 $1 - \delta$:*

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

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

Lemma A.7.5. *(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 A.7.6. *(MLE guarantee, [CHYL23]) Fix $\delta \in (0,1)$. Then, with probability $1-\delta/2$,*

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

$$\mathbb{E}_{(s,a)\sim(\frac{1}{2}\rho_{t,h}(s,a)+\frac{1}{2}\rho'_{t,h}(s,a))}[\|\hat{\mathcal{P}}_{h,t}(\cdot|s,a) - \mathcal{P}^\star_h(\cdot|s,a)\|^2_{\mathrm{TV}}] \leq \zeta_t,$$

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

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

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