Proceedings of Machine Learning Research vol xxx:1–21, 2020

Model-Based Reinforcement Learning with Value-Targeted Regression

Zeyu Jia JIAZY@PKU.EDU.CN
Peking University

Lin F. Yang LINYANG@EE.UCLA.EDU
University of California, Los Angeles

Csaba Szepesvári SZEPESVA@CS.UALBERTA.CA
DeepMind & University of Alberta

Mengdi Wang MENGDIW@PRINCETON.EDU
DeepMind & Princeton University

Abstract

Reinforcement learning (RL) applies to control problems with large state and action spaces, hence it is natural to consider RL with a parametric model. In this paper we focus on finite-horizon episodic RL where the transition model admits the linear parametrization: $P = \sum_{i=1}^{d} (\theta)_i P_i$. This parametrization provides a universal function approximation and capture several useful models and applications. We propose an upper confidence model-based RL algorithm with value-targeted model parameter estimation. The algorithm updates the estimate of θ by recursively solving a regression problem using the latest value estimate as the target. We demonstrate the efficiency of our algorithm by proving its expected regret bound $\tilde{\mathcal{O}}(d\sqrt{H^3T})$, where H, T, d are the horizon, total number of steps and dimension of θ. This regret bound is independent of the total number of states or actions, and is close to a lower bound $\Omega(\sqrt{HdT})$.

1. Introduction

In this paper, we study episodic reinforcement learning in an environment that can be parameterized by finitely many unknown parameters. In particular, we focus on the case where the unknown probability transition law admits a linear parametrization $P = \sum_i \theta_i P_i$ where $P_1, P_2, \ldots, P_d$ are known basis models and $\theta_* = (\theta_1, \ldots, \theta_d)$ are unknown parameters. This is one of the most basic parametrization for transition systems, finding use as discrete-time approximations in robotics (Kober et al., 2013) and queueing systems (Kovalenko, 1968). It can be viewed as a mixture model that aggregates a finite family of known basic dynamical models (Modi et al., 2019), and it is also a linearized approximation to the more general smoothly parameterized system studied by Abbasi-Yadkori and Szepesvári (2015). It contains as an important special case the linear-factor MDP model where, when good feature representations are available, it is sufficient to embed conditional transition distributions of P into a finite-dimensional matrix (Yang and Wang, 2019a).

The main contribution of this paper is an upper confidence RL algorithm, which estimates the model parameter θ_* adaptively using value-targeted regression and learns to act through episodes. The key to the algorithm is estimation of the model parameter, therefore it is a model-based method. However, there are some important differences to existing model-based methods: Model-based RL methods often explicitly estimate transition probabilities so as to predict future observations (or features, or raw states) (e.g. Jaksch et al., 2010; Yang and Wang, 2019a) or update a Bayesian

posterior over a class of transition models given the observations (e.g., Strens, 2000; Osband and Van Roy, 2014; Abbasi-Yadkori and Szepesvári, 2015; Ouyang et al., 2017; Agrawal and Jia, 2017). In contrast, our algorithm estimates the model parameters by setting up a suitably chosen regression problem where the targets in the regression problem are based on the estimated value function that is used by the algorithm.An optimistic bonus is derived by exploiting the linear mixture form of the transition kernel. Value-targeted regression has several advantages: *(i)* The regression target is a one-dimensional variable, so we avoid tuning and transforming a multivariate regression target which is often needed if we were to predict future features or observations; *(ii)* The values used as regression target are also updated throughout the learning process. Using the current estimated values as target is motivated by the theoretical observation that the regret seems to be controlled by the value prediction error. *(iii)* The θ updates admits simple recursive formula. The full algorithm is computationally simple and sample-efficient with regret guarantee.

Despite of the intuitive advantages, one might worry that regression using only next estimated value as the target might miss to capture the full transition model or lead to low sample efficiency. As a result, the estimated $\hat{\theta}$ may not be used to predict the next state, and failing to learn the full transition model might lead to large or even linear regret. Having this concern in mind, in this paper we study the theoretical question:

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

In the model class we study, as expected, the answer turns out to be yes. Our theory suggests that value-targeted regression is indeed sufficient for model-based RL. We prove that the proposed algorithm achieves an expected cumulative regret at most $O(d\sqrt{H^3T})$ after K episodes, where H is the horizon, d is the number of model parameters and $T = HK$ is the total number of steps in K horizons. It is worth noting this regret does not depend on either the size of the state or that of the action space. We also provide a regret lower bound $\Omega(\sqrt{HdT})$ by adapting a known lower bound for tabular RL. Our approach provides a fresh perspective on the use of supervised learning in RL. They hint that one does not need to precisely estimate the state-to-state transition function but instead to fit the state-to-value (value of next state) relation – a much simpler supervised learning task. The estimated value of next state can provide enough information to estimate the model and perform action-value updates, which leads to provable sublinear regret in the worst case.

2. Problem Formulation and Assumption

We study episodic Markov decision processes (MDPs, for short), described by a tuple $(\mathcal{S}, \mathcal{A}, P, r, H)$. Each episode begins at a fixed initial state 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 $a \in \mathcal{A}$ to incur the immediate reward $r(s, a)$, which is also observed. Then the process transitions to the a random next state s' according to the transition law $P(\cdot|s, a)$, which we also denote as a row vector in $\mathbb{R}^{\mathcal{S}}$. A deterministic policy π is a mapping from $\mathcal{S} \times [H]$ into $\mathcal{A}$, where $\pi_h(s)$ denotes the choice of action when encountering state $s \in \mathcal{S}$ at the stage $h \in [H]$. The value function of a policy π is defined via

$$V_h^\pi(s) = \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r(s_{h'}, \pi(s_{h'}))|s_h = s \right], \qquad \forall s\,,$$

where the subscript π (which we will often suppress) signifies that the probabilities underlying the expectation are governed by π. An optimal policy π^* and the optimal value function V^* are defined

to be a policy and the value function such that $V_h^\pi(s)$ achieves the maximum among all possible policies for any $s \in \mathcal{S}$ and $h \in [H]$. In online RL, the agent does not know P and has to learn to make the best decisions based on its past observations. Therefore, an RL agent can be identified with a history-dependent, nonstationary policy, which, colloquially, we also call the algorithm followed by the agent.The pathwise, cumulated (random) regret incurred by algorithm $\mathcal{A}$ across K episodes is defined as

$$R(T) = \sum_{k=1}^{K} \left(V_1^*(s_1^k) - \sum_{h=1}^{H} r(s_h^k, a_h^k) \right),$$

where $T = KH$ is the total number of time steps, s_1^k is the initial state (history independent) at the k-th episode, and $\{(s_1^1, a_1^1, \ldots, s_H^k, a_H^k), \ldots, (s_1^K, a_1^K, \ldots, s_H^K, a_H^K)\}$ denotes the history of K state-action paths generated by $\mathcal{A}$.

In this paper, we aim to design a learning algorithm with provably low expected regret when the transition model belongs to a parametric family described below.

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

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

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.

Models of the form (1) are common in practical complex systems. It can be viewed as a mixture predicted model which is an aggregation of a number of known basis models. We can view each $P_j(\cdot|\cdot)$ is a basis latent "mode" and the actual transition is a probabilistic mixture of these latent modes. For one example, robotic systems are often smoothly parameterized by unknown mechanical parameters such as torque and friction. Our model (1) provides a linearized parameterized model, which can be used to approximate more general smooth parametric robotic systems Abbasi-Yadkori and Szepesvári (2015). For another example, 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 of model (1) is the linear-factored MDP model where P can be embedded in a finite matrix (Yang and Wang (2019a)):

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

where $\phi(s,a) \in \mathbb{R}^{d_1}, \psi(s') \in \mathbb{R}^{d_2}$ are given for every $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$. The matrix $M \in \mathbb{R}^{d_1 \times d_2}$ is an unknown matrix and to be learned. Then it is easy to see that the factored MDP model is a

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

special case of model (1) with each $\psi_j(s')\phi_i(s,a)$ being a basis model. In this case, the number of unknown parameters in the transition model is $d = d_1 \times d_2$. In this setting, without any additional assumption, our regret bound matches the result of Yang and Wang (2019a).

3. Upper Confidence RL with Value-Targeted Model Parameter Regression

We develop a regret minimization algorithm that takes advantage of the linearly parametrized model by following the ideas of linear bandit Dani et al. (2008); Rusmevichientong and Tsitsiklis (2010); Li et al. (2010); Abbasi-Yadkori et al. (2011); Chu et al. (2011). For a more detailed survey on this subject, we refer the readers to Bubeck et al. (2012); Lattimore and Szepesvári (2018) and references therein. To balance the exploration-exploitation tradeoff, the proposed algorithm updates an empirical estimate of θ_* as well as a confidence set. These estimates are used to compute optimistic state-action values and to choose immediate actions greedily. The full form of our algorithm is presented in Algorithm 1.

3.1. Model-Based Upper Confidence RL

Upper confidence methods are prominent in online learning. In our algorithm, we will maintain a confidence ball B_k for estimated parameters θ and construct upper confidence estimates of Q values via optimistic value iteration:

$$\begin{aligned} &Q_{H+1,k}(s,a) = 0, \\ &V_{h,k}(s) = \Pi_{[0,H]}\big[\max_{a\in\mathcal{A}} Q_{h,k}(s,a)\big], \quad \forall 1 \le h \le H+1, \\ &Q_{h,k}(s,a) = r(s,a) + \max_{\theta\in B_k} \sum_{j=1}^{d} (\theta)_j P_j(\cdot|s,a) V_{h+1,k}, \quad \forall 1 \le h \le H. \end{aligned} \tag{2}$$

As long as $\theta_* \in B_k$ with high probability, these value estimates are optimistic estimates of the actual Q values. Next we show how to construct estimates of θ_* and the confidence ball B_k in equation (2) to balance the exploration-exploitation tradeoff.

3.2. Model Parameter Estimation by Value-Targeted Regression

Every time we obtain a sample (s,a,s') from the transition model $P(\cdot|s,a)$, we receive information about the model parameter θ_*. Instead of regression onto fixed target like probabilities or raw states, we will refresh the estimate of θ_* by regression using the estimated value functions as target. At the (h,k)-th time step, suppose $V_{h+1,k}$ is the current estimated value function for the next time step. We let $X_{h,k}^\top \theta$ be the predicted expected value of next state, where $X_{h,k} = \mathbb{E}_\cdot[V_{h+1,k}(s)|s_h^k, a_h^k] \in \mathbb{R}^d$ denotes the vector of predicted value for the basis models, i.e.,

$$(X_{h,k})_j = \mathbb{E}_j[V_{h+1,k}(s)|s_h^k, a_h^k] = \sum_{s\in\mathcal{S}} V_{h+1,k}(s) P_j(s|s_h^k, a_h^k), \qquad j = 1,\dots,d. \tag{3}$$

In situations where the expected value cannot be computed explicitly, one can approximately compute $X_{h,k}$ by using Monte Carlo simulation since P_j's are known. We use the value at the observed next state as the regression target, i.e.,

$$y_{h,k} = V_{h+1,k}(s_{h+1}^k).$$

Then we construct the following empirical loss function to penalize the value prediction error:

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

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

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

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

To solve the above regression problem, we can first calculate $X_{h',k'}$ using (3) and recursively compute estimates of θ_* by letting

$$M_{1,k+1} = M_{1,1} + \sum_{(h',k')\leq(H,k)} X_{h',k'} X_{h',k'}^\top$$
$$w_{1,k+1} = w_{1,1} + \sum_{(h',k')\leq(H,k)} y_{h',k'} \cdot X_{h',k'},$$

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

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

In the above regression update, we see that the regret target keep changing as the algorithm constructs increasingly accurate value estimates. The regression is done adaptively, where the target value functions keep changing as the agent learns. This is in contrast to typical supervised learning for model predictive control, where the regression targets are often fixed objects (such as raw states, features or keypoints; e.g. Doya et al. (2002)) . Our model parameter update 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. (2019) 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.

3.3. Confidence Set and Closed-Form Q-Updates

We construct B_k as follows:

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

where β_k is preselected (see the algorithm). Since the confidence sets are ellipsoids, the Q update given by Eq. (2) have closed-forms solutions:

$$Q_{h,k}(s,a) = r(s,a) + X_{h,k}^\top \theta_k + \sqrt{\beta_k}\sqrt{X_{h,k}^\top M_k^{-1} X_{h,k}}. \tag{4}$$

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.

4. Main Results

In this section we establish the main theorems of the paper. Theorem 1 gives the regret upper bound for Algorithm 1.

Theorem 1 *Let Assumption 1 hold. The T-time-step regret of Algorithm 1 satisfies*

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

where C_θ is a known constant such that $\|\theta_*\| \leq C_\theta$ and $\tilde{\mathcal{O}}$ 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 D.

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 E.

Theorem 2 *For any $H \geq 1$ and $d \geq 8$, there exists an MDP instance $M(\mathcal{S}, \mathcal{A}, P, r, H)$ and d basis models $P_i(\cdot|\cdot), 1 \leq i \leq d$ satisfying Assumption 1 such that any algorithm has regret at least $\Omega(\sqrt{HdT})$ for sufficiently large T.*

The theorems validate that, in the setting we consider, it is sufficient to use the predicted value functions as regression targets. This suggests that it may be unnecessary to apply supervised learning to predict fixed target like raw state in model-based RL. Our regret upper bound is close to the lower bound. Also note Rusmevichientong and Tsitsiklis (2010) gives a regret lower bound $d\sqrt{T}$ for linearly parameterized bandit with actions on the unit sphere. Our regret upper bound matches this bandit lower bound in d, T, although the settings are not exactly the same.

5. Related Work

Reinforcement learning (RL) enables learning to control in complex environments through trial and error. It is a core problem in artificial general intelligence (Goertzel and Pennachin, 2007; Sutton et al., 1998) and recent years have witnessed phenomenal empirical advances such as in games, robotics and science Mnih et al. (2015); Silver et al. (2017); AlQuraishi (2019); Arulkumaran et al. (2019). In online RL, an agent has to learn to act in an unknown environment from the scratch, collect data as she acts and adapt the policy in realtime. An important problem is to design algorithms that provably achieve sublinear regret in a large class of environments. Regret minimization for RL has received considerable attention during recent years (e.g., Jaksch et al. 2010; Osband et al. 2014; Azar et al. 2017; Dann et al. 2017, 2018; Agrawal and Jia 2017; Osband et al. 2017; Jin et al. 2018; Yang and Wang 2019a; Jin et al. 2019). While most of these existing work focus on the tabular or linear-factor MDP, only a handful of prior efforts have studied RL with general model classes, including the seminal paper Strens (2000) and theoretical works (Osband and Van Roy, 2014; Abbasi-Yadkori and Szepesvári, 2015; Theocharous et al., 2017) that adopt a Bayesian, model-based approach. Please see Section A in the appendix for more discussions on related works.

References

Yasin Abbasi-Yadkori and Csaba Szepesvári. Bayesian optimal control of smoothly parameterized systems. In *UAI*, pages 1–11, 2015.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, pages 1184–1194, 2017.

Mohammed AlQuraishi. AlphaFold at CASP13. *Bioinformatics*, 2019.

Kai Arulkumaran, Antoine Cully, and Julian Togelius. Alphastar: An evolutionary computation perspective. *arXiv preprint arXiv:1902.01724*, 2019.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org, 2017.

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pages 30–37. Elsevier, 1995.

Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1-3): 33–57, 1996.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5713–5723, 2017.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. *arXiv preprint arXiv:1811.03056*, 2018.

Kenji Doya, Kazuyuki Samejima, Ken-ichi Katagiri, and Mitsuo Kawato. Multiple model-based reinforcement learning. *Neural computation*, 14(6):1347–1369, 2002.

Simon S Du, Yuping Luo, Ruosong Wang, and Hanrui Zhang. Provably efficient Q-learning with function approximation via distribution shift error checking oracle. *arXiv preprint arXiv:1906.06321*, 2019.

David A Freedman et al. On tail probabilities for martingales. *the Annals of Probability*, 3(1):100–118, 1975.

Ben Goertzel and Cassio Pennachin. *Artificial general intelligence*, volume 2. Springer, 2007.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I. Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388*, 2019.

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

BV Gnedenko Igor Nikolaevich Kovalenko. *Introduction to queueing theory*. Israel Program for Scientific Translation, Jerusalem, 1968.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *preprint*, 2018.

Tor Lattimore and Csaba Szepesvári. Learning with good feature representations in bandits and in RL with a generative model, 2019.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. *arXiv preprint arXiv:1910.10597*, 2019.

Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems*, pages 1466–1474, 2014.

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

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. *arXiv preprint arXiv:1402.0635*, 2014.

Ian Osband, Benjamin Van Roy, Daniel Russo, and Zheng Wen. Deep exploration via randomized value functions. *arXiv preprint arXiv:1703.07608*, 2017.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown Markov decision processes: A Thompson sampling approach. In *Advances in Neural Information Processing Systems*, pages 1333–1342. 2017.

Ronald Parr, Lihong Li, Gavin Taylor, Christopher Painter-Wakefield, and Michael L Littman. An analysis of linear models, linear value-function approximation, and feature selection for reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*, pages 752–759. ACM, 2008.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.

Malcolm J. A. Strens. A Bayesian framework for reinforcement learning. In *ICML*, pages 943–950, 2000.

Richard S Sutton, Andrew G Barto, et al. *Introduction to reinforcement learning*, volume 2. MIT press Cambridge, 1998.

Georgios Theocharous, Zheng Wen, Yasin Abbasi-Yadkori, and Nikos Vlassis. Posterior sampling for large scale reinforcement learning. *arXiv preprint arXiv:1711.07979*, 2017.

John N Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. In *Advances in neural information processing systems*, pages 1075–1081, 1997.

Zheng Wen and Benjamin Van Roy. Efficient exploration and value function generalization in deterministic systems. In *Advances in Neural Information Processing Systems*, pages 3021–3029, 2013.

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

Lin F Yang and Mengdi Wang. Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. *arXiv preprint arXiv:1905.10389*, 2019a.

Lin F Yang and Mengdi Wang. Sample-optimal parametric Q-learning with linear transition models. *International Conference on Machine Learning*, 2019b.

A. Zanette, A. Lazaric, M. J. Kochenderfer, and E. Brunskill. Limiting extrapolation in linear approximate value iteration. In *Advances in Neural Information Processing Systems*, pages 5616–5625. Curran Associates, Inc., 2019.

Appendix A. More 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., Jaksch et al. 2010; Osband et al. 2014; Azar et al. 2017; Dann et al. 2017, 2018; Agrawal and Jia 2017), and on model-free methods (e.g., Osband et al. 2017; Jin et al. 2018). Among these works, the best known regret achieved by a model-based algorithm is $\tilde{\mathcal{O}}(\sqrt{H^2SAT})$ and the best regret achieved by a model-free algorithms is asymptotic $\tilde{\mathcal{O}}(\sqrt{HSAT})$, where T denotes the number of time steps and $\tilde{O}(\cdot)$ hides log factors. Jaksch et al. (2010) established a worst-case regret lower bound of $\Omega(\sqrt{HSAT})$.

Moving beyond tabular MDP, there have been significant theoretical and empirical advances on RL with function approximation, including but not limited to Baird (1995); Tsitsiklis and Van Roy (1997); Parr et al. (2008); Mnih et al. (2013, 2015); Silver et al. (2017); Yang and Wang (2019b); Bradtke and Barto (1996). Among these works, many papers aim to uncover algorithms that are provably efficient. 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. Yang and Wang (2019b) 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, Yang and Wang (2019a) 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 $\tilde{\mathcal{O}}(H^2d\sqrt{T})$ with stronger assumptions and $\tilde{\mathcal{O}}(H^2d^2\sqrt{T})$ in general, where d is the dimension of state representation $\phi(s,a)$. Jin et al. (2019) 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 $\tilde{\mathcal{O}}(\sqrt{H^3d^3T})$. (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 $\frac{d^3K^2H^2}{\epsilon^2}$ 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, the seminal paper 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 (Assumption 1 of this paper), this Bayesian regret becomes $O(d\sqrt{T})$, and our worst-case regret result matches with the Bayesian one. The works Abbasi-Yadkori and Szepesvári (2015); Theocharous et al. (2017) also considered the Bayesian regret and while Abbasi-Yadkori and Szepesvári (2015) considered a smooth parameterization with different notions of smoothness. To the authors' best knowledge, there are no prior works addressing the problem of designing low-regret algorithms for MDPs with linearly parameterized transition models.

Appendix B. Conclusion of the Paper

The paper proposes an episodic upper confidence reinforcement learning method that applies to linearly parameterized transition systems. The proposed method updates the model parameter re-

cursively by regression using the estimated next-state value as the regression target. Then the estimated model is used to update the optimistic state-action values and prescribe actions in the upcoming episode. We show that this simple algorithm achieves a worst-case regret up to $\tilde{O}(H^{3/2}d\sqrt{T})$ where d is the number of model parameters. This result demonstrates the efficacy of value-targeted regression for efficient model-based reinforcement learning.

Appendix C. Main Algorithm

In this section, we provide the model-based reinforcement learning algorithm with value-target regression mentioned in Section 3.

Algorithm 1 UCRL with Value-Targeted Model Estimation

Input: MDP, $d, H, T = KH$;
Initialize: $M_{1,1} \leftarrow H^2 dI, \quad w_{1,1} \leftarrow 0 \in \mathbb{R}^{d\times 1}, \quad \theta_1 \leftarrow M_{1,1}^{-1} w_{1,1} \qquad$ for $1 \le h \le H$;
Initialize: $\delta \leftarrow 1/K, \quad \beta_k \leftarrow 16C_\theta^2 H^2 d \log(1+Hk)\log^2((k+1)^2 H/\delta) \quad$ for $1 \le k \le K$;
Compute Q-function $Q_{h,1}$ using $\theta_{1,1}$ according to (2);
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
    $X_{h,k} \leftarrow \mathbb{E}_\cdot[V_{h+1,k}(s)|s_h^k, a_h^k] = \sum_{s\in\mathcal{S}} V_{h+1,k}(s) \cdot P_\cdot(s|s_h^k, a_h^k)$.
    $y_{h,k} \leftarrow V_{h+1,k}(s_{h+1}^k)$ ▷ 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

$$M_{1,k+1} \leftarrow M_{H+1,k} \quad w_{1,k+1} \leftarrow w_{H+1,k} \quad \text{and} \quad \theta_{k+1} \leftarrow M_{1,k+1}^{-1} w_{1,k+1};$$

  Compute $Q_{h,k+1}, h = H, \ldots, 1$, using θ_{k+1} according to (4) ▷ Computing Q functions
end for

Appendix D. Proof of Theorem 1

Here we will provide the formal proof of Theorem 1. The full proof is divided into five parts in the following five subsections respectively. In the first subsection, we decompose the regret into the sum of bonuses assuming the Q-functions indeed are optimistic estimates. In the second subsection, we discover some important properties of our algorithm. We provide an upper bound to the sum of bonuses in the third subsection. In the fourth subsection, we will prove that the optimism holds with high probability by constructing a particular martingale and showing that it concentrates, and 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 $\mathcal{F}_{h,k}$ be the filtration generated by the random sample path $\{(s_{h'}^{k'}, a_{h'}^{k'}, r_{h'}^{k'})\}_{(h',k') \leq (h,k)}$.

D.1. Regret Analysis

Throughout D.1 to D.3, we assume that $\theta_* \in B_k$ for all $1 \leq k \leq K$. And in subsection D.4 we will prove that this event holds with high probability.

D.1.1. OPTIMISM

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

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

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.

D.1.2. REGRET DECOMPOSITION

Let us denote π_k to be the stationary policy used in the k episode, and let

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

Using the fact that $\pi_k(s_h^k) = a_h^k$ and $\theta_* \in B_k$ and letting 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

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

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.

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

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

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

$$\begin{aligned}
\hat{R}(K) &= \sum_{k=1}^K \left[V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k)\right] \\
&\le \sum_{k=1}^K \sum_{h=1}^H \xi_{h+1}^k + 2 \sum_{k=1}^K \sum_{h=1}^H \sqrt{\beta_k} \cdot \sqrt{\left[P_\cdot(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]^\top M_k^{-1} \left[P_\cdot(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]}.
\end{aligned} \tag{5}$$

Moreover, it is easy to observe that

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

therefore ξ_{h+1}^k is a martingale difference sequence w.r.t. $\mathcal{F}_{h,k}$. Since

$$0 \le V_h^*(s_h^k), V_{h,k}(s_h^k) \le H \quad \text{and} \quad P(\cdot|s_h^k, a_h^k) \text{ is a probability distribution over the state space,}$$

we have $|\xi_h^k| \le H$ with probability 1. By the Azuma-Hoeffding inequality, with probability at least $1 - \delta$, the following inequality holds

$$\sum_{k=1}^K \sum_{h=1}^H \xi_{h+1}^k \le \sqrt{2H^3 K \log(1/\delta)}. \tag{6}$$

It remains to analyze the second term of (5), ie., the sum of bonus given by

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

D.2. Some Properties of Algorithm 1

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

D.2.1.

Note that

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

Denote

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

Denote by $(h,k)+1$ the double index of the next time step after (h,k), that is $(h+1,k)$ if $h<H$ and $(h,k+1)$ otherwise. We can see $\{M_{h,k}\}$ satisfies $M_{1,k}=M_{H+1,k-1}$ and also a recursive formula

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

It implies that

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

D.2.2.

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

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

we have

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

This 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

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

D.2.3.

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

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

where the last inequality uses Assumption 1 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 $M_{(h,k)+1}$:

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

which indicates that

$$\log\det M_{(H,k)+1} - \log\det M_{1,1} \le \log\left((H^2 d + KH^3)^d\right) - \log\left((H^2 d)^d\right) \le d\log(1+HK). \tag{7}$$

Hence we have

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

D.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

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

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

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

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

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

we have

$$\begin{aligned}
&2\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\beta_k}\cdot\sqrt{\left[P_\cdot(\cdot|s_h^k,a_h^k)V_{h+1,k}\right]^\top M_k^{-1}\left[P_\cdot(\cdot|s_h^k,a_h^k)V_{h+1,k}\right]} \\
&\le 2\sqrt{\beta_K}\cdot\sum_{k=1}^{K}\sum_{h=1}^{H}u_{h,k} \le 2\sqrt{\beta_K}\cdot\sum_{k=1}^{K}\sum_{h=1}^{H}\min\{1,u_{h,k}\} \\
&\le 2\sqrt{HK\beta_K}\cdot\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\min\{1,u_{h,k}^2\}} \le 4\sqrt{HK\beta_K}\cdot\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\log(1+u_{h,k}^2)}
\end{aligned} \tag{8}$$

where the third inequality uses the Cauchy-Schwarz inequality. Next we notice that

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

Hence we have

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

We further notice that every eigenvalue of the matrix

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

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

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

This indicates that

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

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

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

where the last inequality uses (7).

D.4. Martingale Concentration Analysis

In this section, we will prove that $\theta_* \in B_k$ for every k with high probability. We define

$$\begin{aligned}
w_{h,k} &= \sum_{(h',k')<(h,k)} V_{h'+1,k'}(s^{k'}_{h'+1}) \cdot \left[P_\bullet(\cdot|s^{k'}_{h'}, a^{k'}_{h'})V_{h'+1,k'}\right], \\
\theta_{h,k} &= M^{-1}_{h,k} w_{h,k}, \\
B_{h,k} &= \left\{(\theta - \theta_{h,k})^T M_{h,k}(\theta - \theta_{h,k}) \le \beta_k\right\},
\end{aligned}$$

and it is easy to see $\theta_k = \theta_{1,k}$ and $B_k = B_{1,k}$. Next, we will prove that $\theta_* \in B_{h,k}$ for every h, k with high probability. Note that all sample transitions are dependent by the nature of our algorithm. We will construct a particular martingale to prove concentration of the proposed estimates.

We define the random vectors $\{Y_{h,k}\}$ as

$$Y_{h,k} = M_{h,k}(\theta_{h,k} - \theta_*) = w_{h,k} - M_{h,k}\theta_*,$$

and define the probabilistic events $\mathcal{E}_{h,k}$ as

$$\mathcal{E}_{h,k} = \{\theta \in B_{h,k}\} = \left\{Y^\top_{h,k} M^{-1}_{h,k} Y_{h,k} \le \beta_k\right\}.$$

In what follows we will prove that, with probability at least $1-\delta$ events $\mathcal{E}_{h,k}$ holds for all (h,k).

For vector $Y_{h,k}$, we have its initial value

$$Y_{1,1} = -H^2 d \cdot \theta_*,$$

and by letting

$$\eta_{h,k} = V_{h+1,k}(s^k_{h+1}) - \mathbb{E}[V_{h+1,k}(s)|s^k_h, a^k_h] = V_{h+1,k}(s^k_{h+1}) - P(\cdot|s^k_h, a^k_h)V_{h+1,k},$$

we have the following iterative formula:

$$\begin{aligned}
Y_{(h,k)+1} &= w_{(h,k)+1} - M_{(h,k)+1}\theta_* \\
&= w_{h,k} + V_{h+1,k}(s^k_{h+1}) \cdot \left[P_\bullet(\cdot|s^k_h, a^k_h)V_{h+1,k}\right] - M_{h,k}\theta_* \\
&\quad - \left[P_\bullet(\cdot|s^k_h, a^k_h)V_{h+1,k}\right]\left[P_\bullet(\cdot|s^k_h, a^k_h)V_{h+1,k}\right]^\top \theta_* \\
&= Y_{h,k} + \left[V_{h+1,k}(s^k_{h+1}) - P(\cdot|s^k_h, a^k_h)V_{h+1,k}\right] \cdot \left[P_\bullet(\cdot|s^k_h, a^k_h)V_{h+1,k}\right] \\
&= Y_{h,k} + \eta_{h,k}\left[P_\bullet(\cdot|s^k_h, a^k_h)V_{h+1,k}\right].
\end{aligned}$$

Since s^k_{h+1} is sampled from distribution $P(\cdot|s^k_h, a^k_h)$, we have

$$\mathbb{E}\left[\eta_{h,k}\middle|\mathcal{F}_{h,k}\right] = 0 \qquad \text{and} \qquad |\eta_{h,k}| \le H.$$

It follows that

$$\begin{aligned}
&Y_{(h,k)+1}^\top M_{(h,k)+1}^{-1} Y_{(h,k)+1} \\
&= Y_{h,k}^\top M_{(h,k)+1}^{-1} Y_{h,k} + 2\eta_{h,k}\left[P_\cdot(\cdot|s_h^k,a_h^k)V_{h+1,k}\right] M_{(h,k)+1}^{-1} Y_{h,k} \\
&\quad + \eta_{h,k}^2 \left[P_\cdot(\cdot|s_h^k,a_h^k)V_{h+1,k}\right]^\top M_{(h,k)+1}^{-1}\left[P_\cdot(\cdot|s_h^k,a_h^k)V_{h+1,k}\right] \\
&\le Y_{h,k}^\top M_{h,k}^{-1} Y_{h,k} + 2\eta_{h,k}\left[P_\cdot(\cdot|s_h^k,a_h^k)V_{h+1,k}\right]^\top M_{(h,k)+1}^{-1} Y_{h,k} \\
&\quad + \eta_{h,k}^2 \left[P_\cdot(\cdot|s_h^k,a_h^k)V_{h+1,k}\right]^\top M_{(h,k)+1}^{-1}\left[P_\cdot(\cdot|s_h^k,a_h^k)V_{h+1,k}\right],
\end{aligned}$$

where the inequality uses the fact $M_{(h,k)+1} \succeq M_{h,k}$. It follows by induction that

$$\begin{aligned}
&Y_{(h,k)+1}^\top M_{(h,k)+1}^{-1} Y_{(h,k)+1} \\
&\le Y_{1,1}^\top M_{1,1}^{-1} Y_{1,1} + 2\sum_{(h',k')\le(h,k)} \eta_{h',k'}\left[P_\cdot(\cdot|s_{h'}^{k'},a_{h'}^{k'})V_{h'+1,k'}\right]^T M_{(h',k')+1}^{-1} Y_{h',k'} \\
&\quad + \sum_{(h',k')\le(h,k)} \eta_{h',k'}^2 \left[P_\cdot(\cdot|s_{h'}^{k'},a_{h'}^{k'})V_{h'+1,k'}\right]^\top M_{(h',k')+1}^{-1}\left[P_\cdot(\cdot|s_{h'}^{k'},a_{h'}^{k'})V_{h'+1,k'}\right].
\end{aligned} \tag{9}$$

For the first term on RHS of (9), according to $M_1 = H^2 dI$ and Assumption 1, we have

$$Y_{1,1}^\top M_{1,1}^{-1} Y_{1,1} = \theta_*^\top M_1 \theta_* \le H^2 d\|\theta_*\|_2^2 = C_\theta^2 \cdot H^2 d.$$

For the third term on RHS of (9), since

$$\left[P_\cdot(\cdot|s_{h'}^{k'},a_{h'}^{k'})V_{h'+1,k'}\right]^\top M_{(h',k')+1}^{-1}\left[P_\cdot(\cdot|s_{h'}^{k'},a_{h'}^{k'})V_{h'+1,k'}\right] = \frac{l_{h',k'}^2}{1+l_{h',k'}^2} \quad \text{and} \quad |\eta_{h',k'}| \le H,$$

we have

$$\begin{aligned}
&\sum_{(h',k')\le(h,k)} \eta_{h',k'}^2 \left[P_\cdot(\cdot|s_{h'}^{k'},a_{h'}^{k'})V_{h'+1,k'}\right]^\top M_{(h',k')+1}^{-1}\left[P_\cdot(\cdot|s_{h'}^{k'},a_{h'}^{k'})V_{h'+1,k'}\right] \\
&\le H^2 \cdot \sum_{(h',k')\le(h,k)} \frac{l_{h',k'}^2}{1+l_{h',k'}^2} \le 2H^2 d\log(1+HK).
\end{aligned}$$

For the second term in (9), we will conduct a martingale concentration analysis. We let $E_{h',k'}$ be the indicator function given by

$$E_{h',k'} = \mathbb{I}\left[\mathcal{E}_{h'',k''} \text{ for all } (h'',k'') \le (h',k')\right]$$

and let

$$G_{h',k'} = E_{h',k'} \cdot \eta_{h',k'}\left[P_\cdot(\cdot|s_{h'}^{k'},a_{h'}^{k'})V_{h'+1,k'}\right] M_{(h',k')+1}^{-1} Y_{h',k'}.$$

We notice that calculating $V_{h'+1,k'}$ only requires samples $(s_{h''}^{k''}, a_{h''}^{k''})$ for $k'' \le k'-1$ and calculating $M_{(h',k')+1}, Y_{h',k'}$ only requires samples $(s_{h''}^{k''}, a_{h''}^{k''})$ for $(h'',k'') \le (h',k')$. We also notice that $\mathcal{E}_{h'',k''} \in \mathcal{F}_{h',k'}$ for every $(h'',k'') \le (h',k')$. These facts indicate

$$V_{h'+1,k'},\ M_{(h',k')+1},\ Y_{h',k'},\ P_\cdot(\cdot|s_{h'}^{k'},a_{h'}^{k'}),\ E_{h',k'} \text{ are all } \mathcal{F}_{h',k'}\text{-measurable.}$$

Therefore, since $\mathbb{E}\left[\eta_{h',k'}\middle|\mathcal{F}_{h',k'}\right]=0$, we have

$$\mathbb{E}\left[G_{h',k'}\middle|\mathcal{F}_{h',k'}\right]=0,$$

therefore $\left\{G_{h',k'}\right\}_{(h',k')\le(h,k)}$ is a *martingale difference sequence* adapted to the filtration $\{\mathcal{F}_{k',h'}\}_{(h',k')\le(h,k)}$. Moreover, when $E_{h',k'}=1$, we have events $\mathcal{E}_{h'',k''}$ hold for $(h'',k'')\le(h',k')$, which implies that

$$E_{h',k'}\cdot Y_{h',k'}M_{h',k'}^{-1}Y_{h',k'}\le\beta_{k'}$$

and when $E_{h',k'}=0$ the left hand side of the above inequality is 0, and this inequality also holds. Therefore, we have the following upper bound that holds for the absolute value of $G_{h',k'}$ with probability 1:

$$\begin{aligned}
&\left|G_{h',k'}\right|\le HE_{h',k'}\cdot\left|\left[P_{\cdot}(\cdot|s_{h'}^{k'},a_{h'}^{k'})V_{h'+1,k'}\right]M_{(h',k')+1}^{-1}Y_{h',k'}\right|\\
&\le H\cdot\sqrt{\left[P_{\cdot}(\cdot|s_{h'}^{k'},a_{h'}^{k'})V_{h'+1,k'}\right]M_{(h',k')+1}^{-1}\left[P_{\cdot}(\cdot|s_{h'}^{k'},a_{h'}^{k'})V_{h'+1,k'}\right]}\cdot\sqrt{E_{h',k'}\cdot Y_{h',k'}M_{(h',k')+1}^{-1}Y_{h',k'}}\\
&=H\cdot\sqrt{\frac{l_{h',k'}^2}{1+l_{h',k'}^2}}\cdot\sqrt{E_{h',k'}\cdot Y_{h',k'}M_{(h',k')+1}^{-1}Y_{h',k'}}\le H\cdot\sqrt{E_{h',k'}\cdot Y_{h',k'}M_{h',k'}^{-1}Y_{h',k'}}\\
&\le H\sqrt{\beta_{k'}}\le H\sqrt{\beta_k},
\end{aligned}$$

It also follows that the sum of conditional variances satisfies with probability 1:

$$\begin{aligned}
&\sum_{(h',k')\le(h,k)}\mathrm{Var}\left(G_{h',k'}\middle|\mathcal{F}_{k',h'}\right)\\
&\le\sum_{(h',k')\le(h,k)}E_{h',k'}\left|\eta_{h',k'}\left[P_{\cdot}(\cdot|s_{h'}^{k'},a_{h'}^{k'})V_{h'+1,k'}\right]M_{(h',k')+1}^{-1}Y_{h',k'}\right|^2\\
&\le\sum_{(h',k')\le(h,k)}H^2\cdot\frac{l_{h',k'}^2}{1+l_{h',k'}^2}\cdot\left(E_{h',k'}\cdot Y_{h',k'}M_{h',k'}^{-1}Y_{h',k'}\right)\le H^2\beta_k\sum_{(h',k')\le(h,k)}\frac{l_{h',k'}^2}{1+l_{h',k'}^2}\\
&\le 4H^2d\log(1+HK)\cdot\beta_k.
\end{aligned}$$

We will use the Freedman concentration inequality as follows.

Theorem 3 (Freedman et al. (1975)) *Let $X_1,\cdots,X_T$ be a martingale difference sequence adapted to filtration $\mathcal{F}_1,\cdots,\mathcal{F}_T$ ($X_1,\cdots,X_{t-1}$ are $\mathcal{F}_t$-measurable and $\mathbb{E}[X_t|\mathcal{F}_t]=0$). Let b be an upper bound on X_i, and let V be the sum of conditioned variances:*

$$V=\sum_{t\le T}Var\left(X_t|\mathcal{F}_t\right).$$

Then for every $a,v>0$, we have

$$\mathbb{P}\left[\sum_{t=1}^{T}X_t\ge a,\text{and }V\le v\right]\le\exp\left(\frac{-a^2}{2v+2ab/3}\right).$$

We apply the above Freedman inequality to our constructed martingale difference sequence and the filtration $\mathcal{F}_{h',k'}$ up to (h,k), by letting

$$\begin{aligned} b &= H\sqrt{\beta_k}, \\ a &= \beta_{k+1}/4 = 4H^2 d\log(1+Hk)\log^2((k+1)^2H/\delta), \\ v &= 4H^2 d\log(1+k)\cdot\beta_k, \end{aligned}$$

then we obtain for every (h,k) that

$$\mathbb{P}\left(\sum_{(h',k')\le(h,k)} G_{h',k'} \le \beta_{k+1}/4\right) \ge 1 - \frac{\delta}{2(k+1)^2 H}.$$

Further by applying an union bound to the above inequality for all (h,k), we obtain

$$\mathbb{P}\left(\sum_{(h',k')\le(h,k)} G_{h',k'} \le \beta_{k+1}/4 \text{ for all } (h,k)\right) \ge 1 - \sum_{k=1}^{\infty} H\cdot\frac{\delta}{2(k+1)^2 H} \ge 1-\delta. \tag{10}$$

In the following we will prove by induction on (h,k) that if inequality $\sum_{(h',k')\le(h,k)} G_{h',k'} \le \beta_{k+1}/4$ holds for every (h,k), then events $\mathcal{E}_{h,k}$ hold for all (h,k). When $(h,k)=(1,1)$, it is easy to verify that event $\mathcal{E}_{h,k}$ holds. Next we assume that events $\mathcal{E}_{h',k'}$ hold for all $(h',k') \le (h,k)$ and consider the event $\mathcal{E}_{(h,k)+1}$.

Since $\mathcal{E}_{h',k'}$ holds for all $(h',k') \le (h,k)$, we have $E_{h',k'} = 1$ for every $(h',k') \le (h,k)$. This indicates that

$$\sum_{(h',k')\le(h,k)} \eta_{h',k'}\left[P_{\cdot}(\cdot|s^{k'}_{h'}, a^{k'}_{h'})V_{h'+1,k'}\right]^T M^{-1}_{(h',k')+1} Y_{h',k'} = \sum_{(h',k')\le(h,k)} G_{h',k'} \le \frac{\beta_{k+1}}{4}$$

Hence by using (9) we obtain the following inequality:

$$Y^\top_{(h,k)+1} M^{-1}_{(h,k)+1} Y_{(h,k)+1} \le C_\theta^2 \cdot H^2 d + \beta_{k+1}/2 + 4H^2 d\log(1+HK) \le \beta_{k+1},$$

in other words $\mathcal{E}_{(h,k)+1}$ holds. This completes the induction of at $(h,k)+1$.

Finally we notice from (10) that with probability at least $1-\delta$, inequality $\sum_{(h',k')\le(h,k)} G_{h',k'} \le \beta_{k+1}/4$ holds for all (h,k). Therefore, with probability at least $1-\delta$, events $\mathcal{E}_{h,k}$ holds for all (h,k), therefore $\theta_* \in B_k$ holds for all $1 \le k \le K$ with probability at least $1-\delta$.

D.5. Expected Regret Analysis

According to Section D.4, we have with probability at least $1-\delta$ that $\theta_* \in B_k$ for all $1 \le k \le K$. When this event happens, we enable the analysis of Sections D.1-D.3. We combine the error bounds (6) and (8) and apply them into the regret bound (5). It follows that, if $T = KN$,

$$\begin{aligned} R(T) &\le 2\sqrt{H^3 K}\log(1/\delta) + 4\sqrt{H^2 dK\beta_K \log(1+HK)} \\ &\le 18 C_\theta H^2 d\sqrt{K}\cdot\log(1+HK)\log((K+1)^2H/\delta) \end{aligned}$$

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

$$\begin{aligned}\mathbb{E}\left[R(T)\right] &\le (1-2\delta)\cdot 18C_\theta H^2 d\sqrt{K}\cdot\log(1+HK)\log((K+1)^2H/\delta) + 2\delta\cdot HK \\ &\le 20C_\theta H^2 d\sqrt{K}\cdot\log(1+HK)\log(K(K+1)^2H) = \tilde{\mathcal{O}}(C_\theta\cdot H^2 d\sqrt{K}) \\ &= \tilde{\mathcal{O}}(C_\theta\cdot d\sqrt{H^3T}).\end{aligned}$$

Thus we have completed the proof of Theorem 1.

Appendix E. Proof of Theorem 2

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 $d \ge 8$. We set $S = 2$ and $A = d/4 \ge 2$. 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 (1). 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. ■