Nonstationary Reinforcement Learning with Linear Function Apporximation

Huozhi Zhou [1] **Jinglin Chen** [2] **Lav R. Varshney** [1] **Ashish Jagmohan** [3]

Abstract

We consider reinforcement learning (RL) in episodic Markov decision processes (MDPs) with linear function approximation under drifting environment. Specifically, both the reward and state transition functions can evolve over time, constrained so their total variations do not exceed a given *variation budget*. We first develop `LSVI-UCB-Restart` algorithm, an optimistic modification of least-squares value iteration combined with periodic restart, and bound its dynamic regret when variation budgets are known. We then propose a parameter-free algorithm that works without knowing the variation budgets, `Ada-LSVI-UCB-Restart`, but with a slightly worse dynamic regret bound. We also derive the first minimax dynamic regret lower bound for nonstationary linear MDPs to show our proposed algorithms are near-optimal. As a byproduct, we establish a minimax regret lower bound for linear MDPs, which was unsolved by Jin et al. (2020).

1. Introduction

Reinforcement learning (RL) is a core control problem in which an agent sequentially interacts with an unknown environment to maximize its cumulative reward (Sutton & Barto, 2018). RL finds enormous applications in real-time bidding in advertisement auctions (Cai et al., 2017), autonomous driving (Shalev-Shwartz et al., 2016), gaming-AI (Silver et al., 2018), and inventory control (Agrawal & Jia, 2019), among others. Recent advances in RL rely on function approximators such as deep neural nets to overcome the curse of dimensionality for large-scale decision making problems, i.e., the value function is approximated by a function which is able to predict the value function for unseen state-action pairs given a few training samples (Mnih et al., 2015; Silver et al., 2017; Akkaya et al., 2019). Motivated by the empirical success of RL algorithms with function approximation, there is growing interest in developing RL algorithms with function approximation that are statistically efficient in the minimax sense (Yang & Wang, 2019; Cai et al., 2020; Jin et al., 2020; Modi et al., 2020; Wang et al., 2020; Wei et al., 2020; Neu & Olkhovskaya, 2020; Zanette et al., 2020). One recent work also studies the instance-dependent sample complexity bound for RL with function approximation, which adapts to the complexity of the specific MDP instance (Foster et al., 2020). The focus of this line of work is to develop statistically efficient algorithms with function approximation for RL. Such efficiency is especially crucial in data-sparse applications such as medical trials (Zhao et al., 2009).

However, all of the aforementioned empirical and theoretical works on RL with function approximation assume the environment is stationary, which is insufficient to model problems with time-varying dynamics. In general nonstationary random processes naturally occur in many settings and are able to characterize larger classes of problems of interest (Cover & Pombra, 1989). In this work, we consider the setting of episodic RL with nonstationary reward and transition functions. To measure the performance of an algorithm, we use the notion of *dynamic regret*, the performance difference between an algorithm and the set of policies optimal for individual episodes in hindsight. For nonstationary RL, dynamic regret is a stronger and more appropriate notion of performance measure than static regret, but is also more challenging for algorithm design and analysis. To incorporate function approximation, we focus on a subclass of MDPs in which the reward and transition dynamics are linear in a known feature map (Melo & Ribeiro, 2007), termed *linear MDP*. For nonstationary linear MDPs, we show that one can design a near-optimal statistically-efficient algorithm to achieve sublinear dynamic regret as long as the total variation of reward and transition dynamics is sublinear.

Our contributions are two-fold. First, we prove the minimax regret lower bound for non-stationary linear MDPs.

[1]Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA [2]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA [3]IBM Thomas J. Watson Research Center, Yorktown Heights, New York, USA. Correspondence to: Huozhi Zhou <hzhou35@illinois.edu>.

Workshop on Reinforcement Learning Theory in the 38^{th} International Conference on Machine Learning, 2021.

As a byproduct, we also derive the minimax regret lower bound for stationary linear MDP, which is unsolved in (Jin et al., 2020). Second, we propose two algorithms to learn the optimal policy for nonstationary linear MDPs that achieve near-optimal regret. The first algorithm is based on combining the periodic restart and optimistic in the face of uncertainty principle, but requires the knowledge of the environmental drift. To overcome the limitation of the first algorithm, the second algorithm leverages the bandit-over-bandit (Cheung et al., 2019) mechanism to become parameter-free, only with a slight performance degradation.

Notation We use $\langle \cdot, \cdot \rangle$ to denote inner products in Euclidean space, $\|\boldsymbol{v}\|_2$ to denote the L_2 norm of vector $\boldsymbol{v}$, and $\|\boldsymbol{v}\|_\Lambda$ to denote the norm induced by a positive definite matrix A for vector $\boldsymbol{v}$, i.e., $\|\boldsymbol{v}\|_\Lambda = \sqrt{\boldsymbol{v}^\top \Lambda \boldsymbol{v}}$. For an integer N, we denote the set of positive integers $\{1, 2, \ldots, N\}$ as $[N]$.

2. Preliminaries

We consider the setting of a nonstationary episodic Markov decision process (MDP), specified by a tuple $(\mathcal{S}, \mathcal{A}, H, K, \mathbb{P} = \{\mathbb{P}_h^k\}_{h\in[H],k\in[K]}, r = \{r_h^k\}_{h\in[H],k\in[K]})$, where the set $\mathcal{S}$ is the collection of states, $\mathcal{A}$ is the collection of actions, H is the length of each episode, K is the total number of episodes, and $\mathbb{P}$ and r are the transition kernel and deterministic reward functions respectively. Moreover, $\mathbb{P}_h^k(\cdot|s,a)$ denotes the transition kernel over the next states if the action a is taken for state s at step h in the k-th episode, and $r_h^k : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the deterministic reward function at step h in the k-th episode. Note that we are considering a nonstationary setting, thus we assume the transition kernel $\mathbb{P}$ and reward function r may change in different episodes. We will explicitly quantify the nonstationarity later.

Given a policy π, a level $h \in [H]$ and a state $s \in \mathcal{S}$, the value function at k-th episode is defined as

$$V_{h,k}^\pi(s) = \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r_{h'}^k(s_{h'}, a_{h'}) | s_h = s \right].$$

Similarly, for a given state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, the Q-funtion for policy π at step h in the k-th episode is defined as

$$Q_{h,k}^\pi(s,a) = r_h^k(s,a) + \mathbb{E}_\pi \left[\sum_{h'=h+1}^{H} r_{h'}^k(s_{h'}, a_{h'}) | s_h = s, a_h = a \right]$$

To measure the convergence to optimality, we consider an equivalent objective of minimizing the *dynamic regret* (Cheung et al., 2020; Jin et al., 2020),

$$\text{Dyn-Reg}(K) = \sum_{k=1}^{K} \left[V_{1,k}^*(s_1^k) - V_{1,k}^{\pi^k}(s_1^k) \right].$$

We consider a special class of MDPs called *linear Markov decision process* (Melo & Ribeiro, 2007; Bradtke & Barto, 1996; Jin et al., 2020), which assumes both transition function $\mathbb{P}$ and reward function r are linear in a known feature map $\boldsymbol{\phi}(\cdot,\cdot)$. The formal definition is as follows.

Definition 1. *(Linear MDP). The MDP $(\mathcal{S}, \mathcal{A}, H, K, \mathbb{P}, r)$ is a linear MDP with the feature map $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, if for any $(h,k) \in [H] \times [K]$, there exist d unknown measures $\boldsymbol{\mu}_{h,k} = (\boldsymbol{\mu}_{h,k}^1, \ldots, \boldsymbol{\mu}_{h,k}^d)^\top$ on $\mathcal{S}$ and a vector $\boldsymbol{\theta}_{h,k} \in \mathbb{R}^d$ such that*

$$\mathbb{P}_h^k(s'|s,a) = \boldsymbol{\phi}(s,a)^\top \boldsymbol{\mu}_{h,k}(s'), \quad r_h^k(s,a) = \boldsymbol{\phi}(s,a)^\top \boldsymbol{\theta}_{h,k}.$$

Without loss of generality, we assume $\|\boldsymbol{\phi}(s,a)\|_2 \le 1$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, and $\max\{\|\boldsymbol{\mu}_{h,k}\|_2, \|\boldsymbol{\theta}_{h,k}\|_2\} \le \sqrt{d}$ for all $(h,k) \in [H] \times [K]$.

Following (Besbes et al., 2014; Cheung et al., 2019; 2020), we quantify the total variation on $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$ in terms of their respective variation budget $B_{\boldsymbol{\theta}}$ and $B_{\boldsymbol{\mu}}$, and define the total variation budget B as the summation of these two variation budgets:

$$B_{\boldsymbol{\theta}} = \sum_{k=2}^{K} \sum_{h=1}^{H} \|\boldsymbol{\theta}_{h,k} - \boldsymbol{\theta}_{h,k-1}\|_2,$$
$$B_{\boldsymbol{\mu}} = \sum_{k=2}^{K} \sum_{h=1}^{H} \|\boldsymbol{\mu}_{h,k}(\mathcal{S}) - \boldsymbol{\mu}_{h,k-1}(\mathcal{S})\|_2,$$

where $\boldsymbol{\mu}_{h,k}(\mathcal{S})$ is the concatenation of $\boldsymbol{\mu}_{h,k}(s)$ for all states.

3. Minimax Regret Lower Bound

In this section, we derive the minimax regret lower bound for nonstationary linear MDPs. All of the detailed proofs for this section are included in Appendix A.

We first derive the minimax regret lower bound for stationary linear MDP by constructing hard instances, which addresses a problem proposed in (Jin et al., 2020).

Theorem 1. *For any algorithm, if $d \ge 4$ and $T \ge 64(d-3)^2 H$, then there exists at least one stationary linear MDP instance that incurs regret at least $\Omega(d\sqrt{HT})$.*

Remark 1. *Note that this lower bound is tighter than simply applying the minimax regret lower bound for tabular episodic MDP. Recall that the minimax regret lower bound for tabular episodic MDP is $\Omega(\sqrt{SAHT})$ (Osband & Van Roy, 2016), and we can convert any tabular MDP*

into a linear MDP by setting the feature $\boldsymbol{\phi}(\cdot,\cdot)$ as an indicator vector with $d = SA$ dimension (Jin et al., 2020). Thus simply applying the regret lower bound for tabular episodic MDP yields $\Omega\left(\sqrt{dHT}\right)$.

The key step of this proof is to construct the hard-to-learn MDP instances. Inspired by lower bound construction for stochastic contextual bandits (Dani et al., 2008; Lattimore & Szepesvári, 2020), we construct an ensemble of hard-to-learn 3-state linear MDPs, which is illustrated in Fig. 1. This construction can be viewed as a generalization of the lower bound construction for linear contextual bandits (Dani et al., 2008; Lattimore & Szepesvári, 2020). The intuition is that the reward distributions under optimal and suboptimal policies for these instances are close: thus it is statistically hard for any learner to identify the optimal policy. Each linear MDP instance in this constructed ensemble has three states s_0, s_1, s_2 (s_1 and s_2 are absorbing states), and it is characterized by a unique $(d-3)$-dimensional vector $\{\pm\sqrt{(d-3)H}/\sqrt{T}\}^{d-3}$. Specifically, the vector $\boldsymbol{v}$ defines the transition function of the corresponding MDP, as illustrated in Fig. 1. Each action a of this MDP instance is encoded by a $(d-3)$ dimensional vector $\boldsymbol{a} \in \left\{\pm 1/\sqrt{d-3}\right\}^{d-3}$. The reward functions for the three states are fixed regardless of the actions, specifically, $r(s_0,a) = r(s_2,a) = 0, r(s_1,a) = 1, \forall a \in \mathcal{A}$. For each episode, the agent starts at s_0, and ends at step H. The transition functions of the linear MDP parametrized by $\boldsymbol{v}$ are defined as follows,

$$\mathbb{P}(s_1|s_0,a) = \delta + \langle \boldsymbol{a}, \boldsymbol{v}\rangle, \quad \mathbb{P}(s_2|s_0,a) = 1 - \delta - \langle \boldsymbol{a}, \boldsymbol{v}\rangle,$$
$$\mathbb{P}(s_1|s_1,a) = 1, \quad \mathbb{P}(s_2|s_2,a) = 1,$$

where $\delta = \frac{1}{4}$. Notice that the optimal policy for the MDP instance parametrized by $\boldsymbol{v}$ is taking the action that maximizes the probability to reach s_1, which is equivalent to taking the action such that its corresponding vector $\boldsymbol{a}$ satisfies $\text{sgn}(\boldsymbol{a}_i) = \text{sgn}(\boldsymbol{v}_i), \forall i \in [d-3]$. Furthermore, it can be verified that the above MDP instance is indeed a linear MDP, by setting:

$$\boldsymbol{\phi}(s_0,a) = (0,1,\delta,\boldsymbol{a}), \boldsymbol{\phi}(s_1,a) = (1,0,0,\vec{0}),$$
$$\boldsymbol{\phi}(s_2,a) = (0,1,0,\vec{0}), \boldsymbol{\mu}(s_0) = (0,0,0,\vec{0}),$$
$$\boldsymbol{\mu}(s_1) = (1,0,1,\boldsymbol{v}), \boldsymbol{\mu}(s_2) = (0,1,-1,-\boldsymbol{v}),$$
$$\theta = (1,0,0,0).$$

Remark 2. *Note that the above parameters violate the normalization assumption in Def. 1, but it is straightforward to normalize them. We ignore the additional rescaling to clarify the presentation.*

After constructing the ensemble of hard instances, we can derive the minimax regret lower bound for stationary linear MDP. For the detailed proof, please refer to Appendix A.

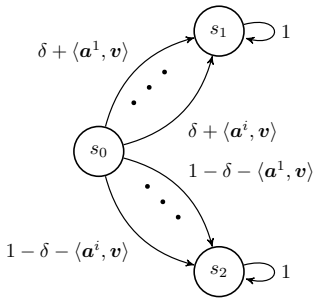

Figure 1. Graphical illustration of the hard-to-learn linear MDP instances with deterministic reward.

Based on Thm. 1, we can derive the dynamic regret lower bound for nonstationary linear MDP.

Theorem 2. *For any algorithm, the dynamic regret is at least $\Omega(B^{1/3}d^{2/3}H^{1/3}T^{2/3})$ for one nonstationary linear MDP instance, if $d \geq 4$, $T \geq 64(d-3)^2H$.*

4. `LSVI-UCB-Restart` Algorithm

In this section, we describe our proposed algorithm `LSVI-UCB-Restart`, and discuss how to tune the hyper-parameters for cases when local variation is known or unknown. For both cases, we present their respective regret bounds. Detailed proofs are deferred to Appendix B.

4.1. Algorithm Description

Our proposed algorithm `LSVI-UCB-Restart` has two key ingredients: least-squares value iteration with upper confidence bound to properly handle the exploration-exploitation trade-off (Jin et al., 2020), and restart strategy to adapt to the unknown nonstationarity. The algorithm is summarized in Alg. 1. From a high-level point of view, our algorithm runs in epochs. At each epoch, we first estimate the action-value function by solving a regularized least-squares problem from historical data, then construct the upper confidence bound for the action-value function, and update the policy greedily w.r.t. action-value function plus the upper confidence bound. Finally, we periodically restart our algorithm to adapt to the nonstationary nature of the environment.

4.2. Regret Analysis

Now we derive the dynamic regret bounds for `LSVI-UCB-Restart`, first introducing additional notation for local variations. We let $B_{\boldsymbol{\theta},\mathcal{E}} = \sum_{k\in\mathcal{E}}\sum_{h=1}^{H}\|\boldsymbol{\theta}_{h,k} - \boldsymbol{\theta}_{h,k-1}\|_2$ and $B_{\boldsymbol{\mu},\mathcal{E}} = \sum_{k\in\mathcal{E}}\sum_{h=1}^{H}\|\boldsymbol{\mu}_{h,k}(\mathcal{S}) - \boldsymbol{\mu}_{h,k-1}(\mathcal{S})\|_2$ be the local variation for $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$ in epoch $\mathcal{E}$.

We proceed to derive the dynamic regret bounds for two

Algorithm 1 `LSVI-UCB-Restart` Algorithm

Require: time horizon T, epoch size W
Set epoch counter $j = 1$.
while $j \leq \lceil \frac{T}{W} \rceil$ **do**
  set $\tau = (j-1)\frac{W}{H}$
  for all $k = \tau, \tau+1, \ldots, \min(\tau + \frac{W}{H} - 1, K)$ **do**
    Receive the initial state s_1^k.
    for all step $h = H, \ldots, 1$ **do**
      $\Lambda_h^k \leftarrow \sum_{l=\tau}^{k-1} \boldsymbol{\phi}(s_h^l, a_h^l)\boldsymbol{\phi}(s_h^l, a_h^l)^\top + I$
      $\boldsymbol{w}_h^k \leftarrow (\Lambda_h^k)^{-1} \sum_{l=\tau}^{k-1} \boldsymbol{\phi}(s_h^l, a_h^l)[r_{h,l}(s_h^l, a_h^l) + \max_a Q_{h+1}^{k-1}(s_{h+1}^l, a)]$
      $Q_h^k(\cdot,\cdot) \leftarrow \min\{(\boldsymbol{w}_h^k)^\top \boldsymbol{\phi}(\cdot,\cdot) + \beta_k \|\boldsymbol{\phi}(\cdot,\cdot)\|_{(\Lambda_h^k)^{-1}}, H\}$
    end for
    for all step $h = 1, \ldots, H$ **do**
      take action $a_h^k \leftarrow \arg\max_a Q_h^k(s_h^k, a)$, and observe s_{h+1}^k
    end for
  end for
  set $j = j + 1$
end while

cases: (1) local variations are known, and (2) local variations are unknown.

4.2.1. Known Local Variations

For the case of known local variations, the dynamic regret upper bound for `LSVI-UCB-Restart` is as follows.

Theorem 3. *If we set* $\beta_k = cdH\sqrt{\log(2dT/p)} + B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d(k-\tau)} + B_{\boldsymbol{\mu},\mathcal{E}} H\sqrt{d(k-\tau)}$, *the dynamic regret of* `LSVI-UCB-Restart` *is* $\tilde{O}(H^{3/2}d^{3/2}TW^{-1/2} + B_{\boldsymbol{\theta}}dW + B_{\boldsymbol{\mu}}dHW)$, *with probability at least* $1-p$.

By properly tuning the epoch size W, we can obtain a tight dynamic regret upper bound.

Corollary 1. *Let* $W = \lceil B^{-2/3}T^{2/3}d^{1/3}H^{-2/3} \rceil H$, *and* $\beta_k = cdH\sqrt{\log(2dW/p)} + B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d(k-\tau)} + B_{\boldsymbol{\mu},\mathcal{E}} H\sqrt{d(k-\tau)}$ *for each epoch.* `LSVI-UCB-Restart` *achieves* $\tilde{O}(B^{1/3}d^{4/3}H^{4/3}T^{2/3})$ *dynamic regret, with probability at least* $1-p$.

Remark 3. *Corollary 1 shows that if local variations are known, we can achieve near-optimal dependency on the the total variation* $B_{\boldsymbol{\theta}}, B_{\boldsymbol{\mu}}$ *and time horizon* T *compared to the lower bound provided in Thm 2. However, the dependency on* d *and* H *is worse. This is not surprising since the dependency on* d *and* H *is not optimal for* `LSVI-UCB` *suggested by Thm 1, thus it is impossible for* `LSVI-UCB-Restart` *to achieve optimal dependency on* d *and* H.

Remark 4. *One concurrent work (Mao et al., 2020) studied nonstationary RL for the tabular setting; their algorithm* `Restart-QUCB` *is also based on combining UCB and periodic restart. When specialized to nonstationary tabular MDP, our algorithm achieves* $\tilde{O}(B^{1/3}S^{4/3}A^{4/3}H^{4/3}T^{2/3})$, *whereas* `Restart-QUCB` *achieves* $\tilde{O}(B^{1/3}S^{1/3}A^{1/3}HT^{2/3})$ *dynamic regret, with better dependency on the size of state space and the planning horizon* H. *This better regret bound is achieved by variance reduction for tabular MDP via reference-advantage decomposition (Zhang et al., 2020).*

4.2.2. Unknown Local Variation

If the local variations are unknown, the dynamic regret bound is as follows.

Theorem 4. *If we set* $\beta = cdH\sqrt{\log(2dT/p)}$, *then the dynamic regret of* `LSVI-UCB-Restart` *is* $\tilde{O}(d^{3/2}H^{3/2}TW^{-1/2} + B_{\boldsymbol{\theta}}d^{1/2}H^{-1/2}W^{3/2} + B_{\boldsymbol{\mu}}d^{1/2}H^{1/2}W^{3/2})$, *with probability at least* $1-p$.

By properly tuning the epoch size W, we can obtain a tight regret bound for the case of unknown local variations as follows.

Corollary 2. *Let* $W = \lceil B^{-1/2}T^{1/2}d^{1/2}H^{-1/2} \rceil H$ *and* $\beta_k = cdH\sqrt{\log(2dW/p)}$. *Then* `LSVI-UCB-Restart` *achieves* $\tilde{O}(B^{1/4}d^{5/4}H^{5/4}T^{3/4})$ *dynamic regret, with probability at least* $1-p$.

Remark 5. *Concurrently, (Touati & Vincent, 2020) propose to combine weighted least-squares value iteration and optimistic principle to solve the same problem, achieving the same regret.*

5. `Ada-LSVI-UCB-Restart`: a Parameter-free Algorithm

In practice, the total variations $B_{\boldsymbol{\theta}}$ and $B_{\boldsymbol{\mu}}$ are unknown. To mitigate this issue, we leverage the bandit-over-bandit mechanism (Cheung et al., 2019) to develop a new algorithm, `ADA-LSVI-UCB-Restart`. `ADA-LSVI-UCB-Restart` keeps running `LSVI-UCB-Restart`, and adaptively choose the time to restart based on the historical rewards. The pseudocode is summarized in Alg. 2, and the choices of hyperparameters are included in the appendix.

Now we present the dynamic regret bound achieved by `Ada-LSVI-UCB-Restart`, and we postpone the detailed proof to the appendix.

Theorem 5. *The dynamic regret of* `Ada-LSVI-UCB-Restart` *is* $\tilde{O}(B^{1/4}d^{5/4}H^{5/4}T^{3/4})$.

Remark 6. *The dynamic regret bound of* `Ada-LSVI-UCB-Restart` *is on the same order as that of* `LSVI-UCB-Restart` *when local variations are unknown. Thus we do not lose too much by not knowing local variations.*

Algorithm 2 `ADA-LSVI-UCB-Restart` Algorithm

Require: time horizon T, block length M, feasible set of epoch size J_w

Initialize α, β, γ and $\{q_{l,1}\}_{l\in[\Delta]}$ according to Eq. 12.
for all $i = 1, 2, \ldots, \lceil T/HM \rceil$ **do**
  Receive the initial state $s_1^{(i-1)H}$
  Update the epoch size selection distribution $\{u_{l,i}\}_{l\in[\Delta]}$ according to Eq. 14
  Sample $l_i \in [\Delta]$ from the updated distribution $\{u_{l,i}\}_{l\in[\Delta]}$, then set the epoch size for block i as $W_i = \lfloor M^{l_i/\lfloor \ln M \rfloor} \rfloor H$.
  for all $t = (i-1)MH + 1, \ldots, \min(iMH, T)$ **do**
    Run `LSVI-UCB-Restart` algorithm with epoch size W_i
  end for
  After observing the total reward for block i, $R_i(W_i, s_1^{(i-1)H})$, update the estimated total reward of running different epoch sizes $\{q_{l,i+1}\}_{l\in[\Delta]}$ according to Eq. 15
end for

6. Conclusion and Future Work

In this paper, we studied nonstationary RL with time-varying reward and transition functions. We focused on the class of nonstationary linear MDPs such that linear function approximation is sufficient to realize any value function. We first incorporated the epoch start strategy into `LSVI-UCB` algorithm (Jin et al., 2020) to propose an algorithm with low dynamic regret when the total variations are known. We then designed a parameter-free algorithm that enjoys a slightly worse dynamic regret bound without knowing the total variations. We derived a minimax regret lower bound is for nonstationary linear MDPs to demonstrate that our proposed algorithms are near-optimal. A number of future directions are of interest. An immediate step is to investigate whether the dependence on the dimension d and planning horizon H in our bounds can be improved, and whether the minimax regret lower bound can also be improved. It would also be interesting to investigate the setting of nonstationary RL under general function approximation (Wang et al., 2020), which is closer to modern RL algorithms in practice.

Acknowledgements

We thank anonymous reviewers for their thoughtful suggestions that helped improve our presentation of the paper. This work was funded in part by the IBM-Illinois Center for Cognitive Computing Systems Research (C3SR), a research collaboration as part of the IBM AI Horizons Network

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Proc. 25th Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, pp. 2312–2320, December 2011.

Agrawal, S. and Jia, R. Learning in structured MDPs with convex cost functions: Improved regret bounds for inventory management. In *Proc. 20th ACM Conf. Electron. Commer. (EC'19)*, pp. 743–744, June 2019.

Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R., Schneider, J., Tezak, N., Tworek, J., Welinder, P., Weng, L., Yuan, Q., Zaremba, W., and Zhang, L. Solving rubik's cube with a robot hand. arXiv:1910.07113 [cs.LG]., October 2019.

Auer, P., Jaksch, T., and Ortner, R. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010.

Besbes, O., Gur, Y., and Zeevi, A. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Proc. 28th Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, pp. 199–207, December 2014.

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

Bretagnolle, J. and Huber, C. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(2):119–137, 1979.

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

Cai, H., Ren, K., Zhang, W., Malialis, K., Wang, J., Yu, Y., and Guo, D. Real-time bidding by reinforcement learning in display advertising. In *Proc. 10th ACM Int. Conf. Web Search Data Min. (WSDM '17)*, pp. 661–670, February 2017.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *Proc. 37th Int. Conf. Mach. Learn. (ICML 2020)*, July 2020.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. Learning to optimize under non-stationarity. In *Proc. 22nd Int. Conf. Artif. Intell. Stat. (AISTATS 2019)*, pp. 1079–1087, April 2019.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. Nonstationary reinforcement learning: The blessing of (more) optimism. In *Proc. 37th Int. Conf. Mach. Learn. (ICML 2020)*, July 2020.

Cover, T. M. and Pombra, S. Gaussian feedback capacity. *IEEE Transactions on Information Theory*, 35(1):37–43, January 1989.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. In *Proc. 21st Annual Conf. Learning Theory (COLT 2008)*, pp. 355–366, July 2008.

Foster, D. J., Rakhlin, A., Simchi-Levi, D., and Xu, Y. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *arXiv preprint arXiv:2010.03104*, 2020.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Proc. Conf. Learning Theory (COLT)*, pp. 2137–2143, July 2020.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, Cambridge, UK, 2020.

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., and Başar, T. Near-optimal regret bounds for model-free rl in non-stationary episodic mdps. *arXiv preprint arXiv:2010.03161*, 2020.

Melo, F. S. and Ribeiro, M. I. Q-learning with linear function approximation. In *Proc. Int. Conf. Computational Learning Theory (COLT 2007)*, pp. 308–322, June 2007.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. Human-level control through deep reinforcement learning. *Nature*, 518 (7540):529–533, February 2015.

Modi, A., Jiang, N., Tewari, A., and Singh, S. Sample complexity of reinforcement learning using linearly combined model ensembles. In *Proc. 23rd Int. Conf. Artif. Intell. Stat. (AISTATS 2020)*, pp. 2010–2020, August 2020.

Neu, G. and Olkhovskaya, J. Online learning in MDPs with linear function approximation and bandit feedback. arXiv:2007.01612 [cs.LG]., July 2020.

Osband, I. and Van Roy, B. On lower bounds for regret in reinforcement learning. arXiv:1608.02732 [stat.ML]., August 2016.

Shalev-Shwartz, S., Shammah, S., and Shashua, A. Safe, multi-agent, reinforcement learning for autonomous driving. arXiv:1610.03295 [cs.AI]., October 2016.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, October 2017.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, December 2018. doi: 10.1126/science.aar6404.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, USA, 2018.

Touati, A. and Vincent, P. Efficient learning in non-stationary linear markov decision processes. *arXiv preprint arXiv:2010.12870*, 2020.

Wang, R., Salakhutdinov, R., and Yang, L. F. Provably efficient reinforcement learning with general value function approximation. In *Proc. 34th Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, May 2020.

Wei, C.-Y., Jafarnia-Jahromi, M., Luo, H., and Jain, R. Learning infinite-horizon average-reward MDPs with linear function approximation. arXiv:2007.11849 [cs.LG]., July 2020.

Yang, L. F. and Wang, M. Sample-optimal parametric Q-learning using linearly additive features. In *Proc. 36th Int. Conf. Mach. Learn. (ICML 2019)*, June 2019.

Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. Learning near optimal policies with low inherent Bellman error. *Proc. 37th Int. Conf. Mach. Learn. (ICML 2020)*, July 2020.

Zhang, Z., Zhou, Y., and Ji, X. Almost optimal model-free reinforcement learningvia reference-advantage decomposition. In *Proc. 34th Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2020.

Zhao, P., Zhang, L., Jiang, Y., and Zhou, Z.-H. A simple approach for non-stationary linear bandits. In *Proc. 23rd Int. Conf. Artif. Intell. Stat. (AISTATS 2020)*, pp. 746–755, August 2020.

Zhao, Y., Kosorok, M. R., and Zeng, D. Reinforcement learning design for cancer clinical trials. *Statistics in Medicine*, 28(26):3294–3315, November 2009.

A. Proofs in Section 3

In this section, we prove the minimax regret lower bound of nonstationary linear MDP. We first prove the regret lower bound of stationary linear MDP.

Proof of Theorem 1. Let $\mathbb{P}^{\pi}_{t,\boldsymbol{v}}$ (assume t is a multiple of H) be the probability distribution of $\{a_1^1, \sum_{h=1}^H r_h^1, a_1^2, \sum_{h=1}^H r_h^2, \ldots, a_1^{t/H}, \sum_{h=1}^H r_h^{t/H}\}$ of running algorithm π on linear MDP parametrized by $\boldsymbol{v}$. First note that by the Markov property of π, we can decompose $D_{KL}(\mathbb{P}^{\pi}_{t,\boldsymbol{v}}||\mathbb{P}^{\pi}_{t,\boldsymbol{v}'})$ as

$$\sum_{l=1}^{t/H} \mathbb{E} D_{KL}\left(\mathbb{P}\left(\sum_{h=1}^H r_h^l | a_1^l, \boldsymbol{v}\right) || \mathbb{P}\left(\sum_{h=1}^H r_h^l | a_1^l, \boldsymbol{v}'\right)\right).$$

Recall that due to our hard cases construction, the first step in every episode determines the distribution of the total reward of that episode, thus

$$\begin{aligned}
&D_{KL}\left(\mathbb{P}\left(\sum_{h=1}^H r_h^l | a_1^l, \boldsymbol{v}\right) || \mathbb{P}\left(\sum_{h=1}^H r_h^l | a_1^l, \boldsymbol{v}'\right)\right) \\
&= \big(\delta + \langle \boldsymbol{a}_1^l, \boldsymbol{v}\rangle\big) \log \frac{\delta + \langle \boldsymbol{a}_1^l, \boldsymbol{v}\rangle}{\delta + \langle \boldsymbol{a}_1^l, \boldsymbol{v}'\rangle} + \big(1 - \delta - \langle \boldsymbol{a}_1^l, \boldsymbol{v}\rangle\big) \log \frac{1 - \delta - \langle \boldsymbol{a}_1^l, \boldsymbol{v}\rangle}{1 - \delta - \langle \boldsymbol{a}_1^l, \boldsymbol{v}'\rangle} \qquad (1)
\end{aligned}$$

We bound the KL divergence in (1) applying the following lemma.

Lemma 1. *(Auer et al., 2010) If $0 \le \delta' \le 1/2$ and $\epsilon' \le 1 - 2\delta'$, then*

$$\delta' \log \frac{\delta'}{\delta' + \epsilon'} + (1 - \delta') \log \frac{1 - \delta'}{1 - \delta' - \epsilon'} \le \frac{2(\epsilon')^2}{\delta'}.$$

To apply Lemma 1, we let $\langle \boldsymbol{a}_1^l, \boldsymbol{v}\rangle + \delta = \delta'$, $\langle \boldsymbol{v} - \boldsymbol{v}', \boldsymbol{a}_1^l\rangle = \epsilon'$. Thus we must ensure the following inequalities hold for any $\boldsymbol{a}, \boldsymbol{v}, \boldsymbol{v}'$:

$$\begin{aligned}
&\langle \boldsymbol{a}, \boldsymbol{v}\rangle + \delta \le \frac{(d-3)\sqrt{H}}{\sqrt{T}} + \delta \le 1/2 \\
&\langle \boldsymbol{v} - \boldsymbol{v}', \boldsymbol{a}\rangle \le \frac{2(d-3)\sqrt{H}}{\sqrt{T}} \le 1 - 2\left(\frac{(d-3)\sqrt{H}}{\sqrt{T}} + \delta\right) \le 1 - 2\delta'.
\end{aligned}$$

To guarantee the above inequalities hold, we can set $\delta = \frac{1}{4}$ and let $\frac{(d-3)\sqrt{H}}{\sqrt{T}} \le \frac{1}{8}$. Now we get back to bounding Eq. 1. Let $\Delta = \frac{(d-3)\sqrt{H}}{\sqrt{T}}$ and suppose $\boldsymbol{v}$ and $\boldsymbol{v}'$ only differ in one coordinate. Then

$$D_{KL}\left(\mathbb{P}\left(\sum_{h=1}^H r_h^l | a_1^l, \boldsymbol{v}\right) || \mathbb{P}\left(\sum_{h=1}^H r_h^l | a_1^l, \boldsymbol{v}'\right)\right) \le \frac{8\Delta^2 \frac{1}{(d-3)^2}}{\delta - \Delta} \le \frac{16\Delta^2}{\delta(d-3)^2} \le \frac{64H}{T}.$$

Furthermore, let $E_{i,b}$ be the the following event:

$$|\{l \in [K] : \mathrm{sgn}(\boldsymbol{a}_1^l)_i \ne \mathrm{sgn}(b)\}| \ge \frac{1}{2}K.$$

Let $q_{i,\boldsymbol{v}} = \mathbb{P}[E_{i,v_i}|\boldsymbol{v}]$, the probability that the agent is taking sub-optimal action for the i-th coordinate for at least half of the episodes given that the underlying linear MDP is parameteriezed by $\boldsymbol{v}$. We can then lower bound the regret of any algorithm when running on linear MDP parameterized by $\boldsymbol{v}$ as:

$$\begin{aligned}
\mathrm{Reg}_{\boldsymbol{v}}(T) &\ge \sum_{i=1}^{d-3} q_{i,\boldsymbol{v}} K(H-1)\sqrt{\frac{H}{T}} \\
&\ge \left(\sqrt{TH} - \sqrt{K}\right) \sum_{i=1}^{d-3} q_{i,\boldsymbol{v}}, \qquad (2)
\end{aligned}$$

since whenever the learner takes a sub-optimal action that differs from the optimal action by one coordinate, it will incur $2\sqrt{\frac{H}{T}}(H-1)$ expected regret. Next we take the average over 2^{d-3} linear MDP instances to show that on average it incurs $\Omega(d\sqrt{HT})$ regret, thus there exists at least one instance incurring $\Omega(d\sqrt{HT})$ regret. Before that, we need to bound the summation of bad events under two close linear MDP instances. Denote the vector which is only different from $\boldsymbol{v}$ in i-th coordinate as $\boldsymbol{v}^{\oplus i}$. Then we have

$$\begin{aligned} q_{i,\boldsymbol{v}} + q_{i,\boldsymbol{v}^{\oplus i}} &= \mathbb{P}[E_{i,\boldsymbol{v}_i}|\boldsymbol{v}] + \mathbb{P}[E_{i,\boldsymbol{v}_i^{\oplus i}}|\boldsymbol{v}^{\oplus i}] \\ &= \mathbb{P}[E_{i,\boldsymbol{v}_i}|\boldsymbol{v}] + \mathbb{P}[\bar{E}_{i,\boldsymbol{v}_i}|\boldsymbol{v}^{\oplus i}] \\ &\geq \frac{1}{2}\exp(-D_{KL}(P_{T,\boldsymbol{v}}||P_{T,\boldsymbol{v}^{\oplus i}})) \\ &\geq \frac{1}{2}\exp(-64), \end{aligned} \tag{3}$$

where the inequality is due to Bretagnolle-Huber inequality (Bretagnolle & Huber, 1979). Now we are ready to lower bound the average regret over all linear MDP instances.

$$\begin{aligned} \frac{1}{2^{d-3}}\sum_{\boldsymbol{v}} \text{Reg}_{\boldsymbol{v}}(T) &\geq \frac{\sqrt{HT}-\sqrt{K}}{2^{d-3}}\sum_{\boldsymbol{v}}\sum_{i=1}^{d-3} q_{i,\boldsymbol{v}} \\ &\geq \frac{\sqrt{HT}-\sqrt{K}}{2^{d-3}}\sum_{i=1}^{d-3}\sum_{\boldsymbol{v}} \frac{q_{i,\boldsymbol{v}}+q_{i,\boldsymbol{v}^{\oplus i}}}{2} \\ &\geq \frac{\sqrt{HT}-\sqrt{K}}{2^{d-3}} 2^{d-3}\frac{1}{4}e^{-64}(d-3) \\ &\gtrsim \Omega(d\sqrt{HT}) \end{aligned}$$

where the first inequality is due to (2), and the third inequality is due to (3). □

Based on Thm. 1, we can derive the minimax dynamic regret for nonstationary linear MDP.

Proof of Thm. 2. We construct the hard instance as follows: We first divide the whole time horizon T into $\lceil \frac{K}{N} \rceil$ intervals, where each interval has $\lceil \frac{K}{N} \rceil$ episodes (the last interval might be shorter if K is not a multiple of N). For each interval, the linear MDP is fixed and parameterized by a $\boldsymbol{v} \in \{\pm\frac{\sqrt{(d-3)}}{\sqrt{N}}\}^{d-3}$ which we define when constructing the hard instances in Thm. 1. Note that different intervals are completely decoupled, thus information is not passed across intervals. For each interval, it incurs regret at least $\Omega(d\sqrt{H^2N})$ by Thm. 1. Thus the total regret is at least

$$\begin{aligned} \text{Dyn-Reg}(T) &\gtrsim (\lceil \frac{K}{N} \rceil - 1)\Omega(d\sqrt{H^2N}) \\ &\gtrsim \Omega(d\sqrt{H^2K^2}N^{-1/2}). \end{aligned} \tag{4}$$

Intuitively, we would like N to be as small as possible to obtain a tight lower bound. However, due to our construction, the total variation for two consecutive blocks is upper-bounded by

$$\sqrt{\sum_{i=1}^{d-3}\frac{4(d-3)}{N}} = \frac{2(d-3)}{\sqrt{N}}.$$

Note that the total time variation for the whole time horizon is B and by definition $B \geq \frac{2(d-3)}{\sqrt{N}}(\lfloor \frac{K}{N} \rfloor - 1)$, which implies$N \gtrsim \Omega(B^{-2/3}d^{2/3}K^{2/3})$. Substituting the lower bound of N into (4), we have

$$\text{Dyn-Reg}(T) \gtrsim \Omega(B^{1/3}d^{2/3}K^{2/3}H) \gtrsim \Omega(B^{1/3}d^{2/3}H^{1/3}T^{2/3})$$

which concludes the proof. □

B. Proofs in Section 4

Here we provide the proofs in Sec. 4. First, we introduce some notations we use throughout the proof. We let $\boldsymbol{w}_h^k$, Λ_h^k and Q_h^k as the parameters and action-value function estimate in episode k for step h. Denote value function estimate as $V_h^k(s) = \max_a Q_h^k(s,a)$. For any policy π, we let $\boldsymbol{w}_{h,k}^\pi$, $Q_{h,k}^\pi$ be the ground-truth parameter and action-value function for that policy in episode k for step h. We also abbreviate $\boldsymbol{\phi}(s_h^l, a_h^l)$ as $\boldsymbol{\phi}_h^l$, and $\mathbb{E}_{s'\sim\mathbb{P}_h^k(\cdot|s,a)}[V_{h+1}(s')] = [\mathbb{P}_h^k V_{h+1}](s,a)$, for notational simplicity.

The regret bound analysis is not just a simple combination of (Jin et al., 2020) and (Zhao et al., 2020), since the estimation error incurred by environmental drift can propagate through the whole episode in an arbitrary manner. Contrarily for the bandit problem, one need not consider the error propagation problem due to unit planing horizon. To derive the dynamic regret upper bounds, we need the following lemma to control the fluctuation of least-squares value iteration.

Lemma 2. *(Modified from (Jin et al., 2020)) Denote τ to be the first episode in the epoch which contains episode k. There exists an absolute constant C such that the following event E,*

$$\begin{aligned}&\left\|\sum_{l=\tau}^{k-1}\boldsymbol{\phi}_h^l[V_{h+1}^k(s_{h+1}^l) - \mathbb{P}_h^l V_{h+1}^k(s_h^l,a_h^l)]\right\|_{(\Lambda_h^k)^{-1}}\\ &\quad\le CdH\sqrt{\log[2(c_\beta+1)dW/p]}, \quad \forall(k,h)\in\mathcal{E}\times[H].\end{aligned}$$

happens with probability at least $1-p/2$.

We first work on the case when local variation is known and then consider the case when local variation is unknown.

B.1. Case 1: Known Local Variation

Before we prove the regret upper bound, we need some additional lemmas.
The first lemma is used to control the fluctuations in least-squares value iteration, when performed on the value function estimate $V_h^k(\cdot)$ maintained in Alg. 1.

Proof of Lemma 2. The lemma is slightly different than Lemma B.3 in (Jin et al., 2020), since they assume $\mathbb{P}_h$ is fixed for different episodes. It can be verified that the proof for stationary case still holds in our case without any modifications since the results in (Jin et al., 2020) holds for least-squares value iteration for arbitrary function in the function class of our interest, i.e., $\{V|V=\{\boldsymbol{\phi}(\cdot,\cdot),\boldsymbol{w}\},\boldsymbol{w}\in\mathbb{R}^d\}$. □

We then proceed to derive the error bound for the action-value function estimate maintained in the algorithm for any policy.

Lemma 3. *Under event E defined in Lemma 2, we have for any policy π, $\forall s,a,h,k\in\mathcal{S}\times\mathcal{A}\times[H]\times\mathcal{E}$,*

$$|\langle\boldsymbol{\phi}(s,a),\boldsymbol{w}_h^k\rangle - Q_{h,k}^\pi(s,a) - \mathbb{P}_h^k(V_{h+1}^k - V_{h+1,k}^\pi)(s,a)| \le \beta_k\left\|\boldsymbol{\phi}(s,a)\right\|_{(\Lambda_h^k)^{-1}},$$

where $\beta_k = C_0 dH\sqrt{\log(2dW/p)} + B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d(k-\tau)} + B_{\boldsymbol{\mu},\mathcal{E}}H\sqrt{d(k-\tau)}$ and τ is the first episode in the current epoch.

Proof of Lemma 3. Note that $Q^{\pi}_{h,k}(s,a) = \langle \phi(s,a), \boldsymbol{w}^{\pi}_{h,k} \rangle$. First we can decompose $\boldsymbol{w}^k_h - \boldsymbol{w}^{\pi}_{h,k}$ as

$$
\begin{aligned}
\boldsymbol{w}^k_h - \boldsymbol{w}^{\pi}_{h,k} &= (\Lambda^k_h)^{-1} \sum_{l=\tau}^{k-1} \boldsymbol{\phi}^l_h [r^l_h(s^l_h, a^l_h) + V^k_{h+1}(s^l_{h+1})] - \boldsymbol{w}^{\pi}_{h,k} \\
&= (\Lambda^k_h)^{-1} \{ -\boldsymbol{w}^{\pi}_{h,k} + \sum_{l=\tau}^{k-1} \boldsymbol{\phi}^l_h [V^k_{h+1}(s^l_{h+1}) - \mathbb{P}^k_h V^{\pi}_{h+1,k}(s^l_h, a^l_h)] + \sum_{l=\tau}^{k-1} \boldsymbol{\phi}^l_h [r^l_h(s^l_h, a^l_h) - r^k_h(s^l_h, a^l_h)] \} \\
&= \underbrace{-(\Lambda^k_h)^{-1} \boldsymbol{w}^{\pi}_{h,k}}_{①} + \underbrace{(\Lambda^k_h)^{-1} \sum_{l=\tau}^{k-1} \boldsymbol{\phi}^l_h [V^k_{h+1}(s^l_{h+1}) - \mathbb{P}^l_h V^k_{h+1}(s^l_h, a^l_h)]}_{②} \\
&\quad + \underbrace{(\Lambda^k_h)^{-1} \sum_{l=\tau}^{k-1} \boldsymbol{\phi}^l_h [(\mathbb{P}^l_h - \mathbb{P}^k_h) V^k_{h+1}(s^l_h, a^l_h)]}_{③} + \underbrace{(\Lambda^k_h)^{-1} \sum_{l=\tau}^{k-1} \boldsymbol{\phi}^l_h \mathbb{P}^k_h (V^k_{h+1} - V^{\pi}_{h+1,k})(s^l_h, a^l_h)}_{④} \\
&\quad + \underbrace{(\Lambda^k_h)^{-1} \sum_{l=\tau}^{k-1} \boldsymbol{\phi}^l_h [r^l_h(s^l_h, a^l_h) - r^k_h(s^l_h, a^l_h)]}_{⑤} .
\end{aligned}
$$

We bound the individual terms on right side one by one. For the first term,

$$
\begin{aligned}
|\langle \boldsymbol{\phi}(s,a), ① \rangle| &= |\langle \boldsymbol{\phi}(s,a), (\Lambda^k_h)^{-1} \boldsymbol{w}^{\pi}_{h,k} \rangle| \\
&\le \left\| \boldsymbol{w}^{\pi}_{h,k} \right\| \left\| \boldsymbol{\phi}(s,a) \right\|_{(\Lambda^k_h)^{-1}} \\
&\le 2H\sqrt{d} \left\| \boldsymbol{\phi}(s,a) \right\|_{(\Lambda^k_h)^{-1}} ,
\end{aligned}
$$

where the last inequality is due to Lemma 9. For the second term, we know that under event E defined in Lemma 2,

$$
|\langle \boldsymbol{\phi}(s,a), ② \rangle| \le CdH \sqrt{\log[2(c_\beta + 1)dW/p]} \left\| \boldsymbol{\phi}(s,a) \right\|_{(\Lambda^k_h)^{-1}} .
$$

For the third term,

$$
\begin{aligned}
\langle \boldsymbol{\phi}(s,a), ③ \rangle &= \langle \boldsymbol{\phi}(s,a), (\Lambda^k_h)^{-1} \sum_{l=\tau}^{k-1} \boldsymbol{\phi}^l_h [(\mathbb{P}^l_h - \mathbb{P}^k_h) V^k_{h+1}(s^l_h, a^l_h)] \rangle \\
&\le \sum_{l=\tau}^{k-1} |\boldsymbol{\phi}(s,a)^\top (\Lambda^k_h)^{-1} \boldsymbol{\phi}^l_h| [(\mathbb{P}^l_h - \mathbb{P}^k_h) V^k_{h+1}(s^l_h, a^l_h)] \\
&\le B_{\boldsymbol{\mu},\mathcal{E}} H \sum_{l=\tau}^{k-1} |\boldsymbol{\phi}(s,a)^\top (\Lambda^k_h)^{-1} \boldsymbol{\phi}^l_h| \\
&\le B_{\boldsymbol{\mu},\mathcal{E}} H \sqrt{\sum_{l=\tau}^{k-1} \left\| \boldsymbol{\phi}(s,a) \right\|^2_{(\Lambda^k_h)^{-1}}} \sqrt{\sum_{l=\tau}^{k-1} (\boldsymbol{\phi}^l_h)^\top (\Lambda^k_h)^{-1} \boldsymbol{\phi}^l_h} \\
&\le \sqrt{d(k-\tau)} B_{\boldsymbol{\mu},\mathcal{E}} H \left\| \boldsymbol{\phi}(s,a) \right\|_{(\Lambda^k_h)^{-1}} ,
\end{aligned}
$$

where the first three inequalities are due to Cauchy-Schwarz inequality and boundedness of $\mathbb{P}^l_h - \mathbb{P}^k_h$ and V^k_{h+1}, and the last inequality is due to Lemma 10.

For the fourth term,

$$\begin{aligned}\langle \boldsymbol{\phi}(s,a), ④\rangle &= \langle \boldsymbol{\phi}(s,a), (\Lambda_h^k)^{-1}\sum_{l=\tau}^{k-1}\boldsymbol{\phi}_h^l \mathbb{P}_h^k(V_{h+1}^k - V_{h+1,k}^\pi)(s_h^l,a_h^l)\rangle \\ &= \langle \boldsymbol{\phi}(s,a), (\Lambda_h^k)^{-1}\sum_{l=\tau}^{k-1}\boldsymbol{\phi}_h^l(\boldsymbol{\phi}_h^l)^\top \int (V_{h+1}^k(s') - V_{h+1,k}^\pi(s'))d\boldsymbol{\mu}_{h,k}(s')\rangle \\ &= \underbrace{\langle \boldsymbol{\phi}(s,a), \int (V_{h+1}^k - V_{h+1,k}^\pi)(s')d\boldsymbol{\mu}_{h,k}(s')\rangle}_{⑥} - \underbrace{\langle \boldsymbol{\phi}(s,a), (\Lambda_h^k)^{-1}\int (V_{h+1}^k - V_{h+1,k}^\pi)(s')d\boldsymbol{\mu}_{h,k}(s')\rangle}_{⑦},\end{aligned}$$

where $⑥ = [\mathbb{P}_h^k(V_{h+1}^k - V_{h+1,k}^\pi)](s,a)$ and $⑦ \le 2H\sqrt{d}\,\|\boldsymbol{\phi}(s,a)\|_{(\Lambda_h^k)^{-1}}$ due to Cauchy-Schwarz inequality.

For the fifth term,

$$\begin{aligned}\langle \boldsymbol{\phi}(s,a), ⑤\rangle &= \langle \boldsymbol{\phi}(s,a), (\Lambda_h^k)^{-1}\sum_{l=\tau}^{k-1}\boldsymbol{\phi}_h^l[r_h^l(s_h^l,a_h^l) - r_h^k(s_h^l,a_h^l)]\rangle \\ &\le \sum_{l=\tau}^{k-1}|\boldsymbol{\phi}(s,a)^\top(\Lambda_h^k)^{-1}\boldsymbol{\phi}_h^l||r_h^l(s_h^l,a_h^l) - r_h^k(s_h^l,a_h^l)| \\ &\le \sqrt{d(k-\tau)}B_{\boldsymbol{\theta},\mathcal{E}}\,\|\boldsymbol{\phi}(s,a)\|_{(\Lambda_h^k)^{-1}},\end{aligned}$$

where the inequalities are derived similarly as bounding the third term. After combining all the upper bounds for these individual terms, we have

$$\begin{aligned}&|\langle \boldsymbol{\phi}(s,a), \boldsymbol{w}_h^k\rangle - Q_{h,k}^\pi(s,a) - \mathbb{P}_h^k\left(V_{h+1}^k - V_{h+1,k}^\pi\right)(s,a)| \\ \le& 4H\sqrt{d}\,\|\boldsymbol{\phi}(s,a)\|_{(\Lambda_h^k)^{-1}} + CdH\sqrt{\log[2(c_\beta+1)dW/p]}\,\|\boldsymbol{\phi}(s,a)\|_{(\Lambda_h^k)^{-1}} \\ &+ B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d(k-\tau)}\,\|\boldsymbol{\phi}(s,a)\|_{(\Lambda_h^k)^{-1}} + B_{\boldsymbol{\mu},\mathcal{E}}H\sqrt{d(k-\tau)}\,\|\boldsymbol{\phi}(s,a)\|_{(\Lambda_h^k)^{-1}} \\ \le& C_0dH\sqrt{\log[2dW/p]}\,\|\boldsymbol{\phi}(s,a)\|_{(\Lambda_h^k)^{-1}} + B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d(k-\tau)}\,\|\boldsymbol{\phi}(s,a)\|_{(\Lambda_h^k)^{-1}} \\ &+ B_{\boldsymbol{\mu},\mathcal{E}}H\sqrt{d(k-\tau)}\,\|\boldsymbol{\phi}(s,a)\|_{(\Lambda_h^k)^{-1}}.\end{aligned}$$

The second inequality holds if we choose a sufficiently large absolute constant C_0. □

The next lemma implies that the action-value function estimate we maintained in Alg. 1 is always an optimistic upper bound of the optimal action-value function with high confidence under event E defined in Lemma 2, if we know the local variation.

Lemma 4. *Under event E defined in Lemma 2, for episode k, if we set $\beta_k = cdH\sqrt{\log(2dW/p)} + B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d(k-\tau)} + B_{\boldsymbol{\mu},\mathcal{E}}H\sqrt{d(k-\tau)}$, we have*

$$Q_h^k(s,a) \ge Q_{h,k}^*, \quad \forall (s,a,h,k) \in \mathcal{S}\times\mathcal{A}\times[H]\times\mathcal{E}.$$

Proof of Lemma 4. We prove this by induction. First prove the base case when $h = H$. According to Lemma 3, we have

$$|\langle \boldsymbol{\phi}(s,a), \boldsymbol{w}_H^k\rangle - Q_{H,k}^*(s,a)| \le \beta_k\,\|\boldsymbol{\phi}(s,a)\|_{(\Lambda_H^k)^{-1}},$$

which implies

$$Q_H^k(s,a) = \min\{\langle \boldsymbol{w}_H^k, \boldsymbol{\phi}(s,a)\rangle + \beta_k\,\|\boldsymbol{\phi}(s,a)\|_{(\Lambda_H^k)^{-1}}, H\} \ge Q_{H,k}^*(s,a).$$

Now suppose the statement holds true at step $h+1$, then for step h, due to Lemma 3, we have

$$|\langle \boldsymbol{\phi}(s,a), \boldsymbol{w}_h^k\rangle - Q_{h,k}^\pi(s,a) - \mathbb{P}_h^k(V_{h+1}^k - V_{h+1,k}^*)(s,a)| \le \beta_k\,\|\boldsymbol{\phi}(s,a)\|_{(\Lambda_h^k)^{-1}}.$$

By the induction hypothesis, we have $\mathbb{P}_h^k(V_{h+1}^k - V_{h+1,k}^*)(s,a) \geq 0$, thus

$$Q_h^k(s,a) = \min\{\langle \boldsymbol{w}_h^k, \boldsymbol{\phi}(s,a)\rangle + \beta_k \left\|\boldsymbol{\phi}(s,a)\right\|_{(\Lambda_H^k)^{-1}}, H\} \geq Q_{h,k}^*(s,a).$$

□

Next we derive the bound for the gap between the value function estimate and the ground-truth value function for the executing policy π^k, $\delta_h^k = V_h^k(s_h^k) - V_{h,k}^{\pi^k}(s_h^k)$, in a recursive manner.

Lemma 5. *Let $\delta_h^k = V_h^k(s_h^k) - V_{h,k}^{\pi^k}(s_h^k)$, $\zeta_{h+1}^k = \mathbb{E}[\delta_{h+1}^k | s_h^k, a_h^k] - \delta_{h+1}^k$. Under event E defined in Lemma 2, we have for all $(k,h) \in \mathcal{E} \times [H]$,*

$$\delta_h^k \leq \delta_{h+1}^k + \zeta_{h+1}^k + 2\beta_k \left\|\boldsymbol{\phi}_h^k\right\|_{(\Lambda_h^k)^{-1}}.$$

Proof. By Lemma 3, for any $(s,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times \mathcal{E}$,

$$Q_h^k(s,a) - Q_h^{\pi^k}(s,a) \leq \mathbb{P}_h^k(V_{h+1}^k - V_{h+1,k}^{\pi^k})(s,a) + 2\beta_k \left\|\boldsymbol{\phi}(s,a)\right\|_{(\Lambda_H^k)^{-1}}.$$

Note that $Q_h^k(s_h^k, a_h^k) = \max_a Q_h^k(s_h^k, a) = V_h^k(s_h^k)$ according to Algorithm 1, and $Q_{h,k}^{\pi^k}(s_h^k, a_h^k) = V_{h,k}^{\pi^k}(s_h^k)$ by the definition. Thus,

$$\delta_h^k \leq \delta_{h+1}^k + \zeta_{h+1}^k + 2\beta_k \left\|\boldsymbol{\phi}_h^k\right\|_{(\Lambda_h^k)^{-1}}.$$

□

Now we are ready to derive the regret bound within one epoch.

Theorem 6. *For each epoch $\mathcal{E}$ with epoch size W, set β in the k-th episode as $\beta_k = cdH\sqrt{\log(2dW/p)} + B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d(k-\tau)} + B_{\boldsymbol{\mu},\mathcal{E}}H\sqrt{d(k-\tau)}$, where c is an absolute constant and $p \in (0,1)$. Then the dynamic regret within that epoch is $\tilde{O}(H^{3/2}d^{3/2}W^{1/2} + B_{\boldsymbol{\theta},\mathcal{E}}dW + B_{\boldsymbol{\mu},\mathcal{E}}dHW)$ with probability at least $1-p$.*

Proof of Theorem 6. We denote the dynamic regret within that epoch as Dyn-Reg($\mathcal{E}$). We define $\delta_h^k = V_h^k(s_h^k) - V_{h,k}^{\pi^k}(s_h^k)$ and $\zeta_{h+1}^k = \mathbb{E}[\delta_{h+1}^k | s_h^k, a_h^k] - \delta_{h+1}^k$ as in Lemma 8. We derive the dynamic regret within a epoch $\mathcal{E}$ (the length of this epoch is W which is equivalent to $\frac{W}{H}$ episodes) conditioned on the event E defined in Lemma 2 which happens with probability at least $1 - p/2$.

$$\begin{aligned}
\text{Dyn-Reg}(\mathcal{E}) &= \sum_{k\in\mathcal{E}} \left[V_{1,k}^*(s_1^k) - V_{1,k}^{\pi^k}\right] \\
&\leq \sum_{k\in\mathcal{E}} \left[V_1^k(s_1^k) - V_{1,k}^{\pi^k}\right] \\
&\leq \sum_{k\in\mathcal{E}} \delta_1^k \\
&\leq \sum_{k\in\mathcal{E}}\sum_{h=1}^H \zeta_h^k + 2\sum_{k\in\mathcal{K}} \beta_k \sum_{h=1}^H \left\|\boldsymbol{\phi}_h^k\right\|_{(\Lambda_h^k)^{-1}},
\end{aligned} \tag{5}$$

where the first inequality is due to Lemma 4, the third inequality is due to Lemma 5. For the first term in the right side, since V_h^k is independent of the new observation s_h^k, $\{\zeta_h^k\}$ is a martingale difference sequence. Applying the Azuma-Hoeffding inequlity, we have for any $t > 0$,

$$\mathbb{P}\left(\sum_{k\in\mathcal{E}}\sum_{h=1}^H \zeta_h^k \geq t\right) \geq \exp(-t^2/(2WH^2)).$$

Hence with probability at least $1 - p/2$, we have

$$\sum_{k\in\mathcal{E}}\sum_{h=1}^{H}\zeta_h^k \le 2H\sqrt{W\log(2dW/p)}. \tag{6}$$

For the second term, we bound via Cauchy-Schwarz inequality:

$$\begin{aligned}
2\sum_{k\in\mathcal{E}}\beta_k\sum_{h=1}^{H}\left\|\boldsymbol{\phi}_h^k\right\|_{(\Lambda_h^k)^{-1}} &= 2C_0dH\sqrt{\log 2(dW/p)}\sum_{k\in\mathcal{E}}\sum_{h=1}^{H}\left\|\boldsymbol{\phi}_h^k\right\|_{(\Lambda_h^k)^{-1}} + 2\sum_{k\in\mathcal{E}}B_{\theta,\mathcal{E}}\sqrt{d(k-\tau)}\sum_{h=1}^{H}\left\|\boldsymbol{\phi}_h^k\right\|_{(\Lambda_h^k)^{-1}} \\
&\quad + 2\sum_{k\in\mathcal{E}}B_{\boldsymbol{\mu}}H\sqrt{d(k-\tau)}\sum_{h=1}^{H}\left\|\boldsymbol{\phi}_h^k\right\|_{(\Lambda_h^k)^{-1}} \\
&\le 2C_0dH\sqrt{\log 2(dW/p)}\sum_{h=1}^{H}\sqrt{W/H}\Big(\sum_{k\in\mathcal{E}}\left\|\boldsymbol{\phi}_h^k\right\|^2_{(\Lambda_h^k)^{-1}}\Big)^{1/2} \\
&\quad + 2\sum_{h=1}^{H}\Big(\sum_{k\in\mathcal{E}}B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d(k-\tau)}\Big)^{1/2}\Big(\sum_{k\in\mathcal{E}}\left\|\boldsymbol{\phi}_h^k\right\|^2_{(\Lambda_h^k)^{-1}}\Big)^{1/2} \\
&\quad + 2\sum_{h=1}^{H}\Big(\sum_{k\in\mathcal{E}}B_{\boldsymbol{\mu},\mathcal{E}}H\sqrt{d(k-\tau)}\Big)^{1/2}\Big(\sum_{k\in\mathcal{E}}\left\|\boldsymbol{\phi}_h^k\right\|^2_{(\Lambda_h^k)^{-1}}\Big)^{1/2} \\
&\le 2C_0dH\sqrt{\log 2(dW/p)}\sum_{h=1}^{H}\sqrt{W/H}\Big(\sum_{k\in\mathcal{E}}\left\|\boldsymbol{\phi}_h^k\right\|^2_{(\Lambda_h^k)^{-1}}\Big)^{1/2} \\
&\quad + 2\sum_{h=1}^{H}B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d}\frac{W}{H}\Big(\sum_{k\in\mathcal{E}}\left\|\boldsymbol{\phi}_h^k\right\|^2_{(\Lambda_h^k)^{-1}}\Big)^{1/2} \\
&\quad + 2\sum_{h=1}^{H}B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d}W\Big(\sum_{k\in\mathcal{E}}\left\|\boldsymbol{\phi}_h^k\right\|^2_{(\Lambda_h^k)^{-1}}\Big)^{1/2}
\end{aligned} \tag{7}$$

By Lemma 11, we have

$$\Big(\sum_{k\in\mathcal{E}}\left\|\boldsymbol{\phi}_h^k\right\|^2_{(\Lambda_h^k)^{-1}}\Big)^{1/2} \le \sqrt{d\log\left(\frac{W}{H}+1\right)}. \tag{8}$$

Finally, by combining Eq. 5–8, we obtain the regret bound within the epoch $\mathcal{E}$ as:

$$\text{Dyn-Reg}(\mathcal{E}) \lesssim \tilde{O}(H^{3/2}d^{3/2}W^{1/2} + B_{\boldsymbol{\theta},\mathcal{E}}dW + B_{\boldsymbol{\mu},\mathcal{E}}dHW).$$

□

By summing over all epochs and applying a union bound, we obtain the regret bound for the whole time horizon.

Theorem 7. *If we set* $\beta = \beta_k = cdH\sqrt{\log(2dT/p)} + B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d(k-\tau)} + B_{\boldsymbol{\mu},\mathcal{E}}H\sqrt{d(k-\tau)}$*, the dynamic regret of* `LSVI-UCB-Restart` *is* $\tilde{O}(H^{3/2}d^{3/2}TW^{-1/2} + B_{\boldsymbol{\theta}}dW + B_{\boldsymbol{\mu}}dHW)$*, with probability at least* $1-p$*.*

Proof. In total there are $N = \lceil \frac{T}{W} \rceil$ epochs. For each epoch $\mathcal{E}_i$ if we set $\delta = \frac{p}{N}$, then it will incur regret $\tilde{O}(d^{3/2}H^{3/2}W^{1/2} + B_{\boldsymbol{\theta},\mathcal{E}_i}dW + B_{\boldsymbol{\mu},\mathcal{E}_i}dHW)$ with probability at least $1 - \frac{p}{N}$. By summing over all epochs and applying the union bound over them, we can obtain the regret upper bound for the whole time horizon. With probability at least $1-p$,

$$\begin{aligned}
\text{Dyn-Reg}(T) &= \sum_{\mathcal{E}_i}\text{Dyn-Reg}(\mathcal{E}_i) \\
&\lesssim \sum_{\mathcal{E}_i}\tilde{O}(d^{3/2}H^{3/2}W^{1/2} + B_{\boldsymbol{\theta},\mathcal{E}_i}dW + B_{\boldsymbol{\mu},\mathcal{E}_i}dHW) \\
&\lesssim \tilde{O}(H^{3/2}d^{3/2}TW^{-1/2} + B_{\boldsymbol{\theta}}dW + B_{\boldsymbol{\mu}}dHW).
\end{aligned}$$

□

B.2. Case 2: Unknown Local Variation

Similar to the case of known local variation, we first derive the error bound for the action-value function estimate maintained in the algorithm for any policy, which is the following technical lemma.

Lemma 6. *Under event E defined in Lemma 2, we have for any policy π, $\forall s, a, h, k \in \mathcal{S} \times \mathcal{A} \times [H] \times \mathcal{E}$,*

$$|\langle \boldsymbol{\phi}(s,a), \boldsymbol{w}_h^k \rangle - Q_{h,k}^{\pi}(s,a) - \mathbb{P}_h^k(V_{h+1}^k - V_{h+1,k}^{\pi})(s,a)| \leq \beta \left\|\boldsymbol{\phi}(s,a)\right\|_{(\Lambda_h^k)^{-1}} + B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d(k-\tau)} + B_{\boldsymbol{\mu},\mathcal{E}} H\sqrt{d(k-\tau)},$$

where $\beta = C_0 dH\sqrt{\log(2dW/p)}$ and τ is the first episode in the current epoch.

Proof. This lemma is a looser upper bound implied by Lemma 3. By Lemma 3, we have

$$\begin{aligned}
&|\langle \boldsymbol{\phi}(s,a), \boldsymbol{w}_h^k \rangle - Q_{h,k}^{\pi}(s,a) - \mathbb{P}_h^k(V_{h+1}^k - V_{h+1,k}^{\pi})(s,a)| \\
\leq & C_o dH\sqrt{\log(2dW/p)}\left\|\boldsymbol{\phi}(s,a)\right\|_{(\Lambda_h^k)^{-1}} + B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d(k-\tau)}\left\|\boldsymbol{\phi}(s,a)\right\|_{(\Lambda_h^k)^{-1}} \\
& + B_{\boldsymbol{\mu},\mathcal{E}} H\sqrt{d(k-\tau)}\left\|\boldsymbol{\phi}(s,a)\right\|_{(\Lambda_h^k)^{-1}} \\
\leq & C_o dH\sqrt{\log(2dW/p)}\left\|\boldsymbol{\phi}(s,a)\right\|_{(\Lambda_h^k)^{-1}} + B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d(k-\tau)} \\
& + B_{\boldsymbol{\mu},\mathcal{E}} H\sqrt{d(k-\tau)},
\end{aligned}$$

where the second inequality is due to $\|\boldsymbol{\phi}(s,a)\| \leq 1$ and $\lambda_{\min}(\Lambda_h^k) \geq 1$, thus $\|\boldsymbol{\phi}(s,a)\|_{(\Lambda_h^k)^{-1}} \leq 1$. □

Different from Lemma 4, when the local variation is unknown, the action-value function estimate we maintained in Algorithm 1 is no longer an optimistic upper bound of the optimal action-value function, but approximately up to some error proportional to the local variation. The rigorous statement is detailed in the following lemma.

Lemma 7. *Under event E defined in Lemma 2, if we set $\beta = cdH\sqrt{\log(2dW/p)}$, we have*

$$\begin{aligned}
&\forall (s,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times \mathcal{E}, \\
&Q_h^k(s,a) \geq Q_{h,k}^* - (H-h+1)(B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d(k-\tau)} + B_{\boldsymbol{\mu},\mathcal{E}} H\sqrt{d(k-\tau)}).
\end{aligned}$$

Proof of Lemma 7. We prove this by induction. First prove the base case when $h = H$. According to Lemma 6, we have

$$|\langle \boldsymbol{\phi}(s,a), \boldsymbol{w}_H^k \rangle - Q_{H,k}^*(s,a)| \leq \beta \left\|\boldsymbol{\phi}(s,a)\right\|_{(\Lambda_H^k)^{-1}} + B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d(k-\tau)} + B_{\boldsymbol{\mu},\mathcal{E}} H\sqrt{d(k-\tau)},$$

which implies

$$\begin{aligned}
Q_H^k(s,a) &= \min\{\langle \boldsymbol{w}_H^k, \boldsymbol{\phi}(s,a)\rangle + \beta \left\|\boldsymbol{\phi}(s,a)\right\|_{(\Lambda_H^k)^{-1}}, H\} \\
&\geq Q_{H,k}^*(s,a) - (B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d(k-\tau)} + B_{\boldsymbol{\mu},\mathcal{E}} H\sqrt{d(k-\tau)}).
\end{aligned}$$

Now suppose the statement holds true at step $h+1$, then for step h, due to Lemma 6, we have

$$\begin{aligned}
&|\langle \boldsymbol{\phi}(s,a), \boldsymbol{w}_h^k \rangle - Q_{h,k}^{\pi}(s,a) - \mathbb{P}_h^k(V_{h+1}^k - V_{h+1,k}^*)(s,a)| \\
\leq & \beta \left\|\boldsymbol{\phi}(s,a)\right\|_{(\Lambda_h^k)^{-1}} + B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d(k-\tau)} + B_{\boldsymbol{\mu},\mathcal{E}} H\sqrt{d(k-\tau)}.
\end{aligned}$$

By the induction hypothesis, we have $[\mathbb{P}_h^k(V_{h+1}^k - V_{h+1,k}^*)](s,a) \geq -(H-h+2)(B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d(k-\tau)} + B_{\boldsymbol{\mu},\mathcal{E}} H\sqrt{d(k-\tau)})$, thus

$$\begin{aligned}
Q_h^k(s,a) &= \min\{\langle \boldsymbol{w}_h^k, \boldsymbol{\phi}(s,a)\rangle + \beta \left\|\boldsymbol{\phi}(s,a)\right\|_{(\Lambda_H^k)^{-1}}, H\} \\
&\geq Q_{h,k}^*(s,a) - (H-h+1)(B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d(k-\tau)} + B_{\boldsymbol{\mu},\mathcal{E}} H\sqrt{d(k-\tau)}).
\end{aligned}$$

□

Similar to Lemma 5, next we derive the bound for the gap between the value function estimate and the ground-truth value function for the executing policy π^k, $\delta_h^k = V_h^k(s_h^k) - V_{h,k}^{\pi^k}(s_h^k)$, in a recursive manner, when the local variation is unknown.

Lemma 8. *Let $\delta_h^k = V_h^k(s_h^k) - V_{h,k}^{\pi^k}(s_h^k)$, $\zeta_{h+1}^k = \mathbb{E}[\delta_{h+1}^k | s_h^k, a_h^k] - \delta_{h+1}^k$. Under event E defined in Lemma 2, we have for all $(k,h) \in \mathcal{E} \times [H]$,*

$$\delta_h^k \leq \delta_{h+1}^k + \zeta_{h+1}^k + 2\beta \left\|\boldsymbol{\phi}_h^k\right\|_{(\Lambda_h^k)^{-1}} + B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d(k-\tau)} + B_{\boldsymbol{\mu},\mathcal{E}} H\sqrt{d(k-\tau)}.$$

Proof. By Lemma 6, for any $(s,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times \mathcal{E}$,

$$Q_h^k(s,a) - Q_h^{\pi^k}(s,a) \leq \mathbb{P}_h^k(V_{h+1}^k - V_{h+1,k}^{\pi^k})(s,a) + 2\beta \left\|\boldsymbol{\phi}(s,a)\right\|_{(\Lambda_H^k)^{-1}} + B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d(k-\tau)} + B_{\boldsymbol{\mu},\mathcal{E}} H\sqrt{d(k-\tau)}.$$

Note that $Q_h^k(s_h^k, a_h^k) = \max_a Q_h^k(s_h^k, a) = V_h^k(s_h^k)$ according to Algorithm 1, and $Q_{h,k}^{\pi^k}(s_h^k, a_h^k) = V_{h,k}^{\pi^k}(s_h^k)$ by the definition. Thus,

$$\delta_h^k \leq \delta_{h+1}^k + \zeta_{h+1}^k + 2\beta \left\|\boldsymbol{\phi}_h^k\right\|_{(\Lambda_h^k)^{-1}} + B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d(k-\tau)} + B_{\boldsymbol{\mu},\mathcal{E}} H\sqrt{d(k-\tau)}.$$

□

Now we are ready to prove Theorem 8, which is the regret upper bound within one epoch.

Theorem 8. *For each epoch $\mathcal{E}$ with epoch size W, if we set $\beta_k = cdH\sqrt{\log(2dW/p)}$, where c is an absolute constant and $p \in (0,1)$, then the dynamic regret within that epoch is $\tilde{O}(\sqrt{d^3H^3W} + B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d/H}W^{3/2} + B_{\boldsymbol{\mu},\mathcal{E}}\sqrt{dH}W^{3/2})$ with probability at least $1-p$, where $B_{\boldsymbol{\theta},\mathcal{E}}$ and $B_{\boldsymbol{\mu},\mathcal{E}}$ are the total variation within that epoch.*

Proof of Theorem 8. We denote the dynamic regret within an epoch as Dyn-Reg($\mathcal{E}$). We define $\delta_h^k = V_h^k(s_h^k) - V_{h,k}^{\pi^k}(s_h^k)$ and $\zeta_{h+1}^k = \mathbb{E}[\delta_{h+1}^k | s_h^k, a_h^k] - \delta_{h+1}^k$ as in Lemma 8. We derive the dynamic regret within a epoch $\mathcal{E}$ (the length of this epoch is W which is equivalent to $\frac{W}{H}$ episodes) conditioned on the event E defined in Lemma 2 which happens with probability at least $1 - p/2$.

$$\begin{aligned}
&\text{Dyn-Reg}(\mathcal{E}) \\
=&\sum_{k\in\mathcal{E}} \left[V_{1,k}^*(s_1^k) - V_{1,k}^{\pi^k}(s_1^k)\right] \\
\leq&\sum_{k\in\mathcal{E}} [V_1^k(s_1^k) + B_{\boldsymbol{\theta},\mathcal{E}} H\sqrt{d(k-\tau)} + B_{\boldsymbol{\mu},\mathcal{E}} H^2\sqrt{d(k-\tau)} - V_{1,k}^{\pi^k}(s_1^k)] \\
\leq&\sum_{k\in\mathcal{E}} [\delta_1^k + B_{\boldsymbol{\theta},\mathcal{E}} H\sqrt{d(k-\tau)} + B_{\boldsymbol{\mu},\mathcal{E}} H^2\sqrt{d(k-\tau)}] \\
\leq&\sum_{k\in\mathcal{E}}\sum_{h=1}^H \zeta_h^k + 2\beta\sum_{k\in\mathcal{E}}\sum_{h=1}^H \left\|\boldsymbol{\phi}_h^k\right\|_{(\Lambda_h^k)^{-1}} + 2\sum_{k\in\mathcal{E}} B_{\boldsymbol{\theta},\mathcal{E}} H\sqrt{d(k-\tau)} + 2\sum_{k\in\mathcal{E}} B_{\boldsymbol{\mu},\mathcal{E}} H^2\sqrt{d(k-\tau)} \\
\leq&\sum_{k\in\mathcal{E}}\sum_{h=1}^H \zeta_h^k + 2\beta\sum_{k\in\mathcal{E}}\sum_{h=1}^H \left\|\boldsymbol{\phi}_h^k\right\|_{(\Lambda_h^k)^{-1}} + B_{\boldsymbol{\theta},\mathcal{E}} W\sqrt{2d(W/H+1)} + B_{\boldsymbol{\mu},\mathcal{E}} W\sqrt{2d(WH+H)} \qquad (9)
\end{aligned}$$

where the first inequality is due to Lemma 7, the third inequality is due to Lemma 8, and the last inequality is due to Jensen's inequality. Now we need to bound the first two terms in the right side. Note that $\{\zeta_h^k\}$ is a martingale difference sequence satisfying $|\zeta_h^k| \leq 2H$ for all (k,h). By Azuma-Hoeffding inequality we have for any $t > 0$,

$$\mathbb{P}\left(\sum_{k\in\mathcal{E}}\sum_{h=1}^H \zeta_h^k \geq t\right) \geq \exp(-t^2/(2WH^2)).$$

Hence with probability at least $1 - p/2$, we have

$$\sum_{k\in\mathcal{E}}\sum_{h=1}^H \zeta_h^k \leq 2H\sqrt{W\log(2dW/p)}. \qquad (10)$$

For the second term, note that by Lemma 11 for any $h \in [H]$, we have

$$\sum_{k\in\mathcal{E}} (\boldsymbol{\phi}_h^k)^\top (\Lambda_h^k)^{-1} \boldsymbol{\phi}_h^k \leq 2\log\left[\frac{\det(\Lambda_h^{k+1})}{\det(\Lambda_h^1)}\right] \leq 2d\log\left(\frac{W}{H}+1\right).$$

By Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\sum_{k\in\mathcal{E}}\sum_{h=1}^{H} \left\|\boldsymbol{\phi}_h^k\right\|_{(\Lambda_h^k)^{-1}} &\leq \sum_{h=1}^{H} \sqrt{W/H}\left[\sum_{k\in\mathcal{E}} (\boldsymbol{\phi}_h^k)^\top (\Lambda_h^k)^{-1}\boldsymbol{\phi}_h^k\right]^{1/2} \\
&\leq H\sqrt{2d\frac{W}{H}\log\left(\frac{W}{H}+1\right)} \\
&\leq H\sqrt{2d\frac{W}{H}\log[2dW/p]}. \qquad (11)
\end{aligned}$$

Finally, combining (9)–(11), we have with probability at least $1-p$,

$$\begin{aligned}
\text{Dyn-Reg}(\mathcal{E}) &\leq 2H\sqrt{W\log(2dW/p)} + C_0 dH^2\sqrt{\log(2dW)/p}\sqrt{2d\frac{W}{H}\log[2dW/p]} \\
&\quad + B_{\boldsymbol{\theta},\mathcal{E}} W\sqrt{2d(W/H+1)} + B_{\boldsymbol{\mu},\mathcal{E}} W\sqrt{2d(WH+H)} \\
&\lesssim \tilde{O}(\sqrt{d^3H^3W} + B_{\boldsymbol{\theta},\mathcal{E}}\sqrt{d/H}W^{3/2} + B_{\boldsymbol{\mu},\mathcal{E}}\sqrt{dH}W^{3/2}).
\end{aligned}$$

□

Now we can derive the regret bound for the whole time horizon by summing over all epochs and applying a union bound. We restate the regret upper bound and provide its detailed proof.

Theorem 9. *If we set $\beta = cdH\sqrt{\log(2dT/p)}$, the dynamic regret of* `LSVI-UCB-Restart` *algorithm is $\tilde{O}(W^{-1/2}Td^{3/2}H^{3/2} + B_{\boldsymbol{\theta}}d^{1/2}H^{-1/2}W^{3/2} + B_{\boldsymbol{\mu}}d^{1/2}H^{1/2}W^{3/2})$, with probability at least $1-p$.*

Proof. In total there are $N = \lceil \frac{T}{W} \rceil$ epochs. For each epoch $\mathcal{E}_i$ if we set $\delta = \frac{p}{N}$, then it will incur regret $\tilde{O}(\sqrt{d^3H^3W} + B_{\boldsymbol{\theta},\mathcal{E}_i}\sqrt{d/H}W^{3/2} + B_{\boldsymbol{\mu},\mathcal{E}_i}\sqrt{dH}W^{3/2})$ with probability at least $1-\frac{p}{N}$. By summing over all epochs and applying a union bound over them, we can obtain the regret upper bound for the whole time horizon. With probability at least $1-p$,

$$\begin{aligned}
\text{Dyn-Reg}(T) = \sum_{\mathcal{E}_i} \text{Dyn-Reg}(\mathcal{E}_i) &\lesssim \sum_{\mathcal{E}_i} \tilde{O}(\sqrt{d^3H^3W} + B_{\boldsymbol{\theta},\mathcal{E}_i}\sqrt{d/H}W^{3/2} + B_{\boldsymbol{\mu},\mathcal{E}_i}\sqrt{dH}W^{3/2}) \\
&\lesssim \tilde{O}(d^{3/2}H^{3/2}TW^{-1/2} + B_{\boldsymbol{\theta}}d^{1/2}H^{-1/2}W^{3/2} + B_{\boldsymbol{\mu}}d^{1/2}H^{1/2}W^{3/2}).
\end{aligned}$$

□

C. Detailed Description of `ADA-LSVI-UCB-Restart`

Inspired by bandit-over-bandit mechanism (Cheung et al., 2019), we develop the `ADA-LSVI-UCB-Restart`. The key idea is to use `LSVI-UCB-Restart` as a subroutine (set $\beta = cdH\sqrt{\log(2dT/p)}$ since we assume total variations are unknown), and periodically update the epoch size based on the historical data under the time-varying $\mathbb{P}$ and r (potentially adversarial). More specifically, `Ada-LSVI-UCB-Restart` (Alg. 2) divides the whole time horizon into $\lceil \frac{T}{HM} \rceil$ blocks of equal length M episodes (the length of the last block can be smaller than M episodes), and specifies a set J_W from which epoch size is drawn. For each block $i \in [\lceil \frac{T}{HM} \rceil]$, `Ada-LSVI-UCB` runs a master algorithm to select the epoch size W_i and runs `LSVI-UCB-Restart` with W_i for the current block. After the end of this block, the total reward of this block is fed back to the master algorithm, and the posteriors of the parameters are updated accordingly.

For the detailed master algorithm, we select `EXP3-P` (Bubeck & Cesa-Bianchi, 2012) since it is able to deal with non-oblivious adversary. Now we present the details of `Ada-LSVI-UCB-Restart`. We set the length of each block M and the feasible set of epoch size J_W as follows:

$$M = \lceil 5T^{1/2}d^{1/2}H^{-1/2} \rceil, J_W = \{H, 2H, 4H, \ldots, MH\}.$$

The intuition of designing the feasible set for epoch size J_W is to guarantee it can well-approximate the optimal epoch size with the knowledge of total variations while on the other hand make it as small as possible, so the learner do not lose much by adaptively selecting the epoch size from J_W. This intuition is more clear when we derive the dynamic regret bound of `Ada-LSVI-UCB-Restart`. Denoting $|J_W|= \Delta$, the master algorithm `EXP3-P` treats each element of J_W as an arm and updates the probabilities of selecting each feasible epoch size based on the reward collected in the past. It begins by initializing

$$\alpha = 0.95\sqrt{\frac{\ln \Delta}{\Delta\lceil T/MH\rceil}}, \beta = \sqrt{\frac{\ln \Delta}{\Delta\lceil T/MH\rceil}}, \tag{12}$$

$$\gamma = 1.05\sqrt{\frac{\ln \Delta}{\Delta\lceil T/MH\rceil}}, q_{l,1} = 0, l \in [\Delta], \tag{13}$$

where α, β, γ are parameters used in `EXP3-P` and $q_{l,1}, l \in [\Delta]$ are the initialization of the estimated total reward of running different epoch lengths. At the beginning of the block i, the agent first sees the initial state $s_1^{(i-1)H}$, and updates the probability of selecting different epoch lengths for block i as

$$u_{l,i} = (1-\gamma)\frac{\exp(\alpha q_{l,i})}{\sum_{l\in[\Delta]}\exp(\alpha q_{l,i})} + \frac{\gamma}{\Delta}. \tag{14}$$

Then the master algorithm samples $l_i \in [\Delta]$ according to the updated distribution $\{u_{l,i}\}_{i\in[\Delta]}$; the epoch size W_i for the block i is chosen as l_i-th element in J_W, $\lfloor M^{l_i/\lfloor \ln M\rfloor}\rfloor H$. After selecting the epoch size W_i, `Ada-LSVI-UCB` runs a new copy of `LSVI-UCB-Restart` with that epoch size. By the end of each block, `Ada-LSVI-UCB-Restart` observes the total reward of the current block, denoted as $R_i(W_i, s_1^{(i-1)H})$, then the algorithm updates the estimated total reward of running different epoch sizes (divide $R_i(W_i, s_1^{(i-1)H})$ by MH to normalize):

$$q_{l,i+1} = q_{l,i} + \frac{\beta + \mathbb{1}\{l = l_i\}R_i(W_i, s_1^{(i-1)H})/MH}{u_{l,i}}. \tag{15}$$

D. Proofs in Section 5

In this section, we derive the regret bound for `Ada-LSVI-UCB-Restart` algorithm.

Proof of Theorem 5. Let $R_i(W, s_1^{(i-1)H})$ be the totol reward recieved in i-th block by running proposed `LSVI-UCB-Restart` with window size W starting at state $s_1^{(i-1)H}$, we can first decompose the regret as follows:

$$\text{Dyn-Reg}(T) = \underbrace{\sum_{k=1}^{K} V_{1,k}^*(s_k^1) - \sum_{i=1}^{\lceil T/MH\rceil} R_i(W^\dagger, s_1^{(i-1)H})}_{①} + \underbrace{\sum_{i=1}^{\lceil T/MH\rceil}(R_i(W^\dagger, s_1^{(i-1)H}) - R_i(W_i, s_1^{(i-1)H})}_{②},$$

where term ① is the regret incurred by always selecting the best epoch size for restart in the feasible set J_W, and term ② is the regret incurred by adaptively tuning epoch size by `EXP3-P`. We denote the optimal epoch size in this case as $W^* = \lceil (B_{\boldsymbol{\theta}} + B_{\boldsymbol{\mu}} + 1)^{-1/2} d^{1/2} H^{1/2} T^{1/2}\rceil H$. It is straightforward to verify that $1 \le W^* \le MH$, thus there exists a $W^\dagger \in J_W$ such that $W^\dagger \le W^* \le 2W^\dagger$, which well-approximates the optimal epoch size up to constant factors. Denote the total variation of $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$ in block i as $B_{\boldsymbol{\theta},i}$ and $B_{\boldsymbol{\mu},i}$ respectively. Now we can bound the regret. For the first term, we have

$$\begin{aligned}
① &\lesssim \sum_{i=1}^{\lceil T/MH\rceil} \tilde{O}(d^{3/2}H^{3/2}MH(W^\dagger)^{-1/2} + B_{\boldsymbol{\theta},i}d^{1/2}H^{-1/2}(W^\dagger)^{3/2} + B_{\boldsymbol{\mu},i}d^{1/2}H^{1/2}(W^\dagger)^{3/2}) \\
&\lesssim \tilde{O}(d^{3/2}H^{3/2}T(W^\dagger)^{-1/2} + B_{\boldsymbol{\theta}}d^{1/2}H^{-1/2}(W^\dagger)^{3/2} + B_{\boldsymbol{\mu}}d^{1/2}H^{1/2}(W^\dagger)^{3/2}) \\
&\lesssim \tilde{O}(d^{3/2}H^{3/2}T(W^*)^{-1/2} + B_{\boldsymbol{\theta}}d^{1/2}H^{-1/2}(W^*)^{3/2} + B_{\boldsymbol{\mu}}d^{1/2}H^{1/2}(W^*)^{3/2}) \\
&\lesssim \tilde{O}((B_{\boldsymbol{\theta}} + B_{\boldsymbol{\mu}} + 1)^{1/4}d^{5/4}H^{5/4}T^{3/4}),
\end{aligned}$$

where the first inequality is due to Theorem 4, and the third inequality is due to $W^\dagger$ differs from W^* up to constant factor. For the second term, we can directly apply the regret bound of `EXP3-P` algorithm (Bubeck & Cesa-Bianchi, 2012). In this case there are $\Delta = \ln M + 1$ arms, number of equivalent time steps is $\lceil \frac{T}{MH} \rceil$, and loss per equivalent time step is bounded within $[0, MH]$. Thus we have

$$② \lesssim \tilde{O}(MH\sqrt{\Delta T/MH}) \leq \tilde{O}(d^{1/4}H^{3/4}T^{3/4}).$$

Combining the bound of ① and ② yields the regret bound of `Ada-LSVI-UCB-Restart`,

$$\text{Dyn-Reg}(T) \lesssim \tilde{O}((B_{\boldsymbol{\theta}} + B_{\boldsymbol{\mu}} + 1)^{1/4} d^{5/4} H^{5/4} T^{3/4}).$$

□

E. Auxiliary Lemmas

In this section, we present some useful auxiliary lemmas.

Lemma 9. *For any fixed policy π, let $\{w^\pi_{h,k}\}_{h\in[H],k\in[K]}$ be the corresponding weights such that $Q^\pi_{h,k}(s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{w}^\pi_{h,k}\rangle$ for all $(s,a,h,k) \in \mathcal{S}\times\mathcal{A}\times[H]\times[K]$. Then we have*

$$\forall (k,h) \in [K]\times[H], \quad \left\|\boldsymbol{w}^\pi_{h,k}\right\| \leq 2H\sqrt{d}.$$

Proof. By the Bellman equation, we know that for any $(h,k) \in [H]\times[K]$,

$$\begin{aligned} Q^\pi_{h,k}(s,a) &= (r^k_h + \mathbb{P}^k_h V^\pi_{h+1,k})(s,a) \\ &= \langle \boldsymbol{\theta}_{h,k} + \int V^\pi_{h+1,k} d\boldsymbol{\mu}_{h,k}(s'), \boldsymbol{\phi}(s,a)\rangle \\ &= \langle \boldsymbol{w}^\pi_{h,k}, \boldsymbol{\phi}(s,a)\rangle, \end{aligned}$$

where the second equality holds due to the linear MDP assumption. Under the normalization assumption in Def. 1, we have $\|\boldsymbol{\theta}_{h,k}\| \leq \sqrt{d}$, $V^\pi_{h+1,k} \leq H$ and $\|\boldsymbol{\mu}_{h,k}(s')\| \leq \sqrt{d}$. Thus,

$$\boldsymbol{w}^\pi_{h,k} \leq \sqrt{d} + H\sqrt{d} \leq 2H\sqrt{d}.$$

□

Lemma 10. *Let $\Lambda_t = I + \sum_{i=1}^t \boldsymbol{\phi}_i^\top \boldsymbol{\phi}_i$, where $\boldsymbol{\phi}_t \in \mathbb{R}^d$, then*

$$\sum_{i=1}^t \boldsymbol{\phi}_i^\top (\Lambda_t)^{-1} \boldsymbol{\phi}_i \leq d.$$

Proof. We have $\sum_{i=1}^t \boldsymbol{\phi}_i^\top (\Lambda_t)^{-1}\boldsymbol{\phi}_i = \sum_{i=1}^t \mathrm{Tr}(\boldsymbol{\phi}_i^\top (\Lambda_t)^{-1}\boldsymbol{\phi}_i) = \mathrm{Tr}((\Lambda_t)^{-1}\sum_{i=1}^t \boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top)$. After apply eigenvalue decomposition, we have $\sum_{i=1}^t \boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top = \mathbf{U}\mathrm{diag}(\lambda_1,\ldots,\lambda_d)$ and $\Lambda_t = \mathbf{U}\mathrm{diag}(\lambda_1+1,\ldots,\lambda_d+1)$. Thus $\sum_{i=1}^t \boldsymbol{\phi}_i^\top (\Lambda_t)^{-1}\boldsymbol{\phi}_i = \sum_{i=1}^d \frac{\lambda_i}{\lambda_i} \leq d$.

□

Lemma 11. *(Abbasi-Yadkori et al., 2011) Let $\{\boldsymbol{\phi}_t\}_{t\geq 0}$ be a bounded sequence in $\mathbb{R}^d$ satisfying $\sup_{t\geq 0}\|\boldsymbol{\phi}_t\| \leq 1$. Let $\Lambda_0 \in \mathbb{R}^{d\times d}$ be a positive definite matrix. For any $t \geq 0$, we define $\Lambda_t = \Lambda_0 + \sum_{j=1}^t \boldsymbol{\phi}_j^\top \boldsymbol{\phi}_j$. Then if the smallest eigenvalue of Λ_0 satisfies $\lambda_{\min}(\Lambda_0) \geq 1$, we have*

$$\log\left[\frac{\det(\Lambda_t)}{\det(\Lambda_0)}\right] \leq \sum_{j=1}^t \boldsymbol{\phi}_j^\top \Lambda_{j-1}^{-1}\boldsymbol{\phi}_j \leq 2\log\left[\frac{\det(\Lambda_t)}{\det(\Lambda_0)}\right].$$