Towards Sample Efficient Reinforcement Learning with Function Approximation

by

Alex Ayoub

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

Abstract

This thesis proposes novel algorithmic ideas in reinforcement learning for regret minimization. These algorithmic ideas enjoy nice theoretical guarantees and are more practical in large problems than their alternatives. We focus on finite-horizon episodic RL. We propose model-based and model-free RL algorithms that are based on the optimism principle which allows us to derive regret bounds for our algorithms. In each episode the model-based algorithm constructs the set of models that are 'consistent' with the data collected. The criterion of consistency is based on the total squared error that the model incurs on the task of predicting state values as determined by the last value estimate along the transitions. The next value function is then chosen by solving the optimistic planning problem with the constructed set of models. We also propose a model-free algorithm inspired by the randomized least squares value iteration algorithm. Unlike existing upper-confidence-bound based approaches this algorithm drives exploration by simply perturbing the training data with judiciously chosen independent and identically distributed scalar noises. To attain optimistic value function estimation without resorting to a UCB-style bonus, we introduce a reward sampling procedure that guarantees optimism in the value estimates. For the model based case, we provide regret bounds on our algorithm and highlight its attractive properties through numerical experiments. For the model-free case, we show that randomizing the history multiple times and adding a regularizer, or data, that ensures the under explored regions have sufficient coverage is enough to get sublinear regret.

Thus, making significant progress toward more computationally efficient RL algorithms that also guarantee sublinear regret.

Preface

This thesis is comprised of two papers (Ayoub et al., 2020; Ishfaq et al., 2021 published at the International Conference on Machine Learning (ICML) 2020 and 2021 respectively. The main results of these works and my individual contributions to these results are outlined in detail in the Chapter 1.1, found in the introduction.

Acknowledgements

I would like to sincerely thank my supervisor Csaba Szepesvári for his time, patience, care, and support. Both professionally and personally, Csaba has supported me during changing times and encouraged me to pursue new and meaningful endeavours. His guidance has kept me focused and honest, and I look forward to continuing under his supervision. Thank you Csaba for constantly taking the time to indulge and nurture my curiosities.

I would also like to acknowledge my friends and labmates whose support and presence has made my life more meaningful. I thank my labmates: Kenny Young, Liam Peet-Pare, Tian Tian, J. Fernando Garcia, Khurram Javed, Roshan Shariff, Andrew Jacobsen, Erfan Miahi, Gábor Mihucz, and Shivam Garg for their conversations, discussions, and in some cases belays. I thank my friends: Connor Stephens, Mikhail Konobeev, Jincheng Mei, and Dhawal Gupta for their support and for their advice during times of uncertainty and during times of jubilation. I wish Mikhail, Jincheng, and Dhawal success as they move on to new opportunities. I thank my roommates Jim Christie-Brooks, Ghada Sayadi, Neha Rao, Claire Brown, and Alyvia Coney for making my home a place of rejuvenation and comfort. I thank Sara and Melad for their great care and for welcoming me into their lives. I wish them happiness has they welcome a new child into their lives. I thank my friends Michael Pedersen, Health Fulmer, and Grant Sennott for their emotional support. I thank my parents, Fadi and Samar Ayoub, for their example and for their love. Finally I thank my sister Amanda Ayoub whose presence in my life keeps me motivated and driven.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In reinforcement learning (RL), a core problem in artificial intelligence Russel and Norvig, 2003; Sutton and Barto, 2018, an agent learns to control a possibly complex, initially unknown environment in a sequential trial and error process. The application of RL algorithms to various domains, such as games, robotics and science, has witnessed phenomenal empirical advances during the last few years (e.g., AlQuraishi, 2019; Arulkumaran et al., 2019; Mnih et al., 2015; Silver et al., 2017). In online RL, an agent has to learn to act in an unknown environment "from scratch", collect data as she acts, and adapt her policy to maximize the reward collected, or, equivalently, to minimize her regret. Designing RL algorithms that provably achieve sublinear regret in some class of environments has been the subject of much research, mainly focusing on the so-called tabular, and linear-factored MDP settings (e.g., S. Agrawal and Jia, 2017; Azar et al., 2017; Dann et al., 2017; Dann et al., 2018; Jaksch et al., 2010; Jin et al., 2018; Jin et al., 2020; Osband et al., 2017; Osband et al., 2014; L. F. Yang and Wang, 2019). An appealing alternative to studying these structured cases is to consider learning and acting when the environment is described by a general model class, the a central topic of this thesis. Despite its appeal, as it appears, prior work, considered this option exclusively in a Bayesian setting. In particular, (M. J. A. Strens, 2000) introduced posterior sampling to RL, which was later analyzed by (Abbasi-Yadkori and Szepesvári, 2015; Osband and Van Roy, 2014; Theocharous et al., 2017) (for a more in-depth discussion of related work, the reader is referred to Chapter 6). As opposed to these works, in the present thesis we are interested in developing algorithms for bounding the worst-case expected regret for both model-based and model-free RL.

In Chapter 4, the specific setting that we adopt is that of episodic reinforcement learning in an environment where the unknown transition probability model that describes the environment's stochastic dynamics belongs to a family of models that is given to the learner. The model family $\mathcal{P}$ is a general set of models, and it may be either finitely parametrized or nonparametric. In particular, our approach accommodates working with smoothly parameterized models (e.g., Abbasi-Yadkori and Szepesvári, 2015), and can find use in both robotics (Kober et al., 2013) and queueing systems (Kovalenko, 1968). An illuminating special case is when elements of $\mathcal{P}$ take the form $P_\theta = \sum_i \theta_i P_i$ where $P_1, P_2, \dots, P_d$ are fixed, known basis models and $\theta = (\theta_1, \dots, \theta_d)$ are unknown, real-valued parameters. Model P_θ can be viewed as a linear mixture model that aggregates a finite family of known basic dynamical models (Modi et al., 2019). As an important special case, linear mixture models include the linear-factor MDP model of (L. F. Yang and Wang, 2019).

The main contribution of this chapter is a new model-based upper confidence RL algorithm. The main novelty is the criterion to select models that are deemed consistent with past data. As opposed to the most common approach where the models are selected based on their ability to predict next states or raw observations (cf. Jaksch et al., 2010; L. F. Yang and Wang, 2019 or (Abbasi-Yadkori and Szepesvári, 2015; S. Agrawal and Jia, 2017; Osband and Van Roy, 2014; Ouyang et al., 2017a; M. J. A. Strens, 2000) in a Bayesian setting), we propose to evaluate models based on their ability to predict the values of a value function at next states, where the value function used is an estimate of the optimal that our algorithm produces based on past information. In essence, our algorithm selects models based on their ability to produce small prediction errors in an appropriately constructed *value-targeted* regression (VTR) problem. Our algorithm combines VTR for constructing sets of plausible models with (standard) optimistic planning. The idea of using a value function estimate to "fit" models has been explored in the context of batch RL by (Farahmand, 2018).

VTR is attractive for multiple reasons: *(i)* Firstly, VTR permits model learning to *focus on task-relevant aspects* of the transition dynamics. This is important as the dynamics can be quite complicated and in a resource bounded setting, modelling irrelevant aspects of the dynamics can draw valuable resources away from modelling task-relevant aspects. *(ii)* Secondly, VTR poses model learning as a real-valued regression problem, which should be easier than the usual approaches to build probabilistic models. In particular, when state-representation available to the agent takes values in a high-dimensional space then building a faithful probability model can be highly demanding. *(iii)* Thirdly, VTR aims to control directly what matters in terms of controlling the regret. Specifically the objective used in value-targeted is obtained from an expression that upper bounds the regret, hence it is natural to expect that minimizing value prediction errors will lead to a small regret.

An additional attractive feature of our algorithm is its modular structure. As a result, advances on the components (faster optimistic planning, tighter confidence sets for VTR) are directly translated into an improved algorithm. On the skeptical side, one may question whether VTR is going "too far" in ignoring details of the dynamics. In particular, since the value functions used in defining the regression targets are derived based on imperfect knowledge, the model may never get sufficiently refined in a way that would keep the regret small. Similarly, one may be worried about that by ignoring details of the observations (i.e., the identity of states), the approach advocated is ignoring information available in the data, which may slow down learning. This leads to central question of Chapter 4:

Is value-targeted regression sufficient and efficient for model-based online RL?

The main contribution of this chapter is a positive answer to this question. In particular, the regret bounds we derive conclusively show that the despite the imperfection and non-stationarity of the value targets, our algorithm will not get "stuck" (i.e., it enjoys sublinear regret). Our results further suggest that in the worst case sense, for common settings, there may be no performance penalty associated with using value-targeted regression. We are careful here as this conclusion is based on comparing worst-case upper bounds, which cannot provide a definitive answer. Finally, it is worth noting that the regret bound does not depend on the size of either the state *or* the action space.

To complement the theoretical findings, results from a number of small-scale, synthetic experiments confirm that our algorithm is competitive in terms of its regret. The experiments also allow us to conclude that it is value-targeted regression *together* with optimistic planning that is effective. In particular, if optimism is taken away (i.e., ϵ-greedy for exploration), value-targeted regression performs worse than using a canonical approach to estimate the model. Similarly, if value-targeted regression is taken away, optimism together with the canonical model-estimation approach is less effective. Note that our results do not rule out that certain combinations of value-targeted regression and canonical model building are more effective than value-targeted regression. In fact, given the vast number of possibilities, we find this to be quite probable. While our proofs can be adjusted to deal with adding simultaneous alternative targets, sadly, our current theoretical tools are unable to capture the tradeoffs that one expects to arise as a result of such modifications.

It is interesting to note that, in an independent and concurrent work, value-targeted regression has also been suggested as the main model building tool of the MuZero algorithm (Schrittwieser et al., 2019), which was empirically evaluated on a number of RL benchmarks, such as the 57 Atari "games", the game of "Go", chess and shogi, and was found to be highly competitive with its state-of-the-art alternatives. This reinforces the conclusion that training models using value-targeted regression is indeed a good approach to build effective model-based RL algorithms. Since MuZero does *not* implement optimistic planning and our results show that optimism is not optional in a worst case sense, the good results of MuZero on these benchmark may seem to contradict our experimental findings that value-targeted regression is ineffective without an appropriate, 'smart' exploration component. However, there is no contradiction: Smart exploration may be optional in some environments; our experiments show that it is not optional on some environments. In short, for robust performance across a wide range of environments, smart exploration is necessary but smart exploration may be optional in some environments. To illustrate the strength of this general technique, we specialize the regret bound for the case of linear mixture models, for which we prove that the expected cumulative regret is at most $O(d\sqrt{H^3T})$, where H is the episode length, d is the number of model parameters and T is the total number of steps that the RL algorithm interacts with its environment. To complement the upper bound, for the linear case we also provide a regret lower bound $\Omega(\sqrt{HdT})$ by adapting a lower bound that has been derived earlier for tabular RL.

In Chapter 5, we investigate a recently rediscovered exploration idea called Thompson sampling (TS) (Osband et al., 2013; Thompson, 1933). It is motivated by the Bayesian perspective on RL, in which we have a prior distribution over the model or the value function; then we draw a sample from this distribution and compute a policy based on this sample. Theoretical guarantees exist for both Bayesian regret (Russo and Van Roy, 2013) and worst-case regret (S. Agrawal and Jia, 2017) for this approach. Although TS is conceptually simple, in many cases the posterior is intractable to compute and the prior may not exist at all. Recently, approximate TS, also known as randomized least squares value iteration (RLSVI) or following the perturbed leader (Kveton et al., 2019), has received significant attention due to its good empirical performance. It has been proven that RLSVI enjoys sublinear worst-case or frequentist regret in tabular RL, by simply adding Gaussian noise on the reward (P. Agrawal et al., 2020; Russo, 2019). However, in the improved bound for tabular MDP (P. Agrawal et al., 2020) and linear MDP (Zanette, Brandfonbrener, et al., 2020), the uncertainty of the estimates still needs to be computed in order to perform optimistic exploration; it is unknown whether this can be removed. Moreover, this computation is difficult to do in the general function approximation setting.

In this chapter, we propose a novel exploration idea called optimistic reward sampling, which combines OFU and TS organically. The algorithm, named Least-Squares Value Iteration with Perturbed History Exploration (LSVI-

PHE), is surprisingly simple: we perturb the reward several times and act greedily with respect to the maximum of the estimated state-action values. The intuition is that after the perturbation, the estimate has a constant probability of being optimistic, and sampling multiple times guarantees that the maximum of these sampled estimates is optimistic with high probability. Thus, our algorithm utilizes approximate TS to achieve optimism.

Similar algorithms have been shown to work empirically, including SUNRISE (Lee et al., 2020), NoisyNet (Fortunato et al., 2017) and bootstrapped DQN (Osband, Blundell, et al., 2016). However, the theoretical analysis of perturbation-based exploration is still missing. We prove that it enjoys near optimal regret $\widetilde{O}(\sqrt{H^3d^3T})$ for linear MDP and the sampling time is only $M = \widetilde{O}(d)$. We also prove similar bounds for the general function approximation case, by using the notion of eluder dimension (Russo and Van Roy, 2013; Wang et al., 2020). In addition, this algorithm is computationally efficient, as we no longer need to compute the upper confidence bound. In the experiments, we find that a small sampling time M is sufficient to achieve good performance, which suggests that the theoretical choice of $M = \widetilde{O}(d)$ is too conservative in practice.

Optimistic reward sampling can be directly plugged into most RL algorithms, improving the sample complexity without harming the computational cost. The algorithm only needs to perform perturbed regression. To our best knowledge, this is the first online RL algorithm that is both computationally and statistically efficient with linear function approximation and general function approximation.

1.1 List of Contributions

- The UCRL-VTR Algorithm, (pg 20)*[1]
- Criterion that selects models that are deemed consistent with past data, (pg 21)*

[1] $*$ denotes major initiative

- Regret bounds for UCRL-VTR under the Assumption of Bounded Eluder Dimension, (pg 24)
- Regret Bound for UCRL-VTR under the Linear Mixture Model Assumption, (pg 24)
- Lower Bound for the Linear Mixture Model Assumption, (pg 25))
- Numerical Experiments for UCRL-VTR, (pg 27)*
- Optimistic Sampling for Model-Free Reinforcement Learning, (pg 33)*
- The LSVI-PHE Algorithm for General Value Function Approximation via Bounded Eluder Dimension, (pg 33)*
- The LSVI-PHE Algorithm for Solving the Linear MDP, (pg 35)*
- Regret Bounds for LSVI-PHE under the Bounded Eluder Dimension Assumption, (pg 40)
- Regret Bounds for LSVI-PHE for the Linear MDP, (pg 41)*
- Numerical Experiments for LSVI-PHE, (pg 41)*

Chapter 2

Problem Setting

We study learning to control episodic Markov decision processes (MDPs, for short), described by a tuple $M = (\mathcal{S}, \mathcal{A}, P, r, H, s_\circ)$. Here, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P = (P_h)_{h=1}^H$ are the state transition probability distributions,, r is a reward function, $H > 0$ is the episode length, or horizon, and $s_\circ \in \mathcal{S}$ is the initial state. In this thesis, both the state and action space are assumed to be discrete. Extra care is required in order to extend these results to the case when both the states and actions are continuous. For each $h \in [H]$, $P_h(\cdot \,|\, s, a)$ is the transition kernel over the next states if action a is taken at state s during the h-th time step of the episode. In the online RL problem, the learning agent is given $\mathcal{S}$, $\mathcal{A}$, H and r but does not know P.[1] The agent interacts with its environment in a number of episodes. Each episode begins at state $s_\circ$ and ends after the agent made H decisions. At state $s \in \mathcal{S}$, the agent, after observing the state s, can choose an action $a \in \mathcal{A}$. As a result, the immediate reward $r(s, a)$ is incurred. Then the process transitions to a random next state $s' \in \mathcal{S}$ according to the transition law $P(\cdot|s, a)$.[2] The agent's goal is to maximize the total expected reward received over time.

If P is known, the behavior that achieves maximum expected reward over any number of episodes can be described by applying a deterministic policy π. Such a policy is a mapping from $\mathcal{S} \times [H]$ into $\mathcal{A}$ (note for a natural number

[1] Our results are easy to extend to the case when r is not known.

[2] The precise definitions require measure-theoretic concepts (Bertsekas and Shreve, 1978), i.e., P is a Markov kernel, mapping from $\mathcal{S} \times \mathcal{A}$ to distributions over $\mathcal{S}$, hence, all these spaces need to be properly equipped with a measurability structure. For the sake of readability and also because they are well understood, we omit these technical details.

n, $[n] = \{1, \ldots, n\}$). Following the policy means that the agent upon encountering state s in stage h will choose action $\pi(s, h)$. In what follows, we will use $\pi_h(s)$ as an alternate notation, as this makes some of the formulae more readable. (We will follow the same convention of moving h to the subindex position when it comes to other functions whose domain is $\mathcal{S} \times [H]$.)

The value function $V^\pi : \mathcal{S} \times [H] \to \mathbb{R}$ of a policy π is defined via

$$V_h^\pi(s) = \mathbb{E}_\pi \left[\sum_{i=h}^{H} r(s_i, \pi(s_i)) \,\middle|\, s_h = s \right], \qquad s \in \mathcal{S}\,,$$

where s_i is the state encountered at stage $i \in [H]$ and the subscript π (which we will often suppress) signifies that the probabilities underlying the expectation are jointly governed by π and P (P is suppressed for clarity). An optimal policy π^* and the optimal value function V^* are defined to be a policy and its value function such that $V_h^\pi(s)$ with $\pi = \pi^*$ achieves the maximum value among all possible policies for any $s \in \mathcal{S}$ and $h \in [H]$. Note that this is well-defined and in fact, as noted above, there is no loss of generality in restricting the search of optimal policies to deterministic policies.

In online RL, a learning agent will use all past observations to come up with its decisions. The performance of such an agent is measured by its regret, which is the total reward the agent misses because she did not follow the optimal policy from the beginning. In particular, the total expected regret of an agent $\mathcal{A}$ across K episodes is given by

$$R(T) = \mathbb{E}\left[\sum_{k=1}^{K} \Big(V_1^*(s_1^k) - \sum_{h=1}^{H} r(s_h^k, a_h^k) \Big) \right], \tag{2.1}$$

where $T = KH$ is the total number of time steps that the agent interacts with its environment, $s_1^k = s_\circ$ is the initial state at the start of the k-th episode, and $s_1^1, a_1^1, \ldots, s_H^k, a_H^k, s_1^2, a_1^2, \ldots, s_H^2, a_H^2, s_1^K, a_1^K, \ldots, s_H^K, a_H^K$ are the $T = KH$ state-action pairs in the order that they are encountered by the agent. The regret is sublinear if $R(T)/T \to 0$ as $T \to \infty$. As is well known, the worst-case value of $R(T)$ over a set of sufficiently large model class, grows at least as fast as $\sqrt{T}$ regardless of the algorithm used (e.g., Jaksch et al., 2010).

In Chapter 4, we aim to design a general model-based reinforcement learning algorithm for the time homogeneous model, $P_1 = P_2 = \cdots = P_h$, with a guaranteed sublinear regret, for any (not too large) family of transition models:

Assumption 1 (Known Transition Model Family)**.** *The unknown transition model P belongs to a family of models $\mathcal{P}$ which is available to the learning agent. The elements of $\mathcal{P}$ are transition kernels mapping state-action pairs to signed distributions over $\mathcal{S}$.*

That we allow signed distributions only increases generality; this may be important when one is given a model class that can be compactly represented but only when it also includes non-probability kernels (see Pires and Szepesvári, 2016 for a discussion of this). Parametric and nonparametric transition models are common in modelling complex stochastic controlled systems. For example, robotic systems are often smoothly parameterized by unknown mechanical parameters such as friction and parameters that describe the geometry of the robot.

An important special case is the class of linear mixture models:

Definition 1 (Linear Mixture Models)**.** *We say that $\mathcal{P}$ is the class of linear mixture models with component models $P_1, \ldots, P_d$ if $P_1, \ldots, P_d$ are transition kernels that map state-action pairs to signed measures and $P \in \mathcal{P}$ if and only if there exists $\theta \in \mathbb{R}^d$ such that for all $(s, a) \in \mathcal{S} \times \mathcal{A}$,*

$$P(ds'|s,a) = \sum_{j=1}^{d} \theta_j P_j(ds'|s,a) = P_\cdot(ds'|s,a)^\top \theta \,. \tag{2.2}$$

The linear mixture model can be viewed as a way of aggregating a number of known basis models as considered by (Modi et al., 2019). We can view each $P_j(\cdot|\cdot)$ as a basis latent "mode". When θ is restricted to lie in the $(d-1)$ simplex, the actual transition is a probabilistic mixture of these latent modes. As an example of when mixture models arise, consider large-scale queueing networks where the arrival rate and job processing speed for each queue is not known. By using a discrete-time Bernoulli approximation, the transition probability matrix from time t to $t + \Delta t$ becomes increasingly close to linear with respect to the unknown arrival/processing rates as $\Delta t \to 0$. In this case, it is common to model the discrete-time state transition as a linear aggregation of arrival/processing processes with unknown parameters (Kovalenko, 1968).

Another interesting special case is the linear-factored MDP model of (L. F. Yang and Wang, 2019) where, assuming a discrete state space for a moment, P takes the form

$$\begin{aligned} P(s'|s,a) &= \phi(s,a)^\top M \psi(s') \\ &= \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} M_{ij} \left[\psi_j(s')\phi_i(s,a)\right], \end{aligned}$$

where $\phi(s,a) \in \mathbb{R}^{d_1}, \psi(s') \in \mathbb{R}^{d_2}$ are given features for every $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$ (when the state space is continuous, ψ becomes an $\mathbb{R}^{d_2}$-valued measure over $\mathcal{S}$). The matrix $M \in \mathbb{R}^{d_1 \times d_2}$ is an unknown matrix and is to be learned. It is easy to see that the factored MDP model is a special case of the linear mixture model (A.11) with each $\psi_j(s')\phi_i(s,a)$ being a basis model (this should be replaced by $\psi_j(ds')\phi_i(s,a)$ when the state space is continuous). In this case, the number of unknown parameters in the transition model is $d = d_1 \times d_2$. In this setting, without additional assumptions, our regret bound will match that of (L. F. Yang and Wang, 2019).

2.1 Additional Notation

For any positive integer n, we denote the set $\{1, 2, \ldots, n\}$ by $[n]$. For any set A, $\langle \cdot, \cdot \rangle_A$ denotes the inner product over set A. For a positive definite matrix $A \in \mathbb{R}^{d \times d}$ and a vector $x \in \mathbb{R}^d$, we denote the norm of x with respect to matrix A by $\|x\|_A = \sqrt{x^T A x}$. We denote the cumulative distribution function of the standard Gaussian by $\Phi(\cdot)$. For function growth, we use $\widetilde{\mathcal{O}}(\cdot)$, ignoring poly-logarithmic factors. The performance of function f on dataset $\mathcal{D} = \{(x_t, y_t)\}_{t \in [|\mathcal{D}|]}$ is defined by $L(f \mid \mathcal{D}) = \left(\sum_{t=1}^{|\mathcal{D}|} (f(x_t) - y_t)^2\right)^{1/2}$. The empirical ℓ_2 norm of function f on input set $\mathcal{Z} = \{x_t\}_{t \in [|\mathcal{Z}|]}$ is defined by $\|f\|_{\mathcal{Z}} = \left(\sum_{t=1}^{|\mathcal{Z}|} f(x_t)^2\right)^{1/2}$. Given a function class $\mathcal{F} \subseteq \{f : X \to \mathbb{R}\}$, we define the width function given some input x as $w(\mathcal{F}, x) = \max_{f, f' \in \mathcal{F}} f(x) - f'(x)$.

Chapter 3

Optimism in the Face of Uncertainty

The work discussed in this chapter comes from Chapter 5 (Lattimore and Szepesvári, 2018 and Chapters 1 and 2 (Boucheron et al., 2013)

Before we can discuss provably and practically efficient algorithms for online reinforcement learning we will need to properly motivate then introduce the principle of Optimism in the Face of Uncertainty (OFU). Optimism in the Face of Uncertainty states that one should act as if the world is a good as plausibly possible. Imagine you are trying to decide which brand of coffee is the most delicious brand of coffee. You would first try some brands and observe some feedback about the brand's deliciousness. Some brands will stand out over other brands, however, maybe one brand was better than the other due to confounding factors such as: when you drank it, what you ate the night before, etc... So you now try the brand which you think is currently the most delicious, turns out you now hate it. You update your internal estimate of this brand's deliciousness, it is not longer the most delicious brand out of the brands you've tried. You then try the brand you now perceive to be the most delicious brand. You repeat this process until you are very certain that you have found the most delicious brand.

In the above example, you only try the brand of coffee you perceive to be the most delicious at a given time interval. You then updates your estimate of a brand's deliciousness. You then repeats this process until you find the brand that you think is most delicious. This is how the OFU principles works at a high level. We will now formulate the OFU principle using tools from probability theory.

3.1 Concentration of Measure

In the above example, the deliciousness of each brand of coffee is initially unknown. Recall that the most delicious brand, or the optimal brand, is the brand that is most consistently delicious, or the brand which has the highest average deliciousness. Since the mean deliciousness, or mean pay-offs, are initially unknown, it must be learned from a stream of observations or data. One may then wonder how many observations is necessary in order for you to confidently learn the mean deliciousness of a brand of coffee. Concentration inequalities allow us to bound the probability that our current estimate of some mean differs from its expected value by a fixed amount. Formally we care about bounding the following

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \varepsilon\right) \leq \delta$$

where X_i is some i.i.d observations with mean μ, $\varepsilon > 0$, and $\delta \in (0, 1]$. We spend the rest of this section bounding the above probability and showing how these bounds can be related back to the OFU principle.

Suppose that $X, X_1, X_2, ..., X_n$ is a sequence of independent and identically distributed (i.i.d.) random variables, and assume that the mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \mathbb{V}[X]$ exist. With just the following information we can bound the probability the sample mean of X_i, denoted $\widehat{\mu}$, is far from its true mean using Chebyshev's inequality.

Lemma 1. *(Chebyshev) For any random variable X and some $\varepsilon > 0$, we have*

$$\mathbb{P}\left(|\widehat{\mu} - \mu| \geq \varepsilon\right) \leq \frac{\sigma^2}{\varepsilon}$$

Chebyshev's inequality is nice because it allows us to bound how far away the sample mean is from the true mean with just the assumption that a random variable has defined first and second moments (mean and variance). However, one will find that with that if we make some more assumptions on our random variable, we could get much tighter bounds than what is given to us by Chebyshev's inequality.

3.1.1 Exponentially Decreasing Tails and the Cramer-Chernoff Method

In this section, we will introduce subgaussianity assumption the use it to derive a tighter inequality on the concentration of a random variable that is also subgaussian.

Definition 2. *(Subguassianity) A random variable, X, is said to be subgaussian with proxy σ, or σ-subgaussian, if for all $\lambda \in \mathbb{R}$ we have*

$$\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\frac{\xi^2\sigma^2}{2}}$$

Note that a subgaussian random variable is a random variable that can be upper bounded by a Gaussian random variable with variance σ^2. Now let us prove some facts about subgaussian random variables

Lemma 2. *Assume X is subgaussian with proxy σ and that X_1 and X_2 are independent and subgaussian with proxy σ_1 and σ_2 respectively. Then the following properties hold*

- *The mean of X, $\mathbb{E}[X] = 0$ and the variance of X, $\mathbb{V}[X] \leq \sigma^2$.*
- *For all $c \in \mathbb{R}$, cX is subgaussian with proxy $|c|\sigma$.*
- *The sum of the the random variables X_1 and X_2, $X_1 + X_2$, is subgaussian with proxy $\sqrt{\sigma_1^2 + \sigma_2^2}$*

Proof. For the first bullet we will use the Taylor expansion of the exponential function and the definition of subgaussianity to write

$$\mathbb{E}[e^{\lambda X}] = \sum_{i=0}^{\infty} \frac{\lambda^i}{i!}\mathbb{E}[X^i] \leq \sum_{i=0}^{\infty} \frac{\lambda^{2i}\sigma^{2i}}{2^i i!} = e^{\frac{\lambda^2\sigma^2}{2}}$$

Expanding both sides we get

$$1 + \lambda\mathbb{E}[X] + \frac{\lambda^2}{2}\mathbb{E}[X^2] + o(\lambda^2) \leq 1 + \frac{\lambda^2\sigma^2}{2} + o(\lambda^2)$$

Now we let $\lambda \to 0$ in order to get rid of the truncation error, $o(\lambda^2)$. So we now evaluate the following

$$\begin{aligned}\lambda\mathbb{E}[X] + \frac{\lambda^2}{2}\mathbb{E}[X^2] &\leq \frac{\lambda^2\sigma^2}{2} \\ 2\frac{\mathbb{E}[X]}{\lambda} + \mathbb{E}[X^2] &\leq \sigma^2\end{aligned}$$

The above expression is only guaranteed to hold for all $\lambda \in \mathbb{R}$ if $\mathbb{E}[X] = 0$. Since $\mathbb{E}[X] = 0$ and $\mathbb{V}[X] = E[X^2] - E[X]^2 = E[X^2] \leq \sigma^2$ we have that $\mathbb{V}[X] \leq \sigma^2$. Thus we have shown the first bullet.

For the second bullet, if $|X|$ is subgaussian with proxy σ and let $\lambda = c\xi$ for all $\xi \in \mathbb{R}$.

$$\mathbb{E}[e^{\xi cX}] = \mathbb{E}[e^{\lambda X}] \leq e^{\frac{\sigma^2\lambda^2}{2}} = e^{\frac{|c\sigma|^2\xi^2}{2}}$$

Thus cX is subgaussian with proxy $|c|\sigma$.

For the third bullet we will use the fact that since X_1, X_2 are independent, we can write $\mathbb{E}[f(X_1)f(X_2)] = \mathbb{E}[f(X_1)]\mathbb{E}[f(X_2)]$ for any function f.

$$\begin{aligned}\mathbb{E}[e^{\lambda(X_1+X_2)}] &= \mathbb{E}[e^{\lambda X_1}e^{\lambda X_2}] = \mathbb{E}[e^{\lambda X_1}]\mathbb{E}[e^{\lambda X_2}] \\ &\leq e^{\frac{\lambda^2\sigma_1^2}{2}}e^{\frac{\lambda^2\sigma_2^2}{2}} = e^{\frac{\lambda^2(\sigma_1^2+\sigma_2^2)}{2}}\end{aligned}$$

Thus $X_1 + X_2$ must be subgaussian with proxy $\sqrt{\sigma_1^2 + \sigma_2^2}$. □

Theorem 3. *Let X be a σ-subgaussian random variable. Then for any $\varepsilon > 0$ we have*

$$\mathbb{P}\left(|X| \geq \varepsilon\right) \leq 2e^{-\frac{\varepsilon^2}{2\sigma^2}}$$

Proof. This proof uses an approach called the Cramer-Chernoff method.

$$\mathbb{P}(X \geq \varepsilon) = \mathbb{P}\left(\mathbb{E}[e^{\lambda X}] \geq e^{\lambda\varepsilon}\right) \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda\varepsilon}} \leq e^{\frac{\lambda\sigma^2}{2}-\lambda\varepsilon}$$

where the first inequality is just Markov's inequality and the second inequality is by the definition of a subgaussian random variable. This bounds the right tail but we must also bound the left tail.

$$\mathbb{P}(X \leq \varepsilon) = \mathbb{P}\left(\mathbb{E}[e^{\lambda X}] \leq e^{\lambda\varepsilon}\right) \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda\varepsilon}} \leq e^{\frac{\lambda\sigma^2}{2} - \lambda\varepsilon}$$

Choosing $\lambda = \frac{\varepsilon}{\sigma^2}$ and using a union bound, $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$, completes the proof. □

Another way to write the result of Theorem 3 is to set $\varepsilon = \sqrt{2\sigma^2 \log(1/\delta)}$, then we get

$$\mathbb{P}\left(|X| \geq \sqrt{2\sigma^2 \log(1/\delta)}\right) \leq 2\delta$$

This bound shows for small enough choices of δ we can bound tail events with high probability.

3.2 Upper Confidence Bounds

In this section we will use the results from the previous section in order to create a high probability upper bound on the true mean μ of a subgaussian random variable X. Ideally we would like this upper bound to shrink proportional to the number of observations we have of X. In sequential decision making problems, like bandits or RL, we often care about learning the mean of some optimal action or value. However, we want to do this as quickly as possible as there is usually a cost to taking bad actions or following a sub-optimal value. Thus upper confidence bounds on empirical estimates, gives us a way to reason about the goodness of an action while taking into account we might be uncertain about that action. We initially incentive estimates we are uncertain of and as the uncertainty decreases we know we must be close enough to the true estimate. However, if the sample mean of a given estimate is much lower than the sample mean of the other estimates then we can be certain that its true mean is also much lower the the other actions true means. Thus we can quickly rule out action that are significantly more bad than other actions.

Corollary 4. *Let $X_i - \mu$ bbe independent, subgaussian random variables with proxy σ. Let $\widehat{\mu} \doteq \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for all $\varepsilon \geq 0$*

$$\mathbb{P}\left(\widehat{\mu} \geq \mu + \varepsilon\right) \leq e^{-\frac{n\varepsilon^2}{2\sigma^2}}$$

and

$$\mathbb{P}\left(\widehat{\mu} \leq \mu - \varepsilon\right) \leq e^{-\frac{n\varepsilon^2}{2\sigma^2}}$$

Proof. By Lemma 2, we have that $\widehat{\mu} - \mu$ is subgaussian with proxy $\frac{\sigma}{\sqrt{n}}$ since $\widehat{\mu} - \mu = \sum_i (X_i - \mu)/n$. We then use Theorem 3 which completes the proof. □

The result of Corollary 4 states that with probability at least $1-\delta$, for any $\delta \in (0,1]$ for a subgaussian random variable X, its empirical mean $\widehat{\mu}$ cannot be too far from its true mean μ,

$$\mu \leq \widehat{\mu} + \sqrt{\frac{2\sigma^2 \log(1+\delta)}{n}}. \tag{3.1}$$

By symmetry, we can say with probability $1-\delta$,

$$\mu \leq \widehat{\mu} - \sqrt{\frac{2\sigma^2 \log(1+\delta)}{n}}.$$

The term $\sqrt{\frac{2\sigma^2 \log(1+\delta)}{n}}$ represents the uncertainty in our estimate of $\widehat{\mu}$. As the number of observations goes to infinity, $n \to \infty$, we are more certain in our estimate of the sample mean. We also have that sample mean converges to the true mean, $\widehat{\mu} \to \mu$, at a rate of $O(1/\sqrt{n})$. This gives us high probability finite time bounds on how much our sample mean can deviate from the true mean. Concentration of empirical or sample estimates is a crucial tool is provably efficient algorithms for sequential decision making processes.

We are now ready to construct an upper confidence bound, which is a high probability overestimate of an unknown mean. The reason why this is a good idea in sequential decision making processes is quite intuitive. Assume that the upper confidence bound assigned to the optimal value is indeed an overestimate. Then another value will only be chosen if its upper confidence bound is larger than the upper confidence bound of the optimal value. However, this can only happen if this sub-optimal value's upper confidence bound is larger than that of the optimal value. This cannot happen too often as choosing a value leads to more observations of that value which in turn shrink its upper confidence bound as the estimate of that value converges to its true value which is less than the true value of the optimal value. We now will define an upper confidence bound using the results provided above. Let $(X_i)_{i=1}^n$ be a sequence of independent subgaussian random variables with proxy σ. Let the mean of these subgaussian random variables be μ and let $\widehat{\mu} = \frac{1}{n}\sum_i X_i$. From Equation 3.1 we have for all $\delta \in (0, 1]$,

$$\mathbb{P}\left(\mu \geq \widehat{\mu} + \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}\right) \leq \delta$$

Thus we have constructed a high probability upper confidence bound on the true mean. These types of arguments and bounds are the essential tools behind showing algorithms are provably efficient for sequential decision making processes. In Sections 4.1 and 5.1, we will use upper confidence bounds in order to construct algorithms for provably efficient RL.

Chapter 4

Upper Confidence Reinforcement Learning with Value-Targeted Regression

The work discussed in this chapter is based on the publication (Ayoub et al., 2020)

4.1 The Algorithm

Our algorithm (Alg 1) can be viewed as a generalization of UCRL (Jaksch et al., 2010), following ideas of (Osband and Van Roy, 2014). In particular, at the beginning of episode $k = 1, 2, \ldots, K$, the algorithm first computes a subset B_k of the model class $\mathcal{P}$ that contains the set of models that are deemed to be consistent with all the data that has been collected in the past. The new idea, value-targeted regression is used in the construction of B_k. The details of how this is done are postponed to a later section. Given B_k, the algorithm needs to find the model that maximizes the optimal value, and the corresponding optimal policy. Denoting by V_P^* the optimal value function under a model P, this amounts to finding the model $P \in B_k$ that maximizes the value $V_{P,1}^*(s_1^k)$. Given the model P_k that maximizes this value, an optimal policy is extracted from the model as in standard dynamic programming, detailed in the next section. At the end of the episode, the data collected is used to refine the confidence set B_k.

Algorithm 1 UCRL-VTR

```
Input:  Family of MDP models $\mathcal{P}$, $d, H, T = KH$, sequence $\{\beta_k\}_{k=1,2,\ldots}$.
$B_1 = \mathcal{P}$
for $k = 1, 2, \ldots, K$ do
    Observe the initial state $s_1^k$ of episode $k$
    Optimistic planning:
        $P_k = \text{argmax}_{P' \in B_k} V^*_{P',1}(s_1^k)$
        Compute $Q_{1,k}, \ldots Q_{H,k}$ for $P_k$ using (4.1)
    for $h = 1, 2, \ldots, H$ do
        Choose the next action greedily with respect to $Q_{h,k}$:
            $a_h^k = \arg\max_{a \in \mathcal{A}} Q_{h,k}(s_h^k, a)$
        Observe state $s_{h+1}^k$
        Compute and store value predictions: $y_{h,k} \leftarrow V_{h+1,k}(s_{h+1}^k)$
    end for
    Construct confidence set using VTR as shown in Sec 4.1.2:
        $B_{k+1} = \{P' \in \mathcal{P} | L_{k+1}(P', \widehat{P}_{k+1}) \le \beta_k\}$
end for
```

4.1.1 Model-Based Optimistic Planning

Upper confidence methods are prominent in sequential online learning. As noted before, we let

$$P_k = \text{argmax}_{P' \in B_k} V^*_{P',1}(s_1^k).$$

Given model P_k, the optimal policy for P_k can be computed using dynamic programming. In particular, for $1 \le h \le H + 1$, define

$$\begin{aligned} &Q_{H+1,k}(s, a) = 0, \\ &V_{h,k}(s) = \max_{a \in \mathcal{A}} Q_{h,k}(s, a), \\ &Q_{h,k}(s, a) = r(s, a) + \langle P_k(\cdot|s, a), V_{h+1,k} \rangle, \end{aligned} \tag{4.1}$$

where for a measure μ and function f that share a common domain, $\langle \mu, f \rangle$ denotes the integral of f with respect to μ. It follows that, taking the action at state s and stage h that maximizes $Q_{h,k}(s, \cdot)$ gives an optimal policy for model P_k. As long as $P \in B_k$ with high probability, the preceding calculation gives an optimistic (that is, upper) estimate of value of an episode. Next, we show how to construct the confidence set B_k.

4.1.2 Value-Targeted Regression for Confidence Set Construction

Every time we observe a transition (s, a, s') with $s' \sim P(\cdot|s, a)$, we receive information about the model P. A standard approach to use this information would be either using a maximum likelihood approach, or regressing "onto" s'. As our goal is not to find the best model, we propose an alternate approach where we set up a regression problem where the model is used to predict the value assigned to s' by our more recent value function estimate:

$$\begin{aligned} \widehat{P}_{k+1} &= \operatorname{argmin}_{P' \in \mathcal{P}} \sum_{k'=1}^{k} \sum_{h=1}^{H} L'(P'), \qquad (4.2) \\ L'(P') &= \Big(\langle P'(\cdot|s_h^{k'}, a_h^{k'}), V_{h+1,k'} \rangle - y_{h,k'} \Big)^2 \\ y_{h,k'} &= V_{h+1,k'}(s_{h+1}^{k'}) \,, \quad h \in [H], k' \in [k] \,. \end{aligned}$$

In the above regression procedure, the regret target keeps changing as the algorithm refines the value estimates. This is in contrast to typical supervised learning for building models, where the regression targets are often fixed objects (such as raw observations, features or keypoints; e.g. Abbasi-Yadkori and Szepesvári, 2015; S. Agrawal and Jia, 2017; Jaksch et al., 2010; Kaiser et al., 2019; Osband and Van Roy, 2014; Xie et al., 2016; L. F. Yang and Wang, 2019). For a confidence set construction, we get inspiration from Proposition 5 in the paper of (Osband and Van Roy, 2014). The set is centered at $\widehat{P}_{k+1}$. Defining

$$\begin{aligned} & L_{k+1}(P, \widehat{P}_{k+1}) \\ &= \sum_{k'=1}^{k} \sum_{h=1}^{H} \Big(\langle P(\cdot|s_h^{k'}, a_h^{k'}) - \widehat{P}_{k+1}(\cdot|s_h^{k'}, a_h^{k'}), V_{h+1,k'} \rangle \Big)^2 \end{aligned}$$

we let

$$B_{k+1} = \{ P' \in \mathcal{P} \mid L_{k+1}(P', \widehat{P}_{k+1}) \le \beta_{k+1} \} \,,$$

where the value of β_k is obtained using a calculation similar to that done in Proposition 5 of the paper of (Osband and Van Roy, 2014). In turn, this calculation is based on the nonlinear least-squares confidence set construction from (Russo and Van Roy, 2014), which we describe and refine in the appendix. It is not hard to see that the confidence set can also be written in the alternative form

$$B_{k+1} = \{P' \in \mathcal{P} \mid \widetilde{L}_{k+1}(P') \leq \widetilde{\beta}_{k+1}\}$$

with a suitably defined $\widetilde{\beta}_{k+1}$ and where

$$\widetilde{L}_{k+1}(P') = \sum_{k'=1}^{k} \sum_{h=1}^{H} \quad \Big(\langle P'(\cdot|s_h^{k'}, a_h^{k'}), V_{h+1,k'} \rangle - y_{h,k'} \Big)^2 \, .$$

Note that the above formulation strongly exploits that the MDP is time-homogeneous: The same transition model is used at all stages of an episode. When the MDP is time-inhomogeneous, the construction can be easily modified to accommodate this.

4.1.3 Implementation of UCRL-VTR

Algorithm 1 gives a general and modular template for model-based RL that is compatible with regression methods/optimistic planners. While the algorithm is conceptually simple, and the optimization and evaluation of the loss in value-targeted regression appears to be at advantage in terms of computation as compared to standard approaches typically used in model-based RL, the implementation of UCRL-VTR is nontrivial in general and for now it requires a case-by-base design.

Computation efficiency of the algorithm depends on the specific family of models chosen. For the linear-factor MDP model considered by (L. F. Yang and Wang, 2019), the regression is linear and admits efficient implementation; further, optimistic planning for this model can be implemented in poly(d) time by using Monte-Carlo simulation and sketching as argued in the cited paper. Other ideas include loosening the confidence set to come up with computationally tractable methods, or relaxing the requirement that the same model is used in all stages. This latter idea is what we use in our experiments. In the general case, optimistic planning is computationally intractable. However, we expect that randomized (e.g. Lu and Van Roy, 2017; Osband et al., 2017;

Osband et al., 2014) and approximate dynamic programming methods (tree search, roll out, see e.g., Bertsekas and Tsitsiklis, 1996) will often lead to tractable and good approximations. As was mentioned above, in some special cases these have been rigorously shown to work. In similar settings, the approximation errors are known to mildly impact the regret (Abbasi-Yadkori and Szepesvári, 2015) and we expect the same to hold in our setting. If we look beyond methods with rigorous guarantees, there are practical deep RL algorithms that implement parts of UCRL-VTR. As mentioned earlier, the MuZero algorithm of (Schrittwieser et al., 2019) is a state-of-the-art algorithm on multiple domains and this algorithm implements value-targeted-regression to learn a model which is fed to a planner that uses Monte Carlo tree search, although the planner does not implement optimistic planning.

4.2 Theoretical Analysis

We will need the concept of Eluder dimension due to Russo and Van Roy, 2014. Let $\mathcal{F}$ be a set of real-valued functions with domain $\mathcal{X}$. For $f \in \mathcal{F}$, $x_1, \dots, x_t \in \mathcal{X}$, introduce the notation $f|_{(x_1,\dots,x_t)} = (f(x_1), \dots, f(x_t))$. We say that $x \in \mathcal{X}$ is ϵ-independent of $x_1, \dots, x_t \in \mathcal{X}$ given $\mathcal{F}$ if there exists $f, f' \in \mathcal{F}$ such that $\|(f - f')|_{(x_1,\dots,x_t)}\|_2 \le \epsilon$ while $f(x) - f'(x) > \epsilon$.

Definition 3. ***(Eluder dimension, [Russo and Van Roy, 2014])*** *The Eluder dimension* $\dim_{\mathcal{E}}(\mathcal{F}, \epsilon)$ *of* $\mathcal{F}$ *at scale* ϵ *is the length of the longest sequence* $(x_1, \dots, x_n)$ *in* $\mathcal{X}$ *such that for some* $\epsilon' \ge \epsilon$*, for any* $2 \le t \le n$*,* x_t *is* ϵ'*-independent of* $(x_1, \dots, x_{t-1})$ *given* $\mathcal{F}$*.*

Let $\mathcal{V}$ be the set of optimal value functions under some model in $\mathcal{P}$: $\mathcal{V} = \{V^*_{P'} : P' \in \mathcal{P}\}$. Note that $\mathcal{V} \subset \mathcal{B}(\mathcal{S}, H)$, where $\mathcal{B}(\mathcal{S}, H)$ denotes the set of real-valued measurable functions with domain $\mathcal{S}$ that are bounded by H. We let $\mathcal{X} = \mathcal{S} \times \mathcal{A} \times \mathcal{V}$. Choose $\mathcal{F}$ to be the collection of functions $f : \mathcal{X} \to \mathbb{R}$ as follows:

$$\mathcal{F} = \left\{ f \,\middle|\, \begin{array}{l} \exists P \in \mathcal{P} \text{ s.t. for any } (s,a,v) \in \mathcal{S} \times \mathcal{A} \times \mathcal{V} \\ f(s,a,v) = \int P(ds'|s,a) v(s') \end{array} \right\}. \tag{4.3}$$

Note that $\mathcal{F} \subset \mathcal{B}(\mathcal{X}, H)$. For a norm $\|\cdot\|$ on $\mathcal{F}$ and $\alpha > 0$ let $\mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|)$ denote the $(\alpha, \|\cdot\|)$-covering number of $\mathcal{F}$. That is, this if $m = \mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|)$ then one can find m elements of $\mathcal{F}$ such that any element in $\mathcal{F}$ is at most α away from one of these elements in norm $\|\cdot\|$. Denote by $\|\cdot\|_\infty$ the supremum norm: $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$.

Define the K-episode *pseudo-regret* as

$$R_K = \sum_{k=1}^{K} \left(V^*(s_0^k) - V^{\pi_k}(s_0^k)\right) .$$

Clearly, $R(KH) = \mathbb{E}R_K$ holds for any $K > 0$ where $R(T)$ is the expected regret after T steps of interaction as defined in 2.1. Thus, to study the expected regret, it suffices to study R_K. Our main result is as follows.

Theorem 5. ***(Regret of Algorithm 1)*** *Let Assumption 1 hold and let* $\alpha \in (0, 1)$. *For* $k > 0$ *let* β_k *be*

$$\begin{aligned} \beta_k &= 2H^2 \log\left(\frac{2\mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|_\infty)}{\delta}\right) \\ &+ 2H(kH-1)\alpha\left\{2 + \sqrt{\log\left(\frac{4kH(kH-1)}{\delta}\right)}\right\} . \end{aligned} \tag{4.4}$$

Then, for any $K > 0$, *with probability* $1 - 2\delta$,

$$\begin{aligned} R_K &\leq \alpha + H(d \wedge K(H-1)) + 4\sqrt{d\beta_K K(H-1)} \\ &+ H\sqrt{2K(H-1)\log(1/\delta)}\,, \end{aligned}$$

where $d = \dim_{\mathcal{E}}(\mathcal{F}, \alpha)$ *is the Eluder dimension with* $\mathcal{F}$ *given by* (4.3).

A typical choice of α is $\alpha = 1/(KH)$. In the special case of linear transition model, Theorem 5 implies a worst-case regret bound that depends linearly on the number of parameters.

Corollary 6. ***(Regret of Algorithm 1 for Linearly-Parametrized Transition Model)*** *Let* $P_1, \ldots, P_d$ *be* d *transition models,* $\Theta \subset \mathbb{R}^d$ *a nonempty set with diameter* R *measured in* $\|\cdot\|_1$ *and let* $\mathcal{P} = \{\sum_j \theta_j P_j : \theta \in \Theta\}$. *Then, for any* $0 < \delta < 1$, *with probability at least* $1 - \delta$, *the pseudo-regret* R_K *of Algorithm 1 when it uses the confidence sets given in Theorem 1 satisfies*

$$R_K = \widetilde{O}(d\sqrt{H^3 K \log(1/\delta)})\,.$$

We also provide a lower bound for the regret in our model. The proof is by reduction to a known lower bound and is left to Appendix A.2.

Theorem 7 (Regret Lower Bound)**.** *For any $H \geq 1$ and $d \geq 8$, there exist a state space $\mathcal{S}$ and action set $\mathcal{A}$, a reward function $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$, d transition models $P_1, \ldots, P_d$ and a set $\Theta \subset \mathbb{R}^d$ of diameter of at most one such that for any algorithm there exists $\theta \in \Theta$ such that for sufficiently large number of episodes K, the expected regret of the algorithm on the H-horizon MDP with reward r and transition model $P = \sum_j \theta_j P_j$ is at least $\Omega(H\sqrt{dK})$.*

(Rusmevichientong and Tsitsiklis, 2010) gave a regret lower bound of $\Omega(d\sqrt{T})$ for linearly parameterized bandit with actions on the unit sphere (see also Section 24.2 of Lattimore and Szepesvári, 2018). Our regret upper bound matches this bandit lower bound in d, T. Whether the upper or lower bound is tight (or none of them) remains to be seen. The theorems validate that, in the setting we consider, it is sufficient to use the predicted value functions as regression targets. That for the special case of linear mixture models the lower bound is close to the upper bound appears to suggest that little benefit if any can be derived from fitting the transition model to predict future observations. We conjecture that this is in fact true when considering the worst-case regret. Of course, a conclusion that is concerned with the worst-case regret has no implication for the behavior of the respective methods on particular MDP instances. We note in passing that by appropriately increasing β_k, the regret upper bounds can be extended to the so-called misspecified case when P can be outside of $\mathcal{P}$ (for related results, see, e.g., Jin et al., 2020; Lattimore and Szepesvári, 2019). However, the details of this are left for future work.

Further, our method applies to handle the case where the linearly parameterized transition model is sparse. Suppose that model parameter θ is known to have at most s nonzero entries. In this case, the class of sparse linear models has a much smaller covering number, and the regret would improve and depend on both d,s. Details of this are left for future work.

4.2.1 Regret Bound with Model Misspecification

Next, we consider the case where the model family $\mathcal{P}$ does not exactly realize the true transition model P:

Assumption 2 (Model with misspecification error)**.** *The model family $\mathcal{P}$ ε-approximates P in the sense that there exists $P^* \in \mathcal{P}$ such that*

$$\sup_{s,a} \|P(\cdot|s,a) - P^*(\cdot|s,a)\|_{TV} \leq \varepsilon, \tag{4.5}$$

where $\|\cdot\|_{TV}$ denotes the total variation distance.

This assumption indicates that the true transition model P of the MDP is close to the family $\mathcal{P}$, and ε measures the worst-case deviation. We handle the misspecification error by slightly enlarging the confidence set. This allows us to obtain a regret bound similar to our previous result with an additional linear term that is proportional to the misspecification error and slightly larger constants:

Theorem 8. *Let Assumption 2 hold, $\alpha, \delta \in (0,1)$. We choose β_k be*

$$\begin{aligned}\beta_k &= 8H^2 \log\left(\frac{4\mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|_\infty)}{\delta}\right) \\ &+ 4H(kH-1)\alpha \left\{2 + \sqrt{\log\left(\frac{8kH(kH-1)}{\delta}\right)}\right\} \\ &+ 8H^3 k\varepsilon^2.\end{aligned}$$

Then, for any $K > 0$, with probability $1 - 2\delta$, the K-episode pseudo-regret R_K of Algorithm 1 satisfies

$$\begin{aligned}R_K &\leq \alpha + H(d \wedge K(H-1)) + 4\sqrt{d\beta_K K(H-1)} \\ &+ H\sqrt{2K(H-1)\log(1/\delta)} + H^2 K\varepsilon,\end{aligned}$$

where $d = \dim_{\mathcal{E}}(\mathcal{F}, \alpha)$ with $\mathcal{F}$ given by (4.3).

The proof of this theorem is given in Appendix A.4.

Exploration/ Targets	**Optimism**	**Dithering**
Next states	UC-MatrixRL	EG-Freq
Values	UCRL-VTR	EG-VTR
Mixed	UCRL-Mixed	EG-Mixed

Table 4.1: Legend to the algorithms compared. Note that UC-MatrixRL of (L. F. Yang and Wang, 2019) in the tabular case essentially becomes UCRL of (Jaksch et al., 2010).

4.3 Numerical Implementation and Results

The goal of our experiments is to provide insight into the benefits and/or pitfalls of using value-targets for fitting models, both with and without optimistic planning. We run our experiments in the tabular setting as it is easy to keep aspects of the test environments under control while avoiding approximate computations. Note that tabular environments are a special case of the linear model where $P_j(s'|s,a) = \mathbb{I}(j = f(s,a,s'))$, where $j \in [S^2A]$ and f is a bijection that maps its arguments to the set $[S^2A]$, making $d = S^2A$. The objective is either to minimize mean-squared error of predicting next states (alternatively, maximize log-likelihood of observed data), which leads to frequency based model estimates, or it is to minimize the value targets as proposed in our paper. The other component of the algorithms is whether they implement optimistic planning, or planning with the nominal model and then implementing an ϵ-greedy policy with respect to the estimated model ("dithering"). In the case of optimistic planning, the algorithm that uses mixed targets uses a union bound and takes the smallest value upper confidence bounds amongst the two bounds obtained with the two model-estimation methods. These leads to six algorithms, as shown in Table 4.1. Results for the "mixed" variants are very similar to the variant that uses VTR and can be found in Appendix A.7 In the experiments we use confidence bounds that are specialized to the linear case. For details of these, see Appendix A.6. For ϵ-greedy, we optimize the value of ϵ in each environment to get the best results. This gives ϵ-greedy an "unfair advantage". As we shall see it soon, despite this advantage, ϵ-greedy will not fair particularly well in our experiments.

4.3.1 Measurements

We report the cumulative regret as a function of the number of episodes and the weighted model error to indicate how well the model is learned. The results are obtained from 30 independent runs for the ϵ-greedy algorithms and 10 independent runs for the UC algorithms. The weighted model error reported is as follows. Given the model estimate $\widehat{P}$, we compute

$$E(\widehat{P}) = \sum_{s,a} \sum_{s'} \frac{N(s,a,s')}{N'(s,a)} |\widehat{P}(s' \mid s,a) - P^*(s' \mid s,a)|, \tag{4.6}$$

where N' is the observation-count of the state-action pair (s,a), N is the count of transitioning to s' from (s,a), and P^* is the true dynamics model. The weighting is introduced so that an algorithm that discards a state-action pair is not (unduly) penalized.

4.3.2 Results for RiverSwim

The schematic diagram of the RiverSwim environment is shown in Figure 4.1. RiverSwim consists of S states arranged in a chain. The agent begins on the far left and has the choice of swimming left or right at each state. There is a current that makes swimming left much easier than swimming right. Swimming left with the current always succeeds in moving the agent left, but swimming right against the current sometimes moves the agent right (with a small probability of moving left as well), but more often than not leaves the agent in the current state. Thus smart exploration is a necessity to learn a good policy in this environment. We experiment with small environments with $S = 5$ and set the horizon to $H = 20$. The optimal value of the initial state is 5.6 for our five-state RiverSwim. The initial state is the leftmost state (s_1 in the diagram). The value that we found to work the best for EGRL-VTR is $\epsilon = 0.2$ and the value that we found to work best for EG-Freq is $\epsilon = 0.12$. The results are shown in Figure 4.2. The regret per episode for an algorithm that "does not learn" is expected to be in the same range as the respective optimal

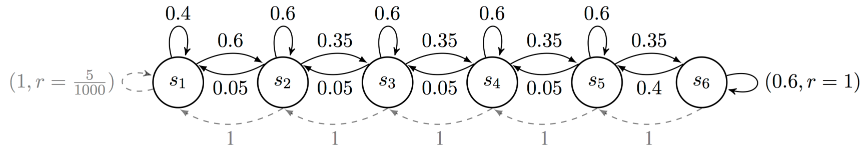

Figure 4.1: The “RiverSwim” environment with 6 states. State s_1 has a small associated reward, state s_6 has a large associated reward. The action whose effect is shown with the dashed arrow deterministically “moves the agent” left. The other action is stochastic, and with relatively high probability moves the agent towards state s_6: This represents swimming “against the current”.

values. Based on this we see that 10^5 episodes is barely sufficient for the algorithms other than UCRL-VTR to learn a good policy. Looking at the model errors we see that EGRL-VTR is doing quite poorly, EG-Freq is also lacking, the others are doing reasonably well. However, this is because EG-Freq visits more uniformly than the other methods the various state-action pairs. The results clearly indicate that *(i)* fitting to the state-value function alone provides enough of a signal for learning as evident by UCRL-VTR obtaining low regret as predicted by our theoretical results, and that *(ii)* optimism is necessary when using VTR to achieve good results, as evident by UCRL-VTR achieving significantly better regret than EGRL-VTR and even in the smaller RiverSwim environment. It is also promising that value-targeted regression with optimistic exploration outperformed optimism based on the “canonical” model estimation procedure. We attribute this to the fact that value-targeted regression will learn a model faster that predicts the optimal values well than the canonical, frequency based approach. That value-targeted regression also learns a model with small weighted error appears to be an accidental feature of this environment. Our next experiments are targeted at further exploring whether VTR can be effective *without* learning a good model.

4.3.3 Results for WideTree

We introduce a novel tabular MDP we call WideTree. The WideTree environment has a fixed horizon $H = 2$ and $S = 11$ states. A visualization of the WideTree environment is shown in Figure 4.3. In WideTree, an agent starts

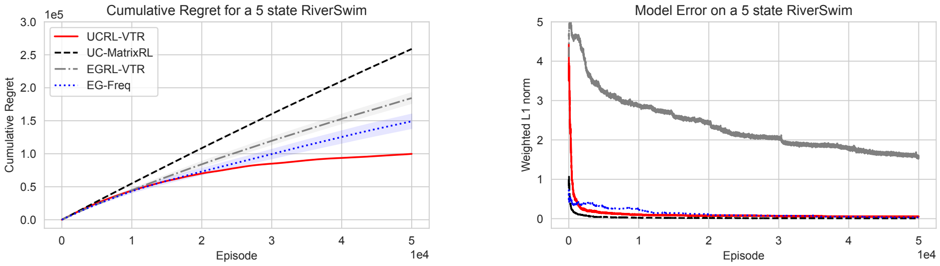

Figure 4.2: The results for the ϵ-greedy algorithms are averaged over thirty runs and error bars are reported for the regret plots.

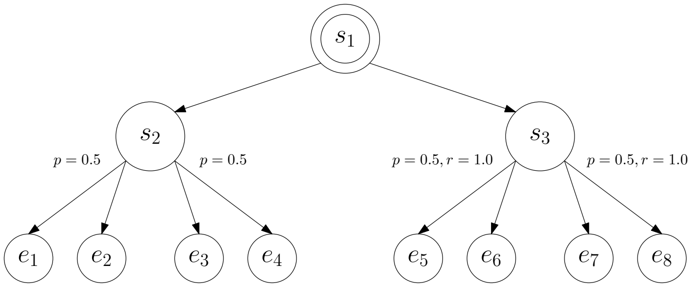

Figure 4.3: An eleven state WideTree MDP. The algorithm starts in the initial state s_1. From the initial state s_1 the algorithm has a choice of either deterministically transitioning to either state s_2 or state s_3. Finally from either state s_2 or state s_3 the algorithm picks one of two possible actions and transitions to one of the terminal states e_i. The choice of the initial action determines the delayed reward the algorithm will observe.

at the initial state s_1. The agent then progresses to one of the many bottom terminal states and collects a reward of either 0 or 1. The only significant action is whether to transition from s_1 to either s_2 or s_3. Note that the model in the second layer is irrelevant for making a good decision: Once in s_3, all actions lead to a reward of one, and once in s_2, all actions lead to a reward of zero. We vary the number of bottom states reachable from states s_2 and s_3 while still maintaining a reward structure but the results here are shown with $S = 11$. We set $\epsilon = 0.1$ in this environment, as this allows the model error of EG-Freq to match that of UC-MatrixRL. The results are shown in Figure 4.4. Both UCRL-VTR and EG-VTR learn equally poor models (their graphs are ‘on the top of each other’). Yet, UCRL-VTR manages to quickly learn a good policy, as attested by its low regret. EG-Freq and EG-VTR perform equally poorly and UC-MatrixRL is even slower as it keeps exploring the environment.

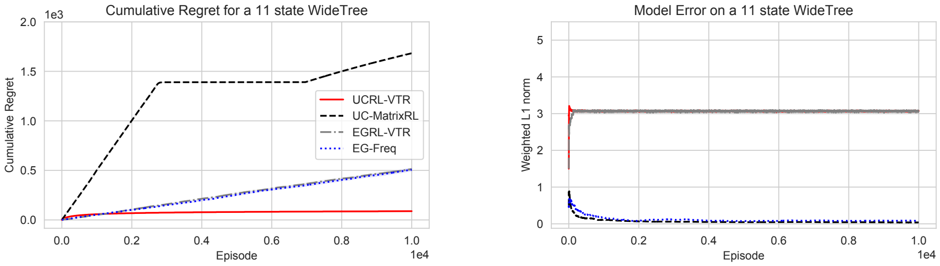

Figure 4.4: The results for the ϵ-greedy algorithms are averaged over thirty runs and error bars are reported for the regret plots.

These experiments clearly illustrate that UCRL-VTR is able to achieve good results without learning a good model – its focus on values makes pays off swiftly in this well-chosen environment.

Chapter 5

Least Squares Value Iteration with Perturbed History Exploration

The work discussed in this chapter is based on the publication (Ishfaq et al., 2021)

5.1 The Algorithm

In this section, we lay out our algorithm (Algorithm 2), an optimistic modification of RLSVI, where the optimism is realized by, what we will call, optimistic reward sampling. To describe our algorithm and facilitate its analysis in Section 5.2, we first define the perturbed least squares regression. We add noises on the regression target and the regularizer to achieve enough randomness in all directions of the regressor.

Definition 4 (Perturbed Least Squares)**.** *Consider a function class $\mathcal{F} : X \to \mathbb{R}$. For an arbitrary dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, a regularizer $R(f) = \sum_{j=1}^D p_j(f)^2$ where $p_j(\cdot)$ are functionals, and positive constant σ, the perturbed dataset and perturbed regularizer are defined as*

$$\widetilde{\mathcal{D}}_\sigma = \{(x_i, y_i + \xi_i)\}_{i=1}^n, \quad \widetilde{R}_\sigma(f) = \sum_{j=1}^D [p_j(f) + \xi'_j]^2,$$

where ξ_i and ξ'_j are i.i.d. zero-mean Gaussian noises with variance σ^2. For a loss function L, the corresponding perturbed least squares regression solution

Algorithm 2 $\mathcal{F}$-LSVI-PHE

```
Set $M$ to be a fixed integer.
For episode $k = 1, 2, \ldots, K$ do
  Receive the initial state $s_1^k$.
  Set $V_{H+1}^k(s) = 0$ for all $s \in \mathcal{S}$.
  For step $h = H, H-1, \ldots, 1$ do
    For $m = 1, 2, \ldots, M$ do
      Sample i.i.d. Gaussian noise $\xi_{h,k}^{\tau,m} \sim \mathcal{N}(0, \sigma_{h,k}^2)$.
      Perturbed dataset: $\widetilde{\mathcal{D}}_h^{k,m} \leftarrow \{(s_h^\tau, a_h^\tau, r_h^\tau + \xi_{h,k}^{\tau,m}$
      $+V_{h+1}^k(s_{h+1}^\tau))\}_{\tau \in [k-1]}$.
      Set $\widetilde{f}_h^{k,m} \leftarrow \mathrm{argmin}_{f \in \mathcal{F}} L(f \,|\, \widetilde{\mathcal{D}}_h^{k,m})^2 + \lambda \widetilde{R}(f)$.
      Set $Q_h^{k,m}(\cdot,\cdot) \leftarrow \widetilde{f}_h^{k,m}(\cdot,\cdot)$.
    Set $Q_h^k(\cdot,\cdot) \leftarrow \min\{\max_{m \in [M]}\{Q_h^{k,m}(\cdot,\cdot)\},$
      $H - h + 1\}$.
    Set $V_h^k(\cdot) \leftarrow \max_{a \in \mathcal{A}} Q_h^k(\cdot, a)$ and
      $\pi_h^k(\cdot) \leftarrow \mathrm{argmax}_{a \in \mathcal{A}} Q_h^k(\cdot, a)$.
  For step $h = 1, 2, \ldots, H$ do
    Take action $a_h^k \leftarrow \mathrm{argmax}_{a \in \mathcal{A}} Q_h^k(s_h^k, a)$.
    Observe reward $r_h^k(s_h^k, a_h^k)$, get next state $s_{h+1}^k$.
```

is

$$\widetilde{f}_\sigma = \mathrm{argmin}_{f \in \mathcal{F}} L(f \,|\, \widetilde{\mathcal{D}}_\sigma)^2 + \lambda \widetilde{R}_\sigma(f).$$

Within each episode $k \in [K]$, at each time-step h, we perturb the dataset by adding zero mean random Gaussian noise to the reward in the replay buffer $\{(s_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau \in [k-1]}$ and the regularizer before we solve the perturbed regularized least-squares regression. At each time step h, we repeat the process for M (to be specified in Section 5.2) times and use the maximum of the regressor as the optimistic estimate of the state-action value function. Concretely, we set $V_{H+1}^k = 0$ and calculate $Q_H^k, Q_{H-1}^k, \ldots, Q_1^k$ iteratively as follows. For each $h \in [H]$ and $m \in [M]$, we solve the following perturbed regression problem,

$$\widetilde{f}_h^{k,m} \leftarrow \mathrm{argmin}_{f \in \mathcal{F}} L(f \,|\, \widetilde{\mathcal{D}}_h^{k,m})^2 + \lambda \widetilde{R}(f). \tag{5.1}$$

We set $Q_h^{k,m}(\cdot,\cdot) = \widetilde{f}_h^{k,m}(\cdot,\cdot)$ and define

$$Q_h^k(\cdot,\cdot) = \min\{\max_{m \in [M]}\{Q_h^{k,m}(\cdot,\cdot)\}, H - h + 1\}. \tag{5.2}$$

We then choose the greedy policy with respect to Q_h^k and collect a trajectory data for the k-th episode. We repeat the procedure until all the K episodes are completed.

5.1.1 LSVI-PHE with Linear Function Class

We now present LSVI-PHE when we consider linear function class (see Algorithm 3). In this case, the following proposition shows that, adding scalar Gaussian noise to the reward is equivalent to perturbing the least-squares estimate using d-dimensional multivariate Gaussian noise.

Proposition 1. *In line 9 of Algorithm 3, conditioned on all the randomness except $\{\epsilon_h^{k,i,j}\}_{(i,j)\in[k-1]\times[M]}$ and $\{\xi_h^{k,j}\}_{j\in[M]}$, the estimated parameter $\widetilde{\theta}_h^{k,j}$ satisfies*

$$\widetilde{\theta}_h^{k,j} - \widehat{\theta}_h^{k,j} \sim N(0, \sigma^2(\Lambda_h^k)^{-1}),$$

where $\widehat{\theta}_h^{k,j} = (\Lambda_h^k)^{-1}(\sum_{\tau=1}^{k-1}[r_h^\tau + V_{h+1}^k(s_{h+1}^\tau)]\phi(s_h^\tau, a_h^\tau))$ is the unperturbed regressor.

Intuitively, adding a zero-mean multivariate Gaussian noise on the parameter $\widehat{\theta}_h^k$ can guarantee that $\widetilde{Q}_h^k$ is optimistic with constant probability. By repeating this procedure multiple times, this constant probability can be amplified to arbitrary high probability.

5.2 Theoretical Analysis

For the analysis we will need the concept of the *eluder dimension*, Definition 3, due to (Russo and Van Roy, 2013). Let $\mathcal{F}$ be a set of real-valued functions with domain $\mathcal{X}$. For $f \in \mathcal{F}, x_1, ..., x_t \in \mathcal{X}$, introduce the notation $f|_{(x_1,...,x_t)} = (f(x_1), ..., f(x_t))$. We say that $x \in \mathcal{X}$ is ϵ-independent of $x_1, ..., x_t \in \mathcal{X}$ given $\mathcal{F}$ if there exists $f, f' \in \mathcal{F}$ such that $||(f-f')|_{(x_1,...,x_t)}||_2 \leq \epsilon$ while $f(x)-f'(x) > \epsilon$. For a more detailed introduction of eluder dimension, readers can refer to Chapter 4 or see (Osband and Van Roy, 2014; Russo and Van Roy, 2013; Wang et al., 2020).

Algorithm 3 LSVI-PHE with Linear function class

```
Set $M$ to be a fixed integer.
**For** episode $k = 1, 2, \dots, K$ **do**
  Receive the initial state $s_1^k$.
  **For** step $h = H, H-1, \dots, 1$ **do**
    $\Lambda_h^k \leftarrow \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top + \lambda I$.
    Sample i.i.d. $\{\epsilon_h^{k,\tau,j}\}_{(\tau,j)\in[k-1]\times[M]} \sim \mathcal{N}(0, \sigma^2)$.
    Sample i.i.d. $\{\xi_h^{k,j}\}_{j\in[M]} \sim \mathcal{N}(0, \sigma^2 \lambda I_d)$.
    $\rho_h^{k,j} \leftarrow \sum_{\tau=1}^{k-1} \Big( [r_h^\tau + V_{h+1}^k(s_{h+1}^\tau) + \epsilon_h^{k,\tau,j}]\phi(s_h^\tau, a_h^\tau) \Big)$.
    $\widetilde{\theta}_h^{k,j} \leftarrow (\Lambda_h^k)^{-1}(\rho_h^k + \xi_h^{k,j})$.
    $\widetilde{Q}_h^{k,j}(\cdot,\cdot) \leftarrow \phi(\cdot,\cdot)^\top \widetilde{\theta}_h^{k,j}$ for $j \in [M]$.
    $Q_h^k(\cdot,\cdot) \leftarrow \min\{\max_{j\in[M]} \widetilde{Q}_h^{k,j}(\cdot,\cdot), H-h+1\}^+$
    $V_h^k(\cdot) \leftarrow \max_{a\in\mathcal{A}} Q_h^k(\cdot, a)$.
  **For** step $h = 1, 2, \dots, H$ **do**
    Take action $a_h^k \leftarrow \operatorname{argmax}_{a\in\mathcal{A}} Q_h^k(s_h^k, a)$.
    Observe reward $r_h^k(s_h^k, a_h^k)$, get next state $s_{h+1}^k$.
```

5.2.1 Assumptions for General Function Approximation

For our general function approximation analysis, we make a few assumptions first. To emphasize the generality of our assumptions, in Section 5.2.1, we show that our assumptions are satisfied by linear function class.

Our algorithm (Algorithm 2) receives a function class $\mathcal{F} \subseteq \{f : \mathcal{S} \times \mathcal{A} \to [0, H]\}$ as input and furthermore, similar to Ayoub et al., 2020; Wang et al., 2020, we assume that for any $V : \mathcal{S} \to [0, H]$, upon applying the Bellman backup operator, the output function lies in the function class $\mathcal{F}$. Concretely, we have the following assumption.

Assumption 3. *For any $V : \mathcal{S} \to [0, H]$ and for any $h \in [H]$, $r_h + P_h V \in \mathcal{F}$, i.e. there exists a function $f_V \in \mathcal{F}$ such that for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ it satisfies*

$$f_V(s, a) = r_h(s, a) + P_h V(s, a). \tag{5.3}$$

We emphasize that many standard assumptions in the RL theory literature such as tabular MDPs (Jaksch et al., 2010; Jin et al., 2018) and Linear MDPs (Jin et al., 2020; L. Yang and Wang, 2019) are special cases of Assumption 3.

In the appendix, we consider a misspecified setting and show that even when (5.3) holds approximately, Algorithm 2 achieves provable regret bounds.

We further assume that our function class has bounded covering number.

Assumption 4. *For any $\varepsilon > 0$, there exists an ε-cover $\mathcal{C}(\mathcal{F}, \varepsilon)$ with bounded covering number $\mathcal{N}(\mathcal{F}, \varepsilon)$.*

Next we define anti-concentration width, which is a function of the function class $\mathcal{F}$, dataset $\mathcal{D}$ and noise variance σ^2.

Definition 5 (Anti-concentration Width Function)**.** *For a loss function $L(\cdot \mid \cdot)$ and dataset $\mathcal{D}$, let $\widehat{f} = \operatorname{argmin}_{f \in \mathcal{F}} L(f \mid \mathcal{D})^2 + \lambda R(f)$ be the regularized least squares solution and $\widetilde{f}_\sigma = \operatorname{argmin}_{f \in \mathcal{F}} L(f \mid \widetilde{\mathcal{D}}_\sigma)^2 + \lambda \widetilde{R}_\sigma(f)$ be the perturbed regularized least-squares solution. For a fixed $v \in (0, 1)$, let $g_\sigma : X \to \mathbb{R}$ be a function such that for any input x:*

$$g_\sigma(x) = \sup_{g \in \mathbb{R}} \mathbb{P}\left(\widetilde{f}_\sigma(x) \geq \widehat{f}(x) + g\right) \geq v.$$

We call $g_\sigma(\cdot)$ the anti-concentration width function.

In plain English, $g_\sigma(\cdot)$ is the largest value some $g \in \mathbb{R}$ can take such that the probability that $\widetilde{f}_\sigma$ is greater than $\widehat{f} + g$ is at least v.

We assume that for a concentrated function class, there exists a σ such that the anti-concentration width is larger than the function class width.

Assumption 5 (Anti-concentration)**.** *Given the input $X = \{x_i\}_{i=1}^n$ of dataset $\mathcal{D}$ and some arbitrary positive constant β, we define a function class $\mathcal{F}_{X,\beta} = \{f : \|f - \widehat{f}\|_X^2 + \lambda R(f - \widehat{f}) \leq \beta\}$. We assume that there exists a σ such that*

$$g_{\sigma'}(x) \geq w(\mathcal{F}_{X,\beta}, x),$$

for all inputs x and $\sigma' \geq \sigma$.

This assumption guarantees that the randomized perturbation over the regression target has large enough probability of being optimistic. This assumption is satisfied by the linear function class. For more details, see Section 5.2.1.

Assumption 6 (Regularization)**.** *We assume that our regularizer $R(\cdot)$ has several basic properties.*

- *$R(f) + R(f') \geq cR(f + f')$ for some positive constant $c > 0$, for all $f, f' \in \mathcal{F}$.*
- *$R(f) = R(-f) \geq 0$, for all $f, f' \in \mathcal{F}$.*
- *For any $V : \mathcal{S} \to [0, H]$, $R(r + PV) \leq B$ for some constant $B \in \mathbb{R}$.*

Here, the first property is nothing but a variation of triangle inequality. The second property is a symmetry property which is natural for norms. Both these properties are satisfied by commonly used regularizers such as ℓ_0, ℓ_1 or ℓ_2 norms. The last property is a boundedness assumption. For the case of ℓ_0 norm B takes the value of the dimension of the space. Moreover, along with the most commonly used (weighted) ℓ_2 regularizer, many other regularizers also satisfy this property. Our final assumption is regarding the boundedness of the Eluder dimension of the function class.

Assumption 7 (Bounded Function Class)**.** *For any $V : \mathcal{S} \to [0, H]$ and any $\mathcal{Z} \in (\mathcal{S} \times \mathcal{A})^{\mathbb{N}}$, let $\mathcal{F}'$ be a subset of function class $\mathcal{F}$, consisting of all $f \in \mathcal{F}$ such that*

$$\|f - v\|_{\mathcal{Z}}^2 + \lambda R(f - v) \leq \beta,$$

where $v = r + PV$. We assume that $\mathcal{F}'$ has bounded Eluder dimension.

Note that in (Wang et al., 2020), they assume that the Eluder dimension of the whole function class $\mathcal{F}$ is bounded. In contrast, ours is a weaker assumption since we only assume a subset $\mathcal{F}'$ to have a bounded Eluder dimension.

Linear Function Class

First, we recall the standard linear MDP definition which was introduced in (Jin et al., 2020; L. Yang and Wang, 2019).

Definition 6 (Linear MDP, Jin et al., 2020; L. Yang and Wang, 2019)**.** *We consider a linear Markov decision process,* $\mathrm{MDP}(\mathcal{S}, \mathcal{A}, H, P, r)$ *with a feature* *map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, where for any $(h, k) \in [H] \times [K]$, there exist d unknown (signed) measures $\mu_h = (\mu_h^{(1)}, \cdots, \mu_h^{(d)})$ over $\mathcal{S}$ and an unknown vector $w_h \in \mathbb{R}^d$, such that for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, the following holds:*

$$P_h(s'|s,a) = \langle \phi(s,a), \mu_h(s') \rangle, \quad r_h(s,a) = \langle \phi(s,a), w_h \rangle.$$

Without loss of generality, we assume, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\|\phi(s,a)\| \leq 1$, and for all $h \in [H]$, $\|w_h\| \leq \sqrt{d}$ and $\|\mu_h(\mathcal{S})\| \leq \sqrt{d}$.

Consider a fixed episode k and step h. We define $\mathcal{F} = \{f_\theta : f_\theta(s,a) = \phi(s,a)^\top \theta\}$ where $\theta \in \mathbb{R}^d$, $\mathcal{D} = \{(s_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau \in [k-1]}$, and $R(f_\theta) = \|\theta\|^2 = \sum_{j=1}^d p_j(f_\theta)^2$ where $p_j(f_\theta) = e_j^\top \theta$ with e_j being the j-th standard basis vector. It is well known that linear function class satisfies Assumption 3 in linear MDP (Jin et al., 2020; L. Yang and Wang, 2019). We set $\widehat{f} = \operatorname{argmin}_{f \in \mathcal{F}} L(f \,|\, \mathcal{D})^2 + \lambda R(f)$ to be $f_{\widehat{\theta}}$. Then we have

$$\begin{aligned}\widehat{\theta} &= \operatorname{argmin}_\theta \sum_{\tau=1}^{k-1} (\phi(s_h^\tau, a_h^\tau)^\top \theta - r_h^\tau)^2 + \lambda \|\theta\|^2 \\ &= (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} r_h^\tau \phi(s_h^\tau, a_h^\tau),\end{aligned}$$

where $\Lambda_h^k = \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau) \phi(s_h^\tau, a_h^\tau)^\top + \lambda I$. Similarly we set $f_{\widetilde{\theta}} = \widetilde{f}_\sigma = \operatorname{argmin}_{f \in \mathcal{F}} L(f \,|\, \widetilde{\mathcal{D}}_\sigma)^2 + \lambda \widetilde{R}_\sigma(f)$. Then we have

$$\begin{aligned}\widetilde{\theta} &= (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} (r_h^\tau + \xi_\tau) \phi(s_h^\tau, a_h^\tau) + (\Lambda_h^k)^{-1} \sum_{j=1}^{d} \xi_j' e_j \\ &\sim \mathcal{N}(\widehat{\theta}, \sigma^2 (\Lambda_h^k)^{-1}).\end{aligned}$$

For Definition 5, we set $v = \Phi(-1)$. Using the anti-concentration property of Gaussian distribution, it is straightforward to show that for any $(s,a) \in \mathcal{S} \times \mathcal{A}$:

$$\mathbb{P}\left(f_{\widetilde{\theta}}(s,a) \geq f_{\widehat{\theta}}(s,a) + \sigma \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} \right) = v.$$

So we have $g_\sigma(s,a) \geq \sigma \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}$ from Definition 5. For Assumption 5, the function class $\mathcal{F}_{\mathcal{D},\beta} = \{f : L(f - \widehat{f} \,|\, \mathcal{D})^2 + \lambda R(f - \widehat{f}) \leq \beta\}$ is equivalent to $\Theta_{\mathcal{D},\beta} = \{\theta : (\theta - \widehat{\theta})^\top \Lambda_h^k (\theta - \widehat{\theta}) \leq \beta\}$. So the width on the state-action pair (s, a) is $2\sqrt{\beta}\|\phi(s, a)\|_{(\Lambda_h^k)^{-1}}$. If we set $\sigma = 2\sqrt{\beta}$, we have

$$g_\sigma(s, a) \geq w(\mathcal{F}_{\mathcal{D},\beta}, s, a).$$

For Assumption 6, as $R(f_\theta) = \|\theta\|^2$ is a ℓ_2 norm function, the first two properties are direct to show with constant $c = 1/2$. For the third property, we have that

$$g(s, a) = r(s, a) + P(s, a)V = \phi(s, a)(w + \sum_{s'} V(s')\mu(s')).$$

So we have $g = g_\theta$ where $\theta = w + \sum_{s'} V(s')\mu(s')$ and $\|\theta\|^2 \leq 2Hd$. For Assumption 7, we set $\theta_f : f = f_{\theta_f}$, $\theta_v : v = f_{\theta_v}$ and $\Theta_{\mathcal{F}'} = \{\theta : f_\theta \in \mathcal{F}'\}$ to be the parameterization. From Assumption 6, we have $\|\theta_v\|^2 \leq 2Hd$. In addition, we have $\lambda R(f - v) = \lambda\|\theta_f - \theta_v\|^2 \leq \beta$. Then we have

$$\begin{aligned} \Theta_{\mathcal{F}'} &\subseteq \{\theta_f : \|\theta_f - \theta_v\|^2 \leq \beta/\lambda, \|\theta_v\|^2 \leq 2Hd\} \\ &= \{\theta_f : \|\theta_f\|^2 \leq 2\beta/\lambda + 4Hd\}. \end{aligned}$$

As shown in (Russo and Van Roy, 2013), this $\mathcal{F}'$ has eluder dimension $\widetilde{O}(d)$.

5.2.2 Regret bound for General Function Approximation

First, we specify our choice of the noise variance σ^2 in the algorithm. We prove certain concentration properties of the regularized regressor $\widehat{f}_h^k$ so that the condition in Assumption 5 holds. Thus we can choose an appropriate σ such that the Assumption 5 is satisfied. A more detailed description is provided in the appendix. Our first lemma is about the concentration of the regressor. A similar argument appears in (Wang et al., 2020) but their result does not include regularization, which is essential in our randomized algorithm to ensure exploration in all directions.

Lemma 9 (Informal Lemma on Concentration). *Under Assumptions 3, 4, 5, 6, and 7, let* $\mathcal{F}_h^{k,m} = \{f \in \mathcal{F} | \|f - \widetilde{f}_h^{k,m}\|^2_{\mathcal{Z}_h^k} + \lambda R(f - \widetilde{f}_h^{k,m}) \leq \beta(\mathcal{F}, \delta)\}$, *where* $\mathcal{Z}_h^k = \{(s_h^\tau, a_h^\tau)\}_{\tau \in [k-1]}$, *and*

$$\beta(\mathcal{F}, \delta) = \widetilde{O}\left((H + \sigma)^2 \log \mathcal{N}(\mathcal{F}, 1/T)\right).$$

With high probability, for all $(k, h, m) \in [K] \times [H] \times [M]$, we have

$$r_h(\cdot, \cdot) + P_h V_{h+1}^k(\cdot, \cdot) \in \mathcal{F}_h^{k,m}.$$

This lemma shows that the perturbed regularized regression still enjoys concentration. Our next lemma shows that LSVI-PHE is optimistic with high probability.

Lemma 10 (Informal Lemma on Optimism)**.** *Let*

$$M = \ln\left(\frac{T|\mathcal{S}||\mathcal{A}|}{\delta}\right) / \ln\left(\frac{1}{1-v}\right).$$

With probability at least $1-\delta$, for all $(s, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, we have

$$Q_h^*(s, a) \leq Q_h^k(s, a).$$

With optimism, the regret is known to be bounded by the sum of confidence width (Wang et al., 2020). As Assumption 7 assumes that all the confidence region is in a bounded function class in the measure of eluder dimension, we can adapt proof techniques from (Wang et al., 2020) and prove our final result.

Theorem 11 (Informal Theorem)**.** *Under Assumptions 3, 4, 5, 6, and 7, with high probability, Algorithm 2 achieves a regret bound of*

$$\text{Regret}(K) \leq \widetilde{O}\left(\sqrt{\dim_{\mathcal{E}}(\mathcal{F}, 1/T)\beta(\mathcal{F}, \delta)HT}\right),$$

where

$$\beta(\mathcal{F}, \delta) = \widetilde{O}\left((H+\sigma)^2 \log \mathcal{N}\left(\mathcal{F}, 1/T\right)\right).$$

The theorem shows that our algorithm enjoys sublinear regret and have polynomial dependence on the horizon H, noise variance σ^2 and eluder dimension $\dim_{\mathcal{E}}(\mathcal{F}, 1/T)$, and have logarithmic dependence on the covering number of the function class $\mathcal{N}\left(\mathcal{F}, 1/T\right)$.

5.2.3 Regret bound for linear function class

Now we present the regret bound for Algorithm 3 under the assumption of linear MDP setting. In the appendix, we provide a simple yet elegant proof of the regret bound.

Theorem 12. *Let $M = d\log(\delta/9)/\log\Phi(1)$, $\sigma = \widetilde{O}(H\sqrt{d})$, and $\delta \in (0,1]$. Under linear MDP assumption from Definition 6, the regret of Algorithm 3 satisfies*

$$\mathit{Regret}(T) \leq \widetilde{\mathcal{O}}(d^{3/2}H^{3/2}\sqrt{T}),$$

with probability at least $1-\delta$.

Remark 1. *Under linear MDP assumption, this regret bound is at the same order as the LSVI-UCB algorithm from (Jin et al., 2020) and $\sqrt{dH}$ better than the state-of-the-art TS-type algorithm (Zanette, Brandfonbrener, et al., 2020). The only work that enjoys a $\sqrt{d}$ better regret is (Zanette, Lazaric, Kochenderfer, et al., 2020), which requires solving an intractable optimization problem.*

Remark 2. *Along with being a competitive algorithm in statistical efficiency, we want to emphasize that our algorithm has good computational efficiency. LSVI-PHE with linear function class only involves linear programming to find the greedy policy while LSVI-UCB (Jin et al., 2020) requires solving a quadratic programming. The optimization problem in OPT-RLSVI (Zanette, Brandfonbrener, et al., 2020) is hard too because the Q-function there is a piecewise continuous function and in one piece, it includes the product of the square root of a quadratic term and a linear term.*

5.3 Numerical Implementation and Results

We run our experiments on RiverSwim (Strehl and Littman, 2008), DeepSea (Osband, Van Roy, and Wen, 2016) and sparse MountainCar (Brockman et al., 2016) environments as these are considered to be hard exploration problems where ε-greedy is known to have poor performance. For both RiverSwim and DeepSea experiments, we make use of linear features. The objective here is to compare an exploration method that randomizes the targets in the history (LSVI-PHE) with an exploration method that computes upper confidence bounds given the history (LSVI-UCB) (Cai et al., 2019; Jin et al., 2020). For the continous control MountainCar environment, we use neural-network as function approximator to implement LSVI-PHE. The objective here is to compare deep RL variant of LSVI-PHE against other popular deep RL algorithms specifically designed to tackle exploration task.

5.3.1 Measurements

We plot the per episode return of each algorithm to benchmark their performance. As the agent begins to act optimally the per episode return begins to converge to the optimal, or baseline, return. The per episodes returns are the sum of all the rewards obtained in an episode. We also report the performance of LSVI-PHE when σ^2 is fixed and M varies.

5.3.2 Results for RiverSwim

A diagram of the RiverSwim environment is shown in the Appendix. RiverSwim consists of $\mathcal{S}$ states lined up in a chain. The agent begins in the leftmost state s_1 and has the choice of swimming to the left or to the right at each state. The agent's goal is to maximize its return by trying to reach the rightmost state which has the highest reward. Swimming to the left, with the current, transitions the agent to the left deterministically. Swimming to the right, against the current, stochastically transitions the agent and has relatively high probability of moving right toward the goal state. However, because the current is strong there is a high chance the agent will stay in the current state and a low chance the agent will get swept up in the current and transition to the left. Thus, smart exploration is required to learn the optimal policy in this environment. We experiment with the variant of RiverSwim where $\mathcal{S} = 12$ and $H = 40$. For this experiment, we swept over the exploration parameters in both LSVI-UCB (Jin et al., 2020) and LSVI-PHE and report the best performing run on a 12 state RiverSwim. LSVI-UCB computes confidence widths of the following form $\beta\|\phi(s,a)\|_{\Sigma^{-1}}$ where $\phi(s,a) \in \mathbb{R}^d$ are the features for a given state-action pair and $\Sigma \in \mathbb{R}^{d\times d}$ is the empirical covariance matrix. We sweep over β for LSVI-UCB and σ^2 for LSVI-PHE, where M is chosen according to our theory (Theorem 12). We sweep over these parameters to speed up learning as choosing the theoretically optimal choices for β and σ^2 often leads

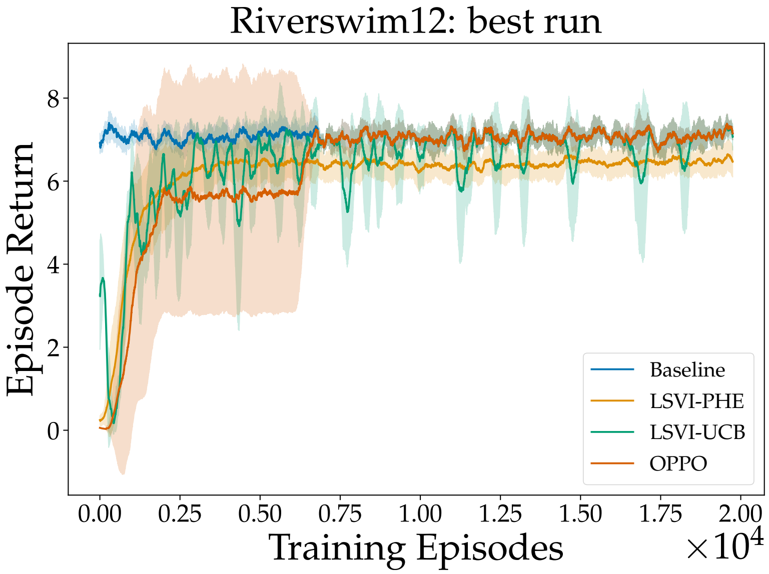

Figure 5.1: The results are averaged over 10 independent runs and error bars are reported for the regret plots. For this plot, $\beta = 5.0$ for LSVI-UCB and $\sigma^2 = 2 \times 10^{-1}$ for LSVI-PHE.

to a more conservative exploration policy which is slow to learn. As shown in Figure 5.1, the best performing LSVI-PHE achieves similar performance to the best performing LSVI-UCB on the 12 state RiverSwim environment.

5.3.3 Results for DeepSea

DeepSea (Osband, Van Roy, and Wen, 2016) consists of $\mathcal{S} = N \times N$ states arranged in a grid, where N is the depth of the sea. The agent begins at the top leftmost state in the grid s_1 and has the choice of moving down and left or down and right at each state. Once the agent reaches the bottom of the sea it transitions back to state s_1. The agent's goal is to maximize its return by reaching the bottom right most state. The agent gets a small negative reward for transitioning to the right while no reward is given if the agent transitions to the left. Thus, smart exploration is required; otherwise the agent will rarely go right the necessary amount of time to reach the goal state. We run our experiments on a 10×10 DeepSea environment. As shown in Figure 5.2, the best performing LSVI-PHE achieves similar performance to the best performing LSVI-UCB on DeepSea. We also vary M given a fixed

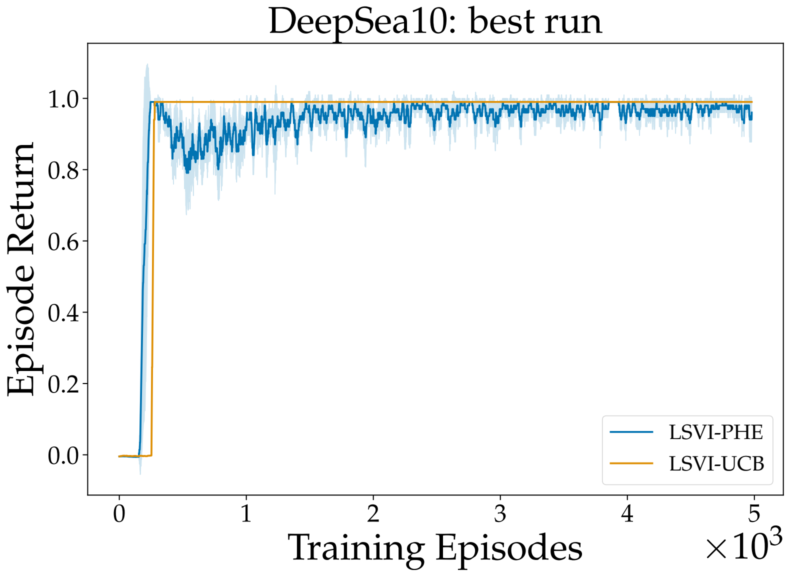

Figure 5.2: The results are averaged over 5 independent runs and error bars are reported for the return per episode plots. For this plot, $\beta = 5 \times 10^{-3}$ for LSVI-UCB and $\sigma^2 = 5 \times 10^{-5}$ for LSVI-PHE.

$\sigma^2 = 5 \times 10^{-4}$. As shown in Figure 5.3, as we increase M, the performance of LSVI-PHE increases.

These experiments on hard exploration problems highlight that we are able to simulate optimistic exploration, as in UCB, by perturbing the targets multiple times and taking the max over the perturbations to boost the probability of an optimistic estimate. If we are willing to sweep over M, the number of times we perturb the history, and σ^2, we can then get a faster algorithm that still performs well in practice. If we let $M = 1$ and $\sigma^2 = 1$ then LSVI-PHE reduces to RLSVI and we would get the same performance as in (Osband, Van Roy, and Wen, 2016).

5.3.4 Results for MountainCar

We further evaluated LSVI-PHE on a continuous control task which requires exploration: sparse reward variant of continuous control MountainCar from OpenAI Gym (Brockman et al., 2016). This environment consists of a 2-dimensional continuous state space and a 1-dimensional continuous action

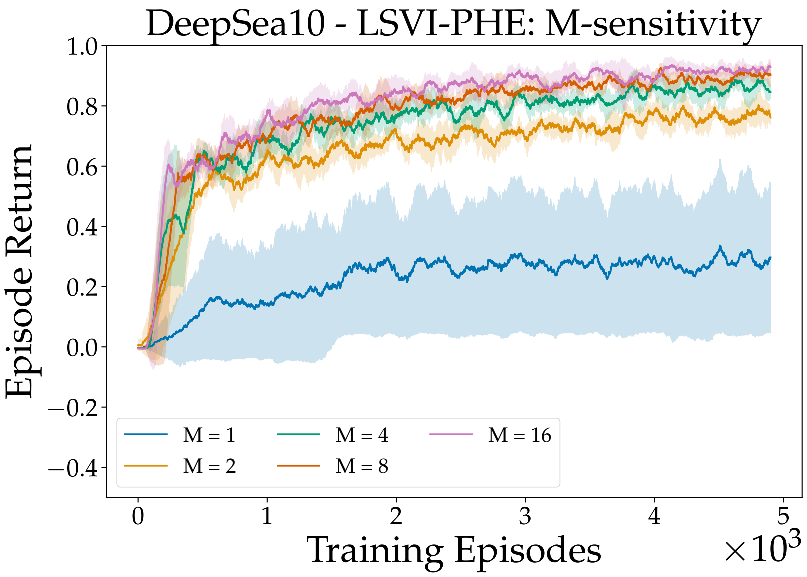

Figure 5.3: The results are averaged over 5 runs and error bars are reported for the return per episode plots. For this plot we fix $\sigma^2 = 5 \times 10^{-4}$.

space $[-1, 1]$. The agent only receives a reward of $+1$ if it reaches the top of the hill and everywhere else it receives a reward of 0. We set the length of the horizon to be 1000 and discount factor $\gamma = 0.99$.

For this setting, we compare four algorithms: LSVI-PHE, DQN with epsilon-greedy exploration, Noisy-Net DQN (Fortunato et al., 2017) and Bootstrapped DQN (Osband, Blundell, et al., 2016). Our experiments are based on the baseline implementations of (Lan, 2019). As neural network, we used a multi-layer perceptron with hidden layers fixed to $[32, 32]$. The size of the replay buffer was $10,000$. The weights of neural networks were optimized by Adam (Kingma and Ba, 2014) with gradient clip 5. We used a batch size of 32. The target network was updated every 100 steps. The best learning rate was chosen from $[10^{-3}, 5 \times 10^{-4}, 10^{-4}]$. For LSVI-PHE, we set $M = 8$ and we chose the best value of σ from $[10^{-4}, 10^{-3}, 10^{-2}]$. Results are shown in Figure 5.4.

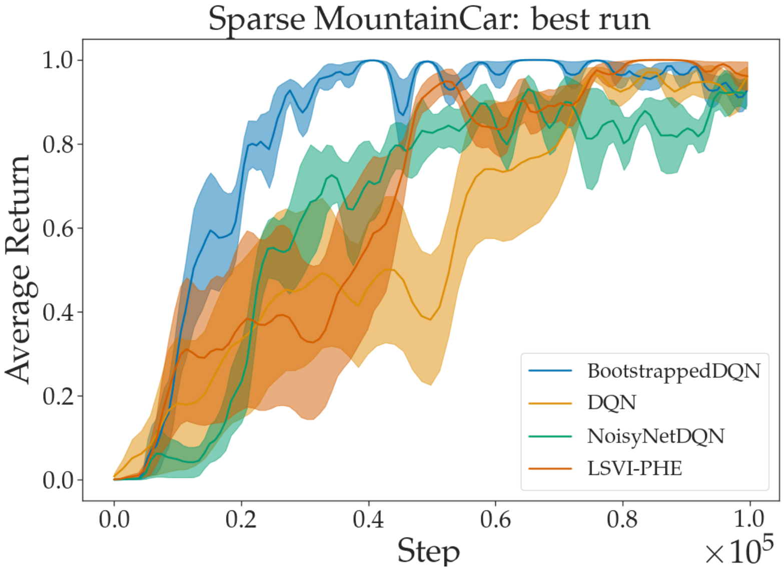

Figure 5.4: : Comparison of four algorithms on sparse MountainCar. The results are averaged over 5 independent runs and error bars are reported for the return per episode plots.

Chapter 6

Related Works

A number of prior efforts have established efficient RL methods with provable regret bounds. For tabular H-horizon MDP with S states and A actions, there have been results on model-based methods (e.g., S. Agrawal and Jia, 2017; Azar et al., 2017; Dann et al., 2017; Dann et al., 2018; Jaksch et al., 2010; Kakade et al., 2018; Osband et al., 2014), and on model-free methods (e.g., Jin et al., 2018; Osband et al., 2017; Zhang et al., 2020). Both model-based and model-free methods are known to achieve a regret of $\widetilde{\mathcal{O}}(\sqrt{H^2SAT})$, where $\widetilde{O}(\cdot)$ hides log factors, T denotes the total number of timesteps and the bound applies to the non-stationary setting (when the transition models are not shared across stages). Moreover, apart from logarithmic factors, this bound is known to be unimprovable (Jaksch et al., 2010; Jin et al., 2018; Osband et al., 2017; Zhang et al., 2020). and the best regret achieved by model-free algorithms is asymptotic $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$, where T denotes the number of time steps and $\widetilde{O}(\cdot)$ hides log factors. (Jaksch et al., 2010) established a worst-case regret lower bound of $\Omega(\sqrt{HSAT})$. There have been significant theoretical and empirical advances on RL with function approximation, including but not limited to(Baird, 1995; Bradtke and Barto, 1996; Mnih et al., 2013; Mnih et al., 2015; Parr et al., 2008; Silver et al., 2017; Tsitsiklis and Van Roy, 1997; L. Yang and Wang, 2019). Under the assumption that the optimal action-value function is captured by linear features, (Zanette et al., 2019) considers the case when the features are "extrapolation friendly" and a simulation oracle is available, (Wen and Van Roy, 2013, 2017) tackle problems where the transition model is deterministic, (Du et al., 2019) deals with a relaxation of the deterministic case when the transition model has low variance. (L. Yang and Wang, 2019) considers the case of linear factor models, while Lattimore and Szepesvári, 2019 considers the case when all the action-value functions of all deterministic policies are well-approximated using a linear function approximator. These latter works handle problems when the algorithm has access to a simulation oracle of the MDP. As for regret minimization in RL using linear function approximation, (L. F. Yang and Wang, 2019) assumed the transition model admits a matrix embedding of the form $P(s'|s,a) = \phi(s,a)^\top M \psi(s')$, and proposed a model-based MatrixRL method with regret bounds $\widetilde{\mathcal{O}}(H^2 d\sqrt{T})$ with stronger assumptions and $\widetilde{\mathcal{O}}(H^2 d^2 \sqrt{T})$ in general, where d is the dimension of state representation $\phi(s,a)$. (Jin et al., 2020) studied the setting of linear-factor MDP and constructed a model-free least-squares action-value iteration algorithm, which was proved to achieve the regret bound $\widetilde{\mathcal{O}}(\sqrt{H^3 d^3 T})$. (Modi et al., 2019) considered a related setting where the transition model is an ensemble involving state-action-dependent features and basis models and proved a sample complexity $\frac{d^3 K^2 H^2}{\epsilon^2}$ where d is the feature dimension, K is the number of basis models and $d \cdot K$ is their total model complexity. As for RL with a general model class, in their seminal work, (Osband and Van Roy, 2014) provided a general posterior sampling RL method that works for any given classes of reward and transition functions. It established a Bayesian regret upper bound $O(\sqrt{d_K d_E T})$, where d_K and d_E are the Kolmogorov and the Eluder dimensions of the model class. In the case of linearly parametrized transition model, this Bayesian regret becomes $O(d\sqrt{T})$, and our worst-case regret result matches with the Bayesian one. (Abbasi-Yadkori and Szepesvári, 2015; Theocharous et al., 2017) also considered the Bayesian regret and, in particular, (Abbasi-Yadkori and Szepesvári, 2015) considered a smooth parameterization. To the authors' best knowledge, there are no prior works addressing the problem of designing low-regret algorithms for MDPs with linearly or non-linearly parameterized transition models. Model based PAC RL algorithms have been studied by (Sun et al., 2019), who essentially adopt the value-aware loss of (Farahmand et al., 2017), who considered this loss in the batch setting. (Farahmand, 2018) refines the work of (Farahmand et al., 2017) by changing the algorithm to be similar to what we use here: In every iteration of fitted Q-iteration, first a model is obtained by minimizing the value prediction loss measured with the last value function, after which this model is used to obtain the next action-value function. The main result bounds the suboptimality of the policy that is greedy with respect to the last action-value function. A preliminary version of (Ayoub et al., 2020) was presented at L4DC.

Thompson Sampling (Thompson, 1933) was proposed almost a century ago and rediscovered several times. (M. Strens, 2000) was the first work to apply TS to RL. (Osband et al., 2013) provides a Bayesian regret bound and (S. Agrawal et al., 2016; Ouyang et al., 2017b) provide worst case regret bounds for TS.

Randomized least-squares value iteration (RLSVI), proposed in (Osband et al., 2019), uses random perturbations to approximate the posterior. Recently, several works focused on the theoretical analysis of RLSVI (P. Agrawal et al., 2020; Russo, 2019; Zanette, Lazaric, Kochenderfer, et al., 2020). (Russo, 2019) provides the first worst-case regret $\widetilde{O}(H^{5/2}S^{3/2}\sqrt{AT})$ for tabular MDP and (P. Agrawal et al., 2020) improves it to $\widetilde{O}(H^2S\sqrt{AT})$ by allowing for a warm-up phase before randomizing the history. (Zanette, Brandfonbrener, et al., 2020) proves $\widetilde{O}(H^2d^2\sqrt{T})$ regret bound for linear MDP. However, (P. Agrawal et al., 2020; Zanette, Brandfonbrener, et al., 2020) both need to compute the confidence width as a warm-up stage, which is complicated and computationally costly.

Chapter 7

Conclusions and Future Directions

In this thesis, we considered online learning in episodic MDPs and proposed an optimistic model-based and model free reinforcement learning methods. The UCRL-VTR algorithm has the unique characteristic to evaluate and select models based on their ability to predict value functions that the algorithm constructs during learning. The regret of the algorithm was shown to be bounded by a quantity that relates to the richness of the model class through the Eluder dimension and the metric entropy of an appropriately constructed function space. For the case of linear mixture models, the regret bound simplifies to $\widetilde{O}(d\sqrt{H^3T})$ where d is the number of model parameters, H is the horizon, and T is the total number of interaction steps. Our experiments for the UCRL-VTR confirmed that the value-targeted regression objective is not only theoretically sound, but also yields a competitive method which allows task-focused model-tuning: In a carefully chosen environment we demonstrated that the algorithm achieves low regret despite that it ignores modeling a major part of the environment.

We also propose an algorithm, LSVI-PHE, that guarantees optimism by adding judiciously chosen random noise to the rewards and then regressing on this perturbed rewards. We then prove the theoretical guarantees of LSVI-PHE and through experiments also demonstrate that it performs competitively against similar algorithms. In the case of the linear MDP, the regret bound of LSVI-PHE simplifies to $\widetilde{O}(d^{3/2}H^{3/2}\sqrt{T})$ where d is the action-value function parameters, H is the horizon, and T is the total number of interaction steps. We then demonstrate the practicallity of LSVI-PHE by using it to solve the sparse mountain car environment with tile coded features. By boosting the level of optimism, LSVI-PHE solves a problem instance that RLSVI cannot. Thus highlighting the need for algorithms with adaptable levels of optimism.

From these works we hope to be able to construct a computationally tractable optimism planner by using optimistic sampling to perturb the rewards in the hopes of outputting an optimistic value-targeted model. While there has been significant progress in designing a computationally tractable optimistic planner for the linear quadratic regulartor (Abeille and Lazaric, 2020), designing one for RL is still an open problem (Lattimore and Szepesvári, 2018, Chapter 38). Another line of work is studying whether Assumption 5 holds with non-linear function approximation. Currently, this assumption is hard to show even in generalized linear bandits (Kveton et al., 2019) without adding even more restrictive assumptions. One possible solution would be to use the new analysis by (Faury et al., 2020) for the logistic bandit to see whether or not Assumption 5 holds in this setting.

References

Abbasi-Yadkori, Y., Pál, D., & Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 2312–2320.

Abbasi-Yadkori, Y., & Szepesvári, C. (2015). Bayesian optimal control of smoothly parameterized systems. *UAI*, 1–11.

Abeille, M., & Lazaric, A. (2020). Efficient optimistic exploration in linear-quadratic regulators via lagrangian relaxation. In H. D. III & A. Singh (Eds.), *Proceedings of the 37th international conference on machine learning* (pp. 23–31). PMLR. https://proceedings.mlr.press/v119/abeille20a.html.

Agrawal, P., Chen, J., & Jiang, N. (2020). Improved worst-case regret bounds for randomized least-squares value iteration. *arXiv preprint arXiv:2010.12163*.

Agrawal, S., Avadhanula, V., Goyal, V., & Zeevi, A. (2016). A near-optimal exploration-exploitation approach for assortment selection. *Proceedings of the 2016 ACM Conference on Economics and Computation*, 599–600.

Agrawal, S., & Jia, R. (2017). Optimistic posterior sampling for reinforcement learning: Worst-case regret bounds. *Advances in Neural Information Processing Systems*, 1184–1194.

AlQuraishi, M. (2019). AlphaFold at CASP13. *Bioinformatics*, *35*(22), 4862–4865.

Arulkumaran, K., Cully, A., & Togelius, J. (2019). Alphastar: An evolutionary computation perspective. *arXiv preprint arXiv:1902.01724*.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., & Yang, L. (2020). Model-based reinforcement learning with value-targeted regression. *International Conference on Machine Learning*, 463–474.

Azar, M. G., Osband, I., & Munos, R. (2017). Minimax regret bounds for reinforcement learning. *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 263–272.

Baird, L. (1995). Residual algorithms: Reinforcement learning with function approximation. *Machine learning proceedings 1995* (pp. 30–37). Elsevier.

Bertsekas, D. P., & Shreve, S. (1978). *Stochastic optimal control: The discrete-time case*. Academic Press.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Athena Scientific.

Boucheron, S., Lugosi, G., & Massart, P. (2013). *Concentration inequalities. A nonasymptotic theory of independence*. Oxford University Press. https://hal.archives-ouvertes.fr/hal-00794821.

Bradtke, S. J., & Barto, A. G. (1996). Linear least-squares algorithms for temporal difference learning. *Machine learning*, *22*(1-3), 33–57.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W. (2016). Openai gym. *arXiv preprint arXiv:1606.01540*.

Cai, Q., Yang, Z., Jin, C., & Wang, Z. (2019). Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*.

Dani, V., Hayes, T. P., & Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback.

Dann, C., Lattimore, T., & Brunskill, E. (2017). Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 5713–5723.

Dann, C., Li, L., Wei, W., & Brunskill, E. (2018). Policy certificates: Towards accountable reinforcement learning. *arXiv preprint arXiv:1811.03056*.

Du, S. S., Luo, Y., Wang, R., & Zhang, H. (2019). Provably efficient Q-learning with function approximation via distribution shift error checking oracle. *arXiv preprint arXiv:1906.06321*.

Farahmand, A.-M. (2018). Iterative value-aware model learning. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 9090–9101.

Farahmand, A.-M., Barreto, A., & Nikovski, D. (2017). Value-aware loss function for model-based reinforcement learning. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, *54*, 1486–1494.

Faury, L., Abeille, M., Calauzenes, C., & Fercoq, O. (2020). Improved optimistic algorithms for logistic bandits. In H. D. III & A. Singh (Eds.), *Proceedings of the 37th international conference on machine learning* (pp. 3052–3060). PMLR. https://proceedings.mlr.press/v119/faury20a.html.

Fortunato, M., Azar, M. G., Piot, B., Menick, J., Osband, I., Graves, A., Mnih, V., Munos, R., Hassabis, D., Pietquin, O., et al. (2017). Noisy networks for exploration. *arXiv preprint arXiv:1706.10295*.

Ishfaq, H., Cui, Q., Nguyen, V., Ayoub, A., Yang, Z., Wang, Z., Precup, D., & Yang, L. F. (2021). Randomized exploration for reinforcement learning with general value function approximation. *ICML*.

Jaksch, T., Ortner, R., & Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, *11*(Apr), 1563–1600.

Jin, C., Allen-Zhu, Z., Bubeck, S., & Jordan, M. I. (2018). Is q-learning provably efficient? *Advances in Neural Information Processing Systems*, 4863–4873.

Jin, C., Yang, Z., Wang, Z., & Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. *Conference on Learning Theory*, 2137–2143.

Kaiser, L., Babaeizadeh, M., Milos, P., Osinski, B., Campbell, R. H., Czechowski, K., Erhan, D., Finn, C., Kozakowski, P., Levine, S., Mohiuddin, A., Sepassi, R., Tucker, G., & Michalewski, H. (2019). Model-based reinforcement learning for Atari. *ICLR*.

Kakade, S., Wang, M., & Yang, L. F. (2018). Variance reduction methods for sublinear reinforcement learning. *arXiv preprint arXiv:1802.09184*.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kober, J., Bagnell, J. A., & Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, *32*(11), 1238–1274.

Kovalenko, B. G. I. N. (1968). *Introduction to queueing theory*. Israel Program for Scientific Translation, Jerusalem.

Kveton, B., Zaheer, M., Szepesvari, C., Li, L., Ghavamzadeh, M., & Boutilier, C. (2019). Randomized exploration in generalized linear bandits.

Lan, Q. (2019). A pytorch reinforcement learning framework for exploring new ideas.

Lattimore, T., & Szepesvári, C. (2018). Bandit algorithms. *preprint*, 28.

Lattimore, T., & Szepesvári, C. (2019). Learning with good feature representations in bandits and in RL with a generative model.

Lee, K., Laskin, M., Srinivas, A., & Abbeel, P. (2020). Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning. *arXiv preprint arXiv:2007.04938*.

Lu, X., & Van Roy, B. (2017). Ensemble sampling. *Advances in neural information processing systems*, 3258–3266.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing Atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529.

Modi, A., Jiang, N., Tewari, A., & Singh, S. (2019). Sample complexity of reinforcement learning using linearly combined model ensembles. *arXiv preprint arXiv:1910.10597*.

Osband, I., Blundell, C., Pritzel, A., & Van Roy, B. (2016). Deep exploration via bootstrapped dqn. *arXiv preprint arXiv:1602.04621*.

Osband, I., Russo, D., & Van Roy, B. (2013). (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 3003–3011.

Osband, I., & Van Roy, B. (2014). Model-based reinforcement learning and the eluder dimension. *arXiv preprint arXiv:1406.1853*.

Osband, I., & Van Roy, B. (2016). On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*.

Osband, I., Van Roy, B., Russo, D., & Wen, Z. (2017). Deep exploration via randomized value functions. *arXiv preprint arXiv:1703.07608*.

Osband, I., Van Roy, B., Russo, D. J., & Wen, Z. (2019). Deep exploration via randomized value functions. *Journal of Machine Learning Research*, *20*(124), 1–62.

Osband, I., Van Roy, B., & Wen, Z. (2014). Generalization and exploration via randomized value functions. *arXiv preprint arXiv:1402.0635*.

Osband, I., Van Roy, B., & Wen, Z. (2016). Generalization and exploration via randomized value functions. *International Conference on Machine Learning*, 2377–2386.

Ouyang, Y., Gagrani, M., Nayyar, A., & Jain, R. (2017a). Learning unknown Markov decision processes: A Thompson sampling approach. *Advances in Neural Information Processing Systems*, 1333–1342.

Ouyang, Y., Gagrani, M., Nayyar, A., & Jain, R. (2017b). Learning unknown markov decision processes: A thompson sampling approach. *Advances in Neural Information Processing Systems*, 1333–1342.

Parr, R., Li, L., Taylor, G., Painter-Wakefield, C., & Littman, M. L. (2008). An analysis of linear models, linear value-function approximation, and feature selection for reinforcement learning. *Proceedings of the 25th international conference on Machine learning*, 752–759.

Pires, B., & Szepesvári, C. (2016). Policy error bounds for model-based reinforcement learning with factored linear models. *COLT*, 121–151.

Rusmevichientong, P., & Tsitsiklis, J. N. (2010). Linearly parameterized bandits. *Mathematics of Operations Research*, *35*(2), 395–411.

Russel, S., & Norvig, P. (2003). *Artificial intelligence – a modern approach*. Prentice Hall.

Russo, D. (2019). Worst-case regret bounds for exploration via randomized value functions. *Advances in Neural Information Processing Systems*, 14410–14420.

Russo, D., & Van Roy, B. (2013). Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 2256–2264.

Russo, D., & Van Roy, B. (2014). Learning to optimize via posterior sampling. *Mathematics of Operations Research*, *39*(4), 1221–1243.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., et al. (2019). Mastering atari, go, chess and shogi by planning with a learned model. *arXiv preprint arXiv:1911.08265*.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354–359.

Strehl, A. L., & Littman, M. L. (2008). An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, *74*(8), 1309–1331.

Strens, M. (2000). A bayesian framework for reinforcement learning. *ICML*, *2000*, 943–950.

Strens, M. J. A. (2000). A Bayesian framework for reinforcement learning. *ICML*, 943–950.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., & Langford, J. (2019). Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. *Conference on Learning Theory*, 2898–2933.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). The MIT Press.

Theocharous, G., Wen, Z., Abbasi-Yadkori, Y., & Vlassis, N. (2017). Posterior sampling for large scale reinforcement learning. *arXiv preprint arXiv:1711.07979*.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, *25*(3/4), 285–294.

Tsitsiklis, J. N., & Van Roy, B. (1997). Analysis of temporal-diffference learning with function approximation. *Advances in neural information processing systems*, 1075–1081.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science* (Vol. 47). Cambridge university press.

Wang, R., Salakhutdinov, R. R., & Yang, L. (2020). Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, *33*.

Wen, Z., & Van Roy, B. (2013). Efficient exploration and value function generalization in deterministic systems. *Advances in Neural Information Processing Systems*, 3021–3029.

Wen, Z., & Van Roy, B. (2017). Efficient reinforcement learning in deterministic systems with value function generalization. *Mathematics of Operations Research*, *42*(3), 762–782.

Xie, C., Patil, S., Moldovan, T., Levine, S., & Abbeel, P. (2016). Model-based reinforcement learning with parametrized physical models and optimism-driven exploration. *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 504–511.

Yang, L., & Wang, M. (2019). Sample-optimal parametric q-learning using linearly additive features. *International Conference on Machine Learning*, 6995–7004.

Yang, L. F., & Wang, M. (2019). Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. *arXiv preprint arXiv:1905.10389*.

Zanette, A., Lazaric, A., Kochenderfer, M. J., & Brunskill, E. (2019). Limiting extrapolation in linear approximate value iteration. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 5616–5625). Curran Associates, Inc.

Zanette, A., Brandfonbrener, D., Brunskill, E., Pirotta, M., & Lazaric, A. (2020). Frequentist regret bounds for randomized least-squares value iteration. *International Conference on Artificial Intelligence and Statistics*, 1954–1964.

Zanette, A., Lazaric, A., Kochenderfer, M., & Brunskill, E. (2020). Learning near optimal policies with low inherent bellman error. *International Conference on Machine Learning*, 10978–10989.

Zhang, Z., Zhou, Y., & Ji, X. (2020). Almost optimal model-free reinforcement learning via reference-advantage decomposition. *arXiv preprint arXiv:2004.10019*.

Zhou, D., Gu, Q., & Szepesvari, C. (2020). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. *ArXiv, abs/2012.08507*.

Appendix A

Upper Confidence Reinforcement Learning with Value-Targeted Regression Appendix

A.1 Proof of Theorem 5

In this section, we provide the regret analysis of the UCRL-VTR Algorithm (Algorithm 1). We will explain the motivation for our construction of confidence sets for general nonlinear squared estimation, and establish the regret bound for a general class of transition models, $\mathcal{P}$.

A.1.1 Preliminaries

Recall that a finite horizon MDP is $M = (\mathcal{S}, \mathcal{A}, P, r, H, s_\circ)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P = (P_a)_{a\in\mathcal{A}}$ is a collection of $P_a : \mathcal{S} \to M_1(\mathcal{S})$ Markov kernels, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, $H > 0$ is the horizon and $s_\circ \in \mathcal{S}$ is the initial state. For a state $s \in \mathcal{S}$ and an action $a \in \mathcal{A}$, $P_a(s)$ gives the distribution of the next state that is obtained when action a is executed in state s. For a bounded (measurable) function $V : \mathcal{S} \to \mathbb{R}$, we will use $\langle P_a(s), V\rangle$ as the shorthand for the expected value of V at a random next state s' whose distribution is $P_a(s)$.

Given any policy π (which may or may not use the history), its value function is

$$V^{\pi}(s) = \mathbb{E}_{\pi,\delta_s}\left[\sum_{i=1}^{H} r(s_i, a_i)\right] ,$$

where E_{π,δ_s} is the expectation operator underlying the probability measure P_{π,δ_s} induced over sequences of state-action pairs of length H by executing policy π starting at state s in the MDP M and s_h is the state visited in stage h and action a_h is the action taken in that stage after visiting s_h. For a nonstationary Markov policy $\pi = (\pi_1, \ldots, \pi_H)$, we also let

$$V_h^{\pi}(s) = \mathbb{E}_{\pi_{h:H},\delta_s}\left[\sum_{i=1}^{H-h+1} r(s_i, a_i)\right]$$

be the value function of π from stage h to H. Here, $\pi_{h:H}$ denotes the policy $(\pi_h, \ldots, \pi_H)$. The optimal value function $V^* = (V_1^*, \ldots, V_H^*)$ is defined via $V_h^*(s) = \max_{\pi} V_h^{\pi}(s)$, $s \in \mathcal{S}$

For simplicity assume that r is known. To indicate the dependence of V^* on the transition model P, we will write $V_P^* = (V_{P,1}^*, \ldots, V_{P,H}^*)$. For convenience, we define $V_{P,H+1}^* = 0$.

Algorithm 1 is an instance of the following general model-based optimistic algorithm: Specific instances of Algorithm 4 differ in terms of how $\mathcal{B}_{k+1}$ is

Algorithm 4 Generic Algorithm 1-Schema for finite horizon problems

```
Input: $\mathcal{P}$ – a set of transition models, $K$ – number of episodes, $s_0$ – initial state
Set $\mathcal{B}_1 = \mathcal{P}$                    ▷ Initial confidence set for transition models
for $k = 1, \ldots, K$ do                                                  ▷ episodes
    $P^k = \text{argmax}\{V_{\widetilde{P}}^*(s_0) \, : \, \widetilde{P} \in \mathcal{B}_k\}$              ▷ Optimistic model
    $V_k = V_{P^k}^*$                             ▷ Optimistic $H$-stage value function
    $s_1^k = s_0$
    for $h = 1, \ldots, H$ do                                               ▷ Acting
        Choose $a_h^k = \text{argmax}_{a \in \mathcal{A}} r(s_h^k, a) + \langle P_a^k(s_h^k), V_{h+1,k} \rangle$
        Observe transition to $s_{h+1}^k$
    end for
    Construct $\mathcal{B}_{k+1}$ based on $(s_1^k, a_1^k, \ldots, s_H^k, a_H^k)$
end for
```

constructed. In particular, UCRL-VTR uses the construction described in Section 4.1.2.

Recall that $V_k = (V_{1,k}, \dots, V_{H,k}, V_{H+1,k})$ (with $V_{H+1,k} = 0$) in Algorithm 4. Let π_k be the nonstationary Markov policy chosen in episode k by Algorithm 4. Let,

$$R_K = \sum_{k=1}^{K} V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k)$$

be the pseudo-regret of Algorithm 1 for K episodes. The following standard lemma bounds the k-th term of the expression on the right-hand side.

Lemma 13. *Assuming that $P \in \mathcal{B}_k$, we have*

$$V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \le \sup_{\widetilde{P} \in \mathcal{B}_k} \sum_{h=1}^{H-1} \langle \widetilde{P}_{a_h^k}(s_h^k) - P_{a_h^k}(s_h^k), V_{h,k} \rangle + \sum_{h=1}^{H-1} \xi_{h+1,k} \,,$$

where

$$\xi_{h+1,k} = \langle P_{a_h^k}(s_h^k), V_{h+1,k} - V_{h+1}^{\pi_k} \rangle - \left(V_{h+1,k}(s_{h+1}^k) - V_{h+1}^{\pi_k}(s_{h+1}^k) \right) \,.$$

Note that $(\xi_{2,1}, \xi_{3,1}, \dots, \xi_{H,1}, \xi_{2,2}, \xi_{3,2}, \dots, \xi_{H,2}, \xi_{2,3}, \dots)$ is a sequence of martingale differences. *Proof* Because $P \in \mathcal{B}_k$, $V_1^*(s_1^k) \le V_{1,k}(s_1^k)$ by the definition of the algorithm. Hence,

$$V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \le V_{1,k}(s_1^k) - V_1^{\pi_k}(s_1^k) \,.$$

Fix $h \in [H]$. In what follows we bound $V_{h,k}(s_h^k) - V_h^{\pi_k}(s_h^k)$. By the definition of π_k, P^k and a_h^k, we have

$$\begin{aligned} V_{h,k}(s_h^k) &= r(s_h^k, a_h^k) + \langle P_{a_h^k}^k(s_h^k), V_{h+1,k} \rangle \text{ and} \\ V_h^{\pi_k}(s_h^k) &= r(s_h^k, a_h^k) + \langle P_{a_h^k}(s_h^k), V_{h+1}^{\pi_k} \rangle \,. \end{aligned}$$

Hence,

$$\begin{aligned} V_{h,k}(s_h^k) - V_h^{\pi_k}(s_h^k) &= \langle P_{a_h^k}^k(s_h^k), V_{h+1,k} \rangle - \langle P_{a_h^k}(s_h^k), V_{h+1}^{\pi_k} \rangle \\ &= \langle P_{a_h^k}^k(s_h^k) - P_{a_h^k}(s_h^k), V_{h+1,k} \rangle + \langle P_{a_h^k}(s_h^k), V_{h+1,k} - V_{h+1}^{\pi_k} \rangle \,. \end{aligned}$$

Therefore, by induction, noting that $V_{H+1,k} = 0$, we get that

$$\begin{aligned} V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) &\leq \sum_{h=1}^{H-1} \langle P^k_{a_h^k}(s_h^k) - P_{a_h^k}(s_h^k), V_{h+1,k}\rangle + \sum_{h=1}^{H-1} \xi_{h+1,k} \\ &\leq \sup_{\widetilde{P} \in \mathcal{B}_k} \sum_{h=1}^{H-1} \langle \widetilde{P}_{a_h^k}(s_h^k) - P_{a_h^k}(s_h^k), V_{h+1,k}\rangle + \sum_{h=1}^{H-1} \xi_{h+1,k} \,. \end{aligned}$$

A.1.2 The confidence sets for Algorithm 1

The previous lemma suggests that at the end of the kth episode, the model could be estimated using

$$\widehat{P}_k = \text{argmin}_{\widetilde{P} \in \mathcal{P}} \sum_{k'=1}^{k} \sum_{h=1}^{H-1} \left(\langle \widetilde{P}_{a_h^{k'}}(s_h^{k'}), V_{h+1,k'}\rangle - V_{h+1,k'}(s_{h+1}^{k'}) \right)^2 \tag{A.1}$$

For a confidence set construction, we get inspiration from Proposition 5 in the paper of (Osband and Van Roy, 2014). The set is centered at $\widehat{P}_k$:

$$\mathcal{B}_k = \{ \widetilde{P} \in \mathcal{P} \, : \, L_k(\widehat{P}_k, \widetilde{P}) \leq \beta_k \} \,, \tag{A.2}$$

where

$$L_k(\widehat{P}, \widetilde{P}) = \sum_{k'=1}^{k} \sum_{h=1}^{H-1} \left(\langle \widetilde{P}_{a_h^{k'}}(s_h^{k'}) - \widehat{P}_{a_h^{k'}}(s_h^{k'}), V_{h+1,k'}\rangle \right)^2 \,.$$

Note that this is the same confidence set as described in Section 4.1.2. To obtain the value of β_k, we now consider the nonlinear least-squares confidence set construction from (Russo and Van Roy, 2014). The next section is devoted to this construction.

A.1.3 Confidence sets for general nonlinear least-squares

Let $(X_p, Y_p)_{p=1,2,\ldots}$ be a sequence of random elements, $X_p \in \mathcal{X}$ for some measurable set $\mathcal{X}$ and $Y_p \in \mathbb{R}$. Let $\mathcal{F}$ be a subset of the set of real-valued measurable functions with domain $\mathcal{X}$. Let $\mathbb{F} = (\mathbb{F}_p)_{p=0,1,\ldots}$ be a filtration such that for all $p \geq 1$, $(X_1, Y_1, \ldots, X_{p-1}, Y_{p-1}, X_p)$ is $\mathbb{F}_{p-1}$ measurable and such that there exists some function $f_* \in \mathcal{F}$ such that $\mathbb{E}[Y_p \mid \mathbb{F}_{p-1}] = f_*(X_p)$ holds for all $p \geq 1$. The (nonlinear) least-squares predictor given $(X_1, Y_1, \ldots, X_t, Y_t)$

is $\widehat{f}_t = \operatorname{argmin}_{f\in\mathcal{F}} \sum_{p=1}^{t}(f(X_p) - Y_p)^2$. We say that Z is conditionally ρ-subgaussian given the σ-algebra $\mathbb{F}$ if for all $\lambda \in \mathbb{R}$, $\log \mathbb{E}[\exp(\lambda Z)|\mathbb{F}] \le \frac{1}{2}\lambda^2\rho^2$. For $\alpha > 0$, let N_α be the $\|\cdot\|_\infty$-covering number of $\mathcal{F}$ at scale α. That is, N_α is the smallest integer for which there exist $\mathcal{G} \subset \mathcal{F}$ with N_α elements such that for any $f \in \mathcal{F}$, $\min_{g\in\mathcal{G}} \|f - g\|_\infty \le \alpha$. For $\beta > 0$, define

$$\mathcal{F}_t(\beta) = \{f \in \mathcal{F} \,:\, \sum_{p=1}^{t}(f(X_p) - \widehat{f}_t(X_p))^2 \le \beta\}\,.$$

We have the following theorem, the proof of which is given in Section A.1.6.

Theorem 14. *Let $\mathbb{F}$ be the filtration defined above and assume that the functions in $\mathcal{F}$ are bounded by the positive constant $C > 0$. Assume that for each $s \ge 1$, $(Y_p - f_*(X_p))_p$ is conditionally σ-subgaussian given $\mathbb{F}_{p-1}$. Then, for any $\alpha > 0$, with probability $1-\delta$, for all $t \ge 1$, $f_* \in \mathcal{F}_t(\beta_t(\delta,\alpha))$, where*

$$\beta_t(\delta,\alpha) = 8\sigma^2 \log(2N_\alpha/\delta) + 4t\alpha\left(C + \sqrt{\sigma^2 \log(4t(t+1)/\delta)}\right)\,.$$

The proof follows that of Proposition 6, (citeRuVR14), with minor improvements, which lead to a slightly better bound. In particular, with our notation, (Russo and Van Roy, 2014) stated their result with

$$\beta_t^{\mathrm{RvR}}(\delta,\alpha) = 8\sigma^2 \log(2N_\alpha/\delta) + 2t\alpha\left(8C + \sqrt{8\sigma^2 \log(8t^2/\delta)}\right)\,.$$

While $\beta_t(\delta,\alpha) \le \beta_t^{\mathrm{RvR}}(\delta,\alpha)$, the improvement is only in terms of smaller constants.

A.1.4 The choice of β_k in Algorithm 1

To use this result in our RL problem recall that $\mathcal{P}$ is the set of transition probabilities parameterized by $\theta \in \Theta$. We index time $t = 1, 2, \ldots$ in a continuous fashion. Episode $k = 1, 2, \ldots$ and stage $h = 1, \ldots, H-1$ corresponds to time $t = (k-1)(H-1) + h$:

episode (k)	1	1	...	1	2	2	...	2	3	...
stage (h)	1	2	...	$H-1$	1	2	...	$H-1$	1	...
time step (t)	1	2	...	$H-1$	H	$H+1$	...	$2H-2$	$2H-1$	...

Note that the transitions at stage $h = H$ are skipped and the time index at the end of episode $k \geq 1$ is $k(H-1)$.

Let $V_{(t)}$ be the value function used by Algorithm 1 at time t ($V_{(t)}$ is constant in periods of length $H-1$), while let $(s_{(t)}, a_{(t)})$ be the state-action pair visited at time t.

Let $\mathcal{V}$ be the set of optimal value functions under some model in $\mathcal{P}$: $\mathcal{V} = \{V^*_{P'} : P' \in \mathcal{P}\}$. Note that $\mathcal{V} \subset \mathcal{B}(\mathcal{S}, H)$, where $\mathcal{B}(\mathcal{S}, H)$ denotes the set of real-valued measurable functions with domain $\mathcal{S}$ that are bounded by H. Note also that for all t, $V_{(t)} \in \mathcal{V}$. Define $\mathcal{X} = \mathcal{S} \times \mathcal{A} \times \mathcal{V}$. We also let $X_t = (s_{(t)}, a_{(t)}, V_{(t)})$, $Y_t = V_{(t)}(s_{(t+1)})$ when $t+1 \notin \{H+1, 2H+1, \dots\}$ and $Y_t = V_{(t)}(s^k_{H+1})$, and choose

$$\mathcal{F} = \left\{ f : \mathcal{X} \to \mathbb{R} \,:\, \exists \widetilde{P} \in \mathcal{P} \text{ s.t. } f(s,a,v) = \int \widetilde{P}_a(ds'|s)v(s') \right\} . \tag{A.3}$$

Note that $\mathcal{F} \subset \mathcal{B}_\infty(\mathcal{X}, H)$.

Let $\phi : \mathcal{P} \to \mathcal{F}$ be the natural surjection to $\mathcal{F}$: $\phi(P) = f$ where $f(s,a,v) = \int P_a(ds'|s)v(s')$ for $(s,a,v) \in \mathcal{X}$. We know show that ϕ is in fact a bijection. If $P \neq P'$, this means that for some $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $U \subset \mathcal{S}$ measurable, $P_a(U|s) \neq P'_a(U|s)$. Choosing v to be the indicator of U, note that $(s,a,v) \in \mathcal{X}$. Hence, $\phi(P)(s,a,v) = P_a(U|s) \neq P'_a(U|s) = \phi(P')(s,a,v)$, and hence $\phi(P) \neq \phi(P')$: ϕ is indeed a bijection. For convenience and to reduce clutter, we will write $f_P = \phi(P)$.

Choose $\mathbb{F} = (\mathbb{F}_t)_{t \geq 0}$ so that $\mathbb{F}_{t-1}$ is generated by $(s_{(1)}, a_{(1)}, V_{(1)}, \dots, s_{(t)}, a_{(t)}, V_{(t)})$. Then $\mathbb{E}[Y_t|\mathbb{F}_{t-1}] = \int P_{a_{(t)}}(ds'|s_{(t)})V_{(t)}(s') = f_P(X_t)$ and by definition $f_P \in \mathcal{F}$. Now, $Y_t \in [0, H]$, hence, $Z_t = Y_t - f_P(X_t)$ is conditionally $H/2$-subgaussian given $\mathbb{F}_{t-1}$.

Let $t = k(H-1)$ for some $k \geq 1$. Thus, this time step corresponds to finishing episode k and thus $V_{(t)} = V_k$. Furthermore, letting $\widehat{f}_t = \operatorname{argmin}_{f \in \mathcal{F}} \sum_{p=1}^{t}(f(X_p) -$

$Y_p)^2$, since ϕ is an injection, we see that $\widehat{f}_t = f_{\widehat{P}_k}$ where $\widehat{P}_k$ is defined using A.1. For $P', P'' \in \mathcal{P}$, we have $L_k(P', P'') = \sum_{p=1}^{t}(f_{P'}(X_p) - f_{P''}(X_p))^2$ and thus

$$\begin{aligned}\mathcal{B}_k = \{\widetilde{P} \in \mathcal{P} \,:\, L_k(\widehat{P}_k, \widetilde{P}) \leq \beta_k\} &= \{\widetilde{P} \in \mathcal{P} \,:\, \sum_{p=1}^{t}(\widehat{f}_t(X_p) - f_{\widetilde{P}}(X_p))^2 \leq \beta_k\} \\ &= \{\phi^{-1}(f) \,:\, f \in \mathcal{F} \text{ and } \sum_{p=1}^{t}(\widehat{f}_t(X_p) - f(X_p))^2 \leq \beta_k\} = \phi^{-1}(\mathcal{F}_t(\beta_k))\,.\end{aligned}$$

Corollary 15. *For $\alpha > 0$ and $k \geq 1$ let*

$$\beta_k = 2H^2 \log\left(\frac{2\mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|_\infty)}{\delta}\right) + 2H(kH-1)\alpha\left\{2 + \sqrt{\log\left(\frac{4kH(kH-1)}{\delta}\right)}\right\}.$$

Then, with probability $1-\delta$, for any $k \geq 1$, $P \in \mathcal{B}_k$ where $\mathcal{B}_k$ is defined by (A.2).

A.1.5 Regret of Algorithm 1

Recall that $\mathcal{X} = \mathcal{S} \times \mathcal{A} \times \mathcal{V}$ where $\mathcal{V} \subset \mathcal{B}_\infty(\mathcal{S}, H)$ is the set of value functions that are optimal under some model in $\mathcal{P}$. We will abbreviate $(x_1, \ldots, x_t) \in \mathcal{X}^t$ as $x_{1:t}$. Further, we let $\mathcal{F}|_{x_{1:t}} = \{(f(x_1), \ldots, f(x_t)) \,:\, f \in \mathcal{F}\} (\subset \mathbb{R}^t)$ and for $S \subset \mathbb{R}^t$, let $\operatorname{diam}(S) = \sup_{u,v \in S} \|u - v\|_2$ be the diameter of S. We will need the following lemma, extracted from (Russo and Van Roy, 2014):

Lemma 16. ***(Lemma 5 of [Russo and Van Roy, 2014])*** *Let $\mathcal{F} \subset B_\infty(\mathcal{X}, C)$ be a set of functions bounded by $C > 0$, $(\mathcal{F}_t)_{t\geq 1}$ and $(x_t)_{t \geq 1}$ be sequences such that $\mathcal{F}_t \subset \mathcal{F}$ and $x_t \in \mathcal{X}$ hold for $t \geq 1$. Then, for any $T \geq 1$ and $\alpha > 0$ it holds that*

$$\sum_{t=1}^{T} \operatorname{diam}(\mathcal{F}_t|_{x_t}) \leq \alpha + C(d \wedge T) + 2\delta_T\sqrt{dT}\,,$$

where $\delta_T = \max_{1 \leq t \leq T} \operatorname{diam}(\mathcal{F}_t|_{x_{1:t}})$ and $d = \dim_{\mathcal{E}}(\mathcal{F}, \alpha)$.

Let

$$W_k = \sup_{\widetilde{P} \in \mathcal{B}_k} \sum_{h=1}^{H-1} \langle \widetilde{P}_{a_h^k}(s_h^k) - P_{a_h^k}(s_h^k), V_{h,k} \rangle\,.$$

From Lemma 13, we get

$$R_K \le \sum_{k=1}^{K} W_k + \sum_{k=1}^{K} \sum_{h=1}^{H-1} \xi_{h+1,k} \,. \tag{A.4}$$

Lemma 17. *Let $\alpha > 0$ and $d = \dim_{\mathcal{E}}(\mathcal{F}, \alpha)$ where $\mathcal{F}$ is given by* (A.3)*. Then, for any nondecreasing sequence $(\beta_k^2)_{k=1}^{K}$, on the event when $P \in \cap_{k\in[K]}\mathcal{B}_k$,*

$$\sum_{k=1}^{K} W_k \le \alpha + H(d \wedge K(H-1)) + 4\sqrt{d\beta_K K(H-1)}\,.$$

Proof. Let $P \in \cap_{k\in[K]}\mathcal{B}_k$ holds. Using the notation of the previous section, letting $\widetilde{\mathcal{F}}_t = \mathcal{F}_t(\beta_k)$ for $(k-1)(H-1)+1 \le t \le k(H-1)$, we have

$$\begin{aligned}
\sum_{k=1}^{K} W_k &\le \sum_{k=1}^{K} \sup_{\widetilde{P}\in\mathcal{B}_k} \sum_{h=1}^{H-1} \left(f_{\widetilde{P}}(s_h^k, a_h^k, V_{h+1,k}) - f_P(s_h^k, a_h^k, V_{h+1,k})\right) \\
&\le \sum_{t=1}^{K(H-1)} \operatorname{diam}(\widetilde{\mathcal{F}}_t|_{X_t}) \qquad\qquad (\text{because } P \in \cap_{k\in[K]}\mathcal{B}_k) \\
&\le \alpha + H(d \wedge K(H-1)) + 2\delta_{K(H-1)}\sqrt{dK(H-1)}\,,
\end{aligned}$$

where X_t is defined in Section A.1.4 and where the last inequality is by Lemma 16, which is applicable because $\mathcal{F} \subset \mathcal{B}_\infty(\mathcal{X}, H)$ holds by choice, and $\delta_{K(H-1)} = \max_{1\le t\le K(H-1)} \operatorname{diam}(\widetilde{\mathcal{F}}_t|_{X_{1:t}})$. Thanks to the definition of $\widetilde{\mathcal{F}}_t$, $\delta_{K(H-1)} \le 2\sqrt{\beta_K}$. Plugging this into the previous display finishes the proof. □

Proof of Theorem 5

Proof. Note that for any $k \in [K]$ and $h \in [H-1]$, $\xi_{h+1}, k \in [-H, H]$. As noted beforehand, $\xi_{2,1}, \xi_{3,1}, \ldots, \xi_{H,1}, \xi_{2,2}, \xi_{3,2}, \ldots, \xi_{H,2}, \xi_{2,3}, \ldots$ is a martingale difference sequence. Thus, with probability $1-\delta$, $\sum_{k=1}^{K}\sum_{h=1}^{H-1} \xi_{h+1,k} \le H\sqrt{2K(H-1)\log(1/\delta)}$. Consider the event when this inequality holds and when $P \in \cap_{k\in[K]}\mathcal{B}_k$. By using Corollary 15 and a union bound, this event holds with probability at least $1-2\delta$. On this event, by (A.4) and Lemma 17, we obtain

$$R_K \le \alpha + H(d \wedge K(H-1)) + 4\sqrt{d\beta_K K(H-1)} + H\sqrt{2K(H-1)\log(1/\delta)}\,.$$

Using $\alpha \le 1$, which holds by assumption, finishes the proof. □

Proof of Corollary 6

Proof. Note that

$$\begin{aligned}\|f_{P'} - f_{P''}\|_\infty &= \sup_{s,a,v} \left| \int (P'_a(ds'|s) - P''_a(ds'|s))v(s')\right| \le H \sup_{s,a} \int |P'_a(ds'|s) - P''_a(ds'|s)| \\ &= H \sup_{s,a} \|P'_a(s) - P''_a(s)\|_1 =: H\|P' - P''\|_{\infty,1}\,.\end{aligned}$$

For $\alpha > 0$ let $\mathcal{N}(\mathcal{P}, \alpha, \|\cdot\|_{\infty,1})$ denote the $(\alpha, \|\cdot\|_{\infty,1})$-covering number of $\mathcal{P}$. Then we have

$$\mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|_\infty) \le \mathcal{N}(\mathcal{P}, \alpha/H, \|\cdot\|_{\infty,1}).$$

Then, by Corollary 15,

$$\beta_K = 2H^2 \log(2\mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|_\infty)/\delta) + C \le 2H^2 \log(2\mathcal{N}(\mathcal{P}, \alpha/H, \|\cdot\|_{\infty,1})/\delta) + C$$

with some universal constant $C > 0$. Let $f : (\Theta, \|\cdot\|) \to (\mathcal{P}, \|\cdot\|_{\infty,1})$ be defined by $\theta \mapsto \sum_j \theta_j P_j$. Note that $\|f(\theta) - f(\theta')\|_{\infty,1} \le \sup_{s,a} \sum_j \|(\theta_j - \theta'_j)P_{j,a}(s)\|_1 = \sum_j |\theta_j - \theta'_j| = \|\theta - \theta'\|_1$. Hence, any $(\epsilon, \|\cdot\|_1)$ covering of Θ induces an $(\epsilon, \|\cdot\|_{\infty,1})$-covering of $\mathcal{P}$ and so $\mathcal{N}(\mathcal{P}, \alpha/H, \|\cdot\|_{\infty,1}) \le \mathcal{N}(\Theta, \alpha/H, \|\cdot\|_1) \le C'(RH/\alpha)^d$ with some universal constant $C' > 0$.

Now, choose $1/\alpha = K\sqrt{\log(KH/\delta)}$. Hence,

$$\beta_K \le 2H^2(\log(2C'/\delta) + d\log(RH/\alpha)) + C\,.$$

Suppressing log factors (e.g., $\log(RH)$), $\log\log$ terms and constants, we have $\beta_K = H^2(d + \log(1/\delta))$.

Let $\mathcal{F}$ be given by (A.3). We now bound $\dim_{\mathcal{E}}(\mathcal{F}, \alpha)$. Let $\mathcal{X} = \mathcal{S} \times \mathcal{A} \times B(\mathcal{S})$ as before. Define $z : \mathcal{S} \times \mathcal{A} \times B(\mathcal{S}) \to \mathbb{R}^d$ using $z(s, a, v)_j = \langle P_{j,a}(s), v\rangle$ and note that if $x \in \mathcal{X}$ is $(\epsilon, \mathcal{F})$-independent of $x_1, \dots, x_k \in \mathcal{X}$ then $z(x) \in \mathbb{R}^d$ is (ϵ, Θ)-independent of $z(x_1), \dots, z(x_k) \in \mathbb{R}^d$. This holds because if $P = \sum_j \theta_j P_j \in \mathcal{P}$ then $f_P(s, a, v) = \langle \theta, z(s, a, v)\rangle$ for any $(s, a, v) \in \mathcal{X}$. Hence, $\dim_{\mathcal{E}}(\mathcal{F}, \alpha) \le \dim_{\mathcal{E}}(\mathrm{Lin}(\mathcal{Z}, \Theta), \alpha)$, where $\mathrm{Lin}(\mathcal{Z}, \Theta)$ is the set of linear maps with domain $\mathcal{Z} = \{z(x) : x \in \mathcal{X}\} \subset \mathbb{R}^d$ and parameter from Θ: $\mathrm{Lin}(\mathcal{Z}, \Theta) = \{h : h : \mathcal{Z} \to \mathbb{R} \text{ s.t. } \exists\theta \in \Theta : h(z) = \langle \theta, z\rangle, z \in \mathcal{Z}\}$. Now, by Proposition 11 of (Russo and Van Roy, 2014), $\dim_{\mathcal{E}}(\mathrm{Lin}(\mathcal{Z}, \Theta), \alpha) = O(d\log(1 + (S\gamma/\alpha)^2)$

where S is the $\|\cdot\|_2$ diameter of Θ and $\gamma = \sup_{z \in \mathcal{Z}} \|z\|_2$. We have

$$\|z\|_2^2 = \sum_j (\langle P_{j,a}(s), v \rangle)^2 \leq H^2 d\,,$$

hence $\gamma \leq H\sqrt{d}$. By the relation between the 1 and 2 norms, the 2-norm diameter of Θ is at most $\sqrt{d}R$. Dropping log terms, $\dim_{\mathcal{E}}(\mathcal{F}, \alpha) = \widetilde{O}(d)$.

Plugging into Theorem 5 gives the desired result. □

A.1.6 Proof of Theorem 14

Recall the following:

Definition 7. *A random variable X is σ-subgaussian if for all $\lambda \in \mathbb{R}$, it holds that* $\mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2\sigma^2/2)$.

The proof of the next couple of statements is standard and is included only for completeness.

Theorem 18. *If X is σ-subgaussian, then for any $\lambda > 0$, with probability at least $1 - \delta$,*

$$X < \frac{1}{\lambda} \log\left(\frac{1}{\delta}\right) + \lambda\frac{\sigma^2}{2}\,. \tag{A.5}$$

Proof. Let $\lambda > 0$. We have, $\{X \geq \epsilon\} = \{\exp(\lambda(X - \epsilon)) \geq 0\}$. Hence, Markov's inequality gives $\mathbb{P}(X \geq \epsilon) \leq \exp(-\lambda\epsilon)\mathbb{E}[\exp(\lambda X)] \leq \exp(-\lambda\epsilon + \frac{1}{2}\lambda^2\sigma^2)$. Equating the right-hand side with δ and solving for ϵ, we get that $\log(\delta) = -\lambda\epsilon + \frac{1}{2}\lambda^2\sigma^2$. Solving for ϵ gives $\epsilon = \log(1/\delta)/\lambda + \frac{\sigma^2}{2}\lambda$, finishing the proof. □

Choosing the λ that minimizes the right-hand side of the bound gives the usual form:

$$\mathbb{P}(X \geq \sqrt{2\sigma^2 \log(1/\delta)}) \leq \delta\,. \tag{A.6}$$

Lemma 19. ***(Lemma 5.4 of Lattimore and Szepesvári, 2018)*** *Suppose that X is σ-subgaussian and X_1 and X_2 are independent and σ_1 and σ_2-subgaussian, respectively, then:*

1. $\mathbb{E}[X] = 0$.

2. cX *is* $|c|\sigma$*-subgaussian for all* $c \in \mathbb{R}$.

3. $X_1 + X_2$ *is* $\sqrt{\sigma_1^2 + \sigma_2^2}$*-subgaussian.*

Let $(Z_p)_p$ be an $\mathbb{F} = (\mathbb{F}_p)_p$-adapted process. Recall that $(Z_p)_p$ is conditionally σ-subgaussian given $\mathbb{F}$ if for all $p \geq 1$,

$$\log \mathbb{E}[\exp(\lambda Z_p)|\mathbb{F}_{p-1}] \leq \frac{1}{2}\lambda^2\sigma^2 \,, \qquad \text{for all } \lambda \in \mathbb{R} \,.$$

A standard calculation gives that $S_t = \sum_{p=1}^t Z_p$ is $\sqrt{t}\sigma$-subgaussian (essentially, a refinement of the calculation that is need to show Part (3) of Lemma 19) and thus, in particular, for any $t \geq 1$ and $\lambda > 0$, with probability $1 - \delta$,

$$S_t < \frac{1}{\lambda} \log\left(\frac{1}{\delta}\right) + \lambda \, \frac{t\sigma^2}{2} \,.$$

In fact, by slightly strengthening the argument, one can show that the above inequality holds simultaneously for all $t \geq 1$:

Theorem 20. ***(E.g., Lemma 7 of [Russo and Van Roy, 2014])*** *Let* $\mathbb{F}$ *be a filtration and let* $(Z_p)_p$ *be an* $\mathbb{F}$*-adapted, conditionally* σ*-subgaussian process. Then for any* $\lambda > 0$*, with probability at least* $1 - \delta$*, for all* $t \geq 1$*,*

$$S_t < \frac{1}{\lambda} \log\left(\frac{1}{\delta}\right) + \lambda \, \frac{t\sigma^2}{2} \,, \tag{A.7}$$

where $S_t = \sum_{p=1}^t Z_p$.

Proof of Theorem 14 Let us introduce the following helpful notation: For vectors $x, y \in \mathbb{R}^t$, let $\langle x, y \rangle_t = \sum_{p=1}^t x_p y_p$, $\|x\|_t^2 = \langle x, x \rangle_t$, and for $f : \mathcal{X} \to \mathbb{R}$, $\|f\|_t^2 = \sum_{p=1}^t f^2(X_p)$. More generally, we will overload addition and subtraction such that for $x \in \mathbb{R}^t$, $x + f \in \mathbb{R}^t$ is the vector whose pth coordinate is $x_p + f(X_p)$ (x_p and X_p both appear on purpose here). We also overload $\langle \cdot, \cdot \rangle_t$ such that $\langle x, f \rangle_t = \langle f, x \rangle_t = \sum_{p=1}^t x_p f(X_p)$.

Define Z_p using $Y_p = f_*(X_p) + Z_p$ and collect $(Y_p)_{p=1}^t$ and $(Z_p)_{p=1}^t$ into the vectors Y and Z. As in the statement of the theorem, let $\mathbb{F} = (\mathbb{F}_p)_{p=0,1,\ldots}$

be such that for any $s \geq 1$, $(X_1, Y_1, \ldots, X_{p-1}, Y_{p-1}, X_p)$ is $\mathbb{F}_{p-1}$-measurable. Note that for any $p \geq 1$, $Z_p = Y_p - f_*(X_p)$ is $\mathbb{F}_p$-measurable, hence $(Z_p)_{p\geq 1}$ is $\mathbb{F}$-adapted.

With this, elementary calculation gives

$$\|Y - f\|_t^2 - \|Y - f_*\|_t^2 = \|f_* - f\|_t^2 + 2\langle Z, f_* - f\rangle_t \,.$$

Splitting $\|f_* - f\|_t^2$ and rearranging gives

$$\frac{1}{2}\|f_* - f\|_t^2 = \|Y - f\|_t^2 - \|Y - f_*\|_t^2 + E(f) \tag{A.8}$$

where

$$E(f) = -\frac{1}{2}\|f_* - f\|_t^2 + 2\langle Z, f - f_*\rangle_t \,.$$

Recall that $\widehat{f}_t = \operatorname{argmin}_{f\in\mathcal{F}} \|Y - f\|_t^2$. Plugging $\widehat{f}_t$ into A.8 in place of f and using that thanks to $f_* \in \mathcal{F}$, $\|Y - \widehat{f}_t\|_t^2 \leq \|Y - f_*\|_t^2$, we get

$$\frac{1}{2}\|f_* - \widehat{f}_t\|_t^2 \leq E(\widehat{f}_t) \,. \tag{A.9}$$

Thus, it remains to bound $E(\widehat{f}_t)$. For this fix some $\alpha > 0$ to be chosen later and let $\mathcal{G}(\alpha) \subset \mathcal{F}$ be an α-cover of $\mathcal{F}$ in $\|\cdot\|_\infty$. Let $g \in \mathcal{G}(\alpha)$ be a random function, also to be chosen later. We have

$$E(\widehat{f}_t) = E(\widehat{f}_t) - E(g) + E(g) \leq E(\widehat{f}_t) - E(g) + \max_{\widetilde{g}\in\mathcal{G}(\alpha)} E(\widetilde{g}) \tag{A.10}$$

We start by bounding the last term above. A simple calculation gives that for any fixed $f \in \mathcal{F}$, w.p. $1 - \delta$, $2\langle Z, f - f_*\rangle_t$ is $2\sigma\|f - f_*\|_t$-subgaussian. Hence, with probability $1 - \delta$, simultaneously for all $t \geq 1$,

$$E(f) \leq -\frac{1}{2}\|f_* - f\|_t^2 + \frac{1}{\lambda}\log\left(\frac{1}{\delta}\right) + \lambda\,\frac{4\sigma^2\|f - f_*\|_t^2}{2} = 4\sigma^2 \log\left(\frac{1}{\delta}\right) \,,$$

where the equality follows by choosing $\lambda = 1/(4\sigma^2)$ (which makes the first and last terms cancel). (Note how splitting $\|f - f_*\|_t^2$ into two halves allowed us to bound the "error term" $E(f)$ independently of t.) Now, by a union bound, it follows that with probability at least $1-\delta$, the second term is bounded by $4\sigma^2 \log(|\mathcal{G}(\alpha)|/\delta)$.

Let us now turn to bounding the first term. We calculate

$$\begin{aligned} E(\widehat{f_t}) - E(g) &= \frac{1}{2}\|g - f_*\|_t^2 - \frac{1}{2}\|\widehat{f_t} - f_*\|_t^2 + 2\langle Z, \widehat{f_t} - g\rangle_t \\ &\leq \frac{1}{2}\left(\langle g - \widehat{f_t}, g + \widehat{f_t} + 2f_*\rangle_t\right) + 2\|Z\|_t\|\widehat{f_t} - g\|_t \\ &\leq \frac{1}{2} 4C\alpha\, t + 2\|Z\|_t \alpha\sqrt{t}\,, \end{aligned}$$

where for the last inequality we chose $g = \operatorname{argmin}_{\widetilde{g}\in\mathcal{G}(\alpha)} \|\widehat{f_t} - \widetilde{g}\|_\infty$ so that $\|\widehat{f_t} - g\|_t \leq \alpha\sqrt{t}$ and used Cauchy-Schwartz, together with that $\|g\|_t, \|\widehat{f_t}\|_t, \|f_*\|_t \leq C\sqrt{t}$, which follows from $g, \widehat{f_t}, f_* \in \mathcal{F}$ and that by assumption all functions in $\mathcal{F}$ are bounded by C.

It remains to bound $\|Z\|_t$. For this, we observe that with probability $1-\delta$, simultaneously for all $t \geq 1$,

$$\|Z\|_t \leq \sigma\sqrt{2t\log(2t(t+1)/\delta)}\,.$$

Indeed, this follows because with probability $1-\delta$, simultaneously for any $s \geq 1$, $|Z_p|^2 \leq 2\sigma^2\log(2s(s+1)/\delta)$ holds because of a union bound and Eq. (A.6). Therefore, for the above choice g, with probability $1-\delta$, simultaneously for all $t \geq 1$, it holds that

$$E(\widehat{f_t}) - E(g) \leq 2C\alpha\, t + 2t\alpha\sqrt{\sigma^2\log(2t(t+1)/\delta)}\,.$$

Merging this with Eqs. (A.9) and (A.10) and with another union bound, we get that with probability $1-\delta$, for any $t \geq 1$,

$$\|f_* - \widehat{f_t}\|_t^2 \leq 8\sigma^2\log(2N_\alpha/\delta) + 4t\alpha\left(C + \sqrt{\sigma^2\log(4t(t+1)/\delta)}\right)\,,$$

where N_α is the $(\alpha, \|\cdot\|_\infty)$-covering number of $\mathcal{F}$. □

A.2 Proof of Theorem 7

In this section we establish a regret lower bound by reduction to a known result for tabular MDP.

Proof We assume without loss of generality that d is a multiple of 4 and $d \geq 8$. We set $S = 2$ and $A = d/4 \geq 2$. According to (Azar et al., 2017; Osband and Van Roy, 2016), there exists an MDP $\mathcal{M}(\mathcal{S}, \mathcal{A}, P, r, H)$ with S states, A actions and horizon H such that any algorithm has regret at least $\Omega(\sqrt{HSAT})$. In this case, we have $|\mathcal{S} \times \mathcal{A} \times \mathcal{S}| = d$. We use $\sigma(s, a, s')$ to denote the index of (s, a, s') in $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$. Letting

$$P_i(s'|s,a) = \begin{cases} 1 & \text{if } \sigma(s,a,s') = i, \\ 0 & \text{otherwise,} \end{cases}$$

and $\theta^i = P(s'|s,a)$ if $\sigma(s,a,s') = i$, we will have $P(s'|s,a) = \sum_{i=1}^d \theta^i P_i(s'|s,a)$. Therefore P can be parametrized using A.11. Therefore, the known lower bound $\Omega(\sqrt{HSAT})$ implies a worst-case lower bound of $\Omega(\sqrt{H \cdot d/2 \cdot T}) = \Omega(\sqrt{HdT})$ for our model.

A.3 The Special Case of Linear Transition Models

We derive a modification of UCRL-VTR when P_θ is a linear model of the form $P_\theta = \sum_{j=1}^d \theta_j P_j$, which is captured in the following assumption:

Assumption 8 (Linear Parameterized Transition Model)**.** *There exists a vector $\theta_* \in \mathbb{R}^d$ such that $\|\theta_*\|_2 \leq C_\theta$ $(C_\theta \geq 1)$ and*

$$P(s'|s,a) = \sum_{j=1}^{d} (\theta_*)_j P_j(s'|s,a) = P_\cdot(s'|s,a)^\top \theta_*, \tag{A.11}$$

where P_j's are known basis models such that $\sup_{j\in[d],(s,a)\in\mathcal{S}\times\mathcal{A}} \|P_j(\cdot|s,a)\|_1 \leq 1$, *and $P_\cdot(s'|s,a)$ denotes the d-dimensional vector $P_\cdot(s'|s,a) = [P_1(s'|s,a), \ldots, P_d(s'|s,a)]^\top$*[1]. *Note that we do not require each basis model P_j to be a probability transition model.*

[1] We also use $P_\cdot(\cdot|s,a)$ to denote a $d \times S$ matrix.

Algorithm 5 UCRL-VTR with linear transition model

```
Input: MDP, $d, H, T = KH$;
Initialize: $M_{1,1} \leftarrow H^2 dI$, $w_{1,1} \leftarrow 0 \in \mathbb{R}^{d\times 1}$, $\theta_1 \leftarrow M_{1,1}^{-1} w_{1,1}$ for $1 \le h \le H$;
Initialize: $\delta \leftarrow 1/K$, and for $1 \le k \le K$,
    $$\beta_k \leftarrow 16 C_\theta^2 H^2 d \log(1 + Hk) \log^2((k+1)^2 H/\delta);$$
Compute Q-function $Q_{h,1}$ using $\theta_{1,1}$ according to (4.1);
for $k = 1 : K$ do
    Obtain initial state $s_1^k$ for episode $k$;
    for $h = 1 : H$ do
        Choose action greedily by
            $$a_h^k = \arg\max_{a\in\mathcal{A}} Q_{h,k}(s_h^k, a)$$
        and observe the next state $s_{h+1}^k$.
        Compute the predicted value vector:    ▷ Evaluate the expected value of next state
            $$X_{h,k} \leftarrow \mathbb{E}.[V_{h+1,k}(s)|s_h^k, a_h^k] = \sum_{s\in\mathcal{S}} V_{h+1,k}(s) \cdot P.(s|s_h^k, a_h^k).$$
        $y_{h,k} \leftarrow V_{h+1,k}(s_{h+1}^k)$    ▷ Update regression parameters
        $M_{h+1,k} \leftarrow M_{h,k} + X_{h,k} X_{h,k}^\top$
        $w_{h+1,k} \leftarrow w_{h,k} + y_{h,k} \cdot X_{h,k}$
    end for
    Update at the end of episode:    ▷ Update Model Parameters
        $$M_{1,k+1} \leftarrow M_{H+1,k},$$
        $$w_{1,k+1} \leftarrow w_{H+1,k},$$
        $$\theta_{k+1} \leftarrow M_{1,k+1}^{-1} w_{1,k+1};$$
    Compute $Q_{h,k+1}, h = H, \ldots, 1$, using $\theta_{k+1}$ according to (A.12)    ▷ Computing Q functions
end for
```

By modifying the algorithm and using optimistic Q-update, we obtain an algorithm that can be implemented using efficient recursive update. See Algorithm 5 for full details of implementation.

Estimating θ_* by recursive regression. We let $X_{h,k}^\top \theta \doteq \mathbb{E}_\cdot[V_{h+1,k}(s)|s_h^k, a_h^k]^\top \theta = \langle P_\theta(\cdot|s,a), V_{h+1,k}\rangle$ be the predicted expected value of next state. In this case, each new observation adds the following loss to regression:

$$\begin{aligned}&\left(X_{h,k}^\top \theta - y_{h,k}\right)^2 :\\ &= \left(\mathbb{E}_\cdot[V_{h+1,k}(s)|s_h^k, a_h^k]^\top \theta - V_{h+1,k}(s_{h+1}^k)\right)^2\end{aligned}$$

By aggregating the value prediction losses constructed from all past experiences, we formulate a ridge regression problem to estimate θ_* by

$$\begin{aligned}&\theta_{k+1}\\ &= \arg\min_{\theta \in \mathbb{R}^d} \left[\theta^\top M_{1,1}\theta + \sum_{(h',k')\le(H,k)} \left(X_{h',k'}^\top \theta - y_{k',h'}\right)^2\right],\end{aligned}$$

where $M_{1,1} = H^2 dI$ acts as a regularization term.

To solve the above regression problem, we can first calculate $X_{h',k'}$ and recursively compute estimates of θ_* by letting

$$\begin{aligned}M_{1,k+1} &= M_{1,1} + \sum_{(h',k')\le(H,k)} X_{h',k'}X_{h',k'}^\top\\ w_{1,k+1} &= w_{1,1} + \sum_{(h',k')\le(H,k)} y_{h',k'} \cdot X_{h',k'},\end{aligned}$$

with $M_{1,1} = H^2 d \cdot I$ and $w_{1,1} = 0$. Then we obtain the estimated θ_{k+1} easily by

$$\theta_{k+1} = M_{1,k+1}^{-1} w_{k+1}.$$

Confidence ball. We construct B_k as follows:

$$B_k = \{\theta | (\theta - \theta_k)^\top M_k(\theta - \theta_k) \le \beta_k\}.$$

where β_k is preselected (see the algorithm).

Our model parameter update, θ_k and M_k, can be via a recursive update in an incremental fashion. In this way, one does not need to re-train the model parameter from scratch every episode. A similarly simple recursion was used in Jin et al., 2020 for model-free Q learning. Our method differs in that our Q functions cannot be parameterized by d parameters and our updates are made on the transition model rather than Q functions.

Optimistic Q-update. Instead of solving the optimistic planning problem $\theta_k = \text{argmax}_\theta \{V_\theta^*(s_1) | \theta \in B_k\}$ as in Algorithm 1, we incorporate optimism into iterative Q-update:

$$
\begin{aligned}
&Q_{H+1,k}(s,a) = 0, \\
&V_{h,k}(s) = \max_{a \in \mathcal{A}} Q_{h,k}(s,a), \\
&Q_{h,k}(s,a) = r(s,a) + \max_{\theta \in B_k} \sum_{j=1}^{d} (\theta)_j P_j(\cdot|s,a) V_{h+1,k}.
\end{aligned}
$$

Since the confidence sets are ellipsoids, the preceding Q update has a closed-forms solution

$$
\begin{aligned}
&Q_{h,k}(s,a) \\
&= r(s,a) + \max_{\theta \in B_k} \langle P_\theta(\cdot|s,a), V_{h+1,k} \rangle \\
&= r(s,a) + X_{h,k}^\top \theta_k + \sqrt{\beta_k} \sqrt{X_{h,k}^\top M_k^{-1} X_{h,k}}.
\end{aligned}
\tag{A.12}
$$

The last term in the above is the “bonus” term that quantifies uncertainty and encourages exploration. This optimistic Q value allows us to greedily pick actions while sufficiently exploring the state space.

Algorithm 5 is a modification of UCRL-VTR and uses a different construction of confidence set. we provide an independent regret analysis using techniques from linear bandit theory. The next theorem gives a egret upper bound for Algorithm 5.

Theorem 21. *Let Assumption 8 hold. If we choose*

$$
\beta_k = \left(H \sqrt{d \log \left(\frac{1 + Hk \cdot H^2 d}{\delta} \right)} + C_\theta H \sqrt{d} \right)^2,
$$

then T-time-step regret of Algorithm 1 satisfies

$$\mathbb{E}\left[R(T)\right] = \widetilde{\mathcal{O}}\left(C_\theta \cdot d\sqrt{H^3T}\right),$$

where $T = HK$ is the total number of steps in K episodes, C_θ ($C_\theta \geq 1$) is a known constant such that $\|\theta_*\| \leq C_\theta$ and $\widetilde{\mathcal{O}}$ hides polylog factors of H, T.

Let us outline the proof ideas. In the first part of the proof, we show that if $\theta_* \in B_{h,k}$, then the estimated Q-functions are optimistic estimates of the true Q-value functions. That is, $Q_{h,k}(s)$ is greater than the true Q-value $Q_h(s)$ for every $s \in \mathcal{S}$. Using this fact, we can bound the regret by the sum of $Q_{1,k}(s_1^k) - Q_1^{\pi_k}(s_1^{\pi_k})$, which can be decomposed into the sum of state-action confidence bounds on the sample path. In the second part, we construct martingale difference sequences and apply a concentration argument to show that $\theta_* \in B_{h,k}$ for all (h,k) with high probability. The full proof is deferred to the Appendix A.5.

A.4 Proof of Theorem 8

In this section, we will present the full proof of Theorem 8. To handle the mispecification error, we will modify the bonus term by replacing it with

$$\beta_k = 8H^2 \log\left(\frac{2\mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|_\infty)}{\delta}\right) + 4H(kH-1)\alpha\left\{2 + \sqrt{\log\left(\frac{4kH(kH-1)}{\delta}\right)}\right\} + 8H^3k\varepsilon^2.$$

The last term in the above choice of β_k can be viewed as an "error tolerance."

Next we show that $P^* \in B_k$ with high probability.

We first present a theorem which is nearly identical to Theorem 14 but tolerates misspecification. We use the same notations as in the proof of Theorem 14.

Theorem 22. *Let $\mathbb{F}$ be the filtration defined above and assume that the functions in $\mathcal{F}$ and also f_* are all bounded by the positive constant $C > 0$ at values*

X_t for all t. Assume that there exists $\widetilde{f} \in \mathcal{F}$ such that $|\widetilde{f}(X) - f_*(X)| \leq \zeta$ for all $X = (s, a, v)$ with $\|v\|_\infty \leq H$, and also for each $s \geq 1$, $(Y_p - f_*(X_p))_p$ is conditionally σ-subgaussian given $\mathbb{F}_{p-1}$. We define

$$\widehat{f_t} = \arg\min_{f \in \mathcal{F}} \sum_{p=1}^{t} (f(X_p) - Y_p)^2$$

and

$$\mathcal{F}_t(\beta) = \left\{ f : \mathcal{X} \to \mathbb{R}, s.t. \sum_{p=1}^{t} (f(X_p) - \widehat{f}(X_p))^2 \leq \beta \right\}.$$

Then, for any $\alpha > 0$, with probability $1 - \delta$, for all $t \geq 1$, $f_* \in \mathcal{F}_t(\beta_t(\delta, \alpha))$, where

$$\beta_t(\delta, \alpha) = 16\sigma^2 \log(4N_\alpha/\delta) + 4t\alpha \left(C + \sqrt{\sigma^2 \log(8t(t+1)/\delta)} \right) + 3t\zeta^2.$$

Note that here the last term is due to the misspecification error.

Proof The proof of this theorem is also nearly identical to Theorem 14, except for the modifications below. Due to model misspecification, we no longer have $f_* \in \mathcal{F}$, and hence we may not have $\|Y - \widehat{f_t}\|_t \leq \|Y - f_*\|_t$ (Here notation $\|\cdot\|_t$ is defined to be the same as the notations in Theorem 14). To handle the misspecification error, we will use the function $\widetilde{f}$ as a bridge to bound the error between $\widehat{f_t}$ and f_*. Hence since $\widehat{f_t} = \arg\min_{f \in \mathcal{F}} \|Y - f\|_t^2$, we have $\|\widehat{f_t} - Y\|_t^2 \leq \|\widetilde{f} - Y\|_t^2$, which indicates that $\|\widehat{f_t} - f_* - Z\|_t^2 \leq \|\widetilde{f} - f_* - Z\|_t^2$. (Recall the notations $Z_p = Y_p - f_*(X_p)$ and $Z = (Z_1, \cdots, Z_p)$.) Therefore, we have

$$\|\widehat{f_t} - f_*\|_t^2 - 2\langle \widehat{f_t} - f_*, Z \rangle_t \leq \|\widetilde{f} - f_*\|_t^2 - 2\langle \widetilde{f} - f_*, Z \rangle_t.$$

We then obtain

$$\begin{aligned} \frac{1}{2}\|\widehat{f_t} - f_*\|_t^2 &\leq -\frac{1}{2}\|\widehat{f_t} - f_*\|_t^2 + 2\langle \widehat{f_t} - f_*, Z \rangle_t + \|\widetilde{f} - f_*\|_t^2 - 2\langle \widetilde{f} - f_*, Z \rangle_t \\ &= E(\widehat{f_t}) + \widetilde{E}(\widetilde{f}) + \frac{3}{2}\|\widetilde{f} - f_*\|_t^2, \end{aligned} \tag{A.13}$$

where we define

$$E(f) = -\frac{1}{2}\|f - f_*\|_t^2 + 2\langle Z, f - f_* \rangle_t \tag{A.14}$$

$$\widetilde{E}(f) = -\frac{1}{2}\|f - f_*\|_t^2 - 2\langle Z, f - f_* \rangle_t \tag{A.15}$$

Next, we will bound $E(\widehat{f}_t)$ and also $\widetilde{E}(\widetilde{f})$. Similar to the proof of Theorem 14, we can show that

$$E(\widehat{f}_t) \leq 4\sigma^2 \log(|N_\alpha|/\delta) + 2C\alpha\, t + 2t\alpha\sqrt{\sigma^2 \log(2t(t+1)/\delta)},$$

holds with probability at least $1-\delta$, where N_α is the α-covering number of $\mathcal{F}$.

Now we analyze $\widetilde{E}(\widetilde{f})$ where $\widetilde{f} \in \mathcal{F}$. Similarly, a simple calculation gives that for any fixed $f \in \mathcal{F}$, $2\langle -Z, f - f_*\rangle_t$ is $2\sigma\|f - f_*\|_t$-subgaussian. Hence, with probability $1-\delta$, simultaneously for all $t \geq 1$,

$$\widetilde{E}(f) \leq -\frac{1}{2}\|f_* - f\|_t^2 + 4\sigma^2 \log\left(\frac{1}{\delta}\right) + \frac{1}{4\sigma^2} \cdot \frac{4\sigma^2\|f - f_*\|_t^2}{2} = 4\sigma^2 \log\left(\frac{1}{\delta}\right),$$

which indicates that with probability at least $1-\delta$, we have

$$\widetilde{E}(\widetilde{f}) \leq 4\sigma^2 \log\left(\frac{1}{\delta}\right).$$

Finally, as for the last term $\|\widetilde{f} - f_*\|_t^2$ in (A.13), we have the following estimation due to the bound of the misspecification error:

$$\|\widetilde{f} - f_*\|_t^2 = \sum_{p=1}^{t}(\widetilde{f}(X_p) - \widetilde{f}(X_p))^2 \leq t \cdot \zeta^2,$$

where we use the fact that $X_p = (s_p, a_p, v_p)$ satisfies that $\|v_p\|_\infty \leq H$.

We combine those bounds on the three terms in (A.13) above, and obtain that with probability at least $1-2\delta$, the following inequality holds:

$$\frac{1}{2}\|\widehat{f} - f_*\|_t^2 \leq 2tC\alpha + 2t\alpha\sqrt{\sigma^2 \log(2t(t+1)/\delta)} + 8\sigma^2 \log(2N_\alpha/\delta) + \frac{3}{2}t\zeta^2.$$

Finally, we switch δ into $\delta/2$ and multiply the above inequality by 2 on both sides. And the proof of Theorem 22 is completed. Next we apply this theorem to prove the following lemma:

Lemma 23. *For any transition model P, we define its corresponding function $f_p : \mathcal{X} \to \mathbb{R}$:*

$$f_P(s, a, v) = \int P(ds'|s, a)v.$$

Then with probability at least $1-\delta$*, we have* $P_* \in \mathcal{B}_k$*, where*

$$\mathcal{B}_k = \left\{ \widetilde{P} \in \mathcal{P} : \sum_{p=1}^{t} (f_{\widetilde{P}}(X_t) - f_{\widehat{P}_t}(X_t))^2 \le \beta_k \right\}, \quad t = k(H-1).$$

Here $\widehat{P}_t$ *is defined in* (A.1) *and we choose*

$$\beta_k = 8H^2 \log\left(\frac{4\mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|_\infty)}{\delta}\right) + 4H(kH-1)\alpha\left\{2 + \sqrt{\log\left(\frac{8kH(kH-1)}{\delta}\right)}\right\} + 8H^3 k\varepsilon^2,$$

Proof In the following proof, the notation of $X_t, Y_t, \mathcal{F}$ are the same as the proof of Corollary 15. We notice that $Y_t - f_P(X_t) \in [-H, H]$ for every $X_t = (s_t, a_t, V_t)$, and

$$\mathbb{E}[Y_t|\mathbb{F}_t] = \mathbb{E}[V_t(s_{t+1})|\mathbb{F}_t] = \int P(ds'|s_t, a_t)V_t(s') = f_P(s_t, a_t, v_t) = f_P(X_t).$$

Hence $Z_t = Y_t - f_P(X_t)$ is $\frac{H}{2}$-subgaussian given $\mathbb{F}_t$.

For every $f \in \mathcal{F}$, there exists some $\widetilde{P} \in \mathcal{P}$ such that $f(s,a,v) = \int \widetilde{P}(ds'|s,a)v(s')$, which indicates that $|f(X_t)| \le H$. Moreover, we also have $|f_P(X_t)| \le H$.

We next apply Theorem 22 with $C = H$ and $\sigma = \frac{H}{2}$ and $f_* = f_P$ and $\zeta = H\epsilon$ and $\widetilde{f} = f_{P^*}$. According to Assumption 2, we notice that, for all $X = (s,a,v)$ with $\|v\|_\infty \le H$, we have

$$|f_*(X) - \widetilde{f}(X)| = \left| \int (P(s'|s,a) - P^*(s'|s,a))v(s')ds' \right| \le \|P(s'|s,a) - P^*(s'|s,a)\|_1 \|v\|_\infty \le H\varepsilon = \zeta.$$

Hence we have verified all the assumptions in Theorem 22. Hence we obtain that: for any $\alpha > 0$, with probability at least $1-\delta$, for all $t \ge 1$, we have

$$\sum_{p=1}^{t} (f_P(X_p) - f_{\widehat{P}_t}(X_p))^2 \le 4H^2 \log\left(\frac{4\mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|_\infty)}{\delta}\right) + 2H(kH-1)\alpha\left\{2 + \sqrt{\log\left(\frac{8kH(kH-1)}{\delta}\right)}\right\} + 3H^3 k\varepsilon^2.$$

Moreover, noticing that

$$(f_P(X_t) - f_{P_*}(X_t))^2 = \left(\int (P(ds'|s_t, a_t) - P^*(ds'|s_t, a_t))V_t \right)^2 \le (H\varepsilon)^2,$$

we have

$$\sum_{p=1}^{t}(f_{P^*}(X_p) - f_{\widehat{P}_t}(X_p))^2 \le \sum_{p=1}^{t} 2(f_P(X_p) - f_{P^*}(X_p))^2 + 2(f_P(X_p) - f_{\widehat{P}_t}(X_p))^2$$

$$\le 8H^2 \log\left(\frac{4\mathcal{N}(\mathcal{F},\alpha,\|\cdot\|_\infty)}{\delta}\right) + 4H(kH-1)\alpha\left\{2+\sqrt{\log\left(\frac{8kH(kH-1)}{\delta}\right)}\right\} + 6H^3k\varepsilon^2 + 2H^2\varepsilon^2$$

$$\le 8H^2 \log\left(\frac{4\mathcal{N}(\mathcal{F},\alpha,\|\cdot\|_\infty)}{\delta}\right) + 4H(kH-1)\alpha\left\{2+\sqrt{\log\left(\frac{8kH(kH-1)}{\delta}\right)}\right\} + 8H^3k\varepsilon^2.$$

which indicates that $P^* \in \mathcal{B}_k$. This finishes the proof of this corollary.

We now provide a lemma similar to Lemma 13, only adding the misspecification analysis.

Lemma 24. *Assuming that $P^* \in \mathcal{B}_k$, we have*

$$V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \le \sup_{\widetilde{P}\in\mathcal{B}_k} \sum_{h=1}^{H-1} \langle \widetilde{P}(\cdot|s_h^k, a_h^k) - P^*(\cdot|s_h^k, a_h^k), V_{h,k}\rangle + \sum_{h=1}^{H-1} \xi_{h+1,k} + H^2\varepsilon\,,$$

where

$$\xi_{h+1,k} = \langle P(\cdot|s_h^k, a_h^k), V_{h+1,k} - V_{h+1}^{\pi_k}\rangle - \left(V_{h+1,k}(s_{h+1}^k) - V_{h+1}^{\pi_k}(s_{h+1}^k)\right)\ .$$

Note that $(\xi_{2,1}, \xi_{3,1}, \ldots, \xi_{H,1}, \xi_{2,2}, \xi_{3,2}, \ldots, \xi_{H,2}, \xi_{2,3}, \ldots)$ is a sequence of martingale differences.

Proof We first prove by induction that

$$V_{h,k}(s_h^k) \ge V_h^*(s_h^k) - (H+1-h)\varepsilon, \quad \forall 1 \le h \le H+1$$

by induction on h according to the fact that $P^* \in \mathcal{B}_k$ (but not $P \in \mathcal{B}_k$). When $h = H+1$, this inequality holds since both sides equal to 0. We assume it holds for $h+1$ and we consider the case of h. Actually we have

$$\begin{aligned} Q_{h,k}(s_h^k) &= r(s_h^k, a_h^k) + \langle P^k(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle \ge r(s_h^k, a_h^k) + \langle P^*(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle \\ &= r(s_h^k, a_h^k) + \langle P(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle - \langle P(\cdot|s_h^k, a_h^k) - P^*(\cdot|s_h^k, a_h^k), V_{h+1,k})\rangle \\ &\ge r(s_h^k, a_h^k) + \langle P(\cdot|s_h^k, a_h^k), V_{h+1}^* - (H-h)\varepsilon\mathbf{1}\rangle - \|P(\cdot|s_h^k, a_h^k) - P^*(\cdot|s_h^k, a_h^k)\|_1 \|V_{h+1,k}\|_\infty \\ &\ge r(s_h^k, a_h^k) + \langle P(\cdot|s_h^k, a_h^k), V_{h+1}^*\rangle - (H+1-h)\xi = Q_h^*(s_h^k, a_h^k) - (H+1-h)\xi, \end{aligned}$$

where in the third line we use the induction and in the last line we use the fact that $\|V_{h+1,k}\|_\infty \leq H$. This indicates that $V_{h,k}(s_h^k) \geq V_h^*(s_h^k) - (H+1-h)\varepsilon$, which completes the induction at h. Hence we know that $V_{h,k}(s_h^k) \geq V_h^*(s_h^k) - (H+1-h)\varepsilon$ holds for all $1 \leq h \leq H+1$.

Therefore,

$$V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \leq V_{1,k}(s_1^k) - V_1^{\pi_k}(s_1^k) + H\varepsilon\,.$$

Fix $h \in [H]$. In what follows we bound $V_{h,k}(s_h^k) - V_h^{\pi_k}(s_h^k)$. By the definition of π_k, P^k and a_h^k, we have

$$\begin{aligned} V_{h,k}(s_h^k) &= r(s_h^k, a_h^k) + \langle P^k(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle \text{ and} \\ V_h^{\pi_k}(s_h^k) &= r(s_h^k, a_h^k) + \langle P(\cdot|s_h^k, a_h^k), V_{h+1}^{\pi_k}\rangle\,. \end{aligned}$$

Hence,

$$\begin{aligned} V_{h,k}(s_h^k) - V_h^{\pi_k}(s_h^k) &= \langle P^k(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle - \langle P_{a_h^k}(s_h^k), V_{h+1}^{\pi_k}\rangle \\ &= \langle P^k(\cdot|s_h^k, a_h^k) - P(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle + \langle P(\cdot|s_h^k, a_h^k), V_{h+1,k} - V_{h+1}^{\pi_k}\rangle\,. \end{aligned}$$

Therefore, by induction, noting that $V_{H+1,k} = 0$, we get that

$$\begin{aligned} V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) &\leq \sum_{h=1}^{H-1} \langle P^k(\cdot|s_h^k, a_h^k) - P(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle + \sum_{h=1}^{H-1} \xi_{h+1,k} + H\varepsilon \\ &\leq \sup_{\widetilde{P} \in \mathcal{B}_k} \sum_{h=1}^{H-1} \langle \widetilde{P}(\cdot|s_h^k, a_h^k) - P(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle + \sum_{h=1}^{H-1} \xi_{h+1,k} + H\varepsilon\,. \end{aligned}$$

Finally noticing that

$$\begin{aligned} \langle \widetilde{P}(\cdot|s_h^k, a_h^k) - P(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle &= \langle \widetilde{P}(\cdot|s_h^k, a_h^k) - P^*(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle + \langle \widetilde{P}^*(\cdot|s_h^k, a_h^k) - P(\cdot|s_h^k, a_h^k), V \\ &\leq \langle \widetilde{P}(\cdot|s_h^k, a_h^k) - P^*(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle + H\varepsilon, \end{aligned}$$

we have

$$\begin{aligned} V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) &\leq \sum_{h=1}^{H-1} \langle \widetilde{P}(\cdot|s_h^k, a_h^k) - P^*(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle + \sum_{h=1}^{H-1} \xi_{h+1,k} + \frac{H(H-1)}{2}\xi + H\varepsilon \\ &\leq \sum_{h=1}^{H-1} \langle \widetilde{P}(\cdot|s_h^k, a_h^k) - P^*(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle + \sum_{h=1}^{H-1} \xi_{h+1,k} + H^2\varepsilon, \end{aligned}$$

which completes the proof of this lemma.

Equipped with these two lemmas, we are ready to prove Theorem 8.

Proof of Theorem 8. According to Lemma 23, we learn that $P^* \in \mathcal{B}_k$ holds with probability at least $1-\delta$. We next assume $P^* \in \mathcal{B}_k$ and bound the error $V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k)$. According to Lemma 24, we have

$$V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \leq \sup_{\widetilde{P} \in \mathcal{B}_k} \sum_{h=1}^{H-1} \langle \widetilde{P}(\cdot|s_h^k, a_h^k) - P^*(\cdot|s_h^k, a_h^k), V_{h,k} \rangle + \sum_{h=1}^{H-1} \xi_{h+1,k} + H^2 \varepsilon\,, \tag{A.16}$$

where

$$\xi_{h+1,k} = \langle P(\cdot|s_h^k, a_h^k), V_{h+1,k} - V_{h+1}^{\pi_k} \rangle - \left(V_{h+1,k}(s_{h+1}^k) - V_{h+1}^{\pi_k}(s_{h+1}^k) \right)\,.$$

We let

$$W_k = \sup_{\widetilde{P} \in \mathcal{B}_k} \sum_{h=1}^{H-1} \langle \widetilde{P}(\cdot|s_h^k, a_h^k) - P^*(\cdot|s_h^k, a_h^k), V_{h,k} \rangle,$$

and summing h from 1 to H in (A.16) we obtain the following bound on the regret up to horizon K:

$$R_K = \sum_{k=1}^K V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \leq \sum_{k=1}^K W_k + \sum_{k=1}^K \sum_{h=1}^{H-1} \xi_{h+1,k} + H^2 K \varepsilon$$

We next bound $\sum_{k=1}^K W_k$. Actually we have

$$\begin{aligned} \sum_{k=1}^K W_k &= \sum_{k=1}^K \sup_{\widetilde{P} \in \mathcal{B}_k} \sum_{h=1}^{H-1} \langle \widetilde{P}(\cdot|s_h^k, a_h^k) - P^*(\cdot|s_h^k, a_h^k), V_{h,k} \rangle \\ &\leq \sum_{k=1}^K \sum_{h=1}^{H-1} \sup_{\widetilde{P} \in \mathcal{B}_k} \langle \widetilde{P}(\cdot|s_h^k, a_h^k) - P^*(\cdot|s_h^k, a_h^k), V_{h,k} \rangle \end{aligned},$$

and each term inside satisfies

$$\sup_{\widetilde{P} \in \mathcal{B}_k} \langle \widetilde{P}(\cdot|s_h^k, a_h^k) - P^*(\cdot|s_h^k, a_h^k), V_{h,k} \rangle \leq \operatorname{diam}(\widetilde{\mathcal{F}}_t|_{X_t})$$

where

$$\widetilde{\mathcal{F}}_t = \left\{ f = f_P : P \in \mathcal{P}, \sum_{p=1}^t (f(X_p) - f_{\widehat{P}_t}(X_p))^2 \leq \beta_k \right\}$$

We notice that $\mathcal{F}_t \subset \mathcal{F} \subset \mathcal{B}_\infty(\mathcal{X}, H)$. Hence we apply Lemma 17 and obtain that

$$\sum_{k=1}^{K} \mathrm{diam}(\widetilde{\mathcal{F}}_t|_{X_t}) \le \alpha + H(d \wedge K(H-1)) + 2\delta_{K(H-1)}\sqrt{dK(H-1)},$$

where $\delta_{K(H-1)} = \max_{1\le t\le K(H-1)} \mathrm{diam}(\widetilde{\mathcal{F}}_t|_{X_t})$. Thanks to the definition of $\widetilde{\mathcal{F}}_t$, $\delta_{K(H-1)} \le 2\sqrt{\beta_K}$. Plugging this into the previous display finishes the proof.

Moreover, we also have $\sum_{k=1}^{K}\sum_{h=1}^{H-1} \xi_{h+1,k} \le H\sqrt{2K(H-1)\log(1/\delta)}$ holds with probability at least $1-\delta$. Hence combine these two inequality together, we obtain that with probability at least $1-2\delta$, the following bound holds

$$\begin{aligned}R_K &= \sum_{k=1}^{K} V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k)\\ &\le \sum_{k=1}^{K} \sup_{\widetilde{P}\in\mathcal{B}_k} \sum_{h=1}^{H-1} \langle \widetilde{P}_{a_h^k}(s_h^k) - P_{a_h^k}(s_h^k), V_{h,k}\rangle + \sum_{k=1}^{K}\sum_{h=1}^{H-1} \xi_{h+1,k} + H^2K\varepsilon\\ &\le \alpha + H(d\wedge K(H-1)) + 4\sqrt{d\beta_K K(H-1)} + H\sqrt{2K(H-1)\log(1/\delta)} + H^2K\varepsilon,\end{aligned}$$

where we use $\alpha \le 1$. □

A.5 Proof of Theorem 21

Here we will provide the formal regret analysis for Algorithm 5, which differs from Algorithm 1. By leveraging the linear structure, we provide an independent proof of Theorem 21 using analysis adapted from the linear bandit literature (Abbasi-Yadkori et al., 2011; Dani et al., 2008).

The full proof is divided into five parts, which each subsection containing a part of the proof. In the first subsection, we decompose the regret into the sum of bonuses assuming the Q-functions are optimistic. In the second subsection, we detail some important properties of our algorithm. In the third subsection, we provide an upper bound on the sum of bonuses introduced in the first subsection. In the fourth subsection, we prove that optimism holds with high probability by constructing a martingale difference sequence and showing that it concentrates. In the final subsection, we will put together all the analysis to finish the proof of upper bound of expected regret.

We say $(h, k) \leq (h', k')$ if $k < k'$ or $k = k', h \leq h'$. Thus, $\leq$ stands for the lexicographic order with k being the variable that takes priority. We say $(h, k) < (h', k')$ if $k < k'$ or $k = k', h < h'$.Let $\mathbb{F}_{h,k}$ be the filtration generated by the random sample path $\{(s_{h'}^{k'}, a_{h'}^{k'}, r_{h'}^{k'})\}_{(h',k')\leq(h,k)}$.

A.5.1 Regret Analysis

The proof in this section is similar to Lemma 13. Throughout A.5.1 to A.5.3, we assume that $\theta_* \in B_k$ for all $1 \leq k \leq K$. Then in A.5.4 we will prove that that the event, $\theta_* \in B_k$ for all $1 \leq k \leq K$, holds with high probability.

Optimism

We show by induction that $Q_h^*(s, a) \leq Q_{h,k}(s, a)$ holds for all (s, a), h and k. When $h = H + 1$, this inequality trivially holds, since both sides of the inequality equal 0. Next suppose that this inequality holds for some $h+1 \leq H$. As a result, we have

$$V_{h+1}^*(s) = \prod\nolimits_{[0,H]} \left[\max_{a\in\mathcal{A}} Q_{h+1}^*(s, a)\right] \leq \prod\nolimits_{[0,H]} \left[\max_{a\in\mathcal{A}} Q_{h+1,k}(s, a)\right] = V_{h+1,k}(s),$$

which indicates that

$$\begin{aligned} Q_h^*(s, a) &= r(s, a) + P(\cdot|s, a)^\top V_{h+1}^* \leq r(s, a) + P(\cdot|s, a)^\top V_{h+1,k} \\ &= r(s, a) + \sum_{j=1}^d (\theta_*)_j P_j(\cdot|s, a)^\top V_{h+1,k} \leq r(s, a) + \max_{\theta\in B_k} \left[\sum_{j=1}^d (\theta)_j P_j(\cdot|s, a)^\top V_{h+1,k}\right] \\ &= Q_{h,k}(s, a). \end{aligned}$$

This completes the induction.

Regret Decomposition

Denote π_k to be the stationary policy used in the k episode and let

$$\bar{\theta}_{h,k}(s, a) = \arg\max_{\theta\in B_k} \sum_{j=1}^d (\theta)_j P_j(\cdot|s, a)^\top V_{h+1,k}.$$

We use the fact that $\pi_k(s_h^k) = a_h^k$ and $\theta_* \in B_k$. Now let ξ_{h+1}^k be

$$\xi_{h+1}^k := P(\cdot|s_h^k, a_h^k)^\top (V_{h+1,k} - V_{h+1}^*) - \left[V_{h+1,k}(s_{h+1}^k) - V_{h+1}^*(s_{h+1}^k)\right],$$

we have

$$\begin{aligned}
V_{h,k}(s_h^k) - V_h^{\pi_k}(s_h^k) &= Q_{h,k}(s_h^k, a_h^k) - Q_h^{\pi_k}(s_h^k, a_h^k) \\
&= r(s_h^k, a_h^k) + \bar{\theta}_{h,k}(s_h^k, a_h^k)^\top P.(\cdot|s_h^k, a_h^k) V_{h+1,k} - r(s_h^k, a_h^k) - \theta_*^\top P.(\cdot|s_h^k, a_h^k) V_{h+1}^{\pi_k} \\
&= \left[\theta_* + \bar{\theta}_{h,k}(s_h^k, a_h^k) - \theta_k + \theta_k - \theta_*\right]^\top P.(\cdot|s_h^k, a_h^k) V_{h+1,k} - \theta_*^\top P.(\cdot|s_h^k, a_h^k) V_{h+1}^{\pi_k} \\
&\le \theta_*^\top P.(\cdot|s_h^k, a_h^k)(V_{h+1,k} - V_{h+1}^{\pi_k}) + 2 \max_{\theta \in B_k} \left|(\theta - \theta_k)^\top P.(\cdot|s_h^k, a_h^k) V_{h+1,k}\right| \\
&\le P(\cdot|s_h^k, a_h^k)^\top (V_{h+1,k} - V_{h+1}^{\pi_k}) + 2 \max_{\theta \in B_k} \sqrt{(\theta - \theta_k)^\top M_k (\theta - \theta_k)} \\
&\qquad \times \sqrt{\left[P.(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]^\top M_k^{-1} \left[P.(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]} \\
&\le V_{h+1,k}(s_{h+1}^k) - V_{h+1}^{\pi_k}(s_{h+1}^k) + \xi_{h+1}^k \\
&\qquad + 2\sqrt{\beta_k} \cdot \sqrt{\left[P.(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]^\top M_k^{-1} \left[P.(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]},
\end{aligned}$$

where the first inequality uses the fact that $\theta_*, \overline{\theta}_{h,k} \in B_k$, the second inequality uses the Cauchy-Schwarz inequality and the third inequality uses the definition of B_k.

Now recall that $V_{h+1,k}(s) = V_{H+1}^*(s) = 0$ for any $s \in \mathcal{S}$. We apply the preceding inequality recursively and obtain

$$\begin{aligned}
V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) &\le V_{1,k}(s_1^k) - V_1^{\pi_k}(s_1^k) \quad \text{(by optimism of value estimates)} \\
&\le \sum_{h=1}^H \xi_{h+1}^k + 2 \sum_{h=1}^H \sqrt{\beta_k} \cdot \sqrt{\left[P.(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]^\top M_k^{-1} \left[P.(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]},
\end{aligned}$$

therefore the expected regret can be bounded by if we bound the expectation of

$$\begin{aligned}
\widehat{R}(K) &= \sum_{k=1}^K \left[V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k)\right] \\
&\le \sum_{k=1}^K \sum_{h=1}^H \xi_{h+1}^k + 2 \sum_{k=1}^K \sum_{h=1}^H \sqrt{\beta_k} \cdot \sqrt{\left[P.(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]^\top M_k^{-1} \left[P.(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]}.
\end{aligned} \tag{A.17}$$

Moreover, observe that

$$\mathbb{E}\left[\xi_{h+1}^k \middle| \mathbb{F}_{h,k}\right] = 0,$$

therefore ξ_{h+1}^k is a martingale difference sequence w.r.t. $\mathcal{F}_{h,k}$. Since $V_h^*(s_h^k), V_{h,k}(s_h^k) \in [0, H]$ and $P(\cdot|s_h^k, a_h^k)$ is a probability distribution over the state space, we have $|\xi_h^k| \leq H$ with probability 1. By the Azuma-Hoeffding inequality, with probability at least $1-\delta$, the following inequality holds

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \xi_{h+1}^k \leq \sqrt{2H^3 K \log(1/\delta)}. \tag{A.18}$$

It remains to analyze the second term of (A.17), ie., the sum of bonus given by

$$2 \sum_{k=1}^{K} \sum_{h=1}^{H} \sqrt{\beta_k} \cdot \sqrt{\left[P.(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]^\top M_k^{-1} \left[P.(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]}.$$

A.5.2 Some Properties of Algorithm 5

In this subsection we establish several useful properties of our algorithm, assuming that optimism holds throughout.

Note that

$$M_{h,k} = M_{1,1} + \sum_{(h',k')<(h,k)} \left[P.(\cdot|s_{h'}^{k'}, a_{h'}^{k'}) V_{h'+1,k'}\right] \left[P.(\cdot|s_{h'}^{k'}, a_{h'}^{k'}) V_{h'+1,k'}\right]^\top.$$

Denote

$$l_{h,k} \doteq \sqrt{\left[P.(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]^\top M_{h,k}^{-1} \left[P.(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]}.$$

Let $(h,k)+1$ denote the tuple index of the next time step after (h,k), that is $(h+1,k)$ if $h < H$ and $(h, k+1)$ otherwise. Thus $\{M_{h,k}\}$ satisfies $M_{1,k} = M_{H+1,k-1}$. We now have

$$\begin{aligned} M_{(h,k)+1}^{-1} &= \left(M_{h,k} + \left[P.(\cdot|s_h^k, a_h^k) V_{h+1,k}\right] \left[P.(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]^\top\right)^{-1} \\ &= M_{h,k}^{-1} - \frac{M_{h,k}^{-1} \left[P.(\cdot|s_h^k, a_h^k) V_{h+1,k}\right] \left[P.(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]^\top M_{h,k}^{-1}}{1 + \left[P.(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]^\top M_{h,k}^{-1} \left[P.(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]}. \end{aligned}$$

Which implies

$$\left[P.(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]^\top M_{(h,k)+1}^{-1} \left[P.(\cdot|s_h^k, a_h^k) V_{h+1,k}\right] = l_{h,k}^2 - \frac{l_{h,k}^2 \cdot l_{h,k}^2}{1 + l_{h,k}^2} = \frac{l_{h,k}^2}{1 + l_{h,k}^2}.$$

Next, we derive an upper bound to the quantity

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{l_{h,k}^2}{1+l_{h,k}^2}.$$

Since

$$M_{(h,k)+1} = M_{h,k} + \left[P_{\cdot}(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]\left[P_{\cdot}(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top,$$

we have that

$$\begin{aligned}\det M_{(h,k)+1} &= \det M_{h,k}\det\Big(I + M_{h,k}^{-1/2}\left[P_{\cdot}(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]\left[P_{\cdot}(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top M_{h,k}^{-1/2}\Big)\\ &= \det M_{h,k}\left(1+l_{h,k}^2\right).\end{aligned}$$

Which indicates that

$$\sum_{(h',k')\le(h,k)}\log\left(1+l_{h',k'}^2\right) = \log\det M_{(h,k)+1} - \log\det M_{1,1}.$$

Furthermore, since

$$\frac{l_{h,k}^2}{1+l_{h,k}^2} \le \min\{1, l_{h,k}^2\} \le 2\log\left(1+l_{h,k}^2\right),$$

we have

$$\begin{aligned}\sum_{(h',k')\le(h,k)}\frac{l_{h',k'}^2}{1+l_{h',k'}^2} &\le \sum_{(h',k')\le(h,k)}\min\left\{1, l_{h',k'}^2\right\}\\ &\le \sum_{(h',k')\le(h,k)}2\log(1+l_{h',k'}^2) = 2\log\det M_{(h,k)+1} - 2\log\det M_{1,1}.\end{aligned}$$

Given the initial value $M_{1,1} = H^2 dI$, we have

$$\begin{aligned}\mathbf{tr}(M_{(h,k)+1}) &= \mathbf{tr}(M_{1,1}) + \sum_{(h',k')\le(h,k)}\|P_{\cdot}(\cdot|s_{h'}^{k'}, a_{h'}^{k'})V_{h'+1,k'}\|^2\\ &= H^2d^2 + \sum_{(h',k')\le(h,k)}\sum_{j=1}^{d}\Big(P_j(\cdot|s_{h'}^{k'}, a_{h'}^{k'})V_{h'+1,k'}\Big)^2\\ &\le H^2d^2 + KdH^3,\end{aligned}$$

where the last inequality uses Assumption 8 and the fact that

$$P_j(\cdot|s_{h'}^{k'}, a_{h'}^{k'})V_{h'+1,k'} \le \|P_j(\cdot|s_{h'}^{k'}, a_{h'}^{k'})\|_1\|V_{h'+1,k'}\|_\infty \le H.$$

Using the inequalities of arithmetic and geometric means, we get the following upper bound for the determinant of $M_{(h,k)+1}$:

$$\det M_{(h,k)+1} \leq \left(\frac{\mathrm{tr}(M_{(h,k)+1})}{d}\right)^d \leq (H^2d + KH^3)^d,$$

which indicates that

$$\log\det M_{(H,k)+1} - \log\det M_{1,1} \leq \log\left((H^2d + KH^3)^d\right) - \log\left((H^2d)^d\right) \leq d\log(1 + HK). \tag{A.19}$$

Hence we have

$$\sum_{(h',k')\leq(h,k)} \frac{l_{h',k'}^2}{1 + l_{h',k'}^2} \leq \sum_{(h',k')\leq(h,k)} \min\left\{1, l_{h',k'}^2\right\} \leq 2d\log(1 + HK). \tag{A.20}$$

A.5.3 Sum-of-Bonus Analysis

In this section, under the assumption that $\theta_* \in B_k$ for every k, we establish an upper bound for the following sum-of-bonus term

$$2\sum_{k=1}^K\sum_{h=1}^H \sqrt{\beta_k} \cdot \sqrt{\left[P.(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top M_k^{-1}\left[P.(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]},$$

where we denote $M_k = M_{1,k}$ for simplicity. We let

$$u_{h,k} = \sqrt{\left[P.(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top M_k^{-1}\left[P.(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]}.$$

Since $\beta_k \leq \beta_K$ for any $1 \leq k \leq K$ and by letting

$$\begin{aligned} u_{h,k}^2 &= \left[P.(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top M_k^{-1}\left[P.(\cdot|s_h^k, a_h^k)V_{h+1,k}\right] \\ &\leq \left[P.(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top M_1^{-1}\left[P.(\cdot|s_h^k, a_h^k)V_{h+1,k}\right] \\ &= \frac{1}{H^2d} \cdot \sum_{j=1}^d \left[P_j(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^2 \leq \frac{1}{H^2d} \cdot H^2d = 1, \end{aligned}$$

we have

$$\begin{aligned} &2\sum_{k=1}^K\sum_{h=1}^H \sqrt{\beta_k} \cdot \sqrt{\left[P.(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top M_k^{-1}\left[P.(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]} \\ &\leq 2\sqrt{\beta_K} \cdot \sum_{k=1}^K\sum_{h=1}^H u_{h,k} \leq 2\sqrt{\beta_K} \cdot \sum_{k=1}^K\sum_{h=1}^H \min\{1, u_{h,k}\} \\ &\leq 2\sqrt{HK\beta_K} \cdot \sqrt{\sum_{k=1}^K\sum_{h=1}^H \min\{1, u_{h,k}^2\}} \leq 4\sqrt{HK\beta_K} \cdot \sqrt{\sum_{k=1}^K\sum_{h=1}^H \log(1 + u_{h,k}^2)} \end{aligned} \tag{A.21}$$

where the third inequality uses the Cauchy-Schwarz inequality. Next recall that

$$M_{k+1} = M_k + \sum_{h=1}^{H} \left[P.(\cdot|s_h^k, a_h^k)V_{h+1,k}\right] \left[P.(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top.$$

Now we have

$$\det(M_{k+1}) = \det(M_k)\cdot\det\left(I + \sum_{h=1}^{H} M_k^{-1/2} \left[P.(\cdot|s_h^k, a_h^k)V_{h+1,k}\right] \left[P.(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top M_k^{-1/2}\right).$$

Now notice that every eigenvalue of the matrix

$$I + \sum_{h=1}^{H} M_k^{-1/2} \left[P.(\cdot|s_h^k, a_h^k)V_{h+1,k}\right] \left[P.(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top M_k^{-1/2}$$

is at least 1, and we have the following bound of its trace:

$$\begin{aligned}
&\operatorname{tr}\left(\sum_{h=1}^{H} M_k^{-1/2} \left[P.(\cdot|s_h^k, a_h^k)V_{h+1,k}\right] \left[P.(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top M_k^{-1/2}\right)\\
&= \sum_{h=1}^{H} \left[P.(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top M_k^{-1} \left[P.(\cdot|s_h^k, a_h^k)V_{h+1,k}\right] = \sum_{h=1}^{H} u_{h,k}^2.
\end{aligned}$$

This indicates that

$$\begin{aligned}
&\det\left(I + \sum_{h=1}^{H} M_k^{-1/2} \left[P.(\cdot|s_h^k, a_h^k)V_{h+1,k}\right] \left[P.(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top M_k^{-1/2}\right)\\
&\geq 1 + \operatorname{tr}\left(I + \sum_{h=1}^{H} M_k^{-1/2} \left[P.(\cdot|s_h^k, a_h^k)V_{h+1,k}\right] \left[P.(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top M_k^{-1/2}\right)\\
&= 1 + \sum_{h=1}^{H} u_{h,k}^2,
\end{aligned}$$

where the first inequality follows from the following fact: $\prod_i (1+w_i) \geq 1 + \sum_i w_i$ provided $w_i \geq 0$. Combining the above inequality with the following inequality

$$1 + \sum_{h=1}^{H} u_{h,k}^2 = \frac{\sum_{h=1}^{H}(1 + Hu_{h,k}^2)}{H} \geq \prod_{h=1}^{H}(1 + Hu_{h,k}^2)^{1/H} \geq \prod_{h=1}^{H}(1 + u_{h,k}^2)^{1/H},$$

we obtain that

$$\sum_{h=1}^{H} \log(1 + u_{h,k}^2) \leq H \log\left(1 + \sum_{h=1}^{H} u_{h,k}^2\right) \leq H \det(M_{k+1}) - H \det(M_k).$$

Therefore, we have

$$
\begin{aligned}
&2\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\beta_k}\cdot\sqrt{\left[P.(\cdot|s_h^k,a_h^k)V_{h+1,k}\right]^\top M_k^{-1}\left[P.(\cdot|s_h^k,a_h^k)V_{h+1,k}\right]}\\
&\le 4\sqrt{HK\beta_K}\cdot\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\log(1+u_{h,h}^2)}\\
&\le 4\sqrt{HK\beta_K}\cdot\sqrt{\sum_{k=1}^{K}H\det(M_{k+1})-H\det(M_k)}\\
&\le 4\sqrt{HK\beta_K}\cdot\sqrt{H\det(M_{(H,k)+1})-H\det(M_{1,1})}\\
&\le 4\sqrt{H^2dK\beta_K\log(1+HK)},
\end{aligned}
$$

where the last inequality uses (A.19).

A.5.4 Confidence Sets for Value Targeted Regression

We now adapt a result from (Abbasi-Yadkori et al., 2011). For $t=H(k-1)+h$, we choose

$$
\begin{aligned}
&\lambda = H^2d,\\
&\overline{V}_t = M_{h,k},\\
&S = C_\theta,\\
&R = H,\\
&L = \sqrt{H^2d}.
\end{aligned}
$$

Then we have

$$\theta_{h,k} = (\mathbf{X}_{1:t}^T\mathbf{X}_{1:t}+\lambda I)^{-1}\mathbf{X}_{1:t}\mathbf{Y}_{1:t} = \widehat{\theta}_t, \quad \text{and} \quad \|\theta_*\|_2 \le C_\theta = S.$$

Moreover, since $|\eta_t| = |Y_t - \langle X_t,\theta_*\rangle| = |Y_{h,k} - P(\cdot|s_h^k,a_h^k)^TV_{h+1,k}| \le H$, η_t is H-subgaussian. We can also verify that

$$\|X_t\|_2^2 = \sum_{i=1}^{d}\left(P_i(\cdot|s_h^k,a_h^k)V_{h+1,k}\right)^2 \le H^2d = L^2.$$

According to Theorem 2 of (Abbasi-Yadkori et al., 2011), we have that with probability at least $1-\delta$, for any $(h,k)\le(H,K)$, the following inequality holds:

$$\|\theta_*-\theta_{h,k}\|_{M_{h,k}} \le H\sqrt{d\log\left(\frac{1+Hk\cdot H^2d}{\delta}\right)}+C_\theta H\sqrt{d}$$

Therefore, if we choose

$$\beta_k = \left(H \sqrt{d \log \left(\frac{1 + Hk \cdot H^2 d}{\delta} \right)} + C_\theta H \sqrt{d} \right)^2 ,$$

then we will have

$$\theta_* \in B_{h,k}$$

for all $(h, k) \leq (H, K)$ with probability at least $1 - \delta$.

A.5.5 Expected Regret Analysis

According to Section A.5.4, we have with probability at least $1-\delta$ that $\theta_* \in B_k$ for all $1 \leq k \leq K$. When this event happens, the results from A.5.1-A.5.3 hold. We combine the error bounds (A.18) and (A.21) and apply them into the regret bound (A.17). It follows that, if $T = KN$,

$$\begin{aligned} R(T) &\leq 2\sqrt{H^3 K} \log(1/\delta) + 4\sqrt{H^2 d K \beta_K \log(1 + HK)} \\ &= 2\sqrt{H^3 K} \log \left(\frac{1}{\delta} \right) + 4H^2 d \sqrt{K \log(1 + HK)} \cdot \left(\sqrt{\log \left(\frac{1 + H^3 K d}{\delta} \right)} + C_\theta \right) \\ &\leq 6H^2 d \sqrt{K} \left(C_\theta \sqrt{\log(1 + HK)} + \log \left(\frac{1 + H^3 K d}{\delta} \right) \right) \end{aligned}$$

with probability at least $1 - 2\delta$. Note the trivial upper bound $R(K) \leq HK$. Therefore, by letting $\delta = 1/K$ and recalling that $T = HK$, we get

$$\begin{aligned} \mathbb{E}\left[R(T)\right] &\leq (1 - 2\delta) \cdot 6H^2 d \sqrt{K} \left(C_\theta \sqrt{\log(1 + HK)} + \log \left(\frac{1 + H^3 K d}{\delta} \right) \right) + 2\delta \cdot HK \\ &\leq 6H^2 d \sqrt{K} \cdot \left(C_\theta \sqrt{\log(1 + HK)} + \log \left(\frac{1 + H^3 K d}{\delta} \right) \right) \\ &= \widetilde{\mathcal{O}}(C_\theta \cdot H^2 d \sqrt{K}) = \widetilde{\mathcal{O}}(C_\theta \cdot d\sqrt{H^3 T}). \end{aligned}$$

Thus we have completed the proof of Theorem 21.

A.6 Implementation

A.6.1 Analysis of Implemented Confidence Bounds

In the implementation of UCRL-VTR used in Section 4.3, we used different confidence intervals then the ones stated in the paper. The confidence intervals used in our implementation are the ones introduced in (Abbasi-Yadkori et al., 2011). These confidence intervals are much tighter in the linear setting than the ones introduced in Sections 4.1 and A.3. This give us better practical performance. The purpose of this section is to formally introduce the confidence intervals used in our implementation of UCRL-VTR as well as show how these confidence intervals were adapted from the linear bandit setting to the linear MDP setting.

Linear Mixture Models

For our implementation of UCRL-VTR we used different confidence then was introduced in the paper. These are the tighter confidence bounds from the seminal work done by (Abbasi-Yadkori et al., 2011) and further expanded upon in Chapter 20 of (Lattimore and Szepesvári, 2018). Firstly, we will restate the assumptions for the linearly mixture model setting. Secondly, we will state the equivalent assumptions for the linear bandit setting. Finally, we will make the connections between the two settings that allow us to use the confidence bounds from the linear bandit setting to the linear mixture model setting.

1. $P^*(s' \mid s, a) = \sum_{i=1}^{d} (\theta_*^{MDP})_i P_i(s' \mid s, a)$
2. $s_{h+1}^k \sim P^*(\cdot \mid s_h^k, a_h^k)$
3. $\mathcal{C}_t^{MDP} = \{\theta^{MDP} \in \mathbb{R}^d : \|\theta^{MDP} - \widehat{\theta}_t^{MDP}\|_{M_k} \leq \beta_t\}$

where t is defined in the table of A.1.4. Also note that in this section $(\cdot)_*$ denotes the true parameter or model, $(\cdot)^{MDP}$ denotes something derived or used in the linear mixture model setting, and $(\cdot)^{LIN}$ denotes something derived or used in the linear bandit setting. Now, under 1-3 of A.6.1 we hope to construct a confidence set $\mathcal{C}_t^{MDP}$ such that

$$\theta^{MDP} \in \bigcap_{t=1}^{\infty} \mathcal{C}_t^{MDP}$$

with high probability. Now the choice of how to choose both $\mathcal{C}_t^{MDP}$ and β_t comes from the linear bandit literature. We will introduce the necessary theorems and assumptions to derive both $\mathcal{C}_t^{LIN}$ and β_t in the linear bandit setting and then adapt the results from the linear bandit setting to the linear MDP setting.

Self Normalized Confidence Bounds for Linear Bandits

The following results are introduced in the paper by (Abbasi-Yadkori et al., 2011) and are further explained in Chapter 20 of the book by (Lattimore and Szepesvári, 2018). In this section, we will introduce the theorems and lemmas that allows us to derive tighter confidence intervals for the linear bandit setting. Then we will carefully adapt the confidence intervals to the linear mixture model setting. Now supposed a bandit algorithm has chosen actions $A_1, ..., A_t \in \mathbb{R}^d$ and received rewards $X_1^{LIN}, ..., X_t^{LIN}$ with $X_s^{LIN} = \langle A_t, \theta_*^{LIN} \rangle + \eta_s$ where η_s is some zero mean noise. The least squares estimator of θ_*^{LIN} is the minimizer of the following loss function

$$L_t(\theta^{LIN}) = \sum_{s=1}^{t} (X_s^{LIN} - \langle A_t, \theta^{LIN} \rangle)^2 + \lambda \|\theta^{LIN}\|_2^2$$

where $\lambda > 0$ is the regularizer. This loss function is minimized by

$$\widehat{\theta}_t^{LIN} = W_t^{-1} \sum_{s=1}^{t} X_s^{LIN} A_s \text{ with } W_t = \lambda I + \sum_{s=1}^{t} A_s A_s^\top$$

notice how this linear bandit problem is very similar to the linear mixture model problem introduced in Section 3 of Ayoub et al., 2020. In our linear mixture model setting, it is convenient to think of M and W as serving equivalent purposes (storing rank one updates) thus it is also convenient to think of A_t and X_t^{MDP} as serving equivalent purposes (the features by which we use to make our predictions), where X_t^{MDP} is defined in Section 3 of Ayoub et al., 2020 with some added notation to distinguish it from the X_t^{LIN} used here in the linear bandit setting. We will now build up some intuition by making some simplifying assumptions.

1. No regularization: $\lambda = 0$ and W_t is invertible.

2. Independent subgaussian noise: $(\eta_s)_s$ are independent and σ-subgaussian

3. Fixed Design: $A_1, ..., A_t$ are deterministically chosen without the knowledge of $X_1^{LIN}, ..., X_t^{LIN}$

finally it is also convenient to think of X_t^{LIN} and $V_{t+1}(s_{t+1})$ as serving equivalent purposes (the target of our predictions). Thus the statements we prove in the linear bandit setting can be easily adapted to the linear mixture model setting. While none of the assumptions stated above is plausible in the bandit setting, the simplifications eases the analysis and provides insight.

Comparing θ_*^{LIN} and $\widehat{\theta}_t^{LIN}$ in the direction $x \in \mathbb{R}^d$, we have

$$\langle \widehat{\theta}_t^{LIN} - \theta_*^{LIN}, x\rangle = \left\langle x, W_t^{-1} \sum_{s=1}^{t} A_s X_s^{LIN} - \theta_*^{LIN} \right\rangle = \left\langle x, W_t^{-1} \sum_{s=1}^{t} A_s (A_s^\top \theta_*^{LIN} + \eta_s) - \theta_*^{LIN} \right\rangle$$
$$= \left\langle x, W_t^{-1} \sum_{s=1}^{t} A_s \eta_s \right\rangle = \sum_{s=1}^{t} \langle x, W_t^{-1} A_s\rangle \eta_s$$

Since $(\eta_s)_s$ are independent and σ-subgaussian, by Lemma 5.4 and Theorem 5.3 (need to be stated),

$$\mathbb{P}\left(\langle \widehat{\theta}_t^{LIN} - \theta_*^{LIN}, x\rangle \geq \sqrt{2\sigma^2 \sum_{s=1}^{t} \langle x, W_t^{-1} A_s\rangle^2 \log\left(\frac{1}{\delta}\right)} \right) \leq \delta$$

A little linear algebra shows that $\sum_{s=1}^{t} \langle x, W_t^{-1} A_s\rangle^2 = \|x\|^2_{W_t^{-1}}$ and so,

$$\mathbb{P}\left(\langle \widehat{\theta}_t^{LIN} - \theta_*^{LIN}, x\rangle \geq \sqrt{2\sigma^2 \|x\|^2_{W_t^{-1}} \log\left(\frac{1}{\delta}\right)} \right) \leq \delta \qquad \text{(A.22)}$$

We now remove the limiting assumptions we stated above and use the newly stated assumptions for the rest of this section

1. There exists a $\theta_*^{LIN} \in \mathbb{R}^d$ such that $X_t^{LIN} = \langle \theta_*^{LIN}, A_t\rangle + \eta_t$ for all $t \geq 1$.

2. The noise is conditionally σ-subgaussian:

   $$\text{for all } \alpha \in \mathbb{R} \text{ and } t \geq 1,\ \mathbb{E}[\exp(\alpha\eta_t) \mid \mathcal{F}_{t-1}] \leq \exp\left(\frac{\alpha\sigma^2}{2}\right) a.s.$$

   where $\mathcal{F}_{t-1}$ is such that $A_1, X_1^{LIN}, ..., A_{t-1}, X_{t-1}^{LIN}$ are $\mathcal{F}_{t-1}$-measurable.

3. In addition, we now assume $\lambda > 0$.

Ideally we would want to use the Cramér-Chernoff method:

$$\mathbb{P}(\|\widehat{\theta}_t^{LIN} - \theta_*^{LIN}\|^2_{W_t} \geq u^2) \leq \inf_{\alpha>0} \mathbb{E}\left[\exp\left(\alpha\|\widehat{\theta}_t^{LIN} - \theta_*^{LIN}\|^2_{W_t} - \alpha u^2\right)\right].$$

However, we cannot bound this expectation. Now consider the special case of $\lambda = 0$. Assuming that $W_t = \sum_{s=1}^t A_s A_s^\top$ is invertible. Let

$$S_t = \sum_{s=1}^{t} \eta_s A_s$$

Recall that $\widehat{\theta}_t^{LIN} = W_t^{-1} \sum_{s=1}^t X_s^{LIN} A_s = \theta_*^{LIN} + W_t^{-1} S_t$. Hence,

$$\frac{1}{2}\|\widehat{\theta}_t^{LIN} - \theta_*^{LIN}\|_{W_t}^2 = \frac{1}{2}\|S_t\|_{W_t^{-1}}^2 = \max_{x \in \mathbb{R}^d} \left(\langle x, S_t \rangle - \frac{1}{2}\|x\|_{W_t}^2 \right).$$

The next lemma shows that the exponential of the term inside the maximum is a supermartingale even when $\lambda \geq 0$.

Lemma 25. *For all $x \in \mathbb{R}^d$ the process $D_t(x) = \exp(\langle x, S_t \rangle - \frac{1}{2}\|x\|_{W_t^2})$ is an $\mathbb{F}$-adapted non-negative supermartingale with $D_0(x) \leq 1$.*

The proof for this Lemma can be found in Chapter 20 of the book by Lattimore and Szepesvári, 2018. Again consider now again the case when $\lambda = 0$. The Cramér–Chernoff method combined with Lemma 25 leads to

$$\mathbb{P}\left(\frac{1}{2}\|\widehat{\theta}_t^{LIN} - \theta_*^{LIN}\|_{W_t}^2 \geq \log(1/\delta) \right) = \mathbb{P}\left(\exp\left(\max_{x \in \mathbb{R}^d} \left(\langle x, S_t \rangle - \frac{1}{2}\|x\|_{W_t}^2 \right) \right) \geq \log(1/\delta) \right) \tag{A.23}$$

$$\leq \delta \mathbb{E}\left[\exp\left(\max_{x \in \mathbb{R}^d} \left(\langle x, S_t \rangle - \frac{1}{2}\|x\|_{W_t}^2 \right) \right) \right] = \delta \mathbb{E}\left[\max_{x \in \mathbb{R}^d} D_t(x) \right] \tag{A.24}$$

Now Lemma 25 shows that $\mathbb{E}[D_t(x)] \leq 1$. Now using Laplace's approximation we write

$$\max_x D_t(x) \approx \int_{\mathbb{R}^d} D_t(x) dh(x),$$

where h is some measure on $\mathbb{R}^d$ chosen so that the integral can be calculated in closed form. We replace the maximum with an integral to obtain the following lemma

Lemma 26. *Let h be a probability measure on $\mathbb{R}^d$; then; $\bar{D}_t = \int_{\mathbb{R}^d} D_t(x) dh(x)$ is an $\mathbb{F}$-adapted non-negative supermartingale with $\bar{D}_0 = 1$.*

The proof of Lemma 26 can, again, be found in Chapter 20 of the book by (Lattimore and Szepesvári, 2018). The following theorem allowa us to derive our confidence sets.

Theorem 27. *For all $\lambda > 0$, and $\delta \in (0,1)$*

$$\mathbb{P}\left(\textit{exists } t \in \mathbb{N} : \|S_t\|^2_{W_t^{-1}} \geq 2\sigma^2 \log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det W_t}{\lambda^d}\right)\right) \leq \delta$$

Furthermore, if $\|\theta_*^{LIN}\|_2 \leq m_2$, then $\mathbb{P}(\textit{exists } t \in \mathbb{N}^+ : \theta_*^{LIN} \notin \mathcal{C}_t^{LIN}) \leq \delta$ with

$$\mathcal{C}_t^{LIN} = \left\{\theta \in \mathbb{R}^d : \|\widehat{\theta}_{t-1}^{LIN} - \theta\|_{W_{t-1}} < m_2\sqrt{\lambda} + \sqrt{2\sigma^2 \log\left(\frac{1}{\delta}\right) + \log\left(\frac{W_{t-1}}{\lambda^d}\right)}\right\}.$$

The proof of Theorem 27 can be found in Chapter 20 of the book by (Lattimore and Szepesvári, 2018).

Adaptation of the Confidence Bounds to our Linear MDP Setting

Now with the analysis introduced in the previous section we are ready to derive the confidence bounds used in our implementation of UCRL-VTR. Now using the notation from the linear bandit setting we set

1. The target $X_t^{MDP} = \int_j V_t(s')P_j(ds' \mid s_t, a_t)$
2. $Y_t = V_t(s_{t+1})$
3. $\mathcal{F}_{t-1} = \sigma(s_1, a_1, ..., s_{t-1}, a_{t-1})$, which just means the filtration is set to be the sigma-algebra generated by all past states and actions observed.
4. $\eta_t = Y_t - \langle X_t^{MDP}, \theta_*^{MDP}\rangle = V_t(s_{t+1}) - \int_j V_t(s')P_j^*(ds' \mid s_t, a_t)$, since θ_*^{MDP} is the true model of the MDP.
5. M_t in the linear mixture model MDP is defined equivalently to W_t in the linear bandit setting, i.e. they are both the sums of a regularizer term and a bunch of rank one updates.

it can be seen that our the noise in our system η_t has zero mean $\mathbb{E}[\eta_t \mid \mathcal{F}_{t-1}] = 0$ finally the noise in our system has variance $H/2$ thus our system in $H/2$-subgaussian.

Lemma 28. *(Hoeffding's lemma) Let $Z = Z - \mathbb{E}[Z]$ be a real centered random variable such that $Z \in [a, b]$ almost surely. Then $\mathbb{E}[\exp(\alpha Z)] \leq \exp(\alpha^2 \frac{(b-a)^2}{8})$ for any $\alpha \in \mathbb{R}$ or Z is subgaussian with variance $\sigma^2 = \frac{(b-a)^2}{4}$.*

Now using Lemma 28 and the fact that Y_t is bounded in the range of $[0, H]$, $\mathbb{E}[Y_t] = \langle X_t^{MDP}, \theta_*^{MDP} \rangle$, and $\eta_t = Y_t - \langle X_t^{MDP}, \theta_*^{MDP} \rangle = Y_t - \mathbb{E}[Y_t]$, the noise η_t for the linear mixture model MDP is $H/2$-subgaussian. This result is also stated in a proof from A.1.4.

Putting this all together we can derive the tighter confidence set for UCRL-VTR in the linear setting,

$$\mathcal{C}_t^{MDP} = \left\{ \theta \in \mathbb{R}^d : \|\widehat{\theta}_{t-1}^{MDP} - \theta\|_{M_{t-1}} < m_2\sqrt{\lambda} + \frac{H}{2}\sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{M_{t-1}}{\lambda^d}\right)} \right\}.$$

where $\|\theta_*^{MDP}\|_2 \leq m_2$. The justification of using these bounds in the linear mixture model MDP follows exactly from the justification given above for using these bounds in the linear bandit setting. In fact an even tighter self-normalized confidence set was proposed by (Zhou et al., 2020 for the linear mixture model MDP.

A.6.2 UCRL-VTR

In the proceeding subsections we discuss the implementation of the algorithms studied in Section 4.3. The first algorithm we present is the algorithm used to generate the results for UCRL-VTR.

Algorithm 6 UCRL-VTR with Tighter Confidence Bounds

```
Input: MDP, $d, H, T = KH$;
Initialize: $M_{1,1} \leftarrow I$, $w_{1,1} \leftarrow 0 \in \mathbb{R}^{d\times 1}$, $\theta_1 \leftarrow M_{1,1}^{-1} w_{1,1}$ for $1 \le h \le H$, $d_1 = |\mathcal{S}| \times |\mathcal{A}|$;
Initialize: $\delta \leftarrow 1/K$, and for $1 \le k \le K$,
Compute Q-function $Q_{h,1}$ using $\theta_{1,1}$ according to (4.1);
for $k = 1 : K$ do
    Obtain initial state $s_1^k$ for episode $k$;
    for $h = 1 : H$ do
        Choose action greedily by
            $$a_h^k = \arg\max_{a\in\mathcal{A}} Q_{h,k}(s_h^k, a)$$
        and observe the next state $s_{h+1}^k$.
        Compute the predicted value vector:    ▷ Evaluate the expected value of next state
            $$X_{h,k} \leftarrow \mathbb{E}.[V_{h+1,k}(s)|s_h^k, a_h^k] \ = \sum_{s\in\mathcal{S}} V_{h+1,k}(s) \cdot P.(s|s_h^k, a_h^k).$$
        $y_{h,k} \leftarrow V_{h+1,k}(s_{h+1}^k)$    ▷ Update regression parameters
        $M_{h+1,k} \leftarrow M_{h,k} + X_{h,k} X_{h,k}^\top$
        $w_{h+1,k} \leftarrow w_{h,k} + y_{h,k} \cdot X_{h,k}$
    end for
    Update at the end of episode:    ▷ Update Model Parameters
        $$\begin{aligned} M_{1,k+1} &\leftarrow M_{H+1,k}, \\ w_{1,k+1} &\leftarrow w_{H+1,k}, \\ \theta_{k+1} &\leftarrow M_{1,k+1}^{-1} w_{1,k+1}; \end{aligned}$$
    Compute $Q_{h,k+1}$ for $h = H, \dots, 1$, using $\theta_{k+1}$ according to (A.25) using
        $$\sqrt{\beta_{h,k}} \leftarrow \sqrt{d_1} + \frac{H-h+1}{2}\sqrt{2\log\left(\frac{1}{\delta}\right) + \log\det(M_{1,k+1})};$$
                                               ▷ Computing Q functions
end for
```

The iterative Q-update for Algorithm 6 is

$$\begin{aligned} V_{h+1,k}(s) &= 0 \\ Q_{h,k}(s,a) &= r(s,a) + X_{h,k}^\top \theta_k + \sqrt{\beta_{h,k}}\sqrt{X_{h,k}^\top M_{1,k+1}^{-1} X_{h,k}} \qquad \text{(A.25)} \\ V_{h,k}(s) &= \max_a Q_{h,k}(s,a) \end{aligned}$$

The choice of the confidence bounds used in Algorithm 6 comes from the tight bounds derived in (Abbasi-Yadkori et al., 2011) for linear bandits and further expanded upon in Chapter 20 of (Lattimore and Szepesvári, 2018). The details of which are shown and stated in A.6.1. We slightly tighten the values for the noise at each stage by using the fact that for each stage in the horizon, $h \in [H]$, the value $V_h^k(\cdot)$ is capped as to never be greater than $H - h + 1$. The appearance of the $\sqrt{d_1}$ comes from the fact that $\|\theta_*\|_2 \leq \sqrt{d_1}$ for all $\theta_* \in \mathbb{R}^d$ in the tabular setting since θ_* in the tabular setting is equal to the true model of the environment.

A.6.3 EGRL-VTR

In this section we discuss the algorithm EGRL-VTR. This algorithm is very similar to UCRL-VTR expect it performs ε-greedy value iteration instead of optimistic value iteration and acts ε-greedy with respect to $Q_{h,k}$.

Algorithm 7 EGRL-VTR

```
Input: MDP, $d, H, T = KH, \varepsilon > 0$;
Initialize: $M_{1,1} \leftarrow I$, $w_{1,1} \leftarrow 0 \in \mathbb{R}^{d\times 1}$, $\theta_1 \leftarrow M_{1,1}^{-1} w_{1,1}$ for $1 \le h \le H$;
Compute Q-function $Q_{h,1}$ using $\theta_{1,1}$ according to (A.26);
for $k = 1 : K$ do
    Obtain initial state $s_1^k$ for episode $k$;
    for $h = 1 : H$ do
        With probability $1 - \varepsilon$ do
            $$a_h^k = \arg\max_{a\in\mathcal{A}} Q_{h,k}(s_h^k, a)$$
        else pick a uniform random action $a_h^k \in \mathcal{A}$. Observe the next state $s_{h+1}^k$.
        Compute the predicted value vector:    ▷ Evaluate the expected value of next state
            $$X_{h,k} \leftarrow \mathbb{E}.[V_{h+1,k}(s)|s_h^k, a_h^k] \quad = \sum_{s\in\mathcal{S}} V_{h+1,k}(s) \cdot P.(s|s_h^k, a_h^k).$$
        $y_{h,k} \leftarrow V_{h+1,k}(s_{h+1}^k)$    ▷ Update regression parameters
        $M_{h+1,k} \leftarrow M_{h,k} + X_{h,k} X_{h,k}^\top$
        $w_{h+1,k} \leftarrow w_{h,k} + y_{h,k} \cdot X_{h,k}$
    end for
    Update at the end of episode:    ▷ Update Model Parameters
        $$M_{1,k+1} \leftarrow M_{H+1,k},$$
        $$w_{1,k+1} \leftarrow w_{H+1,k},$$
        $$\theta_{k+1} \leftarrow M_{1,k+1}^{-1} w_{1,k+1};$$
    Compute $Q_{h,k+1}$ for $h = H, \dots, 1$, using $\theta_{k+1}$ according to (A.26)    ▷ Computing Q functions
end for
```

The iterative value update for EGRL-VTR is

$$
\begin{aligned}
V_{h+1,k}(s) &= 0 \\
Q_{h,k}(s,a) &= r(s,a) + X_{h,k}^\top \theta_k \\
V_{h,k}(s) &= (1-\varepsilon)\Pi_{[0,H]} \max_a Q_{h,k}(s,a) + \frac{\varepsilon}{|\mathcal{A}|} \sum_{a\in\mathcal{A}} Q_{h,k}(s,a)
\end{aligned}
\tag{A.26}
$$

A.6.4 EG-Frequency

In this section we discuss the algorithm EG-Frequency. This algorithm is the ε-greedy version of UC-MatrixRL (L. F. Yang and Wang, 2019).

Algorithm 8 EG-Frequency

```
Input: MDP, Features $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and $\psi : \mathcal{S} \to \mathbb{R}^{|\mathcal{S}|}$, $\varepsilon > 0$, and the total number of episodes $K$;
Initialize: $A_1 \leftarrow I \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$, $M_1 \leftarrow 0 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$, and $K_\psi \leftarrow \sum_{s' \in \mathcal{S}} \psi(s')\psi(s')^\top$;
for $k = 1 : K$ do
    Let $Q_{h,k}$ be given in (A.27) using $M_k$;
    for $h = 1 : H$ do
        Let the current state be $s_h^k$;
        With probability $(1-\varepsilon)$ play action $a_h^k = \arg\max_{a \in \mathcal{A}} Q_{h,k}(s_h^k, a)$ else pick a uniform random action $a_h^k \in \mathcal{A}$.
        Record the next state $s_{h+1}^k$
    end for
    $A_{k+1} \leftarrow A_k + \sum_{h \le H} \phi(s_h^k, a_h^k)\phi(s_h^k, a_h^k)^\top$
    $M_{k+1} \leftarrow M_k + A_{k+1}^{-1} \sum_{h \le H} \phi(s_h^k, a_h^k)\psi(s_{h+1}^k)^\top K_\psi^{-1}$
end for
```

The iterative Q-update for EG-Frequency is

$$\begin{aligned} Q_{h+1,k}(s,a) &= 0 \text{ and} \\ Q_{h,k}(s,a) &= r(s,a) + \phi(s,a)^\top M_k \mathbf{\Psi}^\top V_{h+1,k} \\ V_{h,k} &= (1-\varepsilon)\Pi_{[0,H]} \max_a Q_{h,k}(s,a) + \frac{\varepsilon}{|\mathcal{A}|}\sum_{a \in \mathcal{A}} Q_{h,k}(s,a) \end{aligned} \tag{A.27}$$

Note that $\mathbf{\Psi}$ is a $|\mathcal{S}| \times |\mathcal{S}|$ whose rows are the features $\psi(s')$ and $\mathbf{\Phi}$ is a $|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|$ whose rows are the features $\phi(s,a)$. In the tabular RL setting both $\mathbf{\Psi}$ and $\mathbf{\Phi}$ are the identity matrix which is what we used in our numerical experiments. In the tabular RL setting, EG-Frequency stores the counts of the number of times it transitioned to next state s' from the state-action pair (s,a) and fits the estimated model M_k accordingly.

A.6.5 Further Implementation Notes

In this section, we include some further details on how we implemented Algorithms 6, 7, and 8. All code was written in Python 3 and used the Numpy and Scipy libraries. All plots were generated using MatPlotLib. In Algorithm 6, Numpy's logdet function was used to calculate the determinate in step 15 for numerical stability purposes. No matrix inversion was performed in our code, instead a Sherman-Morrison update was performed for each matrix in which a matrix inversion is performed at each (k, h) in order to save on computation. To read more about the Sherman Morrison update in the context of RL, we refer to the reader to Eqn (9.22) of Sutton and Barto, 2018. When computing the weighted L1-norm, we added a small constant to each summation in the denominator to avoid dividing by zero. Finally, when computing UC-MatrixRL we also used the self-normalize bounds introduced in the beginning of this section. Some pseudocode for using self-normalized bounds with UC-MatrixRL can be found in step 5 of Alg 9.

A.7 Mixture Model

In this section, we introduce, analyze, and evaluate a linear model-based reinforcement learning algorithm that uses both the canonical model and the VTR model for planning. We call this algorithm UCRL-MIX.

A.7.1 UCRL-MIX

Below a meta-algorithm for UCRL-MIX

Algorithm 9 UCRL-MIX

```
Compute Algorithm 6 and UC-MatrixRL L. F. Yang and Wang, 2019 simultaneously.
At end of episode $k$, perform value iteration and set $V_{H+1,k}(s) = 0$.
for $h = H : 1$ do
    for $s \in |\mathcal{S}|$ and $a \in |\mathcal{A}|$ do
        Compute the confidence set bonuses as follows
            $$B_{h,k}^{VTR} \leftarrow \sqrt{d_1} + \frac{H-h+1}{2}\sqrt{2\log\left(\frac{2}{\delta}\right) + \log\det(M_{1,k+1})};$$
            $$B_{h,k}^{MAT} \leftarrow \sqrt{|\mathcal{S}||\mathcal{A}|} + \frac{H-h+1}{2}\sqrt{2\log\left(\frac{2}{\delta}\right) + \log\det(A_{k+1})};$$
        if $B_{h,k}^{VTR}\sqrt{X_{h,k}^\top M_{1,k+1}^{-1} X_{h,k}} \le B_{h,k}^{MAT}\sqrt{\phi^\top(s,a)A_n^{-1}\phi(s,a)}$ then
            Perform one step of value iteration using the VTR model as follows: $Q_{h,k}(s,a) = r(s,a) + X_{h,k}^\top \theta_k + \sqrt{\beta_{h,k}}\sqrt{X_{h,k}^\top M_{1,k+1}^{-1} X_{h,k}}$
        else:
            Update $Q_{h,k}(s,a)$ according to Equation 8 L. F. Yang and Wang, 2019 using the UC-MatrixRL model $A_k$. Note that in L. F. Yang and Wang, 2019 they use $n$ to denote the current episode, in our paper we use $k$ to denote the current episode.
        end if
        $V_{h,k}(s) = \max_a Q_{h,k}(s,a)$
    end for
end for
```

We are now using multiple models instead of a single model, we must adjust our confidence sets accordingly. By using a union bound we replace δ with $\delta/2$ for our confidence parameter. This updated confidence parameter changes the term inside the logarithm. We now have $\log(2/\delta)$ where as before we had $\log(1/\delta)$.

A.7.2 Numerical Results

We will include the cumulative regret and the weighted L1 norm of UCRL-MIX on the RiverSwim environment as in Section 4.3. We also include a bar graph of the relative frequency with which the algorithm used the VTR-model for planning and the canonical model for planning.

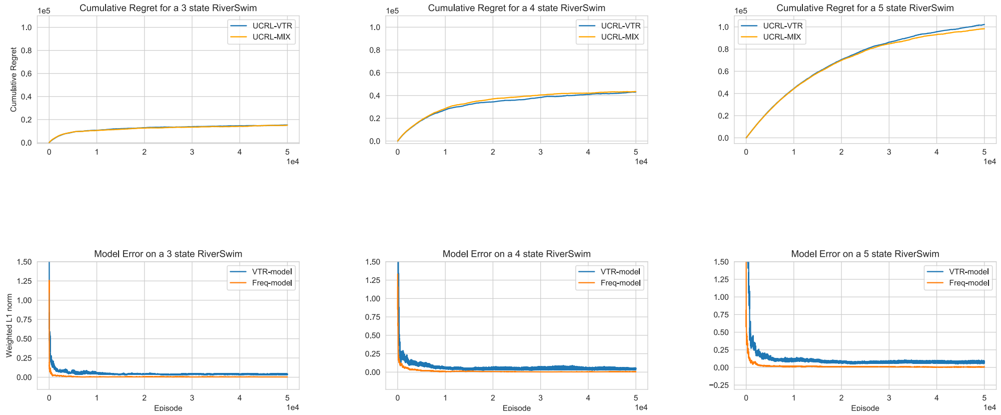

Figure A.1: In the plots for the model error we include model error for both the VTR-model and the canonical model. Even though only one is used during planning both are updated at the end of each episode.

If we compare the results of Figure A.1 with the results of Figure 4.2 from Section 4.3.2 we see that the cumulative regret of UCRL-MIX is almost identical to the cumulative regret of UCRL-VTR. The model errors of both the VTR and the canonical models are almost identical to the model errors of UCRL-VTR and UC-MatrixRL respectively.

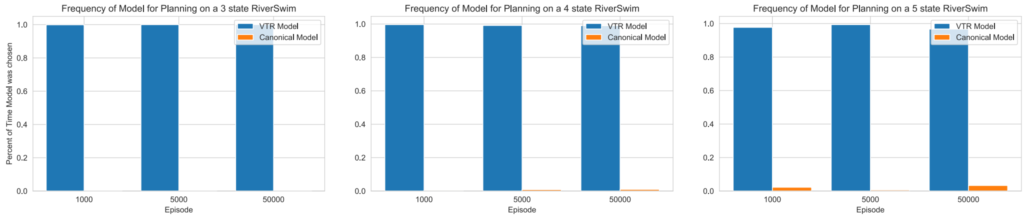

Figure A.2: UCRL-MIX rarely, if ever, chooses the canonical model for planning on the RiverSwim environments.

From Figure A.2, we see that on the RiverSwim environment, UCRL-MIX almost always uses the VTR-model for planning. We calculate this frequency by counting the number of times Step 7 of Alg 9 was observed up until episode k and by counting the number of times Step 9 of Alg 9 was observed up until episode k. We then divide these counts by the sum of the counts to get a percentage. We believe the reason the algorithm overwhelming chose the VTR-model was due to the fact that the confidence intervals for the VTR-model shrink much faster than the confidence intervals for the canonical model.

The canonical model is forced to explore much longer than the VTR-model as its objective is to learn a globally optimal model rather than a model that yields high reward. Thus, the canonical model is forced to explore all state-action-next state tuples, even ones that do not yield high reward, in order to meet its objective of learning a globally optimal model while the VTR-model is only forced to explore state-action-next state tuples that fall in-line with its objective of accumulating high reward. The set of all state-action-next state tuples is much larger then the set of state-action-next state tuples that yield high reward which means the confidence intervals for the canonical model shrink slower than the confidence sets of the VTR-model on the RiverSwim environment.

Appendix B

Least Squares Value Iteration with Perturbed Histories Exploration Appendix

B.1 LSVI-PHE with General Function Approximations

B.1.1 Noise

In the section, we specify how to choose σ in Algorithm 2. Note that we use $\xi_{h,k}^{\tau,m}$ for the noise added in episode k, timestep h, data from episode $\tau < k$ and sampling time m. Similarly, $\xi_{h,k}^{\prime i,m}$ is for episode k, timestep h, regularizer $p_i(\cdot)$ and sampling time m. We set $\lambda = 1$ in our algorithm. By Lemma 32, there exists $\beta'(\mathcal{F}, \delta)$ such that with probability at least $1-\delta$, for all $(k, h) \in [K]\times[H]$, we have

$$f_h^k(\cdot,\cdot) := r(\cdot,\cdot) + P_h V_{h+1}^k(\cdot,\cdot) \in \mathcal{F}_h^k,$$

where $\mathcal{F}_h^k = \{f \in \mathcal{F} \mid \|f - \widehat{f}_h^k\|_{\mathcal{Z}_h^k}^2 + R(f - \widehat{f}_h^k) \le \beta'(\mathcal{F}, \delta)\}$. By Assumption 5, for each $\mathcal{F}_h^k$, there exists a $\sigma_{h,k}$ such that

$$g_{\sigma_{h,k}}(s, a) \ge w(\mathcal{F}_h^k, s, a).$$

We define $\sigma = \max_{k\in[K],h\in[H]} \sigma_{h,k}$ to be the maximum standard deviation of the added noise.

B.1.2 Concentration

We first define few filtrations and good events that we will use in the proof of lemmas in this section.

Definition 8 (Filtrations)**.** *We denote the σ-algbera generated by the set $\mathcal{G}$ using $\sigma(\mathcal{G})$. We define the following filtrations*

$$\mathcal{G}^k \overset{def}{=} \sigma\left(\{(s_t^i, a_t^i, r_t^i)\}_{\{i,t\}\in[k-1]\times[H]} \bigcup \{\xi_{t,l}^{i,j}\}_{i\in[l],\{t,j,l\}\in[H]\times[M]\times[k-1]} \bigcup \{\xi_{t,l}^{\prime i,j}\}_{\{i,t,j,l\}\in[D]\times[H]\times[M]\times[k-1]}\right)$$
$$\mathcal{G}_{h,1}^k \overset{def}{=} \sigma\left(\mathcal{G}^k \bigcup \{(s_t^k, a_t^k, r_t^k)\}_{t\in[h]} \bigcup \{\xi_{t,k}^{i,j}\}_{i\in[k],t\geq h,j\in[M]} \bigcup \{\xi_{t,k}^{\prime i,j}\}_{i\in[D],t\geq h,j\in[M]}\right),$$
$$\mathcal{G}_{h,2}^K \overset{def}{=} \sigma\left(\mathcal{G}^k \bigcup \{(s_t^k, a_t^k, r_t^k)\}_{t\in[h]}\right).$$

Definition 9 (Good events)**.** *For any $\delta > 0$, we define the following random events*

$$\mathcal{G}_h^k(\xi,\delta) \overset{def}{=} \left\{\max_{i\in[k],j\in[M]} |\xi_{h,k}^{i,j}| \leq \sqrt{\gamma_k(\delta)} \bigcap \max_{i\in[D],j\in[M]} |\xi_{h,k}^{\prime i,j}| \leq \sqrt{\gamma_k(\delta)}\right\},$$
$$\mathcal{G}(K,H,\delta) \overset{def}{=} \bigcap_{k\leq K} \bigcap_{h\leq H} \mathcal{G}_h^k(\xi,\delta),$$

where $\gamma_k(\delta)$ is some constant to be specified in Lemma 29.

Notation: To simplify our presentation, in the remaining part of this section, we always denote $\sqrt{\gamma_k} := \sqrt{\gamma_k(\delta)}$.

The next lemma shows that the good event defined in Definition 9 happens with high probability.

Lemma 29. *For good event $\mathcal{G}(K,H,\delta)$ defined in Definition 9, if we set $\sqrt{\gamma_k} = \widetilde{O}(\sigma)$, then it happens with probability at least $1-\delta$.*

Proof. Recall that $\xi_{t,l}^{i,j}$ is a zero-mean Gaussian noise with variance $\sigma_{t,l}^2$. By the concentration of Gaussian distribution (Lemma 51), with probability $1-\delta'$, we have

$$|\xi_{t,l}^{i,j}| \leq \sigma_{t,l}\sqrt{2\log(1/\delta')} \leq \sigma\sqrt{2\log(1/\delta')}.$$

The same result holds for $\xi_{t,l}^{\prime i,j}$. We complete the proof by setting $\delta' = \delta/(K+D)MHK$ and using union bound. □

In Definition 4, for a regularizer $R(f) = \sum_{j=1}^{D} p_j(f)^2$, where $p_j(\cdot)$ are functionals, we defined the perturbed regularizer as $\widetilde{R}_\sigma(f) = \sum_{j=1}^{D}[p_j(f) + \xi'_j]^2$ with ξ'_j being i.i.d. zero-mean Gaussian noise with variance σ^2. Note that in the algorithm, the variance of the noise for the regularizer is the same as the dataset, which is $\sigma_{h,k}^2$. Recall from Assumption 6 that for any $V : \mathcal{S} \to [0, H]$, our regularizer R satisfies $R(r + PV) \leq B$ for some constant $B \in \mathbb{R}$.

Our next lemma establishes a bound on the perturbed estimate of a single backup.

Lemma 30. *Consider a fixed $k \in [K]$ and a fixed $h \in [H]$. Let $\mathcal{Z}_h^k = \{(s_h^\tau, a_h^\tau)\}_{\tau \in [k-1]}$ and $\widetilde{\mathcal{D}}_{h,V}^k = \{(s_h^\tau, a_h^\tau, r_h^\tau + \xi_h^\tau + V(s_{h+1}^\tau))\}_{\tau \in [k-1]}$. Define $\widetilde{f}_{h,V}^k = \operatorname{argmin}_{f \in \mathcal{F}} \|f\|_{\widetilde{\mathcal{D}}_{h,V}^k}^2 + \widetilde{R}(f)$. Conditioned on the good event $\mathcal{G}(K, H, \delta)$, with probability at least $1 - \delta$, for a fixed $V : \mathcal{S} \to [0, H]$ and any $V' : \mathcal{S} \to [0, H]$ with $\|V' - V\|_\infty \leq 1/T$, we have*

$$\left\| \widetilde{f}_{h,V'}(\cdot,\cdot) - r_h(\cdot,\cdot) - P_h V'(\cdot,\cdot) \right\|_{\mathcal{Z}_h^k}^2 + R\left(\widetilde{f}_{h,V'}(\cdot,\cdot) - r_h(\cdot,\cdot) - P_h V'(\cdot,\cdot)\right)$$
$$\leq c' \left[(H + 1 + \sqrt{\gamma_k}) \sqrt{\log(2/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)} + \sqrt{B + \sqrt{\gamma_k B D}} \right]^2,$$

for some constant c'. Here B is the bound on the regularizer (Assumption 6) and D is the number of regularizers (Definition 4). Define this event as $\mathcal{E}_{h,V}(\delta)$.

Proof. Recall that for notational simplicity, we denote $[\mathbb{P}_h V_{h+1}](s, a) = \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot \mid s,a)} V_{h+1}(s')$. Now consider a fixed $V : \mathcal{S} \to [0, H]$, and define

$$f_V(\cdot,\cdot) := r_h(\cdot,\cdot) + P_h V(\cdot,\cdot). \tag{B.1}$$

For any $f \in \mathcal{F}$, we consider $\sum_{\tau \in [k-1]} \chi_h^\tau(f)$ where

$$\chi_h^\tau(f) := 2(f(s_h^\tau, a_h^\tau) - f_V(s_h^\tau, a_h^\tau))(f_V(s_h^\tau, a_h^\tau) - r_h^\tau(s_h^\tau, a_h^\tau) - \xi_h^\tau - V(s_{h+1}^\tau)).$$

Recalling the definition of the filtration $\mathcal{G}_{h,1}^\tau$ from Definition 8, we note

$$\begin{aligned}
\mathbb{E}[\chi_h^\tau(f) | \mathcal{G}_{h,1}^\tau] &= \mathbb{E}[2(f(s_h^\tau, a_h^\tau) - f_V(s_h^\tau, a_h^\tau))(f_V(s_h^\tau, a_h^\tau) - r_h^\tau(s_h^\tau, a_h^\tau) - \xi_h^\tau - V(s_{h+1}^\tau)) | \mathcal{G}_{h,1}^\tau] \\
&= 2(f(s_h^\tau, a_h^\tau) - f_V(s_h^\tau, a_h^\tau))\mathbb{E}[(f_V(s_h^\tau, a_h^\tau) - r_h^\tau(s_h^\tau, a_h^\tau) - \xi_h^\tau - V(s_{h+1}^\tau)) | \mathcal{G}_{h,1}^\tau] \\
&= 2(f(s_h^\tau, a_h^\tau) - f_V(s_h^\tau, a_h^\tau))(f_V(s_h^\tau, a_h^\tau) - r_h^\tau(s_h^\tau, a_h^\tau) - P_h V(s_h^\tau, a_h^\tau)) \\
&= 0.
\end{aligned}$$

In addition, conditioning on the good event $\mathcal{G}(K, H, \delta)$, we have

$$|\chi_h^\tau(f)| \leq 2(H + 1 + \sqrt{\gamma_\tau})|f(s_h^\tau, a_h^\tau) - f_V(s_h^\tau, a_h^\tau)|.$$

As $\chi_h^\tau(f)$ is a martingale difference sequence conditioned on the filtration $\mathcal{G}_{h,1}^\tau$, by Azuma-Hoeffding inequality, we have

$$\mathbb{P}\left[\left|\sum_{\tau\in[k-1]} \chi_h^\tau(f)\right| \geq \epsilon\right] \leq 2\text{exp}\left(-\frac{\epsilon^2}{8(H+1+\sqrt{\gamma_\tau})^2\|f - f_V\|_{\mathcal{Z}_h^k}^2}\right).$$

Now we set

$$\begin{aligned}\epsilon &= \sqrt{8(H+1+\sqrt{\gamma_\tau})^2 \log\left(\frac{2\mathcal{N}(\mathcal{F}, 1/T)}{\delta}\right)\|f - f_V\|_{\mathcal{Z}_h^k}^2} \\ &\leq 4(H+1+\sqrt{\gamma_\tau})\|f - f_V\|_{\mathcal{Z}_h^k}\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)}.\end{aligned}$$

With union bound, for all $g \in \mathcal{C}(\mathcal{F}, 1/T)$, with probability at least $1-\delta$ we have

$$\left|\sum_{(\tau)\in[k-1]} \xi_h^\tau(g)\right| \leq 4(H+1+\sqrt{\gamma_\tau})\|f - f_V\|_{\mathcal{Z}_h^k}\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)}.$$

Thus, for all $f \in \mathcal{F}$, there exists $g \in \mathcal{C}(\mathcal{F}, 1/T)$ such that $\|f - g\|_\infty \leq 1/T$ and

$$\begin{aligned}&\left|\sum_{(\tau)\in[k-1]} \chi_h^\tau(f)\right| \leq \left|\sum_{(\tau)\in[k-1]} \chi_h^\tau(g)\right| + 2(H+1+\sqrt{\gamma_\tau}) \\ &\leq 4(H+1+\sqrt{\gamma_\tau})\|g - f_V\|_{\mathcal{Z}_h^k}\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)} + 2(H+1+\sqrt{\gamma_\tau}) \\ &\leq 4(H+1+\sqrt{\gamma_\tau})(\|f - f_V\|_{\mathcal{Z}_h^k} + 1)\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)} + 2(H+1+\sqrt{\gamma_\tau}).\end{aligned}$$

For $V' : \mathcal{S} \to [0, H]$ such that $\|V - V'\|_\infty \leq 1/T$, we have $\|f_{V'} - f_V\|_\infty \leq \|V' - V\|_\infty \leq 1/T$.

For any $f \in \mathcal{F}$, we have

$$
\begin{aligned}
& \|f\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} - \|f_{V'}\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} \\
=& \|f - f_{V'}\|^2_{\mathcal{Z}^k_h} + 2 \sum_{(s^\tau_h, a^\tau_h) \in \mathcal{Z}^k_h} (f(s^\tau_h, a^\tau_h) - f_{V'}(s^\tau_h, a^\tau_h))(f_{V'}(s^\tau_h, a^\tau_h) - r^\tau_h(s^\tau_h, a^\tau_h) - \xi^\tau_h - V'(s^\tau_{h+1})) \\
\geq& \|f - f_{V'}\|^2_{\mathcal{Z}^k_h} + 2 \sum_{(s^\tau_h, a^\tau_h) \in \mathcal{Z}^k_h} (f(s^\tau_h, a^\tau_h) - f_V(s^\tau_h, a^\tau_h))(f_V(s^\tau_h, a^\tau_h) - r^\tau_h(s^\tau_h, a^\tau_h) - \xi^\tau_h - V(s^\tau_{h+1})) \\
& - 4(H + 1 + \sqrt{\gamma_k})\|V' - V\|_\infty |\mathcal{Z}^k_h| \\
\geq& \|f - f_{V'}\|^2_{\mathcal{Z}^k_h} + \sum_{(\tau,h) \in [k-1]\times[H]} \chi^\tau_h(f) - 4(H + 1 + \sqrt{\gamma_k}) \\
\geq& \|f - f_{V'}\|^2_{\mathcal{Z}^k_h} - 4(H + 1 + \sqrt{\gamma_k})(\|f - f_V\|_{\mathcal{Z}^k_h} + 1)\sqrt{\log(2/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)} - 6(H + 1 + \sqrt{\gamma_k}) \\
\geq& \|f - f_{V'}\|^2_{\mathcal{Z}^k_h} - 4(H + 1 + \sqrt{\gamma_k})(\|f - f_{V'}\|_{\mathcal{Z}^k_h} + 2)\sqrt{\log(2/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)} - 6(H + 1 + \sqrt{\gamma_k}).
\end{aligned}
$$

In addition, using Assumption 6, we have the approximate triangle inequality for the perturbed regularizer:

$$
\begin{aligned}
& \widetilde{R}(f) - \widetilde{R}(f_{V'}) \\
=& \sum_i^D [p_i(f) + \xi'_i]^2 - \sum_i^D [p_i(f_{V'}) + \xi'_i]^2 \\
=& R(f) - R(f_{V'}) + 2\sum_i^D \xi'_i(p_i(f) - p_i(f_{V'})) \\
\geq& cR(f - f_{V'}) - 2R(f_{V'}) - 2\sum_i^D \sqrt{\gamma_k} p_i(f_{V'}) \\
\geq& cR(f - f_{V'}) - 2B - 2\sqrt{\gamma_k}\sqrt{BD}.
\end{aligned}
$$

Summing the above two inequalities we have

$$
\|f\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} + \widetilde{R}(f) - \|f_{V'}\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} - \widetilde{R}(f_{V'}) \geq \|f - f_{V'}\|^2_{\mathcal{Z}^k_h} + cR(f - f_{V'}) - C,
$$

where $C = 4(H + 1 + \sqrt{\gamma_k})(\|f - f_{V'}\|_{\mathcal{Z}^k_h} + 2)\sqrt{\log(2/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)} + 6(H + 1 + \sqrt{\gamma_k}) + 2B + 2\sqrt{\gamma_k}\sqrt{BD}$.

As $\widetilde{f}_{h,V'}$ is the minimizer of $\|f\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} + \widetilde{R}(f)$, we have

$$\begin{aligned}&\|\widetilde{f}_{h,V'} - f_{V'}\|^2_{\mathcal{Z}^k_h} + cR(\widetilde{f}_{h,V'} - f_{V'})\\ &\leq c'\left[(H+1+\sqrt{\gamma_k})\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)} + \sqrt{B + \sqrt{\gamma_k BD}}\right]^2.\end{aligned}$$

To prove the above argument, we use the inequality that if we have $x^2+y \leq ax+b$ for positive a, b, y, then $x \leq a+\sqrt{b}$ and $x^2+y \leq (a+\sqrt{b})^2$. In addition, we can remove c by replacing c' with $c'/\min\{1, c\}$ and then we get our final bound.

□

Lemma 31 (Confidence Region). *Let* $\mathcal{F}^{k,m}_h = \{f \in \mathcal{F} | \|f - \widetilde{f}^{k,m}_h\|^2_{\mathcal{Z}^k_h} + R(f - \widetilde{f}^{k,m}_h) \leq \beta(\mathcal{F}, \delta)\}$, *where*

$$\beta(\mathcal{F}, \delta) = c'\left[(H+1+\sqrt{\gamma_k})\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)} + \sqrt{B + \sqrt{\gamma_k BD}}\right]^2. \tag{B.2}$$

Conditioned on the event $\mathcal{G}(K, H, \delta)$, *with probability at least* $1-\delta$, *for all* $(k, h, m) \in [K] \times [H] \times [M]$, *we have*

$$r_h(\cdot, \cdot) + P_h V^k_{h+1}(\cdot, \cdot) \in \mathcal{F}^{k,m}_h.$$

Proof. First note that for a fixed $(k, h, m) \in [K] \times [H] \times [M]$,

$$\mathcal{Q} = \{\min\{f(\cdot, \cdot), H\} \mid f \in \mathcal{C}(\mathcal{F}, 1/T)\} \cup \{0\}$$

is a $(1/T)$-cover of $Q^{k,m}_{h+1}(\cdot, \cdot)$. This implies $\mathcal{Q}$ is also a $(1/T)$-cover of $Q^k_{h+1}(\cdot, \cdot)$. This further implies

$$\mathcal{V} = \{\max_{a \in \mathcal{A}} q(\cdot, a) \mid q \in \mathcal{Q}\}$$

is a $1/T$ cover of $V^k_{h+1}(\cdot)$ where we have $\log(|\mathcal{V}|) = \log\mathcal{N}(\mathcal{F}, 1/T)$.

For the remaining part of the proof, we condition on $\bigcap_{V \in \mathcal{V}} \mathcal{E}_{h,V}(\delta/|\mathcal{V}|TM)$, where $\mathcal{E}_{h,V}(\delta)$ is the event defined in Lemma 30. By Lemma 30 and union bound, we have $\Pr\left[\bigcap_{V \in \mathcal{V}} \mathcal{E}_{h,V}(\delta/(8|\mathcal{V}|MT)\right] \geq 1 - \delta/(8MT)$.

Let $V \in \mathcal{V}$ such that $\|V - V_{h+1}^k\|_\infty \le 1/T$. By Lemma 30 we have

$$\left\|\widetilde{f}_h^{k,m}(\cdot,\cdot) - r_h(\cdot,\cdot) - P_h V_{h+1}^k(\cdot,\cdot)\right\|_{\mathcal{Z}_h^k}^2 + R(\widetilde{f}_h^{k,m}(\cdot,\cdot) - r_h(\cdot,\cdot) - P_h V_{h+1}^k(\cdot,\cdot))$$
$$\le c' \left[(H+1+\sqrt{\gamma_k})\sqrt{\log(1/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)}\right]^2,$$

where c' is some absolute constant. By union bound, for all $(k,h,m) \in [K]\times[H]\times[M]$ we have $r_h(\cdot,\cdot) + P_h V_{h+1}^k(\cdot,\cdot) \in \mathcal{F}_h^{k,m}$ with probability $1-\delta$. □

The last lemma guarantees that $r_h(\cdot,\cdot) + P_h V_{h+1}^k(\cdot,\cdot)$ lies in the confidence region $\mathcal{F}_h^{k,m}$ with high probability. Note that the confidence region $\mathcal{F}_h^{k,m}$ is centered at $\widetilde{f}_h^{k,m}$, which is the solution to the perturbed regression problem defined in (5.1). For the unperturbed regression problem and its solution as center of the confidence region, we get the following lemma as a direct consequence of Lemma 31.

Lemma 32. *Let* $\mathcal{F}_h^k = \{f \in \mathcal{F} | \|f - \widehat{f}_h^k\|_{\mathcal{Z}_h^k}^2 + R(f - \widehat{f}_h^k) \le \beta'(\mathcal{F},\delta)\}$, *where*

$$\beta'(\mathcal{F},\delta) \ge c'\left[(H+1)\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F},1/T)} + \sqrt{B}\right]^2. \tag{B.3}$$

With probability at least $1-\delta$, *for all* $(k,h,m) \in [K]\times[H]\times[M]$, *we have*

$$r_h(\cdot,\cdot) + P_h V_{h+1}^k(\cdot,\cdot) \in \mathcal{F}_h^k.$$

Proof. This is a direct implication of Lemma 31 with zero perturbance. □

B.1.3 Optimism

In this section, we will show that $\{Q_h^k\}_{(h,k)\in[H]\times[K]}$ is optimistic with high probability. Formally, we have the following lemma.

Lemma 33. *Set* $M = \ln(\frac{T|\mathcal{S}||\mathcal{A}|}{\delta})/\ln(\frac{1}{1-v})$ *in Algorithm 2. Conditioned on the event* $\mathcal{G}(K,H,\delta)$, *with probability at least* $1-\delta$, *for all* $s \in \mathcal{S}$, $a \in \mathcal{A}$, $h \in [H]$, $k \in [K]$, *we have*

$$Q_h^*(s,a) \le Q_h^k(s,a).$$

Proof. For timestep $H+1$, we have $Q_{H+1}^k = Q_{H+1}^* = 0$. By Lemma 32, there exists $\beta'(\mathcal{F},\delta)$ such that with probability at least $1-\delta$, for all $(k,h) \in [K]\times[H]$, we have

$$f_h^k(\cdot,\cdot) := r_h(\cdot,\cdot) + P_h V_{h+1}^k(\cdot,\cdot) \in \mathcal{F}_h^k,$$

where $\mathcal{F}_h^k = \{f \in \mathcal{F} \,|\, \|f - \widehat{f}_h^k\|_{\mathcal{Z}_h^k}^2 + R(f - \widehat{f}_h^k) \leq \beta'(\mathcal{F}, \delta)\}$.

Using notations introduced in Definition 5, let $g_{h,\sigma}^k$ be a function such that $\widetilde{f}_h^{k,m}(s,a) \geq \widehat{f}(s,a) + g_{h,\sigma}^k(s,a)$ holds with probability at least v. We set $M = \ln(\frac{T|\mathcal{S}||\mathcal{A}|}{\delta}) / \ln(\frac{1}{1-v})$ and then $\widetilde{f}_h^{k,m}(s,a) \geq \widehat{f}(s,a) + g_{h,\sigma}^k(s,a)$ with probability at least

$$1 - (1-v)^M = 1 - \frac{\delta}{T|\mathcal{S}||\mathcal{A}|},$$

for any $(k,h) \in [K] \times [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$. By union bound, we have $\widetilde{f}_h^{k,m}(s,a) \geq \widehat{f}(s,a) + g_{h,\sigma}^k(s,a)$ for all $(k,h) \in [K] \times [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$ with probability at least $1 - \delta$ and we have

$$\begin{aligned}
\widetilde{f}_h^k(s,a) &= \max_{m \in [M]} \widetilde{f}_h^{k,m}(s,a) \\
&\geq \widehat{f}_h^k(s,a) + g_{h,\sigma}^k(s,a) \\
&\geq \widehat{f}_h^k(s,a) + w(\mathcal{F}_h^k) \\
&\geq f_h^k(s,a),
\end{aligned}$$

where the second inequality is from Assumption 5 and the choise of σ as discussed in Appendix B.1.1. The last inequality follows from the definition of the width function and the previous observation that $f_h^k(\cdot,\cdot) \in \mathcal{F}_h^k$ with probability at least $1 - \delta$. Now we induct on h from $h = H$ to 1.

$$\begin{aligned}
Q_h^*(s,a) &= \min\{r_h(s,a) + P_h V_{h+1}^*(s,a), H\} \\
&= \min\{f_h^k(s,a) + P_h(V_{h+1}^* - V_{h+1}^k)(s,a), H\} \\
&\leq \min\{\widetilde{f}_h^k(s,a) + P_h(V_{h+1}^* - V_{h+1}^k)(s,a), H\} \\
&\leq \min\{\widetilde{f}_h^k(s,a), H\} \\
&= Q_h^k(s,a).
\end{aligned}$$

Thus,

$$V_h^*(s) = \max_a Q_h^*(s,a) \leq \max_a Q_h^k(s,a) = V_h^k(s).$$

where the second inequality is from $V_{h+1}^* \leq V_{h+1}^k$, which is implied by induction.

□

B.1.4 Regret Bound

We are now ready to provide the regret bound for Algorithm 2. The next lemma upper bounds the regret of the algorithm by the sum of the width functions.

Lemma 34 (Regret decomposition)**.** *Denote $b_h^k(s,a) = w(\mathcal{F}_h^k, s, a)$. Conditioned on the event $\mathcal{G}(K,H,\delta)$, with probability at least $1-\delta$, we have*

$$\text{Regret}(K) \le \sum_{k=1}^{K}\sum_{h=1}^{H} b_h^k(s_h^k, a_h^k) + \sum_{k=1}^{K}\sum_{h=1}^{H} \zeta_h^k,$$

where $\zeta_h^k = P(s_h^k, a_h^k)(V_{h+1}^k - V_{h+1}^{\pi^k}) - (V_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k))$ is a martingale difference sequence with respect to the filtration $\mathcal{G}_{h,2}^k$.

Proof. We condition on the good events in Lemma 31. For all $(k,h,m) \in [K]\times[H]\times[M]$, we have

$$\left\| r_h(\cdot,\cdot) + P_h V_{h+1}^k(\cdot,\cdot) - \widetilde{f}_h^{k,m} \right\|_{\mathcal{Z}_h^k}^2 + R(r_h(\cdot,\cdot) + P_h V_{h+1}^k(\cdot,\cdot) - \widetilde{f}_h^{k,m}) \le \beta(\mathcal{F},\delta).$$

Recall that $\mathcal{F}_h^k = \{f \mid \left\| r_h(\cdot,\cdot) + P_h V_{h+1}^k(\cdot,\cdot) - f\right\|_{\mathcal{Z}_h^k}^2 + R(r_h(\cdot,\cdot) + P_h V_{h+1}^k(\cdot,\cdot) - \widetilde{f}_h^{k,m}) \le \beta(\mathcal{F},\delta)\}$ is the confidence region. Then for $(k,h,m) \in [K]\times[H]\times[M]$, $\widetilde{f}_h^{k,m} \in \mathcal{F}_h^k$. Defining $b_h^k(s,a) = w(\mathcal{F}_h^k, s, a)$, for all $(k,h,m) \in [K]\times[H]\times[M]$ we have,

$$b_h^k(s,a) \ge \left| r(s,a) + P(s,a)V_{h+1}^k - \widetilde{f}_h^{k,m}(s,a) \right|.$$

As $Q_h^k(s,a) = \min\{\max_{m\in[M]}\{\widetilde{f}_h^{k,m}(\cdot,\cdot)\}, H-h+1\}$, we have

$$b_h^k(s,a) \ge \left| r(s,a) + P(s,a)V_{h+1}^k - Q_h^k(s,a) \right|.$$

By Lemma 33 and standard telescoping argument, we have

$$\begin{aligned}
&\text{Regret}(K) \leq \sum_{k=1}^{K} V_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \\
&= \sum_{k=1}^{K} Q_1^k(s_1^k, a_1^k) - Q_1^{\pi^k}(s_1^k, a_1^k) \\
&= \sum_{k=1}^{K} Q_1^k(s_1^k, a_1^k) - (r(s_1^k, a_1^k) + P(s_1^k, a_1^k)V_2^k) + (r(s_1^k, a_1^k) + P(s_1^k, a_1^k)V_2^k) - Q_1^{\pi^k}(s_1^k, a_1^k) \\
&\leq \sum_{k=1}^{K} b_1^k(s_1^k, a_1^k) + P(s_1^k, a_1^k)(V_2^k - V_2^{\pi^k}) \\
&= \sum_{k=1}^{K} b_1^k(s_1^k, a_1^k) + (V_2^k(s_2^k) - V_2^{\pi^k}(s_2^k)) + \zeta_1^k \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H} b_h^k(s_h^k, a_h^k) + \sum_{k=1}^{K}\sum_{h=1}^{H} \zeta_h^k.
\end{aligned}$$

□

Lemma 35 (Time inhomogeneous version of Lemma 10 in (Wang et al., 2020))**.** *Let $\mathcal{F}'$ be a subset of function class $\mathcal{F}$, consisting of all $f \in \mathcal{F}$ such that*

$$\|f - \upsilon\|_{\mathcal{Z}}^2 + R(f - \upsilon) \leq \beta(\mathcal{F}, \delta),$$

where $\upsilon = r + PV$ as in Assumption 7 and $\beta(\mathcal{F}, \delta)$ as defined in Lemma 31. With probability at least $1 - \delta$, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H} b_h^k(s_h^k, a_h^k) \leq H + 4H^3 \dim_{\mathcal{E}}(\mathcal{F}', 1/T) + H\sqrt{c\dim_{\mathcal{E}}(\mathcal{F}', 1/T) K \beta(\mathcal{F}, \delta)},$$

for some absolute constant $c > 0$.

Proof. Define

$$\mathcal{F}_h'^k = \{f \in \mathcal{F}' \,|\, \|f - \widehat{f}_h^k\|_{\mathcal{Z}_h^k}^2 \leq \beta(\mathcal{F}, \delta)\} = \mathcal{F}' \bigcap \{f \in \mathcal{F} \,|\, \|f - \widehat{f}_h^k\|_{\mathcal{Z}_h^k}^2 \leq \beta(\mathcal{F}, \delta)\}.$$

As $\mathcal{F}_h^k \subseteq \mathcal{F}'$ and $\mathcal{F}_h^k \subseteq \bigcap \{f \in \mathcal{F} \,|\, \|f - \widehat{f}_h^k\|_{\mathcal{Z}_h^k}^2 \leq \beta(\mathcal{F}, \delta)\}$, we have $\mathcal{F}_h^k \subseteq \mathcal{F}_h'^k$ and $w(\mathcal{F}_h^k, s, a) \leq w(\mathcal{F}_h'^k, s, a)$ for all s, a. By Assumption 7, $\mathcal{F}'$ has bounded eluder dimension.

Similar to Lemma 10 in (Wang et al., 2020), we have for any h,

$$\sum_{k=1}^{K} b_h^k(s_h^k, a_h^k) \leq \sum_{k=1}^{K} w(\mathcal{F}_h'^k, s, a) \leq 1+4H^2\text{dim}_{\mathcal{E}}(\mathcal{F}', 1/T)+\sqrt{c\text{dim}_{\mathcal{E}}(\mathcal{F}', 1/T)K\beta(\mathcal{F}, \delta)}.$$

Summing over all timestep h and we have the bound in the lemma.

□

Theorem 36. *Under all the assumptions, with probability at least $1-\delta$, Algorithm 2 achieves a regret bound of*

$$\text{Regret}(K) \leq 4H^3\text{dim}_{\mathcal{E}}(\mathcal{F}, 1/T) + \sqrt{\text{dim}_{\mathcal{E}}(\mathcal{F}, 1/T)\beta(\mathcal{F}, \delta)HT},$$

where

$$\beta(\mathcal{F}, \delta) = c' \left[(H+1+\sigma)\sqrt{\log(2/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)} + \sqrt{B + \sigma\sqrt{BD}}\right]^2,$$

for some constant c'.

Proof. By Assumption 7, we can consider $\mathcal{F}' \subseteq \mathcal{F}$ as the whole function class in the analysis because it includes all the $\mathcal{F}_h^k, \forall h, k$. By Azuma-Hoeffding inequality and Lemma 35, With probability at least $1-\delta$, we have

$$\begin{aligned}\text{Regret}(K) &\leq \sum_{k=1}^{K}\sum_{h=1}^{H} b_h^k(s_h^k, a_h^k) + \sum_{k=1}^{K}\sum_{h=1}^{H} \zeta_h^k \\ &\leq c'\left(H + 4H^3\text{dim}_{\mathcal{E}}(\mathcal{F}, 1/T) + H\sqrt{c\text{dim}_{\mathcal{E}}(\mathcal{F}, 1/T)K\beta(\mathcal{F}, \delta)} + H\sqrt{KH\log(1/\delta)}\right),\end{aligned}$$

for some constant c'. We plug in the definition of $\beta(\mathcal{F}, \delta)$ and $\sqrt{\gamma_k} = \widetilde{O}(\sigma)$, then we get the final bound. □

Remark 3. *For linear MDP, as shown in Section 5.2.1, we have*

$$\sigma = 2\sqrt{\beta'(\mathcal{F}, \delta)} = c' \left[(H+1)\sqrt{\log(2/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)} + \sqrt{B}\right]^2,$$

$B = 2Hd$ and $D = d$. In addition, we have $dim_{\mathcal{E}}(\mathcal{F}, 1/T) = \widetilde{O}(d)$ (Russo and Van Roy, 2013) and $\log \mathcal{N}(\mathcal{F}, 1/T) = \widetilde{O}(d)$. As a result, our bound implies a $\widetilde{O}(\sqrt{H^3d^3T})$ regret bound for linear MDP.

B.2 GFA With Model Misspecification

Assumption 9. *(Assumption 3 in Wang et al., 2020) For function class $\mathcal{F}$, there exists a real number ζ, such that for any $V : \mathcal{S} \to [0, H]$, there exists $g_V \in \mathcal{F}$ which satisfies*

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left| g_V(s,a) - r(s,a) - \sum_{s'\in\mathcal{S}} P(s'|s,a)V(s') \right| \le \zeta.$$

In addition, we assume g_V satisfies Assumption 6, i.e. $R(g_V) \le B$.

Lemma 37. *Consider a fixed $k \in [K]$ and a fixed $h \in [H]$. Let $\mathcal{Z}_h^k = \{(s_h^\tau, a_h^\tau)\}_{\tau\in[k-1]}$ and $\widetilde{\mathcal{D}}_{h,V}^k = \{(s_h^\tau, a_h^\tau, r_h^\tau + \xi_h^\tau + V(s_{h+1}^\tau))\}_{\tau\in[k-1]}$. Define $\widetilde{f}_{h,V}^k = \operatorname{argmin}_{f\in\mathcal{F}} \|f\|^2_{\widetilde{\mathcal{D}}_{h,V}^k} + \widetilde{R}(f)$. Conditioned on the good event $\mathcal{G}(K, H, \delta)$, with probability at least $1-\delta$, for a fixed $V : \mathcal{S} \to [0, H]$ and any $V' : \mathcal{S} \to [0, H]$ with $\|V' - V\|_\infty \le 1/T$, we have*

$$\begin{aligned}
&\left\| \widetilde{f}_{h,V'}(\cdot,\cdot) - r_h(\cdot,\cdot) - P_h V'(\cdot,\cdot) \right\|^2_{\mathcal{Z}_h^k} + R(\widetilde{f}_{h,V'}(\cdot,\cdot) - r_h(\cdot,\cdot) - P_h V'(\cdot,\cdot)) \\
\le & c' \left[(H + 1 + \sqrt{\gamma_k})\sqrt{\log(2/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)} + \sqrt{B + \sqrt{\gamma_k BD} + \zeta K(H + \sqrt{\gamma_k})} \right]^2,
\end{aligned}$$

for some constant c'.

Proof. Recall that for notational simplicity, we denote $[\mathbb{P}_h V_{h+1}](s,a) = \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot\,|\,s,a)} V_{h+1}(s')$. Now consider a fixed $V : \mathcal{S} \to [0, H]$, and define

$$f_V(\cdot,\cdot) = r_h(\cdot,\cdot) + P_h V(\cdot,\cdot). \tag{B.4}$$

By Assumption 9, there exists $g_V \in \mathcal{F}$ such that

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} |g_V(s,a) - f_V(s,a)| \le \zeta.$$

For any $f \in \mathcal{F}$, consider

$$\chi_h^\tau = 2(f(s_h^\tau, a_h^\tau) - f_V(s_h^\tau, a_h^\tau))(f_V(s_h^\tau, a_h^\tau) - r_h^\tau(s_h^\tau, a_h^\tau) - \xi_h^\tau - V(s_{h+1}^\tau)).$$

First we show that $\chi_h^\tau(f)$ is a martingale difference sequence with respect to the filtration $\mathcal{G}_{h,1}^\tau$.

$$\begin{aligned}\mathbb{E}[\chi_h^\tau(f)|\mathcal{G}_{h,1}^\tau] &= \mathbb{E}[2(f(s_h^\tau, a_h^\tau) - f_V(s_h^\tau, a_h^\tau))(f_V(s_h^\tau, a_h^\tau) - r_h^\tau(s_h^\tau, a_h^\tau) - \xi_h^\tau - V(s_{h+1}^\tau))|\mathcal{G}_{h,1}^\tau] \\ &= 2(f(s_h^\tau, a_h^\tau) - f_V(s_h^\tau, a_h^\tau))\mathbb{E}[(f_V(s_h^\tau, a_h^\tau) - r_h^\tau(s_h^\tau, a_h^\tau) - \xi_h^\tau - V(s_{h+1}^\tau))|\mathcal{G}_{h,1}^\tau] \\ &= 2(f(s_h^\tau, a_h^\tau) - f_V(s_h^\tau, a_h^\tau))(f_V(s_h^\tau, a_h^\tau) - r_h^\tau(s_h^\tau, a_h^\tau) - P_h V(s_h^\tau, a_h^\tau)) \\ &= 0.\end{aligned}$$

In addition, conditioning on good events $\mathcal{G}(K, H, \delta)$, we have

$$|\chi_h^\tau(f)| \le 2(H + 1 + \sqrt{\gamma_\tau})|f(s_h^\tau, a_h^\tau) - f_V(s_h^\tau, a_h^\tau)|.$$

As $\chi_h^\tau(f)$ is a martingale difference sequence conditioned on the filtration $\mathcal{G}_{h,1}^\tau$, by Azuma-Hoeffding inequality, we have

$$\mathbb{P}\left[\left|\sum_{\tau\in[k-1]} \chi_h^\tau(f)\right| \ge \epsilon\right] \le 2\exp\left(-\frac{\epsilon^2}{8(H+1+\sqrt{\gamma_\tau})^2\|f - f_V\|_{\mathcal{Z}_h^k}^2}\right).$$

Now we set

$$\begin{aligned}\epsilon &= \sqrt{8(H+1+\sqrt{\gamma_\tau})^2 \log\left(\frac{2N(\mathcal{F}, 1/T)}{\delta}\right)\|f - f_V\|_{\mathcal{Z}_h^k}^2} \\ &\le 4(H+1+\sqrt{\gamma_\tau})\|f - f_V\|_{\mathcal{Z}_h^k}\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)}.\end{aligned}$$

With union bound, for all $g \in \mathcal{C}(\mathcal{F}, 1/T)$, with probability at least $1 - \delta$ we have

$$\left|\sum_{(\tau)\in[k-1]} \xi_h^\tau(g)\right| \le 4(H+1+\sqrt{\gamma_\tau})\|f - f_V\|_{\mathcal{Z}_h^k}\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)}.$$

Thus, for all $f \in \mathcal{F}$, there exists $g \in \mathcal{C}(\mathcal{F}, 1/T)$ such that $\|f - g\|_\infty \le 1/T$ and ,

$$\begin{aligned}&\left|\sum_{(\tau)\in[k-1]} \chi_h^\tau(f)\right| \le \left|\sum_{(\tau)\in[k-1]} \chi_h^\tau(g)\right| + 2(H+1+\sqrt{\gamma_\tau}) \\ &\le 4(H+1+\sqrt{\gamma_\tau})\|g - f_V\|_{\mathcal{Z}_h^k}\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)} + 2(H+1+\sqrt{\gamma_\tau}) \\ &\le 4(H+1+\sqrt{\gamma_\tau})(\|f - f_V\|_{\mathcal{Z}_h^k} + 1)\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)} + 2(H+1+\sqrt{\gamma_\tau})\end{aligned}$$

For $V' : \mathcal{S} \to [0, H]$ such that $\|V - V'\|_\infty \le 1/T$, we have $\|f_{V'} - f_V\|_\infty \le \|V' - V\|_\infty \le 1/T$.

For any $f \in \mathcal{F}$, we have

$$
\begin{aligned}
&\|f\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} - \|f_{V'}\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} \\
=&\|f - f_{V'}\|^2_{\mathcal{Z}^k_h} + 2 \sum_{(s^\tau_h, a^\tau_h) \in \mathcal{Z}^k_h} (f(s^\tau_h, a^\tau_h) - f_{V'}(s^\tau_h, a^\tau_h))(f_{V'}(s^\tau_h, a^\tau_h) - r^\tau_h(s^\tau_h, a^\tau_h) - \xi^\tau_h - V'(s^\tau_{h+1})) \\
\ge&\|f - f_{V'}\|^2_{\mathcal{Z}^k_h} + 2 \sum_{(s^\tau_h, a^\tau_h) \in \mathcal{Z}^k_h} (f(s^\tau_h, a^\tau_h) - f_V(s^\tau_h, a^\tau_h))(f_V(s^\tau_h, a^\tau_h) - r^\tau_h(s^\tau_h, a^\tau_h) - \xi^\tau_h - V(s^\tau_{h+1})) \\
&- 4(H + 1 + \sqrt{\gamma_k})\|V' - V\|_\infty |\mathcal{Z}^k_h| \\
\ge&\|f - f_{V'}\|^2_{\mathcal{Z}^k_h} + \sum_{(\tau,h) \in [k-1] \times [H]} \chi^\tau_h(f) - 4(H + 1 + \sqrt{\gamma_k}) \\
\ge&\|f - f_{V'}\|^2_{\mathcal{Z}^k_h} - 4(H + 1 + \sqrt{\gamma_k})(\|f - f_V\|_{\mathcal{Z}^k_h} + 1)\sqrt{\log(2/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)} - 6(H + 1 + \sqrt{\gamma_k}) \\
\ge&\|f - f_{V'}\|^2_{\mathcal{Z}^k_h} - 4(H + 1 + \sqrt{\gamma_k})(\|f - f_{V'}\|_{\mathcal{Z}^k_h} + 2)\sqrt{\log(2/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)} - 6(H + 1 + \sqrt{\gamma_k}).
\end{aligned}
$$

In addition, by Assumption 6, we have

$$
\begin{aligned}
&\widetilde{R}(f) - \widetilde{R}(f_{V'}) = \sum_i [p_i(f) - \xi'_i]^2 - \sum_i [p_i(f_{V'}) - \xi'_i]^2 \\
=&R(f) - R(f_{V'}) - 2\sum_i \xi'_i(p_i(f) - p_i(f_{V'})) \ge cR(f - f_{V'}) - 2R(f_{V'}) - 2\sum_i \sqrt{\gamma_k} p_i(f_{V'}) \\
\ge&cR(f - f_{V'}) - 2B - 2\sqrt{\gamma_k}\sqrt{BD}.
\end{aligned}
$$

Summing the above two inequalities we have

$$
\|f\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} + \widetilde{R}(f) - \|f_{V'}\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} - \widetilde{R}(f_{V'}) \ge \|f - f_{V'}\|^2_{\mathcal{Z}^k_h} + cR(f - f_{V'}) - C,
$$

where $C = 4(H + 1 + \sqrt{\gamma_k})(\|f - f_{V'}\|_{\mathcal{Z}^k_h} + 2)\sqrt{\log(2/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)} + 6(H + 1 + \sqrt{\gamma_k}) + 2B + 2\sqrt{\gamma_k}\sqrt{DB}$.

Now we try to replace the $f_{V'}$ in the RHS with g_V'.

$$\begin{aligned}
&\|f_{V'}\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} - \|g_{V'}\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} \\
=& \sum_{\tau\in[k-1]} (f_{V'}(s^\tau_h, a^\tau_h) - (r^\tau_h + \xi^\tau_h + V(s^\tau_{h+1})))^2 - \sum_{\tau\in[k-1]} (g_{V'}(s^\tau_h, a^\tau_h) - (r^\tau_h + \xi^\tau_h + V(s^\tau_{h+1})))^2 \\
=& \sum_{\tau\in[k-1]} (f_{V'}(s^\tau_h, a^\tau_h) - g_{V'}(s^\tau_h, a^\tau_h))(f_{V'}(s^\tau_h, a^\tau_h) + g_{V'}(s^\tau_h, a^\tau_h) - 2(r^\tau_h + \xi^\tau_h + V(s^\tau_{h+1}))) \\
\geq& -\zeta K(4H + 2\sqrt{\gamma_k}).
\end{aligned}$$

By the boundedness of the regularizer (Assumption 6), we have

$$\|f_{V'}\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} + \widetilde{R}(f_{V'}) - \|g_{V'}\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} - \widetilde{R}(g_{V'}) \geq -\zeta K(4H + 2\sqrt{\gamma_k}) - B.$$

Thus we have

$$\begin{aligned}
&\|f\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} + \widetilde{R}(f) - \|g_{V'}\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} - \widetilde{R}(g_{V'} \\
) \geq& \|f\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} + \widetilde{R}(f) - \|f_{V'}\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} - \widetilde{R}(f_{V'}) - \zeta K(4H + 2\sqrt{\gamma_k}) - B \\
\geq& \|f - f_{V'}\|^2_{\mathcal{Z}^k_h} + cR(f - f_{V'}) - C - \zeta K(4H + 2\sqrt{\gamma_k}) - B.
\end{aligned}$$

As $\widetilde{f}_{h,V'}$ is the minimizer of $\|f\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} + \widetilde{R}(f)$ for $f \in \mathcal{F}$ and note that $g_{V'} \in \mathcal{F}$, we have

$$\begin{aligned}
&\|\widetilde{f}_{h,V'} - f_{V'}\|^2_{\mathcal{Z}^k_h} + cR(\widetilde{f}_{h,V'} - f_{V'}) \\
\leq& c' \left[(H + 1 + \sqrt{\gamma_k})\sqrt{\log(2/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)} + \sqrt{B + \sqrt{\gamma_k BD} + \zeta K(H + \sqrt{\gamma_k})}\right]^2.
\end{aligned}$$

To prove the above argument, we use the inequality that if we have $x^2 + y \leq ax + b$ for positive a, b, y, then $x \leq a + \sqrt{b}$ and $x^2 + y \leq (a + \sqrt{b})^2$. In addition, we can remove c by replacing c' with $c'/\min\{1, c\}$ and then we get the final bound.

□

Lemma 38. *(Misspecified Confidence Region) Let* $\mathcal{F}^{k,m}_h = \{f \in \mathcal{F} \| \|f - \widetilde{f}^{k,m}_h\|^2_{\mathcal{Z}^k_h} + R(f - \widetilde{f}^{k,m}_h) \leq \beta(\mathcal{F}, \delta)\}$*, where*

$$\beta(\mathcal{F}, \delta) = c' \left[(H + 1 + \sqrt{\gamma_k})\sqrt{\log(2/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)} + \sqrt{B + \sqrt{\gamma_k BD} + \zeta K(H + \sqrt{\gamma_k})}\right]^2. \tag{B.5}$$

Conditioned on the event $\mathcal{G}(K,H,\delta)$, with probability at least $1-\delta$, for all $(k,h,m) \in [K] \times [H] \times [M]$, we have

$$r_h(\cdot,\cdot) + P_h V_{h+1}^k(\cdot,\cdot) \in \mathcal{F}_h^{k,m}.$$

Proof. With Lemma 37, the proof is same as Lemma 31. □

Theorem 39. *Under all the assumptions, with probability at least $1-\delta$, Algorithm 2 achieves a regret bound of*

$$\text{Regret}(K) \leq 4H^3 \text{dim}_{\mathcal{E}}(\mathcal{F}, 1/T) + \sqrt{\text{dim}_{\mathcal{E}}(\mathcal{F}, 1/T)\beta(\mathcal{F},\delta)HT},$$

where

$$\beta(\mathcal{F},\delta) = c' \left[(H+1+\sigma)\sqrt{\log(2/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)} + \sqrt{B + \sigma\sqrt{BD} + \zeta K(H+\sigma)}\right]^2,$$

for some constant c'.

Proof. With Lemma 38, the proof is the same as Theorem 36. □

B.3 LSVI-PHE with linear function approximation

In this section, we prove Theorem 12. Our analysis specilized to linear MDP setting is simpler and may provide additional insights. In addition, compared to GFA setting, we improve the bound for M and it no longer depends on $|\mathcal{S}|$ or $|\mathcal{A}|$. We first introduce the notation and few definitions that are used throughout this section. Upon presenting lemmas and their proofs, finally we combine the lemmas to prove Theorem 12.

Definition 10 (Model prediction error)**.** *For all $(k,h) \in [K] \times [H]$, we define the model prediction error associated with the reward r_h^k,*

$$l_h^k(s,a) = r_h^k(s,a) + \mathbb{P}_h V_{h+1}^k(s,a) - Q_h^k(s,a).$$

This depicts the prediction error using V_{h+1}^k instead of $V_{h+1}^{\pi^k}$ in the Bellman equations.

Definition 11 (Unperturbed estimated parameter)**.** *For all $(k,h) \in [K]\times[H]$, we define the unperturbed estimated parameter as*

$$\widehat{\theta}_h^k = (\Lambda_h^k)^{-1}\left(\sum_{\tau=1}^{k-1}[r_h^\tau + V_{h+1}^k(s_{h+1}^\tau)]\phi(s_h^\tau, a_h^\tau)\right).$$

Moreover, we denote the difference between the perturbed estimated parameter $\widetilde{\theta}_h^{k,j}$ and the unperturbed estimated parameter $\widehat{\theta}_h^k$ as

$$\zeta_h^{k,j} = \widetilde{\theta}_h^{k,j} - \widehat{\theta}_h^k.$$

B.3.1 Concentration

Our first lemma characterizes the difference between the perturbed estimated parameter $\widetilde{\theta}_h^{k,j}$ and the unperturbed estimated parameter $\widehat{\theta}_h^k$.

Proposition 2 (restatement of Proposition 1)**.** *In step 9 of Algorithm 3, conditioned on all the randomness except $\{\epsilon_h^{k,i,j}\}_{(i,j)\in[k-1]\times[M]}$ and $\{\xi_h^{k,j}\}_{j\in[M]}$, the estimated parameter $\widetilde{\theta}_h^{k,j}$ satisfies*

$$\zeta_h^{k,j} = \widetilde{\theta}_h^{k,j} - \widehat{\theta}_h^k \sim N(0, \sigma^2(\Lambda_h^k)^{-1}),$$

where $\widehat{\theta}_h^k = (\Lambda_h^k)^{-1}(\sum_{\tau=1}^{k-1}[r_h^\tau + V_{h+1}^k(s_{h+1}^\tau)]\phi(s_h^\tau, a_h^\tau))$ is the unperturbed estimated parameter from Definiton 11.

Proof. From Algorithm 3, note that

$$\begin{aligned}
\widetilde{\theta}_h^{k,j} &= (\Lambda_h^k)^{-1}(\rho_h^k + \xi_h^{k,j}) \\
&= (\Lambda_h^k)^{-1}\left(\sum_{\tau=1}^{k-1}\left([r_h^\tau + V_{h+1}^k(s_{h+1}^\tau) + \epsilon_h^{k,\tau,j}]\phi(s_h^\tau, a_h^\tau)\right) + \xi_h^{k,j}\right) \\
&= (\Lambda_h^k)^{-1}\left(\sum_{\tau=1}^{k-1}[r_h^\tau + V_{h+1}^k(s_{h+1}^\tau)]\phi(s_h^\tau, a_h^\tau)\right) + (\Lambda_h^k)^{-1}\left(\sum_{\tau=1}^{k-1}\epsilon_h^{k,\tau,j}\phi(s_h^\tau, a_h^\tau) + \xi_h^{k,j}\right) \\
&= \widehat{\theta}_h^k + (\Lambda_h^k)^{-1}\left(\sum_{\tau=1}^{k-1}\epsilon_h^{k,\tau,j}\phi(s_h^\tau, a_h^\tau) + \xi_h^{k,j}\right).
\end{aligned}$$

Since $\epsilon_h^{k,\tau,j} \sim N(0,\sigma^2)$, note that for $\tau \in [k-1]$,

$$\epsilon_h^{k,\tau,j}\phi(s_h^\tau, a_h^\tau) \sim N(0, \sigma^2\phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top).$$

Now, since $\xi_h^{k,j} \sim \mathcal{N}(0, \sigma^2 \lambda I_d)$,

$$
\begin{aligned}
(\Lambda_h^k)^{-1} \left(\sum_{\tau=1}^{k-1} \epsilon_h^{k,\tau,j} \phi(s_h^\tau, a_h^\tau) + \xi_h^{k,j} \right) &\sim (\Lambda_h^k)^{-1} \cdot N\left(0, \sigma^2 \left(\sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau) \phi(s_h^\tau, a_h^\tau)^\top + \lambda I_d \right)\right) \\
&\sim (\Lambda_h^k)^{-1} \cdot N\left(0, \sigma^2 \Lambda_h^k\right) \\
&\sim N(0, \sigma^2 (\Lambda_h^k)^{-1}).
\end{aligned}
$$

Thus, we have

$$
\zeta_h^{k,j} = \widetilde{\theta}_h^{k,j} - \widehat{\theta}_h^k \sim N(0, \sigma^2 (\Lambda_h^k)^{-1}).
$$

□

Lemma 40 (Lemma B.1 in Jin et al., 2020)**.** *Under Definition 6 of linear MDP, for any fixed policy π, let $\{\theta_h^\pi\}_{h \in [H]}$ be the corresponding weights such that $Q_h^\pi(s,a) = \langle \phi(s,a), \theta_h^\pi \rangle$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$. Then for all $h \in [H]$, we have*

$$
\|\theta_h^\pi\| \le 2H\sqrt{d}.
$$

Our next lemma states that the unperturbed estimated weight $\widehat{\theta}_h^k$ is bounded.

Lemma 41. *For any $(k,h) \in [K] \times [H]$, the unperturbed estimated weight $\widehat{\theta}_h^k$ in Definition 11 satisfies*

$$
\|\widehat{\theta}_h^k\| \le 2H\sqrt{kd/\lambda}.
$$

Proof. We have

$$
\begin{aligned}
\|\widehat{\theta}_h^k\| &= \left\| (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} [r_h^\tau(s_h^\tau, a_h^\tau) + V_{h+1}^k(s_{h+1}^\tau)] \cdot \phi(s_h^\tau, a_h^\tau) \right\| \\
&= \left\| (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} [r_h^\tau(s_h^\tau, a_h^\tau) + \max_{a \in \mathcal{A}} Q_{h+1}^k(s_{h+1}^\tau, a)] \cdot \phi(s_h^\tau, a_h^\tau) \right\| \\
&\le \frac{1}{\sqrt{\lambda}} \sqrt{k-1} \Big(\sum_{\tau=1}^{k-1} \left\| [r_h^\tau(s_h^\tau, a_h^\tau) + \max_{a \in \mathcal{A}} Q_{h+1}^k(s_{h+1}^\tau, a)] \cdot \phi(s_h^\tau, a_h^\tau) \right\|_{(\Lambda_h^k)^{-1}}^2 \Big)^{1/2} \\
&\le \frac{2H}{\sqrt{\lambda}} \sqrt{k-1} \Big(\sum_{\tau=1}^{k-1} \|\phi(s_h^\tau, a_h^\tau)\|_{(\Lambda_h^k)^{-1}}^2 \Big)^{1/2} \\
&\le 2H\sqrt{kd/\lambda}.
\end{aligned}
$$

Here, the first inequality follows from Lemma 55. The second inequality follows from the truncation of Q_h^k to the range $[0, H-h+1]$ in Line 11 of Algorithm 3. The last inequality is due to Lemma 53. □

For the ease of exposition, we now define the values $\beta_k(\delta)$, $\nu_k(\delta)$ and $\gamma_k(\delta)$ which we use to define our high confidence bounds.

Definition 12 (Noise bounds)**.** *For any $\delta > 0$ and some large enough constants c_1,c_2 and c_3, let*

$$\begin{aligned}\sqrt{\beta_k(\delta)} &\overset{def}{=} c_1 H\sqrt{d\log(Hdk/\delta)},\\ \sqrt{\nu_k(\delta)} &\overset{def}{=} c_2 H\sqrt{d\log(Hdk/\delta)},\\ \sqrt{\gamma_k(\delta)} &\overset{def}{=} c_3\sqrt{d\nu_k(\delta)\log(d/\delta)}.\end{aligned}$$

Definition 13 (Noise distribution)**.** *In Algorithm 3, we set the following values for σ*

$$\sigma_k = 2\sqrt{\nu_k(\delta)}.$$

Thus for all $j \in [M]$, we have,

$$\{\xi_h^{k,j}\} \sim \mathcal{N}\big(0, 4\nu_k(\delta)(\Lambda_h^k)^{-1}\big).$$

Now, we define some events based on the characterization of the random variable $\zeta_h^{k,j}$ as defined in Definition 11.

Definition 14 (Good events)**.** *For any $\delta > 0$, we define the following random events*

$$\begin{aligned}\mathcal{G}_h^k(\zeta,\delta) &\overset{def}{=} \Big\{\max_{j\in[M]} \|\zeta_h^{k,j}\|_{\Lambda_h^k} \le \sqrt{\gamma_k(\delta)}\Big\},\\ \mathcal{G}(K,H,\delta) &\overset{def}{=} \bigcap_{k\le K}\bigcap_{h\le H} \mathcal{G}_h^k(\zeta,\delta).\end{aligned}$$

Next, we present our main concentration lemma in this section.

Lemma 42. *Let $\lambda = 1$ in Algorithm 3. For any fixed $\delta > 0$, conditioned on the event $\mathcal{G}(K,H,\delta)$, we have for all $(k,h) \in [K]\times[H]$,*

$$\Big\|\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)\big[\big(V_{h+1}^k - \mathbb{P}_h V_{h+1}^k\big)(s_h^\tau,a_h^\tau)\big]\Big\|_{(\Lambda_h^k)^{-1}} \le c_1 H\sqrt{d\log\left(Hdk/\delta\right)}, \quad \text{(B.6)}$$

with probability at least $1-\delta$ for some constant $c_1 > 0$.

Proof. From Lemma 41, we know, for all $(k,h) \in [K] \times [H]$, we have $\|\widehat{\theta}_h^k\| \le 2H\sqrt{kd/\lambda}$. In addition, by construction of Λ_{h+1}^k, the minimum eigenvalue of Λ_{h+1}^k is lower bounded by λ. Thus we have $\sqrt{\lambda}\|\zeta_{h+1}^{k,j}\| \le \|\zeta_{h+1}^{k,j}\|_{\Lambda_{h+1}^k} \le \sqrt{\gamma_k(\delta)}$. Finally, triangle inequality implies, $\|\widetilde{\theta}_{h+1}^{k,j}\| = \|\widehat{\theta}_{h+1}^k + \zeta_{h+1}^{k,j}\| \le 2H\sqrt{kd/\lambda} + \sqrt{\gamma_k(\delta)/\lambda}$ for all $j \in [M]$. Combining Lemma 56 and Lemma 58, we have that, for any $\varepsilon > 0$ and $\delta > 0$, with probability at least $1-\delta$,

$$\begin{aligned}
&\Big\|\sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)[(V_{h+1}^k - \mathbb{P}_h V_{h+1}^k)(s_h^\tau, a_h^\tau)]\Big\|_{(\Lambda_h^k)^{-1}} \\
&\le \Big(4H^2\Big[\frac{d}{2}\log\Big(\frac{k+\lambda}{\lambda}\Big) + d\log\Big(1 + \frac{4H\sqrt{kd/\lambda} + 2\sqrt{\gamma_k(\delta)/\lambda}}{\varepsilon}\Big) + \log\frac{1}{\delta}\Big] + \frac{8k^2\varepsilon^2}{\lambda}\Big)^{1/2} \\
&\le \Big(4H^2\Big[\frac{d}{2}\log\Big(\frac{k+\lambda}{\lambda}\Big) + d\log\Big(\frac{3(2H\sqrt{kd/\lambda} + \sqrt{\gamma_k(\delta)/\lambda})}{\varepsilon}\Big) + \log\frac{1}{\delta}\Big] + \frac{8k^2\varepsilon^2}{\lambda}\Big)^{1/2} \\
&\le 2H\Big[\frac{d}{2}\log\Big(\frac{k+\lambda}{\lambda}\Big) + d\log\Big(\frac{3(2H\sqrt{kd/\lambda} + \sqrt{\gamma_k(\delta)/\lambda})}{\varepsilon}\Big) + \log\frac{1}{\delta}\Big]^{1/2} + \frac{2\sqrt{2}k\varepsilon}{\sqrt{\lambda}} \\
&\le 2H\sqrt{d}\Big[\frac{1}{2}\log\Big(\frac{k+\lambda}{\lambda}\Big) + \log\Big(\frac{3(2H\sqrt{kd/\lambda} + \sqrt{\gamma_k(\delta)/\lambda})}{\varepsilon}\Big) + \log\frac{1}{\delta}\Big]^{1/2} + \frac{2\sqrt{2}k\varepsilon}{\sqrt{\lambda}}.
\end{aligned} \tag{B.7}$$

Setting $\lambda = 1$, $\varepsilon = H\sqrt{d}/k$ and substituting $\sqrt{\gamma_k(\delta)} = c_3\sqrt{d\nu_k(\delta)\log(d/\delta)} \le c_4 Hd\log(Hdk/\delta)$ for some constant $c_4 > 0$, we get

$$\begin{aligned}
&\Big\|\sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)[(V_{h+1}^k - \mathbb{P}_h V_{h+1}^k)(s_h^\tau, a_h^\tau)]\Big\|_{(\Lambda_h^k)^{-1}} \\
&\le 2H\sqrt{d}\left[\frac{1}{2}\log(k+1) + \log(1/\delta) + \log\frac{3k[2H\sqrt{dk} + c_4Hd\log(Hdk/\delta)]}{H\sqrt{d}}\right]^{1/2} + 2\sqrt{2}H\sqrt{d} \\
&\le c_1 H\sqrt{d\log(Hdk/\delta)},
\end{aligned} \tag{B.8}$$

for some constant $c_1 > 0$.

□

Lemma 43. *Let $\lambda = 1$ in Algorithm 3. For any $\delta > 0$, conditioned on the event $\mathcal{G}(K,H,\delta)$, for any $(h,k) \in [H] \times [K]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have*

$$\big|\phi(s,a)^\top\widehat{\theta}_h^k - r_h^k(s,a) - \mathbb{P}_h V_{h+1}^k(s,a)\big| \le c_2 H\sqrt{d\log(Hdk/\delta)}\big\|\phi(s,a)\big\|_{(\Lambda_h^k)^{-1}},$$

with probability $1-\delta$, where $c_2 > 0$ is a constant.

Proof. Let us denote the inner product over $\mathcal{S}$ by $\langle\cdot,\cdot\rangle_{\mathcal{S}}$. Using linear MDP assumption for transition kernel from Definition 6, we get

$$\begin{aligned}
\mathbb{P}_h V_{h+1}^k(s,a) &= \phi(s,a)^\top \langle \mu_h, V_{h+1}^k\rangle_{\mathcal{S}} \\
&= \phi(s,a)^\top (\Lambda_h^k)^{-1}\Lambda_h^k \langle \mu_h, V_{h+1}^k\rangle_{\mathcal{S}} \\
&= \phi(s,a)^\top (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau)^\top + \lambda I\Big)\langle \mu_h, V_{h+1}^k\rangle_{\mathcal{S}} \\
&= \phi(s,a)^\top (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau)^\top \langle \mu_h, V_{h+1}^k\rangle_{\mathcal{S}} + \lambda I\langle \mu_h, V_{h+1}^k\rangle_{\mathcal{S}}\Big) \\
&= \phi(s,a)^\top (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)(\mathbb{P}_h V_{h+1}^k)(s_h^\tau,a_h^\tau) + \lambda I\langle \mu_h, V_{h+1}^k\rangle_{\mathcal{S}}\Big),
\end{aligned} \tag{B.9}$$

where in the last line we rely on the definition of $\mathbb{P}_h$.

Using (B.9) we obtain,

$$\phi(s,a)^\top\widehat{\theta}_h^k - r_h^k(s,a) - (\mathbb{P}_h V_{h+1}^k)(s,a) \tag{B.10}$$

$$\begin{aligned}
&= \phi(s,a)^\top (\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\big[r_h^\tau(s_h^\tau,a_h^\tau) + V_{h+1}^k(s_{h+1}^\tau)\big]\cdot\phi(s_h^\tau,a_h^\tau) - r_h^k(s,a) \\
&\qquad - \phi(s,a)^\top (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)(\mathbb{P}_h V_{h+1}^k)(s_h^\tau,a_h^\tau) + \lambda I\langle \mu_h, V_{h+1}^k\rangle_{\mathcal{S}}\Big) \\
&= \underbrace{\phi(s,a)^\top (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)\big[\big(V_{h+1}^k - \mathbb{P}_h V_{h+1}^k\big)(s_h^\tau,a_h^\tau)\big]\Big)}_{\text{(i)}} \\
&\qquad + \underbrace{\phi(s,a)^\top (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1} r_h^\tau(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau)\Big) - r_h^k(s,a)}_{\text{(ii)}} \qquad - \underbrace{\lambda\phi(s,a)^\top(\Lambda_h^k)^{-1}\langle \mu_h, V_{h+1}^k\rangle_{\mathcal{S}}}_{\text{(iii)}}.
\end{aligned} \tag{B.11}$$

In the following we will analyze the each of the three terms in (B.10) separately and derive high probability bound for each of them.

Term (i). Since $(\Lambda_h^k)^{-1} \succ 0$, by Cauchy-Schwarz inequality and Lemma 42, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\phi(s,a)^\top(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)\big[\big(V_{h+1}^k-\mathbb{P}_hV_{h+1}^k\big)(s_h^\tau,a_h^\tau)\big]\Big)\\
&\le \Big\|\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)\big[\big(V_{h+1}^k-\mathbb{P}_hV_{h+1}^k\big)(s_h^\tau,a_h^\tau)\big]\Big\|_{(\Lambda_h^k)^{-1}}\big\|\phi(s,a)\big\|_{(\Lambda_h^k)^{-1}}\\
&\le \sqrt{\beta_k(\delta)}\big\|\phi(s,a)\big\|_{(\Lambda_h^k)^{-1}}. \qquad \text{(B.12)}
\end{aligned}$$

Term (ii). Note that

$$\begin{aligned}
&\phi(s,a)^\top(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1}r_h^\tau(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau)\Big)-r_h^k(s,a)\\
&=\phi(s,a)^\top(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1}r_h^\tau(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau)\Big)-\phi(s,a)^\top w_h\\
&=\phi(s,a)^\top(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1}r_h^\tau(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau)-\Lambda_h^kw_h\Big)\\
&=\phi(s,a)^\top(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1}r_h^\tau(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau)-\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau)^\top w_h-\lambda Iw_h\Big)\\
&=\phi(s,a)^\top(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1}r_h^\tau(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau)-\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)r_h^\tau(s_h^\tau,a_h^\tau)-\lambda Iw_h\Big)\\
&=-\lambda\phi(s,a)^\top(\Lambda_h^k)^{-1}w_h, \qquad \text{(B.13)}
\end{aligned}$$

where in the penultimate step, we used the fact $r_h(s,a)=\langle\phi(s,a),w_h\rangle$ from Definition 6. Applying Cauchy-Schwarz inequality we obtain,

$$\begin{aligned}
-\lambda\phi(s,a)^\top(\Lambda_h^k)^{-1}w_h &\le \lambda\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}\|w_h\|_{(\Lambda_h^k)^{-1}}\\
&\le \sqrt{\lambda}\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}\|w_h\|_2\\
&\le \sqrt{\lambda d}\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}. \qquad \text{(B.14)}
\end{aligned}$$

Here the second inequality follows by observing that the smallest eigenvalue of Λ_h^k is at least λ and thus the largest eigenvalue of $(\Lambda_h^k)^{-1}$ is at most $1/\lambda$. The last inequality follows from Definition 6. Combining (B.13) and (B.14) we get

$$\phi(s,a)^\top(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1}r_h^\tau(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau)\Big)-r_h^k(s,a)\le\sqrt{\lambda d}\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}. \qquad \text{(B.15)}$$

Term (iii). Similar to (B.14), applying Cauchy-Schwarz inequality, we get

$$\begin{aligned}
-\lambda\phi(s,a)^\top(\Lambda_h^k)^{-1}\langle\mu_h, V_{h+1}^k\rangle_{\mathcal{S}} &\le \lambda\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}\|\langle\mu_h, V_{h+1}^k\rangle_{\mathcal{S}}\|_{(\Lambda_h^k)^{-1}} \\
&\le \sqrt{\lambda}\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}\|\langle\mu_h, V_{h+1}^k\rangle_{\mathcal{S}}\|_2 \\
&\le \sqrt{\lambda}\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}\Big(\sum_{\tau=1}^d \|\mu_h^\tau\|_1^2\Big)^{\frac{1}{2}}\|V_{h+1}^k\|_\infty \\
&\le H\sqrt{\lambda d}\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}. \qquad \text{(B.16)}
\end{aligned}$$

Here the second inequality follows using the same observation we did for **term (ii)**. The last inequality follows from $\sum_{\tau=1}^d \|\mu_h^\tau\|_1^2 \le d$ in Definition 6 and the clipping operation performed in Line 12 of Algorithm 3. Now combining (B.12), (B.15) and (B.16), and letting $\lambda = 1$, we get,

$$\begin{aligned}
&\big|\phi(s,a)^\top\widehat{\theta}_h^k - r_h^k(s,a) - \mathbb{P}_h V_{h+1}^k(s,a)\big| && \text{(B.17)} \\
&\le (\sqrt{\beta_k(\delta)} + H\sqrt{d} + \sqrt{d})\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} && \text{(B.18)} \\
&= (c_1 H\sqrt{d\log(Hdk/\delta)} + H\sqrt{d} + \sqrt{d})\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} && \text{(B.19)} \\
&\le c_2 H\sqrt{d\log(Hdk/\delta)}\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}, && \text{(B.20)}
\end{aligned}$$

with probability $1-\delta$ for some constant $c_2 > 0$.

In addition, If we set $\theta_h^k : \phi(\cdot,\cdot)^\top\theta_h^k = r_h^k(\cdot,\cdot) + \mathbb{P}_h V_{h+1}^k(\cdot,\cdot)$ to be the true parameter and $\Delta\theta_h^k = \theta_h^k - \widehat{\theta}_h^k$ to be the regression error, then from the analysis above we can derive that $\|\Delta\theta_h^k\|_{\Lambda_h^k} \le \sqrt{\nu_k(\delta)} = c_2 H\sqrt{d\log(Hdk/\delta)}$. □

Lemma 44 (stochastic upper confidence bound). *Let $\lambda = 1$ in Algorithm 3. For any $\delta > 0$, conditioned on the event $\mathcal{G}(K,H,\delta)$, for any $(h,k) \in [H]\times[K]$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$, with probability at least $1-(\delta + c_0^M)$, we have*

$$l_h^k(s,a) \le 0,$$

and

$$-l_h^k(s,a) \le \Big(\sqrt{\nu_k(\delta)} + \sqrt{\gamma_k(\delta)}\Big)\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}},$$

where $c_0 = \Phi(1)$.

Proof. Applying Lemma 43, for any $(h,k) \in [H] \times [K]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have,

$$\begin{aligned}\left|r_h^k(s,a) + \mathbb{P}_h V_{h+1}^k(s,a) - \phi(s,a)^\top \widehat{\theta}_h^k\right| &\leq c_2 H \sqrt{d \log(Hdk/\delta)} && \text{(B.21)}\\ &= \sqrt{\nu_k(\delta)} \big\|\phi(s,a)\big\|_{(\Lambda_h^k)^{-1}}, && \text{(B.22)}\end{aligned}$$

with probability at least $1-\delta$.

As we are conditioning on the event $\mathcal{G}(K,H,\delta)$, for any $(h,k) \in [H] \times [K]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\max_{j\in[M]} \big|\phi(s,a)^\top \zeta_h^{k,j}\big| \leq \sqrt{\gamma_k(\delta)} \big\|\phi(s,a)\big\|_{(\Lambda_h^k)^{-1}}. \tag{B.23}$$

Now from the definition of model prediction error, using (B.21) and (B.23), we get, with probability $1-\delta$,

$$\begin{aligned}-l_h^k(s,a) &= Q_h^k(s,a) - r_h^k(s,a) - \mathbb{P}_h V_{h+1}^k(s,a)\\ &= \min\{\max_{j\in[M]} \phi(s,a)^\top(\widehat{\theta}_h^k + \zeta_h^{k,j}), H\} - r_h^k(s,a) - \mathbb{P}_h V_{h+1}^k(s,a)\\ &\leq \max_{j\in[M]} \phi(s,a)^\top(\widehat{\theta}_h^k + \zeta_h^{k,j}) - r_h^k(s,a) - \mathbb{P}_h V_{h+1}^k(s,a)\\ &= \max_{j\in[M]} \phi(s,a)^\top \zeta_h^{k,j} - \big(r_h^k(s,a) + \mathbb{P}_h V_{h+1}^k(s,a) - \phi(s,a)^\top \widehat{\theta}_h^k\big)\\ &\leq \big|r_h^k(s,a) + \mathbb{P}_h V_{h+1}^k(s,a) - \phi(s,a)^\top \widehat{\theta}_h^k\big| + \max_{j\in[M]} \big|\phi(s,a)^\top \zeta_h^{k,j}\big|\\ &\leq \Big(\sqrt{\nu_k(\delta)} + \sqrt{\gamma_k(\delta)}\Big) \big\|\phi(s,a)\big\|_{(\Lambda_h^k)^{-1}},\end{aligned} \tag{B.24}$$

Set $\theta_h^k : \phi(\cdot,\cdot)^\top \theta_h^k = r_h^k(\cdot,\cdot) + \mathbb{P}_h V_{h+1}^k(\cdot,\cdot)$ to be the true parameter and $\Delta\theta_h^k = \theta_h^k - \widehat{\theta}_h^k$ to be the regression error. By the concentration part, conditioning on good events, we have $\|\Delta\theta_h^k\|_{\Lambda_h^k} \leq \sqrt{\nu_k(\delta)}$ and $\|\xi_h^{k,j}\|_{\Lambda_h^k} \leq \sqrt{\gamma_k(\delta)}$ for all $j \in [M]$.

For all $(h,k) \in [H] \times [K]$ and any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\begin{aligned}l_h^k(s,a) &= r_h^k(s,a) + \mathbb{P}_h V_{h+1}^k(s,a) - Q_h^k(s,a)\\ &= r_h^k(s,a) + \mathbb{P}_h V_{h+1}^k(s,a) - \min\big\{H, \max_{j\in[M]} \phi(s,a)^\top(\widehat{\theta}_h^k + \xi_h^{k,j})\big\}^+\\ &\leq \max\{\phi(s,a)^\top \Delta\theta_h^k - \max_{j\in[M]} \phi(s,a)^\top \xi_h^{k,j}, 0\}\end{aligned}$$

Now we prove that with high probability, $\max_{j\in[M]} \phi(s,a)^\top \xi_h^{k,j} - \phi(s,a)^\top \Delta\theta_h^k \geq 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Note that the inequality still holds if we scale $\phi(s,a)$.

Now we assume all $\phi(s,a)$ satisfy $\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} = 1$. Define $\mathcal{C}(\epsilon)$ to be a ϵ-cover of the ellipsoid $\{\phi | \|\phi\|_{(\Lambda_h^k)^{-1}} = 1\}$ with respect to norm $\|\cdot\|_{(\Lambda_h^k)^{-1}}$ and $\log|\mathcal{C}(\epsilon)| = \widetilde{O}(d\log(\frac{1}{\epsilon}))$. For all $j \in [M]$, we have,

$$\{\xi_h^{k,j}\} \sim \mathcal{N}\big(0, 4\nu_k(\delta)(\Lambda_h^k)^{-1}\big).$$

Thus, for all $j \in [M]$ and for all $\phi \in \mathcal{C}(\epsilon)$, we have

$$\big\{\phi^\top \xi_h^{k,j}\big\} \sim \mathcal{N}\Big(0, 4\nu_k(\delta)\|\phi\|^2_{(\Lambda_h^k)^{-1}}\Big).$$

Now, for all $j \in [M]$ and for all $\phi \in \mathcal{C}(\epsilon)$, we have

$$\mathbb{P}\Big(\phi^\top \xi_h^{k,j} - 2\sqrt{\nu_k(\delta)}\|\phi\|_{(\Lambda_h^k)^{-1}} \geq 0\Big) = \Phi(-1).$$

Now

$$\begin{aligned}\mathbb{P}\Big(\max_{j\in[M]} \phi^\top \xi_h^{k,j} - 2\sqrt{\nu_k(\delta)}\|\phi\|_{(\Lambda_h^k)^{-1}} \geq 0\Big) &\geq 1 - (1-\Phi(-1))^M \\ &= 1 - \Phi(1)^M \\ &= 1 - c_0^M, \end{aligned} \tag{B.25}$$

By union bound, with probability $1 - |\mathcal{C}(\epsilon)|c_0^M$, the above bound holds for all elements in $\mathcal{C}$ simultaneously.

Now condition on the previous event, for $\phi = \phi(s,a)$, we can find a $\phi' \in \mathcal{C}(\epsilon)$ such that $\|\phi - \phi'\|_{(\Lambda_h^k)^{-1}} \leq \epsilon$. Define $\Delta\phi = \phi - \phi'$.

$$\begin{aligned}&\phi^\top \xi_h^{k,j} - \phi^\top \Delta\theta_h^k = \phi'^\top \xi_h^{k,j} - \phi'^\top \Delta\theta_h^k + \Delta\phi^\top \xi_h^{k,j} + \Delta\phi^\top \Delta\theta_h^k \\ &\geq \phi'^\top \xi_h^{k,j} - 2\sqrt{\nu_k(\delta)}\|\phi'\|_{(\Lambda_h^k)^{-1}} + \sqrt{\nu_k(\delta)}\|\phi'\|_{(\Lambda_h^k)^{-1}} - \epsilon\|\xi_h^{k,j}\|_{\Lambda_h^k} - \epsilon\|\Delta\theta_h^k\|_{\Lambda_h^k} \\ &\geq \phi'^\top \xi_h^{k,j} - 2\sqrt{\nu_k(\delta)}\|\phi'\|_{(\Lambda_h^k)^{-1}} + \sqrt{\nu_k(\delta)}\|\phi'\|_{(\Lambda_h^k)^{-1}} - \epsilon\sqrt{\gamma_k(\delta)} - \epsilon\sqrt{\nu_k(\delta)}\end{aligned}$$

Set $\epsilon = \frac{\sqrt{\nu_k(\delta)}}{\sqrt{\gamma_k(\delta)}+\sqrt{\nu_k(\delta)}} = \widetilde{O}(\frac{1}{\sqrt{d}})$ and we have, with probability $1 - |\mathcal{C}(\epsilon)|c_0^M$,

$$\begin{aligned}\max_{j\in[M]} \phi^\top \xi_h^{k,j} - \phi^\top \Delta\theta_h^k &\geq \max_{j\in[M]} \phi'^\top \xi_h^{k,j} - 2\sqrt{\nu_k(\delta)}\|\phi'\|_{(\Lambda_h^k)^{-1}} \\ &\geq 0.\end{aligned}$$

Finally we have conditioning on good event $\mathcal{G}(K,H,\delta)$, with probability at least $1 - |\mathcal{C}(\epsilon)|c_0^M$, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, $l_h^k(s,a) \leq 0$. As $\log|\mathcal{C}(\epsilon)| = \widetilde{O}(d\log(\frac{1}{\epsilon}))$, we can set $M = \widetilde{O}(\frac{d\log(1/\epsilon\delta)}{\log(1/c_0)}) = \widetilde{O}(d)$ to have probability $1-\delta$.

□

B.3.2 Regret Bound

Definition 15 (Filtrations)**.** *We denote the σ-algbera generated by the set $\mathcal{G}$ using $\sigma(\mathcal{G})$. We define the following filtrations:*

$$\begin{aligned}
\mathcal{F}^k &\overset{def}{=} \sigma\left(\{(s_t^i, a_t^i, r_t^i)\}_{\{i,t\}\in[k-1]\times[H]} \ \bigcup \ \{\xi_t^{i,j}\}_{\{i,t,j\}\in[k-1]\times[H]\times[M]}\right), \\
\mathcal{F}_{h,1}^k &\overset{def}{=} \sigma\left(\mathcal{F}^k \ \bigcup \ \{(s_t^k, a_t^k, r_t^k)\}_{t\in[h]} \ \bigcup \ \{\xi_t^{k,j} : t \le h, \ 1 \le j \le M\}\right), \\
\mathcal{F}_{h,2}^k &\overset{def}{=} \sigma\left(\mathcal{F}_{h,1}^k \ \bigcup \ \{x_{h+1}^k\}\right).
\end{aligned}$$

Lemma 45 (Lemma 4.2 in Cai et al., 2019)**.** *It holds that*

$$\begin{aligned}
\mathrm{Regret}(T) &= \sum_{k=1}^{K}\left(V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)\right) \\
&= \underbrace{\sum_{k=1}^{K}\sum_{t=1}^{H}\mathbb{E}_{\pi^*}\left[\langle Q_h^k(s_h,\cdot), \pi_h^*(\cdot \mid s_h) - \pi_h^k(\cdot \mid s_h)\rangle \,\middle|\, s_1 = s_1^k\right]}_{\textit{(i)}} \\
&\qquad + \underbrace{\sum_{k=1}^{K}\sum_{t=1}^{H}\mathcal{D}_h^k}_{\textit{(ii)}} + \underbrace{\sum_{k=1}^{K}\sum_{t=1}^{H}\mathcal{M}_h^k}_{\textit{(iii)}} \\
&\qquad + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^*}\left[l_h^k(s_h, a_h) \mid s_1 = s_1^k\right] - l_h^k(s_h^k, a_h^k)\right)}_{\textit{(iv)}}, \qquad \text{(B.26)}
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{D}_h^k &:= \langle (Q_h^k - Q_h^{\pi^k})(s_h^k, \cdot), \pi_h^k(\cdot, s_h^k)\rangle - (Q_h^k - Q_h^{\pi^k})(s_h^k, a_h^k), \qquad \text{(B.27)} \\
\mathcal{M}_h^k &:= \mathbb{P}_h((V_{h+1}^k - V_{h+1}^{\pi^k}))(s_h^k, a_h^k) - (V_{h+1}^k - V_{h+1}^{\pi^k})(s_h^k). \qquad \text{(B.28)}
\end{aligned}$$

Lemma 46. *For the policy π_h^k at time-step k of episode h, it holds that*

$$\sum_{k=1}^{K}\sum_{t=1}^{H}\mathbb{E}_{\pi^*}\left[\langle Q_h^k(s_h,\cdot), \pi_h^*(\cdot \mid s_h) - \pi_h^k(\cdot \mid s_h)\rangle \mid s_1 = s_1^k\right] \le 0, \qquad \text{(B.29)}$$

where $T = HK$.

Proof. Obvious from the observation that π_h^k acts greedily with respect to Q_h^k. Note that if $\pi_h^k = \pi_h^*$ then the difference is 0. Else the difference is negative since π_h^k is deterministic with respect to its action-values meaning it takes a value of 1 where π_h^* would take a value of 0 and Q_h^k would have the greatest value at the state-action pair that π_h^k equals one. □

Lemma 47 (Bound on Martingale Difference Sequence). *For any $\delta > 0$, it holds with probability $1 - 2\delta/3$ that*

$$\sum_{k=1}^{K}\sum_{t=1}^{H}\mathcal{D}_h^k + \sum_{k=1}^{K}\sum_{t=1}^{H}\mathcal{M}_h^k \leq 2\sqrt{2H^2T\log(3/\delta)}. \tag{B.30}$$

Proof. Recall that

$$\mathcal{D}_h^k := \langle(Q_h^k - Q_h^{\pi^k})(s_h^k,\cdot), \pi_h^k(\cdot, s_h^k)\rangle - (Q_h^k - Q_h^{\pi^k})(s_h^k, a_h^k),$$
$$\mathcal{M}_h^k := \mathbb{P}_h((V_{h+1}^k - V_{h+1}^{\pi^k}))(s_h^k, a_h^k) - (V_{h+1}^k - V_{h+1}^{\pi^k})(s_h^k).$$

Note that in line 12 of Algorithm 3, we truncate Q_h^k to the range $[0, H - h]$. Thus for any $(k,t) \in [K] \times [H]$, we have, $|\mathcal{D}_h^k| \leq 2H$. Moreover, since $\mathbb{E}[\mathcal{D}_h^k|\mathcal{F}_{h,1}^k] = 0$, $\mathcal{D}_h^k$ is a martingale difference sequence. So, applying Azuma-Hoeffding inequality we have with probability at least $1 - \delta/3$,

$$\sum_{k=1}^{K}\sum_{t=1}^{H}\mathcal{D}_h^k \leq \sqrt{2H^2T\log(3/\delta)}, \tag{B.31}$$

where $T = KH$.

Similarly, $\mathcal{M}_h^k$ is a martingale difference sequence since for any $(k,t) \in [K] \times [H]$, $|\mathcal{M}_h^k| \leq 2H$ and $\mathbb{E}[\mathcal{M}_h^k|\mathcal{F}_{h,1}^k] = 0$. Applying Azuma-Hoeffding inequality we have with probability at least $1 - \delta/3$,

$$\sum_{k=1}^{K}\sum_{t=1}^{H}\mathcal{M}_h^k \leq \sqrt{2H^2T\log(3/\delta)}. \tag{B.32}$$

Applying union bound on (B.31) and (B.32) gives (B.30) and completes the proof.

□

Lemma 48. *Let $\lambda = 1$ in Algorithm 3. For any $\delta > 0$, conditioned on the event $\mathcal{G}(K, H, \delta)$, we have,*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\big(\mathbb{E}_{\pi^*}\left[l_h^k(s_h, a_h)|s_1 = s_1^k\right] - l_h^k(s_h^k, a_h^k)\big) \leq \big(\sqrt{\nu_K(\delta)}+\sqrt{\gamma_K(\delta)}\big)\sqrt{2dHT\log(1+K)}, \tag{B.33}$$

with probability $1-(\delta+c_0^M)$*.*

Proof. By Lemma 44, with probability $1-(\delta+c_0^M)$ it holds that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[l_h^k(s_h,a_h)|s_1=s_1^k\right]\le 0, \tag{B.34}$$

and

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H}-l_h^k(s_h^k,a_h^k) &\le \sum_{k=1}^{K}\sum_{h=1}^{H}\Big(\sqrt{\nu_k(\delta)}+\sqrt{\gamma_k(\delta)}\Big)\big\|\phi(s_h^k,a_h^k)\big\|_{(\Lambda_h^k)^{-1}}\\ &\le \Big(\sqrt{\nu_K(\delta)}+\sqrt{\gamma_K(\delta)}\Big)\sum_{k=1}^{K}\sum_{h=1}^{H}\big\|\phi(s_h^k,a_h^k)\big\|_{(\Lambda_h^k)^{-1}}\\ &\le \Big(\sqrt{\nu_K(\delta)}+\sqrt{\gamma_K(\delta)}\Big)\sum_{h=1}^{H}\sqrt{K}\Big(\sum_{k=1}^{K}\big\|\phi(s_h^k,a_h^k)\big\|^2_{(\Lambda_h^k)^{-1}}\Big)^{1/2}\\ &\le \Big(\sqrt{\nu_K(\delta)}+\sqrt{\gamma_K(\delta)}\Big)H\sqrt{2dK\log(1+K)}\\ &= \Big(\sqrt{\nu_K(\delta)}+\sqrt{\gamma_K(\delta)}\Big)\sqrt{2dHT\log(1+K)}.\end{aligned} \tag{B.35}$$

Here the second inequality follows from the fact that both $\nu_k(\delta)$ and $\gamma_k(\delta)$ are increasinig in k. The third and the fourth inequalities follow from Cauchy-Schwarz inequality and Lemma 54. Combining (B.34) and (B.35) completes the proof. □

Lemma 49 (Good event probability)**.** *For any* $K\in\mathbb{N}$ *and any* $\delta>0$*, we would have the event* $\mathcal{G}(K,H,\delta')$ *with probability at least* $1-\delta$*, where* $\delta'=\delta/MT$*.*

Proof. By Lemma 52, we have, for any fixed t and k, the event $\mathcal{G}_h^k(\xi,\delta')$ occurs with probability at least $1-M\delta'$. Recall from Definition 14 that,

$$\mathcal{G}(K,H,\delta')=\bigcap_{k\le K}\bigcap_{h\le H}\mathcal{G}_h^k(\xi,\delta').$$

Now taking union bound over all $(t,k)\in[H]\times[K]$, we have

$$\mathbb{P}(\bigcap_{k\le K}\bigcap_{h\le H}\mathcal{G}_h^k(\xi,\delta'))\ge 1-MT\delta'=1-\delta,$$

which completes the proof. □

Theorem 50. *Let $\lambda = 1$, $\sigma = \widetilde{O}(H\sqrt{d})$ and $M = d\log(\delta/9)/\log c_0$, where $c_0 = \Phi(1)$ and $\delta \in (0,1]$. Under Definition 6, the regret of Algorithm 3 satisfies*

$$\textit{Regret}(T) \leq \widetilde{\mathcal{O}}(d^{3/2}H^{3/2}\sqrt{T}),$$

with probability at least $1-\delta$.

Proof of Theorem 50. Let $\delta' = \delta/9$. From Lemma 49, the event $\mathcal{G}(K,H,\delta')$ happens with probability $1-\delta'$. Combining Lemma 48 and Lemma 49 we have that the event $\mathcal{G}(K,H,\delta')$ occurs and it holds that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^*}\left[l_h^k(s_h,a_h)|s_1 = s_1^k\right] - l_h^k(s_h^k,a_h^k)\right) \leq \left(\sqrt{\nu_K(\delta')}+\sqrt{\gamma_K(\delta')}\right)\sqrt{2dHT\log(1+K)}, \tag{B.36}$$

with probability at least $(1-\delta')(1-(\delta'+c_0^M))$. Note that $c_0^M = \delta'$ and $(1-\delta')(1-(\delta'+c_0^M)) > 1-3\delta' = 1-\delta/3$. The martingale inequalities from Lemma 47 happens with probability $1-2\delta/3$.

Applying union bound on (B.29), (B.30) and (B.36) gives the final regret bound of $\widetilde{\mathcal{O}}(d^{3/2}H^{3/2}\sqrt{T})$ completes the proof. □

B.4 Auxiliary lemmas

This section presents several auxiliary lemmas and their proofs.

B.4.1 Gaussian Concentration

Lemma 51 (Gaussian Concentration Vershynin, 2018)**.** *Consider a d-dimensional multivariate normal distribution $\eta \sim \mathcal{N}(0, A\Lambda^{-1})$ where A is a scalar. For any $\delta > 0$, with probability $1-\delta$,*

$$\|\eta\|_\Lambda \leq c\sqrt{dA\log(d/\delta)},$$

where c is some absolute constant. For $d = 1$, we have $c = \sqrt{2}$.

Lemma 52. *Consider a d-dimensional multivariate normal distribution $\mathcal{N}(0, A\Lambda^{-1})$ where A is a scalar. Let $\eta_1, \eta_2, \ldots, \eta_M$ be M independent samples from the distribution. Then for any $\delta > 0$*

$$\mathbb{P}\left(\max_{j\in[M]} \|\eta_j\|_\Lambda \le c\sqrt{dA\log(d/\delta)}\right) \ge 1 - M\delta,$$

where c is some absolute constant.

Proof. From Lemma 51, for a fixed $j \in [M]$, with probability at least $1-\delta$ we would have

$$\|\eta\|_\Lambda \le c\sqrt{dA\log(d/\delta)}.$$

Applying union bound over all M samples completes the proof. □

B.4.2 Inequalities for summations

Lemma 53 (Lemma D.1 in (Jin et al., 2020))**.** *Let $\Lambda_h = \lambda I + \sum_{i=1}^t \phi_i\phi_i^\top$, where $\phi_i \in \mathbb{R}^d$ and $\lambda > 0$. Then it holds that*

$$\sum_{i=1}^t \phi_i^\top (\Lambda_h)^{-1}\phi_i \le d.$$

Lemma 54 (Lemma 11 in (Abbasi-Yadkori et al., 2011))**.** *Using the same notation as defined in this paper*

$$\sum_{k=1}^K \big\|\phi(s_h^k, a_h^k)\big\|^2_{(\Lambda_h^k)^{-1}} \le 2d\log\Big(\frac{\lambda+K}{\lambda}\Big).$$

Lemma 55. *Let $A \in \mathbb{R}^{d\times d}$ be a positive definite matrix where its largest eigenvalue $\lambda_{max}(A) \le \lambda$. Let $x_1, \ldots, x_k$ be k vectors in $\mathbb{R}^d$. Then it holds that*

$$\Big\|A\sum_{i=1}^k x_i\Big\| \le \sqrt{\lambda k}\Big(\sum_{i=1}^k \|x_i\|_A^2\Big)^{1/2}.$$

Proof. For any vector $v \in \mathbb{R}^d$,

$$\begin{aligned}\|Av\| &= \|A^{1/2}A^{1/2}v\|\\ &\le \|A^{1/2}\|\|A^{1/2}v\|\\ &= \|A^{1/2}\|\|v\|_A.\end{aligned}$$

Here the inequality follows from the definition of the operator norm $\|A^{1/2}\|$. Moreover, $\|A^{1/2}\| \leq \sqrt{\lambda}$ since $\lambda_{max}(A) \leq \lambda$. Thus,

$$\left\| A \sum_{i=1}^{k} x_i \right\| \leq \sqrt{\lambda} \left\| \sum_{i=1}^{k} x_i \right\|_A . \tag{B.37}$$

Now by Cauchy-Schwarz inequality,

$$\begin{aligned} \left\| \sum_{i=1}^{k} x_i \right\|_A^2 &= \sum_{i=1}^{k} \sum_{j=1}^{k} x_i^\top A x_j \\ &\leq \sum_{i=1}^{k} \sum_{j=1}^{k} \|x_i\|_A \|x_j\|_A \\ &= \left(\sum_{i=1}^{k} \|x_i\|_A \right)^2 \\ &\leq k \sum_{i=1}^{k} \|x_i\|_A^2. \end{aligned} \tag{B.38}$$

Combining (B.37) and (B.38), proves the lemma. □

B.4.3 Covering numbers and self-normalized processes

Lemma 56 (Lemma D.4 in (Jin et al., 2020)). *Let $\{s_i\}_{i=1}^{\infty}$ be a stochastic process on state space $\mathcal{S}$ with corresponding filtration $\{\mathcal{F}_i\}_{i=1}^{\infty}$. Let $\{\phi_i\}_{i=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process where $\phi_i \in \mathcal{F}_{i-1}$, and $\|\phi_i\| \leq 1$. Let $\Lambda_k = \lambda I + \sum_{i=1}^{k} \phi_i \phi_i^\top$. Then for any $\delta > 0$, with probability at least $1 - \delta$, for all $k \geq 0$, and any $V \in \mathcal{V}$ with $\sup_{s \in \mathcal{S}} |V(s)| \leq H$, we have*

$$\Big\| \sum_{i=1}^{k} \phi_i \big\{ V(s_i) - \mathbb{E}[V(s_i) \,|\, \mathcal{F}_{i-1}] \big\} \Big\|_{\Lambda_k^{-1}}^2 \leq 4H^2 \Big[\frac{d}{2} \log\Big(\frac{k+\lambda}{\lambda} \Big) + \log \frac{\mathcal{N}_\varepsilon}{\delta} \Big] + \frac{8k^2 \epsilon^2}{\lambda},$$

where $\mathcal{N}_\varepsilon$ is the ε-covering number of $\mathcal{V}$ with respect to the distance dist$(V, V') =$ $\sup_{s \in \mathcal{S}} |V(s) - V'(s)|$.

Lemma 57 (Covering number of Euclidean ball, (Vershynin, 2018)). *For any $\varepsilon > 0$, the ε-covering number, $\mathcal{N}_\varepsilon$, of the Euclidean ball of radius $B > 0$ in $\mathbb{R}^d$ satisfies*

$$\mathcal{N}_\varepsilon \leq \Big(1 + \frac{2B}{\varepsilon} \Big)^d \leq \Big(\frac{3B}{\varepsilon} \Big)^d .$$

Lemma 58. *Consider a class of functions $\mathcal{V}: \mathcal{S} \to \mathbb{R}$ which has the following parametric form*

$$V(\cdot) = \Big\langle \min\{\phi(\cdot,\cdot)^\top \theta, H\}^+, \pi(\cdot \mid \cdot) \Big\rangle_{\mathcal{A}},$$

where the parameter θ satisfies $\|\theta\| \le B$ and for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have $\|\phi(s,a)\| \le 1$. If $\mathcal{N}_{\mathcal{V},\varepsilon}$ denotes the ε-covering number of $\mathcal{V}$ with respect to the distance dist$(V, V') = \sup_{s\in\mathcal{S}} |V(s) - V'(s)|$, then

$$\log \mathcal{N}_{\mathcal{V},\varepsilon} \le d \log(1 + 2B/\varepsilon) \le d \log(3B/\varepsilon).$$

Proof. Consider any two functions $V_1, V_2 \in \mathcal{V}$ with parameters θ_1 and θ_2, respectively. Note that $\min\{\cdot, H\}$ is a contraction mapping. Thus we have

$$\begin{aligned}
\operatorname{dist}(V_1, V_2) &\le \sup_s \big| \langle \phi(s,\cdot)^\top \theta_1 - \phi(s,\cdot)^\top \theta_2, \pi(\cdot \mid s) \rangle_{\mathcal{A}} \big| \\
&\le \sup_{\phi:\|\phi\|\le 1} \big| \phi^\top \theta_1 - \phi^\top \theta_2 \big| \\
&= \sup_{\phi:\|\phi\|\le 1} \big| \phi^\top \big(\theta_1 - \theta_2\big) \big| \\
&\le \sup_{\phi:\|\phi\|\le 1} \|\theta_1 - \theta_2\|_2 \|\phi\|_2 \\
&= \|\theta_1 - \theta_2\|, \qquad \text{(B.39)}
\end{aligned}$$

where the second inequality follows from the triangle inequality and the third inequality follows from the Cauchy-Schwarz inequality.

If $\mathcal{N}_{\theta,\varepsilon}$ denotes the ε-covering number of $\{\theta \in \mathbb{R}^d \mid \|\theta\| \le B\}$, Lemma 57 implies

$$\mathcal{N}_{\theta,\varepsilon} \le \Big(1 + \frac{2B}{\varepsilon}\Big)^d \le \Big(\frac{3B}{\varepsilon}\Big)^d.$$

Let $\mathcal{C}_{\theta,\varepsilon}$ be an ε-cover of $\{\theta \in \mathbb{R}^d \mid \|\theta\| \le B\}$ with cardinality $\mathcal{N}_{\theta,\varepsilon}$. Consider any $V_1 \in \mathcal{V}$. By (B.39), there exists $\theta_2 \in \mathcal{C}_{\theta,\varepsilon}$ such that V_2 parameterized by θ_2 satisfies $\operatorname{dist}(V_1, V_2) \le \varepsilon$. Thus we have

$$\log \mathcal{N}_{\mathcal{V},\varepsilon} \le \log \mathcal{N}_{\theta,\varepsilon} \le d \log(1 + 2B/\varepsilon) \le d \log(3B/\varepsilon),$$

which concludes the proof. □