University of California
Santa Barbara

On the Data Complexity of Problem-Adaptive Offline Reinforcement Learning

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Computer Science

by

Ming Yin

Committee in charge:

Professor Yu-Xiang Wang, Chair
Professor William Yang Wang
Professor S. Rao Jammalamadaka

December 2023

The Dissertation of Ming Yin is approved.

Professor William Yang Wang

Professor S. Rao Jammalamadaka

Professor Yu-Xiang Wang, Committee Chair

September 2023

On the Data Complexity of Problem-Adaptive Offline Reinforcement Learning

Ming Yin

Acknowledgements

I would like to offer my humble and grateful acknowledgment to all the remarkable individuals who have played pivotal roles in my Ph.D. journey. Their unwavering support and guidance have been instrumental in shaping my academic and personal growth.

First and foremost, my deepest gratitude goes to my advisor, Professor Yu-Xiang Wang, for guiding me in doing research from scratch. You supported me at my hardest time, and the trajectory of my Ph.D. career changed completely since then. It is a privilege to work with you, and there are countless moments when I was inspired by your passion, knowledge, and optimism. Every time I introduce you to someone else, I use the words "Machine Learning Encyclopaedia". Your optimism, and positive attitude toward life, have been a guiding light whenever I encountered difficulties. I still remember when we worked together diligently on my very first AISTATS submission in 2019 and finished it by midnight. I drove to In-and-Out to get dinner around 1AM, full of happiness in my heart. This was the beginning of my publication experiences and also one of my happiest moments during my Ph.D. study. Your profound impact on me over the past few years is hard to describe using words and it will continue to influence my life in immeasurable ways.

Next, I want to thank my committee member, Distinguished Professor S. Rao Jammalamadaka, for your tremendous support throughout my Ph.D. journey. I met you in the initial phase of my Ph.D., and the knowledge I gained from your PSTAT 207 course sequence has been the building blocks for my later research in Statistical Machine Learning. Your captivating explanation of the Cramer-Rao lower bound, attributed to your advisor Calyampudi Radhakrishna Rao, left an indelible impression on me. Such a unique connection lasted and eventually grew into the central topic of my first publication. Your counsel, wisdom, and unwavering guidance have been a consistent source of encouragement during my most challenging moments.

I also want to thank Professor William Wang serve on my dissertation committee. By com-

municating with you, I get to know the frontier of AI from the Natural Language Processing perspective, prompting me to reflect on my research in sequential decision-making and cultivate new ideas. Your insightful feedback has not only enhanced my research but also fortified my long-term career aspirations.

Furthermore, I want to give special thanks to the brilliant researchers I have had the fortune to work with throughout the years, Yu-Xiang Wang, Mengdi Wang, Yu Bai, Yaqi Duan, Dan Qiao, Thanh Nguyen-Tang, Sunil Gupta, Svetha Venkatesh, Raman Arora, Jiachen Li, William Yang Wang, Kaiqi Zhang, Wenjing Chen, Chong Liu, Ming Min, Wenhu Chen, Max Ku, Elaine Wan, Xueguang Ma, Jianyu Xu, Tony Xia, Xinyi Wang, Pan Lu, Songtao Feng, Ruiquan Huang, Jing Yang, Edwin Zhang, Yingbin Liang, Sunil Madhow and Qinxun Bai. My achievements would have remained out of reach without your collaboration and support.

I was fortunate to spend my graduate student life in the Department of Computer Science at UCSB. I would like to thank all my friends, including Andrea Zanette, Chong Liu, Jianyu Xu, Yuqing Zhu, Peng Zhao, Dheeraj Baby, Esha Singh, Rachel Redberg, Dan Qiao, Erchi Wang, Kaiqi Zhang, Xuandong Zhao, Jiachen Li, Shiyang Li, Xinlu Zhang, Kan Wu, Chengsheng Shen, Wenhu Chen, Minshuo Chen, Yuanzhe Xu, Masatoshi Uehara, Yaodong Yu, Zhiyu Chen, Xuezhou Zhang, Songtao Feng, Yu Bai, Tongzheng Ren, Hong Wang, Hengyu Bu, Wenhan Xiong, Fuheng Zhao, Nikki Kuang, Lucy Liu and Min Woo Park. The journey would not have been the same without your presence.

Finally, I want to express my deepest gratitude to my parents, for their unconditional love and support throughout my life. I will always make you proud!

Curriculum Vitæ

Ming Yin

Education

2023	Ph.D. in Computer Science, University of California, Santa Barbara.
2023	Ph.D. in Statistics and Applied Probability, University of California, Santa Barbara.
2016	B.S. in Applied Mathematics, University of Science and Technology of China.

Publications

NeurIPS 2023	Posterior Sampling with Delayed Feedback for Reinforcement Learning with Linear Function Approximation, Nikki Lijing Kuang*, Ming Yin*, Mengdi Wang, Yu-Xiang Wang, Yi-An Ma. *In Proceedings of the 37th Conference on Neural Information Processing Systems, New Orleans, LA, USA*.
EMNLP 2023	TheoremQA: A Theorem-driven Question Answering dataset, Wenhu Chen, Ming Yin, Max Ku, Elaine Wan, Xueguang Ma, Jianyu Xu, Tony Xia, Xinyi Wang, Pan Lu. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Singapore.*
ICLR 2023	Offline Reinforcement Learning with Differentiable Function Approximation is Provably Efficient, Ming Yin, Mengdi Wang, Yu-Xiang Wang. *In Proceedings of the 10th International Conference on Learning Representations, Kigali Rwanda, Africa.*
UAI 2023	No-Regret Linear Bandits beyond Realizability, Chong Liu, Ming Yin, Yu-Xiang Wang. *In Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence, Pittsburgh,* PA, USA.
ICML 2023	Non-stationary Reinforcement Learning under General Function Approximation, Songtao Feng, Ming Yin, Ruiquan Huang, Yu-Xiang Wang, Jing Yang, Yingbin Liang, *In Proceedings of the 40th International Conference on Machine Learning, Honolulu*, HI, USA.
ICML 2023	Offline Reinforcement Learning with Closed-Form Policy Improvement Operators, Jiachen Li, Edwin Zhang, Ming Yin, Qinxun Bai, Yu-Xiang Wang, William Yang Wang. *In Proceedings of the 40th International Conference on Machine Learning, Honolulu*, HI, USA.
ICML WS 2023	Why Quantization Improves Generalization: NTK of Binary Weight Neural Networks, Kaiqi Zhang, Ming Yin, Yu-Xiang Wang, *In ICML workshop in Neural Compression, Honolulu*, HI, USA.

AAAI 2023	On Instance-Dependent Bounds for Offline Reinforcement Learning with Linear Function Approximation, Thanh Nguyen-Tang, Ming Yin, Sunil Gupta, Svetha Venkatesh, Raman Arora. *In Proceedings of Association for the Advancement of Artificial Intelligence, Washtington, DC, USA*.
NeurIPS WS 2022	Offline Policy Evaluation for Reinforcement Learning with Adaptively Collected Data, Sunil Madhow, Dan Qiao, Ming Yin, Yu-Xiang Wang. *In NeurIPS workshop in Offline RL, New Orleans, LA, USA*.
UAI 2022	Offline Stochastic Shortest Path: Learning, Evaluation and Towards Optimality, Ming Yin*, Wenjing Chen*, Mengdi Wang, Yu-Xiang Wang. *In Proceedings of Association for the Advancement of Artificial Intelligence, Washtington, DC, USA*.
ICML 2022	Sample-Efficient Reinforcement Learning with loglog(T) Switching Cost, Dan Qiao, Ming Yin, Ming Min, Yu-Xiang Wang. *In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA*.
ICLR 2022	Near-optimal Offline Reinforcement Learning with Linear Representation: Leveraging Variance Information with Pessimism, Ming Yin, Yaqi Duan, Mengdi Wang, Yu-Xiang Wang. *In Proceedings of the 10th International Conference on Learning Representations, Virtual.*
NeurIPS 2021	Towards Instance-Optimal Offline Reinforcement Learning with Pessimism, Ming Yin, Yu-Xiang Wang. *In Proceedings of the 35th Conference on Neural Information Processing Systems, Vancouver, Canada.*
NeurIPS 2021	Optimal Uniform OPE and Model-based Offline Reinforcement Learning in Time Homogeneous, Reward-Free and Task-Agnostic Settings, Ming Yin, Yu-Xiang Wang. *In Proceedings of the 35th Conference on Neural Information Processing Systems, Vancouver, Canada.*
NeurIPS 2021	Near-Optimal Offline Reinforcement Learning via Double Variance Reduction, Ming Yin, Yu Bai, Yu-Xiang Wang. *In Proceedings of the 35th Conference on Neural Information Processing Systems, Vancouver, Canada.*
AISTATS 2021	Near-Optimal Provable Uniform Convergence in Offline Policy Evaluation for Reinforcement Learning, Ming Yin, Yu Bai, Yu-Xiang Wang **(Oral presentation)** *In Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, Virtual.*
AISTATS 2020	Asymptotically Efficient Off-Policy Evaluation for Tabular Reinforcement Learning, Ming Yin, Yu-Xiang Wang. *In Proceedings of the 23th International Conference on Artificial Intelligence and Statistics, Sicily, Italy.*

Academic Services

Area Chair	**[NeurIPS]** Conference on Neural Information Processing Systems, 2023
Conf. Reviewer	**[ICML]** International Conference on Machine Learning, 2020,2021,2022,2023 **[AISTATS]** International Conference on Artificial Intelligence and Statistics, 2021,2022,2023, 2024 **[NeurIPS]** Conference on Neural Information Processing Systems, 2021,2022 **[ICLR]** International Conference on Learning Representations, 2022,2023,2024 **[AAAI]** AAAI Conference on Artificial Intelligence, 2023, 2024 **[UAI]** Conference on Uncertainty in Artificial Intelligence, 2023 **[EMNLP]** Conference on Empirical Methods in Natural Language Processing, 2023 **[COLT]** Conference on Learning Theory, 2024
Journal Reviewer	**[AOS]** Annals of Statistics **[JASA]** Journal of the American Statistical Association **[JMLR]** Journal of Machine Learning Research **[MACH]** Machine Learning, Journal by Springer **[TMLR]** Transactions on Machine Learning Research **[JDS]** ACM/IMS Journal of Data Science (3-year appointment)

Abstract

On the Data Complexity of Problem-Adaptive Offline Reinforcement Learning

by

Ming Yin

Offline reinforcement learning, a field dedicated to optimizing sequential decision-making strategies using historical data, has found widespread application in real-world scenarios. Recent years have witnessed a surge in research focusing on establishing the statistical foundations for offline reinforcement learning, with many studies achieving near-optimal worst-case performance bounds. However, empirical results often outperform these non-adaptive bounds significantly. A comprehensive understanding of which decision processes and behavior policies are inherently more amenable or challenging for offline RL remains elusive. To address this critical challenge, the first part of this thesis delves into instance-dependent offline learning within tabular Markov Decision Processes. We introduce the Adaptive Pessimistic Value Iteration algorithm, which achieves an instance-dependent guarantee and is also near optimal. This result subsumes a wide spectrum of previous worst-case optimal results, leading to the first instance-dependent guarantee that characterizes the hardness for offline RL.

In the Second chapter of the thesis, we extend our study for tabular reinforcement learning to the function approximation regime. Specifically, within the context of linear function approximation, we present the variance-aware pessimistic value iteration (VAPVI) algorithm, which quantifies the uncertainty of training examples through conditional variance reweighing. VAPVI enhances offline learning bounds compared to the best-known existing results. Crucially, our learning bounds are expressed in terms of system-related quantities, offering natural instance-dependent characterizations that previous studies lacked.

Furthermore, State-Of-The-Art algorithms usually leverage powerful function approxima-

tors (e.g. neural networks) to alleviate the sample complexity hurdle for better empirical performances. In the third chapter, we broaden our focus to function approximation without imposing specific structural constraints on the function class, except for differentiability. This class naturally encompasses a wide range of models with nonlinear and nonconvex structures. Importantly, we demonstrate the provable efficiency of offline RL with differentiable function approximation through an analysis of the pessimistic fitted Q-learning (PFQL) algorithm. Our findings provide the theoretical underpinnings for understanding various practical heuristics relying on Fitted Q-Iteration-style design.

We conclude the thesis by summarizing our work and mentioning other exciting research projects.

On the distinctions between [1] and this thesis. [1] is the thesis submitted in partial satisfaction of the requirements for the Ph.D. degree in Statistics and Applied Probability, and this thesis is submitted in partial satisfaction of the requirements for the Ph.D. degree in Computer Science. Thesis [1] studies tabular offline policy evaluation (OPE) problem, where the target policy is fixed and the environment has finite states and actions. [1] mostly contains the materials from [2, 3, 4] This thesis studies the offline policy learning problem, where the goal is to find a reward-maximizing policy. The environments considered in this thesis include tabular MDPs, linear function approximation, and the general parametric models. This thesis mostly contains the materials from [5, 6, 7]

List of Frequently Used Notations

MDP	Markov Decision Processes
μ	Logging/Behavior policy
$\mathcal{M}_1$	$\max\{2\lambda, 128\log(2d/\delta), 128H^4\log(2d/\delta)/\kappa^2\}$
$\mathcal{M}_2$	$\max\{\frac{\lambda^2}{\kappa\log((\lambda+K)H/\lambda\delta)}, 96^2H^{12}d\log((\lambda+K)H/\lambda\delta)/\kappa^5\}$
$\mathcal{M}_3$	$\max\left\{512H^4/\kappa^2\log\left(\frac{2d}{\delta}\right), 4\lambda H^2/\kappa\right\}$
$\mathcal{M}_4$	$12\sqrt{H^4d\log((\lambda+K)H/\lambda\delta)/\kappa}$
δ	Failure probability
ξ	$\sup_{V\in[0,H],\, s'\sim P_h(s,a),\, h\in[H]}\left\lvert\frac{r_h+V(s')-(\mathcal{T}_hV)(s,a)}{\sigma_V(s,a)}\right\rvert$
$C_{H,d,\kappa,K}$	$36\sqrt{\frac{H^4d^3}{\kappa}\log\left(\frac{(\lambda+K)2KdH^2}{\lambda\delta}\right)}+12\lambda\frac{H^2\sqrt{d}}{\kappa}$
$\Sigma_h^p(\theta)$	$\mathbb{E}_{\mu,h}\left[\nabla f(\theta,\phi(s,a))\cdot\nabla f(\theta,\phi(s,a))^\top\right]$
κ	$\min_{h,\theta}\lambda_{\min}(\Sigma_h^p(\theta))$
$\sigma_V^2(s,a)$	$\max\{1, \mathrm{Var}_{P_h}(V)(s,a)\}$ for any V
K_0	$\max\left\{512\frac{\kappa_1^4}{\kappa^2}\left(\log(\frac{2Hd}{\delta})+d\log(1+\frac{4\kappa_1^3\kappa_2C_\Theta K^3}{\lambda^2})\right), \frac{4\lambda}{\kappa}\right\}$
ζ	$2\max_{s'\sim P(\cdot\lvert s,a),h\in[H]}\frac{(\mathcal{P}_hV^\star_{h+1})(s,a)-r-V^\star_{h+1}(s')}{\sigma_h^\star(s,a)}$
$C_{\mathrm{hot}}=\bar{C}_{\mathrm{hot}}$	$\frac{\kappa_1H}{\sqrt{\kappa}}+\frac{\kappa_1^2H^3d^2}{\kappa}+\sqrt{\frac{d^3H^4\kappa_2^2\kappa_1^2}{\kappa^3}}+\kappa_2\max(\frac{\kappa_1}{\kappa},\frac{1}{\sqrt{\kappa}})d^2H^3+\frac{d^2H^4\kappa_3+\lambda\kappa_1C_\Theta}{\kappa}+\frac{H^3\kappa_2d^2}{\kappa}$
$C'_{\mathrm{hot}}=\bar{C}'_{\mathrm{hot}}$	$C_{\mathrm{hot}}+\frac{\kappa_1\kappa_2H^4d^2}{\kappa^{3/2}}$

Contents

Chapter 1

Introduction

Science and technology should bring better lives for the human race and provide new solutions to some of the biggest challenges we face today — clean energy, climate change, social injustice, changing workforce and more. In particular, sequential decision making could play an important role in these problems as it models the long term impacts of policies. Reinforcement learning, a data-driven framework for the long horizon planing, has been one of the fastest-growing research areas. In the past decade, RL-based applications have led to a few breakthroughs in artificial intelligence, such as defeating human world champions in the game of Go [8] and StarCraft II [9]. The applicability of RL to real-life problems, however, remains limited. The crux of the problem is that most existing RL methods require an environment for the agent to interact with, but in real-life applications, it is rarely feasible to have access to such an environment — deploying an algorithm that learns by trial-and-errors may be costly or have serious legal, ethical and safety issues. In this thesis, we consider offline reinforcement learning, where the goal is to develop general and efficient reinforcement learning algorithms that can learn from offline/historical data with low (optimal) sample and computation complexity.

In *offline Reinforcement Learning* (offline RL [10, 11]), the objective is to identify a reward-maximizing policy in an unknown environment *Markov Decision Process* (MDP) using the

historical data. Unlike an online RL, where the agent can keep interacting with the environment and gain new feedback by exploring unvisited state-action space, offline RL is needed when such online interplays are expensive or even unethical. Since it has no access to interact with the MDP model (which causes distributional mismatches), most of the literature that studies the sample complexity / provable efficiency of offline RL (*e.g.* [12, 13, 14, 15, 3, 16, 17, 18, 19]) relies on making different data-coverage assumptions for making the problem learnable, and provide near-optimal worst-case performance bounds that depend on their data-coverage coefficients. Those results are valuable as they do not depend on the structure of the particular problem, therefore, remain valid even for pathological MDPs. But is this good enough?

In practice, the empirical performances of offline reinforcement learning (*e.g.* [20, 21, 22, 23]) are often far better than what those non-adaptive / problem-independent bounds would indicate. Although empirical evidence can help explain why we may observe better or worse performances on different MDPs, a systematic understanding of what types of decision processes and what kinds of behavior policies are inherently easier or more challenging for offline RL is lacking. Besides, despite the fact that a non-adaptive bound can learn even the pathological examples within the assumption family, there is no guarantee for the instances outside the family. However, practical offline reinforcement learning problems are usually beyond the scope of certain data-coverage assumptions, which limits the applicability of those results.

In this thesis, we derive the first line of instance-dependent bounds for offline reinforcement learning that adapt to the individual instances with weak assumptions. The settings we considered include (bt not limited to) tabular representation, linear function approximation, and parametric differentiable models. Our contribution can be summarized as follows:

- In Chapter 2, we consider the policy learning problem for finite horizon, non-stationary, episodic MDPs with finite states and actions. We propose and analyze the Adaptive Pessimistic Value Iteration algorithm, and derive the suboptimality upper bound that nearly

matches

$$O\left(\sum_{h=1}^{H}\sum_{s_h,a_h} d_h^{\pi^\star}(s_h,a_h)\sqrt{\frac{\mathrm{Var}_{P_{s_h,a_h}}(V_{h+1}^\star + r_h)}{d_h^\mu(s_h,a_h)}}\sqrt{\frac{1}{n}}\right). \tag{1.1}$$

Here $\pi^\star$ is a optimal policy, μ is the behavior policy and d_h^μ is the marginal state-action probability. We name (1.1) the *intrinsic offline reinforcement learning bound* since it indicates all the existing optimal results: minimax rate under uniform data-coverage assumption, horizon-free setting, single policy concentrability, and the tight problem-dependent results. We also study how learning would degrade in the *assumption-free* regime (where we make no assumption on μ) and obtain the assumption-free intrinsic bound.

- In Chapter 3, we present the first near-optimal result for offline reinforcement learning with linear function approximation. We devise the *variance-aware pessimistic value iteration* (VAPVI), which adopts the conditional variance information of the value function for time-inhomogeneous episodic linear *Markov decision processes* (MDPs). VAPVI provides improved guarantees over the best-known existing results. Furthermore, our results are expressed in terms of system quantities, which provide natural instance-dependent characterizations.

- In Chapter 4, we consider offline reinforcement learning with differentiable function approximation. This function class naturally incorporates a wide range of models with nonlinear/nonconvex structures. We propose and analyze the *pessimistic fitted Q-learning* (PFQL) algorithm, which provides the theoretical basis for understanding a variety of practical heuristics that rely on Fitted Q-Iteration style design. In addition, we further improve our guarantee with a tighter instance-dependent characterization.

- In the final Chapter 5 of the thesis, we provide a summary of my other exciting contributions during my Ph.D. career, encompassing topics such as offline posterior sampling

mechanisms, low-adaptive RL, adversarial RL, zero-sum games, Neural Tangent Kernels, and Math Question Answering.

Potential Contributions to the Broader Scientific Regiems. This thesis primarily explores the statistical underpinnings of offline reinforcement learning. However, its findings, including statistical guarantees, algorithms, and designs, have broad implications across numerous scientific fields, underlining offline RL's interdisciplinary impact and its significance in broader scientific discourse.

- Automated driving: offline RL offers a promising avenue to enhance the intelligence and safety of autonomous vehicles. By leveraging vast amounts of driving data, including diverse situations and rare events, offline RL enables the development of more robust and generalizable driving policies. Concretely, our uncertainty reweighting design principle (Chapter 3) has been shown to be effective for improving the empirical performance [24, 25].

- Supply Chain and Logistics: In this domain, offline RL facilitates the formulation of effective decision-making strategies without the expenses and risks associated with real-time trials. For example, it aids in refining inventory management through accurate demand forecasts and optimal stocking methods. In logistics, offline RL enhances route planning and fleet management by analyzing historical delivery data to reduce costs and improve delivery efficiency. Our research conceptualizes logistic challenges within the framework of goal-conditioned offline RL [26], addressing policy evaluation and learning tasks. Adapting these findings to functional approximation settings could significantly benefit real-world applications.

- Healthcare: Offline RL offers a distinct advantage in healthcare by utilizing pre-existing data sets, circumventing the ethical and logistical issues inherent in live experiments in

sensitive medical environments. This method allows for counterfactual analysis and addressing "what if" scenarios. Our marginalized importance sampling method [2] achieves the statistical optimality and has potential applications in personalized medicine [27], patient trajectory prediction, and treatment optimization [28].

Each of these applications demonstrates the far-reaching influence of offline RL, not just within artificial intelligence, but across diverse scientific disciplines, showcasing its capacity to address complex, real-world challenges.

Chapter 2

On the Instance-dependent Tabular Offline Reinforcement Learning

In this chapter, we derive the instance-dependent guarantees for tabular offline reinforcement learning. Concretely, we design the conditional variance based uncertainty, combined with delicate analysis to achieve the near-optimal instance-dependent characterization which we named *intrinsic offline reinforcement learning bound*. Due to its generic form, we believe the intrinsic bound could help illuminate what makes a specific problem hard and reveal the fundamental challenges in offline RL.

2.1 Preliminaries for Offline Reinforcement Learning

Episodic time-inhomogeneous reinforcement learning. A finite-horizon *Markov Decision Process* (MDP) is denoted by a tuple $M = (\mathcal{S}, \mathcal{A}, P, r, H, d_1)$ [29], where $\mathcal{S}$ is the finite state space and $\mathcal{A}$ is the finite action space with $S := |\mathcal{S}| < \infty, A := |\mathcal{A}| < \infty$. A non-stationary transition kernel $P_h : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$ maps each state action(s_h, a_h) to a probability distribution $P_h(\cdot|s_h, a_h)$ and P_h can be different across the time. Besides,

$r : \mathcal{S} \times A \mapsto \mathbb{R}$ is the expected instantaneous reward function satisfying $0 \leq r \leq 1$. d_1 is the initial state distribution. H is the horizon. A policy $\pi = (\pi_1, \ldots, \pi_H)$ assigns each state $s_h \in \mathcal{S}$ a probability distribution over actions according to the map $s_h \mapsto \pi_h(\cdot|s_h)\ \forall h \in [H]$. An MDP together with a policy π induce a random trajectory $s_1, a_1, r_1, \ldots, s_H, a_H, r_H, s_{H+1}$ with $s_1 \sim d_1, a_h \sim \pi(\cdot|s_h), s_{h+1} \sim P_h(\cdot|s_h, a_), \forall h \in [H]$ and r_h is a random realization given the observed s_h, a_h.

Q-values, Bellman (optimality) equations. The value function $V_h^\pi(\cdot) \in \mathbb{R}^S$ and Q-value function $Q_h^\pi(\cdot,\cdot) \in \mathbb{R}^{S\times A}$ for any policy π is defined as: $V_h^\pi(s) = \mathbb{E}_\pi[\sum_{t=h}^H r_t | s_h = s]$, $Q_h^\pi(s,a) = \mathbb{E}_\pi[\sum_{t=h}^H r_t | s_h, a_h = s, a]$, $\forall s, a \in \mathcal{S}, \mathcal{A}, h \in [H]$. The performance is defined as $v^\pi := \mathbb{E}_{d_1}\left[V_1^\pi\right] = \mathbb{E}_{\pi, d_1}\left[\sum_{t=1}^H r_t\right]$, where we denote V_h^π, Q_h^π as column vectors and $P_h \in \mathbb{R}^{SA\times S}$ the transition matrix, then the vector form Bellman (optimality) equations follow $\forall h \in [H]$: $Q_h^\pi = r_h + P_h V_{h+1}^\pi$, $V_h^\pi = \mathbb{E}_{a\sim\pi_h}[Q_h^\pi]$, $Q_h^\star = r_h + P_h V_{h+1}^\star$, $V_h^\star = \max_a Q_h^\star(\cdot, a)$.In addition, we denote the per-step marginal state-action occupancy $d_h^\pi(s,a)$ as: $d_h^\pi(s,a) := \mathbb{P}[s_h = s | s_1 \sim d_1, \pi] \cdot \pi_h(a|s)$, which is the marginal state-action probability at time h.

Offline setting and the goal. The offline RL requires the agent to find a policy π such that the performance v^π is maximized, given only the episodic data $\mathcal{D} = \left\{\left(s_h^\tau, a_h^\tau, r_h^\tau, s_{h+1}^\tau\right)\right\}_{\tau\in[n]}^{h\in[H]}$ rolled out from some behavior policy μ. The offline nature requires we cannot change μ and in particular we do not assume the functional knowledge of μ. That is to say, given the batch data $\mathcal{D}$ and a targeted accuracy $\epsilon > 0$, the offline RL seeks to find a policy π_{alg} such that $v^\star - v^{\pi_{\text{alg}}} \leq \epsilon$.

2.1.1 Assumptions in offline RL

We revise several types of assumptions proposed by existing studies that can yield provably efficient results. Recall $d_h^\mu(s_h, a_h)$ is the marginal state-action probability with respect to μ.

Assumption 2.1.1 (Uniform data coverage [3])**.** *it holds that* $d_m := \min_{h,s_h,a_h} d_h^\mu(s_h, a_h) > 0$. *Here the infimum is over all the states satisfying there exists certain policy so that this state can*

be reached by the current MDP with this policy.

This is the strongest assumption in offline RL as it requires μ to explore each state-action pairs with positive probability. Under 2.1.1, it mostly holds $1/d_m \geq SA$. This reveals offline learning is generically harder than *the generative model setting* [30] in the statistical sense. On the other hand, this is required for the *uniform OPE* task in [3] as it seeks to simultaneously evaluate all the policies within the policy class and it is in general a harder task than offline learning itself.

Assumption 2.1.2 (Uniform concentrability [31, 13]). $C_\mu := \sup_{\pi,h} ||d_h^\pi(\cdot,\cdot)/d_h^\mu(\cdot,\cdot)||_\infty < \infty$.

This is a classical offline RL condition that is commonly assumed in the function approximation scheme (*e.g.* Fitted Q-Iteration). Qualitatively, this is a uniform data-coverage assumption that is similar to Assumption 2.1.1, but quantitatively, the coefficient C_μ can be smaller than $1/d_m$ due the d_h^π term in the numerator.

Assumption 2.1.3 ([32]). *There exists one optimal policy $\pi^\star$, s.t. $\forall s_h, a_h \in \mathcal{S}, \mathcal{A}$, $d_h^\mu(s_h, a_h) > 0$ if $d_h^{\pi^\star}(s_h, a_h) > 0$. We further denote the trackable set as $\mathcal{C}_h := \{(s_h, a_h) : d_h^\mu(s_h, a_h) > 0\}$.*

Assumption 3.3.3 is (arguably) the weakest assumption needed for accurately learning the optimal value $v^\star$ and we will use 3.3.3 for most parts of this thesis. It only requires μ to trace the state-action space of one optimal policy and can be agnostic at other locations. [18, 19] considers this assumption and provide analysis is based on the single concentrability coefficient $C^\star := \max_{s,a} d^{\pi^\star}(s,a)/d^\mu(s,a)$. The dependence on $C^\star$ makes their result less adaptive since there can be lots of locations that have the ratio $d^{\pi^\star}(s,a)/d^\mu(s,a)$ much smaller than $C^\star$.

In the later sections, we will also consider the situation when 3.3.3 might not be true, and this corresponds to the assumption-free regime.

2.1.2 Related prior work

Finite sample analysis for offline reinforcement learning can be traced back to [31, 33, 34] for the *infinite horizon discounted setting* via Fitted Q-Iteration (FQI) type function approximation algorithms. [13, 12, 15, 14] follow this line of research and derive the information-theoretical bounds. Recently, [15] considers the offline RL with only the realizability assumption, [35, 36] considers the offline RL without sufficient coverage and [37, 38] uses the model-based approach for addressing offline RL. Under those weak coverage assumption, their finite sample analysis are suboptimal (*e.g.* in terms of the effective horizon $(1-\gamma)^{-1}$). Recently, [3, 16, 17] study the finite horizon case. In the linear MDP case, [39] studies the pessimistic algorithm for offline policy learning under only the compliance assumption, and, concurrently, [40] proposes the general pessimistic function approximation framework with instantiation in linear MDP and [41] shows actor-critic style algorithm is near-optimal for linear Bellman complete model. In addition, [42, 43] prove some exponential lower bounds under their linear function approximation assumptions.

Among them, there are a few works that achieve the sample optimality under their respective assumptions. Under the uniform data coverage (minimal state-action probability $d_m > 0$), [3] first proves the optimal $\tilde{O}(H^3/d_m\epsilon^2)$ complexity in the time-inhomogeneous MDP. Recently, [16] designs the offline variance reduction algorithm to achieve the optimal $\tilde{O}(H^2/d_m\epsilon^2)$ rate for the time-homogeneous case. Under the setting where the total cumulative reward is bounded by 1, [17] obtains the horizon-free result with $\tilde{O}(1/d_m)$. More recently, [18] considers the single concentrability coefficient $C^\star := \max_{s,a} d^{\pi^\star}(s,a)/d^\mu(s,a)$ and derives the upper bound $\tilde{O}[(1-\gamma)^{-5}SC^\star/\epsilon^2]$ in the infinite horizon setting which is recently improved by the concurrent work [19]. While those worst-case guarantees are desirable, none of them can explain the hardness of the individual problems.[1]

[1] We do mention [41] is near-optimal in their setting, but it is unclear whether it remains optimal in the standard setting where $Q^\pi \in [0, H]$, since there is an additional H factor by rescaling.

2.2 Instrinsic Offline Reinforcement Learning Bound and Adaptive Pessimistic Value Iteration

As a step towards the optimal and strong adaptive offline RL bound, we analyze *the vanilla pessimistic value iteration* (VPVI), a tabular counterpart of *pessimistic value iteration* (PEVI initiated in [39]), to understand what is missing for achieving the fully adaptivity. In particular, VPVI relies on the model-based construction.

Model-based Components. Given data $\mathcal{D} = \left\{ \left(s_h^\tau, a_h^\tau, r_h^\tau, s_{h+1}^\tau \right) \right\}_{\tau\in[n]}^{h\in[H]}$, we denote $n_{s_h,a_h} := \sum_{\tau=1}^n \mathbf{1}[s_h^\tau, a_h^\tau = s_h, a_h]$ be the total counts that visit (s_h, a_h) pair at time h, then we use the offline plug-in estimator to construct the estimators for P_h and r_h as:

$$\widehat{P}_h(s'|s_h,a_h) = \frac{\sum_{\tau=1}^n \mathbf{1}[(s_{h+1}^\tau, a_h^\tau, s_h^\tau) = (s', s_h, a_h)]}{n_{s_h,a_h}},\ \widehat{r}_h(s_h,a_h) = \frac{\sum_{\tau=1}^n \mathbf{1}[(a_h^\tau, s_h^\tau) = (s_h, a_h)] \cdot r_h^\tau}{n_{s_h,a_h}}, \tag{2.1}$$

if $n_{s_h,a_h} > 0$ and $\widehat{P}_h(s'|s_h,a_h) = 1/S, \widehat{r}_h(s_h,a_h) = 0$ if $n_{s_h,a_h} = 0$. In particular, we use the word "vanilla" as it directly mirrors [39] with a pessimistic penalty of order $O(H/\sqrt{n_{s_h,a_h}})$.[2] With $\widehat{P}_h, \widehat{r}_h$ in Algorithm 4 (which we defer to Appendix), VPVI guarantees the following:

Theorem 2.2.1. *Under the Assumption 3.3.3, denote* $\bar{d}_m := \min_{h\in[H]}\{d_h^\mu(s_h,a_h) : d_h^\mu(s_h,a_h) > 0\}$. *For any* $0 < \delta < 1$, *there exists absolute constants* $c_0, C' > 0$, *such that when* $n > c_0 \cdot 1/\bar{d}_m \cdot \iota$ *(*$\iota = \log(HSA/\delta)$*), with probability* $1-\delta$, *the output policy* $\widehat{\pi}$ *of VPVI satisfies*

$$0 \le v^\star - v^{\widehat{\pi}} \le C' H \sum_{h=1}^H \sum_{(s_h,a_h)\in C_h} d_h^{\pi^\star}(s_h,a_h) \cdot \sqrt{\frac{\iota}{n \cdot d_h^\mu(s_h,a_h)}}. \tag{2.2}$$

The full proof can be found in Appendix A.1. Theorem 2.2.1 makes some improvements over the existing works. First, it is more adaptive than the results with uniform data-coverage

[2] This is due to $\sqrt{\phi\left(s_h,a_h\right)^\top \Lambda_h^{-1} \phi\left(s_h,a_h\right)}$ reduces to $\sqrt{1/n_{s_h,a_h}}$ when setting $\phi(s_h,a_h) = \mathbf{1}(s_h,a_h)$ and $\lambda = 0$.

Assumption 2.1.1 ([3, 17]). In addition, by straightforward calculation (2.2) can be bounded by $\tilde{O}(\sqrt{H^4SC^\star/n})$ which improves VI-LCB [18] by a factor of H. Besides, the analysis of VPVI also improves the direct reduction of PEVI [39] in the tabular case by a factor SA since their $\beta = SAH$ when $d = SA$.

However, VPVI is not optimal as the dependence on horizon is H^4 which does not match the optimal worst case guarantee H^3 [3] in the nonstationary setting. Also, the explicit dependence on H in (2.2) possibly hides some key features of the specific offline RL instances. For example, no improvement can be made if the system has the deterministic transition.

Algorithm 1 Adaptive (*assumption-free*) Pessimistic Value Iteration or LCBVI-Bernstein

1: **Input:** Offline dataset $\mathcal{D} = \{(s_h^\tau, a_h^\tau, r_h^\tau, s_{h+1}^\tau)\}_{\tau,h=1}^{n,H}$. Set $C_1 = 2, C_2 = 14$, failure probability δ.

2: **Initialization:** Set $\widehat{V}_{H+1}(\cdot) \leftarrow 0$. Set $\iota = \log(HSA/\delta)$. (if assumption-free, set $M^\dagger, \widehat{M}^\dagger$ as in Section 2.3.)

3: **for** time $h = H, H-1, \ldots, 1$ **do**

4: Set $\widehat{Q}_h(\cdot,\cdot) \leftarrow \widehat{r}_h(\cdot,\cdot) + (\widehat{P}_h \cdot \widehat{V}_{h+1})(\cdot,\cdot)$ (use $\widehat{r}_h^\dagger + (\widehat{P}_h^\dagger \cdot \widehat{V}_{h+1})$ if assumption-free)

5: $\forall s_h, a_h$, set $\Gamma_h(s_h,a_h) = C_1\sqrt{\frac{\mathrm{Var}_{\widehat{P}_{s_h,a_h}}(\widehat{r}_h+\widehat{V}_{h+1})\cdot\iota}{n_{s_h,a_h}}} + \frac{C_2H\cdot\iota}{n_{s_h,a_h}}$ if $n_{s_h,a_h} \geq 1$, o.w. set to $\frac{CH\iota}{1}$.

6: (If assumption-free, use $C_1\sqrt{\mathrm{Var}_{\widehat{P}^\dagger_{s_h,a_h}}(\widehat{r}_h^\dagger + \widehat{V}_{h+1})\cdot\iota/n_{s_h,a_h}} + \frac{C_2H\cdot\iota}{n_{s_h,a_h}}$ if $n_{s_h,a_h} \geq 1$, o.w. use 0.)

7: Set $\widehat{Q}_h^p(\cdot,\cdot) \leftarrow \widehat{Q}_h(\cdot,\cdot) - \Gamma_h(\cdot,\cdot)$. Set $\overline{Q}_h(\cdot,\cdot) \leftarrow \min\{\widehat{Q}_h^p(\cdot,\cdot), H-h+1\}^+$.{Pessmistic update}

8: $\forall s_h$, Select $\widehat{\pi}_h(\cdot|s_h) \leftarrow \mathrm{argmax}_{\pi_h}\langle\overline{Q}_h(s_h,\cdot), \pi_h(\cdot|s_h)\rangle$. Set $\widehat{V}_h(s_h) \leftarrow \langle\overline{Q}_h(s_h,\cdot), \widehat{\pi}_h(\cdot|s_h)\rangle$.

9: **end for**

10: **Output:** $\{\widehat{\pi}_h\}$.

Now we go deeper to understand what is the more intrinsic characterization for offline reinforcement learning. From the study of VPVI, penalizing the Q-function by $\widetilde{O}(H/\sqrt{n_{s_h,a_h}})$ is crude as it estimates the confidence width of $\widehat{Q}_h$ in Algorithm 4 too conservatively therefore loses the accuracy (the bound is suboptimal). This motivates us to use empirical standard deviation instead to create a more adaptive (and also less conservative) Bernstein-type confidence

width as the pessimistic penalty:

$$\Gamma_h(s_h,a_h)=\widetilde{O}\left[\sqrt{\frac{\mathrm{Var}_{\widehat{P}_{s_h,a_h}}(\widehat{r}_h+\widehat{V}_{h+1})}{n_{s_h,a_h}}}+\frac{H}{n_{s_h,a_h}}\right] \text{ (if } n_{s_h,a_h}>0);\ =\widetilde{O}(H) \text{ (if } n_{s_h,a_h}=0). \tag{2.3}$$

and update $\widehat{Q}_h \leftarrow \widehat{Q}_h - \Gamma_h$. On one hand, $\sqrt{\mathrm{Var}_{\widehat{P}_{s_h,a_h}}(\widehat{r}_h+\widehat{V}_{h+1})/n_{s_h,a_h}}$ is a "less pessimistic" penalty than VPVI due to $\sqrt{\mathrm{Var}_{\widehat{P}}(\widehat{r}_h+\widehat{V}_{h+1})} \le H$ and critically this design is more data-adaptive since it holds negative view towards the locations with high uncertainties and recommends the locations that we are confident about, as opposed to the online RL (which encourages exploration in the uncertain locations). Such principles are not reflected by the isotropic design in VPVI. On the other hand, it carries the extremely negative view towards fully agnostic locations $\widetilde{O}(H)$ which in turn causes the agent unlikely to choose them. We summarized the this *adaptive pessimistic value iteration* (APVI) into the Algorithm 1, with $\widehat{P}_h, \widehat{r}_h$ defined in (2.1). APVI has the following guarantee. A sketch of the analysis is presented in Section 2.4 and Appendix A.3 includes the full proof.

Theorem 2.2.2 (Intrinsic offline RL bound)**.** *Under the Assumption 3.3.3, we first denote* $\bar{d}_m := \min_{h\in[H]}\{d_h^\mu(s_h,a_h) : d_h^\mu(s_h,a_h)>0\}$. *For any* $0<\delta<1$*, there exists absolute constants* $c_0, C'>0$*, such that when* $n > c_0\cdot 1/\bar{d}_m\cdot\iota$ *(*$\iota=\log(HSA/\delta)$*), with probability* $1-\delta$*, the output policy* $\widehat{\pi}$ *of APVI (Algorithm 1) satisfies (*$\widetilde{O}$ *hides log factor and higher order terms)*

$$0\le v^\star - v^{\widehat{\pi}} \le C'\sum_{h=1}^{H}\sum_{(s_h,a_h)\in C_h} d_h^{\pi^\star}(s_h,a_h)\cdot\sqrt{\frac{\mathrm{Var}_{P_{s_h,a_h}}(r_h+V_{h+1}^\star)\cdot\iota}{n\cdot d_h^\mu(s_h,a_h)}}+\widetilde{O}\left(\frac{H^3}{n\cdot\bar{d}_m}\right) \tag{2.4}$$

Remark 1. *APVI (Algorithm 1) can also be called* ***LCBVI-Bernstein*** *as it creates the offline counterpart of UCBVI in [44]. However, to highlight that the resulting bound fully adapts to the specific system structure, we use the word "adaptive" instead.*

APVI makes significant improvements in a lot of aspects. First and foremost, the dominate

term is fully expressed by the system quantities that admits no explicit dependence on H, S, A. To the best of our knowledge, this is the first offline RL bound that concretely depicts the interrelations within the problem when the problem instance is a tuple $(M, \pi^\star, \mu)$: an MDP M (coupled with the optimal policy $\pi^\star$) with the data rolling from an offline logging policy μ. As we will discuss later, this result indicates (nearly) all the optimal worst-case non-adaptive bounds (and clearly also the VPVI) under their respective regimes / assumptions. Thus, (2.4) is generic. More interestingly, Theorem 2.2.2 caters to the specific MDP structures and adaptively yields improved sample complexities (*e.g.* faster convergence in deterministic systems) that existing works cannot imply. Such features are crucial as it helps us to understand what type of problems are harder / easier than others, and even more, in a *quantitative* way. Hence, we call the quantity $\sum_{h=1}^{H} \sum_{(s_h,a_h)\in C_h} d_h^{\pi^\star}(s_h,a_h) \cdot \sqrt{\frac{\mathrm{Var}_{P_{s_h,a_h}}(r_h+V_{h+1}^\star)}{n\cdot d_h^\mu(s_h,a_h)}}$ *intrinsic offline reinforcement learning bound*. In the sequel, we provide thorough discussions to explain the intrinsic bound embraces the fundamental challenges in offline RL and the strong adaptivity.

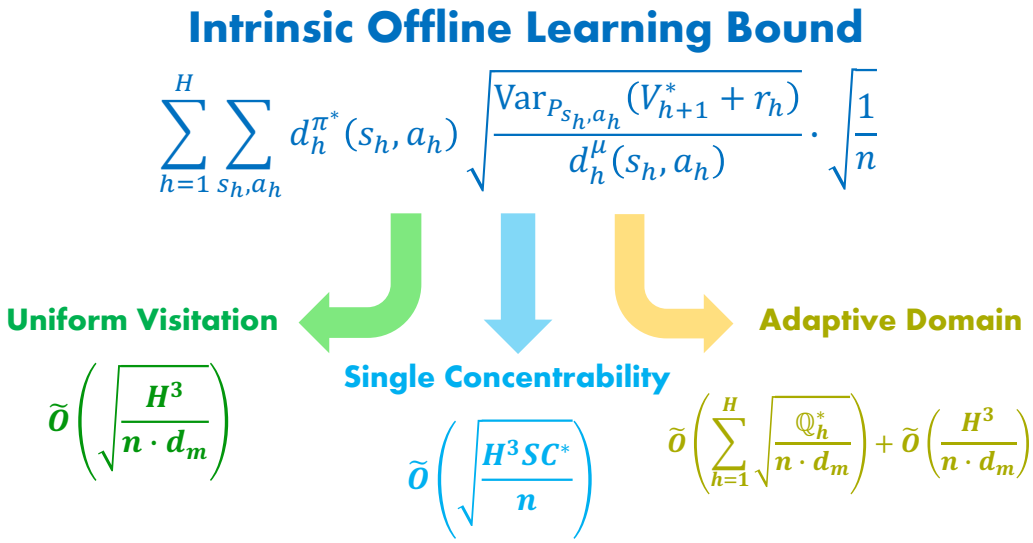

Figure 2.1: A visualization on how intrinsic learning bound subsumes existing best-known results: uniform visitation, single concentrability (partial coverage) and adaptive domain.

2.2.1 Optimality under Uniform data-coverage assumption

Under the uniform exploration Assumption 2.1.1 with parameter $d_m := \min_{h,s_h,a_h} d_h^\mu(s_h, a_h) > 0$, [3] analyzes the model-based plug-in approach and obtains the optimal sample complexity $\widetilde{O}(H^3/d_m\epsilon^2)$ and shows $\Omega(H^3/d_m\epsilon^2)$ is also the lower bound. Indeed, this rate can be directly implied by the intrinsic RL bound via *Cauchy inequality* and *the Sum of Total Variance*:[3]

$$\begin{aligned}
&\sum_{h=1}^{H} \langle d_h^{\pi^\star}(\cdot), \sqrt{\frac{\mathrm{Var}_{P_{(\cdot)}}(r_h + V_{h+1}^\star)}{n \cdot d_h^\mu(\cdot)}} \rangle = \sum_{h=1}^{H} \langle \sqrt{d_h^{\pi^\star}(\cdot)}, \sqrt{\frac{d_h^{\pi^\star}(\cdot) \odot \mathrm{Var}_{P_{(\cdot)}}(r_h + V_{h+1}^\star)}{n \cdot d_m}} \rangle \\
&\leq \sum_{h=1}^{H} \left\| \sqrt{d_h^{\pi^\star}(\cdot)} \right\|_2 \left\| \sqrt{\frac{d_h^{\pi^\star}(\cdot) \odot \mathrm{Var}_{P_{(\cdot)}}(r_h + V_{h+1}^\star)}{n \cdot d_m}} \right\|_2 \leq \sqrt{\frac{H \cdot \mathrm{Var}_{\pi^\star}(\sum_{h=1}^{H} r_h)}{n \cdot d_m}} \leq \sqrt{\frac{H^3}{n \cdot d_m}}
\end{aligned} \tag{2.5}$$

which translates to $\widetilde{O}(H^3/d_m\epsilon^2)$ complexity. Our result maintains the optimal worst-case guarantee when μ has the uniform data-coverage:

Proposition 2.2.1. *Under Assumption 2.1.1 and apply Theorem 2.2.2, APVI achieves the sample complexity of minimax-rate $\widetilde{O}(H^3/d_m\epsilon^2)$ (Theorem 4.1 and Theorem G.2 in [3]).*

Remark 2. *We believe if the MDP is time-invariant, then by a modified construction of $\widehat{P}$, $\widehat{r}$ in* (2.1) *our result will imply the minimax-rate of $\widetilde{O}(H^2/d_m\epsilon^2)$ as achieved in [16].*

2.2.2 Bounded sum of total rewards and the Horizon-Free case

There is another thread of studies that follow the bounded sum of total rewards assumption: *i.e.* $r_h \geq 0$, $\sum_{h=1}^{H} r_h \in [0, 1]$ [45, 46, 47]. Such a setting is much weaker than the uniform bounded instantaneous reward condition, as explained in [48]. In offline RL, [17] derives the nearly horizon-free worst case bound $\widetilde{O}(\sqrt{1/nd_m})$ for the time-invariant MDPs, under the Assumption 2.1.1. As a comparison, our Theorem 2.2.2 achieves the following guarantee for the time-varying (non-stationary) MDPs.

[3]Here $\odot$ denotes element-wise multiplication. Also note under 2.1.1, our $\bar{d}_m = d_m$.

Proposition 2.2.2. *Assume $r_h \geq 0$, $\sum_{h=1}^{H} r_h \leq 1$. Then in the time-varying case AVPI (Theorem 2.2.2) outputs a policy $\widehat{\pi}$ such that the suboptimality gap $v^\star - v^{\widehat{\pi}}$ is bounded by $\widetilde{O}(\sqrt{H/nd_m})$ with high probability under the Assumption 2.1.1.*

The derivation is straightforward by using $\mathrm{Var}_{\pi^\star}(\sum_{h=1}^{H} r_h) \leq 1$ in (2.5). This proposition is interesting since it indicates when the MDP is non-stationary, $\widetilde{O}(H/d_m\epsilon^2)$ is required in the worst case even under $\sum_{h=1}^{H} r_h \leq 1$.[4] The extra H factor resembles the challenge that we have H transitions $(P_1, \ldots, P_H)$ to learn, as opposed to the bandit-type $1/d_m\epsilon^2$ result due to there is only one P throughout (time-invariant). This reveals that one hardness in solving the MDP is in proportion to the number of different transition kernels within the MDP. Such a finding could help researchers understand the special settings like *low switching cost in transitions* [49] or *non-stationarity* [50].

2.2.3 Optimality with Single Concentrability

In the finite horizon discounted setting, [18] proposes the single policy concentrability assumption which is defined as $C^\star := \max_{h,s,a} \frac{d_h^{\pi^\star}(s,a)}{d_h^{\mu}(s,a)} < \infty$ in the current episodic non-stationary MDP setting. Their lower bound translates to $\Omega(\sqrt{\frac{H^3SC^\star}{n}})$ and their VI-LCB algorithm yields $\widetilde{O}(\sqrt{\frac{H^5SC^\star}{n}})$ suboptimality gap in H-horizon case. Since single policy concentrability is strictly weaker than its uniform version (Assumption 2.1.2), we only discuss this set up. In particular, we have the following implication from our Theorem 2.2.2 (whose derivation can be found in Appendix A.4):

Proposition 2.2.3. *Let $\pi^\star$ be a deterministic policy such that $C^\star := \max_{h,s,a} \frac{d_h^{\pi^\star}(s,a)}{d_h^{\mu}(s,a)} < \infty$. Then by Theorem 2.2.2, with high probability the output policy of APVI satisfies the suboptimality gap $\widetilde{O}(\sqrt{\frac{H^3SC^\star}{n}})$ in the time-varying (non-stationary) MDPs.*

[4] Suppose in this case we can achieve $\widetilde{O}(1/d_m\epsilon^2)$ just like [17], then by a rescaling we obtain the $\widetilde{O}(H^2/d_m\epsilon^2)$ under the usual $0 \leq r_h \leq 1$ assumption which violates the $\Omega(H^3/d_m\epsilon^2)$ lower bound.

This can computed similar to (2.5) except we use $\frac{d_h^{\pi^\star}(s,a)}{d_h^{\mu}(s,a)} \leq C^\star$. Our implication improves the VI-LCB by the factor H^2 (in terms of sample complexity) and is optimal (recover the concurrent [19]). Qualitatively, single concentrability is the same as Assumption 3.3.3, but the use of $C^\star$ makes the bound highly problem independent and limits the adaptivity. Problem dependent bound is a more interesting domain as it tailors to each MDP separately. We discuss it now.

2.2.4 Problem dependent domain

We define the *pre-step environmental norm* (the finite horizon counterpart of [51]) as: $\mathbb{Q}_h^\star = \max_{s_h,a_h} \mathrm{Var}_{P_{s_h,a_h}}(r_h + V_{h+1}^\star)$ for all $h \in [H]$, and relax the total sum of rewards to be bounded by any arbitrary value $\mathcal{B}$ (*i.e.* $\sum_{h=1}^H r_h \leq \mathcal{B}$), then Theorem 2.2.2 implies:

Proposition 2.2.4. *Under Assumption 2.1.1, with high probability, subopmality of AVPI is bounded by*

$$\min\left\{\widetilde{O}\Big(\sum_{h=1}^{H}\sqrt{\frac{\mathbb{Q}_h^\star}{n\bar{d}_m}}\Big), \widetilde{O}\Big(\sqrt{\frac{H\cdot\mathcal{B}^2}{n\bar{d}_m}}\Big)\right\} + \widetilde{O}\Big(\frac{H^3}{n\bar{d}_m}\Big).$$

Such a result mirrors the online version of the tight problem-dependent bound [52] but with a more general *pre-step environmental norm* for the non-stationary MDPs.[5] For the problem instances with either small $\mathcal{B}$ or small $\mathbb{Q}_h^\star$, our result yields much better performances, as discussed in the following.

Deterministic systems. For many practical applications of interest, the systems are equipped with low stochasticity, *e.g.* robotics, or even deterministic dynamics, *e.g.* the game of GO. In those scenarios, the agent needs less experience for each state-action therefore the learning procedure could be much faster. In particular, when the system is fully deterministic (in both transitions and rewards) then $\mathbb{Q}_h^\star = 0$ for all h. This enables a faster convergence rate of order

[5] [52] uses the maximal version by maximizing over h.

$\frac{H^3}{n\bar{d}_m}$ and significantly improves over the existing non-adaptive results that have order $\frac{1}{\sqrt{n}}$. The convergence rate $\frac{1}{n}$ matches [53] by translating their constant (in T) regret into the PAC bound.

Partially deterministic systems. Practical worlds are complicated and we could sometimes have a mixture model which contains both deterministic and stochastic steps. In those scenarios, the main complexity is decided by the number of stochastic stages: suppose there are t stochastic P_h, r_h's and $H - t$ deterministic $P_{h'}, r_{h'}$'s, then completing the offline learning guarantees $t \cdot \sqrt{\max Q_h^\star / n\bar{d}_m}$ suboptimality gap, which could be much smaller than $H \cdot \sqrt{\max Q_h^\star / n\bar{d}_m}$ when $t \ll H$.

Fast mixing domains. Consider a class of highly mixing non-stationary MDPs (a variant of [54]) that satisfies the transition $P_h(\cdot|s_h, a_h) := \nu_h(\cdot)$ depends on neither the state s_h nor the action a_h. Define $\bar{s}_t := \arg\max V_t^\star(s)$ and $\underline{s}_t := \arg\max V_t^\star(s)$. Also, denote $\mathrm{rng} V_h^\star$ to be the range of $V_h^\star$. In such cases, Bellman optimality equations have the form

$$V_h^\star\left(\bar{s}_h\right) = \max_a \left(r_h\left(\bar{s}_h, a\right) + \nu_h^\top V_{h+1}^\star\right), \quad V_h^\star\left(\underline{s}_h\right) = \max_a \left(r_h\left(\underline{s}_h, a\right) + \nu_h^\top V_{h+1}^\star\right),$$

which yields $\mathrm{rng} V_h^\star = V_h^\star\left(\bar{s}_h\right) - V_h^\star\left(\underline{s}_h\right) = \max_a r_h\left(\bar{s}_h, a\right) - \min_a r_h\left(\underline{s}_h, a\right) \le 1$, and this in turn gives $\mathbb{Q}_h^\star \le 1 + (\mathrm{rng} V_h^\star)^2 = 2$. As a result, the suboptimality is bounded by $\widetilde{O}(\sqrt{H^2/nd_m})$ in the worst case. This result reveals, although this is a family of stochastic non-stationary MDPs, but it is only as hard as the family of stationary MDPs in the minimax sense ($\Omega(H^2/d_m\epsilon^2)$).

Tabular contextual bandits. Our result also implies $\widetilde{O}(\sum_{x_1,a_1} d_1^{\pi^\star}(x_1, a_1)\sqrt{\frac{\mathrm{Var}(r_1)}{n \cdot d_1^\mu(x_1,a_1)}})$ gap for the *offline tabular contextual bandit* problem and improves to $\widetilde{O}(1/nd_m)$ when the reward is deterministic. In either cases, the result is optimal and this is due to: when r_1 is deterministic, the agent only needs one sample at every location (see [55] for a survey).

2.3 Towards Assumption-Free Offline RL

While assumption 3.3.3 is (arguably) the weakest assumption for correctly learning the optimal value, for the real-world applications even this might not be guaranteed. Can we still learn something meaningful? In this section, we consider this most general setting where the behavior policy μ can be arbitrary. In this case, μ might not cover any optimal policy $\pi^\star$ (*i.e.* there might be high reward location (s, a) that μ can never visit, *e.g.* in the extreme case where a clumsy doctor only uses one treatment all the time), and, irrelevant to the number of episode n, a constant suboptimality gap needs to be suffered. To tackle this problem, we create a fictitious augmented MDP $M^\dagger$ that can help characterize the discrepancy of the values between the original MDP M and the estimated MDP $\widehat{M}^\dagger$. In particular, $M^\dagger$ is negative towards agnostic state-actions s_h, a_h by setting $r_h^\dagger = 0$ and transitions to an absorbing state $s_{h+1}^\dagger$.

Pessimistic augmented MDP. $M^\dagger$ is defined with one extra state $s_h^\dagger$ for all $h \in \{2, \dots, H+1\}$ with the augmented state space $\mathcal{S}^\dagger = \mathcal{S} \cup \{s_h^\dagger\}$. The transition and the reward are defined as follows:

$$P_h^\dagger(\cdot \mid s_h, a_h) = \begin{cases} P_h(\cdot \mid s_h, a_h), \ n_{s_h,a_h} > 0, \\ \delta_{s_{h+1}^\dagger}, \ s_h = s_h^\dagger \text{ or } n_{s_h,a_h} = 0. \end{cases} \quad r^\dagger(s_h, a_h) = \begin{cases} r(s_h, a_h), \ n_{s_h,a_h} > 0, \\ 0, \ s_h = s_h^\dagger \text{ or } n_{s_h,a_h} = 0. \end{cases}$$

here δ_s is the Dirac measure and we denote $V_h^{\dagger\pi}$ and $v^{\dagger\pi}$ to be the values under $M^\dagger$. $\widehat{M}^\dagger$ is the empirical counterpart of $M^\dagger$ with $\widehat{P}, \widehat{r}$ (the same as (2.1)) replacing P, r. By Algorithm 1, we have

Theorem 2.3.1 (Assumption-free offline reinforcement learning)**.** *Let us make no assumption for μ and still denote $\bar{d}_m := \min_{h\in[H]}\{d_h^\mu(s_h, a_h) : d_h^\mu(s_h, a_h) > 0\}$. For any $0 < \delta < 1$, there exists absolute constants $c_0, C' > 0$, such that when $n > c_0 \cdot 1/\bar{d}_m \cdot \iota$ ($\iota = \log(HSA/\delta)$), with*

probability $1-\delta$*, the output policy* $\widehat{\pi}$ *of APVI satisfies (recall* $C_h := \{(s_h, a_h) : d_h^\mu(s_h, a_h) > 0\}$*)*

$$v^\star - v^{\widehat{\pi}} \le \sum_{h=2}^{H+1} d_h^{\dagger\pi^\star}(s_h^\dagger) + C' \sum_{h=1}^{H} \sum_{(s_h,a_h)\in C_h} d_h^{\dagger\pi^\star}(s_h, a_h) \cdot \sqrt{\frac{\mathrm{Var}_{P^\dagger_{s_h,a_h}}(r_h^\dagger + V_{h+1}^{\dagger\pi^\star}) \cdot \iota}{n \cdot d_h^\mu(s_h, a_h)}} + \widetilde{O}\left(\frac{H^3}{n\bar{d}_m}\right), \quad (2.6)$$

where $d_h^{\dagger\pi^\star}(s_h, a_h) \le d_h^{\pi^\star}(s_h, a_h), V_h^{\dagger\pi^\star}(s_h) \le V_h^\star(s_h)$ *for all* $s_h, a_h \in \mathcal{S}\times\mathcal{A}$*, and for all* $h \in [H]$*,* $d_h^{\dagger\pi^\star}(s_h^\dagger) = \sum_{t=1}^{h-1} \sum_{(s_t,a_t)\in\mathcal{S}\times\mathcal{A}\setminus C_t} d_t^{\dagger\pi^\star}(s_t, a_t)$*. The proof is in Appendix A.2.*

Take-aways of Theorem 2.3.1. In $M^\dagger$, there is no agnostic location any more since the original unknown spaces now all have *known* deterministic transitions to $s^\dagger$ in $M^\dagger$. At a price, the algorithm has to suffer the constant suboptimality $\sum_{h=2}^{H+1} d_h^{\dagger\pi^\star}(s_h^\dagger)$ due to no data in the region. The quantity $\sum_{h=2}^{H+1} d_h^{\dagger\pi^\star}(s_h^\dagger)$ helps characterize the hardness when nothing is assumed about μ: it is always less than H (cannot suffer more than H suboptimality); under Assumption 2.1.1, it is 0 since $M^\dagger = M$ with high probability (by Chernoff bound) and this causes $\mathcal{S}\times\mathcal{A}\setminus C_h = \emptyset$; under Assumption 3.3.3, it is also 0 and 2.3.1 reduces to Theorem 2.2.2 (see Appendix A.3).

2.3.1 Assumption Free vs Without Great Coverage (Partial Coverage)

Recently there is a surge of studies that aim at weakening the assumptions of provable offline / batch RL. Those learning bounds are derived (mostly) under the insufficient data coverage assumptions. One type of works consider the assumption *without great coverage* (or partial coverage): [36, 38] assume $\max_{s,a} d^{\pi_e}(s, a)/\mu(s, a) < \infty$ where π_e is either an expert policy or a policy of great quality and they further compete against with this policy π_e. Those assumptions are similar to 3.3.3 and therefore are stronger than the assumption-free RL we considered in 2.3.1.

In addition, there are other studies that apply to the case where μ can be arbitrary: [35] considers the behavior policy with insufficient coverage probability ϵ_ζ (see their Definition 1),

and they end up with the constant suboptimality gap $\frac{V_{\max}\epsilon_\zeta}{1-\gamma}$ (their Theorem 1), when the insufficient coverage probability $\epsilon_\zeta > 0$, this gap has order $(1-\gamma)^{-2}$, which is larger in order than the biggest possible suboptimality gap $(1-\gamma)^{-1}$ therefore unable to characterize the essential statistical gap over the region that can never be visited by the behavior policy (and this happens similarly in [37], see their Theorem 1); [39] derive the nice assumption-free result via regularization and their bound can incur $O(H^2)$ constant gap when there is at least one (s_h, a_h) cannot be obtained by μ for all $h \in [H]$ (*i.e.* replacing $nd_h^\mu(s_h, a_h)$ by 1 in (2.2)). The concurrent work [40] provides a better characterization (and they call it *off-support error*) with roughly $\frac{1}{1-\gamma}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left(d_\pi \backslash \nu\right)(s,a)\left[\Delta f_\pi(s,a) - \left(\mathcal{T}^\pi \Delta f_\pi\right)(s,a)\right]$, however, in the worst case $\Delta f_\pi(s,a) - \left(\mathcal{T}^\pi \Delta f_\pi\right)(s,a)$ might be large (which depends on the quality (assumption) of the function approximation class).

In contrast, our $\sum_{h=2}^{H+1} d_h^{\dagger\pi^\star}(s_h^\dagger)$ quantity (with $d_h^{\dagger\pi^\star}(s_h^\dagger) = \sum_{t=1}^{h-1}\sum_{(s_t,a_t)\in\mathcal{S}\times\mathcal{A}\setminus C_t} d_t^{\dagger\pi^\star}(s_t,a_t) \leq 1$) describes the "must-suffer" gap in a more precise way by absorbing all the agnostic probabilities into $s^\dagger$ and it is always bounded between 0 and H. It reduces to 0 when $\pi^\star$ is covered. The gap is always of order H (as opposed to $O(H^2)$).

2.4 Sketch of the Analysis for APVI

We sketch the key proving ideas in Section 2.2. Our analysis of the intrinsic learning bound in Section 2.2 leverage the key design feature of APVI that $\widehat{V}_{h+1}$ only depends on the transition data from time $h+1$ to H while $\widehat{P}_h$ only uses transition pairs at time h. This enables concentration inequalities due the *conditional* independence. To cater for the data-adaptive bonus (2.3), we use *Empirical* Bernstein inequality to get $(\widehat{P}_h - P_h)\widehat{V}_{h+1} \lesssim \sqrt{\mathrm{Var}_{\hat{P}}(\widehat{V}_{h+1})/n_{s_h,a_h}}$. Especially, to recover the $\sqrt{\mathrm{Var}_P(V^\star_{h+1})}$ structure to we use a self-bounding reduction as follows. First, $\sqrt{\mathrm{Var}_{\hat{P}}(\widehat{V}_{h+1})} - \sqrt{\mathrm{Var}_P(\widehat{V}_{h+1})} \lesssim H/\sqrt{n\bar{d}_m}$ and $\sqrt{\mathrm{Var}_P(\widehat{V}_{h+1})} - \sqrt{\mathrm{Var}_P(V^\star_{h+1})} \leq ||\widehat{V}_{h+1} - V^\star_{h+1}||_\infty$. Next, we use (2.2) as the intermediate step to crude bounding $||\widehat{V}_{h+1} - V^\star_{h+1}||_\infty \lesssim H^2/\sqrt{n\bar{d}_m}$

(where "the use of (2.2)" is the more intricate self-bounding Lemma A.2.6 in the actual proof) and this yields the desired structure of $\sqrt{\mathrm{Var}_P(V^\star_{h+1})} + H^2/\sqrt{n\bar{d}_m}$. Lastly, we can combine this with *the extended value difference lemma* in [56] to bound $V^\star_1 - \widehat{V}_1$ and leverage the pessimistic design for bounding $\widehat{V}_1 - V^{\widehat{\pi}}_1$.

2.5 Conclusion

This work studies the offline reinforcement learning problem and contributes the intrinsic offline learning bound which is a near-optimal and strong adaptive bound that subsumes existing worst-case bounds under various assumptions. The adaptive characterization of the intrinsic bound abandons the explicit dependence on $H, S, A, C^\star, d_m$ and helps reveal the fundamental hardness of each individual instances. In this sense, it draws a clearer picture of what offline reinforcement learning looks like and serves as a step towards instance optimality in offline RL.

Nevertheless, it is still unclear whether (2.4) is optimal over all the instances. For example, for fully deterministic systems, our bound provides a faster convergence $H^3/n\bar{d}_m$, however, H^3 might be very suboptimal comparing to algorithms that are designed specifically for deterministic MDPs, since the agent only need to experience each location (s, a) once to fully acquire the dynamic $P(\cdot|s, a)$ and $r(s, a)$. Recently, [57] goes beyond the minimax (worst case) optimality and studies the instance optimality behavior for the simplified batch bandit setting. One of their findings is: for "easy enough" tasks, different type of algorithms can be equally good, provably. This seems to suggest instance optimality only matters for problems that are hard to learn. How to formally define the instance optimality metric for different problems remains an open problem and how to design a single algorithm that can achieve optimality for all instances could be challenging (or even infeasible). We leave those as the future works.

Chapter 3

Near-optimal Offline Reinforcement Learning with Linear Representation

In this Chapter of the thesis, we consider offline reinforcement learning with function approximation. Due to the advantage that appropriate function approximators can help mitigate the sample complexity burden in modern reinforcement learning problems, existing endeavors usually enforce powerful function representation models (*e.g.* neural networks) to learn the optimal policies. However, a precise understanding of the statistical limits with function representations, remains elusive, even when such a representation is linear.

Towards this goal, we study the statistical limits of offline reinforcement learning with linear model representations. To derive the tight offline learning bound, we design the *variance-aware pessimistic value iteration* (VAPVI), which adopts the conditional variance information of the value function for time-inhomogeneous episodic linear *Markov decision processes* (MDPs). VAPVI leverages estimated variances of the value functions to reweight the Bellman residuals in the least-square pessimistic value iteration and provides improved offline learning bounds over the best-known existing results (whereas the Bellman residuals are equally weighted by design). More importantly, our learning bounds are expressed in terms of system quantities,

which provide natural instance-dependent characterizations that previous results are short of. We hope our results draw a clearer picture of what offline learning should look like when linear representations are provided.

3.1 Motivation and Related Prior Works

Offline reinforcement learning (offline RL or batch RL [11, 10]) is the framework for learning a reward-maximizing policy in an unknown environment (*Markov Decision Process* or MDP)[1] using the logged data coming from some behavior policy μ. Function approximations, on the other hand, are well-known for generalization in the standard supervised learning. Offline RL with function representation/approximation, as a result, provides generalization across large state-action spaces for the challenging sequential decision-making problems when no iteration is allowed (as opposed to online learning). This paradigm is crucial to the success of modern RL problems as many deep RL algorithms find their prototypes in the literature of offline RL. For example, [14] provides a view that *Fitted Q-Iteration* [58, 59] can be considered as the theoretical prototype of the deep Q-networks algorithm (DQN) [60] with neural networks being the function representors. On the empirical side, there are a huge body of deep RL-based algorithms [60, 8, 61, 62, 63, 37, 64, 65, 23, 66, 67] that utilize function approximations to achieve respective successes in the offline regime. However, it is also realized that practical function approximation schemes can be quite sample inefficient (*e.g.* millions of samples are needed for deep Q-network to solve certain Atari games [60]).

To understand this phenomenon, there are numerous studies consider how to achieve sample efficiency with function approximation from the theoretical side, as researchers find sample efficient algorithms are possible with particular model representations, in either online RL (*e.g.*

[1] The environment could have other forms as well, *e.g. partially-observed MDP* (POMDP) or *non-markovian decision process* (NMDP).

[68, 69, 70, 71, 72, 46, 73, 74, 75, 76, 77, 78]) or offline RL (*e.g.* [79, 13, 14, 80, 40, 81, 82, 83, 41]).

Among them, the linear MDP model [69, 71], where the transition is represented as a linear combinations of the given d-dimensional feature, is (arguably) the most studied setting in function approximation and there are plenty of extensions based upon it (*e.g.* generalized linear model [84], reward-free RL [85], gap-dependent analysis [86] or generative adversarial learning [87]). Given its prosperity, however, there are still unknowns for understanding function representations in RL, especially in the offline case.

- While there are surging researches in showing provable sample efficiency (polynomial sample complexity is possible) under a variety of function approximation schemes, how to improve the sample efficiency for a given class of function representations remains understudied. For instance, given a neural network approximation class, an algorithm that learns the optimal policy with complexity $O(H^{10})$ is far worse than the one that can learn in $O(H^3)$ sample complexity, despite that both algorithms are considered sample efficient. Therefore, how to achieve the optimal/tight sample complexity when function approximation is provided is a valuable question to consider. On the other hand, it is known that tight sample complexity, due to the limit of the existing statistical analysis tools, can be very tough to establish when function representation has a very complicated form. However, does this mean tight analysis is not hopeful even when the representation is linear?

- Second, in the existing analysis of offline RL (with function approximation or simply the tabular MDPs), the learning bounds depend either explicitly on the data-coverage quantities (*e.g.* uniform concentrability coefficients [13, 14], uniform visitation measure [3, 4] and single concentrability [18, 19]) or the horizon length H [80, 38]. While those results are valuable as they do not depend on the structure of the particular problem (therefore,

remain valid even for pathological MDPs), in practice, the empirical performances of offline reinforcement learning are often far better than those non-adaptive bounds would indicate. Can the learning bounds reflect the nature of individual MDP instances when the MDP model has a certain function representation?

Those observation motivates us to consider the following question in offline RL: ***Can we achieve the statistical limits for offline RL when models have linear representations?***

3.1.1 Related works

Offline RL with general function representations. The finite sample analysis of offline RL with function approximation is initially conducted by Fitted Q-Iteration (FQI) type algorithms and can be dated back to [79, 31, 33, 34]. Later, [13, 12, 14] follow this line of research and derive the improved learning results. However, owing to the aim for tackling general function approximation, those learning bounds are expressed in terms of the stringent *concentrability coefficients* (therefore, are less adaptive to individual instances) and are usually only *information-theoretical*, due to the computational intractability of the optimization procedure over the general function classes. Other works impose weaker assumptions (*e.g.* partial coverage [35, 37, 38]), and their finite sample analysis are generally suboptimal in terms of H or the effective horizon $(1-\gamma)^{-1}$.

Offline RL with tabular models. For tabular MDPs, tight learning bounds can be achieved under several data-coverage assumptions. For the class of problems with uniform data-visitation measure d_m, the near-optimal sample complexity bound has the rate $O(H^3/d_m\epsilon^2)$ for time-inhomogeneous MDPs [3] and $O(H^2/d_m\epsilon)$ for time-homogeneous MDPs [4, 17]. Under the single concentrability assumption, the tight rate $O(H^3SC^\star/\epsilon^2)$ is obtained by [19]. In particular, the recent study [5] introduces the *intrinsic offline learning bound* that is not only instance-dependent but also subsumes previous optimal results. More recently, [88] uses the model-free

approaches to achieve the minimax rate with a $[0, H^{-1}]$ ϵ-range.

Offline RL with linear model representations. Recently, there is more focus on studying the provable efficient offline RL under the linear model representations. [80] first shows offline RL with linear MDP is provably efficient by *the pessimistic value iteration*. Their analysis deviates from their lower bound by a factor of $d \cdot H$ (check their Theorem 4.4 and 4.6). Later, [40] considers function approximation under the Bellman-consistent assumptions, and, when realized to linear MDP setting, improves the sample complexity guarantee of [80] by an order $O(d)$ (Theorem 3.2).[2] However, their improvement only holds for finite action space (due to the dependence $\log |\mathcal{A}|$) and by the direct reduction (from Theorem 3.1) their result does not imply a computationally tractable algorithm with the same guarantee.

Concurrently, [41] considers the Linear Bellman Complete model and designs the *actor-critic* style algorithm that achieves tight result under the assumption that the value function is bounded by 1. While their algorithm is efficient (which is based on solving a sequence of second-order cone programs), the resulting learning bound requires the action space to be finite due to the mirror descent updates in the *Actor* procedure [89]. Besides, assuming the value function to be less than 1 simplifies the challenges in dealing with horizon H since when rescaling their result to $[0, H]$, there is a H factor blow-up, which makes no horizon improvement comparing to [80]. As a result, none of the existing algorithms can achieve the statistical limit for the well-structured linear MDP model with the general (infinite or continuous) state-action spaces. On the other hand, [90, 91] study the statistical hardness of offline RL with linear representations by proving the exponential lower bounds. Recently, [92] shows realizability and concentrability are not sufficient for offline learning when state space is arbitrary large.

Variance-aware studies. [93] first incorporates the variance structure in online tabular MDPs and [52] tightens the result. For linear MDPs, [76] first uses variance structure to achieve near-optimal result and the *Weighted OFUL* incorporates the variance structure explicitly in the

[2] This comparison is based on translating their infinite horizon discounted setting to the finite-horizon case.

regret bound. Recently, Variance-awareness is also considered in [94] for horizon-free setting and for OPE problem [82]. In particular, We point out that [82] is the first work that uses variance reweighting for policy evaluation in offline RL, which inspires our study for policy optimization problem. The guarantee of [82] strictly improves over [95] for OPE problem.

Our contribution. We design the *variance-aware pessimistic value iteration* (VAPVI, Algorithm 2) which incorporates the conditional variance information of the value function and, by the variance structure, Theorem 3.3.1 is able to improve over the aforementioned state-of-the-art guarantees. In addition, we further improve the state-action guarantee by designing an even tighter bonus (3.4). VAPVI-Improved (Theorem 3.3.2) is near-minimax optimal as indicated by our lower bound (Theorem 3.3.4). Importantly, the resulting learning bounds from VAPVI/VAPVI-Improved are able to characterize the adaptive nature of individual instances and yield different convergence rates for different problems. Algorithmically, our design builds upon the nice [82] with pessimism as we use the estimated variances to reweight the Bellman residual learning objective so that the (training) samples with high uncertainty get less attention. This is the key to obtaining instance-adaptive guarantees.

3.2 Preliminaries for Linear Markov Decision Processes

Episodic time-homogeneous linear Markov decision process. A finite-horizon *Markov Decision Process* (MDP) is denoted as $M = (\mathcal{S}, \mathcal{A}, P, r, H, d_1)$ [29], where $\mathcal{S}$ is the arbitrary state space and $\mathcal{A}$ is the arbitrary action space which can be infinite or even continuous. A time-inhomogeneous transition kernel $P_h : \mathcal{S} \times \mathcal{A} \mapsto \Delta^{\mathcal{S}}$ ($\Delta^{\mathcal{S}}$ represents a probability simplex) maps each state action(s_h, a_h) to a probability distribution $P_h(\cdot|s_h, a_h)$ and P_h can be different across time. In addition, $r : \mathcal{S} \times A \mapsto \mathbb{R}$ is the mean reward function satisfying $0 \le r \le 1$. d_1 is the initial state distribution. H is the horizon. A policy $\pi = (\pi_1, \ldots, \pi_H)$ assigns each state $s_h \in \mathcal{S}$ a probability distribution over actions according to the map $s_h \mapsto \pi_h(\cdot|s_h)\ \forall h \in [H]$ and

induces a random trajectory $s_1, a_1, r_1, \ldots, s_H, a_H, r_H, s_{H+1}$ with $s_1 \sim d_1, a_h \sim \pi(\cdot|s_h), s_{h+1} \sim P_h(\cdot|s_h, a_h), \forall h \in [H]$. In particular, we adopts the linear MDP protocol from [71, 80], meaning that the transition kernel and the mean reward function admit linear structures in the feature map.[3]

Definition 3.2.1 (Linear MDPs)**.** [4] *An episodic MDP $(\mathcal{S}, \mathcal{A}, H, P, r)$ is called a linear MDP with a known (unsigned) feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ if there exist d unknown (unsigned) measures $\nu_h = (\nu_h^{(1)}, \ldots, \nu_h^{(d)})$ over $\mathcal{S}$ and an unknown vector $\theta_h \in \mathbb{R}^d$ such that*

$$P_h\left(s' \mid s, a\right) = \left\langle \phi(s,a), \nu_h\left(s'\right)\right\rangle, \quad r_h(s,a) = \langle \phi(x,a), \theta_h\rangle, \quad \forall s', s \in \mathcal{S},\ a \in \mathcal{A},\ h \in [H].$$

where $\left\|\nu_h(\mathcal{S})\right\|_2 \leq \sqrt{d}$ and $\max(\|\phi(s,a)\|_2, \left\|\theta_h\right\|_2) \leq 1$ for all $h \in [H]$ and $\forall s, a \in \mathcal{S} \times \mathcal{A}$. $\left\|\mu_h(\mathcal{S})\right\| = \int_{\mathcal{S}} \left\|\mu_h(s)\right\| ds$.

V-values and Q-values. For any policy π, the V-value functions $V_h^\pi(\cdot) \in \mathbb{R}^S$ and Q-value functions $Q_h^\pi(\cdot,\cdot) \in \mathbb{R}^{S\times A}$ are defined as: $V_h^\pi(s) = \mathbb{E}_\pi[\sum_{t=h}^H r_t | s_h = s]$, $Q_h^\pi(s,a) = \mathbb{E}_\pi[\sum_{t=h}^H r_t | s_h, a_h = s, a]$, $\forall s, a, h \in \mathcal{S}, \mathcal{A}, [H]$. The performance measure is defined as $v^\pi := \mathbb{E}_{d_1}\left[V_1^\pi\right] = \mathbb{E}_{\pi, d_1}\left[\sum_{t=1}^H r_t\right]$. The Bellman (optimality) equations follow $\forall h \in [H]$: $Q_h^\pi = r_h + P_h V_{h+1}^\pi$, $V_h^\pi = \mathbb{E}_{a\sim\pi_h}[Q_h^\pi]$, $Q_h^\star = r_h + P_h V_{h+1}^\star$, $V_h^\star = \max_a Q_h^\star(\cdot, a)$ (where Q_h, V_h, P_h are vectors). By Definition 3.2.1, the Q-values also admit linear structures, *i.e.* $Q_h^\pi = \langle \phi, w_h^\pi\rangle$ for some $w_h^\pi \in \mathbb{R}^d$ (Lemma B.5.9). Lastly, for a policy π, we denote the induced occupancy measure over the state-action space at any time $h \in [H]$ to be: for any $E \subseteq \mathcal{S} \times \mathcal{A}$, $d_h^\pi(E) := \mathbb{E}[(s_h, a_h) \in E | s_1 \sim d_1, a_i \sim \pi(\cdot|s_i), s_i \sim P_{i-1}(\cdot|s_{i-1}, a_{i-1}), 1 \leq i \leq h]$ and $\mathbb{E}_{\pi,h}[f(s,a)] := \int_{\mathcal{S}\times\mathcal{A}} f(s,a) d_h^\pi(s,a) \cdot ds da$. Here for notation simplicity we abuse $d_h^\pi(\cdot)$ to denote either probability measure or density function.

[3] For completeness, we also provide a brief discussion for the related Linear mixture model [56] setting (Appendix B.4.2).

[4] This definition is a standard extension over the tabular MDPs by referencing the similar notions from the bandit literature, *i.e.* from *Multi-armed Bandit* to *Linear Bandit* [96].

Offline learning setting. Offline RL requires the agent to learn the policy π that maximizes v^{π}, provided with the historical data $\mathcal{D} = \left\{ \left(s_h^\tau, a_h^\tau, r_h^\tau, s_{h+1}^\tau \right) \right\}_{\tau \in [K]}^{h \in [H]}$ rolled out from some behavior policy μ. The offline nature requires we cannot change μ and in particular we do not know the data generating distribution of μ. To sum up, the agent seeks to find a policy π_{alg} such that $v^\star - v^{\pi_{\text{alg}}} \leq \epsilon$ for the given batch data $\mathcal{D}$ and a given targeted accuracy $\epsilon > 0$.

3.2.1 Assumptions

It is known that learning a near-optimal policy from the offline data $\mathcal{D}$ cannot be sample efficient without certain data-coverage assumptions [90, 5]. To begin with, we define the population covariance matrix under the behavior policy μ for all $h \in [H]$:

$$\Sigma_h^p := \mathbb{E}_{\mu,h} \left[\phi(s,a)\phi(s,a)^\top \right], \tag{3.1}$$

since Σ_h^p measure the coverage of state-action space for data $\mathcal{D}$, we make the following assumption.

Assumption 3.2.1 (Feature Coverage)**.** *The data distributions μ satisfy the minimum eigenvalue condition: $\forall h \in [H]$, $\kappa_h := \lambda_{\min}(\Sigma_h^p) > 0$ and denote $\kappa = \min_h \kappa_h$. Note κ is a system-dependent (non-universal) quantity as it is upper bounded by $1/d$ (Assumption 2 in [90]).*

We make this assumption for the following reasons. First of all, our offline learning guarantee (Theorem 3.3.1) provides simultaneously comparison to all the policies, which is stronger than only competing with the optimal policy (whereas relaxed assumption suffices, for example $\sup_{x \in \mathbb{R}^d} \frac{x\Sigma_{\pi^\star} x^\top}{x\Sigma_\mu x^\top} < \infty$ [38]). As a consequence, the behavior distribution μ must be able to explore each feature dimension for the result to be valid.

Even if Assumption 4.2.3 does not hold, we can always restrict our algorithmic design to the *effective subspan* of Σ_h^p, which causes the alternative notion of $\kappa := \min_{h \in [H]} \{ \kappa_h : s.t.\ \kappa_h =$

smallest positive eigenvalue at time h} (see Appendix B.4.1 for detailed discussions). In this scenario, learning the optimal policy cannot be guaranteed as a constant suboptimality gap needs to be suffered due to the lack of coverage and this is formed as *assumption-free RL* in [5]. Lastly, previous works analyzing the linear MDPs impose very similar assumptions, *e.g.* [40] Theorem 3.2 where Σ_D^{-1} exists and [82] for the OPE problem.

Next, for any function $V_{h+1}(\cdot) \in [0, H-h]$, we define the conditional variance $\sigma_{V_{h+1}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$ as $\sigma_{V_{h+1}}(s,a)^2 := \max\{1, \mathrm{Var}_{P_h}(V_{h+1})(s,a)\}$.[5] Based on this definition, we can define the variance-involved population covariance matrices as:

$$\Lambda_h^p := \mathbb{E}_{\mu,h}\left[\sigma_{V_{h+1}}(s,a)^{-2}\phi(s,a)\phi(s,a)^\top\right].$$

In particular, when $V_h = V_h^\star$, we use the notation $\Lambda_h^{\star p}$ instead. Since $\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sigma_{V_h}(s,a)^2 \le H^2$, then by Assumption 4.2.3 we directly have the following corollary.

Corollary 3.2.1. *Define* $\iota_h := \lambda_{\min}(\Lambda_h^p)$, $\iota := \min_h \iota_h$. *Then* $\iota_h \ge \frac{\kappa_h}{H^2} > 0\ \forall h \in [H]$, *and* $\iota > 0$.

3.3 Algorithm and Main Results

Least square regression is usually considered as one of the "default" tools for handling problems with linear structures (*e.g.* LinUCB algorithm for linear Bandits) and finds its popularity in RL as well since *Least-Square Value Iteration* (LSVI, [71]) is shown to be provably efficient for linear MDPs, due to that $V_{h+1}(s')$ is an unbiased estimator of $[P_h V_{h+1}](s,a)$. Concretely, it solves the ridge regression problems at each time steps (with $\lambda > 0$ being the regularization parameter):

$$\widehat{w}_h := \operatorname*{argmin}_{w\in\mathbb{R}^d} \lambda\|w\|_2^2 + \sum_{k=1}^{K}\left[\langle\phi(s_h^k, a_h^k), w\rangle - r_h^k - V_{h+1}(s_{h+1}'^k)\right]^2 \tag{3.2}$$

[5] The $\max(1,\cdot)$ applied here is for technical reason only. In general, it suffices to think $\sigma^2_{V_{h+1}} \approx \mathrm{Var}_h V_{h+1}$.

and has the closed-form solution $\widehat{w}_h = \Sigma_h^{-1} \sum_{k=1}^K \phi(s_h^k, a_h^k)[r_{k,h} + V_{h+1}(s_h'^k)]$ with the cumulative sample covariance $\Sigma_h^{-1} = \sum_{k=1}^K \phi(s_h^k, a_h^k)\phi(s_h^k, a_h^k)^\top + \lambda I$. In offline RL, this has also been leveraged in *pessimistic value iteration* [80] and *fitted Q-evaluation* [95]. Nevertheless, LSVI could only yield suboptimal guarantees, as illustrated by the following example.

Example. Instantiate PEVI (Theorem 4.4 in [80]) with $\phi(s, a) = \mathbf{1}_{s,a}$ (*i.e.* tabular MDPs)[6], by direct calculation the learning bound has the form $O(dH \cdot \sum_{h,s,a} d_h^{\pi^\star}(s, a)\sqrt{\frac{1}{K \cdot d_h^\mu(s,a)}})$ and the optimal result ([5] Theorem 4.1) gives $O(\sum_{h,s,a} d_h^{\pi^\star}(s, a)\sqrt{\frac{\mathrm{Var}_{P_{s,a}}(r+V_{h+1}^\star)}{K \cdot d_h^\mu(s,a)}})$. The former has the horizon dependence H^2 and the latter is $H^{3/2}$ by law of total variance.

Motivation. By comparing the above two expressions, it can be seen that PEVI cannot get rid of the explicit H factor due to missing the variance information (*w.r.t* $V^\star$). If we go deeper, one could find that it might not be all that ideal to put equal weights on all the training samples in the least square objective (3.2), since, unlike linear regression where the randomness coming from one source distribution, we are regressing over a sequence of distributions in RL (*i.e.* each s_h, a_h corresponds to a different distribution $P(\cdot|s_h, a_h)$ and there are possibly infinite many of them). Therefore, conceptually, the sample piece (s_h, a_h, s_{h+1}) that has higher variance distribution $P(\cdot|s_h, a_h)$ tends to be less "reliable" than the one (s_h', a_h', s_{h+1}') with lower variance (hence should not have equal weight in (3.2)). This suggests reweighting scheme might help improve the learning guarantee and reweighting over the variance of the value function stands as a natural choice.

3.3.1 Variance-Aware Pessimistic Value Iteration

Now we explain our framework that incorporates the variance information. Our design is motivated by the previous [76] (for online learning) and [82] (for policy evaluation). By the offline nature, we can use the independent episodic data $\mathcal{D}' = \{(\bar{s}_h^\tau, \bar{a}_h^\tau, \bar{r}_h^\tau, \bar{s}_h^{\tau\prime})\}_{\tau\in[K]}^{h\in[H]}$ (from μ)

[6]This provides a valid illustration since tabular MDP is a special case of linear MDPs.

to estimate the conditional variance of any V-values V_{h+1} via the definition $[\mathrm{Var}_h V_{h+1}](s,a) = [P_h(V_{h+1})^2](s,a) - ([P_h V_{h+1}](s,a))^2$. For the second order moment, by Definition 3.2.1, it holds

$$\left[P_h V_{h+1}^2\right](s,a) = \int_{\mathcal{S}} V_{h+1}^2\left(s'\right)\ \mathrm{d}P_h\left(s' \mid s,a\right) = \phi(s,a)^\top \int_{\mathcal{S}} V_{h+1}^2\left(s'\right)\ \mathrm{d}\nu_h\left(s'\right).$$

Denote $\beta_h := \int_{\mathcal{S}} V_{h+1}^2(s')\ \mathrm{d}\nu_h(s')$, then $P_h V_{h+1}^2 = \langle \phi, \beta_h \rangle$ and we can estimator it via:

$$\bar{\beta}_h = \operatorname*{argmin}_{\beta\in\mathbb{R}^d} \sum_{k=1}^K \left[\left\langle \phi(\bar{s}_h^k, \bar{a}_h^k), \beta \right\rangle - V_{h+1}^2\left(\bar{s}_{h+1}^k\right)\right]^2 + \lambda\|\beta\|_2^2 = \bar{\Sigma}_h^{-1} \sum_{k=1}^K \phi(\bar{s}_h^k, \bar{a}_h^k) V_{h+1}^2\left(\bar{s}_{h+1}^k\right)$$

and, similarly, the first order moment $P_h V_{h+1} := \langle \phi, \theta_h \rangle$ can be estimated via:

$$\bar{\theta}_h = \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \sum_{k=1}^K \left[\left\langle \phi(\bar{s}_h^k, \bar{a}_h^k), \theta \right\rangle - V_{h+1}\left(\bar{s}_{h+1}^k\right)\right]^2 + \lambda\|\theta\|_2^2 = \bar{\Sigma}_h^{-1} \sum_{k=1}^K \phi(\bar{s}_h^k, \bar{a}_h^k) V_{h+1}\left(\bar{s}_{h+1}^k\right)$$

The final estimator is defined as $\widehat{\sigma}_{V_h}^2(\cdot,\cdot) := \max\{1, \widehat{\mathrm{Var}}_h V_{h+1}(\cdot,\cdot)\}$ with the estimated conditional variance $\widehat{\mathrm{Var}}_h V_{h+1}(\cdot,\cdot) = \langle \phi(\cdot,\cdot), \bar{\beta}_h \rangle_{[0,(H-h+1)^2]} - \left[\langle \phi(\cdot,\cdot), \bar{\theta}_h \rangle_{[0,H-h+1]}\right]^2$.[7] In particular, when setting $V_{h+1} = \widehat{V}_{h+1}$, it recovers $\widehat{\sigma}_h$ in Algorithm 2 line 8. Here $\bar{\Sigma}_h = \sum_{\tau=1}^K \phi(\bar{s}_h^\tau, \bar{a}_h^\tau)\phi(\bar{s}_h^\tau, \bar{a}_h^\tau)^\top + \lambda I_d$.

Variance-weighted LSVI. The idea of LSVI (3.2) is based on approximate the Bellman updates: $\mathcal{T}_h(V)(s,a) = r_h(s,a) + (P_h V)(s,a)$. With variance estimator $\widehat{\sigma}_h$ at hand, we can modify (3.2) to solve the variance-weighted LSVI instead (Line 10 of Algorithm 2)

$$\widehat{\omega}_h := \operatorname*{argmin}_{w\in\mathbb{R}^d} \lambda\|w\|_2^2 + \sum_{k=1}^K \frac{\left[\langle \phi(s_h^k, a_h^k), w \rangle - r_h^k - \widehat{V}_{h+1}(s_{h+1}'^k)\right]^2}{\widehat{\sigma}_h^2(s_h^k, a_h^k)} = \widehat{\Lambda}_h^{-1} \sum_{k=1}^K \frac{\phi\left(s_h^k, a_h^k\right) \cdot \left[r_h^k + \widehat{V}_{h+1}\left(s_{h+1}^k\right)\right]}{\widehat{\sigma}^2(s_h^k, a_h^k)}$$

where $\widehat{\Lambda}_h = \sum_{k=1}^K \phi(s_h^k, a_h^k)\phi(s_h^k, a_h^k)^\top / \widehat{\sigma}_h^2(s_h^k, a_h^k) + \lambda I_d$. The estimated Bellman update $\widehat{\mathcal{T}}_h$ (acts on $\widehat{V}_{h+1}$) is defined as: $(\widehat{\mathcal{T}}_h \widehat{V}_{h+1})(\cdot,\cdot) = \phi(\cdot,\cdot)^\top \widehat{\omega}_h$ and the pessimism Γ_h is assigned to update $\widehat{Q}_h \approx \widehat{\mathcal{T}}_h \widehat{V}_{h+1} - \Gamma_h$, *i.e.* Bellman update + Pessimism (Line 10-12 in Algorithm 2).

[7] The truncation used here is a standard treatment for making the estimator to be within the valid range.

Algorithm 2 **V**ariance-**A**ware **P**essimistic **V**alue **I**teration (VAPVI)

Input: Dataset $\mathcal{D} = \left\{\left(s_h^\tau, a_h^\tau, r_h^\tau\right)\right\}_{\tau,h=1}^{K,H}$ $\mathcal{D}' = \left\{\left(\bar{s}_h^\tau, \bar{a}_h^\tau, \bar{r}_h^\tau\right)\right\}_{\tau,h=1}^{K,H}$. Universal constant C.
Initialization: Set $\widehat{V}_{H+1}(\cdot) \leftarrow 0$.
for $h = H, H-1, \ldots, 1$ **do**
  Set $\bar{\Sigma}_h \leftarrow \sum_{\tau=1}^K \phi(\bar{s}_h^\tau, \bar{a}_h^\tau)\phi(\bar{s}_h^\tau, \bar{a}_h^\tau)^\top + \lambda I$
  Set $\bar{\beta}_h \leftarrow \bar{\Sigma}_h^{-1} \sum_{\tau=1}^K \phi(\bar{s}_h^\tau, \bar{a}_h^\tau) \cdot \widehat{V}_{h+1}(\bar{s}_{h+1}^\tau)^2$
  Set $\bar{\theta}_h \leftarrow \bar{\Sigma}_h^{-1} \sum_{\tau=1}^K \phi(\bar{s}_h^\tau, \bar{a}_h^\tau) \cdot \widehat{V}_{h+1}(\bar{s}_{h+1}^\tau)$
  Set $\left[\widehat{\mathrm{Var}}_h \widehat{V}_{h+1}\right](\cdot,\cdot) = \left\langle \phi(\cdot,\cdot), \bar{\beta}_h \right\rangle_{[0,(H-h+1)^2]} - \left[\left\langle \phi(\cdot,\cdot), \bar{\theta}_h \right\rangle_{[0,H-h+1]}\right]^2$
  Set $\widehat{\sigma}_h(\cdot,\cdot)^2 \leftarrow \max\{1, \widehat{\mathrm{Var}}_{P_h} \widehat{V}_{h+1}(\cdot,\cdot)\}$
  Set $\widehat{\Lambda}_h \leftarrow \sum_{\tau=1}^K \phi\left(s_h^\tau, a_h^\tau\right) \phi\left(s_h^\tau, a_h^\tau\right)^\top / \widehat{\sigma}^2(s_h^\tau, a_h^\tau) + \lambda \cdot I$,
  Set $\widehat{\omega}_h \leftarrow \widehat{\Lambda}_h^{-1} \left(\sum_{\tau=1}^K \phi\left(s_h^\tau, a_h^\tau\right) \cdot \left(r_h^\tau + \widehat{V}_{h+1}\left(s_{h+1}^\tau\right)\right) / \widehat{\sigma}^2(s_h^\tau, a_h^\tau)\right)$
  Set $\Gamma_h(\cdot,\cdot) \leftarrow C\sqrt{d} \cdot \left(\phi(\cdot,\cdot)^\top \widehat{\Lambda}_h^{-1} \phi(\cdot,\cdot)\right)^{1/2} + \frac{2H^3\sqrt{d}}{K}$ (Use Γ_h^I for the improved version)
  Set $\bar{Q}_h(\cdot,\cdot) \leftarrow \phi(\cdot,\cdot)^\top \widehat{\omega}_h - \Gamma_h(\cdot,\cdot)$
  Set $\widehat{Q}_h(\cdot,\cdot) \leftarrow \min\left\{\bar{Q}_h(\cdot,\cdot), H-h+1\right\}^+$
  Set $\widehat{\pi}_h(\cdot \mid \cdot) \leftarrow \arg\max_{\pi_h} \left\langle \widehat{Q}_h(\cdot,\cdot), \pi_h(\cdot \mid \cdot)\right\rangle_{\mathcal{A}}$, $\widehat{V}_h(\cdot) \leftarrow \max_{\pi_h} \left\langle \widehat{Q}_h(\cdot,\cdot), \pi_h(\cdot \mid \cdot)\right\rangle_{\mathcal{A}}$
end for
Output: $\left\{\widehat{\pi}_h\right\}_{h=1}^H$.

Tighter Pessimistic Design. To improve the learning guarantee, we create a tighter penalty design that includes $\widehat{\Lambda}_h^{-1}$ rather than $\bar{\Sigma}_h^{-1}$ and an extra higher order $O(\frac{1}{K})$ term:

$$\Gamma_h \leftarrow O\left(\sqrt{d} \cdot (\phi(\cdot,\cdot)^\top \widehat{\Lambda}_h^{-1} \phi(\cdot,\cdot))^{1/2}\right) + \frac{2H^3\sqrt{d}}{K}$$

Note such a design admits no explicit factor in H in the main term (as opposed to [80]) therefore is the key for achieving adaptive/problem-dependent results (as we shall discuss later). The full algorithm VAPVI is stated in Algorithm 2. In particular, we halve the offline data into two independent parts with $\mathcal{D} = \{(s_h^\tau, a_h^\tau, r_h^\tau, s_h^{\tau\prime})\}_{\tau\in[K]}^{h\in[H]}$ and $\mathcal{D}' = \{(\bar{s}_h^\tau, \bar{a}_h^\tau, \bar{r}_h^\tau, \bar{s}_h^{\tau\prime})\}_{\tau\in[K]}^{h\in[H]}$ for different purposes (estimating variance and updating Q-values).

3.3.2 Main result

We denote quantities $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4$ as in the Notation List. Then VAPVI provides the following result. The complete proof is provided in Appendix B.1.

Theorem 3.3.1. *Let K be the number of episodes. If $K > \max\{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4\}$ and $\sqrt{d} > \xi$, where $\xi := \sup_{V\in[0,H],\, s'\sim P_h(s,a),\, h\in[H]} \left|\frac{r_h+V(s')-(\mathcal{T}_hV)(s,a)}{\sigma_V(s,a)}\right|$. Then for any $0 < \lambda < \kappa$, with probability $1-\delta$, for all policy π simultaneously, the output $\widehat{\pi}$ of Algorithm 2 satisfies*

$$v^{\pi} - v^{\widehat{\pi}} \leq \widetilde{O}\Big(\sqrt{d}\cdot\sum_{h=1}^{H}\mathbb{E}_{\pi}\left[\sqrt{\phi(\cdot,\cdot)^{\top}\Lambda_h^{-1}\phi(\cdot,\cdot)}\right]\Big) + \frac{2H^4\sqrt{d}}{K}$$

where $\Lambda_h = \sum_{k=1}^{K}\frac{\phi(s_h^k,a_h^k)\cdot\phi(s_h^k,a_h^k)^{\top}}{\sigma^2_{\widehat{V}_{h+1}(s_h^k,a_h^k)}} + \lambda I_d$. In particular, we have with probability $1-\delta$,

$$v^{\star} - v^{\widehat{\pi}} \leq \widetilde{O}\Big(\sqrt{d}\cdot\sum_{h=1}^{H}\mathbb{E}_{\pi^{\star}}\left[\sqrt{\phi(\cdot,\cdot)^{\top}\Lambda_h^{\star-1}\phi(\cdot,\cdot)}\right]\Big) + \frac{2H^4\sqrt{d}}{K} \tag{3.3}$$

where $\Lambda_h^{\star} = \sum_{k=1}^{K}\frac{\phi(s_h^k,a_h^k)\cdot\phi(s_h^k,a_h^k)^{\top}}{\sigma^2_{V^{\star}_{h+1}(s_h^k,a_h^k)}} + \lambda I_d$ and $\widetilde{O}$ hides universal constants and the Polylog terms.

Theorem 3.3.1 provides improvements over the existing best-known results and we now explain it. However, before that, we first discuss about our theorem condition.

Comparing to [76]. In the online regime, [76] is the first result that achieves optimal regret rate with $O(dH\sqrt{T})$ in the linear (mixture) MDPs. However, this result requires the condition $d \geq H$ (their Theorem 6 and Remark 7). In offline RL, VAPVI only requires a milder condition $\sqrt{d} > \xi$ comparing to $d \geq H$ (since for any fixed $V \in [0, H]$, the standardized quantity $\frac{r+V(s')-(\mathcal{T}_hV)(s,a)}{\sigma_V(s,a)}$ is bounded by constant with high probability, *e.g.* by *chebyshev* inequality), which makes our result apply to a wider range of linear MDPs.

Comparing to [80]. [80] first shows *pessimistic value iteration* (PEVI) is provably efficient for Linear MDPs in offline RL. VAPVI improves PEVI over $O(\sqrt{d})$ on the feature dimension,

and improves the horizon dependence as $\Lambda_h \succcurlyeq \frac{1}{H^2}\Sigma_h$ implies $\Lambda_h^{-1} \preccurlyeq H^2\Sigma_h^{-1}$. In addition, when instantiate to the tabular case, *i.e.* $\phi(s,a) = \mathbf{1}_{s,a}$, VAPVI gives $O(\sqrt{d}\sum_{h,s,a} d_h^{\pi^\star}(s,a)\sqrt{\frac{\mathrm{Var}_{P_{s,a}}(r+V_{h+1}^\star)}{K\cdot d_h^\mu(s,a)}})$, which enjoys $O(\sqrt{H})$ improvement over PEVI and the order $O(H^{3/2})$ is tight (check Section B.4 for the detailed derivation).

Comparing to [40]. Their linear MDP guarantee in Theorem 3.2. enjoys the same rate as VAPVI in feature dimension but the horizon dependence is essentially the same as [80] (by translating $H \approx O(\frac{1}{1-\gamma})$) therefore is not optimal. The general function approximation scheme in [40] provides elegant characterizations for on-support error and off-support error, but the algorithmic framework is information-theoretical only (and the practical version PSPI will not yield the same learning guarantee). Also, due to the use finite function class and policy class, the reduction to linear MDP only works with finite action space. As a comparison, VAPVI has no constraints on any of these.

Comparing to [41]. Concurrently, [41] considers offline RL with the linear Bellman complete model, which is more general than linear MDPs and, with the assumption $Q^\pi \le 1$, their PACLE algorithm provides near-minimax optimal guarantee in this setting. However, when recovering to the standard setting $Q^\pi \in [0, H]$, their bound will rescale by an H factor,[8] which could be suboptimal due to the variance-unawareness. The reason behind this is: when $Q^\pi \le 1$, lack of variance information encoding will not matter, since in this case $\mathrm{Var}_P(V^\pi) \le 1$ has constant order (therefore will not affect the optimal rate); when $Q^\pi \in [0, H]$, $\mathrm{Var}_P(V^\pi)$ can be as large as H^2, effectively leveraging the variance information can help improve the sample efficiency, *e.g.* via *law of total variances*, just like VAPVI does. On the other hand, their guarantee also requires finite action space, due to the mirror descent style analysis. Nevertheless, we do point out [41] has improved state-action measure than VAPVI, as $\left\|E_\pi[\phi(\cdot,\cdot)]\right\|_{M^{-1}} \le \mathbb{E}_\pi[\|\phi(\cdot,\cdot)\|_{M^{-1}}]$ by Jensen's inequality and that norm $\|\cdot\|_{M^{-1}}$ is convex for some positive-definite matrix M.

[8] Check their Footnote 2 in Page 9.

Adaptive characterization and faster convergence. Comparing to existing works, one major improvement is that the main term for VAPVI $\sqrt{d}\sum_{h=1}^{H}\mathbb{E}_{\pi^\star}\left[\sqrt{\phi(\cdot,\cdot)^\top\Lambda_h^{\star-1}\phi(\cdot,\cdot)}\right]$ admits no explicit dependence on H, which provides a more adaptive/instance-dependent characterization. For instance, if we ignore the technical treatment by taking $\lambda = 0$ and $\sigma_h^\star \approx \mathrm{Var}_P(V_{h+1}^\star)$, then for the **partially deterministic systems** (where there are t stochastic P_h's and $H-t$ deterministic P_h's), the main term diminishes to $\sqrt{d}\sum_{i=1}^{t}\mathbb{E}_{\pi^\star}\left[\sqrt{\phi(\cdot,\cdot)^\top\Lambda_{h_i}^{\star-1}\phi(\cdot,\cdot)}\right]$ with $h_i \in \{h : \ s.t.\ P_h \text{ is stochastic}\}$ and can be a much smaller quantity when $t \ll H$. Furthermore, for the **fully deterministic system**, VAPVI automatically provides faster convergence rate $O(\frac{1}{K})$ from the higher order term, given that the main term degenerates to 0. Those adaptive/instance-dependent features are not enjoyed by [40, 41], as they always provide the standard statistical rate $O(\frac{1}{\sqrt{K}})$ (also check Remark 7 for a related discussion).

3.3.3 VAPVI-Improved: Further improvement in state-action dimension

Can we further improve the VAPVI? Indeed, by deploying a carefully tuned tighter penalty, we are able to further improve the state-action dependence if the feature is non-negative ($\phi \geq 0$). Concretely, we replace the following Γ_h^I in Algorithm 2 instead, and call the algorithm VAPVI-Improved (or VAPVI-I for short). The proof can be found in Appendix B.2.

$$\Gamma_h^I(s,a) \leftarrow \phi(s,a)^\top\left|\widehat{\Lambda}_h^{-1}\sum_{\tau=1}^{K}\frac{\phi\left(s_h^\tau,a_h^\tau\right)\cdot\left(r_h^\tau+\widehat{V}_{h+1}\left(s_{h+1}^\tau\right)-\left(\widehat{\mathcal{T}}_h\widehat{V}_{h+1}\right)\left(s_h^\tau,a_h^\tau\right)\right)}{\widehat{\sigma}_h^2(s_h^\tau,a_h^\tau)}\right|+\widetilde{O}(\frac{H^3d/\kappa}{K}) \tag{3.4}$$

Theorem 3.3.2. *Suppose the feature is non-negative ($\phi \geq 0$). Let K be the number of episodes. If $K > \max\{\mathcal{M}_1,\mathcal{M}_2,\mathcal{M}_3,\mathcal{M}_4\}$ and $\sqrt{d} > \xi$. Deploying Γ_h^I (3.4) in Algorithm 2. Then for any $0 < \lambda < \kappa$, with probability $1-\delta$, for all policy π simultaneously, the output $\widehat{\pi}$ of Algorithm 2*

(VAPVI-I) satisfies

$$v^{\pi} - v^{\hat{\pi}} \leq \widetilde{O}\big(\sqrt{d} \cdot \sum_{h=1}^{H} \sqrt{\mathbb{E}_{\pi}[\phi(\cdot,\cdot)]^{\top}\Lambda_h^{-1}\mathbb{E}_{\pi}[\phi(\cdot,\cdot)]}\big) + \widetilde{O}(\frac{H^4 d/\kappa}{K})$$

In particular, when choosing $\pi = \pi^{\star}$, the above guarantee holds true with Λ_h^{-1} replaced by $\Lambda_h^{\star-1}$. Here Λ_h^{-1}, $\Lambda_h^{\star-1}$, ξ are defined the same as Theorem 3.3.1.

Theorem 3.3.2 maintains nearly all the features of Theorem 3.3.1 (except higher order term is slightly worse) and the dominate term evolves from $\mathbb{E}_{\pi}\|\phi\|_{\Lambda_h^{-1}}$ to $\big\|\mathbb{E}_{\pi}[\phi]\big\|_{\Lambda_h^{-1}}$. Clearly, the two bounds differ by the magnitude of Jensen's inequality. To provide a concrete view of how much improvement is made, we check the parameter dependence in the context of tabular MDPs (where we ignore the higher order term for conciseness). In particular, we compare the results under the single-policy concentrability.

Assumption 3.3.3 ([18, 19])**.** *There exists a optimal policy $\pi^{\star}$, s.t. $\sup_{h,s,a} d_h^{\pi^{\star}}(s,a)/d_h^{\mu}(s,a) := C^{\star} < \infty$, where d^{π} is the marginal state-action probability under π.*

In tabular RL, $\phi(s,a) = \mathbf{1}_{s,a}$ and $d = S \cdot A$ (S, A be the finite state, action cardinality), then

$$\begin{aligned}
\text{Theorem 3.3.1} &\to \sqrt{SA}\sum_{h}^{H}\sum_{s,a} d_h^{\pi^{\star}}(s,a)\sqrt{\frac{\mathrm{Var}_{P_{s,a}}(r + V_{h+1}^{\star})}{K \cdot d_h^{\mu}(s,a)}} \leq \sqrt{\frac{H^3 C^{\star} S^2 A}{K}};\\
\text{Theorem 3.3.2} &\to \sqrt{SA}\sum_{h}^{H}\sqrt{\sum_{s,a} d_h^{\pi^{\star}}(s,a)^2\frac{\mathrm{Var}_{P_{s,a}}(r + V_{h+1}^{\star})}{K \cdot d_h^{\mu}(s,a)}} \leq \sqrt{\frac{H^3 C^{\star} SA}{K}}.
\end{aligned} \tag{3.5}$$

Theorem 3.3.2 enjoys a S state improvement over Theorem 3.3.1 and nearly recovers the minimax rate $\sqrt{\frac{H^3 C^{\star} S}{K}}$ [19]. The detailed derivation can be found in Appendix B.4. Also, to show our result is near-optimal, we provide the corresponding lower bound. The proof is in Appendix B.3.

Theorem 3.3.4 (Minimax lower bound)**.** *There exist a pair of universal constants $c, c' > 0$ such that given dimension d, horizon H and sample size $K > c'd^3$, one can always find a family of linear MDP instances $\mathcal{M}$ such that (where $\Lambda_h^\star = \sum_{k=1}^K \frac{\phi(s_h^k,a_h^k)\cdot\phi(s_h^k,a_h^k)^\top}{\mathrm{Var}_h(V_{h+1}^\star)(s_h^k,a_h^k)}$ satisfies $(\Lambda_h^\star)^{-1}$ exists and $\mathrm{Var}_h(V_{h+1}^\star)(s_h^k,a_h^k) > 0\ \forall M \in \mathcal{M}$)*

$$\inf_{\hat{\pi}} \sup_{M\in\mathcal{M}} \mathbb{E}_M\left[v^\star - v^{\hat{\pi}}\right] \Big/ \left(\sqrt{d}\cdot\sum_{h=1}^H \sqrt{\mathbb{E}_{\pi^\star}[\phi]^\top(\Lambda_h^\star)^{-1}\mathbb{E}_{\pi^\star}[\phi]}\right) \geq c. \tag{3.6}$$

Theorem 3.3.4 nearly matches the main term in VAPVI-I (Theorem 3.3.2) and certifies it is near-optimal. On the other hand, it is worth understanding how the above lower bound compares to the lower bound in [80]. In general, they are not directly comparable since both results are global minimax (not instance-dependent/local-minimax) lower bounds as the hardness only hold for a family of hard instances (which makes comparison outside of the family instances vacuum). However, for all the instances within the family, we can verify our lower bound 3.3.4 is tighter (see Appendix B.3.6 for detailed discussion).

3.4 Proof Overview

In this section, we provide a brief overview of the key proving ideas of the theorems. We begin with Theorem 3.3.1. First, by *the extended value difference lemma* (Lemma D.0.7), we can convert bounding the suboptimality gap of $v^\star - v^{\hat{\pi}}$ to bounding $\sum_{h=1}^H 2\cdot\mathbb{E}_\pi\left[\Gamma_h(s_h,a_h)\right]$, given that $|(\mathcal{T}_h\widehat{V}_{h+1} - \widehat{\mathcal{T}}_h\widehat{V}_{h+1})(s,a)| \leq \Gamma_h(s,a)$ for all s,a,h. To bound $\mathcal{T}_h\widehat{V}_{h+1} - \widehat{\mathcal{T}}_h\widehat{V}_{h+1}$, by decomposing it reduces to bounding the key quantity

$$\phi(s,a)^\top\widehat{\Lambda}_h^{-1}\left[\sum_{\tau=1}^K \phi\left(s_h^\tau,a_h^\tau\right)\cdot\left(r_h^\tau + \widehat{V}_{h+1}\left(s_{h+1}^\tau\right) - \left(\mathcal{T}_h\widehat{V}_{h+1}\right)\left(s_h^\tau,a_h^\tau\right)\right)/\widehat{\sigma}_h^2(s_h^\tau,a_h^\tau)\right] \tag{3.7}$$

The term is treated in two steps. First, we bound the gap of $\left\|\sigma^2_{\widehat{V}_{h+1}} - \widehat{\sigma}^2_h\right\|$ so we can convert $\widehat{\sigma}^2_h$ to $\sigma^2_{\widehat{V}_{h+1}}$. Next, since Var $\left[r^\tau_h + \widehat{V}_{h+1}\left(s^\tau_{h+1}\right) - \left(\mathcal{T}_h\widehat{V}_{h+1}\right)\left(s^\tau_h, a^\tau_h\right) \mid s^\tau_h, a^\tau_h\right] \approx \sigma^2_{\widehat{V}_{h+1}}$, therefore by the variance-weighted scheme in ((3.7)), we can leverage the recent technical development *Bernstein inequality for self-normalized martingale* (Lemma C.11.4) for acquiring the tight result, in contrast to the previous treatment of Hoeffding inequality for self-normalized martingale + Covering.[9] For the second part, one needs to further convert $\sigma^2_{\widehat{V}_{h+1}}$ to $\sigma^{\star 2}_h$ (Λ^{-1}_h to $\Lambda^{\star -1}_h$) with appropriate concentrations. The proof of Theorem 3.3.2 is similar but with more complicated computations and relies on using the linear representation of ϕ in Γ^I_h (3.4), so that the expectation over π is inside the square root by taking expectation over the linear representation at the beginning. The lower bound proof uses a simple modification of [41] which consists of the reduction from learning to testing with Assouad's method, and the use of standard information inequalities (*e.g.* from total variation to KL divergence). For completeness, we provide the full proof in Appendix B.3.

3.5 Conclusion

This chapter studies offline RL with linear MDP representation and contributes *Variance Aware Pessimistic Value Iteration* (VAPVI) which adopts the conditional variance information of the value function. VAPVI uses the estimated variances to reweight the Bellman residuals in the least-square pessimistic value iteration and provides improved offline learning bounds over the existing best-known results. VAPVI-I further improves over VAPVI in the state-action dimension and is near-minimax optimal. One highlight of the theorems is that our learning bounds are expressed in terms of system quantities, which automatically provide natural instance-dependent characterizations that previous results are short of.

[9] Variance-reweighting in (3.7) is important, since applying *Bernstein inequality for self-normalized martingale* (Lemma C.11.4) without variance-reweighting cannot provide any improvement.

On the other hand, while VAPVI/VAPVI-I close the existing gap from previous literature [80, 40], the optimal guarantee is in the minimax sense. Although our upper bounds possess instance-dependent characterizations, the lower bound only holds true for a class of hard instances. In this sense, whether "instance-dependent optimality" can be achieved remains elusive in the current linear MDP setting (such a discussion is recently initiated in MAB problems [57]). Furthermore, removing the dependence on κ in the higher order terms (*e.g.* Theorem 3.3.2) is challenging and the recent development [97] using robust estimation has the potential to address this issue. We leave these as future works.

Chapter 4

Provably Efficient Offline Reinforcement Learning with Differentiable Function Approximation

State-Of-The-Art offline reinforcement learning algorithms usually leverage powerful function approximators (*e.g.* neural networks) to alleviate the sample complexity hurdle for better empirical performances. Despite the successes, a more systematic understanding of the statistical complexity for function approximation remains lacking. Towards bridging the gap, in this Chapter we study offline reinforcement learning with *differentiable function class approximation* (DFA). This function class naturally incorporates a wide range of models with nonlinear/nonconvex structures. Most importantly, we show offline RL with differentiable function approximation is provably efficient by analyzing the *pessimistic fitted Q-learning* (PFQL) algorithm, and our results provide the theoretical basis for understanding a variety of practical heuristics that rely on Fitted Q-Iteration style design. In addition, we further improve our guarantee with a tighter instance-dependent characterization.

4.1 Introduction, Related Work, and Our Contribution

Offline reinforcement learning [11, 10] refers to the paradigm of learning a policy in the sequential decision making problems, where only the logged data are available and were collected from an unknown environment (*Markov Decision Process* / MDP). Inspired by the success of scalable supervised learning methods, modern reinforcement learning algorithms (*e.g.* [8]) incorporate high-capacity function approximators to acquire generalization across large state-action spaces and have achieved excellent performances along a wide range of domains. For instance, there are a huge body of deep RL-based algorithms that tackle challenging problems such as the game of Go and chess [8, 98], Robotics [99, 100], energy control [101] and Biology [102, 103]. Nevertheless, practitioners also noticed that algorithms with general function approximators can be quite data/sample inefficient, especially for deep neural networks where the models may require million of steps for tuning the large number of parameters they contain.[1]

On the other hand, statistical analysis has been actively conducted to understand the sample/statistical efficiency for reinforcement learning with function approximation, and fruitful results have been achieved under the respective model representations [79, 13, 68, 73, 74, 70, 71, 72, 75, 77, 78, 80, 76, 40, 82, 83, 105, 41, 106, 107, 108]. However, most works consider *linear* model approximators (*e.g.* linear (mixture) MDPs) or its variants. While the explicit linear structures make the analysis trackable (linear problems are easier to analyze), they are unable to reveal the sample/statistical complexity behaviors of practical algorithms that apply powerful function approximations (which might have complex structures).

In addition, there is an excellent line of works tackling provably efficient offline RL with general function approximation (*e.g.* [13, 40, 109]). Due to the generic function approximation class considered, those complexity bounds are usually expressed in the standard worst-case fashion $O(V_{\max}^2\sqrt{\frac{1}{n}})$ which lack the characterizations of individual instance behaviors. However,

[1] Check [104] and the references therein for an overview.

as mentioned in [52], practical reinforcement learning algorithms often perform far better than what these problem-independent bounds would suggest.

These observations motivate us to consider function approximation schemes that can help address the existing limitations. In particular, in this work we consider offline reinforcement learning with *differentiable function class* approximations. Its definition is given in below.

Definition 4.1.1 (Parametric Differentiable Function Class)**.** *Let $\mathcal{S}, \mathcal{A}$ be arbitrary state, action spaces and a feature map $\phi(\cdot,\cdot) : \mathcal{S} \times \mathcal{A} \to \Psi \subset \mathbb{R}^m$. The parameter space $\Theta \in \mathbb{R}^d$. Both Θ and Ψ are compact spaces. Then the parametric function class (for a model $f : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$) is defined as*

$$\mathcal{F} := \{f(\theta, \phi(\cdot,\cdot)) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}, \theta \in \Theta\}$$

that satisfies differentiability/smoothness condition: 1. for any $\phi \in \mathbb{R}^m$, $f(\theta, \phi)$ is third-time differentiable with respect to θ; 2. $f, \partial_\theta f, \partial^2_{\theta,\theta} f, \partial^3_{\theta,\theta,\theta} f$ are jointly continuous for (θ, ϕ).

Remark 3. *Differentiable Function Class was recently proposed for studying Off-Policy Evaluation (OPE) Problem [110] and we adopt it here for the policy learning task. Note by the compactness of Θ, Ψ and continuity, there exists constants $C_\Theta, B_\mathcal{F}, \kappa_1, \kappa_2, \kappa_3 > 0$ that bounds: $\|\theta\|_2 \le C_\Theta, |f(\theta, \phi(s,a))| \le B_\mathcal{F}, \|\nabla_\theta f(\theta, \phi(s,a))\|_2 \le \kappa_1, \|\nabla^2_{\theta\theta} f(\theta, \phi(s,a))\|_2 \le \kappa_2$, and $\|\nabla^3_{\theta\theta\theta} f(\theta, \phi(s,a))\|_2 \le \kappa_3$ for all $\theta \in \Theta,\ s, a \in \mathcal{S} \times \mathcal{A}$.*[2]

Why consider differentiable function class (Definition 4.1.1)? There are two main reasons why differentiable function class is worth studying for reinforcement learning.

- Due to the limitation of statistical tools, existing analysis in reinforcement learning usually favor basic settings such as *tabular MDPs* (where the state space and action space are finite [111, 44, 112, 113, 114, 30, 3, 16, 115, 17, 19, 116, 117, 118, 119]) or linear

[2]Here $\|\nabla^3_{\theta\theta\theta} f(\theta, \phi(s,a))\|_2$ is defined as the 2-norm for 3-d tensor and in the finite horizon setting we simply instantiate $\mathcal{B}_\mathcal{F} = H$.

MDPs [69, 71, 85, 80, 120, 90, 82] / linear Mixture MDPs [70, 56, 121, 122, 76] (where the transition dynamic admits linear structures) so that well-established techniques (*e.g.* from linear regression) can be applied. In addition, subsequent extensions are often based on linear models (*e.g.* Linear Bellman Complete models [75] and Eluder dimension [123, 77]). Differentiable function class strictly generalizes over the previous popular choices, *i.e.* by choosing $f(\theta, \phi) = \langle \theta, \phi \rangle$ or specifying ϕ to be one-hot representations, and is far more expressive as it encompasses nonlinear approximators.

- Practically speaking, the flexibility of selecting model f provides the possibility for handling a variety of tasks. For instance, when f is specified to be neural networks, θ corresponds to the weights of each network layers and $\phi(\cdot, \cdot)$ corresponds to the state-action representations (which is induced by the network architecture). When facing with easier tasks, we can deploy simpler model f such as polynomials. Yet, our statistical guarantee is not affected by the specific choices as we can plug the concrete form of model f into Theorem 4.3.2 to obtain the respective bounds (we do not need separate analysis for different tasks).

4.1.1 Related works

Reinforcement learning with function approximation. RL with function approximation has a long history that can date back to [124, 125]. Later, it draws significant interest for the finite sample analysis [71, 68]. Since then, people put tremendous efforts towards generalizing over linear function approximations and examples include Linear Bellman complete models [75], Eluder dimension [123, 77], linear deterministic $Q^\star$ [53] or Bilinear class [78]. While those extensions are valuable, the structure conditions assumed usually make the classes hard to track beyond the linear case. For example, the practical instances of Eluder Dimension are based on the linear-in-feature (or its transformation) representations (Section 4.1 of [53]). As a

Algorithm	Assumption	Suboptimality Gap $v^\star - v^{\widehat{\pi}}$
VFQL, Theorem 4.3.1	Concentrability 4.2.2	$\sqrt{C_{\text{eff}}}H \cdot \sqrt{\frac{H^2 d + \lambda C_\Theta^2}{K}} + \sqrt[4]{\frac{H^3 d \epsilon_{\mathcal{F}}}{K}} + \sqrt{C_{\text{eff}} H^3 \epsilon_{\mathcal{F}}} + H\epsilon_{\mathcal{F}}$
PFQL, Theorem 4.3.2	Uniform Coverage 4.2.3	$\sum_{h=1}^H 16 d H \cdot \mathbb{E}_{\pi^\star}\left[\sqrt{\nabla_\theta^\top f(\theta_h^\star, \phi(s_h, a_h)) \Sigma_h^{\star -1} \nabla_\theta f(\theta_h^\star, \phi(s_h, a_h))}\right]$
VAFQL, Theorem 4.4.1	Uniform Coverage 4.2.3	$16 d \cdot \sum_{h=1}^H \mathbb{E}_{\pi^\star}\left[\sqrt{\nabla_\theta^\top f(\theta_h^\star, \phi(s_h, a_h)) \Lambda_h^{\star -1} \nabla_\theta f(\theta_h^\star, \phi(s_h, a_h))}\right]$

Table 4.1: Suboptimality gaps for different algorithms with differentiable function class 4.1.1. Here we omit the higher order term for clear comparison. With Concentrability, we can only achieve the worst case bound that does not explicit depend on the function model f. With the stronger uniform coverage 4.2.3, better instance-dependent characterizations become available. Here C_{eff} is in 4.2.2, $\Sigma^\star$ in 4.3.2, $\Lambda^\star$ in 4.4.1 and $\epsilon_{\mathcal{F}}$ in 4.2.1.

comparison, differentiable function class contains a range of functions that are widely used in practical algorithms [126].

Offline RL with general function approximation (GFA). Another interesting thread of work considered offline RL with general function approximation [59, 13, 35, 40] which only imposes realizability and completeness/concentrability assumptions. The major benefit is that the function hypothesis can be arbitrary with no structural assumptions and it has been shown that offline RL with GFA is provably efficient. However, the generic form of functions in GFA makes it hard to go beyond worst-case analysis and obtain fine-grained instance-dependent learning bounds similar to those under linear cases. On the contrary, our results with DFA can be more problem adaptive by leveraging gradients and higher order information.

In addition to the above, there are more connected works. [110] first considers the differentiable function approximation (DFA) for the off-policy evaluation (OPE) task and builds the asymptotic theory, [127] analyzes the *deep Q-learning* with the specific ReLu activations, and [128] considers semi-parametric / nonparametric methods for offline RL (as opposed to our parametric DFA in 4.1.1). These are nice complementary studies to our work.

4.1.2 Our contribution

We provide the first Instance-dependent offline learning bound under non-linear function approximation. Informally, we show that (up to a lower order term) the natural complexity measure is proportional to $\sum_{h=1}^{H} \mathbb{E}_{\pi^\star,h}[\sqrt{g_\theta(s,a)^\top \Sigma_h^{-1} g_\theta(s,a)}]$ where $g_\theta(s,a) := \nabla f(\theta, \phi(s,a))$ is the gradient *w.r.t.* the parameter $\theta^\star$ at feature ϕ and $\Sigma_h = \sum_i g(s_{i,h}, a_{i,h}) g(s_{i,h}, a_{i,h})^\top$ is the Fisher information matrix of the observed data at $\widehat{\theta}$. This is achieved by analyzing the *pessimistic fitted Q-learning* (PFQL) algorithm (Theorem 4.3.2). In addition, we further analyze its variance-reweighting variant, which recovers the variance-dependent structure and can yield faster sample convergence rate. Last but not least, existing offline RL studies with tabular models, linear models and GLM models can be directly indicated by the appropriate choice of our model $\mathcal{F}$.

4.2 Preliminaries

Episodic Markov decision process. Let $M = (S, \mathcal{A}, P, r, H, d_1)$ to denote a finite-horizon *Markov Decision Process* (MDP), where S is the arbitrary state space and $\mathcal{A}$ is the arbitrary action space which can be infinite or continuous. The transition kernel $P_h : S \times \mathcal{A} \mapsto \Delta^S$ (Δ^S represents a distribution over states) maps each state action(s_h, a_h) to a probability distribution $P_h(\cdot|s_h, a_h)$ and P_h can be different for different h (time-inhomogeneous). H is the planning horizon and d_1 is the initial state distribution. Besides, $r : S \times A \mapsto \mathbb{R}$ is the mean reward function satisfying $0 \le r \le 1$. A policy $\pi = (\pi_1, \ldots, \pi_H)$ assigns each state $s_h \in S$ a probability distribution over actions by mapping $s_h \mapsto \pi_h(\cdot|s_h)\ \forall h \in [H]$ and induces a random trajectory $s_1, a_1, r_1, \ldots, s_H, a_H, r_H, s_{H+1}$ with $s_1 \sim d_1, a_h \sim \pi(\cdot|s_h), s_{h+1} \sim P_h(\cdot|s_h, a_h), \forall h \in [H]$.

Given a policy π, the V-value functions and state-action value function (Q-functions) $Q_h^\pi(\cdot,\cdot) \in \mathbb{R}^{S\times A}$ are defined as: $V_h^\pi(s) = \mathbb{E}_\pi[\sum_{t=h}^{H} r_t | s_h = s]$, $Q_h^\pi(s,a) = \mathbb{E}_\pi[\sum_{t=h}^{H} r_t | s_h, a_h = s, a]$, $\forall s, a, h \in$

$\mathcal{S}, \mathcal{A}, [H]$. The Bellman (optimality) equations follow $\forall h \in [H], s, a \in \mathcal{S} \times \mathcal{A}$:

$$Q_h^\pi(s,a) = r_h(s,a) + \int_{\mathcal{S}} V_{h+1}^\pi(\cdot) dP_h(\cdot|s,a), \quad V_h^\pi(s) = \mathbb{E}_{a\sim\pi_h(s)}[Q_h^\pi(s,a)],$$
$$Q_h^\star(s,a) = r_h(s,a) + \int_{\mathcal{S}} V_{h+1}^\star(\cdot) dP_h(\cdot|s,a), \quad V_h^\star(s) = \max_a Q_h^\star(s,a).$$

We define Bellman operator $\mathcal{P}_h$ for any function $V \in \mathbb{R}^S$ as $\mathcal{P}_h(V) = r_h + \int_{\mathcal{S}} V dP_h$, then $\mathcal{P}_h(V_{h+1}^\pi) = Q_h^\pi$ and $\mathcal{P}_h(V_{h+1}^\star) = Q_h^\star$. The performance measure is $v^\pi := \mathbb{E}_{d_1}\left[V_1^\pi\right] = \mathbb{E}_{\pi,d_1}\left[\sum_{t=1}^H r_t\right]$. Lastly, the induced state-action marginal occupancy measure for any $h \in [H]$ is defined to be: for any $E \subseteq \mathcal{S} \times \mathcal{A}$, $d_h^\pi(E) := \mathbb{E}[(s_h, a_h) \in E | s_1 \sim d_1, a_i \sim \pi(\cdot|s_i), s_i \sim P_{i-1}(\cdot|s_{i-1}, a_{i-1}), 1 \le i \le h]$ and $\mathbb{E}_{\pi,h}[f(s,a)] := \int_{\mathcal{S}\times\mathcal{A}} f(s,a) d_h^\pi(s,a) ds da$.

Offline Reinforcement Learning. The goal of Offline RL is to learn the policy $\pi^\star := \text{argmax}_\pi v^\pi$ using only the historical data $\mathcal{D} = \left\{\left(s_h^\tau, a_h^\tau, r_h^\tau, s_{h+1}^\tau\right)\right\}_{\tau\in[K]}^{h\in[H]}$. The data generating behavior policy is denoted as μ. In the offline regime, we have neither the knowledge about μ nor the access to further exploration for a different policy. The agent is asked to find a policy $\hat{\pi}$ such that $v^\star - v^{\hat{\pi}} \le \epsilon$ for the given batch data $\mathcal{D}$ and a specified accuracy $\epsilon > 0$.

4.2.1 Assumptions

Function approximation in offline RL requires sufficient expressiveness of $\mathcal{F}$. In fact, even under the *realizability* and *concentrability* conditions, sample efficient offline RL might not be achievable [92]. Therefore, under the differentiable function setting (Definition 4.1.1), we make the following assumptions.

Assumption 4.2.1 (Realizability+Bellman Completeness)**.** *The differentiable function class $\mathcal{F}$ in Definition 4.1.1 satisfies:*

- *Realizability: for optimal $Q_h^\star$, there exists $\theta_h^\star \in \Theta$ such that $Q_h^\star(\cdot,\cdot) = f(\theta_h^\star, \phi(\cdot))\ \forall h$;*
- *Bellman Completeness: Let $\mathcal{G} := \{V(\cdot) \in \mathbb{R}^S : \text{such that } \|V\|_\infty \le H\}$. Then in this*

case $\sup_{V \in \mathcal{G}} \inf_{f \in \mathcal{F}} \left\| f - \mathcal{P}_h(V) \right\|_\infty \leq \epsilon_{\mathcal{F}}$ *for some* $\epsilon_{\mathcal{F}} \geq 0$.

Realizability and *Bellman Completeness* are widely adopted in the offline RL analysis with general function approximations [13, 40] and Assumption 4.2.1 states its differentiable function approximation version. There are other forms of completeness, *e.g.* optimistic closure defined in [129].

Data coverage assumption. Furthermore, in the offline regime, it is known that function approximation cannot be sample efficient for learning a ϵ-optimal policy without data-coverage assumptions when ϵ is small (*i.e.* high accuracy) [90]. In particular, we consider two types of coverage assumptions and provide guarantees for them separately.

Assumption 4.2.2 (Concentrability Coverage)**.** *For any fixed policy* π*, define the marginal state-action occupancy ratio as* $d_h^\pi(s,a)/d_h^\mu(s,a)\ \forall s, a$*. Then the concentrability coefficient is defined as* $C_{\mathit{eff}} := \sup_\pi \sup_{h \in [H]} \left\| d_h^\pi / d_h^\mu \right\|_{2, d_h^\mu}^2$*, where* $\|g(\cdot,\cdot)\|_{2,d^\mu} := \sqrt{\mathbb{E}_{d^\mu}[g(\cdot,\cdot)^2]}$ *and* $C_{\mathit{eff}} < \infty$.

This is the standard coverage assumption that has has been widely assumed in [59, 31, 13, 15]. In the above, it requires the occupancy ratio to be finitely bounded for all the policies. In the recent work [40], they prove offline learning with GFA is efficient with only single policy concentrability, we believe similar results can be derived for DFA by modifying their main algorithm (3.2). However, chances are it will end up with a computational intractable algorithm. We leave this as the future work.

Assumption 4.2.2 is fully characterized by the MDPs. In addition, we can make an alternative assumption 4.2.3 that depends on both the MDPs and the function approximation class $\mathcal{F}$.[3] It assumes a curvature condition for $\mathcal{F}$.

Assumption 4.2.3 (Uniform Coverage)**.** *We have* $\forall h \in [H]$*, there exists* $\kappa > 0$*,*

[3]Generally speaking, 4.2.2 and 4.2.3 are not directly comparable. However, for the specific function class $f = \langle \theta, \phi \rangle$ with $\phi = \mathbf{1}(s,a)$ and tabular MDPs, it is easy to check 4.2.3 is strong than 4.2.2.

- $\mathbb{E}_{\mu,h}\left[\left(f(\theta_1,\phi(\cdot,\cdot))-f(\theta_2,\phi(\cdot,\cdot))\right)^2\right]\geq\kappa\left\|\theta_1-\theta_2\right\|_2^2,\ \ \forall\theta_1,\theta_2\in\Theta;\ (\star)$
- $\mathbb{E}_{\mu,h}\left[\nabla f(\theta,\phi(s,a))\cdot\nabla f(\theta,\phi(s,a))^\top\right]\succ\kappa I,\forall\theta\in\Theta.\ (\star\star)$

In the linear function approximation regime, Assumption 4.2.3 reduces to 4.2.4 since $(\star)$ and $(\star\star)$ are identical assumptions. If $f(\theta,\phi)=\langle\theta,\phi\rangle$, then $(\star)\ \mathbb{E}_{\mu,h}[\left(f(\theta_1,\phi(\cdot,\cdot))-f(\theta_2,\phi(\cdot,\cdot))\right)^2]=(\theta_1-\theta_2)^\top\mathbb{E}_{\mu,h}[\phi\phi^\top](\theta_1-\theta_2)\geq\kappa\left\|\theta_1-\theta_2\right\|_2^2\forall\theta_1,\theta_2\Leftrightarrow 4.2.4\Leftrightarrow(\star\star)\mathbb{E}_{\mu,h}\left[\nabla f(\theta,\phi(s,a))\cdot\nabla f(\theta,\phi(s,a))^\top\right]\succ\kappa I$. Therefore, 4.2.3 can be considered as a natural extension of 4.2.4 for differentiable class. We do point that 4.2.3 can be violated for function class $\mathcal{F}$ that is "not identifiable" by the data distribution μ (*i.e.*, there exists $f(\theta_1),f(\theta_2)\in\mathcal{F},\theta_1\neq\theta_2$ s.t. $\mathbb{E}_{\mu,h}[\left(f(\theta_1,\phi(\cdot,\cdot))-f(\theta_2,\phi(\cdot,\cdot))\right)^2]=0$). Nevertheless, there are representative non-linear differentiable classes (*e.g.* generalized linear model (GLM)) satisfying 4.2.3.

Example 4.2.4 (Linear function coverage assumption [90, 82, 106, 130])**.** *It satisfies that* $\Sigma_h^{\text{feature}}:=\mathbb{E}_{\mu,h}\left[\phi(s,a)\phi(s,a)^\top\right]\succ\kappa I,\ \forall h\in[H]$ *with some* $\kappa>0$.

Example 4.2.5 (offline generalized linear model [131, 129])**.** *For a known feature map* $\phi:\mathcal{S}\times\mathcal{A}\to\mathcal{B}_d$ *and link function* $f:[-1,1]\mapsto[-1,1]$ *the class of GLM is* $\mathcal{F}_{\text{GLM}}:=\{(s,a)\mapsto f(\langle\phi(s,a),\theta\rangle):\theta\in\Theta\}$ *satisfying* $\mathbb{E}_{\mu,h}\left[\phi(s,a)\phi(s,a)^\top\right]\succ\kappa I$. *Furthermore,* $f(\cdot)$ *is either monotonically increasing or decreasing and* $0<\kappa\leq|f'(z)|\leq K<\infty,|f''(z)|\leq M<\infty$ *for all* $|z|\leq 1$ *and some* κ,K,M. *Then* $\mathcal{F}_{\text{GLM}}$ *satisfies 4.2.3, see Appendix C.1.*

4.3 Differentiable Function Approximation is Provably Efficient

In this section, we present our solution for offline reinforcement learning with differentiable function approximation. As a warm-up, we first analyze the *vanilla fitted Q-learning* (VFQL,

Algorithm 5), which only requires the concentrability Assumption 4.2.2. The algorithm is presented in Appendix C.8.

Theorem 4.3.1. *Choose* $0 < \lambda \leq 1/2C_\Theta^2$ *in Algorithm 5. Suppose Assumption 4.2.1,4.2.2. Then if* $K \geq \max\left\{512\frac{\kappa_1^4}{\kappa^2}\left(\log(\frac{2Hd}{\delta}) + d\log(1 + \frac{4\kappa_1^3\kappa_2 C_\Theta K^3}{\lambda^2})\right), \frac{4\lambda}{\kappa}\right\}$, *with probability* $1-\delta$, *the output* $\widehat{\pi}$ *of VFQL guarantees*

$$v^\star - v^{\widehat{\pi}} \leq \sqrt{C_{\text{eff}}} H \cdot \widetilde{O}\left(\sqrt{\frac{H^2 d + \lambda C_\Theta^2}{K}} + \sqrt[4]{\frac{H^3 d \epsilon_{\mathcal{F}}}{K}}\right) + O(\sqrt{C_{\text{eff}} H^3 \epsilon_{\mathcal{F}}} + H\epsilon_{\mathcal{F}})$$

If the model approximation capacity is insufficient, 4.3.1 will induce extra error due to the large $\epsilon_{\mathcal{F}}$. If $\epsilon_{\mathcal{F}} \to 0$, the standard statistical rate $\frac{1}{\sqrt{K}}$ can be recovered and similar results are derived with general function approximation (GFA) [13, 15]. However, using concentrability coefficient conceals the problem-dependent structure and omits the specific information of differentiable functions in the complexity measure. Owing to this, we switch to the stronger "uniform" coverage 4.2.3 and analyze the *pessimistic fitted Q-learning* (PFQL, Algorithm 3) to arrive at the conclusion that offline RL with differentiable function approximation is provably efficient.

Motivation of PFQL. The PFQL algorithm mingles the two celebrated algorithmic choices: Fitted Q-Iteration (FQI) and Pessimism. However, before going into the technical details, we provide some interesting insights that motivate our analysis.

First of all, the square error loss used in FQI [58, 59] naturally couples with the differentiable function class as the resulting optimization objective is more computationally tractable (since *stochastic gradient descent* (SGD) can be readily applied) comparing to other information-theoretical algorithms derived with general function approximation (*e.g.* the *maxmin* objective in [40], eqn (3.2)).[4] In particular, FQI with differentiable function approximation resembles

[4]We mention [40] has a nice practical version PSPI, but the convergence is slower (the rate $O(n^{-\frac{1}{3}})$).

the theoretical prototype of neural FQI algorithm [126] and DQN algorithm [60, 127] when instantiating the model f to be deep neural networks. Furthermore, plenty of practical algorithms leverage fitted-Q subroutines for updating the *critic* step (*e.g.* [132, 133]) with different differentiable function choices.

In addition, we also incorporate pessimism for the design. Indeed, one of the fundamental challenges in offline RL comes from the *distributional shift*. When such a mismatch occurs, the estimated/optimized Q-function (using batch data $\mathcal{D}$) may witness severe overestimation error due to the extrapolation of model f [10]. Pessimism is the scheme to mitigate the error / overestimation bias via penalizing the Q-functions at state-action locations that have high uncertainties (as opposed to the *optimism* used in the online case), and has been widely adopted (*e.g.* [134, 37, 80]).

Algorithm 3 description. Inside the backward iteration of PFQL, Fitted Q-update is performed to optimize the parameter (Line 4). $\widehat{\theta}_h$ is the root of the first-order stationarity equation $\sum_{k=1}^{K}\left(f(\theta,\phi_{h,k}) - r_{h,k} - \widehat{V}_{h+1}(s_{h+1}^k)\right)\cdot\nabla_\theta^\top f(\theta,\phi_{h,k}) + \lambda\theta = 0$ and Σ_h is the Gram matrix with respect to $\nabla_\theta f|_{\theta=\widehat{\theta}_h}$. Note for any $s,a\in\mathcal{S}\times\mathcal{A}$, $m(s,a) := (\nabla_\theta f(\widehat{\theta}_h,\phi(s,a))^\top\Sigma_h^{-1}\nabla_\theta f(\widehat{\theta}_h,\phi(s,a)))^{-1}$ measures the effective sample size that explored s,a along the gradient direction $\nabla_\theta f|_{\theta=\widehat{\theta}_h}$, and $\beta/\sqrt{m(s,a)}$ is the estimated uncertainty at (s,a). However, the quantity $m(s,a)$ depends on $\widehat{\theta}_h$, and $\widehat{\theta}_h$ needs to be close to the true $\theta_h^\star$ (*i.e.* $\widehat{Q}_h \approx f(\widehat{\theta}_h,\phi)$ needs to be close to $Q_h^\star$) for the uncertainty estimation Γ_h to be valid, since putting a random θ into $m(s,a)$ can cause an arbitrary Γ_h that is useless (or might even deteriorate the algorithm). Such an "implicit" constraint over $\widehat{\theta}_h$ imposes the extra difficulty for the theoretical analysis due to that general differentiable functions encode nonlinear structures. As a direct comparison, in the simpler linear MDP case, the uncertainty measure $\Gamma_h := \sqrt{\phi(\cdot,\cdot)^\top(\Sigma_h^{\text{linear}})^{-1}\phi(\cdot,\cdot)}$ is always valid since it does not depend on the least-square regression weight $\widehat{\omega}_h$ [80].[5] Besides, the choice of β is set to be $\widetilde{O}(dH)$ in Theorem 4.3.2 and the extra higher order term $\widetilde{O}(\frac{1}{K})$ in Γ_h is for theoretical reason only.

[5]Here $\Sigma_h^{\text{linear}} := \sum_{k=1}^{K}\phi_{h,k}\phi_{h,k}^\top + \lambda I_d$.

Algorithm 3 Pessimistic Fitted Q-Learning (PFQL)

Input: Offline Dataset $\mathcal{D} = \left\{\left(s_h^k, a_h^k, r_h^k, s_{h+1}^k\right)\right\}_{k,h=1}^{K,H}$. Require β. Denote $\phi_{h,k} := \phi(s_h^k, a_h^k)$.

Initialization: Set $\widehat{V}_{H+1}(\cdot) \leftarrow 0$ and $\lambda > 0$.

for $h = H, H-1, \ldots, 1$ **do**

  Set $\widehat{\theta}_h \leftarrow \operatorname{argmin}_{\theta\in\Theta} \left\{ \sum_{k=1}^K \left[f\left(\theta, \phi_{h,k}\right) - r_{h,k} - \widehat{V}_{h+1}(s_{h+1}^k) \right]^2 + \lambda \cdot \|\theta\|_2^2 \right\}$

  Set $\Sigma_h \leftarrow \sum_{k=1}^K \nabla_\theta f(\widehat{\theta}_h, \phi_{h,k}) \nabla_\theta^\top f(\widehat{\theta}_h, \phi_{h,k}) + \lambda I_d$.

  Set $\Gamma_h(\cdot,\cdot) \leftarrow \beta \sqrt{\nabla_\theta f(\widehat{\theta}_h, \phi(\cdot,\cdot))^\top \Sigma_h^{-1} \nabla_\theta f(\widehat{\theta}_h, \phi(\cdot,\cdot))} \left(+\widetilde{O}(\frac{1}{K})\right)$

  Set $\bar{Q}_h(\cdot,\cdot) \leftarrow f(\widehat{\theta}_h, \phi(\cdot,\cdot)) - \Gamma_h(\cdot,\cdot)$

  Set $\widehat{Q}_h(\cdot,\cdot) \leftarrow \min\left\{\bar{Q}_h(\cdot,\cdot), H-h+1\right\}^+$

  Set $\widehat{\pi}_h(\cdot \mid \cdot) \leftarrow \arg\max_{\pi_h} \left\langle \widehat{Q}_h(\cdot,\cdot), \pi_h(\cdot \mid \cdot) \right\rangle_{\mathcal{A}}$, $\widehat{V}_h(\cdot) \leftarrow \max_{\pi_h} \left\langle \widehat{Q}_h(\cdot,\cdot), \pi_h(\cdot \mid \cdot) \right\rangle_{\mathcal{A}}$

end for

Output: $\left\{\widehat{\pi}_h\right\}_{h=1}^H$.

Model-Based vs. Model-Free. PFQL can be viewed as the strict generalization over the previous value iteration based algorithms, *e.g.* PEVI algorithm ([80], linear MDPs) and the VPVI algorithm ([5], tabular MDPs). On one hand, *approximate value iteration* (AVI) algorithms [135] are usually model-based algorithms (for instance the tabular algorithm VPVI uses empirical model $\widehat{P}$ for planning). On the other hand, FQI has the form of batch Q-learning update (*i.e.* Q-learning is a special case with batch size equals to one), therefore is more of model-free flavor. Since FQI is a concrete instantiation of the abstract AVI procedure [136], PFQL draws a unified view of model-based and model-free learning.

Now we are ready to state our main result for PFQL and the full proof can be found in Appendix C.3,C.4,C.5.

Theorem 4.3.2. *Let $\beta = 8dH\iota$ and choose $0 < \lambda \le 1/2C_\Theta^2$ in Algorithm 3. Suppose Assump-*

tion 4.2.1,4.2.3 with $\epsilon_{\mathcal{F}} = 0$.[6] *Then if* $K \geq \max\left\{512\frac{\kappa_1^4}{\kappa^2}\left(\log(\frac{2Hd}{\delta}) + d\log(1 + \frac{4\kappa_1^3\kappa_2 C_\Theta K^3}{\lambda^2})\right), \frac{4\lambda}{\kappa}\right\}$, *with probability* $1-\delta$, *for all policy* π *simultaneously, the output of PFQL guarantees*

$$v^{\pi} - v^{\hat{\pi}} \leq \sum_{h=1}^{H} 8dH \cdot \mathbb{E}_{\pi}\left[\sqrt{\nabla_\theta^\top f(\widehat{\theta}_h, \phi(s_h,a_h))\Sigma_h^{-1}\nabla_\theta f(\widehat{\theta}_h, \phi(s_h,a_h))}\right] \cdot \iota + \widetilde{O}(\frac{C_{\text{hot}}}{K}),$$

where ι *is a Polylog term and the expectation of* π *is taken over* s_h, a_h. *In particular, if further* $K \geq \max\{\widetilde{O}(\frac{(\kappa_1^2+\lambda)^2\kappa_2^2\kappa_1^2H^4d^2}{\kappa^6}), \frac{128\kappa_1^4\log(2d/\delta)}{\kappa^2}\}$ *we have*

$$0 \leq v^{\pi^\star} - v^{\hat{\pi}} \leq \sum_{h=1}^{H} 16dH \cdot \mathbb{E}_{\pi^\star}\left[\sqrt{\nabla_\theta^\top f(\theta_h^\star, \phi(s_h,a_h))\Sigma_h^{\star-1}\nabla_\theta f(\theta_h^\star, \phi(s_h,a_h))}\right] \cdot \iota + \widetilde{O}(\frac{C'_{\text{hot}}}{K}).$$

Here $\Sigma_h^\star = \sum_{k=1}^K \nabla_\theta f(\theta_h^\star, \phi(s_h^k,a_h^k))\nabla_\theta^\top f(\theta_h^\star, \phi(s_h^k,a_h^k)) + \lambda I_d$ *and the definition of higher order parameter* C_{hot}, C'_{hot} *can be found in the Notation List.*

Corollary 4.3.1 (Offline Generalized Linear Models (GLM))**.** *Consider the GLM function class defined in 4.2.5. Suppose* β, λ, K *are defined the same as Theorem 4.3.2.* $\epsilon_{\mathcal{F}} = 0$. *Then with probability* $1-\delta$, *for all policy* π *simultaneously, PFQL guarantees*

$$v^{\pi} - v^{\hat{\pi}} \leq \sum_{h=1}^{H} 8dH \cdot \mathbb{E}_{\pi}\left[\sqrt{f'(\langle\widehat{\theta}_h, \phi(s_h,a_h)\rangle)^2 \cdot \phi^\top(s_h,a_h)\Sigma_h^{-1}\phi(s_h,a_h)}\right] \cdot \iota + \widetilde{O}(\frac{C_{\text{hot}}}{K}).$$

PFQL is provably efficient. Theorem 4.3.2 verifies PFQL is statistically efficient. In particular, by Lemma C.11.5 we have $\left\|\nabla_\theta f(\theta_h^\star, \phi)\right\|_{\Sigma_h^{-1}} \lesssim \frac{2\kappa_1}{\sqrt{\kappa K}}$, resulting the main term to be bounded by $\frac{32dH^2\kappa_1}{\sqrt{\kappa K}}$ that recovers the standard statistical learning convergence rate $\frac{1}{\sqrt{K}}$.

Comparing to [80]. Theorem 4.3.2 strictly subsumes the linear MDP learning bound in [80]. Indeed, 4.3.2 reduces to $O(dH\sum_{h=1}^H \mathbb{E}_{\pi^\star}[\sqrt{\phi(s_h,a_h)^\top(\Sigma_h^{\text{linear}})^{-1}\phi(s_h,a_h)}])$ since $\nabla_\theta f(\theta,\phi) = \nabla_\theta\langle\theta,\phi\rangle = \phi$.

Instance-dependent learning. Previous studies for offline RL with general function ap-

[6]Here we assume model capacity is sufficient to make the presentation concise. If $\epsilon_{\mathcal{F}} > 0$, the complexity bound will include the term $\epsilon_{\mathcal{F}}$. We include more discussion in Appendix C.7.

proximation (GFA) [13, 14] are more of worst-case flavors as they usually rely on the *concentrability* coefficient C. The resulting learning bounds are expressed in the form[7] $O(V_{\max}^2\sqrt{\frac{C}{n}})$ that is unable to depict the behavior of individual instances. In contrast, the guarantee with differentiable function approximation is more adaptive due to the instance-dependent structure $\sum_{h=1}^{H}\mathbb{E}_{\pi^\star}\left[\sqrt{\nabla_\theta^\top f(\theta_h^\star,\phi)\Sigma_h^{\star-1}\nabla_\theta f(\theta_h^\star,\phi)}\right]$. This Fisher-Information style quantity characterizes the learning hardness of separate problems explicitly as for different MDP instances M_1, M_2, via $\sum_{h=1}^{H}\mathbb{E}_{\pi^\star}\left[\sqrt{\nabla_\theta^\top f(\theta_{h,M_i}^\star,\phi)\Sigma_h^{\star-1}\nabla_\theta f(\theta_{h,M_i}^\star,\phi)}\right]$ $(i=1,2)$, the coupled $\theta_{h,M_1}^\star,\theta_{h,M_2}^\star$ will generate different performances. Standard worst-case bounds (*e.g.* from GFA approximation) cannot explicitly differentiate between problem instances.

Feature representation vs. Parameters. One interesting observation from Theorem 4.3.2 is that the learning complexity does not depend on the feature representation dimension m but only on parameter dimension d as long as function class $\mathcal{F}$ satisfies differentiability definition 4.1.1 (not even in the higher order term). This seems to suggest, when changing the model f with more complex representations, the learning hardness will not grow as long as the number of parameters need to be learned does not increase. Note in the linear MDP analysis this phenomenon is not captured since the two dimensions are coupled ($d=m$). Therefore, this heuristic might help people rethink about what is the more essential element (feature representation vs. parameter space) in the *representation learning RL* regime (*e.g.* low rank MDPs [107]). We leave the concrete understanding the connection between features and parameters as the future work.

Technical challenges with differentiable function approximation (DFA). Informally, one key step for the analysis is to bound $|f(\widehat{\theta}_h,\phi)-f(\theta_h^\star,\phi)|$. This can be estimated by the first order approximation $\nabla f(\widehat{\theta}_h,\phi)^\top\cdot(\widehat{\theta}_h-\theta_h^\star)$. However, different from the least-square value iteration (LSVI) objective [71, 80], the *fitted Q-update* (Line 4, Algorithm 3) no longer admits a closed-form solution for $\widehat{\theta}_h$. Instead, we can only leverage $\widehat{\theta}_h$ is a stationary point of $Z_h(\theta):=$

[7]Here n is the number of samples used in the infinite horizon discounted setting and is similar to K in the episodic setting.

$\sum_{k=1}^{K}\left[f\left(\theta,\phi_{h,k}\right)-r_{h,k}-\widehat{V}_{h+1}\left(s_{h+1}^{k}\right)\right]\nabla f(\theta,\phi_{h,k})+\lambda\cdot\theta$ (since $Z_h(\widehat{\theta}_h)=0$). To measure the difference $\widehat{\theta}_h-\theta_h^\star$, for any $\theta\in\Theta$, we do the *Vector Taylor expansion* $Z_h(\theta)-Z_h(\widehat{\theta}_h)=\Sigma_h^s(\theta-\widehat{\theta}_h)+R_K(\theta)$ (where $R_K(\theta)$ is the higher-order residuals) at the point $\widehat{\theta}_h$ with

$$\begin{aligned}\Sigma_h^s &:= \frac{\partial}{\partial\theta}Z_h(\theta)\Big|_{\theta=\widehat{\theta}_h}=\frac{\partial}{\partial\theta}\left(\sum_{k=1}^{K}\left[f\left(\theta,\phi_{h,k}\right)-r_{h,k}-\widehat{V}_{h+1}(s_{h+1}^{k})\right]\nabla f(\theta,\phi_{h,k})+\lambda\cdot\theta\right)_{\theta=\widehat{\theta}_h}\\ &=\underbrace{\sum_{k=1}^{K}\left(f(\widehat{\theta}_h,\phi_{h,k})-r_{h,k}-\widehat{V}_{h+1}(s_{h+1}^{k})\right)\cdot\nabla_{\theta\theta}^2 f(\widehat{\theta}_h,\phi_{h,k})}_{:=\Delta_{\Sigma_h^s}}+\underbrace{\sum_{k=1}^{K}\nabla_\theta f(\widehat{\theta}_h,\phi_{h,k})\nabla_\theta^\top f(\widehat{\theta}_{h,k},\phi_{h,k})+\lambda I_d}_{:=\Sigma_h}.\end{aligned}\tag{4.1}$$

The perturbation term $\Delta_{\Sigma_h^s}$ encodes one key challenge for solving $\widehat{\theta}_h-\theta_h^\star$ since it breaks the positive definiteness of Σ_h^s, and, as a result, we cannot invert the Σ_h^s in the Taylor expansion of Z_h. This is due to DFA (Definition 4.1.1) is a rich class that incorporates *nonlinear* curvatures. In the linear function approximation regime, this hurdle will not show up since $\nabla_{\theta\theta}^2 f\equiv 0$ and $\Delta_{\Sigma_h^s}$ is always invertible as long as $\lambda>0$. Moreover, for the *off-policy evaluation* (OPE) task, one can overcome this issue by expanding over the population counterpart of Z_h at underlying true parameter of the given behavior target policy [110].[8] However, for the policy learning task, we cannot use either population quantity or the true parameter $\theta_h^\star$ since we need a computable/data-based pessimism Γ_h to make the algorithm practical. Check the following section for more discussions of the analysis.

4.3.1 Sketch of the PFQL Analysis

Due to the space constraint, here we only overview the key components of the analysis. To begin with, by following the result of general MDP in [80], the suboptimality gap can be bounded by (Appendix C.3) $\sum_{h=1}^{H}2\mathbb{E}_\pi\left[\Gamma_h(s_h,a_h)\right]$ if $|(\mathcal{P}_h\widehat{V}_{h+1}-f(\widehat{\theta}_h,\phi))(s,a)|\le\Gamma_h(s,a)$. To

[8] *i.e.* expanding over $Z_h^p(\theta):=\mathbb{E}_{s,a,s'}[(f(\theta,\phi(s,a))-r-V_{h+1}^\pi(s'))\nabla f(\theta,\phi(s,a))]$, and the corresponding $\Delta_{\Sigma_h^s}$ in $\frac{\partial}{\partial\theta}Z_h(\theta)|_{\theta=\theta_h^\pi}$ is zero by Bellman equation.

deal with $\mathcal{P}_h\widehat{V}_{h+1}$, by Assumption 4.2.1 we can leverage the *parameter Bellman operator* $\mathbb{T}$ (Definition C.3.1) so that $\mathcal{P}_h\widehat{V}_{h+1} = f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi)$. Next, we apply the second-order approximation to obtain $\mathcal{P}_h\widehat{V}_{h+1} - f(\widehat{\theta}_h, \phi) \approx \nabla f(\widehat{\theta}_h, \phi)^\top(\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h) + \frac{1}{2}(\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h)^\top\nabla^2_{\theta\theta}f(\widehat{\theta}_h, \phi)(\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h)$. Later, we use (4.1) to represent

$$Z_h(\theta_{\mathbb{T}\widehat{V}_{h+1}}) - Z_h(\widehat{\theta}_h) = \Sigma^s_h(\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h) + R_K(\theta_{\mathbb{T}\widehat{V}_{h+1}}) = \Sigma_h(\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h) + \widetilde{R}_K(\theta_{\mathbb{T}\widehat{V}_{h+1}})$$

by denoting $\widetilde{R}_K(\theta_{\mathbb{T}\widehat{V}_{h+1}}) = \Delta_{\Sigma^s_h}(\widehat{\theta}_h - \theta_{\mathbb{T}\widehat{V}_{h+1}}) + R_K(\theta_{\mathbb{T}\widehat{V}_{h+1}})$. Now that Σ_h^{-1} is invertible thus provides the estimation (note $Z_h(\widehat{\theta}_h) = 0$)

$$\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h = \Sigma_h^{-1} \cdot Z_h(\theta_{\mathbb{T}\widehat{V}_{h+1}}) - \Sigma_h^{-1}\widetilde{R}_K(\theta_{\mathbb{T}\widehat{V}_{h+1}}).$$

However, to handle the higher order terms, we need the explicit finite sample bound for $\|\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h\|_2$ (or $\|\theta^\star_h - \widehat{\theta}_h\|_2$). In the OPE literature, [110] uses *asymptotic theory* (Prohorov's Theorem) to show the existence of $B(\delta)$ such that $\|\widehat{\theta}_h - \theta^\star_h\| \leq \frac{B(\delta)}{\sqrt{K}}$. However, this is insufficient for *finite sample/non-asymptotic* guarantees since the abstraction of $B(\delta)$ might prevent the result from being sample efficient. For example, if $B(\delta)$ has the form $e^H \log(\frac{1}{\delta})$, then $\frac{e^H \log(\frac{1}{\delta})}{\sqrt{K}}$ is an inefficient bound since K needs to be e^H/ϵ^2 large to guarantee ϵ accuracy.

To address this technicality, we use a novel reduction to *general function approximation* (GFA) learning proposed in [13]. Concretely, we first bound the loss objective $\mathbb{E}_\mu[\ell_h(\widehat{\theta}_h)] - \mathbb{E}_\mu[\ell_h(\theta_{\mathbb{T}\widehat{V}_{h+1}})]$ via a "orthogonal" decomposition and by solving a quadratic equation. The resulting bound can be directly used to further bound $\|\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h\|_2$ for obtaining efficient guarantee $\widetilde{O}(\frac{dH}{\sqrt{\kappa K}})$. During the course, the covering technique is applied to extend the finite function hypothesis in [13] to all the differentiable functions in Definition 4.1.1. See Appendix C.6 for the complete proofs. The full proof can be found in Appendix C.3,C.4,C.5.

4.4 Improved Learning via Variance Awareness

In addition to knowing the provable efficiency for differentiable function approximation (DFA), it is of great interest to understand what is the statistical limit with DFA, or equivalently, what is the "optimal" sample/statistical complexity can be achieved in DFA (measured by minimaxity criteria)? Towards this goal, we further incorporate *variance awareness* to improve our learning guarantee. Variance awareness is first designed for linear Mixture MDPs [93, 76] to achieve the near-minimax sample complexity and it uses estimated conditional variances $\mathrm{Var}_{P(\cdot|s,a)}(V^{\star}_{h+1})$ to reweight each training sample in the LSVI objective.[9] Later, such a technique is leveraged by [82, 106] to obtained the instance-dependent results. Intuitively, conditional variances $\sigma^2(s,a) := \mathrm{Var}_{P(\cdot|s,a)}(V^{\star}_{h+1})$ serves as the uncertainty measure of the sample (s,a,r,s') that comes from the distribution $P(\cdot|s,a)$. If $\sigma^2(s,a)$ is large, then the distribution $P(\cdot|s,a)$ has high variance and we should put less weights in a single sample (s,a,r,s') rather than weighting all the samples equally. In the differentiable function approximation regime, the update is modified to

$$\widehat{\theta}_h \leftarrow \operatorname*{argmin}_{\theta\in\Theta} \left\{ \sum_{k=1}^{K} \frac{\left[f\left(\theta, \phi_{h,k}\right) - r_{h,k} - \widehat{V}_{h+1}(s^k_{h+1}) \right]^2}{\sigma^2_h(s^k_h, a^k_h)} + \lambda \cdot \|\theta\|_2^2 \right\}$$

with $\sigma^2_h(\cdot,\cdot)$ estimated by the offline data. Notably, empirical algorithms have also shown uncertainty reweighting can improve the performances for both online RL [138] and offline RL [139]. These motivates our *variance-aware fitted Q-learning* (VAFQL) algorithm 6.

Theorem 4.4.1. *Suppose Assumption 4.2.1,4.2.3 with $\epsilon_{\mathcal{F}} = 0$. Let $\beta = 8d\iota$ and choose $0 < \lambda \leq 1/2C^2_{\Theta}$ in Algorithm 6. Then if $K \geq K_0$ and $\sqrt{d} \geq \widetilde{O}(\zeta)$, with probability $1-\delta$, for all policy*

[9] We mention [137] uses variance-aware confidence sets in a slightly different way.

π simultaneously, the output of VAFQL guarantees

$$v^{\pi} - v^{\widehat{\pi}} \leq \sum_{h=1}^{H} 8d \cdot \mathbb{E}_{\pi}\left[\sqrt{\nabla_{\theta}^{\top} f(\widehat{\theta}_h, \phi(s_h, a_h))\Lambda_h^{-1} \nabla_{\theta} f(\widehat{\theta}_h, \phi(s_h, a_h))}\right] \cdot \iota + \widetilde{O}(\frac{\bar{C}_{\text{hot}}}{K}),$$

where ι is a Polylog term and the expectation of π is taken over s_h, a_h. In particular, we have

$$0 \leq v^{\pi^\star} - v^{\widehat{\pi}} \leq 16d \cdot \sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\left[\sqrt{\nabla_{\theta}^{\top} f(\theta_h^\star, \phi(s_h, a_h))\Lambda_h^{\star -1} \nabla_{\theta} f(\theta_h^\star, \phi(s_h, a_h))}\right] \cdot \iota + \widetilde{O}(\frac{\bar{C}'_{\text{hot}}}{K}).$$

Here $\Lambda_h^\star = \sum_{k=1}^{K} \nabla_\theta f(\theta_h^\star, \phi_{h,k}) \nabla_\theta^\top f(\theta_h^\star, \phi_{h,k}) / \sigma_h^\star(s_h^k, a_h^k)^2 + \lambda \boldsymbol{I}_d$ and the conditional variance quantity $\sigma_h^\star(\cdot,\cdot)^2 := \max\{1, \text{Var}_{P_h} V_{h+1}^\star(\cdot,\cdot)\}$. The definition of $K_0, \bar{C}_{\text{hot}}, \bar{C}'_{\text{hot}}, \zeta$ can be found in the Notation List.

In particular, to bound the error for $\boldsymbol{u}_h, \boldsymbol{v}_h$ and $\widehat{\sigma}_h^2$, we need to define an operator $\mathbb{J}$ that is similar to the *parameter Bellman operator* C.3.1. The Full proof of Theorem 4.4.1 can be found in Appendix C.9. Comparing to Theorem 4.3.2, VAFQL enjoys a net improvement of the horizon dependence since $\text{Var}_P(V_h^\star) \leq H^2$. Moreover, VAFQL provides better instance-dependent characterizations as the main term is fully depicted by the system quantities except the feature dimension d. For instance, when the system is fully deterministic (transition P_h's are deterministic), $\sigma_h^\star \approx \text{Var}_{P_h} V_{h+1}^\star(\cdot,\cdot) \equiv 0$ (if ignore the truncation) and $\Lambda^{\star -1} \to 0$. This yields a faster convergence with rate $O(\frac{1}{K})$. Lastly, when reduced to linear MDPs, 4.4.1 recovers the results of [106] except an extra factor of $\sqrt{d}$.

On the statistical limits. To complement the study, we incorporate a minimax lower bound via a reduction to [41, 106]. The following theorem reveals we cannot improve over Theorem 4.4.1 by more than a factor of $\sqrt{d}$ in the most general cases. The full discussion is deterred to Appendix C.10.

Theorem 4.4.2 (Minimax lower bound)**.** *Specifying the model to have linear representation $f = \langle \theta, \phi \rangle$. There exist a pair of universal constants $c, c' > 0$ such that given dimension d,*

horizon H and sample size $K > c'd^3$, one can always find a family of MDP instances such that for any algorithm $\widehat{\pi}$

$$\inf_{\widehat{\pi}} \sup_{M \in \mathcal{M}} \mathbb{E}_M \left[v^\star - v^{\widehat{\pi}} \right] \geq c \sqrt{d} \cdot \sum_{h=1}^{H} \mathbb{E}_{\pi^\star} \left[\sqrt{\nabla_\theta^\top f(\theta_h^\star, \phi(\cdot,\cdot))(\Lambda_h^{\star,p})^{-1} \nabla_\theta f(\theta_h^\star, \phi(\cdot,\cdot))} \right], \tag{4.2}$$

where $\Lambda_h^{\star,p} = \mathbb{E}\Big[\sum_{k=1}^{K} \frac{\nabla_\theta f(\theta_h^\star, \phi(s_h^k, a_h^k)) \cdot \nabla_\theta f(\theta_h^\star, \phi(s_h^k, a_h^k))^\top}{\mathrm{Var}_h(V_{h+1}^\star)(s_h^k, a_h^k)} \Big]$.

4.5 Conclusion

In this chapter [7], we study offline reinforcement learning with differentiable function approximation and show the sample efficiency for differentiable function learning. We further improve the sample complexity with respect to the horizon dependence via a variance aware variant. However, the dependence of the parameter space still scales with d (whereas for linear function approximation this dependence is $\sqrt{d}$), and this is due to applying covering argument for the rich class of differentiable functions. For large deep models, the dimension of the parameter can be huge, therefore it would be interesting to know if certain algorithms can further improve the parameter dependence, or whether this d is essential.

Also, how to relax uniform coverage assumption 4.2.3 is unknown under the current analysis. In addition, due to the technical reason, we require the third-order smoothness in Definition 4.1.1. If only the second-order or the first-order derivative information is provided, whether learning efficiency can be achieved remains an interesting question. In addition, understanding the connections between the differentiable function approximation and overparameterized neural networks approximation [140, 141] is important.

Lastly, the differentiable function approximation setting provides a general framework that is not confined to offline RL. Understanding the sample complexity behaviors of online reinforcement learning [71, 129], reward-free learning [142, 85] and representation learning [107]

might provide new and unified views over the existing studies.

Chapter 5

Conclusions and Summary

In this thesis, we analyzed the sample complexity for offline RL with problem-adaptive guarantees. In particular, we propose the Adaptive Pessimistic Value Iteration for tabular RL in Chapter 1, Variane-Aware Pessimistic Value Iteration (VAPVI) for linear function approximation in Chapter 2, and Pessimistic Fitted Q-Learning (PFQL) for differentiable function approximation in Chapter 3. Beyond that, our study also covers a wide range of topics that are not included in the previous chapters.

- **Offline Policy Evaluation.** We proposed Tabular Marginalized Importance Sampling (TMIS) estimator [2], whose MSE nearly-matches the cramer-rao lower bound in [143], and this reveals TMIS estimator is asymptotically, locally, uniformly minimax optimal, namely, optimal for every problem instance separately. Later, we propose the uniform convergence problem in OPE, and obtained the near-optimal sample complexity in the time-homogeneous and time-inhomogeneous settings respectively [3, 4, 144].

- **Offline Policy Learning.** For the policy learning task, we propose the *Double Variance Reduction* algorithm (DVR)[3] for the tabular reinforcement learning, which attains the near-optimal minimax sample complexity guarantees for finite-horizon time-

homogenuous, time-inhomogenuous, and infinite horizon discounted settings respectively. Besides, we also show linear function approximation [145] with partial coverage condition is also provably efficient.

- **Stochastic Shortest Path and Posterior Sampling RL.** We initiated the stochastic shortest path setting in the offline regime under the tabular setting (there are finite number of states and actions) [26]. We consider both the offline policy learning and the offline policy evaluation tasks for this goal-orientated setting. Very recently, we propose the posterior sampling algorithm for RL with delayed feedback and obtain the $\sqrt{T}$-regret [146].

- **Low-switching RL and ρ-gap-adjusted misspecification.** In many real-world reinforcement learning (RL) tasks, it is costly to run fully adaptive algorithms that update the exploration policy frequently. Instead, collecting data in large batches using the current policy deployment is usually cheaper. Those problems can be cast as the low-switching RL problem, and [147] first achieves the $\log\log T$ switching cost with $\sqrt{T}$ regret. Later, we further derive the logarithmic switching cost for the Linear Bellman Complete model and generalized linear model in [148]. For bandit problem, we define the new ρ-gap-adjusted misspecification, which does not require the function class to be uniformly misspecified. Under the mild assumptions, we apply the same LinUCB algorithm to achieve the $\sqrt{T}$ regret for this new notion [149].

- **Non-stationary RL and Zero-Sum Markov Games.** We made the first attempt for Non-stationary RL with general function approximation, and proposed a new complexity metric called dynamic Bellman Eluder (DBE) dimension for non-stationary MDPs, which subsumes majority of existing tractable RL problems in static MDPs as well as non-stationary MDPs [150]. Recently, for the model-free zero-sum Markov Games, the sample complexity of our algorithm for identifying ϵ-optimal Nash Equilibrium (NE) is upper bounded by $O(H^3SAB/\epsilon^2)$, which is optimal in the dependence of the horizon

H and the number of states S (where A and B denote the number of actions of the two players, respectively) [151].

- **Deep Reinforcement Learning.** We design the Closed-Form Policy Improvement (CFPI) [152] operator for tackling the locomotion tasks. We initiate offline RL algorithms with our novel policy improvement operators and empirically demonstrate their effectiveness over state-of-the-art algorithms on the standard D4RL benchmark [21].

- **MathAI and Quantization for Generalization.** In [153], we introduce TheoremQA, the first theorem-driven question-answering dataset designed to evaluate AI models' capabilities to apply theorems to solve challenging science problems. We evaluate a wide spectrum of 16 large language and code models with different prompting strategies like Chain-of-Thoughts and Program-of-Thoughts. Given the diversity and broad coverage of TheoremQA, we believe it can be used as a better benchmark to evaluate LLMs' capabilities to solve challenging science problems. Lastly, in [154], we explain why quantization improves generalization by proposing a quasi-neural network to approximate the distribution propagation.

Appendix A

Supplementary Material in Chapter 2

Algorithm 4 Vanilla Pessimistic Value Iteration

Input: Offline dataset $\mathcal{D} = \{(s_h^\tau, a_h^\tau, r_h^\tau, s_{h+1}^\tau)\}_{\tau,h=1}^{n,H}$. Absolute Constant C, failure probability δ.

Initialization: Set $\widehat{V}_{H+1}(\cdot) \leftarrow 0$.

for time $h = H, H-1, \ldots, 1$ **do**

  Set $\widehat{Q}_h(\cdot,\cdot) \leftarrow \widehat{r}_h(\cdot,\cdot) + (\widehat{P}_h \cdot \widehat{V}_{h+1})(\cdot,\cdot)$

  $\forall s_h, a_h$, set $\Gamma_h(s_h, a_h) = \frac{CH\log(HSA/\delta)}{\sqrt{n_{s_h,a_h}}}$ if $n_{s_h,a_h} \geq 1$, o.w. set to $\frac{CH\log(HSA/\delta)}{1}$.

  Set $\widehat{Q}_h^p(\cdot,\cdot) \leftarrow \widehat{Q}_h(\cdot,\cdot) - \Gamma_h(\cdot,\cdot)$.{Pessmistic update}

  Set $\overline{Q}_h(\cdot,\cdot) \leftarrow \min\{\widehat{Q}_h^p(\cdot,\cdot), H-h+1\}^+$.

  Select $\widehat{\pi}_h(\cdot|s_h) \leftarrow \operatorname{argmax}_{\pi_h}\langle \overline{Q}_h(s_h,\cdot), \pi_h(\cdot|s_h)\rangle, \forall s_h$.

  Set $\widehat{V}_h(s_h) \leftarrow \langle \overline{Q}_h(s_h,\cdot), \widehat{\pi}_h(\cdot|s_h)\rangle, \forall s_h$.

end for

Output: $\{\widehat{\pi}_h\}$.

A.1 Proof of VPVI (Theorem 2.2.1)

We begin with the following helpful lemma.

Lemma A.1.1. *For any $0 < \delta < 1$, there exists an absolute constant c_1 such that when total episode $n > c_1 \cdot 1/\bar{d}_m \cdot \log(HSA/\delta)$, then with probability $1-\delta$, $\forall h \in [H]$*

$$n_{s_h,a_h} \geq n \cdot d_h^{\mu}(s_h, a_h)/2, \quad \forall\, (s_h, a_h) \in \mathcal{C}_h.$$

Furthermore, we denote

$$\mathcal{E} := \{n_{s_h,a_h} \geq n \cdot d_h^{\mu}(s_h, a_h)/2,\ \forall\, (s_h, a_h) \in \mathcal{C}_h,\ h \in [H].\} \tag{A.1}$$

then equivalently $P(\mathcal{E}) > 1-\delta$.

In addition, we denote

$$\mathcal{E}' := \{n_{s_h,a_h} \leq \frac{3}{2} n \cdot d_h^{\mu}(s_h, a_h),\ \forall\, (s_h, a_h) \in \mathcal{C}_h,\ h \in [H].\} \tag{A.2}$$

then similarly $P(\mathcal{E}') > 1-\delta$.

Proof: [Proof of Lemma A.1.1] Define $E := \{\exists h, (s_h, a_h) \in \mathcal{C}_h \text{ s.t. } n_{s_h,a_h} < n d_h^{\mu}(s_h, a_h)/2\}$. Then combining the first part of multiplicative Chernoff bound (Lemma D.0.1 in the Appendix) and a union bound, we obtain

$$\begin{aligned}\mathbb{P}[E] &\leq \sum_h \sum_{(s_h,a_h)\in\mathcal{C}_h} \mathbb{P}[n_{s_h,a_h} < n d_h^{\mu}(s_h, a_h)/2] \\ &\leq HSA \cdot e^{-\frac{n\cdot d_m}{8}} := \delta\end{aligned}$$

solving this for n then provides the stated result.

For $\mathcal{E}'$ we can similarly use the second part of Lemma D.0.1 to prove.

Now in Lemma D.0.8, take $\pi = \pi^\star$, $\widehat{Q}_h = \overline{Q}_h$ and $\widehat{\pi} = \widehat{\pi}$ in Algorithm 4, we have

$$V_1^{\pi^\star}(s) - V_1^{\widehat{\pi}}(s) \leq \sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\left[\xi_h(s_h, a_h) \mid s_1 = s\right] - \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}\left[\xi_h(s_h, a_h) \mid s_1 = s\right] \tag{A.3}$$

here $\xi_h(s,a) = (\mathcal{T}_h \widehat{V}_{h+1})(s,a) - \overline{Q}_h(s,a)$. This is true since by the definition of $\widehat{\pi}$ in Algorithm 4 $\langle \overline{Q}_h\left(s_h, \cdot\right), \pi_h\left(\cdot|s_h\right) - \widehat{\pi}_h\left(\cdot|s_h\right)\rangle \leq 0$ almost surely. Next we prove the asymmetric bound for ξ_h, which is the key lemma for the proof.

Lemma A.1.2. *Denote $\xi_h(s,a) = (\mathcal{T}_h \widehat{V}_{h+1})(s,a) - \overline{Q}_h(s,a)$, where $\widehat{V}_{h+1}$ and $\overline{Q}_h$ are the quantities in Algorithm 4 and $\mathcal{T}_h(V) := r_h + P_h \cdot V$ for any V. Then with probability $1-\delta$, then for any h, s_h, a_h such that $d_h^\mu(s_h, a_h) > 0$, we have (C' is an absolute constant)*

$$0 \leq \xi_h(s_h, a_h) = (\mathcal{T}_h \widehat{V}_{h+1})(s_h, a_h) - \overline{Q}_h(s_h, a_h) \leq C' \cdot \sqrt{\frac{H^2 \log(HSA/\delta)}{n \cdot d_h^\mu(s_h, a_h)}}.$$

Proof: [Proof of Lemma A.1.2] Let us first consider the case where $n_{s_h, a_h} \geq 1$ for all $(s_h, a_h) \in \mathcal{C}_h$. In this case, by Hoeffding's inequality and a union bound, w.p. $1-\delta$, since $0 \leq r_h \leq 1$,

$$|\widehat{r}_h(s_h, a_h) - r_h(s_h, a_h)| \leq 2\sqrt{\frac{\log(HSA/\delta)}{n_{s_h, a_h}}} \; \forall (s_h, a_h) \in \mathcal{C}_h, h \in [H]. \tag{A.4}$$

Next, recall $\widehat{\pi}_{h+1}$ in Algorithm 4 is computed backwardly therefore only depends on sample tuple from time $h+1$ to H. Aa a result $\widehat{V}_{h+1} = \langle \overline{Q}_{h+1}, \widehat{\pi}_{h+1}\rangle$ also only depends on the sample tuple from time $h+1$ to H. On the other side, by our construction $\widehat{P}_h$ only depends on the transition pairs from h to $h+1$. Therefore $\widehat{V}_{h+1}$ and $\widehat{P}_h$ are *Conditionally* independent (This trick is also use in [3]) so by Hoeffding's inequality again[1] (note $||\widehat{V}_h||_\infty \leq ||\overline{Q}_h|| \leq H$ by

[1] It is worth mentioning if sub-policy $\widehat{\pi}_{h+1:t}$ depends on the data from all time steps $1, 2, \ldots, H$, then $\widehat{V}_{h+1}$ and $\widehat{P}_h$ are no longer conditionally independent and Hoeffding's inequality cannot be applied.

VPVI)

$$\left|\left((\widehat{P}_h - P_h)\widehat{V}_{h+1}\right)(s_h, a_h)\right| \leq 2\sqrt{\frac{H^2 \cdot \log(HSA/\delta)}{n_{s_h,a_h}}}, \quad \forall (s_h, a_h) \in C_h. \tag{A.5}$$

Now apply Lemma A.1.1, we have with high probability the event $\mathcal{E}$ (A.1) is true, combining this with (A.4), (A.5) and rescaling the constants we obtain with probability $1-\delta$, for all $h \in [H]$,

$$\begin{aligned} |\widehat{r}_h(s_h, a_h) - r_h(s_h, a_h)| &\leq C\sqrt{\frac{\log(HSA/\delta)}{6n \cdot d_h^{\mu}(s_h, a_h)}} \\ \left|\left((\widehat{P}_h - P_h)\widehat{V}_{h+1}\right)(s_h, a_h)\right| &\leq C\sqrt{\frac{H^2 \cdot \log(HSA/\delta)}{6n \cdot d_h^{\mu}(s_h, a_h)}}, \quad \forall (s_h, a_h) \in C_h. \end{aligned} \tag{A.6}$$

Now we are ready to prove the Lemma.

Step1: we prove $\xi_h(s_h, a_h) \geq 0$ for all $(s_h, a_h) \in C_h$, $h \in [H]$ with probability $1-\delta$.

We can condition on $\mathcal{E}'$ and (A.6) is true since our lemma is high probability version. Indeed, if $\widehat{Q}_h^p(s_h, a_h) < 0$, then $\overline{Q}_h(s_h, a_h) = 0$. In this case, $\xi_h(s_h, a_h) = (\mathcal{T}_h\widehat{V}_{h+1})(s_h, a_h) \geq 0$. If $\widehat{Q}_h^p(s_h, a_h) \geq 0$, then by definition $\overline{Q}_h(s_h, a_h) = \min\{\widehat{Q}_h^p(s_h, a_h), H-h+1\}^+ \leq \widehat{Q}_h^p(s_h, a_h)$ and this implies

$$\begin{aligned} \xi_h(s_h, a_h) \geq& (\mathcal{T}_h\widehat{V}_{h+1})(s_h, a_h) - \widehat{Q}_h^p(s_h, a_h) \\ =& (r_h - \widehat{r}_h)(s_h, a_h) + (P_h - \widehat{P}_h)\widehat{V}_{h+1}(s_h, a_h) + \Gamma_h(s_h, a_h) \\ \geq& -2C\sqrt{\frac{H^2 \cdot \log(HSA/\delta)}{6n \cdot d_h^{\mu}(s_h, a_h)}} + \Gamma_h(s_h, a_h) \\ \geq& -C\sqrt{\frac{2H^2 \cdot \log(HSA/\delta)}{3n \cdot d_h^{\mu}(s_h, a_h)}} + C\sqrt{\frac{H^2 \cdot \log(HSA/\delta)}{3/2 \cdot n \cdot d_h^{\mu}(s_h, a_h)}} = 0 \end{aligned}$$

where the second inequality uses (A.6) and the third inequality uses $\mathcal{E}'$.

Step2: we prove $\xi_h(s_h, a_h) \leq C' \cdot \sqrt{\frac{H^2\log(HSA/\delta)}{n \cdot d_h^{\mu}(s_h, a_h)}}$ for all $h \in [H], (s_h, a_h) \in C_h$ with probability $1-\delta$.

First, since the construction $\widehat{V}_h \leq H - h + 1$ for all $h \in [H]$, this implies

$$\widehat{Q}_h^p = \widehat{Q}_h - \Gamma_h \leq \widehat{Q}_h = \widehat{r}_h + (\widehat{P}_h \widehat{V}_{h+1}) \leq 1 + (H - h) = H - h + 1$$

which uses $\widehat{r}_h \leq 1$ almost surely and $\widehat{P}_h$ is row-stochastic. Due to this, we have the equivalent definition

$$\overline{Q}_h := \min\{\widehat{Q}_h^p, H - h + 1\}^+ = \max\{\widehat{Q}_h^p, 0\} \geq \widehat{Q}_h^p.$$

Therefore

$$\begin{aligned}
\xi_h(s_h, a_h) =& (\mathcal{T}_h \widehat{V}_{h+1})(s_h, a_h) - \overline{Q}_h(s_h, a_h) \leq (\mathcal{T}_h \widehat{V}_{h+1})(s_h, a_h) - \widehat{Q}_h^p(s_h, a_h) \\
=& (\mathcal{T}_h \widehat{V}_{h+1})(s_h, a_h) - \widehat{Q}_h(s_h, a_h) + \Gamma_h(s_h, a_h) \\
=& (r_h - \widehat{r}_h)(s_h, a_h) + (P_h - \widehat{P}_h)\widehat{V}_{h+1}(s_h, a_h) + \Gamma_h(s_h, a_h) \\
\leq& 2C\sqrt{\frac{H^2 \cdot \log(HSA/\delta)}{6n \cdot d_h^\mu(s_h, a_h)}} + \Gamma_h(s_h, a_h) \\
\leq& C\sqrt{\frac{2H^2 \cdot \log(HSA/\delta)}{3n \cdot d_h^\mu(s_h, a_h)}} + C\sqrt{\frac{2H^2 \cdot \log(HSA/\delta)}{n \cdot d_h^\mu(s_h, a_h)}} \\
=& (\sqrt{\frac{2}{3}} + \sqrt{2})C\sqrt{\frac{H^2 \cdot \log(HSA/\delta)}{n \cdot d_h^\mu(s_h, a_h)}} := C'\sqrt{\frac{H^2 \cdot \log(HSA/\delta)}{n \cdot d_h^\mu(s_h, a_h)}}
\end{aligned}$$

where the first inequality uses (A.6) and the second one uses $P(\mathcal{E}) \geq 1 - \delta$ (A.1).

Combining Step 1 and Step 2 we finish the proof.

Now we can finish proving the Theorem 2.2.1.

Proof: [Proof of Theorem 2.2.1]

Indeed, applying Lemma A.1.2 to (A.3) and average over initial distribution s_1, we obtain

with probability $1-\delta$

$$
\begin{aligned}
v^{\pi^\star} - v^{\widehat{\pi}} \leq & \sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\left[\xi_h(s_h, a_h)\right] - \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}\left[\xi_h(s_h, a_h)\right] \\
\leq & \sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\left[\xi_h(s_h, a_h)\right] - \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}[0] \\
\leq & C' H \sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\left[\sqrt{\frac{\log(HSA/\delta)}{n \cdot d_h^\mu(s_h, a_h)}}\right] - 0 \\
= & C' H \sum_{h=1}^{H} \sum_{(s_h, a_h) \in C_h} d_h^{\pi^\star}(s_h, a_h) \cdot \sqrt{\frac{\log(HSA/\delta)}{d_h^\mu(s_h, a_h)}} \cdot \sqrt{\frac{1}{n}}
\end{aligned}
$$

Note the second inequality is valid since by Line 5 of Algorithm 4 the Q-value at locations with $n_{s_h, a_h} = 0$ are heavily penalized with $O(H)$, hence the greedy $\widehat{\pi}$ will search at locations where $n_{s_h, a_h} > 0$ (which implies $d_h^\mu(s_h, a_h) > 0$). The third inequality is valid since $d_h^{\pi^\star}(s_h, a_h) > 0$ only if $d_h^\mu(s_h, a_h) > 0$. Therefore the expectation over $\pi^\star$, instead of summing over all $(s_h, a_h) \in \mathcal{S} \times \mathcal{A}$, is a sum over (s_h, a_h) s.t. $d_h^\mu(s_h, a_h) > 0$. This completes the proof.

A.2 Proof of Assumption-Free Offline Reinforcement Learning (Theorem 2.3.1)

Due to the assumption-free setting, the behavior policy μ is on longer guaranteed to trace any optimal policy $\pi^\star$. Therefore, in order to characterize the gap for the state-action agnostic space, we design the *pessimistic augmented MDP* $M^\dagger$ to reformulate the system so that the stat-actions that are agnostic to the behavior policy are subsumed into new state $s^\dagger$. Indeed, it comes from its optimistic counterpart which has a long history (*e.g.* RMAX exploration [155, 156]). Recently, [32, 37, 134] leverage this idea for continuous offline policy optimization, but their use either does not follow the assumption-free regime (see Assumption 1 of [32]) or is more empirically orientated [134, 37]. We find this helps to characterize the statistical gap when no

assumption is made in offline RL, which provides a formal understanding of the hardness in distributional mismatches.

A.2.1 Pessimistic augmented MDP

Let us define $M^\dagger$ use one extra state $s_h^\dagger$ for all $h \in \{2, \dots, H\}$ with augmented state space $S^\dagger = S \cup \{s_h^\dagger\}$ and the transition and reward is defined as follows: (recall $\mathcal{C}_h := \{(s_h, a_h) : d_h^\mu(s_h, a_h) > 0\}$)

$$P_h^\dagger(\cdot \mid s_h, a_h) = \begin{cases} P_h(\cdot \mid s_h, a_h) & s_h, a_h \in \mathcal{C}_h, \\ \delta_{s_{h+1}^\dagger} & s_h = s_h^\dagger \text{ or } s_h, a_h \notin \mathcal{C}_h, \end{cases} \quad r^\dagger(s_h, a_h) = \begin{cases} r(s_h, a_h) & s_h, a_h \in \mathcal{C}_h \\ 0 & s_h = s_h^\dagger \text{ or } s_h, a_h \notin \mathcal{C}_h \end{cases}$$

and we further define for any π

$$V_h^{\dagger\pi}(s) = \mathbb{E}_\pi^\dagger \left[\sum_{t=h}^H r_t^\dagger \middle| s_h = s \right], v^{\dagger\pi} = \mathbb{E}_\pi^\dagger \left[\sum_{t=1}^H r_t^\dagger \right] \ \forall h \in [H]. \tag{A.7}$$

Furthermore, denote $\mathcal{K}_h := \{(s_h, a_h) : n_{s_h, a_h} > 0\}$, we also create a fictitious version $\widetilde{M}^\dagger$ with:

$$\widetilde{P}_h^\dagger(\cdot \mid s_h, a_h) = \begin{cases} P_h(\cdot \mid s_h, a_h) & s_h, a_h \in \mathcal{K}_h, \\ \delta_{s_{h+1}^\dagger} & s_h = s_h^\dagger \text{ or } s_h, a_h \notin \mathcal{K}_h, \end{cases} \quad \widetilde{r}^\dagger(s_h, a_h) = \begin{cases} r(s_h, a_h) & s_h, a_h \in \mathcal{K}_h \\ 0 & s_h = s_h^\dagger \text{ or } s_h, a_h \notin \mathcal{C}_h \end{cases} \tag{A.8}$$

and the value functions under $\widetilde{M}^\dagger$ is similarly defined. Note in Section 2.3, we call (A.8) $M^\dagger$. However, it does not really matter since $\widetilde{M}^\dagger = M^\dagger$ with high probability, as stated in the following.

Lemma A.2.1. *For any $0 < \delta < 1$, there exists absolute constant c s.t. when $n \geq c \cdot 1/\bar{d}_m \cdot \log(HSA/\delta)$,*

$$\mathbb{P}(\widetilde{M}^\dagger = M^\dagger) \geq 1 - \delta.$$

Proof: Note $\{\widetilde{M}^\dagger \neq M^\dagger\} \subset \{\exists\ d_h^\mu(s_h, a_h) > 0 \text{ and } n_{s_h,a_h} = 0\}$. Similar to Lemma A.1.1, this happens with probability less than δ under the condition of n.

We have the following theorem to characterize the difference between the augmented MDP $M^\dagger$ and the original MDP M.

Theorem A.2.1. *Denote $M^\dagger = \{S, \mathcal{A}, H, r^\dagger, P^\dagger, d_1\}$ and for any π denote $V_h^{\dagger\pi}$ be the value under $M^\dagger$. Then*

$$v^\pi - \sum_{h=2}^{H+1}\sum_{t=1}^{h-1}\sum_{(s_t,a_t)\in\mathcal{S}\times\mathcal{A}\setminus\mathcal{C}_h} d_t^\pi(s_t, a_t) \leq v^\pi - \sum_{h=2}^{H+1} d_h^{\dagger\pi}(s_h^\dagger) \leq v^{\dagger\pi} \leq v^\pi \tag{A.9}$$

Before proving Theorem A.2.1, we first prove the following helper Lemmas A.2.2, A.2.3.

Lemma A.2.2. $\forall h \in [H], (s_h, a_h) \in \mathcal{S}\times\mathcal{A},\ d_h^\pi(s_h, a_h) \geq d_h^{\dagger\pi}(s_h, a_h)$.

Proof: [Proof of Lemma A.2.2] There are two cases for $(s_h, a_h) \in \mathcal{S}\times\mathcal{A}$: either $(s_h, a_h) \in \mathcal{C}_h$ or $(s_h, a_h) \notin \mathcal{C}_h$.

Step1: by the definition of $P_h^\dagger$, it directly holds: for all $s_{h+1} \in \mathcal{S}$ and $(s_h, a_h) \in \mathcal{S}\times\mathcal{A}$, $P_h^\dagger(s_{h+1}|s_h, a_h) \leq P_h(s_{h+1}|s_h, a_h)$.

Step2: we prove the argument by induction. It is clear when $h = 1$ $d_1^\pi(s_1, a_1) = d_1^{\dagger\pi}(s_1, a_1)$

(since there is no $s_1^\dagger$). Then for any $(s_h, a_h) \in \mathcal{S} \times \mathcal{A}$,

$$
\begin{aligned}
d_{h+1}^{\pi}(s_{h+1}, a_{h+1}) &= \sum_{s_h, a_h \in \mathcal{S}\times\mathcal{A}} P^{\pi}(s_{h+1}, a_{h+1}|s_h, a_h) d_h^{\pi}(s_h, a_h) \\
&= \sum_{s_h, a_h \in \mathcal{S}\times\mathcal{A}} \pi(a_{h+1}|s_{h+1}) P_h^{\pi}(s_{h+1}|s_h, a_h) d_h^{\pi}(s_h, a_h) \\
&\geq \sum_{s_h, a_h \in \mathcal{S}\times\mathcal{A}} \pi(a_{h+1}|s_{h+1}) P_h^{\dagger\pi}(s_{h+1}|s_h, a_h) d_h^{\pi}(s_h, a_h) \\
&\geq \sum_{s_h, a_h \in \mathcal{S}\times\mathcal{A}} \pi(a_{h+1}|s_{h+1}) P_h^{\dagger\pi}(s_{h+1}|s_h, a_h) d_h^{\dagger\pi}(s_h, a_h) \\
&= \sum_{s_h, a_h \in \mathcal{S}\times\mathcal{A}, s_h = s_h^\dagger} \pi(a_{h+1}|s_{h+1}) P_h^{\dagger\pi}(s_{h+1}|s_h, a_h) d_h^{\dagger\pi}(s_h, a_h) = d_{h+1}^{\dagger\pi}(s_{h+1}, a_{h+1}).
\end{aligned}
$$

where the first inequality uses Step1, the second inequality uses induction assumption and the second to last equal sign uses $P_h^{\dagger\pi}(s_{h+1}|s_h^\dagger, a_h) = 0$ for $s_{h+1} \in \mathcal{S}$. By induction we conclude the proof for this lemma.

Next we prove the second lemma that measures $d_h^{\dagger\pi}(s_h^\dagger)$.

Lemma A.2.3. *For all $h \in [2, H+1]$, $d_h^{\dagger\pi}(s_h^\dagger) = \sum_{t=1}^{h-1} \sum_{(s_t, a_t) \in \mathcal{S}\times\mathcal{A}\setminus\mathcal{C}_t} d_t^{\dagger\pi}(s_t, a_t)$.*

Proof: [Proof of Lemma A.2.3] Indeed,

$$
\begin{aligned}
&d_{h+1}^{\dagger\pi}(s_{h+1}^{\dagger}) = \sum_{a_{h+1}} d_{h+1}^{\dagger\pi}(s_{h+1}^{\dagger}, a_{h+1}) \\
=&\sum_{a_{h+1}} \sum_{(s_h,a_h)\notin C_h, s_h=s_h^{\dagger}} P^{\dagger}(s_{h+1}^{\dagger}, a_{h+1} \mid s_h, a_h) d_h^{\dagger\pi}(s_h, a_h) \\
=&\sum_{a_{h+1}} \left(\sum_{(s_h,a_h)\notin C_h} P^{\dagger}(s_{h+1}^{\dagger}, a_{h+1} \mid s_h, a_h) d_h^{\dagger\pi}(s_h, a_h) + \sum_{a_h} P^{\dagger}(s_{h+1}^{\dagger}, a_{h+1} \mid s_h^{\dagger}, a_h) d_h^{\dagger\pi}(s_h^{\dagger}, a_h) \right) \\
=&\sum_{a_{h+1}} \left(\sum_{(s_h,a_h)\notin C_h} P^{\dagger}(s_{h+1}^{\dagger}, a_{h+1} \mid s_h, a_h) d_h^{\dagger\pi}(s_h, a_h) + \sum_{a_h} \pi(a_{h+1} \mid s_{h+1}^{\dagger}) d_h^{\dagger\pi}(s_h^{\dagger}, a_h) \right) \\
=&\sum_{a_{h+1}} \left(\sum_{(s_h,a_h)\notin C_h} P^{\dagger}(s_{h+1}^{\dagger}, a_{h+1} \mid s_h, a_h) d_h^{\dagger\pi}(s_h, a_h) \right) + d_h^{\dagger\pi}(s_h^{\dagger}) \\
=&\sum_{a_{h+1}} \left(\sum_{(s_h,a_h)\notin C_h} \pi(a_{h+1} \mid s_{h+1}^{\dagger}) d_h^{\dagger\pi}(s_h, a_h) \right) + d_h^{\dagger\pi}(s_h^{\dagger}) = \sum_{(s_h,a_h)\notin C_h} d_h^{\dagger\pi}(s_h, a_h) + d_h^{\dagger\pi}(s_h^{\dagger}).
\end{aligned}
$$

Apply the above recursively we obtain the result.

Now we are ready to prove Theorem A.2.1.

Proof: [Proof of Theorem A.2.1] **Step1:** we first show $v^{\dagger\pi} \le v^{\pi}$.

Consider the stopping time $T = \inf\{t : s.t.\ (s_t, a_t) \notin C_h\} \wedge H$. Then $1 \le T \le H$.

$$
\begin{aligned}
v^{\pi} =& E_{\pi}\left[\sum_{h=1}^{H} r\left(s_h, a_h\right)\right] = E_{\pi}\left[\sum_{h=1}^{T-1} r\left(s_h, a_h\right) + \sum_{h=T}^{H} r\left(s_h, a_h\right)\right] \\
=& E_{\pi}^{\dagger}\left[\sum_{h=1}^{T-1} r\left(s_h, a_h\right)\right] + E_{\pi}\left[\sum_{h=T}^{H} r\left(s_h, a_h\right)\right] \ge E_{\pi}^{\dagger}\left[\sum_{h=1}^{T-1} r\left(s_h, a_h\right)\right] + E_{\pi}\left[\sum_{h=T}^{H} 0\right] \\
=& E_{\pi}^{\dagger}\left[\sum_{h=1}^{T-1} r\left(s_h, a_h\right)\right] + E_{\pi}^{\dagger}\left[\sum_{h=T}^{H} 0\right] = E_{\pi}^{\dagger}\left[\sum_{h=1}^{T-1} r\left(s_h, a_h\right)\right] + E_{\pi}^{\dagger}\left[\sum_{h=T}^{H} r(s_h, a_h)\right] = v^{\dagger\pi},
\end{aligned}
$$

where the third and the fourth equal signs use the distribution of T is identical under either M

or $M^\dagger$ by construction. The fifth equal sign uses the definition of pessimistic reward.

Step2: Next we show

$$v^\pi \le v^{\dagger\pi} + \sum_{h=2}^{H+1} d_h^{\dagger\pi}(s_h^\dagger) \le v^{\dagger\pi} + \sum_{h=2}^{H+1}\sum_{t=1}^{h-1}\sum_{(s_t,a_t)\in\mathcal{S}\times\mathcal{A}\setminus\mathcal{C}_t} d_t^\pi(s_t,a_t). \tag{A.10}$$

Indeed,

$$\begin{aligned}
v^\pi &= \sum_{h=1}^{H}\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d_h^\pi(s_h,a_h) r(s_h,a_h) \\
&= \sum_{h=1}^{H}\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} \left(d_h^\pi(s_h,a_h) - d_h^{\dagger\pi}(s_h,a_h)\right) r(s_h,a_h) + \sum_{h=1}^{H}\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d_h^{\dagger\pi}(s_h,a_h) r(s_h,a_h) \\
&\le \sum_{h=1}^{H}\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} \left(d_h^\pi(s_h,a_h) - d_h^{\dagger\pi}(s_h,a_h)\right)\cdot 1 + \sum_{h=1}^{H}\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d_h^{\dagger\pi}(s_h,a_h) r(s_h,a_h) \\
&= \sum_{h=1}^{H}\left(1 - \sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d_h^{\dagger\pi}(s_h,a_h)\right) + \sum_{h=1}^{H}\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d_h^{\dagger\pi}(s_h,a_h) r(s_h,a_h) \\
&= \sum_{h=2}^{H} d_h^{\dagger\pi}(s_h^\dagger) + \sum_{h=1}^{H}\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d_h^{\dagger\pi}(s_h,a_h) r(s_h,a_h) \\
&= \sum_{h=2}^{H} d_h^{\dagger\pi}(s_h^\dagger) + \sum_{h=1}^{H}\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d_h^{\dagger\pi}(s_h,a_h)\left(r(s_h,a_h) - r^\dagger(s_h,a_h)\right) + \sum_{h=1}^{H}\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d_h^{\dagger\pi}(s_h,a_h) r^\dagger(s_h,a_h) \\
&= \sum_{h=2}^{H} d_h^{\dagger\pi}(s_h^\dagger) + \sum_{h=1}^{H}\sum_{(s_h,a_h)\notin\mathcal{C}_h} d_h^{\dagger\pi}(s_h,a_h)\left(r(s_h,a_h) - r^\dagger(s_h,a_h)\right) + \sum_{h=1}^{H}\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d_h^{\dagger\pi}(s_h,a_h) r^\dagger(s_h,a_h)
\end{aligned}$$

$$\begin{aligned}
&= \sum_{h=2}^{H} d_h^{\dagger\pi}(s_h^\dagger) + \sum_{h=1}^{H}\sum_{(s_h,a_h)\notin\mathcal{C}_h} d_h^{\dagger\pi}(s_h,a_h)\left(r(s_h,a_h) - r^\dagger(s_h,a_h)\right) + v^{\dagger\pi} \\
&\le \sum_{h=2}^{H} d_h^{\dagger\pi}(s_h^\dagger) + \sum_{h=1}^{H}\sum_{(s_h,a_h)\notin\mathcal{C}_h} d_h^{\dagger\pi}(s_h,a_h)\cdot 1 + v^{\dagger\pi} = \sum_{h=2}^{H+1} d_h^{\dagger\pi}(s_h^\dagger) + v^{\dagger\pi}
\end{aligned}$$

The first inequality is due to Lemma A.2.2. The fourth equal sign uses $d_1^\dagger(s_1^\dagger) = 0$. The sixth equal sign is due to $r(s_h, a_h) = r^\dagger(s_h, a_h)$ when $(s_h, a_h) \in \mathcal{C}_h$. The seventh equal sign is due to $r^\dagger(s_h^\dagger, a_h) = 0$. The last equal sign uses Lemma A.2.3. The right inequality in (A.10) uses Lemma A.2.2. Step 1 and Step 2 conclude the proof of Theorem A.2.1.

Strong adaptive assumption-free bound

Now we are ready to launch the *assumption-free* AVPI (Algorithm 1) with the following model-based construction $\widehat{M}^\dagger$ (recall $\mathcal{K}_h := \{(s_h, a_h) : n_{s_h,a_h} > 0\}$):

$$\widehat{P}_h^\dagger(\cdot \mid s_h, a_h) = \begin{cases} \widehat{P}_h(\cdot \mid s_h, a_h) & s_h, a_h \in \mathcal{K}_h, \\ \delta_{s_{h+1}^\dagger} & s_h = s_h^\dagger \text{ or } s_h, a_h \notin \mathcal{K}_h, \end{cases} \quad \widehat{r}^\dagger(s_h, a_h) = \begin{cases} \widehat{r}(s_h, a_h) & s_h, a_h \in \mathcal{S} \times \mathcal{A} \\ 0 & s_h = s_h^\dagger \text{ or } s_h, a_h \notin \mathcal{C}_h \end{cases}$$

where $\widehat{P}, \widehat{r}$ is defined as

$$\widehat{P}_h(s'|s_h, a_h) = \frac{\sum_{\tau=1}^n \mathbf{1}[(s_{h+1}^\tau, a_h^\tau, s_h^\tau) = (s', s_h, a_h)]}{n_{s_h,a_h}}, \ \widehat{r}_h(s_h, a_h) = \frac{\sum_{\tau=1}^n \mathbf{1}[(a_h^\tau, s_h^\tau) = (s_h, a_h)] \cdot r_h^\tau}{n_{s_h,a_h}}, \tag{A.11}$$

The benefit of using $\widetilde{M}^\dagger$ (A.8) is that in $\widetilde{M}^\dagger$ there is no agnostic location even no assumption is made. The $\widehat{M}^\dagger$ creates a empirical estimate for $\widetilde{M}^\dagger$. In this case, the pessimistic bonus is designed as

$$\Gamma_h(s_h, a_h) = 2\sqrt{\frac{\mathrm{Var}_{\widehat{P}_{s_h,a_h}^\dagger}(\widehat{r}_h^\dagger + \widehat{V}_{h+1}) \cdot \iota}{n_{s_h,a_h}}} + \frac{14H \cdot \iota}{3n_{s_h,a_h}}$$

if $n_{s_h,a_h} \in \mathcal{K}_h$ and 0 otherwise (here $\widehat{V}_{h+1}$ is computed backwardly from the next time step in Algorithm 1). Now let us start the proof. First of all, let us assume $\widetilde{M}^\dagger = M^\dagger$ for the moment so we can get rid of the tilde expression for notation convenience. We will formally recover the result for $M^\dagger$ at the end by Lemma A.2.1.

In particular, while we always use $\pi^\star$ to denote the optimal policy in the *Original* MDP, we

augment it in the $M^{\dagger}(\widetilde{M}^{\dagger})$ arbitrarily and abuse the notation as:

$$\pi^{\star}(\cdot|s_h) = \begin{cases} \pi^{\star}(\cdot|s_h) & s_h \in \mathcal{S} \\ \textit{arbitrary distribution} & s_h = s_h^{\dagger} \end{cases} \tag{A.12}$$

and always use $\widehat{\pi}$ to denote the output of Algorithm 1. We rely on the following lemma that characterize the suboptimality gap.

Lemma A.2.4. *Recall $\pi^{\star}$ in* (A.12) *and define $(\mathcal{T}_h^{\dagger}V)(\cdot,\cdot) := r_h^{\dagger}(\cdot,\cdot) + (P_h^{\dagger}V)(\cdot,\cdot)$ for any $V \in \mathbb{R}^{S+1}$. Note $\widehat{\pi}$, $\overline{Q}_h$, $\widehat{V}_h$ are defined in Algorithm 1 and denote $\xi_h^{\dagger}(s,a) = (\mathcal{T}_h^{\dagger}\widehat{V}_{h+1})(s,a) - \overline{Q}_h(s,a)$.*

$$V_1^{\dagger\pi^{\star}}(s) - V_1^{\dagger\widehat{\pi}}(s) \le \sum_{h=1}^{H} \mathbb{E}_{\pi^{\star}}^{\dagger}\left[\xi_h^{\dagger}(s_h,a_h) \mid s_1 = s\right] - \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}^{\dagger}\left[\xi_h^{\dagger}(s_h,a_h) \mid s_1 = s\right]. \tag{A.13}$$

where $V_1^{\dagger\pi}$ is defined in (A.7). *Furthermore,* (A.13) *holds for all $V_h^{\dagger\pi^{\star}}(s) - V_h^{\dagger\widehat{\pi}}(s)$.*

Proof: [Proof of Lemma A.2.4] Apply Lemma D.0.8 with $\mathcal{T}_h = \mathcal{T}_h^{\dagger}$, $\pi = \pi^{\star}$, $\widehat{Q}_h = \overline{Q}_h$ and $\widehat{\pi} = \widehat{\pi}$ in Algorithm 1, we can obtain the result since by the definition of $\widehat{\pi}$ in Algorithm 1 $\langle \overline{Q}_h\left(s_h,\cdot\right), \pi_h\left(\cdot|s_h\right) - \widehat{\pi}_h\left(\cdot|s_h\right)\rangle \le 0$ almost surely for any π. The proof for $V_h^{\dagger\pi^{\star}}(s) - V_h^{\dagger\widehat{\pi}}(s)$ is identical.

Next we prove the adaptive asymmetric bound for $\xi_h^{\dagger}$, which is the key for recover the structure of intrinsic bound.

Lemma A.2.5. *Denote $\xi_h^{\dagger}(s,a) = (\mathcal{T}_h^{\dagger}\widehat{V}_{h+1})(s,a) - \overline{Q}_h(s,a)$, where $\widehat{V}_{h+1}$ and $\overline{Q}_h$ are the quantities in Algorithm 1 and $\mathcal{T}_h^{\dagger}(V) := r_h^{\dagger} + P_h^{\dagger}\cdot V$ for any $V \in \mathbb{R}^{S+1}$. Then with probability $1-\delta$, then for any h, s_h, a_h such that $n_{s_h,a_h} > 0$, we have*

$$\begin{aligned} 0 \le& \xi_h^{\dagger}(s_h,a_h) = (\mathcal{T}_h^{\dagger}\widehat{V}_{h+1})(s_h,a_h) - \overline{Q}_h(s_h,a_h) \\ \le& 4\sqrt{\frac{\mathrm{Var}_{\widehat{P}_{s_h,a_h}^{\dagger}}(\widehat{r}_h^{\dagger} + \widehat{V}_{h+1})\cdot\log(HSA/\delta)}{n_{s_h,a_h}}} + \frac{28H\cdot\log(HSA/\delta)}{3n_{s_h,a_h}} \end{aligned}$$

Proof: [Proof of Lemma A.2.5] Recall we are under $M^\dagger$ ($\widehat{M}^\dagger$). For all $(s_h, a_h) \in \mathcal{K}_h$, by Empirical Bernstein inequality (Lemma D.0.4) and a union bound[2], w.p. $1-\delta$, since $0 \le r_h^\dagger \le 1$,

$$|\widehat{r}_h^\dagger(s_h,a_h) - r_h^\dagger(s_h,a_h)| \le \sqrt{\frac{2\mathrm{Var}_{\widehat{P}^\dagger}(\widehat{r}_h^\dagger)\log(HSA/\delta)}{n_{s_h,a_h}}} + \frac{7\log(HSA/\delta)}{3n_{s_h,a_h}} \ \forall (s_h,a_h) \in \mathcal{K}_h, h \in [H]. \tag{A.14}$$

Next, recall $\widehat{\pi}_{h+1}$ in Algorithm 1 is computed backwardly therefore only depends on sample tuple from time $h+1$ to H. Aa a result $\widehat{V}_{h+1} = \langle \overline{Q}_{h+1}, \widehat{\pi}_{h+1} \rangle$ also only depends on the sample tuple from time $h+1$ to H. On the other side, by our construction $\widehat{P}_h^\dagger$ only depends on the transition pairs from h to $h+1$. Therefore $\widehat{V}_{h+1}$ and $\widehat{P}_h^\dagger$ are *Conditionally* independent (This trick is also use in [3]) so by Empirical Bernstein inequality again[3] and a union bound (note $||\widehat{V}_h||_\infty \le ||\overline{Q}_h|| \le H$ by APVI) for all $(s_h, a_h) \in \mathcal{K}_h$, w.p. $1-\delta$,

$$\left| \left((\widehat{P}_h^\dagger - P_h^\dagger)\widehat{V}_{h+1} \right)(s_h,a_h) \right| \le \sqrt{\frac{2\mathrm{Var}_{\widehat{P}^\dagger_{s_h,a_h}}(\widehat{V}_{h+1}) \cdot \log(HSA/\delta)}{n_{s_h,a_h}}} + \frac{7H \cdot \log(HSA/\delta)}{3n_{s_h,a_h}}. \tag{A.15}$$

Now we are ready to prove the Lemma.

Step1: we prove $\xi_h(s_h,a_h) \ge 0$ for all $(s_h,a_h) \in \mathcal{K}_h$, $h \in [H]$ with probability $1-\delta$.

Indeed, if $\widehat{Q}_h^p(s_h,a_h) < 0$, then $\overline{Q}_h(s_h,a_h) = 0$. In this case, $\xi_h(s_h,a_h) = (\mathcal{T}_h\widehat{V}_{h+1})(s_h,a_h) \ge 0$ (note $\widehat{V}_h \ge 0$ by the definition). If $\widehat{Q}_h^p(s_h,a_h) \ge 0$, then by definition $\overline{Q}_h(s_h,a_h) = \min\{\widehat{Q}_h^p(s_h,a_h), H-$

[2]Here note even though $|\mathcal{S}^\dagger| = S+1$, for state $s_h^\dagger$ we always have $n_{s_h^\dagger,a_h} = 0$ for any a_h. Therefore apply the union bound only provides HSA in th log term instead of $H(S+1)A$.

[3]It is worth mentioning if sub-policy $\widehat{\pi}_{h+1:t}$ depends on the data from all time steps $1, 2, \ldots, H$, then $\widehat{V}_{h+1}$ and $\widehat{P}_h$ are no longer conditionally independent and Hoeffding's inequality cannot be applied.

$h+1\}^+ \le \widehat{Q}_h^p(s_h,a_h)$ and this implies

$$\begin{aligned}
&\xi_h^\dagger(s_h,a_h) \ge (\mathcal{T}_h^\dagger \widehat{V}_{h+1})(s_h,a_h) - \widehat{Q}_h^p(s_h,a_h)\\
=&(r_h^\dagger - \widehat{r}_h^\dagger)(s_h,a_h) + (P_h^\dagger - \widehat{P}_h^\dagger)\widehat{V}_{h+1}(s_h,a_h) + \Gamma_h(s_h,a_h)\\
\ge& -2\sqrt{\frac{\mathrm{Var}_{\widehat{P}^\dagger_{s_h,a_h}}(\widehat{r}_h^\dagger + \widehat{V}_{h+1})\cdot\log(HSA/\delta)}{n_{s_h,a_h}}} - \frac{14H\cdot\log(HSA/\delta)}{3n_{s_h,a_h}} + \Gamma_h(s_h,a_h) = 0
\end{aligned}$$

where the inequality uses (A.14), (A.15) and $\sqrt{a}+\sqrt{b} \le \sqrt{2(a+b)}$ and r_h and s_{h+1} are conditionally independent given s_h, a_h. The last equal sign uses Line 6 of Algorithm 1.

Step2: we prove $\xi_h^\dagger(s_h,a_h) \le 4\sqrt{\frac{\mathrm{Var}_{\widehat{P}^\dagger_{s_h,a_h}}(\widehat{r}_h^\dagger+\widehat{V}_{h+1})\cdot\log(HSA/\delta)}{n_{s_h,a_h}}} + \frac{28H\cdot\log(HSA/\delta)}{3n_{s_h,a_h}}$ for all $h \in [H], (s_h,a_h) \in \mathcal{K}_h$ with probability $1-\delta$.

First, since by construction $\widehat{V}_h \le H-h+1$ for all $h \in [H]$, this implies

$$\widehat{Q}_h^p = \widehat{Q}_h - \Gamma_h \le \widehat{Q}_h = \widehat{r}_h^\dagger + (\widehat{P}_h^\dagger \widehat{V}_{h+1}) \le 1 + (H-h) = H-h+1$$

which uses $\widehat{r}_h^\dagger \le 1$ almost surely and $\widehat{P}_h^\dagger$ is row-stochastic. Due to this, we have the equivalent definition

$$\overline{Q}_h := \min\{\widehat{Q}_h^p, H-h+1\}^+ = \max\{\widehat{Q}_h^p, 0\} \ge \widehat{Q}_h^p.$$

Therefore

$$\begin{aligned}
&\xi_h^\dagger(s_h,a_h) = (\mathcal{T}_h^\dagger \widehat{V}_{h+1})(s_h,a_h) - \overline{Q}_h(s_h,a_h) \le (\mathcal{T}_h^\dagger \widehat{V}_{h+1})(s_h,a_h) - \widehat{Q}_h^p(s_h,a_h)\\
=&(\mathcal{T}_h^\dagger \widehat{V}_{h+1})(s_h,a_h) - \widehat{Q}_h(s_h,a_h) + \Gamma_h(s_h,a_h)\\
=&(r_h^\dagger - \widehat{r}_h^\dagger)(s_h,a_h) + (P_h^\dagger - \widehat{P}_h^\dagger)\widehat{V}_{h+1}(s_h,a_h) + \Gamma_h(s_h,a_h)\\
\le&2\sqrt{\frac{\mathrm{Var}_{\widehat{P}^\dagger_{s_h,a_h}}(\widehat{r}_h^\dagger + \widehat{V}_{h+1})\cdot\log(HSA/\delta)}{n_{s_h,a_h}}} + \frac{14H\cdot\log(HSA/\delta)}{3n_{s_h,a_h}} + \Gamma_h(s_h,a_h)\\
=&4\sqrt{\frac{\mathrm{Var}_{\widehat{P}^\dagger_{s_h,a_h}}(\widehat{r}_h^\dagger + \widehat{V}_{h+1})\cdot\log(HSA/\delta)}{n_{s_h,a_h}}} + \frac{28H\cdot\log(HSA/\delta)}{3n_{s_h,a_h}}.
\end{aligned}$$

Combining Step 1 and Step 2 we finish the proof.

Proof of Theorem 2.3.1

Now we are ready to prove the Theorem 2.3.1.

First of all, by Lemma A.2.4 and Lemma A.2.5, for all $t \in [H]$, $s \in \mathcal{S}$ (excluding $s^\dagger$) w.p. $1-\delta$

$$\begin{aligned}
&V_t^{\dagger\pi^\star}(s) - V_t^{\dagger\widehat{\pi}}(s) \le \sum_{h=t}^H \mathbb{E}^\dagger_{\pi^\star}\left[\xi_h^\dagger(s_h,a_h) \mid s_t = s\right] - \sum_{h=t}^H \mathbb{E}^\dagger_{\widehat{\pi}}\left[\xi_h^\dagger(s_h,a_h) \mid s_t = s\right]\\
\le&\sum_{h=t}^H \mathbb{E}^\dagger_{\pi^\star}\left[\xi_h^\dagger(s_h,a_h) \mid s_t = s\right] - 0\\
\le&\sum_{h=t}^H \mathbb{E}^\dagger_{\pi^\star}\left[4\sqrt{\frac{\mathrm{Var}_{\widehat{P}^\dagger_{s_h,a_h}}(\widehat{r}_h^\dagger + \widehat{V}_{h+1})\cdot\iota}{n_{s_h,a_h}}} + \frac{28H\cdot\iota}{3n_{s_h,a_h}} \mid s_t = s\right] \qquad (A.16)\\
\le&\sum_{h=t}^H \mathbb{E}^\dagger_{\pi^\star}\left[4\sqrt{\frac{2\mathrm{Var}_{\widehat{P}^\dagger_{s_h,a_h}}(\widehat{r}_h^\dagger + \widehat{V}_{h+1})\cdot\iota}{nd_h^\mu(s_h,a_h)}} + \frac{56H\cdot\iota}{3nd_h^\mu(s_h,a_h)} \mid s_t = s\right]
\end{aligned}$$

here recall the expectation is only taken over s_h, a_h. Note by the Pessimistic MDP $\widetilde{M}^\dagger$ ($\widehat{M}^\dagger$),

for all $(s_h, a_h) \notin \mathcal{K}_h$ and $s_h^\dagger$, the pessimistic reward leads to $Q^{\dagger\pi}(s_h, a_h), V^{\dagger\pi}(s_h^\dagger) = 0$ for any π, therefore Lemma A.2.5 can be applied. Moreover, the last inequality is by Lemma A.1.1.

Lemma A.2.6 (self-bounding)**.** *We prove, for all $t \in [H]$, w.p. $1-\delta$, for all $s \in \mathcal{S}$ (excluding $s^\dagger$),*

$$\left|V_t^{\dagger\pi^\star}(s) - \widehat{V}_t(s)\right| \le \frac{8\sqrt{2\iota}H^2}{\sqrt{n \cdot \bar{d}_m}} + \frac{112H^2 \cdot \iota}{3n \cdot \bar{d}_m}.$$

where $\bar{d}_m$ is defined in Theorem 2.3.1.

Remark 4. *The self-bounding lemma essentially provides a crude high probability bound for $|V_t^{\dagger\pi^\star} - \widehat{V}_t|$ (or $|V_t^{\dagger\pi^\star} - V_t^{\dagger\widehat{\pi}}|$) with suboptimal order $\widetilde{O}(\frac{H^2}{\sqrt{n\bar{d}_m}})$ and we can use it to further bound the higher order term in the main result.*

Proof: [Proof of Lemma A.2.6] Indeed, by (A.16), since $\mathrm{Var}_{\widehat{P}^\dagger_{s_h,a_h}}(\widehat{r}_h^\dagger + \widehat{V}_{h+1}) \le H^2$, we have w.p. $1-\delta$,

$$\left|V_t^{\dagger\pi^\star}(s) - V_t^{\dagger\widehat{\pi}}(s)\right| \le \frac{4\sqrt{2\iota}H^2}{\sqrt{n \cdot \bar{d}_m}} + \frac{56H^2 \cdot \iota}{3n \cdot \bar{d}_m} \tag{A.17}$$

for all $t \in [H]$. Next, when apply Lemma D.0.8 to Lemma A.2.4, by (D.2) and (D.3) we essentially obtain

$$\begin{aligned} V_t^{\dagger\pi^\star}(s) - \widehat{V}_t(s) &= \sum_{h=t}^{H} \mathbb{E}^\dagger_{\pi^\star}\left[\xi_h^\dagger(s_h, a_h) \mid s_t = s\right] + \sum_{h=t}^{H} \mathbb{E}^\dagger_{\pi^\star}\left[\left\langle \widehat{Q}_h\left(s_h, \cdot\right), \pi_h^\star\left(\cdot|s_h\right) - \widehat{\pi}_h\left(\cdot|s_h\right)\right\rangle \mid s_t = s\right] \\ &\le \frac{4\sqrt{2\iota}H^2}{\sqrt{n \cdot \bar{d}_m}} + \frac{56H^2 \cdot \iota}{3n \cdot \bar{d}_m} + 0 \end{aligned}$$

and

$$\widehat{V}_t(s) - V_t^{\dagger\widehat{\pi}}(s) = -\sum_{h=t}^{H} \mathbb{E}^\dagger_{\widehat{\pi}}\left[\xi_h^\dagger(s_h, a_h) \mid s_t = s\right] \ge 0.$$

Combing those two with (A.17) we obtain the result.

Lemma A.2.7. *For all $(a_h, a_h) \in \mathcal{K}_h$ and any $||V||_\infty \leq H$, w.p. $1-\delta$,*

$$\sqrt{\mathrm{Var}_{\widehat{P}^\dagger_{s_h,a_h}}(V)} \leq 6H\sqrt{\frac{\iota}{n \cdot d_h^\mu(s_h,a_h)}} + \sqrt{\mathrm{Var}_{P^\dagger_{s_h,a_h}}(V)}.$$

Proof:

This is a direct application of Lemma D.0.6 with a union bound. Specifically, we apply $\frac{n-1}{n} \leq 1$.

Now by Lemma A.2.6 and Lemma A.2.7, for all $(s_h, a_h) \in \mathcal{K}_h$, w.p. $1-\delta$,

$$\begin{aligned}
&\sqrt{\mathrm{Var}_{\widehat{P}^\dagger_{s_h,a_h}}(\widehat{r}^\dagger_h + \widehat{V}_{h+1})} \leq \sqrt{\mathrm{Var}_{P^\dagger_{s_h,a_h}}(\widehat{r}^\dagger_h + \widehat{V}_{h+1})} + 6H\sqrt{\frac{\iota}{n \cdot d_h^\mu(s_h,a_h)}} \\
\leq &\sqrt{\mathrm{Var}_{P^\dagger_{s_h,a_h}}(r^\dagger_h + V^{\dagger\pi^\star}_{h+1})} + \left\|(\widehat{r}^\dagger_h + \widehat{V}_{h+1}) - (r^\dagger_h + V^{\dagger\pi^\star}_{h+1})\right\|_{\infty,s\in\mathcal{S}} + 6H\sqrt{\frac{\iota}{n \cdot d_h^\mu(s_h,a_h)}} \\
\leq &\sqrt{\mathrm{Var}_{P^\dagger_{s_h,a_h}}(r^\dagger_h + V^{\dagger\pi^\star}_{h+1})} + \frac{10\sqrt{2\iota}H^2}{\sqrt{n \cdot \bar{d}_m}} + \frac{112H^2 \cdot \iota}{3n \cdot \bar{d}_m} + 6H\sqrt{\frac{\iota}{n \cdot d_h^\mu(s_h,a_h)}}
\end{aligned}$$

Therefore plug this into (A.16), and average over s_1, we finally get, w.p. $1-\delta$,

$$\begin{aligned}
v^{\dagger\pi^\star} - v^{\dagger\widehat{\pi}} \leq &\sum_{h=1}^H \mathbb{E}^\dagger_{\pi^\star}\left[4\sqrt{\frac{2\mathrm{Var}_{\widehat{P}^\dagger_{s_h,a_h}}(\widehat{r}^\dagger_h + \widehat{V}_{h+1}) \cdot \iota}{n d_h^\mu(s_h,a_h)}} + \frac{56H \cdot \iota}{3n d_h^\mu(s_h,a_h)} \mid s_1 = s\right] \\
\leq &C' \sum_{h=1}^H \mathbb{E}^\dagger_{\pi^\star}\left[\sqrt{\frac{\mathrm{Var}_{P^\dagger_{s_h,a_h}}(r^\dagger_h + V^{\dagger\pi^\star}_{h+1}) \cdot \iota}{n d_h^\mu(s_h,a_h)}}\right] + \widetilde{O}(\frac{H^3}{n \cdot \bar{d}_m}) \\
= &C' \sum_{h=1}^H \sum_{(s_h,a_h)\in\mathcal{K}_h} d^{\dagger\pi^\star}(s_h,a_h)\sqrt{\frac{\mathrm{Var}_{P^\dagger_{s_h,a_h}}(r^\dagger_h + V^{\dagger\pi^\star}_{h+1}) \cdot \iota}{n d_h^\mu(s_h,a_h)}} + \widetilde{O}(\frac{H^3}{n \cdot \bar{d}_m})
\end{aligned}$$

here $\widetilde{O}$ absorbs log factor and even higher orders.

Note throughout the section we assume $\widetilde{M}^{\dagger} = M^{\dagger}$. Now be Lemma A.2.1, we can replace the $\mathcal{K}_h$ in above by C_h so the result holds in high probability.

Lastly, we end up with w.p. $1-\delta$

$$\begin{aligned} 0 \leq & v^{\pi^\star} - v^{\widehat{\pi}} \leq \sum_{h=2}^{H+1} d_h^{\dagger\pi^\star}(s_h^\dagger) + v^{\dagger\pi^\star} - v^{\widehat{\pi}} \leq \sum_{h=2}^{H+1} d_h^{\dagger\pi^\star}(s_h^\dagger) + v^{\dagger\pi^\star} - v^{\dagger\widehat{\pi}} \\ & \leq \sum_{h=2}^{H+1} d_h^{\dagger\pi^\star}(s_h^\dagger) + C' \sum_{h=1}^{H} \sum_{(s_h,a_h)\in C_h} d_h^{\dagger\pi^\star}(s_h,a_h) \sqrt{\frac{\mathrm{Var}_{P^\dagger_{s_h,a_h}}(r_h^\dagger + V_{h+1}^{\dagger\pi^\star})\cdot\iota}{n d_h^{\mu}(s_h,a_h)}} + \widetilde{O}(\frac{H^3}{n\cdot \bar{d}_m}) \end{aligned} \tag{A.18}$$

where the first inequality uses Lemma A.2.1 with $\pi = \pi^\star$ and the second one uses Lemma A.2.1 with $\pi = \widehat{\pi}$. This concludes the proof of Theorem 2.3.1. The rest of the results are coming from Lemma A.2.2,A.2.3.

Remark 5. *We mention the summation of the main term in* (A.18) *does not include* $s_h^\dagger$ *since* $V_h^{\dagger\pi}(s_h^\dagger) = 0$ *for any* π *due to the pessimistic MDP design. In particular, this state contributes nothing to neither* $v^{\dagger\pi^\star}$ *nor* $v^{\dagger\widehat{\pi}}$.

A.2.2 Interpretation of Theorem 2.3.1

The constant (in n) gap, which is incurred by the behavior agnostic space $\bigcup_{h=1}^{H}\{(s_h,a_h) : d_h^{\mu}(s_h,a_h) = 0\}$, is bounded by

$$\sum_{h=2}^{H+1} d_h^{\dagger\pi^\star}(s_h^\dagger) = \sum_{h=2}^{H+1}\sum_{t=1}^{h-1}\sum_{(s_t,a_t)\in\mathcal{S}\times\mathcal{A}\backslash C_t} d_t^{\dagger\pi^\star}(s_t,a_t) \leq \sum_{h=2}^{H+1}\sum_{t=1}^{h-1}\sum_{(s_t,a_t)\in\mathcal{S}\times\mathcal{A}\backslash C_t} d_t^{\pi^\star}(s_t,a_t),$$

Note for quantity $d_t^{\dagger\pi^\star}(s_t,a_t)$ (where $(s_t,a_t)\in\mathcal{S}\times\mathcal{A}\backslash C_t$), it is equivalently defined as

$$d_t^{\dagger\pi^\star}(s_t,a_t) = \mathbb{P}_{M^\dagger}\left[S_t, A_t = s_t, a_t \middle| (S_{t-1},A_{t-1})\in C_{t-1},\ldots,(S_1,A_1)\in C_1\right]$$

is probability for the first time the trajectory exits the reachable regions and enters $(s_t, a_t) \notin C_t$. Therefore, $d_t^{\dagger\pi^\star}(s_t, a_t)$ is much smaller than $d_t^{\pi^\star}(s_t, a_t)$ for $s_t, a_t \notin C_h$ (since $d_t^{\pi^\star}(s_t, a_t)$ includes the probability that trajectory s_t, a_t). Such a feature is reflected by the quantity that express the gap using the mass of the absorbing state: $\sum_{h=2}^{H+1} d_h^{\dagger\pi^\star}(s_h^\dagger)(= \sum_{h=2}^{H+1}\sum_{t=1}^{h-1}\sum_{(s_t,a_t)\in\mathcal{S}\times\mathcal{A}\setminus C_t} d_t^{\dagger\pi^\star}(s_t, a_t))$. Especially, this gap can vary between 0 and H, depending on the exploratory ability of μ. Also, different from AVPI, the *assumption-free* AVPI set 0 penalty at locations where $n_{s_t,a_t} = 0$. The interpretation is: the locations with $n_{s_t,a_t} = 0$ in $M^\dagger$ are the fully aware locations (with deterministic transition to $s^\dagger$ and reward 0 by design) therefore we are certain about the behaviors in those places.

A.3 Proof of Theorem 2.2.2

Indeed, Theorem 2.2.2 can be implied by Theorem 2.3.1 as a special case. *Proof:* [Proof of Theorem 2.2.2] Under Assumption 3.3.3, $d_h^{\pi^\star}(s_h, a_h) = 0$ if $d_h^{\mu}(s_h, a_h) = 0$. In this case,

$$
\begin{aligned}
0 \le \sum_{h=2}^{H+1} d_h^{\dagger\pi^\star}(s_h^\dagger) &= \sum_{h=2}^{H+1}\sum_{t=1}^{h-1}\sum_{(s_t,a_t)\in\mathcal{S}\times\mathcal{A}\setminus C_t} d_t^{\dagger\pi^\star}(s_t, a_t) \le \sum_{h=2}^{H+1}\sum_{t=1}^{h-1}\sum_{(s_t,a_t)\in\mathcal{S}\times\mathcal{A}\setminus C_t} d_t^{\pi^\star}(s_t, a_t) \\
&= \sum_{h=2}^{H+1}\sum_{t=1}^{h-1}\sum_{(s_t,a_t):\, d_t^{\mu}(s_t,a_t)=0} d_t^{\pi^\star}(s_t, a_t) = 0
\end{aligned}
$$

due to Lemma A.2.2,A.2.3. Therefore, the gap $\sum_{h=1}^{H} d_h^{\dagger\pi^\star}(s_h^\dagger)$ vanishes when Assumption 3.3.3 is true. Also, in this case $M^\dagger$ can be replaced by a M', where M' is the sub-MDP induced by μ. *i.e.*, $M' = \bigcup_{h=1}^{H} \mathcal{S}_h \times \mathcal{A}_h$ with $\mathcal{S}_h \times \mathcal{A}_h = C_h$.[4] The transitions and the rewards remain the same in $M^\dagger$.

Since there is certain $\pi^\star$ that is fully covered by μ, for such $\pi^\star$ we have $V_h^{\pi^\star}|_M = V_h^{\pi^\star}|_{M'}$

[4] In this sub-MDP, each state might have different number of actions!

for all $h \in [H]$. Also, in M', μ can explore all the locations, therefore the probability transition to $s_h^\dagger$ is 0. Hence, all the $d^\dagger, P^\dagger, r^\dagger, V^\dagger$ in Theorem 2.2.2 are replaced by its original version.

Remark 6. *Note even though the proof can essentially leverage the reduction of the proving procedure of Theorem 2.3.1, for clear presentation of the algorithm design we still include the locations with no observation and set the severe penalty $\tilde{O}(H)$. This is different from its assumption-free version with 0 penalty (also see Section A.2.2 for related discussions).*

A.4 Discussions and missing derivations in Section 2.2

We omit the $\widetilde{O}$ notation in the derivations for the simplicity.

A.4.1 Derivation in Section 2.2.1

When the uniform data-coverage is satisfied,

$$
\begin{aligned}
v^\star - v^{\widehat{\pi}} \lesssim & \sum_{h=1}^{H} \sum_{(s_h,a_h)\in\mathcal{C}_h} d_h^{\pi^\star}(s_h,a_h) \cdot \sqrt{\frac{\mathrm{Var}_{P_{s_h,a_h}}(r_h + V_{h+1}^\star)}{n \cdot d_h^\mu(s_h,a_h)}} \\
\leq & \sqrt{\frac{1}{nd_m}} \sum_{h=1}^{H} \sum_{(s_h,a_h)\in\mathcal{C}_h} d_h^{\pi^\star}(s_h,a_h) \cdot \sqrt{\mathrm{Var}_{P_{s_h,a_h}}(r_h + V_{h+1}^\star)} \\
\leq & \sqrt{\frac{1}{nd_m}} \sum_{h=1}^{H} \sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s_h,a_h) \cdot \sqrt{\mathrm{Var}_{P_{s_h,a_h}}(r_h + V_{h+1}^\star)} \\
\leq & \sqrt{\frac{1}{nd_m}} \sum_{h=1}^{H} \sqrt{\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s_h,a_h)} \cdot \sqrt{\sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s_h,a_h)\mathrm{Var}_{P_{s_h,a_h}}(r_h + V_{h+1}^\star)} \\
\leq & \sqrt{\frac{1}{nd_m}} \sqrt{\sum_{h=1}^{H} 1} \cdot \sqrt{\sum_{h=1}^{H} \sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s_h,a_h)\mathrm{Var}_{P_{s_h,a_h}}(r_h + V_{h+1}^\star)} \\
\leq & \sqrt{\frac{1}{nd_m}} \sqrt{H} \cdot \sqrt{\mathrm{Var}_\pi\left[\sum_{t=1}^{H} r_t\right]} \leq \sqrt{\frac{H^3}{nd_m}},
\end{aligned}
$$

where we use the Cauchy inequality and Sum of total variance.

A.4.2 Uniform data-coverage in the time-invariant setting (Remark 2)

In the time-invariant setting, P is identical, therefore given data $\mathcal{D} = \left\{ \left(s_h^\tau, a_h^\tau, r_h^\tau, s_{h+1}^\tau \right) \right\}_{\tau \in [n]}^{h \in [H]}$, we should modify $n_{s,a} := \sum_{h=1}^H \sum_{\tau=1}^n \mathbf{1}[s_h^\tau, a_h^\tau = s_h, a_h]$ and

$$\widehat{P}(s'|s,a) = \frac{\sum_{h=1}^H \sum_{\tau=1}^n \mathbf{1}[(s_{h+1}^\tau, a_h^\tau, s_h^\tau) = (s', s, a)]}{n_{s,a}}, \ \widehat{r}(s,a) = \frac{\sum_{h=1}^H \sum_{\tau=1}^n \mathbf{1}[(a_h^\tau, s_h^\tau) = (s,a)] \cdot r_h^\tau}{n_{s,a}},$$

if $n_{s_h,a_h} > 0$ and $\widehat{P}(s'|s,a) = 1/S, \widehat{r}(s,a) = 0$ if $n_{s,a} = 0$. Define $\bar{d}^\mu(s,a) = \frac{1}{H} \sum_{h=1}^H d_h^\mu(s,a)$, then since in this case

$$\mathbb{E}[n_{s,a}] = \sum_{h=1}^H \sum_{\tau=1}^n d_h^\mu(s_h, a_h) = nH\bar{d}^\mu(s,a),$$

A similar algorithm should yield

$$\sqrt{\frac{1}{nHd_m}} \sqrt{H} \cdot \sqrt{\mathrm{Var}_\pi \left[\sum_{t=1}^H r_t \right]} \leq \sqrt{\frac{H^2}{nd_m}}.$$

Formalizing this result depends on decoupling the dependence between $\widehat{P}$ and $\widehat{V}_h$, which could be more tricky (see [4, 17] for two treatments under the uniform data coverage assumption). We leave this as the future work.

A.4.3 Derivation in Section 2.2.2

This follows from the derivation of Section 2.2.1 by bounding

$$v^\star - v^{\hat{\pi}} \lesssim \sqrt{\frac{1}{nd_m}} \sqrt{H} \cdot \sqrt{\mathrm{Var}_\pi \left[\sum_{t=1}^H r_t \right]} \leq \sqrt{\frac{H}{nd_m}}.$$

A.4.4 Derivation in Section 2.2.3

Using the single concentrability coefficient $C^\star$, when $\pi^\star$ is deterministic,

$$\begin{aligned}
v^\star - v^{\widehat{\pi}} \lesssim & \sum_{h=1}^{H} \sum_{(s_h,a_h)\in\mathcal{C}_h} d_h^{\pi^\star}(s_h,a_h) \cdot \sqrt{\frac{\mathrm{Var}_{P_{s_h,a_h}}(r_h + V_{h+1}^\star)}{n \cdot d_h^\mu(s_h,a_h)}} \\
\leq & \sqrt{\frac{C^\star}{n}} \sum_{h=1}^{H} \sum_{(s_h,a_h)\in\mathcal{C}_h} \sqrt{d_h^{\pi^\star}(s_h,a_h) \cdot \mathrm{Var}_{P_{s_h,a_h}}(r_h + V_{h+1}^\star)} \\
\leq & \sqrt{\frac{C^\star}{n}} \sum_{h=1}^{H} \sum_{(s_h,a_h)\in\mathcal{S}\times\mathcal{A}} \sqrt{d_h^{\pi^\star}(s_h,a_h) \cdot \mathrm{Var}_{P_{s_h,a_h}}(r_h + V_{h+1}^\star)} \\
= & \sqrt{\frac{C^\star}{n}} \sum_{h=1}^{H} \sum_{s_h\in\mathcal{S}} \sqrt{d_h^{\pi^\star}(s_h,\pi_h^\star(s_h)) \cdot \mathrm{Var}_{P_{s_h,\pi_h^\star(s_h)}}(r_h + V_{h+1}^\star)} \\
\leq & \sqrt{\frac{C^\star}{n}} \sum_{h=1}^{H} \sqrt{\sum_{s_h\in\mathcal{S}} 1} \sqrt{\sum_{s_h\in\mathcal{S}} d_h^{\pi^\star}(s_h,\pi_h^\star(s_h)) \cdot \mathrm{Var}_{P_{s_h,\pi_h^\star(s_h)}}(r_h + V_{h+1}^\star)} \\
\leq & \sqrt{\frac{SC^\star}{n}} \sum_{h=1}^{H} \sqrt{\sum_{s_h\in\mathcal{S}} d_h^{\pi^\star}(s_h,\pi_h^\star(s_h)) \cdot \mathrm{Var}_{P_{s_h,\pi_h^\star(s_h)}}(r_h + V_{h+1}^\star)} \\
\leq & \sqrt{\frac{SC^\star}{n}} \sqrt{H} \cdot \sqrt{\mathrm{Var}_\pi\left[\sum_{t=1}^{H} r_t\right]} \leq \sqrt{\frac{H^3 SC^\star}{n}}.
\end{aligned}$$

where we use the Cauchy inequality and Sum of total variance. This is minimax rate optimal.

A.4.5 Derivation in Section 2.2.4

The derivation of Proposition 2.2.4 is similar to the previous cases except we use the bounds $\mathrm{Var}_{P_h}(V_{h+1}^\star) \leq \mathbb{Q}_h^\star$ and $\sum_{h=1}^{H} r_h \leq \mathcal{B}$. The derivations for the deterministic system or the partially deterministic system are straightforward. For the fast mixing example, we leverage the fact that for any random variable X, $|X - \mathbb{E}[X]| \leq \mathrm{rng}(X)$, hence $\mathbb{Q}^\star \leq 1 + (\mathrm{rng}V^\star)^2 \leq 2$.

Last but not least, we mention the *per-step environmental norm* $\mathbb{Q}_h^\star := \max_{s_h,a_h} \mathrm{Var}_{P_{s_h,a_h}}(V_{h+1}^\star)$ is more general than its maximal version in [52] with $\mathbb{Q}^\star := \max_{s_h,a_h,h} \mathrm{Var}_{P_{s_h,a_h}}(V_{h+1}^\star)$. Improvement can be made for the $\mathbb{Q}_h^\star$ version, *e.g.* for the partially deterministic systems, $t\sqrt{\mathbb{Q}^\star/n\bar{d}_m}$

vs $H\sqrt{\mathbb{Q}^\star/n\bar{d}_m}$. Even though [52] considers the time-invariant setting, *i.e.* P is identical, the quantity $\mathbb{Q}_h^\star := \max_{s,a} \mathrm{Var}_{P_{s,a}}(V_{h+1}^\star)$ can still be much smaller than $\mathbb{Q}^\star$, *e.g.* when the range of $V_t^\star, \ldots, V_H^\star$ is relatively small and the range of $V_1^\star, \ldots, V_{t-1}^\star$ is relatively large.

In this sense, beyond the current adaptive regret $\sqrt{\mathbb{Q}^\star SAT}$ [52], the more adaptive regret should have a form like either

$$\sqrt{\frac{\sum_{h=1}^H \mathbb{Q}_h^\star SAT}{H}} \quad \text{or} \quad \sum_{h=1}^H \frac{\sqrt{\mathbb{Q}_h^\star SAT}}{H}.$$

This remains an open question in online RL.

Appendix B

Supplementary Material in Chapter 3

B.1 Proofs in Section 3.3.2

Instead of proofing the result for $v^\star - v^{\widehat{\pi}}$, in most parts of the proof we deal with $V_1^\star - V_1^{\widehat{\pi}}$, which is more general.

B.1.1 Some preparations

Define the Bellman update error $\zeta_h(s,a) := (\mathcal{T}_h \widehat{V}_{h+1})(s,a) - \widehat{Q}_h(s,a)$ and recall $\widehat{\pi}_h(s) = \operatorname{argmax}_{\pi_h} \langle \widehat{Q}_h(s,\cdot), \pi_h(\cdot \mid s) \rangle_{\mathcal{A}}$, then by the direct application of Lemma D.0.8

$$V_1^{\pi}(s) - V_1^{\widehat{\pi}}(s) \le \sum_{h=1}^{H} \mathbb{E}_{\pi}\left[\zeta_h(s_h,a_h) \mid s_1 = s\right] - \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}\left[\zeta_h(s_h,a_h) \mid s_1 = s\right]. \tag{B.1}$$

The next lemma shows it is sufficient to bound the pessimistic penalty, which is the key in the proof.

Lemma B.1.1. *Suppose with probability* $1-\delta$*, it holds for all* $h,s,a \in [H] \times \mathcal{S} \times A$ *that* $|(\mathcal{T}_h \widehat{V}_{h+1} - \widehat{\mathcal{T}}_h \widehat{V}_{h+1})(s,a)| \le \Gamma_h(s,a)$*, then it implies* $\forall s,a,h \in \mathcal{S} \times \mathcal{A} \times [H]$*,* $0 \le \zeta_h(s,a) \le$

$2\Gamma_h(s,a)$. *Furthermore, it holds for any policy* π *simultaneously, with probability* $1-\delta$,

$$V_1^{\pi}(s) - V_1^{\widehat{\pi}}(s) \leq \sum_{h=1}^{H} 2 \cdot \mathbb{E}_{\pi}\left[\Gamma_h(s_h,a_h) \mid s_1 = s\right].$$

Proof: [Proof of Lemma C.3.2]

We first show given $|(\mathcal{T}_h\widehat{V}_{h+1} - \widehat{\mathcal{T}}_h\widehat{V}_{h+1})(s,a)| \leq \Gamma_h(s,a)$, then $0 \leq \zeta_h(s,a) \leq 2\Gamma_h(s,a)$, $\forall s,a,h \in \mathcal{S}\times\mathcal{A}\times[H]$.

Step1: we first show $0 \leq \zeta_h(s,a)$, $\forall s,a,h \in \mathcal{S}\times\mathcal{A}\times[H]$.

Indeed, if $\bar{Q}_h(s,a) \leq 0$, then by definition $\widehat{Q}_h(s,a) = 0$ and in this case $\zeta_h(s,a) := (\mathcal{T}_h\widehat{V}_{h+1})(s,a) - \widehat{Q}_h(s,a) = (\mathcal{T}_h\widehat{V}_{h+1})(s,a) \geq 0$; if $\bar{Q}_h(s,a) > 0$, then $\widehat{Q}_h(s,a) \leq \bar{Q}_h(s,a)$ and

$$\begin{aligned}\zeta_h(s,a) :=& (\mathcal{T}_h\widehat{V}_{h+1})(s,a) - \widehat{Q}_h(s,a) \geq (\mathcal{T}_h\widehat{V}_{h+1})(s,a) - \bar{Q}_h(s,a)\\ =& (\mathcal{T}_h\widehat{V}_{h+1})(s,a) - (\widehat{\mathcal{T}}_h\widehat{V}_{h+1})(s,a) + \Gamma_h(s,a) \geq 0.\end{aligned}$$

Step2: next we show $\zeta_h(s,a) \leq 2\Gamma_h(s,a)$, $\forall s,a,h \in \mathcal{S}\times\mathcal{A}\times[H]$.

Indeed, we have $\widehat{Q}_h(s,a) = \max(\bar{Q}_h(s,a),0)$ and this is because: $\bar{Q}_h(x,a) = (\widehat{\mathcal{T}}_h\widehat{V}_{h+1})(x,a) - \Gamma_h(x,a) \leq (\mathcal{T}_h\widehat{V}_{h+1})(x,a) \leq H-h+1$. Therefore, in this case we have:

$$\begin{aligned}\zeta_h(s,a) :=& (\mathcal{T}_h\widehat{V}_{h+1})(s,a) - \widehat{Q}_h(s,a) \leq (\mathcal{T}_h\widehat{V}_{h+1})(s,a) - \bar{Q}_h(s,a)\\ =& (\mathcal{T}_h\widehat{V}_{h+1})(s,a) - (\widehat{\mathcal{T}}_h\widehat{V}_{h+1})(s,a) + \Gamma_h(s,a) \leq 2\cdot\Gamma_h(s,a).\end{aligned}$$

For the last statement, denote $\mathfrak{F} := \{0 \leq \zeta_h(s,a) \leq 2\Gamma_h(s,a),\ \forall s,a,h \in \mathcal{S}\times\mathcal{A}\times[H]\}$. Note conditional on $\mathfrak{F}$, then by (B.1), $V_1^{\pi}(s) - V_1^{\widehat{\pi}}(s) \leq \sum_{h=1}^{H} 2\cdot\mathbb{E}_{\pi}[\Gamma_h(s_h,a_h) \mid s_1 = s]$ holds

for any policy π almost surely. Therefore,

$$
\begin{aligned}
&\mathbb{P}\left[\forall\pi,\ \ V_1^{\pi}(s)-V_1^{\widehat{\pi}}(s)\le\sum_{h=1}^{H}2\cdot\mathbb{E}_{\pi}[\Gamma_h(s_h,a_h)\mid s_1=s].\right]\\
=&\mathbb{P}\left[\forall\pi,\ \ V_1^{\pi}(s)-V_1^{\widehat{\pi}}(s)\le\sum_{h=1}^{H}2\cdot\mathbb{E}_{\pi}[\Gamma_h(s_h,a_h)\mid s_1=s]\middle|\mathfrak{F}\right]\cdot\mathbb{P}[\mathfrak{F}]\\
+&\mathbb{P}\left[\forall\pi,\ \ V_1^{\pi}(s)-V_1^{\widehat{\pi}}(s)\le\sum_{h=1}^{H}2\cdot\mathbb{E}_{\pi}[\Gamma_h(s_h,a_h)\mid s_1=s]\middle|\mathfrak{F}^c\right]\cdot\mathbb{P}[\mathfrak{F}^c]\\
\ge&\mathbb{P}\left[\forall\pi,\ \ V_1^{\pi}(s)-V_1^{\widehat{\pi}}(s)\le\sum_{h=1}^{H}2\cdot\mathbb{E}_{\pi}[\Gamma_h(s_h,a_h)\mid s_1=s]\middle|\mathfrak{F}\right]\cdot\mathbb{P}[\mathfrak{F}]\ge 1\cdot\mathbb{P}[\mathfrak{F}]\ge 1-\delta,
\end{aligned}
$$

which finishes the proof.

B.1.2 Bounding $\left|(\mathcal{T}_h\widehat{V}_{h+1})(s,a)-(\widehat{\mathcal{T}}_h\widehat{V}_{h+1})(s,a)\right|$.

By Lemma C.3.2, it remains to bound $|(\mathcal{T}_h\widehat{V}_{h+1})(s,a)-(\widehat{\mathcal{T}}_h\widehat{V}_{h+1})(s,a)|$. Suppose w_h is the coefficient corresponding to the $\mathcal{T}_h\widehat{V}_{h+1}$ (such w_h exists by Lemma B.5.9), *i.e.* $\mathcal{T}_h\widehat{V}_{h+1}=\phi^{\top}w_h$, and recall $(\widehat{\mathcal{T}}_h\widehat{V}_{h+1})(s,a)=\phi(s,a)^{\top}\widehat{w}_h$, then:

$$
\begin{aligned}
&\left(\mathcal{T}_h\widehat{V}_{h+1}\right)(s,a)-\left(\widehat{\mathcal{T}}_h\widehat{V}_{h+1}\right)(s,a)=\phi(s,a)^{\top}\left(w_h-\widehat{w}_h\right)\\
=&\phi(s,a)^{\top}w_h-\phi(s,a)^{\top}\widehat{\Lambda}_h^{-1}\left(\sum_{\tau=1}^{K}\phi\left(s_h^{\tau},a_h^{\tau}\right)\cdot\left(r_h^{\tau}+\widehat{V}_{h+1}\left(s_{h+1}^{\tau}\right)\right)/\widehat{\sigma}_h^2(s_h^{\tau},a_h^{\tau})\right)\\
=&\underbrace{\phi(s,a)^{\top}w_h-\phi(s,a)^{\top}\widehat{\Lambda}_h^{-1}\left(\sum_{\tau=1}^{K}\phi\left(s_h^{\tau},a_h^{\tau}\right)\cdot\left(\mathcal{T}_h\widehat{V}_{h+1}\right)\left(s_h^{\tau},a_h^{\tau}\right)/\widehat{\sigma}_h^2(s_h^{\tau},a_h^{\tau})\right)}_{\text{(i)}}\\
&\underbrace{+\phi(s,a)^{\top}\widehat{\Lambda}_h^{-1}\left(\sum_{\tau=1}^{K}\phi\left(s_h^{\tau},a_h^{\tau}\right)\cdot\left(r_h^{\tau}+\widehat{V}_{h+1}\left(s_{h+1}^{\tau}\right)-\left(\mathcal{T}_h\widehat{V}_{h+1}\right)\left(s_h^{\tau},a_h^{\tau}\right)\right)/\widehat{\sigma}_h^2(s_h^{\tau},a_h^{\tau})\right)}_{\text{(ii)}}.
\end{aligned}
\tag{B.2}
$$

The term (i) is dealt by the following lemma.

Lemma B.1.2. *Recall κ in Assumption 4.2.3. Suppose $K \geq \max\left\{512H^4/\kappa^2 \log\left(\frac{2d}{\delta}\right), 4\lambda H^2/\kappa\right\}$, then with probability $1-\delta$, for all $s, a, h \in \mathcal{S}\times\mathcal{A}\times[H]$*

$$\left|\phi(s,a)^\top w_h - \phi(s,a)^\top \widehat{\Lambda}_h^{-1}\left(\sum_{\tau=1}^K \phi\left(s_h^\tau, a_h^\tau\right)\cdot\left(\mathcal{T}_h\widehat{V}_{h+1}\right)\left(s_h^\tau, a_h^\tau\right)/\widehat{\sigma}^2(s_h^\tau, a_h^\tau)\right)\right| \leq \frac{2\lambda H^3\sqrt{d}/\kappa}{K}.$$

Proof: Recall $\mathcal{T}_h\widehat{V}_{h+1} = \phi^\top w_h$ and apply Lemma C.11.5, we obtain with probability $1-\delta$, for all $s, a, h \in \mathcal{S}\times\mathcal{A}\times[H]$,

$$\begin{aligned}
&\phi(s,a)^\top w_h - \phi(s,a)^\top \widehat{\Lambda}_h^{-1}\left(\sum_{\tau=1}^K \phi\left(s_h^\tau, a_h^\tau\right)\cdot\left(\mathcal{T}_h\widehat{V}_{h+1}\right)\left(s_h^\tau, a_h^\tau\right)/\widehat{\sigma}^2(s_h^\tau, a_h^\tau)\right)\\
=&\phi(s,a)^\top w_h - \phi(s,a)^\top \widehat{\Lambda}_h^{-1}\left(\sum_{\tau=1}^K \phi\left(s_h^\tau, a_h^\tau\right)\cdot\phi(s_h^\tau, a_h^\tau)^\top w_h/\widehat{\sigma}^2(s_h^\tau, a_h^\tau)\right)\\
=&\phi(s,a)^\top w_h - \phi(s,a)^\top \widehat{\Lambda}_h^{-1}\left(\widehat{\Lambda}_h - \lambda I\right) w_h = \lambda\cdot\phi(s,a)^\top\widehat{\Lambda}_h^{-1} w_h\\
\leq&\lambda\left\|\phi(s,a)\right\|_{\widehat{\Lambda}_h^{-1}}\cdot\left\|w_h\right\|_{\widehat{\Lambda}_h^{-1}} \leq \frac{\lambda}{K}\left\|\phi(s,a)\right\|_{(\tilde{\Lambda}_h^p)^{-1}}\cdot\left\|w_h\right\|_{(\tilde{\Lambda}_h^p)^{-1}}\\
\leq&\frac{\lambda}{K}1\cdot\sqrt{\left\|(\tilde{\Lambda}_h^p)^{-1}\right\|}\cdot 2H\sqrt{d}\cdot\sqrt{\left\|(\tilde{\Lambda}_h^p)^{-1}\right\|}
\end{aligned}$$

where $\tilde{\Lambda}_h^p := \mathbb{E}_{\mu,h}\left[\widehat{\sigma}_h(s,a)^{-2}\phi(s,a)\phi(s,a)^\top\right]$ and the second inequality is by Lemma C.11.5 (with $\phi' = \phi/\widehat{\sigma}_h$ and $\left\|\phi/\widehat{\sigma}_h\right\| \leq \|\phi\| \leq 1 := C$) and the third inequality uses $\sqrt{a^\top\cdot A\cdot a} \leq \sqrt{\|a\|_2\|A\|_2\|a\|_2} = \|a\|_2\sqrt{\|A\|_2}$ with a to be either ϕ or w_h. Moreover, $\lambda_{\min}(\tilde{\Lambda}_h^p) \geq \kappa/\max_{h,s,a}\widehat{\sigma}_h(s,a)^2 \geq \kappa/H^2$ implies $\left\|(\tilde{\Lambda}_h^p)^{-1}\right\| \leq H^2/\kappa$, therefore for all $s, a, h \in \mathcal{S}\times\mathcal{A}\times[H]$, with probability $1-\delta$

$$\left|\phi(s,a)^\top w_h - \phi(s,a)^\top \widehat{\Lambda}_h^{-1}\left(\sum_{\tau=1}^K \phi\left(s_h^\tau, a_h^\tau\right)\cdot\left(\mathcal{T}_h\widehat{V}_{h+1}\right)\left(s_h^\tau, a_h^\tau\right)/\widehat{\sigma}^2(s_h^\tau, a_h^\tau)\right)\right| \leq \frac{2\lambda H^3\sqrt{d}/\kappa}{K}.$$

For term (ii), denote: $x_\tau = \frac{\phi(s_h^\tau, a_h^\tau)}{\widehat{\sigma}(s_h^\tau, a_h^\tau)}$, $\quad \eta_\tau = \left(r_h^\tau + \widehat{V}_{h+1}\left(s_{h+1}^\tau\right) - \left(\mathcal{T}_h\widehat{V}_{h+1}\right)\left(s_h^\tau, a_h^\tau\right)\right)/\widehat{\sigma}(s_h^\tau, a_h^\tau)$,

then by Cauchy inequality it follows

$$\begin{aligned}&\left|\phi(s,a)^\top\widehat{\Lambda}_h^{-1}\left(\sum_{\tau=1}^{K}\phi\left(s_h^\tau,a_h^\tau\right)\cdot\left(r_h^\tau+\widehat{V}_{h+1}\left(s_{h+1}^\tau\right)-\left(\mathcal{T}_h\widehat{V}_{h+1}\right)\left(s_h^\tau,a_h^\tau\right)\right)/\widehat{\sigma}_h^2(s_h^\tau,a_h^\tau)\right)\right|\\ \leq&\sqrt{\phi(s,a)^\top\widehat{\Lambda}_h^{-1}\phi(s,a)}\cdot||\sum_{\tau=1}^{K}x_\tau\eta_\tau||_{\widehat{\Lambda}_h^{-1}}\end{aligned}\tag{B.3}$$

Analyzing the term $\sqrt{\phi(s,a)\widehat{\Lambda}_h^{-1}\phi(s,a)}$

Recall (in Theorem 3.3.1) the estimated $\widehat{\Lambda}_h=\sum_{\tau=1}^{K}\phi\left(s_h^\tau,a_h^\tau\right)\phi\left(s_h^\tau,a_h^\tau\right)^\top/\widehat{\sigma}^2(s_h^\tau,a_h^\tau)+\lambda\cdot\boldsymbol{I}$ and $\Lambda_h=\sum_{\tau=1}^{K}\phi(s_h^\tau,a_h^\tau)^\top\phi(s_h^\tau,a_h^\tau)/\sigma^2_{\widehat{V}_{h+1}}(s_h^\tau,a_h^\tau)+\lambda\boldsymbol{I}$. Then we have the following lemma to control the term $\sqrt{\phi(s,a)\widehat{\Lambda}_h^{-1}\phi(s,a)}$.

Lemma B.1.3. *Denote the quantities $\boldsymbol{C}_1=\max\{2\lambda,128\log(2d/\delta),128H^4\log(2d/\delta)/\kappa^2\}$ and $\boldsymbol{C}_2=\max\{\frac{\lambda^2}{\kappa\log((\lambda+K)H/\lambda\delta)},96^2H^{12}d\log((\lambda+K)H/\lambda\delta)/\kappa^5\}$. Suppose the number of episode $\boldsymbol{K}$ satisfies $K>\max\{\boldsymbol{C}_1,\boldsymbol{C}_2\}$, then with probability $1-\delta$,*

$$\sqrt{\phi(s,a)\widehat{\Lambda}_h^{-1}\phi(s,a)}\leq 2\sqrt{\phi(s,a)\Lambda_h^{-1}\phi(s,a)},\quad\forall s,a\in\mathcal{S}\times\mathcal{A}.$$

Proof: [Proof of Lemma B.1.3]

By definition $\sqrt{\phi(s,a)\widehat{\Lambda}_h^{-1}\phi(s,a)}=\|\phi(s,a)\|_{\widehat{\Lambda}_h^{-1}}$. Then denote

$$\widehat{\Lambda}'_h=\frac{1}{K}\widehat{\Lambda}_h,\quad\Lambda'_h=\frac{1}{K}\Lambda_h,$$

where $\Lambda_h=\sum_{\tau=1}^{K}\phi(s_h^\tau,a_h^\tau)^\top\phi(s_h^\tau,a_h^\tau)/\sigma^2_{\widehat{V}_{h+1}}(s_h^\tau,a_h^\tau)+\lambda\boldsymbol{I}$. Under the condition of K, by Lemma B.1.6,

with probability $1-\delta$

$$\begin{aligned}
\left\|\widehat{\Lambda}'_h - \Lambda'_h\right\| &\le \sup_{s,a}\left\|\frac{\phi(s,a)\phi(s,a)^\top}{\widehat{\sigma}_h^2(s,a)} - \frac{\phi(s,a)\phi(s,a)^\top}{\sigma^2_{\widehat{V}_{h+1}}(s,a)}\right\| \\
&\le \sup_{s,a}\left|\frac{\widehat{\sigma}_h^2(s,a) - \sigma^2_{\widehat{V}_{h+1}}(s,a)}{\widehat{\sigma}_h^2(s,a)\sigma^2_{\widehat{V}_{h+1}}(s,a)}\right| \cdot \|\phi(s,a)\|^2 \le \sup_{s,a}\left|\frac{\widehat{\sigma}_h^2(s,a) - \sigma^2_{\widehat{V}_{h+1}}(s,a)}{1}\right| \cdot 1 \qquad \text{(B.4)} \\
&\le 12\sqrt{\frac{H^4 d}{\kappa K}\log\left(\frac{(\lambda+K)H}{\lambda\delta}\right)} + 12\lambda\frac{H^2\sqrt{d}}{\kappa K}.
\end{aligned}$$

Next by Lemma C.11.6 (with ϕ to be $\phi/\sigma_{\widehat{V}_{h+1}}$ and $C=1$), it holds with probability $1-\delta$,

$$\left\|\Lambda'_h - \left(\mathbb{E}_{\mu,h}[\phi(s,a)\phi(s,a)^\top/\sigma^2_{\widehat{V}_{h+1}}(s,a)] + \frac{\lambda}{K}I_d\right)\right\| \le \frac{4\sqrt{2}}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2}.$$

Therefore by *Weyl's spectrum theorem* and the condition $K > \max\{2\lambda, 128\log(2d/\delta), 128H^4\log(2d/\delta)/\kappa^2\}$, the above implies

$$\begin{aligned}
\left\|\Lambda'_h\right\| =& \lambda_{\max}(\Lambda'_h) \le \lambda_{\max}\left(\mathbb{E}_{\mu,h}[\phi(s,a)\phi(s,a)^\top/\sigma^2_{\widehat{V}_{h+1}}(s,a)]\right) + \frac{\lambda}{K} + \frac{4\sqrt{2}}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2} \\
=& \left\|\mathbb{E}_{\mu,h}[\phi(s,a)\phi(s,a)^\top/\sigma^2_{\widehat{V}_{h+1}}(s,a)]\right\|_2 + \frac{\lambda}{K} + \frac{4\sqrt{2}}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2} \\
\le& \|\phi(s,a)\|^2 + \frac{\lambda}{K} + \frac{4\sqrt{2}}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2} \le 1 + \frac{\lambda}{K} + \frac{4\sqrt{2}}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2} \le 2, \\
\lambda_{\min}(\Lambda'_h) \ge& \lambda_{\min}\left(\mathbb{E}_{\mu,h}[\phi(s,a)\phi(s,a)^\top/\sigma^2_{\widehat{V}_{h+1}}(s,a)]\right) + \frac{\lambda}{K} - \frac{4\sqrt{2}}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2} \\
\ge& \lambda_{\min}\left(\mathbb{E}_{\mu,h}[\phi(s,a)\phi(s,a)^\top/\sigma^2_{\widehat{V}_{h+1}}(s,a)]\right) - \frac{4\sqrt{2}}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2} \\
\ge& \frac{\kappa}{H^2} - \frac{4\sqrt{2}}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2} \ge \frac{\kappa}{2H^2}.
\end{aligned}$$

Hence with probability $1-\delta$, $\left\|\Lambda'_h\right\| \le 2$ and $\left\|\Lambda'^{-1}_h\right\| = 1/\lambda_{\min}(\Lambda'_h) \le 2H^2/\kappa$. Similarly, one

can show $\left\|\widehat{\Lambda}_h'^{-1}\right\| \leq 2H^2/\kappa$ with high probability.

Now apply Lemma C.11.7 to $\widehat{\Lambda}_h'$ and Λ_h' and a union bound, we obtain with probability $1-\delta$, for all s, a

$$\begin{aligned}
\|\phi(s,a)\|_{\widehat{\Lambda}_h'^{-1}} &\leq \left[1+\sqrt{\left\|\Lambda_h'^{-1}\right\|\left\|\Lambda_h'\right\|\cdot\left\|\widehat{\Lambda}_h'^{-1}\right\|\cdot\left\|\widehat{\Lambda}_h'-\Lambda_h'\right\|}\right]\cdot\|\phi(s,a)\|_{\Lambda_h'^{-1}} \\
&\leq \left[1+\sqrt{\frac{2H^2}{\kappa}\cdot 1\cdot\frac{2H^2}{\kappa}\cdot\left\|\widehat{\Lambda}_h'-\Lambda_h'\right\|}\right]\cdot\|\phi(s,a)\|_{\Lambda_h'^{-1}} \\
&\leq \left[1+\sqrt{\frac{48H^4}{\kappa^2}\left(\sqrt{\frac{H^4 d}{\kappa K}\log\left(\frac{(\lambda+K)H}{\lambda\delta}\right)}+\lambda\frac{H^2\sqrt{d}}{\kappa K}\right)}\right]\cdot\|\phi(s,a)\|_{\Lambda_h'^{-1}} \\
&\leq \left[1+\sqrt{\frac{96H^4}{\kappa^2}\sqrt{\frac{H^4 d}{\kappa K}\log\left(\frac{(\lambda+K)H}{\lambda\delta}\right)}}\right]\cdot\|\phi(s,a)\|_{\Lambda_h'^{-1}} \leq 2\|\phi(s,a)\|_{\Lambda_h'^{-1}}
\end{aligned}$$

where the third inequality uses (B.4) and the last and the second last inequality use $K > \max\{\frac{\lambda^2}{\kappa\log((\lambda+K)H/\lambda\delta)}, 96^2H^{12}d\log((\lambda+K)H/\lambda\delta)/\kappa^5\}$. Note the above is equivalent to $\sqrt{\phi(s,a)\widehat{\Lambda}_h^{-1}\phi(s,a)} \leq 2\sqrt{\phi(s,a)\Lambda_h^{-1}\phi(s,a)}$ by multiplying $1/\sqrt{K}$ on both sides.

Analyzing the term $||\sum_{\tau=1}^K x_\tau\eta_\tau||_{\hat{\Lambda}^{-1}}$

Lemma B.1.4. *Recall $x_\tau = \frac{\phi(s_h^\tau,a_h^\tau)}{\widehat{\sigma}(s_h^\tau,a_h^\tau)}$ and $\eta_\tau = \left(r_h^\tau + \widehat{V}_{h+1}\left(s_{h+1}^\tau\right) - \left(\mathcal{T}_h\widehat{V}_{h+1}\right)\left(s_h^\tau,a_h^\tau\right)\right)/\widehat{\sigma}(s_h^\tau,a_h^\tau)$. Let $C_{H,d,\kappa,K} := 36\sqrt{\frac{H^4d^3}{\kappa}\log\left(\frac{(\lambda+K)2KdH^2}{\lambda\delta}\right)} + 12\lambda\frac{H^2\sqrt{d}}{\kappa}$ and denote*

$$\xi := \sup_{V\in[0,H],\ s'\sim P_h(s,a),\ h\in[H]}\left|\frac{r_h+V(s')-\left(\mathcal{T}_hV\right)(s,a)}{\sigma_V(s,a)}\right|.$$

If $K \geq 4C_{H,d,\kappa,K}^2$ *and* $K \geq \widetilde{O}(H^6 d/\kappa)$*, then with probability* $1-\delta$*,*

$$\left\|\sum_{\tau=1}^{K} x_\tau \eta_\tau\right\|_{\widehat{\Lambda}^{-1}} \leq 16\sqrt{d\log\left(1+\frac{K}{\lambda d}\right)\cdot\log\left(\frac{4K^2}{\delta}\right)} + 4\xi\log\left(\frac{4K^2}{\delta}\right) \leq \widetilde{O}\max\left\{\sqrt{d},\xi\right\},$$

where $\widetilde{O}$ *absorbs the constants and Polylog terms.*

Proof: [Proof of Lemma B.1.4] By construction, we have $\|x_\tau\| \leq \|\phi/\widehat{\sigma}\| \leq 1$ and by Lemma B.1.6, with probability $1-\delta/3$,

$$\left\|\sigma_{\widehat{V}_{h+1}} - \widehat{\sigma}_h\right\|_\infty = \sup_{s,a}\frac{\left|\sigma^2_{\widehat{V}_{h+1}}(s,a) - \widehat{\sigma}_h^2(s,a)\right|}{\left|\sigma_{\widehat{V}_{h+1}}(s,a) + \widehat{\sigma}_h(s,a)\right|} \leq \frac{1}{2}\left\|\sigma^2_{\widehat{V}_{h+1}} - \widehat{\sigma}_h^2\right\|_\infty \leq C_{H,d,\kappa,K}\sqrt{\frac{1}{K}}$$

Therefore, when $K \geq 4C_{H,d,\kappa,K}^2$, $C_{H,d,\kappa,K}\sqrt{\frac{1}{K}} \leq 1/2 \leq \sigma_{\widehat{V}_{h+1}}(s_h^\tau, a_h^\tau)/2$ and hence

$$\begin{aligned}
|\eta_\tau| &\leq \left|\frac{r_h^\tau + \widehat{V}_{h+1}\left(s_{h+1}^\tau\right) - \left(\mathcal{T}_h\widehat{V}_{h+1}\right)\left(s_h^\tau, a_h^\tau\right)}{\sigma_{\widehat{V}_{h+1}}(s_h^\tau, a_h^\tau) - \frac{C_{H,d,\kappa,K}}{K^{1/2}}}\right| \leq 2\left|\frac{r_h^\tau + \widehat{V}_{h+1}\left(s_{h+1}^\tau\right) - \left(\mathcal{T}_h\widehat{V}_{h+1}\right)\left(s_h^\tau, a_h^\tau\right)}{\sigma_{\widehat{V}_{h+1}}(s_h^\tau, a_h^\tau)}\right| \\
&\leq 2\sup_{V\in[0,H],\, s'\sim P_h(s,a)}\left|\frac{r + V(s') - \left(\mathcal{T}_h V\right)(s,a)}{\sigma_V(s,a)}\right| := \xi.
\end{aligned}$$

Next, for a fixed function V, we define the Bellman error as $\mathcal{B}_h(V)(s,a) = r_h + V(s') -$

$(\mathcal{T}_h V)(s,a)$, then

$$
\begin{aligned}
\mathrm{Var}\left[\eta_\tau|\mathcal{F}_{\tau-1}\right] =& \frac{\mathrm{Var}\left[r_h^\tau + \widehat{V}_{h+1}\left(s_{h+1}^\tau\right) - \left(\mathcal{T}_h\widehat{V}_{h+1}\right)\left(s_h^\tau,a_h^\tau\right)\middle|\mathcal{F}_{\tau-1}\right]}{\widehat{\sigma}^2(s_h^\tau,a_h^\tau)} \\
=& \frac{\mathrm{Var}\left[\mathcal{B}_h\widehat{V}_{h+1}(s_h^\tau,a_h^\tau) - \mathcal{B}_h V_{h+1}^\star(s_h^\tau,a_h^\tau) + \mathcal{B}_h V_{h+1}^\star(s_h^\tau,a_h^\tau)\middle|\mathcal{F}_{\tau-1}\right]}{\widehat{\sigma}^2(s_h^\tau,a_h^\tau)} \\
\le& \frac{\mathrm{Var}\left[\mathcal{B}_h V_{h+1}^\star(s_h^\tau,a_h^\tau)\middle|\mathcal{F}_{\tau-1}\right] + 8H\left\|\mathcal{B}_h\widehat{V}_{h+1} - \mathcal{B}_h V_{h+1}^\star\right\|_\infty}{\widehat{\sigma}^2(s_h^\tau,a_h^\tau)} \\
\le& \frac{\mathrm{Var}\left[\mathcal{B}_h V_{h+1}^\star(s_h^\tau,a_h^\tau)\middle|\mathcal{F}_{\tau-1}\right] + 16H\left\|\widehat{V}_{h+1} - V_{h+1}^\star\right\|_\infty}{\widehat{\sigma}^2(s_h^\tau,a_h^\tau)} \\
\le& \frac{\mathrm{Var}\left[\mathcal{B}_h V_{h+1}^\star(s_h^\tau,a_h^\tau)\middle|\mathcal{F}_{\tau-1}\right] + \widetilde{O}(\frac{H^3\sqrt{d}}{\sqrt{\kappa K}})}{\widehat{\sigma}^2(s_h^\tau,a_h^\tau)} \\
=& \frac{\mathrm{Var}\left[\mathcal{B}_h V_{h+1}^\star(s_h^\tau,a_h^\tau)\middle|s_h^\tau,a_h^\tau\right] + \widetilde{O}(\frac{H^3\sqrt{d}}{\sqrt{\kappa K}})}{\widehat{\sigma}^2(s_h^\tau,a_h^\tau)} \\
=& \frac{\mathrm{Var}_{V_{h+1}^\star}(s_h^\tau,a_h^\tau) + \widetilde{O}(\frac{H^3\sqrt{d}}{\sqrt{\kappa K}})}{\widehat{\sigma}^2(s_h^\tau,a_h^\tau)} \le \frac{2\mathrm{Var}_{V_{h+1}^\star}(s_h^\tau,a_h^\tau) + \widetilde{O}(\frac{H^3\sqrt{d}}{\sqrt{\kappa K}})}{\sigma^{\star 2}(s_h^\tau,a_h^\tau)} \le 2 + \frac{\widetilde{O}(\frac{H^3\sqrt{d}}{\sqrt{\kappa K}})}{\sigma^{\star 2}(s_h^\tau,a_h^\tau)} \\
\le& \widetilde{O}(1)
\end{aligned}
$$

where the first inequality is by Lemma B.5.11, the second inequality is by $\mathcal{T}_h$ is non-expansive, the third inequality is by Lemma B.1.7, the next equality is by Markovian property, and the fourth inequality is by Lemma B.1.6 and Lemma B.1.8. The fifth inequality uses definition $\sigma_{h,V}(s,a)^2 := \max\{1, \mathrm{Var}_{P_h}(V)(s,a)\}$ and the last one is by condition $K \ge \widetilde{O}(H^6 d/\kappa)$ and $\sigma_{h,V^\star}(s,a)^2 := \max\{1, \mathrm{Var}_{P_h}(V^\star)(s,a)\} \ge 1$. Thus, by Bernstein inequality for self-normalized

martingale (Lemma C.11.4),[1] with probability $1-\delta$,

$$\left\|\sum_{\tau=1}^{K} x_\tau \eta_\tau\right\|_{\hat{\Lambda}^{-1}} \leq \widetilde{O}\left(\sqrt{d\log\left(1+\frac{K}{\lambda d}\right)\cdot\log\left(\frac{4K^2}{\delta}\right)}\right)+4\xi\log\left(\frac{4K^2}{\delta}\right)\leq \widetilde{O}\max\left\{\sqrt{d},\xi\right\}$$

where $\widetilde{O}$ absorbs the constants and Polylog terms.

Recall $\mathcal{M}_1,\mathcal{M}_2,\mathcal{M}_3,\mathcal{M}_4$ in List . Based on the above results, we have the following key lemma:

Lemma B.1.5. *Assume $K>\max\{\mathcal{M}_1,\mathcal{M}_2,\mathcal{M}_3,\mathcal{M}_4\}$, for any $0<\lambda<\kappa$, suppose $\sqrt{d}>\xi$, where $\xi := \sup_{V\in[0,H],\, s'\sim P_h(s,a),\, h\in[H]}\left|\frac{r_h+V(s')-(\mathcal{T}_hV)(s,a)}{\sigma_V(s,a)}\right|$. Then with probability $1-\delta$, for all $h,s,a\in[H]\times\mathcal{S}\times\mathcal{A}$,*

$$\left|(\mathcal{T}_h\widehat{V}_{h+1}-\widehat{\mathcal{T}}_h\widehat{V}_{h+1})(s,a)\right|\leq\widetilde{O}\left(\sqrt{d}\sqrt{\phi(s,a)\Lambda_h^{-1}\phi(s,a)}\right)+\frac{2H^3\sqrt{d}}{K},$$

where $\Lambda_h=\sum_{\tau=1}^{K}\phi(s_h^\tau,a_h^\tau)^\top\phi(s_h^\tau,a_h^\tau)/\sigma^2_{\widehat{V}_{h+1}}(s_h^\tau,a_h^\tau)+\lambda I$ and $\widetilde{O}$ absorbs the universal constants and Polylog terms.

Proof: [Proof of Lemma B.1.5] Combing (B.2), Lemma B.1.2, (B.3), Lemma B.1.3 and B.1.4 and a union bound to finish the proof.

B.1.3 Proof of the first part of Theorem 3.3.1

Theorem B.1.1 (First part of Theorem 3.3.1)**.** *Let K be the number of episodes. Suppose $\sqrt{d}>\xi$, where $\xi := \sup_{V\in[0,H],\, s'\sim P_h(s,a),\, h\in[H]}\left|\frac{r_h+V(s')-(\mathcal{T}_hV)(s,a)}{\sigma_V(s,a)}\right|$ and $K>\max\{\mathcal{M}_1,\mathcal{M}_2,\mathcal{M}_3,\mathcal{M}_4\}$[2]. Then for any $0<\lambda<\kappa$, with probability $1-\delta$, for all policy π simultaneously, the output $\widehat{\pi}$ of*

[1] To be rigorous, Lemma C.11.4 needs to be modified since the absolute value bound and the variance bound here are in the high probability sense. However, this will not affect the validity of the result as the weaker version can also be obtained (see [157] and a related discussion in [3] Remark E.7.) To make the proof more readable, we do not include them here to avoid over-technicality.

[2] The definition of $\mathcal{M}_i$ is in List .

Algorithm 2 satisfies

$$v^{\pi} - v^{\widehat{\pi}} \leq \widetilde{O}\left(\sqrt{d} \cdot \sum_{h=1}^{H} \mathbb{E}_{\pi}\left[\left(\phi(\cdot,\cdot)^{\top}\Lambda_h^{-1}\phi(\cdot,\cdot)\right)^{1/2}\right]\right) + \frac{2H^4\sqrt{d}}{K}$$

where $\Lambda_h = \sum_{\tau=1}^{K} \frac{\phi(s_h^{\tau},a_h^{\tau})\cdot\phi(s_h^{\tau},a_h^{\tau})^{\top}}{\sigma^2_{\widehat{V}_{h+1}(s_h^{\tau},a_h^{\tau})}} + \lambda I_d$ *and* $\widetilde{O}$ *absorbs the universal constants and the Polylog terms.*

Proof: [Proof of Theorem B.1.1] Combing Lemma C.3.2 and Lemma B.1.5, we directly have with probability $1-\delta$, for all policy π simultaneously,

$$V_1^{\pi}(s) - V_1^{\widehat{\pi}}(s) \leq \widetilde{O}\left(\sqrt{d} \cdot \sum_{h=1}^{H} \mathbb{E}_{\pi}\left[\left(\phi(\cdot,\cdot)^{\top}\Lambda_h^{-1}\phi(\cdot,\cdot)\right)^{1/2}\Big| s_1 = s\right]\right) + \frac{2H^4\sqrt{d}}{K}, \tag{B.5}$$

now take the initial distribution d_1 on both sides to get the stated result.

B.1.4 Two Intermediate results

The next two lemmas provide intermediate results in finishing the whole proofs.

Bounding the variance

Lemma B.1.6. *Recall the definition* $\widehat{\sigma}_h(\cdot,\cdot)^2 = \max\{1, \widehat{\mathrm{Var}}_{P_h}\widehat{V}_{h+1}(\cdot,\cdot)\} + 1$ *and* $\sigma_{\widehat{V}_{h+1}}(\cdot,\cdot)^2 := \max\{1, \mathrm{Var}_{P_h}\widehat{V}_{h+1}(\cdot,\cdot)\}+1$. *Moreover,* $\left[\widehat{\mathrm{Var}}_h\widehat{V}_{h+1}\right](\cdot,\cdot) = \left\langle \phi(\cdot,\cdot), \bar{\beta}_h\right\rangle_{[0,(H-h+1)^2]} - \left[\left\langle \phi(\cdot,\cdot), \bar{\theta}_h\right\rangle_{[0,H-h+1]}\right]^2$ *(where* $\bar{\beta}_h$ *and* $\bar{\theta}_h$ *are defined in Algorithm 2). Let* $K \geq \max\left\{512(1/\kappa)^2\log\left(\frac{4Hd}{\delta}\right), 4\lambda/\kappa\right\}$, *then with probability* $1-\delta$,

$$\sup_h ||\widehat{\sigma}_h^2 - \sigma^2_{\widehat{V}_{h+1}}||_{\infty} \leq 36\sqrt{\frac{H^4d^3}{\kappa K}\log\left(\frac{(\lambda+K)2KdH^2}{\lambda\delta}\right)} + 12\lambda\frac{H^2\sqrt{d}}{\kappa K}.$$

Proof: **Step1:** we first show for all $h, s, a \in [H] \times \mathcal{S} \times \mathcal{A}$, with probability $1-\delta$

$$\left|\langle \phi(s,a), \bar{\beta}_h \rangle_{[0,(H-h+1)^2]} - \mathbb{P}_h(\widehat{V}_{h+1})^2(s,a)\right| \leq 12\sqrt{\frac{H^4 d^3}{\kappa K}\log\left(\frac{(\lambda+K)2KdH^2}{\lambda\delta}\right)} + 4\lambda\frac{H^2\sqrt{d}}{\kappa K}.$$

Proof of Step1. Note

$$\begin{aligned}
&\left|\langle \phi(s,a), \bar{\beta}_h \rangle_{[0,(H-h+1)^2]} - \mathbb{P}_h(\widehat{V}_{h+1})^2(s,a)\right| \leq \left|\langle \phi(s,a), \bar{\beta}_h \rangle - \mathbb{P}_h(\widehat{V}_{h+1})^2(s,a)\right| \\
=&\left|\phi(s,a)^\top \bar{\Sigma}_h^{-1} \sum_{\tau=1}^K \phi(\bar{s}_h^\tau, \bar{a}_h^\tau)\cdot \widehat{V}_{h+1}(\bar{s}_{h+1}^\tau)^2 - \mathbb{P}_h(\widehat{V}_{h+1})^2(s,a)\right| \\
=&\left|\phi(s,a)^\top \bar{\Sigma}_h^{-1} \sum_{\tau=1}^K \phi(\bar{s}_h^\tau, \bar{a}_h^\tau)\cdot \widehat{V}_{h+1}(\bar{s}_{h+1}^\tau)^2 - \phi(s,a)^\top \int_{\mathcal{S}} (\widehat{V}_{h+1})^2(s')d\nu_h(s')\right| \\
=&\left|\phi(s,a)^\top \bar{\Sigma}_h^{-1} \sum_{\tau=1}^K \phi(\bar{s}_h^\tau, \bar{a}_h^\tau)\cdot \widehat{V}_{h+1}(\bar{s}_{h+1}^\tau)^2 - \phi(s,a)^\top \bar{\Sigma}_h^{-1}(\sum_{\tau=1}^K \phi(\bar{s}_h^\tau, \bar{a}_h^\tau)\phi(\bar{s}_h^\tau, \bar{a}_h^\tau)^\top + \lambda I)\int_{\mathcal{S}} (\widehat{V}_{h+1})^2(s')d\nu_h(s')\right| \\
\leq&\underbrace{\left|\phi(s,a)^\top \bar{\Sigma}_h^{-1} \sum_{\tau=1}^K \phi(\bar{s}_h^\tau, \bar{a}_h^\tau)\cdot \left(\widehat{V}_{h+1}(\bar{s}_{h+1}^\tau)^2 - \mathbb{P}_h(\widehat{V}_{h+1})^2(\bar{s}_h^\tau, \bar{a}_h^\tau)\right)\right|}_{1} + \underbrace{\lambda\left|\phi(s,a)^\top \bar{\Sigma}_h^{-1}\int_{\mathcal{S}} (\widehat{V}_{h+1})^2(s')d\nu_h(s')\right|}_{2}
\end{aligned}$$

For 2, since $K \geq \max\left\{512(1/\kappa)^2\log\left(\frac{4Hd}{\delta}\right), 4\lambda/\kappa\right\}$, by Lemma C.11.5 and a union bound over $h \in [H]$, with probability $1-\delta$ for all $h, s, a \in [H]\times\mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}
2 \leq& \lambda \left\|\phi(s,a)\right\|_{\bar{\Sigma}_h^{-1}} \left\|\int_{\mathcal{S}} (\widehat{V}_{h+1})^2(s')d\nu_h(s')\right\|_{\bar{\Sigma}_h^{-1}} \\
\leq& \lambda\frac{2}{\sqrt{K}}\left\|\phi(s,a)\right\|_{(\Sigma_h^p)^{-1}} \frac{2}{\sqrt{K}}\left\|\int_{\mathcal{S}} (\widehat{V}_{h+1})^2(s')d\nu_h(s')\right\|_{(\Sigma_h^p)^{-1}} \leq 4\lambda\left\|(\Sigma_h^p)^{-1}\right\|\frac{H^2\sqrt{d}}{K} \leq 4\lambda\frac{H^2\sqrt{d}}{\kappa K}.
\end{aligned} \tag{B.6}$$

For 1, we have

$$1 \leq \left\|\phi(s,a)\right\|_{\bar{\Sigma}_h^{-1}}\left\|\sum_{\tau=1}^K \phi(\bar{s}_h^\tau, \bar{a}_h^\tau)\cdot\left(\widehat{V}_{h+1}(\bar{s}_{h+1}^\tau)^2 - \mathbb{P}_h(\widehat{V}_{h+1})^2(\bar{s}_h^\tau, \bar{a}_h^\tau)\right)\right\|_{\bar{\Sigma}_h^{-1}} \tag{B.7}$$

Bounding using covering. Note for any fix V_{h+1}, we can define $x_\tau = \phi(\bar{s}_h^\tau, \bar{a}_h^\tau)$ ($\|\phi\|_2 \leq 1$) and

$\eta_\tau = V_{h+1}(\bar{s}^\tau_{h+1})^2 - \mathbb{P}_h(V_{h+1})^2(\bar{s}^\tau_h, \bar{a}^\tau_h)$ is H^2-subgaussian, by Lemma C.11.3 (where $t = K$ and $L = 1$) with probability $1-\delta$,

$$\left\| \sum_{\tau=1}^{K} \phi(\bar{s}^\tau_h, \bar{a}^\tau_h) \cdot \left(V_{h+1}(\bar{s}^\tau_{h+1})^2 - \mathbb{P}_h(V_{h+1})^2(\bar{s}^\tau_h, \bar{a}^\tau_h) \right) \right\|_{\bar{\Sigma}_h^{-1}} \leq \sqrt{8H^4 \cdot \frac{d}{2} \log\left(\frac{\lambda + K}{\lambda\delta} \right)}$$

let $\mathcal{N}_h(\epsilon)$ be the minimal ϵ-cover (with respect the supremum norm) of $\mathcal{V}_h := \{V_h : V_h(\cdot) = \max_{a\in\mathcal{A}} \left\{ \min\{\phi(s,a)^\top \theta - C_1 \sqrt{d \cdot \phi(\cdot,\cdot)^\top \widehat{\Lambda}_h^{-1} \phi(\cdot,\cdot)} - C_2, H - h + 1\}^+ \right\}$. That is, for any $V \in \mathcal{V}_h$, there exists a value function $V' \in \mathcal{N}_h(\epsilon)$ such that $\sup_{s\in\mathcal{S}} |V(s) - V'(s)| < \epsilon$. Now by a union bound, we obtain with probability $1-\delta$

$$\sup_{V_{h+1}\in\mathcal{N}_{h+1}(\epsilon)} \left\| \sum_{\tau=1}^{K} \phi(\bar{s}^\tau_h, \bar{a}^\tau_h) \cdot \left(V_{h+1}(\bar{s}^\tau_{h+1})^2 - \mathbb{P}_h(V_{h+1})^2(\bar{s}^\tau_h, \bar{a}^\tau_h) \right) \right\|_{\bar{\Sigma}_h^{-1}} \leq \sqrt{8H^4 \cdot \frac{d}{2} \log\left(\frac{\lambda + K}{\lambda\delta} |\mathcal{N}_{h+1}(\epsilon)| \right)}$$

which implies

$$\begin{aligned} &\left\| \sum_{\tau=1}^{K} \phi(\bar{s}^\tau_h, \bar{a}^\tau_h) \cdot \left(\widehat{V}_{h+1}(\bar{s}^\tau_{h+1})^2 - \mathbb{P}_h(\widehat{V}_{h+1})^2(\bar{s}^\tau_h, \bar{a}^\tau_h) \right) \right\|_{\bar{\Sigma}_h^{-1}} \\ \leq &\sqrt{8H^4 \cdot \frac{d}{2} \log\left(\frac{\lambda + K}{\lambda\delta} |\mathcal{N}_{h+1}(\epsilon)| \right)} + 4H^2 \sqrt{\epsilon^2 K^2/\lambda} \end{aligned}$$

choosing $\epsilon = d\sqrt{\lambda}/K$, applying Lemma B.3 of [80][3] to the covering number $\mathcal{N}_{h+1}(\epsilon)$ w.r.t. $\mathcal{V}_{h+1}$, we can further bound above by

$$\leq \sqrt{8H^4 \cdot \frac{d^3}{2} \log\left(\frac{\lambda + K}{\lambda\delta} 2dHK \right)} + 4H^2\sqrt{d^2} \leq 6\sqrt{H^4 \cdot d^3 \log\left(\frac{\lambda + K}{\lambda\delta} 2dHK \right)}$$

[3] Note the same result in [80] applies even though we have an extra constant C_2.

Apply a union bound for $h \in [H]$, we have with probability $1-\delta$, for all $h \in [H]$,

$$\left\| \sum_{\tau=1}^{K} \phi(\bar{s}_h^\tau, \bar{a}_h^\tau) \cdot \left(\widehat{V}_{h+1}(\bar{s}_{h+1}^\tau)^2 - \mathbb{P}_h(\widehat{V}_{h+1})^2(\bar{s}_h^\tau, \bar{a}_h^\tau) \right) \right\|_{\bar{\Sigma}_h^{-1}} \leq 6\sqrt{H^4 d^3 \log\left(\frac{(\lambda+K)2KdH^2}{\lambda\delta} \right)} \tag{B.8}$$

and similar to 2, with probability $1-\delta$ for all $h, s, a \in [H] \times \mathcal{S} \times \mathcal{A}$,

$$\|\phi(s,a)\|_{\bar{\Sigma}_h^{-1}} \leq \frac{2\left\|(\Sigma_h^p)^{-1}\right\|^{1/2}}{\sqrt{K}} \leq \frac{2}{\sqrt{\kappa K}}. \tag{B.9}$$

Combing (B.6), (B.7), (B.8) and (B.9) we obtain with probability $1-\delta$ for all $h, s, a \in [H] \times \mathcal{S} \times \mathcal{A}$,

$$\left| \langle \phi(s,a), \bar{\beta}_h \rangle_{[0,(H-h+1)^2]} - \mathbb{P}_h(\widehat{V}_{h+1})^2(s,a) \right| \leq 12\sqrt{\frac{H^4 d^3}{\kappa K} \log\left(\frac{(\lambda+K)2KdH^2}{\lambda\delta} \right)} + 4\lambda \frac{H^2\sqrt{d}}{\kappa K}.$$

Step2: we show for all $h, s, a \in [H] \times \mathcal{S} \times \mathcal{A}$, with probability $1-\delta$

$$\left| \langle \phi(s,a), \bar{\theta}_h \rangle_{[0,H-h+1]} - \mathbb{P}_h(\widehat{V}_{h+1})(s,a) \right| \leq 12\sqrt{\frac{H^2 d^3}{\kappa K} \log\left(\frac{(\lambda+K)2KdH^2}{\lambda\delta} \right)} + 4\lambda \frac{H\sqrt{d}}{\kappa K}. \tag{B.10}$$

The proof of Step2 follows nearly the identical way as Step1 except $\widehat{V}_h^2$ is replaced by $\widehat{V}_h$.

Step3: We prove $\sup_h ||\widehat{\sigma}_h^2 - \sigma_{\widehat{V}_h}^2||_\infty \leq 36\sqrt{\frac{H^4 d^3}{\kappa K} \log\left(\frac{(\lambda+K)2KdH^2}{\lambda\delta} \right)} + 12\lambda \frac{H^2\sqrt{d}}{\kappa K}$.

Proof of Step3. By (B.10),

$$\begin{aligned}
&\left| \left[\left\langle \phi(\cdot,\cdot), \bar{\theta}_h \right\rangle_{[0,H-h+1]} \right]^2 - \left[\mathbb{P}_h(\widehat{V}_{h+1})(s,a) \right]^2 \right| \\
=&\left| \langle \phi(s,a), \bar{\theta}_h \rangle_{[0,H-h+1]} + \mathbb{P}_h(\widehat{V}_{h+1})(s,a) \right| \cdot \left| \langle \phi(s,a), \bar{\theta}_h \rangle_{[0,H-h+1]} - \mathbb{P}_h(\widehat{V}_{h+1})(s,a) \right| \\
\leq& 2H \cdot \left| \langle \phi(s,a), \bar{\theta}_h \rangle_{[0,H-h+1]} - \mathbb{P}_h(\widehat{V}_{h+1})(s,a) \right| \leq 24\sqrt{\frac{H^4 d^3}{\kappa K} \log\left(\frac{(\lambda+K)2KdH^2}{\lambda\delta} \right)} + 8\lambda \frac{H^2\sqrt{d}}{\kappa K}.
\end{aligned}$$

Combining this with Step1 we receive $\forall h, s, a \in [H] \times \mathcal{S} \times \mathcal{A}$, with probability $1-\delta$

$$\left|\widehat{\mathrm{Var}}_h \widehat{V}_{h+1}(s,a) - \mathrm{Var}_{P_h} \widehat{V}_{h+1}(s,a)\right| \leq 36\sqrt{\frac{H^4 d^3}{\kappa K} \log\left(\frac{(\lambda+K)2KdH^2}{\lambda\delta}\right)} + 12\lambda\frac{H^2\sqrt{d}}{\kappa K}.$$

Finally, by the non-expansiveness of operator $\max\{1, \cdot\}$, we have the stated result.

A crude bound on $\sup_h ||V_h^\star - \widehat{V}_h||_\infty$**.**

Lemma B.1.7. *Define* $\widehat{\sigma}_h(s,a) = \sqrt{\max\left\{1, \widehat{\mathrm{Var}}_{P_h} \widehat{V}_{h+1}(s,a)\right\} + 1}$*, if* $K \geq \max\{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4\}$ *and* $K > C \cdot H^4\kappa^2$*, then with probability at least* $1-\delta$*,*

$$\sup_h \left\|V_h^\star - \widehat{V}_h\right\|_\infty \leq \widetilde{O}\left(\frac{H^2\sqrt{d}}{\sqrt{\kappa K}}\right).$$

Proof: **Step1:** We show with probability at least $1-\delta$, $\sup_h \left\|V_h^\star - V_h^{\widehat{\pi}}\right\|_\infty \leq \widetilde{O}\left(\frac{H^2\sqrt{d}}{\sqrt{\kappa K}}\right)$.

Indeed, combing Lemma C.3.2 and Lemma B.1.5, similar to the proof of Theorem B.1.1, we directly have with probability $1-\delta$, for all policy π simultaneously, and for all $s \in \mathcal{S}$, $h \in [H]$

$$V_h^\pi(s) - V_h^{\widehat{\pi}}(s) \leq \widetilde{O}\left(\sqrt{d} \cdot \sum_{t=h}^{H} \mathbb{E}_\pi\left[\left(\phi(\cdot,\cdot)^\top \Lambda_t^{-1} \phi(\cdot,\cdot)\right)^{1/2} \middle| s_h = s\right]\right) + \frac{2H^4\sqrt{d}}{K}, \quad \text{(B.11)}$$

Next, since $K \geq \max\left\{512(1/\kappa)^2 \log\left(\frac{4Hd}{\delta}\right), 4\lambda/\kappa\right\}$, by Lemma C.11.5 and a union bound over $h \in [H]$, with probability $1-\delta$

$$\sup_{s,a} \|\phi(s,a)\|_{\widehat{\Lambda}_h^{-1}} \leq \frac{2}{\sqrt{K}} \sup_{s,a} \|\phi(s,a)\|_{\Lambda_h^{p-1}} \leq \frac{2H}{\sqrt{\kappa K}}, \quad \forall h \in [H].$$

Lastly, taking $\pi = \pi^\star$ in (B.11) to obtain

$$\begin{aligned} 0 \le V_h^{\pi^\star}(s) - V_h^{\hat{\pi}}(s) \le & \widetilde{O}\left(\sqrt{d} \cdot \sum_{t=h}^{H} \mathbb{E}_{\pi^\star}\left[\left(\phi(\cdot,\cdot)^\top \Lambda_t^{-1} \phi(\cdot,\cdot) \right)^{1/2} \Big| s_h = s \right] \right) + \frac{2H^4\sqrt{d}}{K} \\ \le & \widetilde{O}\left(\frac{H^2\sqrt{d}}{\sqrt{\kappa K}} \right) + \frac{2H^4\sqrt{d}}{K}. \end{aligned} \tag{B.12}$$

This implies by using the condition $K > C \cdot H^4\kappa^2$, we finish the proof of Step1.

Step2: We show with probability $1-\delta$, $\sup_h \left\| \widehat{V}_h - V_h^{\hat{\pi}} \right\|_\infty \le \widetilde{O}\left(\frac{H^2\sqrt{d}}{\sqrt{\kappa K}} \right)$.

Indeed, applying Extended Value Difference Lemma D.0.7 for $\pi = \pi' = \hat{\pi}$, then with probability $1-\delta$, for all s, h

$$\begin{aligned} \left| \widehat{V}_h(s) - V_h^{\hat{\pi}}(s) \right| &= \left| \sum_{t=h}^{H} \mathbb{E}_{\hat{\pi}} \left[\widehat{Q}_h(s_h,a_h) - \left(\mathcal{T}_h \widehat{V}_{h+1} \right)(s_h,a_h) \Big| s_h = s \right] \right| \\ \le & \sum_{t=h}^{H} \left\| (\widehat{\mathcal{T}}_h \widehat{V}_{h+1} - \mathcal{T}_h \widehat{V}_{h+1})(s,a) \right\| + \left\| \Gamma_h(s,a) \right\| \\ \le & \widetilde{O}\left(H\sqrt{d} \left\| \sqrt{\phi(s,a)\Lambda_h^{-1}\phi(s,a)} \right\| \right) + \frac{4H^4\sqrt{d}}{K} \le \widetilde{O}\left(\frac{H^2\sqrt{d}}{\sqrt{\kappa K}} \right) \end{aligned}$$

where the second inequality uses Lemma B.1.5[4] and the last inequality follows the same procedure as Step1.

Step3: Combine Step1 and Step2, by triangular inequality and a union bound we finish the proof of the lemma.

Remark 7. *Note as an intermediate calculation,* (B.12) *ensures a learning bound with order*

[4]To be absolutely rigorous, we cannot directly apply Lemma B.1.5 here since the crude bound has already been used in Lemma B.1.4. However, this can be resolved completely by first deriving an even cruder bound for $\sup_h ||V_h^\star - \widehat{V}_h||_\infty$ that has $1/\sqrt{K}$ rate without using Lemma C.5 (which we call it Lemma C.8*), and we can use Lemma C.8* to show a similar result Lemma C.5*. Finally, we can use Lemma C.5* here to finish the proof of this Lemma B.1.7. However, we avoid explicitly doing this to prevent over-technicality.

$\widetilde{O}(\frac{H^2\sqrt{d}}{\sqrt{\kappa K}})$. *Here, the convergence rate is the standard statistical rate* $\frac{1}{\sqrt{K}}$ *and the* H^2 *dependence is loose. However, the feature dependence* $\sqrt{d/\kappa}$ *is roughly tight, since, in the well-explored case (Assumption 2 of [90]),* $\kappa = 1/d$ *and the* $\sqrt{d/\kappa} = \sqrt{d^2}$ *recovers the optimal feature dependence* $dH\sqrt{T}$ *in the online setting [76]. If* $\kappa \ll 1/d$*, then doing offline learning requires sample size proportional to* d/κ*, which reveals offline RL is harder when the exploration of behavior policy is insufficient. When* $\kappa = 0$*, learning the optimal policy accurately cannot be guaranteed even if the sample/episode size* $K \to \infty$*.*

B.1.5 Proof of the second part of Theorem 3.3.1

Lemma B.1.8. *Recall* $\widehat{\sigma}_h = \sqrt{\max\left\{1, \widehat{\mathrm{Var}}_{P_h}\widehat{V}_{h+1}\right\} + 1}$ *and* $\sigma_h^\star = \sqrt{\max\left\{1, \mathrm{Var}_{P_h}V_{h+1}^\star\right\} + 1}$. *Let* $K \geq \max\left\{512(1/\kappa)^2\log\left(\frac{4Hd}{\delta}\right), 4\lambda/\kappa\right\}$ *and* $K \geq \max\{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4\}$*, then with probability* $1-\delta$*,*

$$\sup_h ||\widehat{\sigma}_h^2 - \sigma_h^{\star 2}||_\infty \leq \widetilde{O}\left(\frac{H^3\sqrt{d}}{\sqrt{\kappa K}}\right).$$

Proof: By definition and the non-expansiveness of $\max\{1, \cdot\} + 1$, we have

$$\begin{aligned}
&\left\|\sigma^2_{\widehat{V}_{h+1}} - \sigma_h^{\star 2}\right\|_\infty \leq \left\|\mathrm{Var}\widehat{V}_{h+1} - \mathrm{Var}V_{h+1}^\star\right\|_\infty \\
\leq &\left\|\mathbb{P}_h\left(\widehat{V}_{h+1}^2 - V_{h+1}^{\star 2}\right)\right\|_\infty + \left\|(\mathbb{P}_h\widehat{V}_{h+1})^2 - (\mathbb{P}_hV_{h+1}^\star)^2\right\|_\infty \\
\leq &\left\|\widehat{V}_{h+1}^2 - V_{h+1}^{\star 2}\right\|_\infty + \left\|(\mathbb{P}_h\widehat{V}_{h+1} + \mathbb{P}_hV_{h+1}^\star)(\mathbb{P}_h\widehat{V}_{h+1} - \mathbb{P}_hV_{h+1}^\star)\right\|_\infty \\
\leq &2H\left\|\widehat{V}_{h+1} - V_{h+1}^\star\right\|_\infty + 2H\left\|\mathbb{P}_h\widehat{V}_{h+1} - \mathbb{P}_hV_{h+1}^\star\right\|_\infty \leq \widetilde{O}\left(\frac{H^3\sqrt{d}}{\sqrt{\kappa K}}\right).
\end{aligned}$$

with probability $1-\delta$ for all $h \in [H]$, where the last inequality comes from Lemma B.1.7. Combining this with Lemma B.1.6, we have the stated result.

Lemma B.1.9. *Denote the quantities* $C_1 = \max\{2\lambda, 128\log(2d/\delta), 128H^4\log(2d/\delta)/\kappa^2\}$ *and* $C_2 = \max\{\frac{\lambda^2}{\kappa\log((\lambda+K)H/\lambda\delta)}, 96^2H^{12}d\log((\lambda+K)H/\lambda\delta)/\kappa^5\}$*. Suppose the number of episode* K

satisfies $K > \max\{C_1, C_2\}$*, then with probability* $1-\delta$*,*

$$\sqrt{\phi(s,a)\Lambda_h^{-1}\phi(s,a)} \le 2\sqrt{\phi(s,a)\Lambda_h^{\star-1}\phi(s,a)}, \quad \forall s,a \in \mathcal{S}\times\mathcal{A},$$

Proof: [Proof of Lemma B.1.9]

By definition $\sqrt{\phi(s,a)\Lambda_h^{-1}\phi(s,a)} = \|\phi(s,a)\|_{\Lambda_h^{-1}}$. Then denote

$$\Lambda_h' = \frac{1}{K}\Lambda_h, \quad \Lambda_h^{\star\prime} = \frac{1}{K}\Lambda_h^{\star},$$

where $\Lambda_h = \sum_{\tau=1}^{K} \phi(s_h^\tau, a_h^\tau)^\top \phi(s_h^\tau, a_h^\tau)/\sigma^2_{V^\star_{h+1}}(s_h^\tau, a_h^\tau) + \lambda I$. Under the condition of K, by Lemma B.1.8, with probability $1-\delta$

$$\begin{aligned}
\left\|\Lambda_h^{\star\prime} - \Lambda_h'\right\| &\le \sup_{s,a}\left\|\frac{\phi(s,a)\phi(s,a)^\top}{\sigma_h^{\star 2}(s,a)} - \frac{\phi(s,a)\phi(s,a)^\top}{\sigma^2_{\widehat{V}_{h+1}}(s,a)}\right\| \\
&\le \sup_{s,a}\left|\frac{\sigma_h^{\star 2}(s,a) - \sigma^2_{\widehat{V}_{h+1}}(s,a)}{\sigma_h^{\star 2}(s,a)\sigma^2_{\widehat{V}_{h+1}}(s,a)}\right| \cdot \|\phi(s,a)\|^2 \le \sup_{s,a}\left|\frac{\sigma_h^{\star 2}(s,a) - \sigma^2_{\widehat{V}_{h+1}}(s,a)}{1}\right| \cdot 1 \qquad \text{(B.13)} \\
&\le \widetilde{O}\left(\frac{H^3\sqrt{d}}{\sqrt{\kappa K}}\right).
\end{aligned}$$

Next by Lemma C.11.6 (with ϕ to be $\phi/\sigma_{V^\star_{h+1}}$ and $C=1$), it holds with probability $1-\delta$,

$$\left\|\Lambda_h^{\star\prime} - \left(\mathbb{E}_{\mu,h}[\phi(s,a)\phi(s,a)^\top/\sigma^2_{V^\star_{h+1}}(s,a)] + \frac{\lambda}{K}I_d\right)\right\| \le \frac{4\sqrt{2}}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2}.$$

Therefore by *Weyl's spectrum theorem* and the condition $K > \max\{2\lambda, 128\log(2d/\delta), 128H^4\log(2d/\delta)/\kappa^2\}$,

the above implies

$$
\begin{aligned}
\left\|\Lambda_h^{\star\prime}\right\| =& \lambda_{\max}(\Lambda_h^{\star\prime}) \leq \lambda_{\max}\left(\mathbb{E}_{\mu,h}[\phi(s,a)\phi(s,a)^\top/\sigma^2_{V^\star_{h+1}}(s,a)]\right) + \frac{\lambda}{K} + \frac{4\sqrt{2}}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2} \\
&\leq \left\|\mathbb{E}_{\mu,h}[\phi(s,a)\phi(s,a)^\top/\sigma^2_{V^\star_{h+1}}(s,a)]\right\| + \frac{\lambda}{K} + \frac{4\sqrt{2}}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2} \\
&\leq \|\phi(s,a)\|^2 + \frac{\lambda}{K} + \frac{4\sqrt{2}}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2} \leq 1 + \frac{\lambda}{K} + \frac{4\sqrt{2}}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2} \leq 2, \\
\lambda_{\min}(\Lambda_h^{\star\prime}) \geq& \lambda_{\min}\left(\mathbb{E}_{\mu,h}[\phi(s,a)\phi(s,a)^\top/\sigma^2_{V^\star_{h+1}}(s,a)]\right) + \frac{\lambda}{K} - \frac{4\sqrt{2}}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2} \\
&\geq \lambda_{\min}\left(\mathbb{E}_{\mu,h}[\phi(s,a)\phi(s,a)^\top/\sigma^2_{V^\star_{h+1}}(s,a)]\right) - \frac{4\sqrt{2}}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2} \\
&\geq \frac{\kappa}{H^2} - \frac{4\sqrt{2}}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2} \geq \frac{\kappa}{2H^2}.
\end{aligned}
$$

Hence with probability $1-\delta$, $\left\|\Lambda_h^{\star\prime}\right\| \leq 2$ and $\left\|\Lambda_h^{\star\prime-1}\right\| = 1/\lambda_{\min}(\Lambda_h^{\star\prime}) \leq 2H^2/\kappa$. Similarly, $\left\|\Lambda_h^{\prime-1}\right\| \leq 2H^2/\kappa$ with high probability.

Now apply Lemma C.11.7 to $\Lambda_h^{\star\prime}$ and Λ_h' and a union bound, we obtain with probability $1-\delta$, for all s, a

$$
\begin{aligned}
\|\phi(s,a)\|_{\Lambda_h^{\prime-1}} &\leq \left[1 + \sqrt{\left\|\Lambda_h^{\star\prime-1}\right\|\left\|\Lambda_h^{\star\prime}\right\| \cdot \left\|\Lambda_h^{\prime-1}\right\| \cdot \left\|\Lambda_h^{\star\prime} - \Lambda_h'\right\|}\right] \cdot \|\phi(s,a)\|_{\Lambda_h^{\star\prime-1}} \\
&\leq \left[1 + \sqrt{\frac{2H^2}{\kappa} \cdot 1 \cdot \frac{2H^2}{\kappa} \cdot \left\|\Lambda_h^{\star\prime} - \Lambda_h'\right\|}\right] \cdot \|\phi(s,a)\|_{\Lambda_h^{\star\prime-1}} \\
&\leq \left[1 + \sqrt{\frac{H^4}{\kappa^2}\left[\widetilde{O}\left(\frac{H^3\sqrt{d}}{\sqrt{\kappa K}}\right)\right]}\right] \cdot \|\phi(s,a)\|_{\Lambda_h^{\star\prime-1}} \leq 2\|\phi(s,a)\|_{\Lambda_h^{\star\prime-1}}
\end{aligned}
$$

where the third inequality uses (B.13) and the last inequality uses $K > \max\{\frac{\lambda^2}{\kappa\log((\lambda+K)H/\lambda\delta)}, 96^2H^{12}d\log((\lambda+K)H/\lambda\delta)/\kappa^5\}$. The claimed result follows straightforwardly by multiplying $1/\sqrt{K}$ on both

sides of the above.

Proof: [Proof of Theorem 3.3.1] The first part of the theorem has been shown in Theorem B.1.1. For the second part, apply Theorem B.1.1 with $\pi = \pi^\star$, then with probability $1-\delta$,

$$v^{\pi^\star} - v^{\hat{\pi}} \le \widetilde{O}\left(\sqrt{d}\cdot\sum_{h=1}^{H}\mathbb{E}_{\pi^\star}\left[\left(\phi(\cdot,\cdot)^\top\Lambda_h^{-1}\phi(\cdot,\cdot)\right)^{1/2}\right]\right) + \frac{2H^4\sqrt{d}}{K},$$

Now apply Lemma B.1.9 and a union bound, with probability $1-\delta$,

$$0 \le v^\star - v^{\hat{\pi}} \le \widetilde{O}\left(\sqrt{d}\cdot\sum_{h=1}^{H}\mathbb{E}_{\pi^\star}\left[\left(\phi(\cdot,\cdot)^\top\Lambda_h^{\star-1}\phi(\cdot,\cdot)\right)^{1/2}\right]\right) + \frac{2H^4\sqrt{d}}{K}.$$

B.2 Proof of Theorem 3.3.2

First of all, we show the following lemma.

Lemma B.2.1. *Suppose* $K > \max\{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4\}$. *Plug*

$$\Gamma_h^I(s,a) \leftarrow \phi(s,a)^\top\left|\widehat{\Lambda}_h^{-1}\sum_{\tau=1}^{K}\frac{\phi\left(s_h^\tau,a_h^\tau\right)\cdot\left(r_h^\tau+\widehat{V}_{h+1}\left(s_{h+1}^\tau\right)-\left(\widehat{\mathcal{T}}_h\widehat{V}_{h+1}\right)\left(s_h^\tau,a_h^\tau\right)\right)}{\widehat{\sigma}_h^2(s_h^\tau,a_h^\tau)}\right| + \widetilde{O}(\frac{H^3d/\kappa}{K})$$

in Algorithm 2 and let $\mathcal{T}_h$ *be the Bellman operator and* $\widehat{\mathcal{T}}_h$ *be the approximated Bellman operator. Then we have with probability* $1-\delta$*:*

$$|(\mathcal{T}_h\widehat{V}_{h+1} - \widehat{\mathcal{T}}_h\widehat{V}_{h+1})(s,a)| \le \Gamma_h^I(s,a), \quad \forall s,a \in \mathcal{S}\times\mathcal{A}.$$

Proof: [Proof of Lemma B.2.1] Suppose w_h is the coefficient corresponding to the $\mathcal{T}_h\widehat{V}_{h+1}$ (such w_h exists by Lemma B.5.9), *i.e.* $\mathcal{T}_h\widehat{V}_{h+1} = \phi^\top w_h$, and recall $(\widehat{\mathcal{T}}_h\widehat{V}_{h+1})(s,a) = \phi(s,a)^\top\widehat{w}_h$,

then:

$$
\begin{aligned}
&\left(\mathcal{T}_h\widehat{V}_{h+1}\right)(s,a)-\left(\widehat{\mathcal{T}}_h\widehat{V}_{h+1}\right)(s,a)=\phi(s,a)^\top\left(w_h-\widehat{w}_h\right)\\
=&\phi(s,a)^\top w_h-\phi(s,a)^\top\widehat{\Lambda}_h^{-1}\left(\sum_{\tau=1}^{K}\phi\left(s_h^\tau,a_h^\tau\right)\cdot\left(r_h^\tau+\widehat{V}_{h+1}\left(s_{h+1}^\tau\right)\right)/\widehat{\sigma}_h^2(s_h^\tau,a_h^\tau)\right)\\
=&\underbrace{\phi(s,a)^\top w_h-\phi(s,a)^\top\widehat{\Lambda}_h^{-1}\left(\sum_{\tau=1}^{K}\phi\left(s_h^\tau,a_h^\tau\right)\cdot\left(\mathcal{T}_h\widehat{V}_{h+1}\right)\left(s_h^\tau,a_h^\tau\right)/\widehat{\sigma}_h^2(s_h^\tau,a_h^\tau)\right)}_{\text{(i)}}\\
&\underbrace{+\phi(s,a)^\top\widehat{\Lambda}_h^{-1}\left(\sum_{\tau=1}^{K}\phi\left(s_h^\tau,a_h^\tau\right)\cdot\left(r_h^\tau+\widehat{V}_{h+1}\left(s_{h+1}^\tau\right)-\left(\widehat{\mathcal{T}}_h\widehat{V}_{h+1}\right)\left(s_h^\tau,a_h^\tau\right)\right)/\widehat{\sigma}_h^2(s_h^\tau,a_h^\tau)\right)}_{\text{(ii)}}\\
&\underbrace{+\phi(s,a)^\top\widehat{\Lambda}_h^{-1}\left(\sum_{\tau=1}^{K}\phi\left(s_h^\tau,a_h^\tau\right)\cdot\left(\left(\widehat{\mathcal{T}}_h\widehat{V}_{h+1}\right)\left(s_h^\tau,a_h^\tau\right)-\left(\mathcal{T}_h\widehat{V}_{h+1}\right)\left(s_h^\tau,a_h^\tau\right)\right)/\widehat{\sigma}_h^2(s_h^\tau,a_h^\tau)\right)}_{\text{(iii)}}
\end{aligned}
\tag{B.14}
$$

For term (i), by Lemma B.1.2 it is bounded by $\frac{2\lambda H^3\sqrt{d}/\kappa}{K}$ with probability $1-\delta/2$.[5]

For term (ii), it is bounded by

$$
\phi(s,a)^\top\left|\widehat{\Lambda}_h^{-1}\sum_{\tau=1}^{K}\frac{\phi\left(s_h^\tau,a_h^\tau\right)\cdot\left(r_h^\tau+\widehat{V}_{h+1}\left(s_{h+1}^\tau\right)-\left(\widehat{\mathcal{T}}_h\widehat{V}_{h+1}\right)\left(s_h^\tau,a_h^\tau\right)\right)}{\widehat{\sigma}_h^2(s_h^\tau,a_h^\tau)}\right|.
$$

[5]Note Here Lemma B.1.2 still applies even if the Γ_h changes since it works for all $\widehat{V}_h\in[0,H]$ so that $\left\|w_h\right\|_2\leq 2H\sqrt{d}$ and the truncation (Line 13 in Algorithm 2) guarantees this.

For term (iii), by Cauchy inequality

$$
\begin{aligned}
&\phi(s,a)^\top \widehat{\Lambda}_h^{-1} \left(\sum_{\tau=1}^{K} \phi\left(s_h^\tau, a_h^\tau\right) \cdot \left(\left(\widehat{\mathcal{T}}_h \widehat{V}_{h+1}\right)\left(s_h^\tau, a_h^\tau\right) - \left(\mathcal{T}_h \widehat{V}_{h+1}\right)\left(s_h^\tau, a_h^\tau\right) \right) / \widehat{\sigma}_h^2(s_h^\tau, a_h^\tau) \right) \\
\leq & \|\phi(s,a)\|_{\widehat{\Lambda}_h^{-1}} \cdot \left\| \sum_{\tau=1}^{K} \phi\left(s_h^\tau, a_h^\tau\right) \cdot \left(\left(\widehat{\mathcal{T}}_h \widehat{V}_{h+1}\right)\left(s_h^\tau, a_h^\tau\right) - \left(\mathcal{T}_h \widehat{V}_{h+1}\right)\left(s_h^\tau, a_h^\tau\right) \right) / \widehat{\sigma}_h^2(s_h^\tau, a_h^\tau) \right\|_{\widehat{\Lambda}_h^{-1}} \\
\leq & \frac{2H}{\sqrt{\kappa K}} \cdot \left\| \sum_{\tau=1}^{K} \phi\left(s_h^\tau, a_h^\tau\right) \cdot \left(\left(\widehat{\mathcal{T}}_h \widehat{V}_{h+1}\right)\left(s_h^\tau, a_h^\tau\right) - \left(\mathcal{T}_h \widehat{V}_{h+1}\right)\left(s_h^\tau, a_h^\tau\right) \right) / \widehat{\sigma}_h^2(s_h^\tau, a_h^\tau) \right\|_{\widehat{\Lambda}_h^{-1}} \\
\leq & \frac{2H}{\sqrt{\kappa K}} \cdot \widetilde{O}\left(\frac{H^2\sqrt{d/\kappa}}{\sqrt{K}}\right) \cdot \sqrt{d} = \widetilde{O}\left(\frac{H^3 d/\kappa}{K}\right)
\end{aligned}
$$

where the first inequality is by Lemma C.11.5 (with $\phi' = \phi/\widehat{\sigma}_h$ and $\left\|\phi/\widehat{\sigma}_h\right\| \leq \|\phi\| \leq 1 := C$) and the third inequality uses $\sqrt{a^\top \cdot A \cdot a} \leq \sqrt{\|a\|_2 \|A\|_2 \|a\|_2} = \|a\|_2 \sqrt{\|A\|_2}$ with a to be either ϕ or w_h. Moreover, $\lambda_{\min}(\tilde{\Lambda}_h^p) \geq \kappa / \max_{h,s,a} \widehat{\sigma}_h(s,a)^2 \geq \kappa/H^2$ implies $\left\|(\tilde{\Lambda}_h^p)^{-1}\right\| \leq H^2/\kappa$.

The second inequality is true by denoting $x_\tau = \phi(s_h^\tau, a_h^\tau)/\widehat{\sigma}(s_h^\tau, a_h^\tau)$ and

$$
\eta_\tau = \left(\left(\widehat{\mathcal{T}}_h \widehat{V}_{h+1}\right)\left(s_h^\tau, a_h^\tau\right) - \left(\mathcal{T}_h \widehat{V}_{h+1}\right)\left(s_h^\tau, a_h^\tau\right) \right) / \widehat{\sigma}_h(s_h^\tau, a_h^\tau)
$$

and use Lemma B.5.10 as the condition for applying Lemma C.11.3. By collecting those three terms together we have the result.

B.2.1 Proof of Theorem 3.3.2

Proof: Use Lemma B.2.1 as the condition for Lemma C.3.2 and average over initial distribution d_1, we obtain with probability $1-\delta$,

$$v^{\pi} - v^{\hat{\pi}} \leq$$
$$\sum_{h=1}^{H} \mathbb{E}_{\pi_h}[\phi(s,a)]^{\top} \left| \widehat{\Lambda}_h^{-1} \sum_{\tau=1}^{K} \frac{\phi\left(s_h^{\tau}, a_h^{\tau}\right) \cdot \left(r_h^{\tau} + \widehat{V}_{h+1}\left(s_{h+1}^{\tau}\right) - \left(\widehat{\mathcal{T}}_h \widehat{V}_{h+1}\right)\left(s_h^{\tau}, a_h^{\tau}\right)\right)}{\widehat{\sigma}_h^2(s_h^{\tau}, a_h^{\tau})} \right| + \widetilde{O}(\frac{H^4 d/\kappa}{K}) \tag{B.15}$$

Denote $A_h := \sum_{\tau=1}^{K} \frac{\phi\left(s_h^{\tau}, a_h^{\tau}\right) \cdot \left(r_h^{\tau} + \widehat{V}_{h+1}\left(s_{h+1}^{\tau}\right) - \left(\mathcal{T}_h \widehat{V}_{h+1}\right)\left(s_h^{\tau}, a_h^{\tau}\right)\right)}{\widehat{\sigma}_h^2(s_h^{\tau}, a_h^{\tau})}$, then

$$\mathbb{E}_{\pi_h}[\phi(s,a)]^{\top} \left| \widehat{\Lambda}_h^{-1} \sum_{\tau=1}^{K} \frac{\phi\left(s_h^{\tau}, a_h^{\tau}\right) \cdot \left(r_h^{\tau} + \widehat{V}_{h+1}\left(s_{h+1}^{\tau}\right) - \left(\widehat{\mathcal{T}}_h \widehat{V}_{h+1}\right)\left(s_h^{\tau}, a_h^{\tau}\right)\right)}{\widehat{\sigma}_h^2(s_h^{\tau}, a_h^{\tau})} \right|$$
$$\leq \mathbb{E}_{\pi_h}[\phi]^{\top} \cdot \left| \widehat{\Lambda}_h^{-1} A_h \right| + \mathbb{E}_{\pi_h}[\phi]^{\top} \left| \widehat{\Lambda}_h^{-1} \sum_{\tau=1}^{K} \frac{\phi\left(s_h^{\tau}, a_h^{\tau}\right) \cdot \left(\mathcal{T}_h \widehat{V}_{h+1}\left(s_h^{\tau}, a_h^{\tau}\right) - \widehat{\mathcal{T}}_h \widehat{V}_{h+1}\left(s_h^{\tau}, a_h^{\tau}\right)\right)}{\widehat{\sigma}_h^2(s_h^{\tau}, a_h^{\tau})} \right|$$

For the second term, it can be bounded similar to term (iii) in Lemma B.2.1 and for the first term we have the following:

$$\mathbb{E}_{\pi_h}[\phi]^{\top} \cdot \left| \widehat{\Lambda}_h^{-1} A_h \right| = \mathbb{E}_{\pi_h}[\phi]^{\top} \cdot \widehat{\Lambda}_h^{-1} \cdot \widehat{\Lambda}_h \left| \widehat{\Lambda}_h^{-1} A_h \right| \leq \left\| \mathbb{E}_{\pi_h}[\phi] \right\|_{\widehat{\Lambda}_h^{-1}} \cdot \left\| \widehat{\Lambda}_h | \widehat{\Lambda}_h^{-1} A_h | \right\|_{\widehat{\Lambda}_h^{-1}}$$
$$\leq \left\| \mathbb{E}_{\pi_h}[\phi] \right\|_{\widehat{\Lambda}_h^{-1}} \cdot \left\| |A_h| \right\|_{\widehat{\Lambda}_h^{-1}} \leq \widetilde{O}(\sqrt{d}) \left\| \mathbb{E}_{\pi_h}[\phi] \right\|_{\widehat{\Lambda}_h^{-1}} \leq \widetilde{O}(\sqrt{d} \left\| \mathbb{E}_{\pi_h}[\phi] \right\|_{\Lambda_h^{-1}}),$$

where the first inequality uses Cauchy's inequality, the second inequality uses $\widehat{\Lambda}_h$ is coordinate-wise positive (since we assume here $\phi \geq 0$), the third inequality is identical to the analysis in Section B.1.2 and the fourth inequality is identical to the analysis in Section B.1.2 with ϕ replaced by $\mathbb{E}[\phi]$. Plug this back to (B.15) we finish the proof for the first part. For the second

part, converting Λ_h^{-1} to $\Lambda_h^{\star-1}$ is identical to Section B.1.5. This finishes the proof.

B.3 Proof of Minimax Lower bound Theorem 3.3.4

The proof follows the lower bound proof of zanette2021provable. For completeness, we provide all the details in below.

B.3.1 Construction

Similar to the proof of [zanette2021provable, Theorem 2], we construct a family of MDPs, each parameterized by a Boolean vector $u = (u_1, \dots, u_H)$ with each $u_h \in \{-1, +1\}^{d-2}$ for $h \in [H]$. The MDPs share the same transition kernel and are only different in the reward observations.

State space: At each time step h, there are two states $\mathcal{S} = \{+1, -1\}$.

Action space: The action space $\mathcal{A} = \{-1, 0, +1\}^{d-2}$.

Feature map: The feature map $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ is given by

$$\phi(+1, a) = \begin{pmatrix} \frac{a}{\sqrt{2d}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{pmatrix} \in \mathbb{R}^d, \qquad \phi(-1, a) = \begin{pmatrix} \frac{a}{\sqrt{2d}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{pmatrix} \in \mathbb{R}^d.$$

The construction ensures the condition $\|\phi(s, a)\|_2 \leq 1$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Transition kernel: The transition probability $P_h(s' \mid s, a)$ is independent of action a. In other words, the Markov decision process reduces to a homogeneous Markov chain with tran-

sition matrix

$$\mathbf{P} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \in \mathbb{R}^2.$$

By letting

$$\nu_h(+1) = \nu_h(-1) = \begin{pmatrix} \mathbf{0}_{d-2} \\ \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \in \mathbb{R}^d,$$

we have $P_h(s' \mid s, a) = \langle \phi(s,a), \nu_h(s') \rangle$ to be a valid probability transition.

Reward observations: For any MDP M_u, at each times step h, the reward follows a Gaussian distribution with

$$R_{u,h}(s,a) \sim \mathcal{N}\left(\frac{s}{\sqrt{6}} + \frac{\delta}{\sqrt{2d}} \langle a, u_h \rangle, 1 \right),$$

where $\delta \in \left[0, \frac{1}{\sqrt{3d}}\right]$ determines to what extent the MDP models are different from each other. The mean reward function satisfies $r_{u,h}(s,a) = \langle \phi(s,a), \theta_{u,h} \rangle$ with

$$\theta_{u,h} = \begin{pmatrix} \delta u_h \\ \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{3}} \end{pmatrix} \in \mathbb{R}^d.$$

Offline data collection Scheme: The dataset $\mathcal{D} = \{(s_h^\tau, a_h^\tau, r_h^\tau, s_{h+1}^\tau)\}_{\tau \in [K]}^{h \in [H]}$ consist of K i.i.d. trajectories. All the trajectories initiate from uniform distribution. We take a behavior policy $\mu(\cdot \mid s)$ that is independent of state s. Let $\{e_1, e_2, \ldots, e_{d-2}\}$ be the canonical bases

of $\mathbb{R}^{d-2}$ and $\mathbf{0}_{d-2} \in \mathbb{R}^{d-2}$ be the zero vector. The behavior policy μ is set as

$$\mu(e_j \mid s) = \tfrac{1}{d} \quad \text{for any } j \in [d-2] \qquad \text{and} \qquad \mu(\mathbf{0}_{d-2} \mid s) = \tfrac{2}{d}.$$

B.3.2 Overview of proof

The proof of the theorem is based on Assouad's method, where we first reduce the problem to binary hypothesis tests and then connect the testing error to the uncertainty quantity in the upper bound.

Lemma B.3.1 (Reduction to testing)**.** *There exists a universal constant $c_1 > 0$ such that*

$$\inf_{\widehat{\pi}} \max_{u \in \mathcal{U}} \mathbb{E}_u\big[V_u^\star - V_u^{\widehat{\pi}}\big] \geq c_1\, \delta \sqrt{d}\, H \min_{u,u' \in \mathcal{U}:\, D_H(u';u)=1} \inf_{\psi} \big[\mathbb{P}_u(\psi \neq u) + \mathbb{P}_{u'}(\psi \neq u')\big],$$

where $\widehat{\pi}$ denotes the output of any algorithm that maps from observations to an estimated policy. ψ is any test function for parameter u and D_H is the hamming distance.

Lemma B.3.2. *There exists a universal constant $c_2 > 0$ such that when taking $\delta := \frac{c_2\, d}{\sqrt{K}}$, we have*

$$\min_{u,u' \in \mathcal{U}:\, D_H(u';u)=1} \inf_{\psi} \big[\mathbb{P}_u(\psi \neq u) + \mathbb{P}_{u'}(\psi \neq u')\big] \geq \frac{1}{2}. \tag{B.16}$$

When $K \gtrsim d^3$, $\delta := \frac{c_2\, d}{\sqrt{K}}$ ensures that $\delta \leq 1/\sqrt{3d}$. Combining the above two lemmas yields a lower bound

$$\inf_{\widehat{\pi}} \max_{u \in \mathcal{U}} \mathbb{E}_u\big[V_u^\star - V_u^{\widehat{\pi}}\big] \geq c\, \frac{d\sqrt{d}H}{\sqrt{K}}, \tag{B.17}$$

where $c > 0$ is a universal constant. We then use the following B.3.3 to connect the above lower bound to the uncertainty term $\sqrt{d} \cdot \sum_{h=1}^{H} \sqrt{\mathbb{E}_{\pi^\star}[\phi]^\top (\Lambda_h^\star)^{-1} \mathbb{E}_{\pi^\star}[\phi]}$ for the chosen linear MDP

instances class $\mathcal{M}$.

Lemma B.3.3. *There exists a universal constant $c_3 > 0$ such that for all $M \in \mathcal{M}$,*

$$\sum_{h=1}^{H} \sqrt{\mathbb{E}_{\pi^\star}[\phi]^\top (\Lambda_h^\star)^{-1} \mathbb{E}_{\pi^\star}[\phi]} \leq c_3 \frac{d\,H}{\sqrt{K}}. \tag{B.18}$$

Plugging inequality (B.3.3) into the bound (B.17), we obtain the minimax lower bound (3.6) in the statement of theorem.

B.3.3 Reduction to testing via Assouad's method

Proof: [Proof of Lemma B.3.1] For any index vector $u = (u_1, \ldots, u_H) \in \mathcal{U} = \{-1, +1\}^{(d-2)\times H}$, the optimal policy for MDP instance M_u is simply

$$\pi_h^\star(\cdot) = u_h \qquad \text{for } h \in [H].$$

Similar to the proof of Lemma 9 in [41], we can show that the value suboptimality of policy π on MDP M_u is given by

$$V_u^\star - V_u^\pi = \frac{\delta}{\sqrt{2d}} \sum_{h=1}^{H} \left\| u_h - \mathbb{E}_\pi[a_h] \right\|_1.$$

Define $u^\pi = (u_1^\pi, \ldots, u_H^\pi)$ with $u_h^\pi := \operatorname{sign}\big(\mathbb{E}_\pi[a_h]\big)$, then the ℓ_1-norm is lower bounded as

$$\left\| u_h - \mathbb{E}_\pi[a_h] \right\|_1 \geq D_H(u_h^\pi; u_h),$$

where $D_H(\cdot;\cdot)$ denotes the Hamming distance. It follows that

$$V_u^\star - V_u^\pi \geq \frac{\delta}{\sqrt{2d}} D_H(u^\pi; u). \tag{B.19}$$

We then apply Assouad's method (Lemma 2.12 in [158]) and obtain that

$$\inf_{\hat{u}\in\mathcal{U}} \max_{u\in\mathcal{U}} \mathbb{E}_u\big[D_H(\hat{u};u)\big] \geq \frac{(d-2)H}{2} \min_{u,u'\in\mathcal{U}:\, D_H(u';u)=1} \inf_{\psi} \big[\mathbb{P}_u(\psi \neq u) + \mathbb{P}_{u'}(\psi \neq u')\big], \quad \text{(B.20)}$$

where ψ is any test functions mapping from observations to $\{u, u'\}$. Combining inequalities (B.19) and (B.20), we finish the proof.

B.3.4 Lower bound on the testing error

Proof: [Proof of Lemma B.3.2] The proof of Lemma B.3.2 is similar to that of Lemma 10 in [41]. We first apply Theorem 2.12 in [158] to lower bound the testing error using Kullback–Leibler divergence and obtain

$$\min_{u,u'\in\mathcal{U}:\, D_H(u';u)=1} \inf_{\psi} \big[\mathbb{P}_u(\psi \neq u) + \mathbb{P}_{u'}(\psi \neq u')\big] \geq 1 - \left(\frac{1}{2} \max_{u,u'\in\mathcal{U}:\, D_H(u';u)=1} D_{\mathrm{KL}}(\mathcal{Q}_u \| \mathcal{Q}_{u'})\right)^{1/2}. \quad \text{(B.21)}$$

It only remains to estimate $D_{\mathrm{KL}}(\mathcal{Q}_u \| \mathcal{Q}_{u'})$.

The probability density $\mathcal{Q}_u$ takes the form

$$\mathcal{Q}_u(\mathcal{D}) = \prod_{k=1}^{K} \xi_1(s_1^k) \prod_{h=1}^{H} \mu\big(a_h^k \mid s_h^k\big)\, \big[R_{u,h}(s_h^k, a_h^k)\big](r_h^k)\, \mathbb{P}_h(s_{h+1}^k \mid s_h^k, a_h^k)$$

where $\xi_1 = \left[\frac{1}{2}, \frac{1}{2}\right]$ is the initial distribution. It follows that

$$
\begin{aligned}
&D_{KL}(\mathcal{Q}_u \| \mathcal{Q}_{u'}) = \mathbb{E}_u\Big[\log(\mathcal{Q}_u/\mathcal{Q}_{u'})\Big] \\
&= K \cdot \sum_{h=1}^{H} \mathbb{E}_u\Big[\log\Big(\big[R_{u,h}(s_h^1, a_h^1)\big](r_h^1)\big/\big[R_{u',h}(s_h^1, a_h^1)\big](r_h^1)\Big)\Big] \\
&= \frac{K}{d}\sum_{j=1}^{d-2} D_{\mathrm{KL}}\Big(\mathcal{N}\big(\tfrac{\delta}{\sqrt{2d}}\langle e_j, u_h\rangle, 1\big) \,\Big\|\, \mathcal{N}\big(\tfrac{\delta}{\sqrt{2d}}\langle e_j, u'_h\rangle, 1\big)\Big).
\end{aligned}
$$

If we take $\delta = \frac{c_2\, d}{\sqrt{K}}$, then inequality (B.16) is ensured, as claimed in the statement of the lemma.

B.3.5 Connection to the uncertainty term

Proof: [Proof of Lemma B.3.3] We first calculate the explicit form of the inverse of variance-rescaled covariance matrix $\Lambda_h^{\star,p}$. For each time step $h \in [H]$, the value function $V_{u,h+1}^{\star}$ takes the form

$$
V_{u,h+1}^{\star} = \mathbb{E}_{\pi^\star} r_{u,h+1} + \big(\mathbb{P}_{h+1}^{\pi^\star} V_{u,h+2}^{\star}\big).
$$

Since $\big(\mathbb{P}_{h+1} V_{u,h+2}^{\star}\big)(+1) = \big(\mathbb{P}_{h+1} V_{u,h+2}^{\star}\big)(-1)$ and $r_{u,h+1}(+1, a) - r_{u,h+1}(-1, a) = 2/\sqrt{6}$, we have

$$
\mathrm{Var}_{P_h}(V_{u,h+1}^{\star})(+1, a) = \mathrm{Var}_{P_h}(\mathbb{E}_{\pi^\star} r_{u,h+1})(+1, a) = \frac{1}{6}.
$$

Similarly,

$$
\mathrm{Var}_{P_h}(V_{u,h+1}^{\star})(-1, a) = \mathrm{Var}_{P_h}(V_{u,h+1}^{\star})(+1, a) = \frac{1}{6}.
$$

By routine calculation, we find that the population-level rescaled covariance matrix takes the form

$$\Lambda_h^{\star,p} = \frac{3K}{2}\begin{pmatrix} \frac{2}{d^2}\mathbf{I}_{d-2} & \frac{1}{d\sqrt{d}}\mathbf{1}_{(d-2)\times 2} \\ \frac{1}{d\sqrt{d}}\mathbf{1}_{2\times(d-2)} & \mathbf{I}_2 \end{pmatrix} \in \mathbb{R}^{d\times d}$$

for any $h \in [H]$. Applying Gaussian elimination on $\Lambda_h^{\star,p}$, we have

$$(\Lambda_h^{\star,p})^{-1} = \frac{2}{3K}\begin{pmatrix} \frac{d^2}{2}\left\{\mathbf{I}_{d-2} + \frac{1}{d-2}\mathbf{1}_{(d-2)\times(d-2)}\right\} & -\frac{d\sqrt{d}}{2(d-2)}\mathbf{1}_{(d-2)\times 2} \\ -\frac{d\sqrt{d}}{2(d-2)}\mathbf{1}_{2\times(d-2)} & \frac{1}{d-2}\begin{pmatrix} d-1 & 1 \\ 1 & d-1 \end{pmatrix} \end{pmatrix}.$$

For each time step $h \in [H]$, we have (by Jensen's inequality)

$$\sqrt{\mathbb{E}_{\pi^\star}[\phi]^\top(\Lambda_h^\star)^{-1}\mathbb{E}_{\pi^\star}[\phi]} \le \frac{1}{2}\left\|\phi(+1,u_h)\right\|_{(\Lambda_h^{\star,p})^{-1}} + \frac{1}{2}\left\|\phi(-1,u_h)\right\|_{(\Lambda_h^{\star,p})^{-1}}.$$

Recall that by our construction,

$$\phi(+1,u_h) = \begin{pmatrix} \frac{u_h}{\sqrt{2d}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{pmatrix} \in \mathbb{R}^d, \qquad \phi(-1,u_h) = \begin{pmatrix} \frac{u_h}{\sqrt{2d}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{pmatrix} \in \mathbb{R}^d.$$

It follows that

$$\begin{aligned}
&\left\|\phi(+1,u_h)\right\|^2_{(\Lambda_h^{\star,p})^{-1}} = \left\|\phi(-1,u_h)\right\|^2_{(\Lambda_h^{\star,p})^{-1}} \\
&= \frac{2}{3K}\left\{\frac{d}{4}u_h^\top\left\{\mathbf{I}_{d-2}+\tfrac{1}{d-2}\mathbf{1}_{(d-2)\times(d-2)}\right\}u_h - \frac{d}{2(d-2)}\mathbf{1}_{d-2}^\top u_h + \frac{d-1}{2(d-2)}\right\} \\
&= \frac{2}{3K}\left\{\frac{d^2}{4}+\frac{d}{4(d-2)}\left(1-\mathbf{1}_{d-2}^\top u_h\right)^2+\frac{1}{4}\right\} \\
&\le \frac{2}{3K}\left\{\frac{d^2}{4}+\frac{d(d-1)^2}{4(d-2)}+\frac{1}{4}\right\} = \frac{2}{3K}\left\{\frac{d^2}{2}+\frac{d-1}{2(d-2)}\right\} \lesssim d^2/K.
\end{aligned}$$

Therefore,

$$\sqrt{\mathbb{E}_{\pi^\star}[\phi]^\top(\Lambda_h^\star)^{-1}\mathbb{E}_{\pi^\star}[\phi]} \lesssim d/\sqrt{K}.$$

Taking the summation over $h \in [H]$, we obtain the bound (B.18) as claimed in the lemma statement.

B.3.6 Comparison to Lower bound in [80]

Generally speaking, Theorem 3.3.4 and lower bound in [80] are not directly comparable since both results are global minimax (not instance-dependent/local-minimax) lower bounds and their hardness only hold for a family of hard instances (which makes comparison outside of the family instances vacuum). However, for all the instances within the family, we can compare them. Since both papers use tabular hard instance constructions, we only compare $\sqrt{d}\sum_{h=1}^H\sqrt{\mathbb{E}_{\pi^\star}[\phi]^\top(\Lambda_h^\star)^{-1}\mathbb{E}_{\pi^\star}[\phi]}$ and $\sum_{h=1}^H\mathbb{E}_{\pi^\star}[\sqrt{\phi(s_h,a_h)^\top\Lambda_h^{\star-1}\phi(s_h,a_h)}]$ under the tabular setting.

Indeed, under the tabular setting $\phi(s,a) = \mathbf{1}_{s,a}$, $d = SA$, we have

$$\begin{aligned}
&\sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\left[\sqrt{\phi(\cdot,\cdot)^\top \Lambda_h^{\star-1} \phi(\cdot,\cdot)}\right] = \sum_{h=1}^{H}\sum_{s,a} d_h^{\pi^\star}(s,a)\sqrt{\mathbf{1}_{s,a}^\top \Lambda_h^{\star-1}\mathbf{1}_{s,a}} \\
&= \sum_{h=1}^{H}\sum_{s,a} d_h^{\pi^\star}(s,a)\sqrt{\mathbf{1}_{s,a}^\top \mathrm{diag}\left\{\frac{\mathrm{Var}_{P_{\cdot,\cdot}}(V_{h+1}^\star)}{n_{h,\cdot,\cdot}}\right\}\mathbf{1}_{s,a}} \\
&= \sum_{h=1}^{H}\sum_{s,a} d_h^{\pi^\star}(s,a)\sqrt{\frac{\mathrm{Var}_{P_{s,a}}(V_{h+1}^\star)}{n_{h,s,a}}} \quad n_{h,s,a} := \sum_{\tau=1}^{K}\mathbf{1}[s_h^\tau, a_h^\tau = s,a]
\end{aligned}$$

and

$$\begin{aligned}
&\sqrt{d}\sum_{h=1}^{H}\sqrt{\mathbb{E}_{\pi^\star}[\phi]^\top(\Lambda_h^\star)^{-1}\mathbb{E}_{\pi^\star}[\phi]} = \sqrt{SA}\sum_{h=1}^{H}\sqrt{\mathbb{E}_{\pi^\star}[\mathbf{1}_{s,a}]^\top \mathrm{diag}\left\{\frac{\mathrm{Var}_{P_{\cdot,\cdot}}(V_{h+1}^\star)}{n_{h,\cdot,\cdot}}\right\}\mathbb{E}_{\pi^\star}[\mathbf{1}_{s,a}]} \\
=&\sqrt{SA}\sum_{h=1}^{H}\sqrt{\mathrm{Vec}(d_h^{\pi^\star}(\cdot,\cdot))^\top \mathrm{diag}\left\{\frac{\mathrm{Var}_{P_{\cdot,\cdot}}(V_{h+1}^\star)}{n_{h,\cdot,\cdot}}\right\}\mathrm{Vec}(d_h^{\pi^\star}(\cdot,\cdot))} \\
=&\sqrt{SA}\sum_{h=1}^{H}\sqrt{\sum_{s,a} d_h^{\pi^\star}(s,a)^2\frac{\mathrm{Var}_{P_{s,a}}(V_{h+1}^\star)}{n_{h,s,a}}} = \sum_{h=1}^{H}\sqrt{SA}\sqrt{\sum_{s,a} d_h^{\pi^\star}(s,a)^2\frac{\mathrm{Var}_{P_{s,a}}(V_{h+1}^\star)}{n_{h,s,a}}} \\
\geq &\sum_{h=1}^{H}\sum_{s,a} d_h^{\pi^\star}(s,a)\sqrt{\frac{\mathrm{Var}_{P_{s,a}}(V_{h+1}^\star)}{n_{h,s,a}}},
\end{aligned}$$

where the last step uses C-S inequality. This finishes verification.

B.4 Some missing derivations and discussions

B.4.1 Regarding coverage assumption

Now we discuss the feature coverage assumption. Indeed, even if Assumption 4.2.3 is not satisfied, we can still learn in the effective subspan of $\Sigma_h^p := \mathbb{E}_{\mu,h}\left[\phi(s,a)\phi(s,a)^\top\right]$. Concretely, since Σ_h^p is symmetric, by orthogonal decomposition we have $\Sigma_h^p = Z_h \Lambda Z_h^\top$, where Z_h (can be estimated using the samples for practical purpose) consists of orthogonal basis and Λ consists

of eigenvalues of Σ_h^p in the diagonal. Suppose we do not have a full coverage, i.e.

$$\Lambda = \text{diag}[\lambda_1, \lambda_2, ..., \lambda_{d'}, 0, ..., 0] \quad \text{with} \quad d' < d,$$

then we can create transformed features $\phi'_h(s,a) = Z_h \cdot \phi_h(s,a)$, and then

$$\mathbb{E}_{\mu,h}\left[\phi'_h(s,a)\phi'_h(s,a)^\top\right] = \Lambda = \text{diag}[\lambda_1, \lambda_2, ..., \lambda_{d'}, 0, ..., 0].$$

Then we can do learning w.r.t. the truncated features $\phi'_h|_{1:d'}$'s instead of the original ϕ. It reduces to the weaker notion of $\kappa := \min_{h\in[H]}\{\kappa_h : s.t.\ \kappa_h = \text{smallest positive eigenvalue at time } h\}$.

B.4.2 On Variance-Awareness for Linear Mixture MDP

In this section we provide a short discussion of applying VAPVI for the setting where the model is described by a mixture of linear kernels. We first recall the definition of linear mixture MDP models.

Definition B.4.1 (Linear Mixture Models)**.** *We assume the MDP is linear w.r.t. feature map $\psi : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}^d$, i.e., for any $h \in [H]$, there exists $\theta_h \in \mathbb{R}^d$ with $\|\theta_h\|_2 \le B$ for B bounded such that*

$$P_h(s'|s,a) = \psi(s,a,s')^\top\theta_h$$

for all $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$. Also, for any bounded function $V : \mathcal{S} \to [0,1]$,

$$\left\|\int_{\mathcal{S}} \psi(s,a,s')\cdot V(s')ds'\right\|_2 \le \sqrt{d}$$

for any $(s,a) \in \mathcal{S}\times\mathcal{A}$.

In this setting, similar to [56], one can create a value-dependent state-action feature as

$$\phi_h^V(\cdot,\cdot) = \int_S \psi(\cdot,\cdot,s')V_{h+1}(s')ds'$$

Then replacing the all the feature mapping ϕ in Algorithm 2 by ϕ_h^V, VAPVI can be similarly conducted for Linear mixture MDPs. However, there are three differences:

- ϕ_h^V depends on V, which could incur extra randomness when instantiated with VAPVI (since V is plugged by $\widehat{V}_{h+1}$);
- Unlike ϕ in linear MDP, ϕ_h is different for all the time step h (since it is coupled with V_{h+1}).
- The reward might not admit linear in feature structure, which requires modifications of the algorithm (*e.g.* regressing over $P_hV_{h+1}(\cdot,\cdot)$ instead of $P_hV_{h+1}(\cdot,\cdot) + r_h(\cdot,\cdot) = Q_h(\cdot,\cdot)$ in [56]).

We leave how to analyze VAPVI-style algorithm for linear mixture MDPs as the future works.

B.4.3 Derivation of (3.5)

When reducing Theorem 3.3.1,3.3.2 to the tabular case, set $\phi(s,a) = \mathbf{1}_{s,a}$, $d = SA$, $\lambda = 0$, and recall by Assumption 3.3.3 (let's assume $\pi^\star$ is a deterministic policy as it always exists in

tabular MDP) $C^\star := \sup_{h,s,a} d_h^{\pi^\star}(s,a)/d_h^{\mu}(s,a)$, then for Theorem 3.3.1

$$
\begin{aligned}
&\sqrt{d}\cdot\sum_{h=1}^{H}\mathbb{E}_{\pi^\star}\left[\sqrt{\phi(\cdot,\cdot)^\top\Lambda_h^{\star-1}\phi(\cdot,\cdot)}\right] = \sqrt{d}\cdot\sum_{h=1}^{H}\sum_{s,a} d_h^{\pi^\star}(s,a)\sqrt{\mathbf{1}_{s,a}^\top\Lambda_h^{\star-1}\mathbf{1}_{s,a}}\\
&= \sqrt{SA}\cdot\sum_{h=1}^{H}\sum_{s,a} d_h^{\pi^\star}(s,a)\sqrt{\mathbf{1}_{s,a}^\top \mathrm{diag}\left\{\frac{\mathrm{Var}_{P_{\cdot,\cdot}}(V_{h+1}^\star)}{n_{h,\cdot,\cdot}}\right\}\mathbf{1}_{s,a}}\\
&= \sqrt{SA}\cdot\sum_{h=1}^{H}\sum_{s,a} d_h^{\pi^\star}(s,a)\sqrt{\frac{\mathrm{Var}_{P_{s,a}}(V_{h+1}^\star)}{n_{h,s,a}}} \quad n_{h,s,a} := \sum_{\tau=1}^{K}\mathbf{1}[s_h^\tau, a_h^\tau = s,a]\\
&\lesssim \sqrt{SA}\cdot\sum_{h=1}^{H}\sum_{s,a} d_h^{\pi^\star}(s,a)\sqrt{\frac{\mathrm{Var}_{P_{s,a}}(V_{h+1}^\star)}{K\cdot d_h^{\mu}(s,a)}}\\
&\le \sqrt{SAC^\star/K}\cdot\sum_{h=1}^{H}\sum_{s,a}\sqrt{d_h^{\pi^\star}(s,a)\mathrm{Var}_{P_{s,a}}(V_{h+1}^\star)}\\
&= \sqrt{SAC^\star/K}\cdot\sum_{h=1}^{H}\sum_{s}\sqrt{d_h^{\pi^\star}(s,\pi^\star(s))\mathrm{Var}_{P_{s,\pi^\star(s)}}(V_{h+1}^\star)}\\
&\le \sqrt{SAC^\star/K}\cdot\sum_{h=1}^{H}\sqrt{S\cdot\sum_{s} d_h^{\pi^\star}(s,\pi^\star(s))\mathrm{Var}_{P_{s,\pi^\star(s)}}(V_{h+1}^\star)}\\
&\le \sqrt{S^2AC^\star/K}\cdot\sqrt{H\sum_{h=1}^{H}\sum_{s} d_h^{\pi^\star}(s,\pi^\star(s))\mathrm{Var}_{P_{s,\pi^\star(s)}}(V_{h+1}^\star)}\\
&= \sqrt{S^2AC^\star/K}\cdot\sqrt{H\cdot\sum_{h=1}^{H}\mathbb{E}_{\pi_h^\star}[\mathrm{Var}_{P_{(\cdot,\cdot)}}(V_{h+1}^\star)]} \le \sqrt{H^3S^2AC^\star/K}
\end{aligned}
$$

where the first inequality is by Chernoff bound and the last one is by Lemma 3.4. of [2] (Law of total variances). The rest of them are from Cauchy's inequality. Similarly, for Theorem 3.3.2,

we also have

$$\begin{aligned}
&\sqrt{d}\cdot\sum_{h=1}^{H}\sqrt{\mathbb{E}_{\pi^\star}[\phi]^\top\Lambda_h^{\star-1}\mathbb{E}_{\pi^\star}[\phi]} = \sqrt{d}\cdot\sum_{h=1}^{H}\sqrt{\mathrm{Vec}\{d^{\pi^\star}\}\Lambda_h^{\star-1}\mathrm{Vec}\{d^{\pi^\star}\}}\\
&= \sqrt{d}\cdot\sum_{h=1}^{H}\sqrt{\mathrm{Vec}\{d^{\pi^\star}\}\mathrm{diag}\left\{\frac{\mathrm{Var}_{P_{\cdot,\cdot}}(V_{h+1}^\star)}{n_{h,\cdot,\cdot}}\right\}\mathrm{Vec}\{d^{\pi^\star}\}}\\
&= \sqrt{SA}\cdot\sum_{h=1}^{H}\sqrt{\sum_{s,a}d_h^{\pi^\star}(s,a)^2\frac{\mathrm{Var}_{P_{s,a}}(V_{h+1}^\star)}{n_{h,s,a}}}\\
&\lesssim \sqrt{SA}\cdot\sum_{h=1}^{H}\sqrt{\sum_{s,a}d_h^{\pi^\star}(s,a)^2\frac{\mathrm{Var}_{P_{s,a}}(V_{h+1}^\star)}{K\cdot d_h^{\mu}(s,a)}}\\
&\leq \sqrt{SAC^\star/K}\cdot\sum_{h=1}^{H}\sqrt{\sum_{s,a}d_h^{\pi^\star}(s,a)\mathrm{Var}_{P_{s,a}}(V_{h+1}^\star)}\\
&= \sqrt{SAC^\star/K}\cdot\sum_{h=1}^{H}\sqrt{\sum_{s}d_h^{\pi^\star}(s,\pi^\star(s))\mathrm{Var}_{P_{s,\pi^\star(s)}}(V_{h+1}^\star)}\\
&\leq \sqrt{SAC^\star/K}\cdot\sqrt{H\cdot\sum_{h=1}^{H}\mathbb{E}_{\pi_h^\star}[\mathrm{Var}_{P_{(\cdot,\cdot)}}(V_{h+1}^\star)]} \leq \sqrt{H^3SAC^\star/K}.
\end{aligned}$$

B.5 Related Concentration Results and Decompositions

Lemma B.5.1 (Matrix McDiarmid inequality / Matrix Chernoff bound [159])**.** *Let z_k, $k = 1,\dots,K$ be independent random vectors in $\mathbb{R}^d$, and let H be a mapping that maps K vectors to a $d\times d$ symmetric matrix. Assume there exists a sequence of fixed symmetric matrices $\{A_k\}_{k\in[K]}$ such that for z_k, z_k' ranges over all possible values for each $k\in[K]$, it holds*

$$(H(z_1,\dots,z_k,\dots,z_K) - H(z_1,\dots,z_k',\dots,z_K))^2 \preceq A_k^2.$$

Define $\sigma^2 := \left\|\sum_k A_k^2\right\|$. *Then for any* $t > 0$,

$$\mathbb{P}\left\{\left\|H(z_1, \ldots, z_K) - \mathbb{E}H(z_1, \ldots, z_K)\right\| \geq t\right\} \leq d \cdot \exp\left(\frac{-t^2}{8\sigma^2}\right)$$

Lemma B.5.2 (Hoeffding inequality for self-normalized martingales [160])**.** *Let* $\{\eta_t\}_{t=1}^{\infty}$ *be a real-valued stochastic process. Let* $\{\mathcal{F}_t\}_{t=0}^{\infty}$ *be a filtration, such that* η_t *is* $\mathcal{F}_t$*-measurable. Assume* η_t *also satisfies* η_t *given* $\mathcal{F}_{t-1}$ *is zero-mean and* R*-subgaussian,* i.e.

$$\forall \lambda \in \mathbb{R}, \quad \mathbb{E}\left[e^{\lambda \eta_t} \mid \mathcal{F}_{t-1}\right] \leq e^{\lambda^2 R^2/2}$$

Let $\{x_t\}_{t=1}^{\infty}$ *be an* $\mathbb{R}^d$*-valued stochastic process where* x_t *is* $\mathcal{F}_{t-1}$ *measurable and* $\left\|x_t\right\| \leq L$. *Let* $\Lambda_t = \lambda I_d + \sum_{s=1}^{t} x_s x_s^\top$. *Then for any* $\delta > 0$*, with probability* $1 - \delta$*, for all* $t > 0$,

$$\left\|\sum_{s=1}^{t} x_s \eta_s\right\|_{\Lambda_t^{-1}}^2 \leq 8R^2 \cdot \frac{d}{2} \log\left(\frac{\lambda + tL}{\lambda \delta}\right).$$

Lemma B.5.3 (Bernstein inequality for self-normalized martingales [76])**.** *Let* $\{\eta_t\}_{t=1}^{\infty}$ *be a real-valued stochastic process. Let* $\{\mathcal{F}_t\}_{t=0}^{\infty}$ *be a filtration, such that* η_t *is* $\mathcal{F}_t$*-measurable. Assume* η_t *also satisfies*

$$\left|\eta_t\right| \leq R, \mathbb{E}\left[\eta_t \mid \mathcal{F}_{t-1}\right] = 0, \mathbb{E}\left[\eta_t^2 \mid \mathcal{F}_{t-1}\right] \leq \sigma^2.$$

Let $\{x_t\}_{t=1}^{\infty}$ *be an* $\mathbb{R}^d$*-valued stochastic process where* x_t *is* $\mathcal{F}_{t-1}$ *measurable and* $\left\|x_t\right\| \leq L$. *Let* $\Lambda_t = \lambda I_d + \sum_{s=1}^{t} x_s x_s^\top$. *Then for any* $\delta > 0$*, with probability* $1 - \delta$*, for all* $t > 0$,

$$\left\|\sum_{s=1}^{t} \mathbf{x}_s \eta_s\right\|_{\mathbf{\Lambda}_t^{-1}} \leq 8\sigma \sqrt{d \log\left(1 + \frac{tL^2}{\lambda d}\right) \cdot \log\left(\frac{4t^2}{\delta}\right)} + 4R \log\left(\frac{4t^2}{\delta}\right)$$

Lemma B.5.4 (Converting the variance under the matrix norm)**.** *Let* Λ_1 *and* $\Lambda_2 \in \mathbb{R}^{d \times d}$ *are*

two positive semi-definite matrices. Then:

$$\left\|\Lambda_1^{-1}\right\| \leq \left\|\Lambda_2^{-1}\right\| + \left\|\Lambda_1^{-1}\right\| \cdot \left\|\Lambda_2^{-1}\right\| \cdot \|\Lambda_1 - \Lambda_2\|$$

and

$$\|\phi\|_{\Lambda_1^{-1}} \leq \left[1 + \sqrt{\left\|\Lambda_2^{-1}\right\| \|\Lambda_2\| \cdot \left\|\Lambda_1^{-1}\right\| \cdot \|\Lambda_1 - \Lambda_2\|}\right] \cdot \|\phi\|_{\Lambda_2^{-1}}.$$

for all $\phi \in \mathbb{R}^d$.

Proof: For the first part, note

$$\left\|\Lambda_1^{-1}\right\| \leq \left\|\Lambda_2^{-1}\right\| + \left\|\Lambda_1^{-1} - \Lambda_2^{-1}\right\| \leq \left\|\Lambda_2^{-1}\right\| + \left\|\Lambda_2^{-1}\right\| \|\Lambda_1 - \Lambda_2\| \left\|\Lambda_1^{-1}\right\|$$

For the second one,

$$\begin{aligned}
&\|\phi\|_{\Lambda_1^{-1}} = \sqrt{\phi^\top \Lambda_1^{-1} \phi} = \sqrt{\phi^\top \left(\Lambda_1^{-1} - \Lambda_2^{-1}\right) \phi + \phi^\top \Lambda_2^{-1} \phi} \\
=&\sqrt{\phi^\top \Lambda_2^{-1/2} \left(\Lambda_2^{1/2} \Lambda_1^{-1} \Lambda_2^{1/2} - I + I\right) \Lambda_2^{-1/2} \phi} \leq \sqrt{\|\phi\|_{\Lambda_2^{-1}} \cdot \left(1 + \left\|\Lambda_2^{1/2} \Lambda_1^{-1} \Lambda_2^{1/2} - I\right\|\right) \|\phi\|_{\Lambda_2^{-1}}} \\
\leq& \left(1 + \left\|\Lambda_2^{1/2} \Lambda_1^{-1} \Lambda_2^{1/2} - I\right\|^{1/2}\right) \cdot \|\phi\|_{\Lambda_2^{-1}} = \left(1 + \left\|\Lambda_2^{1/2} \Lambda_1^{-1} \left(\Lambda_2 - \Lambda_1\right) \Lambda_2^{-1} \Lambda_2^{1/2}\right\|^{1/2}\right) \cdot \|\phi\|_{\Lambda_2^{-1}} \\
\leq& \left(1 + \sqrt{\|\Lambda_2\| \left\|\Lambda_1^{-1}\right\| \left\|\Lambda_2^{-1}\right\| \|\Lambda_1 - \Lambda_2\|}\right) \cdot \|\phi\|_{\Lambda_2^{-1}}
\end{aligned}$$

Lemma B.5.5 (Lemma H.4 of [82])**.** *let* $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ *satisfies* $\|\phi(s,a)\| \leq C$ *for all* $s, a \in \mathcal{S} \times \mathcal{A}$*. For any* $K > 0, \lambda > 0$*, define* $\bar{G}_K = \sum_{k=1}^K \phi(s_k, a_k)\phi(s_k, a_k)^\top + \lambda I_d$ *where*

(s_k, a_k)'s are i.i.d samples from some distribution ν. Then with probability $1-\delta$,

$$\left\| \frac{\bar{G}_K}{K} - \mathbb{E}_\nu \left[\frac{\bar{G}_K}{K} \right] \right\| \leq \frac{4\sqrt{2}C^2}{\sqrt{K}} \left(\log \frac{2d}{\delta} \right)^{1/2}.$$

Proof: [Proof of Lemma C.11.6] For completeness, we provide the proof of Lemma C.11.6. Let $x_k = \phi(s_k, a_k)$. Denote $\widetilde{\Sigma}_h$ as the matrix obtained by replacing the k-th vector x_k in $\widehat{\Sigma}_h$ by $\widetilde{x}_k$ and leaving the rest $K-1$ vectors unchanged. Then

$$\left(\frac{\widehat{\Sigma}_h}{K} - \frac{\widetilde{\Sigma}_h}{K} \right)^2 = \left(\frac{x_k x_k^\top - \tilde{x}_k \tilde{x}_k^\top}{K} \right) \preceq \frac{1}{K^2} \left(2x_k x_k^\top x_k x_k^\top + 2\tilde{x}_k \tilde{x}_k^\top \tilde{x}_k \tilde{x}_k^\top \right) \preceq \frac{4C^4}{K^2} I_d := A_k^2.$$

Notice that $\left\| \sum_k^K A_k^2 \right\| = \frac{4C^4}{K}$, by Lemma B.5.1 we have the result.

Lemma B.5.6 (Lemma H.5. of [82])**.** *Let $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ be a bounded function s.t. $\|\phi\|_2 \leq C$. Define $\bar{G}_K = \sum_{k=1}^K \phi(s_k, a_k)\phi(s_k, a_k)^\top + \lambda I_d$ where (s_k, a_k)'s are i.i.d samples from some distribution ν. Let $G = \mathbb{E}_\nu[\phi(s, a)\phi(s, a)^\top]$. Then for any $\delta \in (0, 1)$, if K satisfies*

$$K \geq \max \left\{ 512C^4 \left\| \mathbf{G}^{-1} \right\|^2 \log \left(\frac{2d}{\delta} \right), 4\lambda \left\| \mathbf{G}^{-1} \right\| \right\}.$$

Then with probability at least $1-\delta$, it holds simultaneously for all $u \in \mathbb{R}^d$ that

$$\|u\|_{\bar{G}_K^{-1}} \leq \frac{2}{\sqrt{K}} \|u\|_{G^{-1}}.$$

Lemma B.5.7 (Extended Value Difference (Section B.1 in [56]))**.** *Let $\pi = \{\pi_h\}_{h=1}^H$ and $\pi' = \{\pi'_h\}_{h=1}^H$ be two arbitrary policies and let $\{\widehat{Q}_h\}_{h=1}^H$ be any given Q-functions. Then define $\widehat{V}_h(s) := \langle \widehat{Q}_h(s, \cdot), \pi_h(\cdot \mid s) \rangle$ for all $s \in \mathcal{S}$. Then for all $s \in \mathcal{S}$,*

$$\begin{aligned}\widehat{V}_1(s) - V_1^{\pi'}(s) = &\sum_{h=1}^{H} \mathbb{E}_{\pi'}\left[\langle \widehat{Q}_h\left(s_h, \cdot\right), \pi_h\left(\cdot \mid s_h\right) - \pi'_h\left(\cdot \mid s_h\right)\rangle \mid s_1 = s\right] \\ &+ \sum_{h=1}^{H} \mathbb{E}_{\pi'}\left[\widehat{Q}_h\left(s_h, a_h\right) - \left(\mathcal{T}_h \widehat{V}_{h+1}\right)\left(s_h, a_h\right) \mid s_1 = s\right]\end{aligned} \tag{B.22}$$

where $(\mathcal{T}_h V)(\cdot,\cdot) := r_h(\cdot,\cdot) + (P_h V)(\cdot,\cdot)$ *for any* $V \in \mathbb{R}^S$.

Proof:

Denote $\xi_h = \widehat{Q}_h - \mathcal{T}_h \widehat{V}_{h+1}$. For any $h \in [H]$, we have

$$\begin{aligned}\widehat{V}_h - V_h^{\pi'} &= \langle \widehat{Q}_h, \pi_h\rangle - \langle Q_h^{\pi'}, \pi'_h\rangle \\ &= \langle \widehat{Q}_h, \pi_h - \pi'_h\rangle + \langle \widehat{Q}_h - Q_h^{\pi'}, \pi'_h\rangle \\ &= \langle \widehat{Q}_h, \pi_h - \pi'_h\rangle + \langle P_h(\widehat{V}_{h+1} - V_{h+1}^{\pi'}) + \xi_h, \pi'_h\rangle \\ &= \langle \widehat{Q}_h, \pi_h - \pi'_h\rangle + \langle P_h(\widehat{V}_{h+1} - V_{h+1}^{\pi'}), \pi'_h\rangle + \langle \xi_h, \pi'_h\rangle\end{aligned}$$

recursively apply the above for $\widehat{V}_{h+1} - V_{h+1}^{\pi'}$ and use the $\mathbb{E}_{\pi'}$ notation (instead of the inner product of P_h, π'_h) we can finish the prove of this lemma.

Lemma B.5.8. *Let* $\widehat{\pi} = \left\{\widehat{\pi}_h\right\}_{h=1}^{H}$ *and* $\widehat{Q}_h(\cdot,\cdot)$ *be the arbitrary policy and Q-function and also* $\widehat{V}_h(s) = \langle \widehat{Q}_h(s,\cdot), \widehat{\pi}_h(\cdot|s)\rangle \ \forall s \in \mathcal{S}$. *and* $\zeta_h(s,a) := (\mathcal{T}_h \widehat{V}_{h+1})(s,a) - \widehat{Q}_h(s,a)$ *(element-wisely) to be the Bellman update error. Then for any arbitrary* π, *we have*

$$\begin{aligned}V_1^{\pi}(s) - V_1^{\widehat{\pi}}(s) = &\sum_{h=1}^{H} \mathbb{E}_{\pi}\left[\zeta_h(s_h, a_h) \mid s_1 = s\right] - \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}\left[\zeta_h(s_h, a_h) \mid s_1 = s\right] \\ &+ \sum_{h=1}^{H} \mathbb{E}_{\pi}\left[\langle \widehat{Q}_h\left(s_h, \cdot\right), \pi_h\left(\cdot|s_h\right) - \widehat{\pi}_h\left(\cdot|s_h\right)\rangle \mid s_1 = x\right]\end{aligned}$$

where the expectation are taken over s_h, a_h.

Proof: Note the gap can be rewritten as

$$V_1^{\pi}(s) - V_1^{\widehat{\pi}}(s) = V_1^{\pi}(s) - \widehat{V}_1(s) + \widehat{V}_1(s) - V_1^{\widehat{\pi}}(s).$$

By Lemma D.0.7 with $\pi = \widehat{\pi}$, $\pi' = \pi$, we directly have

$$V_1^{\pi}(s) - \widehat{V}_1(s) = \sum_{h=1}^{H} \mathbb{E}_{\pi}\left[\zeta_h(s_h, a_h) \mid s_1 = s\right] + \sum_{h=1}^{H} \mathbb{E}_{\pi}\left[\langle \widehat{Q}_h\left(s_h, \cdot\right), \pi_h\left(\cdot|s_h\right) - \widehat{\pi}_h\left(\cdot|s_h\right)\rangle \mid s_1 = s\right] \tag{B.23}$$

Next apply Lemma D.0.7 again with $\pi = \pi' = \widehat{\pi}$, we directly have

$$\widehat{V}_1(s) - V_1^{\widehat{\pi}}(s) = -\sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}\left[\zeta_h(s_h, a_h) \mid s_1 = s\right]. \tag{B.24}$$

Combine the above two results we prove the stated result.

Lemma B.5.9. *For a linear MDP, for any $0 \leq V(\cdot) \leq H$, then there exists a $w_h \in \mathbb{R}^d$ s.t. $\mathcal{T}_h V = \langle \phi, w_h \rangle$ and $\|w_h\|_2 \leq 2H\sqrt{d}$ for all $h \in [H]$. Here $\mathcal{T}_h(V)(s,a) = r_h(x,a) + (P_h V)(s,a)$. Similarly, for any π, there exists $w_h^{\pi} \in \mathbb{R}^d$, such that $Q_h^{\pi} = \langle \phi, w_h^{\pi} \rangle$ with $\|w_h^{\pi}\|_2 \leq 2(H - h + 1)\sqrt{d}$.*

Proof: By definition,

$$\mathcal{T}_h V = r_h + (P_h V) = \langle \phi, \theta_h \rangle + \langle \phi, \int_{\mathcal{S}} V(s) d\nu_h(s) \rangle$$
$$\Rightarrow w_h = \theta_h + \int_{\mathcal{S}} V(s) d\nu_h(s),$$

therefore $\|w_h\|_2 \leq \|\theta_h\|_2 + H \cdot \|\nu_h(\mathcal{S})\| \leq 1 + H\sqrt{d} \leq 2H\sqrt{d}$. The proof of the second part is similar by backward induction and the fact $V_h^{\pi} \leq H - h + 1$ for any π.

Lemma B.5.10. *For any pessimistic bonus design Γ_h, suppose $K > \max\{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4\}$,*

then with probability $1-\delta$*, Algorithm 2 yields*

$$\left\|\mathcal{T}_h\widehat{V}_{h+1}-\widehat{\mathcal{T}}_h\widehat{V}_{h+1}\right\|_\infty \le \widetilde{O}(\frac{H^2\sqrt{d/\kappa}}{\sqrt{K}})$$

Proof: [Proof of Lemma B.5.10] Suppose w_h is the coefficient corresponding to the $\mathcal{T}_h\widehat{V}_{h+1}$ (such w_h exists by Lemma B.5.9), *i.e.* $\mathcal{T}_h\widehat{V}_{h+1}=\phi^\top w_h$, and recall $(\widehat{\mathcal{T}}_h\widehat{V}_{h+1})(s,a)=\phi(s,a)^\top\widehat{w}_h$, then:

$$\begin{aligned}
&\left(\mathcal{T}_h\widehat{V}_{h+1}\right)(s,a)-\left(\widehat{\mathcal{T}}_h\widehat{V}_{h+1}\right)(s,a)=\phi(s,a)^\top\left(w_h-\widehat{w}_h\right)\\
=&\phi(s,a)^\top w_h-\phi(s,a)^\top\widehat{\Lambda}_h^{-1}\left(\sum_{\tau=1}^K\phi\left(s_h^\tau,a_h^\tau\right)\cdot\left(r_h^\tau+\widehat{V}_{h+1}\left(s_{h+1}^\tau\right)\right)/\widehat{\sigma}_h^2(s_h^\tau,a_h^\tau)\right)\\
=&\underbrace{\phi(s,a)^\top w_h-\phi(s,a)^\top\widehat{\Lambda}_h^{-1}\left(\sum_{\tau=1}^K\phi\left(s_h^\tau,a_h^\tau\right)\cdot\left(\mathcal{T}_h\widehat{V}_{h+1}\right)\left(s_h^\tau,a_h^\tau\right)/\widehat{\sigma}_h^2(s_h^\tau,a_h^\tau)\right)}_{\text{(i)}}\\
&\qquad\underbrace{+\phi(s,a)^\top\widehat{\Lambda}_h^{-1}\left(\sum_{\tau=1}^K\phi\left(s_h^\tau,a_h^\tau\right)\cdot\left(r_h^\tau+\widehat{V}_{h+1}\left(s_{h+1}^\tau\right)-\left(\mathcal{T}_h\widehat{V}_{h+1}\right)\left(s_h^\tau,a_h^\tau\right)\right)/\widehat{\sigma}_h^2(s_h^\tau,a_h^\tau)\right)}_{\text{(ii)}}.
\end{aligned}\tag{B.25}$$

For term (i), it is bounded by $\frac{2\lambda H^3\sqrt{d}/\kappa}{K}$ with probability $1-\delta$ by Lemma B.1.2.

For term (ii), by Cauchy inequality it is bounded by

$$\begin{aligned}
&\|\phi(s,a)\|_{\widehat{\Lambda}_h^{-1}}\cdot\left\|\sum_{\tau=1}^K\phi\left(s_h^\tau,a_h^\tau\right)\cdot\left(r_h^\tau+\widehat{V}_{h+1}\left(s_{h+1}^\tau\right)-\left(\mathcal{T}_h\widehat{V}_{h+1}\right)\left(s_h^\tau,a_h^\tau\right)\right)/\widehat{\sigma}_h^2(s_h^\tau,a_h^\tau)\right\|_{\widehat{\Lambda}_h^{-1}}\\
\le&\frac{2H}{\sqrt{\kappa K}}\left\|\sum_{\tau=1}^K\phi\left(s_h^\tau,a_h^\tau\right)\cdot\left(r_h^\tau+\widehat{V}_{h+1}\left(s_{h+1}^\tau\right)-\left(\mathcal{T}_h\widehat{V}_{h+1}\right)\left(s_h^\tau,a_h^\tau\right)\right)/\widehat{\sigma}_h^2(s_h^\tau,a_h^\tau)\right\|_{\widehat{\Lambda}_h^{-1}}\\
\le&\frac{2H}{\sqrt{\kappa K}}\cdot\tilde{O}(H\sqrt{d})=\widetilde{O}(\frac{H^2\sqrt{d/\kappa}}{\sqrt{K}}),
\end{aligned}$$

where the first inequality is by Lemma C.11.5 (with $\phi'=\phi/\widehat{\sigma}_h$ and $\left\|\phi/\widehat{\sigma}_h\right\|\le\|\phi\|\le 1:=C$)

and the third inequality uses $\sqrt{a^\top \cdot A \cdot a} \le \sqrt{\|a\|_2 \|A\|_2 \|a\|_2} = \|a\|_2 \sqrt{\|A\|_2}$ with a to be either ϕ or w_h. Moreover, $\lambda_{\min}(\tilde{\Lambda}_h^p) \ge \kappa / \max_{h,s,a} \widehat{\sigma}_h(s,a)^2 \ge \kappa / H^2$ implies $\left\|(\tilde{\Lambda}_h^p)^{-1}\right\| \le H^2/\kappa$. The second inequality comes from Lemma C.11.3 with $R = H$ since $|\eta_\tau| = |(r_h^\tau + \widehat{V}_{h+1}\left(s_{h+1}^\tau\right) - (\mathcal{T}_h \widehat{V}_{h+1})(s_h^\tau, a_h^\tau))/\widehat{\sigma}_h(s_h^\tau, a_h^\tau)| \le H$ and $|x_\tau| = |\phi(s_h^\tau, a_h^\tau)/\widehat{\sigma}_h(s_h^\tau, a_h^\tau)| \le 1$.

The final result is obtained by absorbing the term (i) via the condition $K > \max\{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4\}$.

Lemma B.5.11. *Suppose random variables $\|X\|_\infty \le 2H$, $\|Y\|_\infty \le 2H$, then*

$$|\mathrm{Var}(X) - \mathrm{Var}(Y)| \le 8H \cdot \|X - Y\|_\infty .$$

Proof: [Proof of Lemma B.5.11]

$$\begin{aligned}
|\mathrm{Var}(X) - \mathrm{Var}(Y)| =& |\mathbb{E}[X^2] - \mathbb{E}[Y^2] - (\mathbb{E}[X]^2 - \mathbb{E}[Y]^2)| = |\mathbb{E}[(X+Y)(X-Y)] - (\mathbb{E}[X+Y])(\mathbb{E}[X-Y])| \\
\le& \mathbb{E}[|X+Y| \cdot |X-Y|] + 4H \cdot \|X - Y\|_\infty \\
\le& 4H\mathbb{E}[|X-Y|] + 4H \cdot \|X-Y\|_\infty = 8H \cdot \|X-Y\|_\infty .
\end{aligned}$$

Appendix C

Supplementary Material in Chapter 4

C.1 Further Illustration that Generalized Linear Model Example satisfies 4.2.3

Recall the definition in 4.2.5, then:

For ($\star\star$),

$$\mathbb{E}_{\mu,h}\left[\nabla f(\theta,\phi(s,a))\cdot\nabla f(\theta,\phi(s,a))^{\top}\right]=\mathbb{E}_{\mu,h}\left[f'(\langle\theta,\phi(s,a)\rangle)^2\phi(\cdot,\cdot)\cdot\phi(\cdot,\cdot)^{\top}\right]$$
$$\succ\kappa^2\mathbb{E}_{\mu,h}\left[\phi(\cdot,\cdot)\cdot\phi(\cdot,\cdot)^{\top}\right]\succ\kappa^3 I,\quad\forall\theta\in\Theta$$

For ($\star$), by Taylor's Theorem,

$$\mathbb{E}_{\mu,h}\left[\left(f(\theta_1,\phi(\cdot,\cdot))-f(\theta_2,\phi(\cdot,\cdot))\right)^2\right]=\mathbb{E}_{\mu,h}[f'(\theta_{s,a},\phi(\cdot,\cdot))^2(\theta_1-\theta_2)^{\top}\phi(\cdot,\cdot)\phi(\cdot,\cdot)^{\top}(\theta_1-\theta_2)]$$
$$\geq\kappa^2\mathbb{E}_{\mu,h}[(\theta_1-\theta_2)^{\top}\phi(\cdot,\cdot)\phi(\cdot,\cdot)^{\top}(\theta_1-\theta_2)]=\kappa^2(\theta_1-\theta_2)^{\top}\mathbb{E}_{\mu,h}[\phi(\cdot,\cdot)\phi(\cdot,\cdot)^{\top}](\theta_1-\theta_2)\geq\kappa^3\left\|\theta_1-\theta_2\right\|_2^2$$

and choose κ^3 as κ in 4.2.3.

C.2 On the computational complexity

For storage of Pessimistic Fitted Q-learning, at each time step $h \in [H]$ in Algorithm 3, we need to store $\widehat{\theta}_h$, Σ_h and $\nabla f(\widehat{\theta}_h, \phi_{h,k})$. Therefore, the total space complexity is $O(dH + d^2H + dKH)$. For computation, assuming $\widehat{\theta}_h$ is solved via SGD and let M denote the number of gradient steps, then the complexity is dominated by computing $\widehat{\theta}_h$, Σ_h and Σ_h^{-1}, which results in $O(MH + KdH + d^3H)$ complexity (where H comes from $h = H, \ldots, 1$).

The space complexity and computational complexity for VAFQL has the same order as PFQL except that the constant factors are larger.

C.3 Some basic constructions

First of all, Recall in the first-order condition, we have

$$\nabla_\theta \left\{ \sum_{k=1}^{K} \left[f\left(\theta, \phi_{h,k}\right) - r_{h,k} - \widehat{V}_{h+1}\left(s_{h+1}^k\right) \right]^2 + \lambda \cdot \|\theta\|_2^2 \right\} \Bigg|_{\theta=\widehat{\theta}_h} = 0, \quad \forall h \in [H].$$

Therefore, if we define the quantity $Z_h(\cdot,\cdot) \in \mathbb{R}^d$ as

$$Z_h(\theta|V) = \sum_{k=1}^{K} \left[f\left(\theta, \phi_{h,k}\right) - r_{h,k} - V\left(s_{h+1}^k\right) \right] \nabla f(\theta, \phi_{h,k}) + \lambda \cdot \theta, \quad \forall \theta \in \Theta, \|V\|_2 \le H,$$

then we have (recall $\widehat{\theta}_h \in \mathrm{Int}(\Theta)$)

$$Z_h(\widehat{\theta}_h|\widehat{V}_{h+1}) = 0.$$

In addition, according to Bellman completeness Assumption 4.2.1, for any bounded $V(\cdot) \in \mathbb{R}^S$ with $\|V\|_\infty \le H$, $\inf_{f\in\mathcal{F}} \left\|f - \mathcal{P}_h(V)\right\|_\infty \le \epsilon_{\mathcal{F}}$, $\forall h$ (recall $\mathcal{P}_h(V) = r_h + \int_S V\, dP_h$). Therefore, we can define the *parameter Bellman operator* $\mathbb{T}$ as follows.

Definition C.3.1 (parameter Bellman operator)**.** *By the Bellman completeness Assumption 4.2.1,*

for any $\|V\|_\infty \le H$*, we can define the* parameter Bellman operator $\mathbb{T} : V \to \theta_{\mathbb{T}V} \in \Theta$ *such that*

$$\theta_{\mathbb{T}V} = \operatorname*{argmin}_{\theta\in\Theta} \left\| f(\theta,\phi) - \mathcal{P}_h(V) \right\|_\infty$$

Denote $\delta_V := f(\theta_{\mathbb{T}V},\phi) - \mathcal{P}_h(V)$*, then we have* $\left\|f(\theta_{\mathbb{T}V},\phi) - \mathcal{P}_h(V)\right\|_\infty = \left\|\delta_V\right\|_\infty \le \epsilon_{\mathcal{F}}$*. In particular, by realizability Assumption 4.2.1 it holds* $\theta_{\mathbb{T}V^\star_{h+1}} = \theta^\star_h$ *and this is due to* $f(\theta_{\mathbb{T}V^\star_{h+1}},\phi) = \mathcal{P}_h(V^\star_{h+1}) = V^\star_h = f(\theta^\star_h,\phi)$.[1]

C.3.1 Suboptimality decomposition

Denote $\iota_h(s,a) := \mathcal{P}_h\widehat{V}_{h+1}(s,a) - \widehat{Q}_h(s,a)$, by [80] we have the following decomposition.

Lemma C.3.1 (Lemma 3.1 of [80])**.** *Let* $\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^H$ *a policy and* $\widehat{Q}_h$ *be any estimates with* $\widehat{V}_h = \langle \widehat{Q}_h(s,\cdot), \widehat{\pi}_h(\cdot \mid s)\rangle_{\mathcal{A}}$*. Then for any policy* π*, we have*

$$v^\pi - v^{\widehat{\pi}} = -\sum_{h=1}^H E_{\widehat{\pi}}[\iota_h(s_h,a_h)] + \sum_{h=1}^H E_\pi[\iota_h(s_h,a_h)] + \sum_{h=1}^H E_\pi[\langle \widehat{Q}_h\left(s_h,\cdot\right), \pi_h\left(\cdot \mid s_h\right) - \widehat{\pi}_h\left(\cdot \mid s_h\right)\rangle_{\mathcal{A}}].$$

In particular, if we choose $\widehat{\pi}_h(\cdot|s) := \operatorname{argmax}_\pi \langle \widehat{Q}_h(s,\cdot), \pi(\cdot \mid s)\rangle_{\mathcal{A}}$*, then*

$$v^\pi - v^{\widehat{\pi}} = -\sum_{h=1}^H E_{\widehat{\pi}}[\iota_h(s_h,a_h)] + \sum_{h=1}^H E_\pi[\iota_h(s_h,a_h)].$$

Lemma C.3.2. *Let* $\widehat{\mathcal{P}}_h$ *be the general estimated Bellman operator. Suppose with probability* $1-\delta$*, it holds for all* $h,s,a \in [H]\times\mathcal{S}\times A$ *that* $|(\mathcal{P}_h\widehat{V}_{h+1} - \widehat{\mathcal{P}}_h\widehat{V}_{h+1})(s,a)| \le \Gamma_h(s,a)$*, then it implies* $\forall s,a,h \in \mathcal{S}\times\mathcal{A}\times[H]$*,* $0 \le \zeta_h(s,a) \le 2\Gamma_h(s,a)$*. Furthermore, it holds for any policy*

[1] Here without loss of generality we assume $Q^\star_h$ can be uniquely identified, *i.e.* there is a unique $\theta^\star$ such that $f(\theta^\star_h,\phi) = Q^\star_h$.

π simultaneously, with probability $1-\delta$*,*

$$V_1^{\pi}(s) - V_1^{\widehat{\pi}}(s) \leq \sum_{h=1}^{H} 2 \cdot \mathbb{E}_{\pi}\left[\Gamma_h(s_h, a_h) \mid s_1 = s\right].$$

Proof: [Proof of Lemma C.3.2] This is a generic result that holds true for the general MDPs and was first raised by Theorem 4.2 of [80]. Later, it is summarized in Lemma C.1 of [106].

With Lemma C.3.2, we need to bound the term $|\mathcal{P}_h\widehat{V}_{h+1}(s,a) - \widehat{\mathcal{P}}_h\widehat{V}_{h+1}(s,a)|$.

C.4 Analyzing $|\mathcal{P}_h\widehat{V}_{h+1}(s,a) - \widehat{\mathcal{P}}_h\widehat{V}_{h+1}(s,a)|$ for PFQL.

Throughout this section, we suppose $\epsilon_{\mathcal{F}} = 0$, *i.e.* $f(\theta_{\mathbb{T}V}, \phi) = \mathcal{P}_h(V)$. According to the regression oracle (Line 4 of Algorithm 3), the estimated Bellman operator $\widehat{\mathcal{P}}_h$ maps $\widehat{V}_{h+1}$ to $\widehat{\theta}_h$, *i.e.* $\widehat{\mathcal{P}}_h\widehat{V}_{h+1} = f(\widehat{\theta}_h, \phi)$. Therefore (recall Definition C.3.1)

$$\begin{aligned}
&\mathcal{P}_h\widehat{V}_{h+1}(s,a) - \widehat{\mathcal{P}}_h\widehat{V}_{h+1}(s,a) = \mathcal{P}_h\widehat{V}_{h+1}(s,a) - f(\widehat{\theta}_h, \phi(s,a)) \\
=&f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi(s,a)) - f(\widehat{\theta}_h, \phi(s,a)) \\
=&\nabla f(\widehat{\theta}_h, \phi(s,a))\left(\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h\right) + \mathrm{Hot}_{h,1},
\end{aligned} \tag{C.1}$$

where we apply the first-order Taylor expansion for the differentiable function f at point $\widehat{\theta}_h$ and $\mathrm{Hot}_{h,1}$ is a higher-order term. Indeed, the following Lemma C.4.1 bounds the $\mathrm{Hot}_{h,1}$ term with $\widetilde{O}(\frac{1}{K})$.

Lemma C.4.1. *Recall the definition (from the above decomposition)* $\mathrm{Hot}_{h,1} := f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi(s,a)) - f(\widehat{\theta}_h, \phi(s,a)) - \nabla f(\widehat{\theta}_h, \phi(s,a))\left(\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h\right)$*, then with probability* $1-\delta$*,*

$$\left|\mathrm{Hot}_{h,1}\right| \leq \frac{18H^2\kappa_2(\log(H/\delta) + C_{d,\log K}) + \kappa_2\lambda C_{\Theta}^2}{\kappa K}, \quad \forall h \in [H].$$

Proof: [Proof of Lemma C.4.1] By second-order Taylor's Theorem, there exists a point ξ (lies in the line segment of $\widehat{\theta}_h$ and $\theta_{\mathbb{T}\widehat{V}_{h+1}}$) such that

$$f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi(s,a)) - f(\widehat{\theta}_h, \phi(s,a)) = \nabla f(\widehat{\theta}_h, \phi(s,a))^\top \left(\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h\right) + \frac{1}{2}\left(\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h\right)^\top \nabla^2_{\theta\theta} f(\xi, \phi(s,a)) \left(\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h\right)$$

Therefore, by directly applying Theorem C.6.1, with probability $1-\delta$, for all $h \in [H]$,

$$\begin{aligned}
\left|\mathrm{Hot}_{h,1}\right| =& \frac{1}{2}\left|\left(\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h\right)^\top \nabla^2_{\theta\theta} f(\xi, \phi(s,a)) \left(\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h\right)\right| \\
\leq& \frac{1}{2}\kappa_2 \cdot \left\|\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h\right\|_2^2 \leq \frac{18H^2\kappa_2(\log(H/\delta) + C_{d,\log K}) + \kappa_2\lambda C_\Theta^2}{\kappa K}
\end{aligned}$$

C.4.1 Analyzing $\nabla f(\widehat{\theta}_h, \phi(s,a)) \left(\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h\right)$ via Z_h.

From (C.1) and Lemma C.4.1, the problem further reduces to bounding $\nabla f(\widehat{\theta}_h, \phi(s,a)) \left(\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h\right)$. To begin with, we first provide a characterization of $\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h$. Indeed, by first-order Vector Taylor expansion (Lemma C.11.1), we have (note $Z_h(\widehat{\theta}_h|\widehat{V}_{h+1}) = 0$) for any $\theta \in \Theta$,

$$Z_h(\theta|\widehat{V}_{h+1}) - Z_h(\widehat{\theta}_h|\widehat{V}_{h+1}) = \Sigma_h^s(\theta - \widehat{\theta}_h) + R_K(\theta), \tag{C.2}$$

where $R_K(\theta)$ is the higher-order residuals and $\Sigma_h^s := \frac{\partial}{\partial\theta} Z_h(\theta|\widehat{\theta}_{h+1})\Big|_{\theta=\widehat{\theta}_h}$ with

$$\begin{aligned}\Sigma_h^s :=& \frac{\partial}{\partial\theta} Z_h(\theta|\widehat{V}_{h+1})\Big|_{\theta=\widehat{\theta}_h} = \frac{\partial}{\partial\theta}\left(\sum_{k=1}^{K}\left[f\left(\theta,\phi_{h,k}\right) - r_{h,k} - \widehat{V}_{h+1}(s_{h+1}^k)\right]\nabla f(\theta,\phi_{h,k}) + \lambda\cdot\theta\right)_{\theta=\widehat{\theta}_h}\\ &= \underbrace{\sum_{k=1}^{K}\left\{\left(f(\widehat{\theta}_h,\phi_{h,k}) - r_{h,k} - \widehat{V}_{h+1}(s_{h+1}^k)\right)\cdot\nabla_{\theta\theta}^2 f(\widehat{\theta}_h,\phi_{h,k})\right\}}_{:=\Delta_{\Sigma_h^s}}\\ &+ \underbrace{\sum_{k=1}^{K}\nabla_\theta f(\widehat{\theta}_h,\phi_{h,k})\nabla_\theta^\top f(\widehat{\theta}_{h,k},\phi_{h,k}) + \lambda I_d}_{:=\Sigma_h},\end{aligned} \tag{C.3}$$

here $\nabla^2 = \nabla\bigotimes\nabla$ denotes outer product of gradients.

Note $\Delta_{\Sigma_h^s}$ is not desirable since it could prevent Σ_h^s from being positive-definite (and it could cause Σ_h^s to be singular). Therefore, we first deal with $\Delta_{\Sigma_h^s}$ in below.

Lemma C.4.2. *With probability* $1-\delta$*, for all* $h\in[H]$*,*

$$\begin{aligned}&\frac{1}{K}\left\|\Delta_{\Sigma_h^s}\right\|_2 = \left\|\frac{1}{K}\sum_{k=1}^{K}\left(f(\widehat{\theta}_h,\phi_{h,k}) - r_{h,k} - \widehat{V}_{h+1}(s_{h+1}^k)\right)\cdot\nabla_{\theta\theta}^2 f(\widehat{\theta}_h,\phi_h)\right\|_2\\ \le& 9\kappa_2\max(\frac{\kappa_1}{\sqrt{\kappa}},1)\sqrt{\frac{dH^2(\log(2H/\delta) + d\log(1+2C_\Theta H\kappa_3 K) + C_{d,\log K})}{K}} + \frac{1}{K}.\end{aligned}$$

Proof: [Proof of Lemma C.4.2]

Step1: We prove for fixed $\bar{\theta}\in\Theta$, with probability $1-\delta$, for all $h\in[H]$,

$$\left\|\frac{1}{K}\sum_{k=1}^{K}\left(f(\widehat{\theta}_h,\phi_{h,k}) - r_{h,k} - \widehat{V}_{h+1}(s_{h+1}^k)\right)\cdot\nabla_{\theta\theta}^2 f(\bar{\theta},\phi_h)\right\|_2 \le 9\kappa_2\max(\frac{\kappa_1}{\sqrt{\kappa}},1)\sqrt{\frac{H^2(\log(2H/\delta) + C_{d,\log K})}{K}}.$$

Indeed, we have

$$\begin{aligned}
&\left\| \frac{1}{K}\sum_{k=1}^{K}\left(f(\widehat{\theta}_h,\phi_{h,k}) - r_{h,k} - \widehat{V}_{h+1}(s_{h+1}^k)\right)\cdot\nabla^2_{\theta\theta}f(\bar{\theta},\phi_h)\right\|_2 \\
\leq &\left\| \frac{1}{K}\sum_{k=1}^{K}\left(f(\widehat{\theta}_h,\phi_{h,k}) - f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi_{h,k})\right)\cdot\nabla^2_{\theta\theta}f(\bar{\theta},\phi_h)\right\|_2 \\
+ &\left\| \frac{1}{K}\sum_{k=1}^{K}\left(f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi_{h,k}) - r_{h,k} - \widehat{V}_{h+1}(s_{h+1}^k)\right)\cdot\nabla^2_{\theta\theta}f(\bar{\theta},\phi_h)\right\|_2 .
\end{aligned} \tag{C.4}$$

On one hand, by Theorem C.6.1 with probability $1-\delta/2$ for all $h\in[H]$

$$\begin{aligned}
&\left\| \frac{1}{K}\sum_{k=1}^{K}\left(f(\widehat{\theta}_h,\phi_{h,k}) - f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi_{h,k})\right)\cdot\nabla^2_{\theta\theta}f(\bar{\theta},\phi_h)\right\|_2 \leq \kappa_2\cdot\max_{\theta,s,a}\|\nabla f(\theta,\phi(s,a))\|_2\left\|\widehat{\theta}_h - \theta_{\mathbb{T}\widehat{V}_{h+1}}\right\|_2 \\
\leq &\kappa_2\kappa_1\left\|\widehat{\theta}_h - \theta_{\mathbb{T}\widehat{V}_{h+1}}\right\|_2 \leq \kappa_2\kappa_1\left(\sqrt{\frac{36H^2(\log(H/\delta)+C_{d,\log K})+2\lambda C_\Theta^2}{\kappa K}} + \sqrt{\frac{b_{d,K,\epsilon_{\mathcal{F}}}}{\kappa}} + \sqrt{\frac{2H\epsilon_{\mathcal{F}}}{\kappa}}\right).
\end{aligned} \tag{C.5}$$

On other hand, recall the definition of $\mathbb{T}$, we have

$$\begin{aligned}
&\mathbb{E}\left[(f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi_{h,k}) - r_{h,k} - \widehat{V}_{h+1}(s_{h+1}^k))\cdot\nabla^2_{\theta\theta}f(\bar{\theta},\phi_{h,k})\middle| s_h^k,a_h^k\right] \\
=&\mathbb{E}\left[f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi_{h,k}) - r_{h,k} - \widehat{V}_{h+1}(s_{h+1}^k)\middle| s_h^k,a_h^k\right]\cdot\nabla^2_{\theta\theta}f(\bar{\theta},\phi_{h,k}) \\
=&\left((\mathcal{P}_h\widehat{V}_{h+1})(s_h^k,a_h^k) - \mathbb{E}\left[r_{h,k}+\widehat{V}_{h+1}(s_{h+1}^k)\middle| s_h^k,a_h^k\right]\right)\cdot\nabla^2_{\theta\theta}f(\bar{\theta},\phi_{h,k}) \\
=&\left((\mathcal{P}_h\widehat{V}_{h+1})(s_h^k,a_h^k) - (\mathcal{P}_h\widehat{V}_{h+1}(s_{h+1}^k))(s_h^k,a_h^k)\right)\cdot\nabla^2_{\theta\theta}f(\bar{\theta},\phi_{h,k}) = 0.
\end{aligned}$$

Also, since $\left\|\left(f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi_{h,k}) - r_{h,k} - \widehat{V}_{h+1}(s_{h+1}^k))\right)\cdot\nabla^2_{\theta\theta}f(\bar{\theta},\phi_h)\right\|_2 \leq H\kappa_2$, denote $\sigma^2 := K\cdot H^2\kappa_2^2$, then by Vector Hoeffding's inequality (Lemma C.11.2),

$$\mathbb{P}\left(\left\| \frac{1}{K}\sum_{k=1}^{K}\left(f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi_{h,k}) - r_{h,k} - \widehat{V}_{h+1}(s_{h+1}^k)\right)\cdot\nabla^2_{\theta\theta}f(\bar{\theta},\phi_h)\right\|_2 \geq t/K\middle|\{s_h^k,a_h^k\}_{k=1}^K\right) \leq d\cdot e^{-t^2/8dKH^2\kappa_2^2} := \delta$$

which is equivalent to

$$\mathbb{P}\left(\left\|\frac{1}{K}\sum_{k=1}^{K}\left(f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi_{h,k})-r_{h,k}-\widehat{V}_{h+1}(s_{h+1}^k)\right)\cdot\nabla^2_{\theta\theta}f(\bar{\theta},\phi_h)\right\|_2\leq\sqrt{\frac{8dH^2\kappa_2^2\log(d/\delta)}{K}}\middle|\{s_h^k,a_h^k\}_{k=1}^K\right)\geq 1-\delta$$

Define $A=\{\left\|\frac{1}{K}\sum_{k=1}^{K}\left(f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi_{h,k})-r_{h,k}-\widehat{V}_{h+1}(s_{h+1}^k)\right)\cdot\nabla^2_{\theta\theta}f(\bar{\theta},\phi_h)\right\|_2\leq\sqrt{\frac{8dH^2\kappa_2^2\log(d/\delta)}{K}}\}$, then by law of total expectation $\mathbb{P}(A)=\mathbb{E}[\mathbf{1}_A]=\mathbb{E}[\mathbb{E}[\mathbf{1}_A|\{s_h^k,a_h^k\}_{k=1}^K]]=\mathbb{E}[\mathbb{P}[A|\{s_h^k,a_h^k\}_{k=1}^K]]\geq\mathbb{E}[1-\delta]=1-\delta$, i.e. with probability at least $1-\delta/2$ (and a union bound),

$$\left\|\frac{1}{K}\sum_{k=1}^{K}\left(f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi_{h,k})-r_{h,k}-\widehat{V}_{h+1}(s_{h+1}^k)\right)\cdot\nabla^2_{\theta\theta}f(\bar{\theta},\phi_h)\right\|_2\leq\sqrt{\frac{8dH^2\kappa_2^2\log(2Hd/\delta)}{K}},\ \forall h\in[H].$$

Using above and (C.4), (C.5) and a union bound, w.p. $1-\delta$, for all $h\in[H]$,

$$\begin{aligned}\left\|\frac{1}{K}\sum_{k=1}^{K}\left(f(\widehat{\theta}_h,\phi_{h,k})-r_{h,k}-\widehat{V}_{h+1}(s_{h+1}^k)\right)\cdot\nabla^2_{\theta\theta}f(\bar{\theta},\phi_h)\right\|_2&\leq 6\kappa_2\kappa_1\sqrt{\frac{H^2(\log(2H/\delta)+C_{d,\log K})}{\kappa K}}\\&+\sqrt{\frac{8dH^2\kappa_2^2\log(2Hd/\delta)}{K}}\leq 9\kappa_2\max(\frac{\kappa_1}{\sqrt{\kappa}},1)\sqrt{\frac{dH^2(\log(2H/\delta)+C_{d,\log K})}{K}}\end{aligned}$$

Step2: we finish the proof of the lemma.

Consider $\left\{f(\bar{\theta}):=\left\|\frac{1}{K}\sum_{k=1}^{K}\left(f(\widehat{\theta}_h,\phi_{h,k})-r_{h,k}-\widehat{V}_{h+1}(s_{h+1}^k)\right)\cdot\nabla^2_{\theta\theta}f(\bar{\theta},\phi_h)\right\|_2\middle|\bar{\theta}\in\Theta\right\}$, then by triangular inequality

$$\begin{aligned}|f(\bar{\theta}_1)-f(\bar{\theta}_2)|\leq&\left\|\frac{1}{K}\sum_{k=1}^{K}\left(f(\widehat{\theta}_h,\phi_{h,k})-r_{h,k}-\widehat{V}_{h+1}(s_{h+1}^k)\right)\cdot\left[\nabla^2_{\theta\theta}f(\bar{\theta}_1,\phi_h)-\nabla^2_{\theta\theta}f(\bar{\theta}_2,\phi_h)\right]\right\|_2\\\leq&H\cdot\sup_{s,a}\left\|\nabla^2_{\theta\theta}f(\bar{\theta}_1,\phi_h)-\nabla^2_{\theta\theta}f(\bar{\theta}_2,\phi_h)\right\|_2\leq H\kappa_3\left\|\bar{\theta}_1-\bar{\theta}_2\right\|_2.\end{aligned}$$

By Lemma C.11.8, the covering number $\mathcal{C}$ of the ϵ-net of the above function class satisfies $\log\mathcal{C}\leq d\log(1+\frac{2C_\Theta H\kappa_3}{\epsilon})$. By choosing $\epsilon=1/K$, by a union bound over $\mathcal{C}$ cases we obtain for

all $h \in [H]$

$$\begin{aligned}&\left\|\frac{1}{K}\sum_{k=1}^{K}\left(f(\widehat{\theta}_h,\phi_{h,k})-r_{h,k}-\widehat{V}_{h+1}(s_{h+1}^k)\right)\cdot\nabla^2_{\theta\theta}f(\widehat{\theta}_h,\phi_h)\right\|_2\\ \leq& 9\kappa_2\max(\frac{\kappa_1}{\sqrt{\kappa}},1)\sqrt{\frac{dH^2(\log(2H/\delta)+d\log(1+2C_\Theta H\kappa_3 K)+C_{d,\log K})}{K}}+\frac{1}{K}.\end{aligned}$$

Combing Lemma C.4.2 and Theorem C.6.1 (and a union bound), we directly have

Corollary C.4.1. *With probability* $1-\delta$*,*

$$\left\|\frac{1}{K}\Delta_{\Sigma_h^s}(\widehat{\theta}_h-\theta_{\mathbb{T}\widehat{V}_{h+1}})\right\|_2\leq\left\|\frac{1}{K}\Delta_{\Sigma_h^s}\right\|_2\left\|\widehat{\theta}_h-\theta_{\mathbb{T}\widehat{V}_{h+1}}\right\|_2\leq\widetilde{O}(\frac{\kappa_2\max(\frac{\kappa_1}{\kappa},\frac{1}{\sqrt{\kappa}})d^2H^2}{K})$$

Here $\widetilde{O}$ *absorbs all the constants and Polylog terms.*

Now we select $\theta=\theta_{\mathbb{T}\widehat{V}_{h+1}}$ in (C.2), and denote $\widetilde{R}_K(\theta_{\mathbb{T}\widehat{V}_{h+1}})=\Delta_{\Sigma_h^s}(\widehat{\theta}_h-\theta_{\mathbb{T}\widehat{V}_{h+1}})+R_K(\theta_{\mathbb{T}\widehat{V}_{h+1}})$, then (C.2) is equivalent to

$$Z_h(\theta_{\mathbb{T}\widehat{V}_{h+1}}|\widehat{V}_{h+1})-Z_h(\widehat{\theta}_h|\widehat{V}_{h+1})=\Sigma_h^s(\theta_{\mathbb{T}\widehat{V}_{h+1}}-\widehat{\theta}_h)+R_K(\theta_{\mathbb{T}\widehat{V}_{h+1}})=\Sigma_h(\theta_{\mathbb{T}\widehat{V}_{h+1}}-\widehat{\theta}_h)+\widetilde{R}_K(\theta_{\mathbb{T}\widehat{V}_{h+1}})$$

Note $\lambda>0$ implies Σ_h is invertible, then we have (recall $Z_h(\widehat{\theta}_h|\widehat{\theta}_{h+1})=0$)

$$\begin{aligned}\theta_{\mathbb{T}\widehat{V}_{h+1}}-\widehat{\theta}_h=&\Sigma_h^{-1}[Z_h(\theta_{\mathbb{T}\widehat{V}_{h+1}}|\widehat{V}_{h+1})-Z_h(\widehat{\theta}_h|\widehat{V}_{h+1})]-\Sigma_h^{-1}\widetilde{R}_K(\theta_{\mathbb{T}\widehat{V}_{h+1}})\\ =&\Sigma_h^{-1}[Z_h(\theta_{\mathbb{T}\widehat{V}_{h+1}}|\widehat{V}_{h+1})]-\Sigma_h^{-1}\widetilde{R}_K(\theta_{\mathbb{T}\widehat{V}_{h+1}})\end{aligned}$$

Plug it back to (C.1) to get

$$
\begin{aligned}
&\nabla f(\widehat{\theta}_h, \phi(s,a)) \left(\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h\right) \\
=&\nabla f(\widehat{\theta}_h, \phi(s,a))\Sigma_h^{-1}[Z_h(\theta_{\mathbb{T}\widehat{V}_{h+1}}|\widehat{V}_{h+1})] - \nabla f(\widehat{\theta}_h, \phi(s,a))\Sigma_h^{-1}\widetilde{R}_K(\theta_{\mathbb{T}\widehat{V}_{h+1}}) \\
=&\nabla f(\widehat{\theta}_h, \phi(s,a))\Sigma_h^{-1}[\sum_{k=1}^K \left(f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi_{h,k}) - r_{h,k} - \widehat{V}_{h+1}(s_{h+1}^k)\right)\cdot \nabla_\theta^\top f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi_{h,k}) + \lambda\theta_{\mathbb{T}\widehat{V}_{h+1}}] \\
&-\nabla f(\widehat{\theta}_h, \phi(s,a))\Sigma_h^{-1}\widetilde{R}_K(\theta_{\mathbb{T}\widehat{V}_{h+1}}) \\
=&\underbrace{\nabla f(\widehat{\theta}_h, \phi(s,a))\Sigma_h^{-1}[\sum_{k=1}^K \left(f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi_{h,k}) - r_{h,k} - \widehat{V}_{h+1}(s_{h+1}^k)\right)\cdot \nabla_\theta^\top f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi_{h,k})]}_{:=I} \\
&\underbrace{-\nabla f(\widehat{\theta}_h, \phi(s,a))\Sigma_h^{-1}\left[\widetilde{R}_K(\theta_{\mathbb{T}\widehat{V}_{h+1}}) + \lambda\theta_{\mathbb{T}\widehat{V}_{h+1}}\right]}_{:=\mathrm{Hot}_2}
\end{aligned}
\tag{C.6}
$$

We will bound second term Hot_2 to have higher order $O(\frac{1}{K})$ in Section C.4.5 and focus on the first term. By direct decomposition,

$$
\begin{aligned}
I :=&\nabla f(\widehat{\theta}_h, \phi(s,a))\Sigma_h^{-1}[\sum_{k=1}^K \left(f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi_{h,k}) - r_{h,k} - \widehat{V}_{h+1}(s_{h+1}^k)\right)\cdot \nabla_\theta^\top f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi_{h,k})] \\
=&\underbrace{\nabla f(\widehat{\theta}_h, \phi(s,a))\Sigma_h^{-1}[\sum_{k=1}^K \left(f(\theta_{\mathbb{T}V^\star_{h+1}}, \phi_{h,k}) - r_{h,k} - V^\star_{h+1}(s_{h+1}^k)\right)\cdot \nabla_\theta^\top f(\widehat{\theta}_h, \phi_{h,k})]}_{:=I_1} \\
&+\underbrace{\nabla f(\widehat{\theta}_h, \phi(s,a))\Sigma_h^{-1}[\sum_{k=1}^K \left(f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi_{h,k}) - f(\theta_{\mathbb{T}V^\star_{h+1}}, \phi_{h,k}) - \widehat{V}_{h+1}(s_{h+1}^k) + V^\star_{h+1}(s_{h+1}^k)\right)\cdot \nabla_\theta^\top f(\widehat{\theta}_h, \phi_{h,k})]}_{:=I_2} \\
&+\underbrace{\nabla f(\widehat{\theta}_h, \phi(s,a))\Sigma_h^{-1}\left[\sum_{k=1}^K \left(f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi_{h,k}) - r_{h,k} - \widehat{V}_{h+1}(s_{h+1}^k)\right)\cdot \left(\nabla_\theta^\top f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi_{h,k}) - \nabla_\theta^\top f(\widehat{\theta}_h, \phi_{h,k})\right)\right]}_{:=I_3}
\end{aligned}
$$

C.4.2 Bounding the term I_3

We first bound the term I_3. We have the following Lemma.

Lemma C.4.3. *For any fixed $V(\cdot) \in \mathbb{R}^S$ with $\|V\|_\infty \leq H$ and any fixed θ such that $\left\|\theta_{\mathbb{T}V} - \theta\right\|_2 \leq \sqrt{\frac{36H^2(\log(H/\delta)+C_{d,\log K})+2\lambda C_\Theta^2}{\kappa K}}$. Let*

$$\widetilde{I}_3 := \nabla f(\widehat{\theta}_h, \phi(s,a))^\top \Sigma_h^{-1} \left[\sum_{k=1}^K \left(f(\theta_{\mathbb{T}V}, \phi_{h,k}) - r_{h,k} - V(s_{h+1}^k)\right) \cdot \left(\nabla_\theta f(\theta_{\mathbb{T}V}, \phi_{h,k}) - \nabla_\theta f(\theta, \phi_{h,k})\right)\right],$$

and if $K \geq \max\left\{512\frac{\kappa_1^4}{\kappa^2}\left(\log(\frac{2d}{\delta}) + d\log(1+\frac{4\kappa_1 D^2\kappa_2 C_\Theta K^3}{\lambda^2})\right), \frac{4\lambda}{\kappa}\right\}$, then with probability $1-\delta$, (where $D = \max\{\kappa_1, \sqrt{\frac{(144dH^2\kappa_2^2\left(H^2\log(H/\delta)+C_{d,\log K}\right)+8dH^2\kappa_2^2\lambda C_\Theta^2)\log(d/\delta)}{\kappa}}\}$)

$$|\widetilde{I}_3| \leq 4\kappa_1\sqrt{\frac{(144dH^2\kappa_2^2\left(H^2\log(H/\delta)+C_{d,\log K}\right)+8dH^2\kappa_2^2\lambda C_\Theta^2)\log(d/\delta)}{\kappa^3}}\frac{1}{K} + O(\frac{1}{K^{3/2}}).$$

Proof: [Proof of Lemma C.4.3] Indeed, with probability $1-\delta/2$,

$$\begin{aligned}
|\widetilde{I}_3| &= \left\|\nabla f(\widehat{\theta}_h, \phi(s,a))^\top \Sigma_h^{-1}\left[\sum_{k=1}^K \left(f(\theta_{\mathbb{T}V}, \phi_{h,k}) - r_{h,k} - V(s_{h+1}^k)\right)\cdot\left(\nabla_\theta f(\theta_{\mathbb{T}V}, \phi_{h,k}) - \nabla_\theta f(\theta, \phi_{h,k})\right)\right]\right\| \\
&\leq \left\|\nabla f(\widehat{\theta}_h, \phi(s,a))\right\|_{\Sigma_h^{-1}} \left\|\sum_{k=1}^K \left(f(\theta_{\mathbb{T}V}, \phi_{h,k}) - r_{h,k} - V(s_{h+1}^k)\right)\cdot\left(\nabla_\theta f(\theta_{\mathbb{T}V}, \phi_{h,k}) - \nabla_\theta f(\theta, \phi_{h,k})\right)\right\|_{\Sigma_h^{-1}} \\
&\leq \left(\frac{2\kappa_1}{\sqrt{\kappa K}} + O(\frac{1}{K})\right)\left\|\sum_{k=1}^K \left(f(\theta_{\mathbb{T}V}, \phi_{h,k}) - r_{h,k} - V(s_{h+1}^k)\right)\cdot\left(\nabla_\theta f(\theta_{\mathbb{T}V}, \phi_{h,k}) - \nabla_\theta f(\theta, \phi_{h,k})\right)\right\|_{\Sigma_h^{-1}}
\end{aligned}$$

where, under the condition $K \geq \max\left\{512\frac{\kappa_1^4}{\kappa^2}\left(\log(\frac{2d}{\delta}) + d\log(1+\frac{4\kappa_1^3\kappa_2 C_\Theta K^3}{\lambda^2})\right), \frac{4\lambda}{\kappa}\right\}$, we applied Lemma C.11.5 .

Next, on one hand, $\left\|\nabla_\theta f(\theta_{\mathbb{T}V}, \phi_{h,k}) - \nabla_\theta f(\theta, \phi_{h,k})\right\|_2 \leq \kappa_2 \cdot \left\|\theta_{\mathbb{T}V} - \theta\right\|_2 \leq \kappa_2\sqrt{\frac{36H^2(\log(H/\delta)+C_{d,\log K})+2\lambda C_\Theta^2}{\kappa K}}$.

On the other hand,

$$\begin{aligned}
&\mathbb{E}\left[\left(f(\theta_{\mathbb{T}V},\phi_{h,k})-r_{h,k}-V(s_{h+1}^k)\right)\cdot\left(\nabla_\theta^\top f(\theta_{\mathbb{T}V},\phi_{h,k})-\nabla_\theta^\top f(\theta,\phi_{h,k})\right)\middle| s_h^k,a_h^k\right]\\
=&\mathbb{E}\left[\left(f(\theta_{\mathbb{T}V},\phi_{h,k})-r_{h,k}-V(s_{h+1}^k)\right)\middle| s_h^k,a_h^k\right]\cdot\left(\nabla_\theta^\top f(\theta_{\mathbb{T}V},\phi_{h,k})-\nabla_\theta^\top f(\theta,\phi_{h,k})\right)\\
=&\left((\mathcal{P}_hV)(s_h^k,a_h^k)-(\mathcal{P}_hV)(s_h^k,a_h^k)\right)\cdot\left(\nabla_\theta^\top f(\theta_{\mathbb{T}V},\phi_{h,k})-\nabla_\theta^\top f(\theta,\phi_{h,k})\right)=0
\end{aligned}$$

Therefore by Vector Hoeffding's inequality (Lemma C.11.2) (also note the condition for boundedness $\left\|\left(f(\theta_{\mathbb{T}V},\phi_{h,k})-r_{h,k}-V(s_{h+1}^k)\right)\cdot\left(\nabla_\theta f(\theta_{\mathbb{T}V},\phi_{h,k})-\nabla_\theta f(\theta,\phi_{h,k})\right)\right\|_2\le H\kappa_2\cdot\left\|\theta_{\mathbb{T}V}-\theta\right\|_2\le H\kappa_2\sqrt{\frac{36H^2(\log(H/\delta)+C_{d,\log K})+2\lambda C_\Theta^2}{\kappa K}}$) with probability $1-\delta/2$,

$$\begin{aligned}
&\left\|\frac{1}{K}\sum_{k=1}^K\left(f(\theta_{\mathbb{T}V},\phi_{h,k})-r_{h,k}-V(s_{h+1}^k)\right)\cdot\left(\nabla_\theta f(\theta_{\mathbb{T}V},\phi_{h,k})-\nabla_\theta f(\theta,\phi_{h,k})\right)\right\|_2\\
\le&\sqrt{\frac{4d\left(H\kappa_2\sqrt{\frac{36H^2(\log(H/\delta)+C_{d,\log K})+2\lambda C_\Theta^2}{\kappa K}}\right)^2\log(\frac{d}{\delta})}{K}}\\
=&\sqrt{\frac{(144dH^2\kappa_2^2\left(H^2\log(H/\delta)+C_{d,\log K}\right)+8dH^2\kappa_2^2\lambda C_\Theta^2)\log(d/\delta)}{\kappa}}\cdot\frac{1}{K}
\end{aligned}$$

and this implies with probability $1-\delta/2$,

$$\begin{aligned}
&\left\|\sum_{k=1}^K\left(f(\theta_{\mathbb{T}V},\phi_{h,k})-r_{h,k}-V(s_{h+1}^k)\right)\cdot\left(\nabla_\theta f(\theta_{\mathbb{T}V},\phi_{h,k})-\nabla_\theta f(\theta,\phi_{h,k})\right)\right\|_2\\
\le&\sqrt{\frac{(144dH^2\kappa_2^2\left(H^2\log(H/\delta)+C_{d,\log K}\right)+8dH^2\kappa_2^2\lambda C_\Theta^2)\log(d/\delta)}{\kappa}}
\end{aligned}$$

choose $u=\sum_{k=1}^K\left(f(\theta_{\mathbb{T}V},\phi_{h,k})-r_{h,k}-V(s_{h+1}^k)\right)\cdot\left(\nabla_\theta f(\theta_{\mathbb{T}V},\phi_{h,k})-\nabla_\theta f(\theta,\phi_{h,k})\right)$ in Lemma C.11.5,

by a union bound we obtain with probability $1-\delta$

$$
\begin{aligned}
|\widetilde{I}_3| &= \left\| \nabla f(\widehat{\theta}_h, \phi(s,a))^\top \Sigma_h^{-1} \left[\sum_{k=1}^{K} \left(f(\theta_{\mathbb{T}V}, \phi_{h,k}) - r_{h,k} - V(s_{h+1}^k) \right) \cdot \left(\nabla_\theta f(\theta_{\mathbb{T}V}, \phi_{h,k}) - \nabla_\theta f(\theta, \phi_{h,k}) \right) \right] \right\| \\
&\leq \left(\frac{2\kappa_1}{\sqrt{\kappa K}} + O(\frac{1}{K}) \right) \left\| \sum_{k=1}^{K} \left(f(\theta_{\mathbb{T}V}, \phi_{h,k}) - r_{h,k} - V(s_{h+1}^k) \right) \cdot \left(\nabla_\theta f(\theta_{\mathbb{T}V}, \phi_{h,k}) - \nabla_\theta f(\theta, \phi_{h,k}) \right) \right\|_{\Sigma_h^{-1}} \\
&\leq \left(\frac{2\kappa_1}{\sqrt{\kappa K}} + O(\frac{1}{K}) \right) \left(2\sqrt{\frac{(144 d H^2 \kappa_2^2 \left(H^2 \log(H/\delta) + C_{d,\log K} \right) + 8 d H^2 \kappa_2^2 \lambda C_\Theta^2) \log(d/\delta)}{\kappa^2 K}} + O(\frac{1}{K}) \right) \\
&= 4\kappa_1 \sqrt{\frac{(144 d H^2 \kappa_2^2 \left(H^2 \log(H/\delta) + C_{d,\log K} \right) + 8 d H^2 \kappa_2^2 \lambda C_\Theta^2) \log(d/\delta)}{\kappa^3}} \frac{1}{K} + O(\frac{1}{K^{3/2}}).
\end{aligned}
$$

Lemma C.4.4. *Under the same condition as Lemma C.4.3. With probability* $1-\delta$*,*

$$
|I_3| \leq 4\kappa_1 \sqrt{\frac{(144 d H^2 \kappa_2^2 \left(H^2 \log(H/\delta) + D_{d,\log K} + C_{d,\log K} \right) + 8 d H^2 \kappa_2^2 \lambda C_\Theta^2)(\log(d/\delta) + D_{d,\log K})}{\kappa^3}} \frac{1}{K} + O(\frac{1}{K^{3/2}}).
$$

Here $D_{d,\log K} := d \cdot \log(1+6C_\Theta(2\kappa_1^2 + H\kappa_2)K) + d \log(1+6C_\Theta H \kappa_2 K) + d \log\left(1 + 288 C_\Theta \kappa_1^2 (\kappa_1 \sqrt{C_\Theta} + 2\sqrt{B\kappa_1\kappa_2})^2 K^2\right) + d^2 \log\left(1 + 288\sqrt{d} B \kappa_1^4 K^2\right) = \widetilde{O}(d^2)$ *with* $\widetilde{O}$ *absorbs Polylog terms.*

Proof: [Proof of Lemma C.4.4] Define

$$
h(V, \widetilde{\theta}, \theta) = \sum_{k=1}^{K} \left(f(\widetilde{\theta}, \phi_{h,k}) - r_{h,k} - V(s_{h+1}^k) \right) \cdot \left(\nabla_\theta f(\widetilde{\theta}, \phi_{h,k}) - \nabla_\theta f(\theta, \phi_{h,k}) \right),
$$

then

$$
\begin{aligned}
&|h(V_1,\widetilde{\theta}_1,\theta_1)-h(V_2,\widetilde{\theta}_2,\theta_2)|\\
\leq&\left|\sum_{k=1}^{K}\left([f(\widetilde{\theta}_1,\phi_{h,k})-V_1(s_{h+1}^k)]-[f(\widetilde{\theta}_2,\phi_{h,k})-V_2(s_{h+1}^k)]\right)\cdot\left(\nabla_\theta f(\widetilde{\theta}_1,\phi_{h,k})-\nabla_\theta f(\theta_1,\phi_{h,k})\right)\right|\\
+&\left|\sum_{k=1}^{K}\left(f(\widetilde{\theta}_2,\phi_{h,k})-r_{h,k}-V_2(s_{h+1}^k)\right)\cdot\left([\nabla_\theta f(\widetilde{\theta}_1,\phi_{h,k})-\nabla_\theta f(\theta_1,\phi_{h,k})]-[\nabla_\theta f(\widetilde{\theta}_2,\phi_{h,k})-\nabla_\theta f(\theta_2,\phi_{h,k})]\right)\right|\\
\leq&K\sup_{s,a,s'}\left|[f(\widetilde{\theta}_1,\phi(s,a))-f(\widetilde{\theta}_2,\phi(s,a))]-[V_1(s')-V_2(s')]\right|_2\cdot 2\kappa_1\\
+&KH\cdot\sup_{s,a}\left\|[\nabla_\theta f(\widetilde{\theta}_1,\phi(s,a))-\nabla_\theta f(\theta_1,\phi(s,a))]-[\nabla_\theta f(\widetilde{\theta}_2,\phi(s,a))-\nabla_\theta f(\theta_2,\phi(s,a))]\right\|_2\\
\leq&K2\kappa_1^2\left\|\widetilde{\theta}_1-\widetilde{\theta}_2\right\|_2+2K\kappa_1\left\|V_1-V_2\right\|_\infty+HK\kappa_2\left\|\widetilde{\theta}_1-\widetilde{\theta}_2\right\|_2+HK\kappa_2\left\|\theta_1-\theta_2\right\|_2\\
=&(2\kappa_1^2+H\kappa_2)K\left\|\widetilde{\theta}_1-\widetilde{\theta}_2\right\|_2+2\kappa_1K\left\|V_1-V_2\right\|_\infty+HK\kappa_2\left\|\theta_1-\theta_2\right\|_2.
\end{aligned}
$$

Let $\mathcal{C}_a$ be the $\frac{\epsilon/3}{(2\kappa_1^2+H\kappa_2)K}$-covering net of $\{\theta:\|\theta\|_2\leq C_\Theta\}$, $\mathcal{C}_V$ be the $\frac{\epsilon}{6\kappa_1 K}$-covering net of $\mathcal{V}$ defined in Lemma C.11.9 and $\mathcal{C}_b$ be the $\frac{\epsilon}{3H\kappa_2 K}$-covering net of $\{\theta:\|\theta\|_2\leq C_\Theta\}$, then by Lemma C.11.8 and Lemma C.11.9,

$$
\begin{aligned}
\log|\mathcal{C}_a|\leq d\cdot\log(1+\frac{6C_\Theta(2\kappa_1^2+H\kappa_2)K}{\epsilon}),\ \log|\mathcal{C}_b|\leq d\log(1+\frac{6C_\Theta H\kappa_2 K}{\epsilon})\\
\log\mathcal{C}_V\leq d\log\left(1+\frac{288C_\Theta\kappa_1^2(\kappa_1\sqrt{C_\Theta}+2\sqrt{B\kappa_1\kappa_2})^2K^2}{\epsilon^2}\right)+d^2\log\left(1+\frac{288\sqrt{d}B\kappa_1^4K^2}{\epsilon^2}\right).
\end{aligned}
$$

Further notice with probability $1-\delta/2$ (by Lemma C.11.5), for all fixed sets of parameters θ, V satisfies $\left\|\theta_{\mathbb{T}V}-\theta\right\|_2\leq\sqrt{\frac{36H^2(\log(2H/\delta)+C_{d,\log K})+2\lambda C_\Theta^2}{\kappa K}}$ simultaneously,

$$
\begin{aligned}
|I_3-\widetilde{I}_3|&\leq\left\|\nabla f(\widehat{\theta}_h,\phi(s,a))\right\|_{\Sigma_h^{-1}}\cdot\left\|h(\widehat{V}_{h+1},\theta_{\mathbb{T}\widehat{V}_{h+1}},\widehat{\theta}_h)-h(V,\theta_{\mathbb{T}V},\theta)\right\|_{\Sigma_h^{-1}}\\
&\leq\left(\frac{2\kappa_1}{\sqrt{\kappa K}}+O(\frac{1}{K})\right)\cdot\left\|h(\widehat{V}_{h+1},\theta_{\mathbb{T}\widehat{V}_{h+1}},\widehat{\theta}_h)-h(V,\theta_{\mathbb{T}V},\theta)\right\|_{\Sigma_h^{-1}}
\end{aligned}
$$

and $\left\|\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h\right\|_2 \leq \sqrt{\frac{36H^2(\log(2H/\delta)+C_{d,\log K})+2\lambda C_\Theta^2}{\kappa K}}$ with probability $1-\delta/2$ by Theorem C.6.1. Now, choosing $\epsilon = O(1/K^2)$ and by Lemma C.4.3 and union bound over covering instances, we obtain with probability $1-\delta$

$$|I_3| \leq 4\kappa_1\sqrt{\frac{(144dH^2\kappa_2^2\left(H^2\log(H/\delta)+D_{d,\log K}+C_{d,\log K}\right)+8dH^2\kappa_2^2\lambda C_\Theta^2)(\log(d/\delta)+D_{d,\log K})}{\kappa^3}}\frac{1}{K}+O(\frac{1}{K^{3/2}}).$$

C.4.3 Bounding the second term I_2

In this section, we bound the term

$$I_2 := \nabla f(\widehat{\theta}_h,\phi(s,a))\Sigma_h^{-1}[\sum_{k=1}^{K}\left(f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi_{h,k}) - f(\theta_{\mathbb{T}V^\star_{h+1}},\phi_{h,k}) - \widehat{V}_{h+1}(s^k_{h+1}) + V^\star_{h+1}(s^k_{h+1})\right)\cdot\nabla^\top_\theta f(\widehat{\theta}_h,\phi_{h,k})].$$

The following Lemma shows that I_2 is a higher-order error term with rate $\widetilde{O}(\frac{1}{K})$.

Lemma C.4.5 (Bounding I_2)**.** *If K satisfies $K \geq 512\frac{\kappa_1^4}{\kappa^2}\left(\log(\frac{2d}{\delta}) + d\log(1+\frac{4\kappa_1^3\kappa_2 C_\Theta K}{\lambda^2})\right)$, and $K \geq 4\lambda/\kappa$, then with probability $1-\delta$*

$$|I_2| \leq \widetilde{O}(\frac{\kappa_1^2H^2d^2}{\kappa K}) + \widetilde{O}(\frac{1}{K^{3/2}}).$$

Here $\widetilde{O}$ absorbs constants and Polylog terms.

Proof: [Proof of Lemma C.4.5] **Step1.** Define $\eta_k(V) := f(\theta_{\mathbb{T}V},\phi_{h,k}) - f(\theta_{\mathbb{T}V^\star_{h+1}},\phi_{h,k}) - V(s^k_{h+1}) + V^\star_{h+1}(s^k_{h+1})$ and let $\|V(\cdot)\|_\infty \leq H$ be any fixed function such that $\sup_{s^k_h,a^k_h,s^k_{h+1}}|\eta_k(V)| \leq \widetilde{O}(\kappa_1 H^2\sqrt{\frac{d^2}{\kappa K}})$, *i.e.* arbitrary fixed V function in the neighborhood (measured by η_k) of $V^\star_{h+1}$. Then by definition of $\mathbb{T}$ it holds $\mathbb{E}[\eta_k(V,\theta)|s^k_h,a^k_h] = 0$. Let the fixed $\theta\in\Theta$ be arbitrary and define $x_k(\theta) = \nabla_\theta f(\theta,\phi_{h,k})$. Next, define $G_h(\theta) = \sum_{k=1}^K \nabla f(\theta,\phi(s^k_h,a^k_h))\cdot\nabla f(\theta,\phi(s^k_h,a^k_h))^\top + \lambda I_d$, since $\|x_k\|_2 \leq \kappa_1$ and $|\eta_k| \leq \widetilde{O}(\kappa_1 H^2\sqrt{\frac{d^2}{\kappa K}})$, by self-normalized Hoeffding's inequality

(Lemma C.11.3), with probability $1-\delta$ (recall $t := K$ in Lemma C.11.3),

$$\left\|\sum_{k=1}^{K} x_k(\theta)\eta_k(V)\right\|_{G_h(\theta)^{-1}} \leq \widetilde{O}(\kappa_1 H^2\sqrt{\frac{d^2}{\kappa K}})\sqrt{d\log\left(\frac{\lambda+K\kappa_1}{\lambda\delta}\right)}.$$

Step2. Define $h(V,\theta) := \sum_{k=1}^{K} x_k(\theta)\eta_k(V)$ and $H(V,\theta) := \left\|\sum_{k=1}^{K} x_k(\theta)\eta_k(V)\right\|_{G_h(\theta)^{-1}}$, then note by definition $|\eta_k(V)| \leq 2H$, which implies $\|h(V,\theta)\|_2 \leq 2KH\kappa_1$ and

$$|\eta_k(V_1)-\eta_k(V_2)| \leq |\mathcal{P}_h V_1 - \mathcal{P}_h V_2| + \left\|V_1-V_2\right\|_\infty \leq 2\left\|V_1-V_2\right\|_\infty$$

and

$$\begin{aligned}\left\|h(V_1,\theta_1)-h(V_2,\theta_2)\right\|_2 \leq& K\max_k\left(2H\left\|x_k(\theta_1)-x_k(\theta_2)\right\|_2 + \kappa_1|\eta_k(V_1)-\eta_k(V_2)|\right)\\ \leq& K(2H\kappa_2\left\|\theta_1-\theta_2\right\|_2 + 2\kappa_1\left\|V_1-V_2\right\|_\infty).\end{aligned}$$

Furthermore,

$$\begin{aligned}&\left\|G_h(\theta_1)^{-1}-G_h(\theta_2)^{-1}\right\|_2 \leq \left\|G_h(\theta_1)^{-1}\right\|_2\left\|G_h(\theta_1)-G_h(\theta_2)\right\|_2\left\|G_h(\theta_2)^{-1}\right\|_2\\ \leq& \frac{1}{\lambda^2}K\sup_k\left\|\nabla f(\theta_1,\phi_{h,k})\cdot\nabla f(\theta_1,\phi_{h,k})^\top - \nabla f(\theta_2,\phi_{h,k})\cdot\nabla f(\theta_2,\phi_{h,k})^\top\right\|_2\\ \leq& \frac{1}{\lambda^2}K\sup_k\left[\left\|(\nabla f(\theta_1,\phi_{h,k})-\nabla f(\theta_2,\phi_{h,k}))\cdot\nabla f(\theta_1,\phi_{h,k})^\top\right\|_2 + \left\|\nabla f(\theta_2,\phi_{h,k})\cdot(\nabla f(\theta_1,\phi_{h,k})^\top - \nabla f(\theta_2,\phi_{h,k})^\top)\right\|_2\right]\\ \leq& \frac{2\kappa_1 K}{\lambda^2}\kappa_2\left\|\theta_1-\theta_2\right\|_2 = \frac{2\kappa_1\kappa_2 K}{\lambda^2}\left\|\theta_1-\theta_2\right\|_2.\end{aligned}$$

All the above imply

$$
\begin{aligned}
&|H(V_1,\theta_1)-H(V_2,\theta_2)| \le \sqrt{\left|h(V_1,\theta_1)^\top G_h(\theta_1)^{-1}h(V_1,\theta_1)-h(V_2,\theta_2)^\top G_h(\theta_2)^{-1}h(V_2,\theta_2)\right|}\\
\le&\sqrt{\left\|h(V_1,\theta_1)-h(V_2,\theta_2)\right\|_2\cdot\frac{1}{\lambda}\cdot 2KH\kappa_1}+\sqrt{2KH\kappa_1\cdot\left\|G_h(\theta_1)^{-1}-G_h(\theta_2)^{-1}\right\|_2\cdot 2KH\kappa_1}\\
+&\sqrt{2KH\kappa_1\cdot\frac{1}{\lambda}\cdot\left\|h(V_1,\theta_1)-h(V_2,\theta_2)\right\|_2}\\
\le&2\sqrt{K(2H\kappa_2\left\|\theta_1-\theta_2\right\|_2+2\kappa_1\left\|V_1-V_2\right\|_\infty)\cdot\frac{1}{\lambda}\cdot 2KH\kappa_1}+\sqrt{2KH\kappa_1\cdot\frac{2\kappa_1\kappa_2K}{\lambda^2}\left\|\theta_1-\theta_2\right\|_2\cdot 2KH\kappa_1}\\
\le&\left(4\sqrt{K^3H^2\kappa_1\kappa_2\frac{1}{\lambda}}+\sqrt{8K^3H^2\kappa_1^3\kappa_2\frac{1}{\lambda^2}}\right)\sqrt{\left\|\theta_1-\theta_2\right\|_2}+4\sqrt{K^3\kappa_1^2H\frac{1}{\lambda}\left\|V_1-V_2\right\|_\infty}
\end{aligned}
$$

Then a ϵ-covering net of $\{H(V,\theta)\}$ can be constructed by the union of $\frac{\epsilon^2}{4\left(4\sqrt{K^3H^2\kappa_1\kappa_2\frac{1}{\lambda}}+\sqrt{8K^3H^2\kappa_1^3\kappa_2\frac{1}{\lambda^2}}\right)^2}$-covering net of $\{\theta\in\Theta\}$ and $\frac{\epsilon^2}{4(4\sqrt{K^3\kappa_1^2H\frac{1}{\lambda}})^2}$-covering net of $\mathcal{V}$ in Lemma C.11.9. The covering number $\mathcal{N}_\epsilon$ satisfies

$$
\begin{aligned}
\log\mathcal{N}_\epsilon \le& d\log\left(1+\frac{8C_\Theta\left(4\sqrt{K^3H^2\kappa_1\kappa_2\frac{1}{\lambda}}+\sqrt{8K^3H^2\kappa_1^3\kappa_2\frac{1}{\lambda^2}}\right)^2}{\epsilon^2}\right)\\
&+d\log\left(1+\frac{8C_\Theta(\kappa_1\sqrt{C_\Theta}+2\sqrt{B\kappa_1\kappa_2})^2}{\frac{\epsilon^4}{16(4\sqrt{K^3\kappa_1^2H\frac{1}{\lambda}})^4}}\right)+d^2\log\left(1+\frac{8\sqrt{d}B\kappa_1^2}{\frac{\epsilon^4}{16(4\sqrt{K^3\kappa_1^2H\frac{1}{\lambda}})^4}}\right).
\end{aligned}
$$

Step3. First note by definition in Step2

$$
\left\|\sum_{k=1}^K\left(f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi_{h,k})-f(\theta_{\mathbb{T}V^\star_{h+1}},\phi_{h,k})-\widehat{V}_{h+1}(s^k_{h+1})+V^\star_{h+1}(s^k_{h+1})\right)\cdot\nabla_\theta^\top f(\widehat{\theta}_h,\phi_{h,k})\right\|_{\Sigma_h^{-1}}=H(\widehat{V}_{h+1},\widehat{\theta}_h)
$$

and with probability $1-\delta$

$$\begin{aligned}|\eta_k(\widehat{V}_{h+1})| =& |f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi_{h,k}) - f(\theta_{\mathbb{T}V^\star_{h+1}}, \phi_{h,k}) - \widehat{V}_{h+1}(s^k_{h+1}) + V^\star_{h+1}(s^k_{h+1})| \\ \leq& \kappa_1 \cdot \left\|\theta_{\mathbb{T}\widehat{V}_{h+1}} - \theta^\star_h\right\|_2 + \left\|\widehat{V}_{h+1} - V^\star_{h+1}\right\|_\infty \\ \leq& \kappa_1 \sqrt{\frac{36H^2(\log(H/\delta) + C_{d,\log K}) + 2\lambda C^2_\Theta}{\kappa K}} + C\left(\kappa_1 H^2\sqrt{\frac{d^2}{\kappa K}}\right) = \widetilde{O}\left(\kappa_1 H^2\sqrt{\frac{d^2}{\kappa K}}\right)\end{aligned} \tag{C.7}$$

where the second inequality uses $\theta_{\mathbb{T}V^\star_{h+1}} = \theta^\star_h$ and the third inequality uses Theorem C.6.1 and Theorem C.6.2. The last equal sign is due to $C_{d,\log K} \leq \widetilde{O}(d^2)$ (recall Lemma C.6.1).

Now choosing $\epsilon = O(1/K)$ in Step2 and union bound over both (C.7) and covering number in Step2, we obtain with probability $1-\delta$,

$$H(\widehat{V}_{h+1}, \widehat{\theta}_h) = \left\|\sum_{k=1}^K x_k(\widehat{\theta}_h)\eta_k(\widehat{V}_{h+1})\right\|_{G_h(\widehat{\theta}_h)^{-1}} \leq \widetilde{O}(\kappa_1 H^2\sqrt{\frac{d^2}{\kappa K}})\sqrt{d+d^2} = \widetilde{O}(\frac{\kappa_1 H^2 d^2}{\sqrt{\kappa K}}) \tag{C.8}$$

where we absorb all the Polylog terms. Meanwhile, by Lemma C.11.5 with probability $1-\delta$,

$$\left\|\nabla f(\widehat{\theta}_h, \phi_{s,a})\right\|_{\Sigma_h^{-1}} \leq \frac{2\kappa_1}{\sqrt{\kappa K}} + O(\frac{1}{K}). \tag{C.9}$$

Finally, by (C.8) and (C.9) and a union bound, we have with probability $1-\delta$,

$$\begin{aligned}|I_2| :=& \left|\nabla f(\widehat{\theta}_h, \phi(s,a))\Sigma_h^{-1}[\sum_{k=1}^K \left(f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi_{h,k}) - f(\theta_{\mathbb{T}V^\star_{h+1}}, \phi_{h,k}) - \widehat{V}_{h+1}(s^k_{h+1}) + V^\star_{h+1}(s^k_{h+1})\right)\cdot \nabla^\top_\theta f(\widehat{\theta}_h, \phi_{h,k})]\right| \\ \leq& \left\|\nabla f(\widehat{\theta}_h, \phi_{s,a})\right\|_{\Sigma_h^{-1}} \left\|\sum_{k=1}^K \left(f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi_{h,k}) - f(\theta_{\mathbb{T}V^\star_{h+1}}, \phi_{h,k}) - \widehat{V}_{h+1}(s^k_{h+1}) + V^\star_{h+1}(s^k_{h+1})\right)\cdot \nabla^\top_\theta f(\widehat{\theta}_h, \phi_{h,k})\right\|_{\Sigma_h^{-1}} \\ =& \left\|\nabla f(\widehat{\theta}_h, \phi_{s,a})\right\|_{\Sigma_h^{-1}} \cdot H(\widehat{V}_{h+1}, \widehat{\theta}_h) \leq \left(\frac{2\kappa_1}{\sqrt{\kappa K}} + O(\frac{1}{K})\right)\cdot \widetilde{O}(\frac{\kappa_1 H^2 d^2}{\sqrt{\kappa K}}) = \widetilde{O}(\frac{\kappa_1^2 H^2 d^2}{\kappa K}) + \widetilde{O}(\frac{1}{K^{3/2}})\end{aligned}$$

where the first inequality is Cauchy–Schwarz inequality.

C.4.4 Bounding the main term I_1

In this section, we bound the dominate term

$$I_1 := \nabla f(\widehat{\theta}_h, \phi(s,a))\Sigma_h^{-1}[\sum_{k=1}^{K}\left(f(\theta_{\mathbb{T}V_{h+1}^\star}, \phi_{h,k}) - r_{h,k} - V_{h+1}^\star(s_{h+1}^k)\right)\cdot\nabla_\theta^\top f(\widehat{\theta}_h, \phi_{h,k})].$$

First of all, by Cauchy–Schwarz inequality, we have

$$|I_1| \le \left\|\nabla f(\widehat{\theta}_h, \phi(s,a))\right\|_{\Sigma_h^{-1}} \cdot \left\|\sum_{k=1}^{K}\left(f(\theta_{\mathbb{T}V_{h+1}^\star}, \phi_{h,k}) - r_{h,k} - V_{h+1}^\star(s_{h+1}^k)\right)\cdot\nabla_\theta^\top f(\widehat{\theta}_h, \phi_{h,k})\right\|_{\Sigma_h^{-1}}. \tag{C.10}$$

Then we have the following Lemma to bound I_1.

Lemma C.4.6. *With probability* $1-\delta$*,*

$$|I_1| \le 4Hd\left\|\nabla f(\widehat{\theta}_h, \phi(s,a))\right\|_{\Sigma_h^{-1}} \cdot C_{\delta,\log K} + \widetilde{O}(\frac{\kappa_1}{\sqrt{\kappa}K}),$$

where $C_{\delta,\log K}$ *only contains Polylog terms.*

Proof: [Proof of Lemma C.4.6] **Step1.** Let the fixed $\theta \in \Theta$ be arbitrary and define $x_k(\theta) = \nabla_\theta f(\theta, \phi_{h,k})$. Next, define $G_h(\theta) = \sum_{k=1}^{K}\nabla f(\theta, \phi(s_h^k, a_h^k))\cdot\nabla f(\theta, \phi(s_h^k, a_h^k))^\top + \lambda I_d$, then $\|x_k\|_2 \le \kappa_1$. Also denote $\eta_k := f(\theta_{\mathbb{T}V_{h+1}^\star}, \phi_{h,k}) - r_{h,k} - V_{h+1}^\star(s_{h+1}^k)$, then $\mathbb{E}[\eta_k | s_h^k, a_h^k] = 0$ and $|\eta_k| \le H$. Now by self-normalized Hoeffding's inequality (Lemma C.11.3), with probability $1-\delta$ (recall $t := K$ in Lemma C.11.3),

$$\left\|\sum_{k=1}^{K} x_k(\theta)\eta_k\right\|_{G_h(\theta)^{-1}} \le 2H\sqrt{d\log\left(\frac{\lambda + K\kappa_1}{\lambda\delta}\right)}.$$

Step2. Define $h(\theta) := \sum_{k=1}^{K} x_k(\theta)\eta_k$ and $H(\theta) := \left\|\sum_{k=1}^{K} x_k(\theta)\eta_k\right\|_{G_h(\theta)^{-1}}$, then note by defini-

tion $|\eta_k| \le H$, which implies $\|h(\theta)\|_2 \le KH\kappa_1$ and by $x_k(\theta_1) - x_k(\theta_2) = \nabla^2_{\theta\theta} f(\xi, \phi) \cdot (\theta_1 - \theta_2)$,

$$\left\|h(\theta_1) - h(\theta_2)\right\|_2 \le K \max_k \left(H \left\|x_k(\theta_1) - x_k(\theta_2)\right\|_2\right) \le HK\kappa_2 \left\|\theta_1 - \theta_2\right\|_2.$$

Furthermore,

$$\begin{aligned}
&\left\|G_h(\theta_1)^{-1} - G_h(\theta_2)^{-1}\right\|_2 \le \left\|G_h(\theta_1)^{-1}\right\|_2 \left\|G_h(\theta_1) - G_h(\theta_2)\right\|_2 \left\|G_h(\theta_2)^{-1}\right\|_2 \\
\le &\frac{1}{\lambda^2} K \sup_k \left\|\nabla f(\theta_1, \phi_{h,k}) \cdot \nabla f(\theta_1, \phi_{h,k})^\top - \nabla f(\theta_2, \phi_{h,k}) \cdot \nabla f(\theta_2, \phi_{h,k})^\top\right\|_2 \\
\le &\frac{2\kappa_1 K}{\lambda^2} \kappa_2 \left\|\theta_1 - \theta_2\right\|_2 = \frac{2\kappa_1\kappa_2 K}{\lambda^2} \left\|\theta_1 - \theta_2\right\|_2.
\end{aligned}$$

All the above imply

$$\begin{aligned}
&|H(\theta_1) - H(\theta_2)| \le \sqrt{\left|h(\theta_1)^\top G_h(\theta_1)^{-1} h(\theta_1) - h(\theta_2)^\top G_h(\theta_2)^{-1} h(\theta_2)\right|} \\
\le &\sqrt{\left\|h(\theta_1) - h(\theta_2)\right\|_2 \cdot \frac{1}{\lambda} \cdot KH\kappa_1} + \sqrt{KH\kappa_1 \cdot \left\|G_h(\theta_1)^{-1} - G_h(\theta_2)^{-1}\right\|_2 \cdot KH\kappa_1} \\
&+ \sqrt{KH\kappa_1 \cdot \frac{1}{\lambda} \cdot \left\|h(\theta_1) - h(\theta_2)\right\|_2} \\
\le &2\sqrt{KH\kappa_2 \left\|\theta_1 - \theta_2\right\|_2 \cdot \frac{1}{\lambda} \cdot KH\kappa_1} + \sqrt{KH\kappa_1 \cdot \frac{2\kappa_1\kappa_2 K}{\lambda^2} \left\|\theta_1 - \theta_2\right\|_2 \cdot KH\kappa_1} \\
\le &\left(\sqrt{4K^2H^2\kappa_1\kappa_2/\lambda} + \sqrt{2K^3H^2\kappa_1^3\kappa_2/\lambda^2}\right) \sqrt{\left\|\theta_1 - \theta_2\right\|_2}
\end{aligned}$$

Then a ϵ-covering net of $\{H(\theta)\}$ can be constructed by the union of $\frac{\epsilon^2}{\left(\sqrt{4K^2H^2\kappa_1\kappa_2/\lambda} + \sqrt{2K^3H^2\kappa_1^3\kappa_2/\lambda^2}\right)^2}$-covering net of $\{\theta \in \Theta\}$. By Lemma C.11.8, the covering number $\mathcal{N}_\epsilon$ satisfies

$$\log \mathcal{N}_\epsilon \le d \log \left(1 + \frac{2C_\Theta \left(\sqrt{4K^2H^2\kappa_1\kappa_2/\lambda} + \sqrt{2K^3H^2\kappa_1^3\kappa_2/\lambda^2}\right)^2}{\epsilon^2}\right) = \widetilde{O}(d)$$

Step3. First note by definition in Step2

$$\left\|\sum_{k=1}^{K}\left(f(\theta_{\mathbb{T}V^{\star}_{h+1}},\phi_{h,k})-r_{h,k}-V^{\star}_{h+1}(s^{k}_{h+1})\right)\cdot\nabla^{\top}_{\theta}f(\widehat{\theta}_h,\phi_{h,k})\right\|_{\Sigma^{-1}_h}=H(\widehat{\theta}_h)$$

Now choosing $\epsilon = O(1/K)$ in Step2 and union bound over the covering number in Step2, we obtain with probability $1-\delta$,

$$H(\widehat{\theta}_h)=\left\|\sum_{k=1}^{K}x_k(\widehat{\theta}_h)\eta_k\right\|_{G_h(\widehat{\theta}_h)^{-1}}\leq 2H\sqrt{d\left[\log\left(\frac{\lambda+K\kappa_1}{\lambda\delta}\right)+\widetilde{O}(d)\right]}+O(\frac{1}{K}). \tag{C.11}$$

where we absorb all the Polylog terms. Combing above with (C.10), we obtain with probability $1-\delta$,

$$\begin{aligned}
|I_1|\leq&\left\|\nabla f(\widehat{\theta}_h,\phi(s,a))\right\|_{\Sigma^{-1}_h}\cdot\left\|\sum_{k=1}^{K}\left(f(\theta_{\mathbb{T}V^{\star}_{h+1}},\phi_{h,k})-r_{h,k}-V^{\star}_{h+1}(s^{k}_{h+1})\right)\cdot\nabla^{\top}_{\theta}f(\widehat{\theta}_h,\phi_{h,k})\right\|_{\Sigma^{-1}_h}\\
\leq&\left\|\nabla f(\widehat{\theta}_h,\phi(s,a))\right\|_{\Sigma^{-1}_h}\cdot\left(2H\sqrt{d\left[\log\left(\frac{\lambda+K\kappa_1}{\lambda\delta}\right)+\widetilde{O}(d)\right]}+O(\frac{1}{K})\right)\\
\leq&4Hd\left\|\nabla f(\widehat{\theta}_h,\phi(s,a))\right\|_{\Sigma^{-1}_h}\cdot C_{\delta,\log K}+\widetilde{O}(\frac{\kappa_1}{\sqrt{\kappa}K}),
\end{aligned}$$

where $C_{\delta,\log K}$ only contains Polylog terms.

C.4.5 Analyzing Hot_2 in (C.6)

Lemma C.4.7. *Recall* $\mathrm{Hot}_2 := \nabla f(\widehat{\theta}_h,\phi(s,a))\Sigma^{-1}_h\left[\widetilde{R}_K(\theta_{\mathbb{T}\widehat{V}_{h+1}})+\lambda\theta_{\mathbb{T}\widehat{V}_{h+1}}\right]$. *If the number of episode* K *satisfies* $K\geq\max\left\{512\frac{\kappa_1^4}{\kappa^2}\left(\log(\frac{2d}{\delta})+d\log(1+\frac{4\kappa_1^3\kappa_2 C_\Theta K^3}{\kappa\lambda^2})\right),\frac{4\lambda}{\kappa}\right\}$, *then with proba-*

bility $1-\delta$,

$$\left|\nabla f(\widehat{\theta}_h, \phi(s,a))\Sigma_h^{-1}\left[\widetilde{R}_K(\theta_{\mathbb{T}\widehat{V}_{h+1}}) + \lambda\theta_{\mathbb{T}\widehat{V}_{h+1}}\right]\right| \leq \widetilde{O}\left(\frac{\kappa_2 \max(\frac{\kappa_1}{\kappa}, \frac{1}{\sqrt{\kappa}})d^2H^2 + \frac{d^2H^3\kappa_3 + \lambda\kappa_1 C_\Theta}{\kappa}}{K}\right)$$

where $\widetilde{O}$ *absorbs all the constants and Polylog terms.*

Proof: [Proof of Lemma C.4.7]

Step1: we first show with probability $1-\delta$

$$\left|\nabla f(\widehat{\theta}_h, \phi(s,a))\Sigma_h^{-1}\widetilde{R}_K(\theta_{\mathbb{T}\widehat{V}_{h+1}})\right| \leq \widetilde{O}(\frac{1}{K}).$$

Recall by plug in $\theta_{\mathbb{T}\widehat{V}_{h+1}}$ in (C.2), we have

$$Z_h(\theta_{\mathbb{T}\widehat{V}_{h+1}}|\widehat{V}_{h+1}) - Z_h(\widehat{\theta}_h|\widehat{V}_{h+1}) = \frac{\partial}{\partial\theta}Z_h(\widehat{\theta}_h|\widehat{V}_{h+1})(\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h) + R_K(\theta_{\mathbb{T}\widehat{V}_{h+1}}), \tag{C.12}$$

and by second-order Taylor's Theorem we have

$$\begin{aligned}\left\|R_K(\theta_{\mathbb{T}\widehat{V}_{h+1}})\right\|_2 =& \left\|Z_h(\theta_{\mathbb{T}\widehat{V}_{h+1}}|\widehat{V}_{h+1}) - Z_h(\widehat{\theta}_h|\widehat{V}_{h+1}) - \frac{\partial}{\partial\theta}Z_h(\widehat{\theta}_h|\widehat{V}_{h+1})(\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h)\right\|_2 \\ =& \frac{1}{2}\left\|(\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h)^\top \frac{\partial^2}{\partial\theta\partial\theta}Z_h(\xi|\widehat{V}_{h+1})(\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h)\right\|_2 \\ \leq& \frac{1}{2}\kappa_{z_2}\left\|\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h\right\|_2^2\end{aligned} \tag{C.13}$$

Note

$$\begin{aligned}\frac{\partial^2}{\partial\theta\theta}Z_h(\theta|\widehat{V}_{h+1})\Big|_{\theta=\xi} =& \frac{\partial}{\partial\theta}\Sigma_h^s = \sum_{k=1}^K \frac{\partial}{\partial\theta}\left\{\left(f(\xi,\phi_{h,k}) - r_{h,k} - \widehat{V}_{h+1}(s_{h+1}^k)\right)\cdot\nabla_{\theta\theta}^2 f(\xi,\phi_{h,k})\right\} \\ &+ \sum_{k=1}^K \frac{\partial}{\partial\theta}\left(\nabla_\theta f(\xi,\phi_{h,k})\nabla_\theta^\top f(\xi,\phi_{h,k}) + \lambda I_d\right)\end{aligned} \tag{C.14}$$

Therefore, we can bound κ_{z_2} with $\kappa_{z_2} \leq (H\kappa_3 + 3\kappa_1\kappa_2)K$ and this implies with probability

$1-\delta/2$,

$$\begin{aligned}\left\|R_K(\theta_{\mathbb{T}\widehat{V}_{h+1}})\right\|_2 \leq& \frac{1}{2}\kappa_{z_2}\left\|\theta_{\mathbb{T}\widehat{V}_{h+1}}-\widehat{\theta}_h\right\|_2^2 \leq \frac{1}{2}(H\kappa_3+3\kappa_1\kappa_2)K\cdot\left\|\theta_{\mathbb{T}\widehat{V}_{h+1}}-\widehat{\theta}_h\right\|_2^2\\ \leq& \frac{1}{2}(H\kappa_3+3\kappa_1\kappa_2)K\cdot\frac{36H^2(\log(H/\delta)+C_{d,\log K})+2\lambda C_\Theta^2}{\kappa K}\\ \leq& \widetilde{O}((H\kappa_3+3\kappa_1\kappa_2)H^2d^2/\kappa).\end{aligned}$$

And by Corollary C.4.1 with probability $1-\delta/2$,

$$\left\|\Delta_{\Sigma_h^s}(\widehat{\theta}_h-\theta_{\mathbb{T}\widehat{V}_{h+1}})\right\|_2 \leq \widetilde{O}(1),$$

Therefore, by Lemma C.11.5 and a union bound with probability $1-\delta$,

$$\begin{aligned}&|\nabla f(\widehat{\theta}_h,\phi(s,a))^\top\Sigma_h^{-1}\widetilde{R}_K(\theta_{\mathbb{T}\widehat{V}_{h+1}})| = \left|\nabla f(\widehat{\theta}_h,\phi(s,a))^\top\Sigma_h^{-1}\left(\Delta_{\Sigma_h^s}(\widehat{\theta}_h-\theta_{\mathbb{T}\widehat{V}_{h+1}})+R_K(\theta_{\mathbb{T}\widehat{V}_{h+1}})\right)\right|\\ &\leq \left\|\nabla f(\widehat{\theta}_h,\phi(s,a))\right\|_{\Sigma_h^{-1}}\left\|\Delta_{\Sigma_h^s}(\widehat{\theta}_h-\theta_{\mathbb{T}\widehat{V}_{h+1}})+R_K(\theta_{\mathbb{T}\widehat{V}_{h+1}})\right\|_{\Sigma_h^{-1}}\\ &\leq \left(\frac{2\kappa_1}{\sqrt{\kappa K}}+O(\frac{1}{K})\right)\left\|\Delta_{\Sigma_h^s}(\widehat{\theta}_h-\theta_{\mathbb{T}\widehat{V}_{h+1}})+R_K(\theta_{\mathbb{T}\widehat{V}_{h+1}})\right\|_{\Sigma_h^{-1}}\\ &\leq \left(\frac{2\kappa_1}{\sqrt{\kappa K}}+O(\frac{1}{K})\right)\left(\frac{C}{\sqrt{K}}+O(\frac{1}{K})\right)=\widetilde{O}\left(\frac{\kappa_2\max(\frac{\kappa_1}{\kappa},\frac{1}{\sqrt{\kappa}})d^2H^2+\frac{d^2H^3\kappa_3}{\kappa}}{K}\right)\end{aligned}$$

where $\widetilde{O}$ absorbs all the constants and Polylog terms. Here the last inequality uses bound for $\left\|R_K(\theta_{\mathbb{T}\widehat{V}_{h+1}})\right\|_2$ and $\left\|\Delta_{\Sigma_h^s}(\widehat{\theta}_h-\theta_{\mathbb{T}\widehat{V}_{h+1}})\right\|_2$.

Step2: By Lemma C.11.5, with probability $1-\delta$,

$$\begin{aligned}&\left|\nabla f(\widehat{\theta}_h,\phi(s,a))\Sigma_h^{-1}\lambda\theta_{\mathbb{T}\widehat{V}_{h+1}}\right| \leq \lambda\left\|\nabla f(\widehat{\theta}_h,\phi(s,a))\right\|_{\Sigma_h^{-1}}\left\|\theta_{\mathbb{T}\widehat{V}_{h+1}}\right\|_{\Sigma_h^{-1}}\\ &\leq \lambda\left(\frac{2\kappa_1}{\sqrt{\kappa K}}+O(\frac{1}{K})\right)\cdot\left(\frac{2C_\Theta}{\sqrt{\kappa K}}+O(\frac{1}{K})\right)=\frac{4\lambda\kappa_1C_\Theta}{\kappa K}+O(\frac{1}{K^{\frac{3}{2}}})\end{aligned}$$

C.5 Proof of Theorem 4.3.2

Now we are ready to prove Theorem 4.3.2. In particular, we prove the first part. Also, recall that we consider the exact Bellman completeness ($\epsilon_{\mathcal{F}} = 0$).

C.5.1 The first part

Proof: [Proof of Theorem 4.3.2 (first part)] First of all, from the previous calculation (C.1), (C.6), we have

$$\begin{aligned}\left|\mathcal{P}_h\widehat{V}_{h+1}(s,a) - \widehat{\mathcal{P}}_h\widehat{V}_{h+1}(s,a)\right| \leq& \left|\nabla f(\widehat{\theta}_h, \phi(s,a))\left(\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h\right)\right| + \left|\mathrm{Hot}_{h,1}\right| \\ \leq& |I_1| + |I_2| + |I_3| + |\mathrm{Hot}_{h,2}| + \left|\mathrm{Hot}_{h,1}\right|\end{aligned}$$

Now by Lemma C.4.4, Lemma C.4.5, Lemma C.4.6, Lemma C.4.7 and Lemma C.4.1 (and a union bound), with probability $1-\delta$,

$$\begin{aligned}|I_3| \leq& \widetilde{O}(\sqrt{\frac{d^3H^2\kappa_2^2\kappa_1^2}{\kappa^3}})\frac{1}{K}, \\ |I_2| \leq& \widetilde{O}(\frac{\kappa_1^2H^2d^2}{\kappa K}) + \widetilde{O}(\frac{1}{K^{3/2}}), \\ |I_1| \leq& 4Hd\left\|\nabla f(\widehat{\theta}_h, \phi(s,a))\right\|_{\Sigma_h^{-1}} \cdot C_{\delta,\log K} + \widetilde{O}(\frac{\kappa_1}{\sqrt{\kappa}K}), \\ |\mathrm{Hot}_{2,h}| \leq& \widetilde{O}\left(\frac{\kappa_2\max(\frac{\kappa_1}{\kappa}, \frac{1}{\sqrt{\kappa}})d^2H^2 + \frac{d^2H^3\kappa_3 + \lambda\kappa_1 C_\Theta}{\kappa}}{K}\right), \\ |\mathrm{Hot}_{1,h}| \leq& \widetilde{O}(\frac{H^2\kappa_2 d^2}{\kappa})\frac{1}{K}.\end{aligned}$$

Finally, Plug the above into Lemma C.3.2, by a union bound over all $h \in [H]$, we have with

probability $1-\delta$, for any policy π,

$$\begin{aligned} v^{\pi}-v^{\widehat{\pi}} &\leq \sum_{h=1}^{H} 2\cdot \mathbb{E}_{\pi}\left[|I_1|+|I_2|+|I_3|+|\mathrm{Hot}_{h,2}|+\left|\mathrm{Hot}_{h,1}\right|\right] \\ &\leq \sum_{h=1}^{H} 8dH\mathbb{E}_{\pi}\left[\sqrt{\nabla^{\top} f(\widehat{\theta}_h,\phi(s_h,a_h))\Sigma_h^{-1}\nabla f(\widehat{\theta}_h,\phi(s_h,a_h))}\right]\cdot \iota + \widetilde{O}(\frac{C_{\mathrm{hot}}}{K}). \end{aligned}$$

where $\iota = C_{\delta,\log K}$ only contains Polylog terms and

$$C_{\mathrm{hot}} = \frac{\kappa_1 H}{\sqrt{\kappa}}+\frac{\kappa_1^2 H^3 d^2}{\kappa}+\sqrt{\frac{d^3H^4\kappa_2^2\kappa_1^2}{\kappa^3}}+\kappa_2\max(\frac{\kappa_1}{\kappa},\frac{1}{\sqrt{\kappa}})d^2H^3+\frac{d^2H^4\kappa_3+\lambda\kappa_1 C_{\Theta}}{\kappa}+\frac{H^3\kappa_2 d^2}{\kappa}$$

C.5.2 The second part

Next we prove the second part of Theorem 4.3.2. *Proof:* [Proof of Theorem 4.3.2 (second part)] **Step1.** Choose $\pi=\pi^{\star}$ in the first part, we have

$$0\leq v^{\pi^{\star}}-v^{\widehat{\pi}}\leq \sum_{h=1}^{H} 8dH\cdot\mathbb{E}_{\pi^{\star}}\left[\sqrt{\nabla_{\theta}^{\top} f(\widehat{\theta}_h,\phi(s_h,a_h))\Sigma_h^{-1}\nabla_{\theta} f(\widehat{\theta}_h,\phi(s_h,a_h))}\right]\cdot\iota+\widetilde{O}(\frac{C_{\mathrm{hot}}}{K}),$$

Next, by the triangular inequality of the norm to obtain

$$\begin{aligned} &\left|\left\|\nabla_{\theta} f(\widehat{\theta}_h,\phi(s_h,a_h))\right\|_{\Sigma_h^{-1}}-\left\|\nabla_{\theta} f(\theta_h^{\star},\phi(s_h,a_h))\right\|_{\Sigma_h^{-1}}\right| \\ \leq &\left\|\nabla_{\theta} f(\widehat{\theta}_h,\phi(s_h,a_h))-\nabla_{\theta} f(\theta_h^{\star},\phi(s_h,a_h))\right\|_{\Sigma_h^{-1}} \\ = &\left\|\nabla_{\theta\theta}^2 f(\xi,\phi(s_h,a_h))\cdot\left(\widehat{\theta}_h-\theta_h^{\star}\right)\right\|_{\Sigma_h^{-1}}, \end{aligned}$$

since with probability $1-\delta$,

$$\left\|\nabla^2_{\theta\theta} f(\xi, \phi(s_h, a_h)) \cdot \left(\widehat{\theta}_h - \theta^\star_h\right)\right\|_2 \le \kappa_2 \left\|\widehat{\theta}_h - \theta^\star_h\right\|_2 \le \widetilde{O}\left(\frac{\kappa_1\kappa_2 H^2 d}{\kappa}\sqrt{\frac{1}{K}}\right),$$

where the last inequality uses part three of Theorem C.6.2. Then by a union bound and Lemma C.11.5,

$$\left\|\nabla^2_{\theta\theta} f(\xi, \phi(s_h, a_h)) \cdot \left(\widehat{\theta}_h - \theta^\star_h\right)\right\|_{\Sigma_h^{-1}} \le \widetilde{O}\left(\frac{\kappa_1\kappa_2 H^2 d}{\kappa^{3/2}} \cdot \frac{1}{K}\right).$$

Step2. Next, we show with probability $1-\delta$,

$$\left\|\nabla_\theta f(\theta^\star_h, \phi(s_h, a_h))\right\|_{\Sigma_h^{-1}} \le 2\left\|\nabla_\theta f(\theta^\star_h, \phi(s_h, a_h))\right\|_{\Sigma_h^{\star-1}}.$$

First of all,

$$\begin{aligned}
&\left\|\frac{1}{K}\Sigma_h - \frac{1}{K}\Sigma^\star_h\right\|_2 = \left\|\frac{1}{K}\left(\sum_{k=1}^K \nabla f(\widehat{\theta}_h, \phi(s,a))\nabla f(\widehat{\theta}_h, \phi(s,a))^\top - \nabla f(\theta^\star_h, \phi(s,a))\nabla f(\theta^\star_h, \phi(s,a))^\top\right)\right\|_2 \\
\le &\sup_{s,a}\left(\left\|\left(\nabla f(\widehat{\theta}_h, \phi(s,a)) - \nabla f(\theta^\star_h, \phi(s,a))\right)\nabla f(\widehat{\theta}_h, \phi(s,a))\right\|_2\right. \\
&\left. + \left\|\left(\nabla f(\widehat{\theta}_h, \phi(s,a)) - \nabla f(\theta^\star_h, \phi(s,a))\right)\nabla f(\widehat{\theta}_h, \phi(s,a))\right\|_2\right) \\
\le &2\kappa_2\kappa_1\left\|\widehat{\theta}_h - \theta^\star_h\right\|_2 \le \widetilde{O}\left(\frac{\kappa_2\kappa_1^2 H^2 d}{\kappa}\sqrt{\frac{1}{K}}\right)
\end{aligned}$$

Second, by Lemma C.11.6 with probability $1-\delta$

$$\left\|\frac{\Sigma^\star_h}{K} - \mathbb{E}_\mu[\nabla_\theta f(\theta^\star_h, \phi)\nabla_\theta f(\theta^\star_h, \phi)^\top] - \frac{\lambda}{K}\right\| \le \frac{4\sqrt{2}\kappa_1^2}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2}$$

This implies

$$\begin{aligned}\left\|\frac{\Sigma_h^\star}{K}\right\| \leq& \left\|\mathbb{E}_\mu[\nabla_\theta f(\theta_h^\star,\phi)\nabla_\theta f(\theta_h^\star,\phi)^\top]\right\| + \frac{\lambda}{K} + \frac{4\sqrt{2}\kappa_1^2}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2}\\ \leq& \kappa_1^2 + \lambda + 4\sqrt{2}\kappa_1^2\left(\log\frac{2d}{\delta}\right)^{1/2}\end{aligned}$$

and also by *Weyl's spectrum theorem* and under the condition $K \geq \frac{128\kappa_1^4\log(2d/\delta)}{\kappa^2}$, with probability $1-\delta$

$$\begin{aligned}\lambda_{\min}(\frac{\Sigma_h^\star}{K}) \geq& \lambda_{\min}\left(\mathbb{E}_\mu[\nabla_\theta f(\theta_h^\star,\phi)\nabla_\theta f(\theta_h^\star,\phi)^\top]\right) + \frac{\lambda}{K} - \frac{4\sqrt{2}\kappa_1^2}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2}\\ \geq& \kappa + \frac{\lambda}{K} - \frac{4\sqrt{2}\kappa_1^2}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2} \geq \frac{\kappa}{2}\end{aligned}$$

then $\left\|(\frac{\Sigma_h^\star}{K})^{-1}\right\| \leq \frac{2}{\kappa}$. Similarly, with probability $1-\delta$, $\left\|(\frac{\Sigma_h}{K})^{-1}\right\| \leq \frac{2}{\kappa}$. Then by Lemma C.11.7,

$$\begin{aligned}\left\|\nabla_\theta f(\theta_h^\star,\phi(s,a))\right\|_{K\Sigma_h^{-1}} \leq& \left[1+\sqrt{\left\|K\Sigma_h^{\star-1}\right\|\left\|\Sigma_h^\star/K\right\|\cdot\left\|K\Sigma_h^{-1}\right\|\cdot\left\|\Sigma_h/K-\Sigma_h^\star/K\right\|}\right]\cdot\left\|\nabla_\theta f(\theta_h^\star,\phi(s,a))\right\|_{K\Sigma_h^{\star-1}}\\ \leq& \left[1+\sqrt{\frac{4}{\kappa^2}O(\kappa_1^2+\lambda)\widetilde{O}\left(\frac{\kappa_2\kappa_1^2H^2d}{\kappa}\sqrt{\frac{1}{K}}\right)}\right]\cdot\left\|\nabla_\theta f(\theta_h^\star,\phi(s,a))\right\|_{K\Sigma_h^{\star-1}}\\ \leq& 2\left\|\nabla_\theta f(\theta_h^\star,\phi(s,a))\right\|_{K\Sigma_h^{\star-1}}\end{aligned}$$

as long as $K \geq \widetilde{O}(\frac{(\kappa_1^2+\lambda)^2\kappa_2^2\kappa_1^2H^4d^2}{\kappa^6})$. The above is equivalently to

$$\left\|\nabla_\theta f(\theta_h^\star,\phi(s_h,a_h))\right\|_{\Sigma_h^{-1}} \leq 2\left\|\nabla_\theta f(\theta_h^\star,\phi(s_h,a_h))\right\|_{\Sigma_h^{\star-1}}.$$

Combining Step1, Step2 and a union bound, we have with probability $1-\delta$,

$$
\begin{aligned}
0 \leq & v^{\pi^\star} - v^{\widehat{\pi}} \leq \sum_{h=1}^{H} 8dH \cdot \mathbb{E}_{\pi^\star}\left[\sqrt{\nabla_\theta^\top f(\widehat{\theta}_h, \phi(s_h,a_h))\Sigma_h^{-1}\nabla_\theta f(\widehat{\theta}_h, \phi(s_h,a_h))}\right] \cdot \iota + \widetilde{O}(\frac{C_{\text{hot}}}{K}) \\
& \leq \sum_{h=1}^{H} 8dH \cdot \mathbb{E}_{\pi^\star}\left[\sqrt{\nabla_\theta^\top f(\theta_h^\star, \phi(s_h,a_h))\Sigma_h^{-1}\nabla_\theta f(\theta_h^\star, \phi(s_h,a_h))}\right] \cdot \iota + \widetilde{O}(\frac{C_{\text{hot}}}{K}) + \widetilde{O}\left(\frac{\kappa_1\kappa_2 H^4 d^2}{\kappa^{3/2}} \cdot \frac{1}{K}\right) \\
& \leq \sum_{h=1}^{H} 16dH \cdot \mathbb{E}_{\pi^\star}\left[\sqrt{\nabla_\theta^\top f(\theta_h^\star, \phi(s_h,a_h))\Sigma_h^{\star-1}\nabla_\theta f(\theta_h^\star, \phi(s_h,a_h))}\right] \cdot \iota + \widetilde{O}(\frac{C'_{\text{hot}}}{K})
\end{aligned}
$$

where $C'_{\text{hot}} = C_{\text{hot}} + \frac{\kappa_1\kappa_2 H^4 d^2}{\kappa^{3/2}}$.

C.6 Provable Efficiency by reduction to General Function Approximation

In this section, we bound the accuracy of the parameter difference $\left\|\widehat{\theta}_h - \theta_{\mathbb{T}\widehat{V}_{h+1}}\right\|_2$ via a reduction to General Function Approximation scheme in [13].

Recall the objective

$$
\ell_h(\theta) := \frac{1}{K}\sum_{k=1}^{K}\left[f\left(\theta, \phi(s_h^k, a_h^k)\right) - r(s_h^k, a_h^k) - \widehat{V}_{h+1}\left(s_{h+1}^k\right)\right]^2 + \frac{\lambda}{K}\cdot\|\theta\|_2^2 \tag{C.15}
$$

Then by definition, $\widehat{\theta}_h := \operatorname{argmin}_{\theta\in\Theta}\ell_h(\theta)$ and $\theta_{\mathbb{T}\widehat{V}_{h+1}}$ satisfies $f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi) = \mathcal{P}_h\widehat{V}_{h+1} + \delta_{\widehat{V}_{h+1}}$. Therefore, in this case, we have the following lemma:

Lemma C.6.1. *Fix $h \in [H]$. With probability $1-\delta$,*

$$
\mathbb{E}_\mu[\ell_h(\widehat{\theta}_h)] - \mathbb{E}_\mu[\ell_h(\theta_{\mathbb{T}\widehat{V}_{h+1}})] \leq \frac{36H^2(\log(1/\delta) + C_{d,\log K}) + \lambda C_\Theta^2}{K} + \sqrt{\frac{16H^3\epsilon_{\mathcal{F}}(\log(1/\delta) + C_{d,\log K})}{K}} + 4H\epsilon_{\mathcal{F}}.
$$

where the expectation over μ is taken w.r.t. $(s_h^k, a_h^k, s_{h+1}^k)$ $k = 1, ..., K$ only (i.e., first compute $\mathbb{E}_\mu[\ell_h(\theta)]$ for a fixed θ, then plug-in either $\widehat{\theta}_{h+1}$ or $\theta_{\mathbb{T}\widehat{V}_{h+1}}$). Here $C_{d,\log(K)} := d\log(1+$

$24C_\Theta(H+1)\kappa_1 K)+d\log\left(1+288H^2C_\Theta(\kappa_1\sqrt{C_\Theta}+2\sqrt{\kappa_1\kappa_2/\lambda})^2K^2\right)+d^2\log\left(1+288H^2\sqrt{d}\kappa_1^2K^2/\lambda\right).$

Proof: [Proof of Lemma C.6.1] **Step1:** we first prove the case where $\lambda = 0$.

Indeed, fix $h \in [H]$ and any function $V(\cdot) \in \mathbb{R}^S$. Similarly, define $f_V(s,a) := f(\theta_{\mathbb{T}V}, \phi) = \mathcal{P}_h V + \delta_V$. For any fixed $\theta \in \Theta$, denote $g(s,a) = f(\theta, \phi(s,a))$. Then define[2]

$$X(g, V, f_V) := (g(s,a) - r - V(s'))^2 - (f_V(s,a) - r - V(s'))^2.$$

Since all episodes are independent of each other, $X_k(g, V, f_V) := X(g(s_h^k, a_h^k), V(s_{h+1}^k), f_V(s_h^k, a_h^k))$ are independent r.v.s and it holds

$$\frac{1}{K}\sum_{k=1}^{K} X_k(g, V, f_V) = \ell(g) - \ell(f_V). \tag{C.16}$$

Next, the variance of X is bounded by:

$$\begin{aligned}
&\mathrm{Var}[X(g,V,f_V)] \le \mathbb{E}_\mu[X(g,f,f_V)^2] \\
=&\mathbb{E}_\mu\left[\left(\left(g(s_h,a_h) - r_h - V(s_{h+1})\right)^2 - \left(f_V(s_h,a_h) - r_h - V(s_{h+1})\right)^2\right)^2\right] \\
=&\mathbb{E}_\mu\left[(g(s_h,a_h) - f_V(s_h,a_h))^2(g(s_h,a_h) + f_V(s_h,a_h) - 2r_h - 2V(s_{h+1}))^2\right] \\
\le&4H^2\cdot\mathbb{E}_\mu[(g(s_h,a_h) - f_V(s_h,a_h))^2] \\
\le&4H^2\cdot\mathbb{E}_\mu\left[\left(g(s_h,a_h) - r_h - V(s_{h+1})\right)^2 - \left(f_V(s_h,a_h) - r_h - V(s_{h+1})\right)^2\right] + 8H^3\epsilon_{\mathcal{F}} \quad (*) \\
=&4H^2\cdot\mathbb{E}_\mu[X(g,f,f_V)] + 8H^3\epsilon_{\mathcal{F}}
\end{aligned}$$

[2]We abuse the notation here to use either $X(g,V,f_V)$ or $X(\theta,V,f_V)$. They mean the same quantity.

where the step $(*)$ comes from

$$\begin{aligned}
&\mathbb{E}_\mu\left[\left(g(s_h,a_h)-r_h-V(s_{h+1})\right)^2-\left(f_V(s_h,a_h)-r_h-V(s_{h+1})\right)^2\right]\\
=&\mathbb{E}_\mu\left[\left(g(s_h,a_h)-f_V(s_h,a_h)\right)\cdot\left(g(s_h,a_h)+f_V(s_h,a_h)-2r_h-2V(s_{h+1})\right)\right]\\
=&\mathbb{E}_\mu\left[\left(g(s_h,a_h)-f_V(s_h,a_h)\right)\cdot\left(g(s_h,a_h)-f_V(s_h,a_h)+2f_V(s_h,a_h)-2r_h-2V(s_{h+1})\right)\right]\\
=&\mathbb{E}_\mu\left[\left(g(s_h,a_h)-f_V(s_h,a_h)\right)^2\right]+\mathbb{E}_\mu\left[2(g(s_h,a_h)-f_V(s_h,a_h))\mathbb{E}_{P_h}[f_V(s_h,a_h)-r_h-V(s_{h+1})\mid s_h,a_h]\right]\\
\geq&\mathbb{E}_\mu\left[\left(g(s_h,a_h)-f_V(s_h,a_h)\right)^2\right]-2H\left\|\delta_V\right\|_\infty\geq\mathbb{E}_\mu\left[\left(g(s_h,a_h)-f_V(s_h,a_h)\right)^2\right]-2H\epsilon_{\mathcal{F}}
\end{aligned}\tag{C.17}$$

where the last step uses law of total expectation and the definition of f_V.

Therefore, by Bernstein inequality, with probability $1-\delta$,

$$\begin{aligned}
&\mathbb{E}_\mu[X(g,f,f_V)]-\frac{1}{K}\sum_{k=1}^K X_k(g,f,f_V)\\
\leq&\sqrt{\frac{2\mathrm{Var}[X(g,f,f_V)]\log(1/\delta)}{K}}+\frac{4H^2\log(1/\delta)}{3K}\\
\leq&\sqrt{\frac{8H^2\mathbb{E}_\mu[X(g,f,f_V)]\log(1/\delta)}{K}}+\sqrt{\frac{16H^3\epsilon_{\mathcal{F}}\log(1/\delta)}{K}}+\frac{4H^2\log(1/\delta)}{3K}.
\end{aligned}$$

Now, if we choose $g(s,a):=f(\widehat{\theta}_h,\phi(s,a))$, then $\widehat{\theta}_h$ minimizes $\ell_h(\theta)$, therefore, it also minimizes $\frac{1}{K}\sum_{k=1}^K X_i(\theta,\widehat{V}_{h+1},f_{\widehat{V}_{h+1}})$ and this implies

$$\frac{1}{K}\sum_{k=1}^K X_k(\widehat{\theta}_h,\widehat{V}_{h+1},f_{\widehat{V}_{h+1}})\leq\frac{1}{K}\sum_{k=1}^K X_k(\theta_{\mathbb{T}\widehat{V}_{h+1}},\widehat{V}_{h+1},f_{\widehat{V}_{h+1}})=0.$$

Therefore, we obtain

$$\mathbb{E}_\mu[X(\widehat{\theta}_h,\widehat{V}_{h+1},f_{\widehat{V}_{h+1}})]\leq\sqrt{\frac{8H^2\cdot\mathbb{E}_\mu[X(\widehat{\theta}_h,\widehat{V}_{h+1},f_{\widehat{V}_{h+1}})]\log(1/\delta)}{K}}+\sqrt{\frac{16H^3\epsilon_{\mathcal{F}}\log(1/\delta)}{K}}+\frac{4H^2\log(1/\delta)}{3K}.$$

However, the above does not hold with probability $1-\delta$ since $\widehat{\theta}_h$ and $\widehat{V}_{h+1}:=\min\{\max_a f(\widehat{\theta}_{h+1},\phi(\cdot,a))-$

$\sqrt{\nabla f(\widehat{\theta}_{h+1}, \phi(\cdot,a))^\top A \cdot \nabla f(\theta, \phi(\cdot,a))}, H\}$ (where A is certain symmetric matrix with bounded norm) depend on $\widehat{\theta}_h$ and $\widehat{\theta}_{h+1}$ which are data-dependent. Therefore, we need to further apply covering Lemma C.11.10 and choose $\epsilon = O(1/K)$ and a union bound to obtain with probability $1-\delta$,

$$\mathbb{E}_\mu[X(\widehat{\theta}_h, \widehat{V}_{h+1}, f_{\widehat{V}_{h+1}})] \le \sqrt{\frac{8H^2 \cdot \mathbb{E}_\mu[X(\widehat{\theta}_h, \widehat{V}_{h+1}, f_{\widehat{V}_{h+1}})](\log(1/\delta) + C_{d,\log K})}{K}} + \frac{7H^2(\log(1/\delta) + C_{d,\log K})}{3K} + \sqrt{\frac{16H^3\epsilon_{\mathcal{F}}(\log(1/\delta) + C_{d,\log K})}{K}} + 4H\epsilon_{\mathcal{F}}$$

where $C_{d,\log(K)} := \log(1+24C_\Theta(H+1)\kappa_1 K) + d\log\left(1 + 288H^2C_\Theta(\kappa_1\sqrt{C_\Theta} + 2\sqrt{\kappa_1\kappa_2/\lambda})^2K^2\right) + d^2\log\left(1 + 288H^2\sqrt{d}\kappa_1^2K^2/\lambda\right)$.[3] Solving this quadratic equation to obtain with probability $1-\delta$,

$$\mathbb{E}_\mu[X(\widehat{\theta}_h, \widehat{V}_{h+1}, f_{\widehat{V}_{h+1}})] \le \frac{36H^2(\log(1/\delta) + C_{d,\log K})}{K} + \sqrt{\frac{16H^3\epsilon_{\mathcal{F}}(\log(1/\delta) + C_{d,\log K})}{K}} + 4H\epsilon_{\mathcal{F}}$$

Now according to (C.16), by definition we finally have with probability $1-\delta$ (recall the expectation over μ is taken w.r.t. $(s_h^k, a_h^k, s_{h+1}^k)$ $k = 1, ..., K$ only)

$$\begin{aligned} &\mathbb{E}_\mu[\ell_h(\widehat{\theta}_{h+1})] - \mathbb{E}_\mu[\ell_h(\theta_{\mathbb{T}\widehat{V}_{h+1}})] = \mathbb{E}_\mu[X(\widehat{\theta}_h, \widehat{V}_{h+1}, f_{\widehat{V}_{h+1}})] \\ &\le \frac{36H^2(\log(1/\delta) + C_{d,\log K})}{K} + \sqrt{\frac{16H^3\epsilon_{\mathcal{F}}(\log(1/\delta) + C_{d,\log K})}{K}} + 4H\epsilon_{\mathcal{F}}. \end{aligned} \tag{C.18}$$

Step2. If $\lambda > 0$, there is only extra term $\frac{\lambda}{K}\left(\left\|\widehat{\theta}_h\right\|_2 - \left\|\theta_{\mathbb{T}\widehat{V}_{h+1}}\right\|_2\right) \le \frac{\lambda}{K}\left\|\widehat{\theta}_h\right\|_2 \le \frac{\lambda C_\Theta^2}{K}$ in addition to above. This finishes the proof.

Theorem C.6.1 (Provable efficiency (Part I)). *Let $C_{d,\log K}$ be the same as Lemma C.6.1. Then*

[3] Here in our realization of Lemma C.11.9, we set $B = 1/\lambda$ (since $\left\|\Sigma_h^{-1}\right\|_2 \le 1/\lambda$).

denote $b_{d,K,\epsilon_{\mathcal{F}}} := \sqrt{\frac{16H^3\epsilon_{\mathcal{F}}(\log(1/\delta)+C_{d,\log K})}{K}} + 4H\epsilon_{\mathcal{F}}$, *with probability* $1-\delta$

$$\left\|\widehat{\theta}_h - \theta_{\mathbb{T}\widehat{V}_{h+1}}\right\|_2 \le \sqrt{\frac{36H^2(\log(H/\delta)+C_{d,\log K})+2\lambda C_\Theta^2}{\kappa K}} + \sqrt{\frac{b_{d,K,\epsilon_{\mathcal{F}}}}{\kappa}} + \sqrt{\frac{2H\epsilon_{\mathcal{F}}}{\kappa}},\ \forall h\in[H].$$

Proof: [Proof of Theorem C.6.1] Apply a union bound in Lemma C.6.1, we have with probability $1-\delta$,

$$\begin{aligned}
&\mathbb{E}_\mu[\ell_h(\widehat{\theta}_h)] - \mathbb{E}_\mu[\ell_h(\theta_{\mathbb{T}\widehat{V}_{h+1}})] \le \frac{36H^2(\log(H/\delta)+C_{d,\log K})+\lambda C_\Theta^2}{K} + b_{d,K,\epsilon_{\mathcal{F}}},\ \ \forall h\in[H]\\
\Rightarrow&\mathbb{E}_\mu[\ell_h(\widehat{\theta}_h) - \frac{\lambda}{K}\left\|\widehat{\theta}_h\right\|_2^2] - \mathbb{E}_\mu[\ell_h(\theta_{\mathbb{T}\widehat{V}_{h+1}}) - \frac{\lambda}{K}\left\|\theta_{\mathbb{T}\widehat{V}_{h+1}}\right\|_2^2] \le \frac{36H^2(\log(H/\delta)+C_{d,\log K})+2\lambda C_\Theta^2}{K} + b_{d,K,\epsilon_{\mathcal{F}}}
\end{aligned} \tag{C.19}$$

Now we prove for all $h\in[H]$,

$$\mathbb{E}_\mu\left[\left(f(\widehat{\theta}_h,\phi(\cdot,\cdot)) - f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(\cdot,\cdot))\right)^2\right] \le \mathbb{E}_\mu\left[\ell_h(\widehat{\theta}_h) - \frac{\lambda\left\|\widehat{\theta}_h\right\|_2^2}{K}\right] - \mathbb{E}_\mu\left[\ell_h(\theta_{\mathbb{T}\widehat{V}_{h+1}}) - \frac{\lambda\left\|\theta_{\mathbb{T}\widehat{V}_{h+1}}\right\|_2^2}{K}\right] + 2H\epsilon_{\mathcal{F}}. \tag{C.20}$$

Indeed, similar to (C.18), by definition we have

$$\begin{aligned}
&\mathbb{E}_\mu\left[\ell_h(\widehat{\theta}_h) - \frac{\lambda\left\|\widehat{\theta}_h\right\|_2^2}{K}\right] - \mathbb{E}_\mu\left[\ell_h(\theta_{\mathbb{T}\widehat{V}_{h+1}}) - \frac{\lambda\left\|\theta_{\mathbb{T}\widehat{V}_{h+1}}\right\|_2^2}{K}\right] = \mathbb{E}_\mu[X(\widehat{\theta}_h,\widehat{V}_{h+1},f_{\widehat{V}_{h+1}})]\\
=&\mathbb{E}_\mu\left[\left(f(\widehat{\theta}_h,\phi(\cdot,\cdot)) - f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(\cdot,\cdot))\right)^2\right]\\
+&\mathbb{E}_\mu\left[\left(f(\widehat{\theta}_h,\phi(s_h,a_h)) - f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(s_h,a_h))\right)\cdot\left(f\left(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(s_h,a_h)\right) - r_h - \widehat{V}_{h+1}(s_{h+1})\right)\right]\\
=&\mathbb{E}_\mu\left[\left(f(\widehat{\theta}_h,\phi(\cdot,\cdot)) - f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(\cdot,\cdot))\right)^2\right]\\
+&\mathbb{E}_\mu\left[\left(f(\widehat{\theta}_h,\phi(s_h,a_h)) - f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(s_h,a_h))\right)\cdot\mathbb{E}\left(f\left(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(s_h,a_h)\right) - r_h - \widehat{V}_{h+1}(s_{h+1})\middle| s_h,a_h\right)\right]\\
\ge&\mathbb{E}_\mu\left[\left(f(\widehat{\theta}_h,\phi(\cdot,\cdot)) - f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(\cdot,\cdot))\right)^2\right] - 2H\epsilon_{\mathcal{F}}
\end{aligned}$$

where the third identity uses μ is taken w.r.t. s_h, a_h, s_{h+1} (recall Lemma C.6.1) and law of total expectation. The first inequality uses the definition of $\theta_{\mathbb{T}\widehat{V}_{h+1}}$.

Now apply Assumption 4.2.3, we have

$$\mathbb{E}_\mu\left[\left(f(\widehat{\theta}_h, \phi(\cdot,\cdot)) - f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi(\cdot,\cdot))\right)^2\right] \geq \kappa \left\|\widehat{\theta}_h - \theta_{\mathbb{T}\widehat{V}_{h+1}}\right\|_2^2,$$

Combine the above with (C.19) and (C.20), we obtain the stated result.

Theorem C.6.2 (Provable efficiency (Part II))**.** *Let $C_{d,\log K}$ be the same as Lemma C.6.1 and suppose $\epsilon_{\mathcal{F}} = 0$. Furthermore, suppose $\lambda \leq 1/2C_\Theta^2$ and $K \geq \max\left\{512\frac{\kappa_1^4}{\kappa^2}\left(\log(\frac{2d}{\delta}) + d\log(1 + \frac{4\kappa_1^3\kappa_2 C_\Theta K^3}{\lambda^2})\right), \frac{4\lambda}{\kappa}\right\}$. Then, with probability $1-\delta$, $\forall h \in [H]$,*

$$\sup_{s,a}\left|f(\widehat{\theta}_h, \phi(s,a)) - f(\theta_h^\star, \phi(s,a))\right| \leq \left(\kappa_1 H\sqrt{\frac{36H^2(\log(H^2/\delta) + C_{d,\log K}) + 2\lambda C_\Theta^2}{\kappa}} + \frac{2H^2 d\kappa_1}{\sqrt{\kappa}}\right)\sqrt{\frac{1}{K}} + O(\frac{1}{K}).$$

Furthermore, we have with probability $1-\delta$,

$$\begin{aligned}\sup_h\left\|\widehat{V}_h - V_h^\star\right\|_\infty \leq& \left(\kappa_1 H\sqrt{\frac{36H^2(\log(H^2/\delta) + C_{d,\log K}) + 2\lambda C_\Theta^2}{\kappa}} + \frac{2H^2 d\kappa_1}{\sqrt{\kappa}}\right)\sqrt{\frac{1}{K}} + O(\frac{1}{K}) \\ =& \widetilde{O}\left(\kappa_1 H^2\sqrt{\frac{d^2}{\kappa}}\sqrt{\frac{1}{K}}\right)\end{aligned}$$

where $\widetilde{O}$ absorbs Polylog terms and higher order terms. Lastly, it also holds for all $h \in [H]$,

w.p. $1-\delta$

$$\begin{aligned}\left\|\widehat{\theta}_h-\theta_h^\star\right\|_2 \leq & \left(\kappa_1 H \frac{\sqrt{72H^2(\log(H^2/\delta)+C_{d,\log K})+4\lambda C_\Theta^2}}{\kappa}+\frac{4H^2 d\kappa_1}{\kappa}\right)\sqrt{\frac{1}{K}}+O(\frac{1}{K}) \\ =&\widetilde{O}\left(\frac{\kappa_1 H^2 d}{\kappa}\sqrt{\frac{1}{K}}\right)\end{aligned}$$

Proof: [Proof of Theorem C.6.2] **Step1:** we show the first result.

We prove this by backward induction. When $h=H+1$, by convention $f(\widehat{\theta}_h,\phi(s,a))=f(\theta_h^\star,\phi(s,a))=0$ so the base case holds. Suppose for $h+1$, with probability $1-(H-h)\delta$, it holds true that $\sup_{s,a}\left|f(\widehat{\theta}_{h+1},\phi(s,a))-f(\theta_{h+1}^\star,\phi(s,a))\right| \leq C_{h+1}\sqrt{\frac{1}{K}}+a(h+1)$, we next consider the case for $t=h$.

On one hand, by Theorem C.6.1, we have with probability $1-\delta/2$,

$$\begin{aligned}&\sup_{s,a}\left|f(\widehat{\theta}_h,\phi(s,a))-f(\theta_h^\star,\phi(s,a))\right| \\ \leq&\sup_{s,a}\left|f(\widehat{\theta}_h,\phi(s,a))-f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(s,a))\right|+\sup_{s,a}\left|f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(s,a))-f(\theta_h^\star,\phi(s,a))\right| \\ =&\sup_{s,a}\left|\nabla f(\xi,\phi(s,a))^\top(\widehat{\theta}_h-\theta_{\mathbb{T}\widehat{V}_{h+1}})\right|+\sup_{s,a}\left|f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(s,a))-f(\theta_{\mathbb{T}V_{h+1}^\star},\phi(s,a))\right| \\ \leq&\kappa_1\cdot\left\|\widehat{\theta}_h-\theta_{\mathbb{T}\widehat{V}_{h+1}}\right\|_2+\sup_{s,a}\left|\mathcal{P}_{h,s,a}\widehat{V}_{h+1}-\mathcal{P}_{h,s,a}V_{h+1}^\star\right| \\ \leq&\kappa_1\sqrt{\frac{36H^2(\log(H/\delta)+C_{d,\log K})+2\lambda C_\Theta^2}{\kappa K}}+\left\|\widehat{V}_{h+1}-V_{h+1}^\star\right\|_\infty,\end{aligned}$$

Recall $\widehat{V}_{h+1}(\cdot):=\min\{\max_a f(\widehat{\theta}_{h+1},\phi(\cdot,a))-\Gamma_h(\cdot,a),H\}$ and $V_{h+1}^\star(\cdot)=\max_a f(\theta_{h+1}^\star,\phi(\cdot,a))=\min\{\max_a f(\theta_{h+1}^\star,\phi(\cdot,a)),H\}$, we obtain

$$\left\|\widehat{V}_{h+1}-V_{h+1}^\star\right\|_\infty \leq \sup_{s,a}\left|f(\widehat{\theta}_{h+1},\phi(s,a))-f(\theta_{h+1}^\star,\phi(s,a))\right|+\sup_{h,s,a}\Gamma_h(s,a) \tag{C.21}$$

Note the above holds true for any generic $\Gamma_h(s,a)$. In particular, according to Algorithm 3, we specify

$$\Gamma_h(\cdot,\cdot) = dH\sqrt{\nabla_\theta f(\widehat{\theta}_h,\phi(\cdot,\cdot))^\top \Sigma_h^{-1}\nabla_\theta f(\widehat{\theta}_h,\phi(\cdot,\cdot))}\left(+\widetilde{O}(\frac{1}{K})\right)$$

and by Lemma C.11.5, with probability $1-\delta$,

$$\Gamma_h \le \frac{2dH\kappa_1}{\sqrt{\kappa K}} + \widetilde{O}(\frac{1}{K})$$

and by a union bound this implies with probability $1-(H-h+1)\delta$,

$$\begin{aligned}
&\sup_{s,a}\left|f(\widehat{\theta}_h,\phi(s,a)) - f(\theta_h^\star,\phi(s,a))\right| \\
\le& C_{h+1}\sqrt{\frac{1}{K}} + \kappa_1\sqrt{\frac{36H^2(\log(H/\delta)+C_{d,\log K})+2\lambda C_\Theta^2}{\kappa K}} + \frac{2dH\kappa_1}{\sqrt{\kappa K}} + \widetilde{O}(\frac{1}{K}) := C_h\sqrt{\frac{1}{K}} + \widetilde{O}(\frac{1}{K})
\end{aligned}$$

Solving for C_h, we obtain $C_h \le \kappa_1 H\sqrt{\frac{36H^2(\log(H/\delta)+C_{d,\log K})+2\lambda C_\Theta^2}{\kappa}} + H\frac{2dH\kappa_1}{\sqrt{\kappa}}$ for all H. By a union bound (replacing δ by δ/H), we obtain the stated result.

Step2: Utilizing the intermediate result (C.21), we directly have with probability $1-\delta$,

$$\sup_h\left\|\widehat{V}_h - V_h^\star\right\|_\infty \le \sup_{s,a}\left|f(\widehat{\theta}_h,\phi(s,a)) - f(\theta_h^\star,\phi(s,a))\right| + \frac{2dH\kappa_1}{\sqrt{\kappa K}} + O(\frac{1}{K}),$$

where $\sup_{s,a}\left|f(\widehat{\theta}_h,\phi(s,a)) - f(\theta_h^\star,\phi(s,a))\right|$ can be bounded using Step1.

Step3: Denote $M := \left(\kappa_1 H\sqrt{\frac{36H^2(\log(H^2/\delta)+C_{d,\log K})+2\lambda C_\Theta^2}{\kappa}} + \frac{2H^2 d\kappa_1}{\sqrt{\kappa}}\right)\sqrt{\frac{1}{K}} + O(\frac{1}{K})$, then by Step1 we have with probability $1-\delta$ (here ξ is some point between $\widehat{\theta}_h$ and $\theta_h^\star$) for all $h\in[H]$

$$\begin{aligned}
M^2 \ge& \sup_{s,a}\left|f(\widehat{\theta}_h,\phi(s,a)) - f(\theta_h^\star,\phi(s,a))\right|^2 \\
\ge& \mathbb{E}_{\mu,h}[(f(\widehat{\theta}_h,\phi(s,a)) - f(\theta_h^\star,\phi(s,a)))^2] \ge \kappa\left\|\widehat{\theta}_h - \theta_h^\star\right\|_2^2
\end{aligned}$$

where the last inequality is by Assumption 4.2.3. Solve this to obtain the stated result.

C.7 With positive Bellman completeness coefficient $\epsilon_{\mathcal{F}} > 0$

In Theorem 4.3.2, we consider the case where $\epsilon_{\mathcal{F}} = 0$. If $\epsilon_{\mathcal{F}} > 0$, similar guarantee can be achieved with the measurement of model misspecification. For instance, the additional error $\sqrt{\frac{16H^3\epsilon_{\mathcal{F}}(\log(1/\delta)+C_{d,\log K})}{K}} + 4H\epsilon_{\mathcal{F}}$ will show up in Lemma C.6.1 (as stated in the current version), $\sqrt{\frac{b_{d,K,\epsilon_{\mathcal{F}}}}{\kappa}} + \sqrt{\frac{2H\epsilon_{\mathcal{F}}}{\kappa}}$ will show up in Lemma C.6.1. Then the decomposition in (C.1) will incur the extra $\delta_{\widehat{V}_{h+1}}$ term with $\delta_{\widehat{V}_{h+1}}$ might not be 0. The analysis with positive $\epsilon_{\mathcal{F}} > 0$ will make the proofs more intricate but incurs no additional technical challenge. Since the inclusion of this quantity is not our major focus, as a result, we only provide the proof for the case where $\epsilon_{\mathcal{F}} = 0$ so the readers can focus on the more critical components that characterize the hardness of differentiable function class.

C.8 VFQL and its analysis

We present the *vanilla fitted Q-learning* (VFQL) Algorithm 5 as follows. For VFQL, no pessimism is used and we assume $\widehat{\theta}_h \in \Theta$ without loss of generality.

Algorithm 5 Vanilla Fitted Q-Learning (VFQL)

1: **Input:** Offline Dataset $\mathcal{D} = \left\{\left(s_h^k, a_h^k, r_h^k, s_{h+1}^k\right)\right\}_{k,h=1}^{K,H}$. Denote $\phi_{h,k} := \phi(s_h^k, a_h^k)$.

2: **Initialization:** Set $\widehat{V}_{H+1}(\cdot) \leftarrow 0$ and $\lambda > 0$.

3: **for** $h = H, H-1, \ldots, 1$ **do**

4: Set $\widehat{\theta}_h \leftarrow \operatorname{argmin}_{\theta\in\Theta}\left\{\sum_{k=1}^K \left[f\left(\theta, \phi_{h,k}\right) - r_{h,k} - \widehat{V}_{h+1}(s_{h+1}^k)\right]^2 + \lambda\cdot\|\theta\|_2^2\right\}$

5: Set $\widehat{Q}_h(\cdot,\cdot) \leftarrow \min\left\{f(\widehat{\theta}_h, \phi(\cdot,\cdot)), H-h+1\right\}^+$

6: Set $\widehat{\pi}_h(\cdot \mid \cdot) \leftarrow \arg\max_{\pi_h}\left\langle \widehat{Q}_h(\cdot,\cdot), \pi_h(\cdot\mid\cdot)\right\rangle_{\mathcal{A}}$, $\widehat{V}_h(\cdot) \leftarrow \max_{\pi_h}\left\langle \widehat{Q}_h(\cdot,\cdot), \pi_h(\cdot\mid\cdot)\right\rangle_{\mathcal{A}}$

7: **end for**

8: **Output:** $\left\{\widehat{\pi}_h\right\}_{h=1}^H$.

C.8.1 Analysis for VFQL (Theorem 4.3.1)

Recall $\iota_h(s,a) := \mathcal{P}_h\widehat{V}_{h+1}(s,a) - \widehat{Q}_h(s,a)$ and the definition of Bellman operator C.3.1. Note $\min\{\cdot, H-h+1\}^+$ is a non-expansive operator, therefore we have

$$\begin{aligned}
|\iota_h(s,a)| =& |\mathcal{P}_h\widehat{V}_{h+1}(s,a) - \widehat{Q}_h(s,a)| = \left|\min\left\{\mathcal{P}_h\widehat{V}_{h+1}(s,a), H-h+1\right\}^+ - \min\left\{f(\widehat{\theta}_h, \phi(\cdot,\cdot)), H-h+1\right\}^+\right| \\
\le& \left|\mathcal{P}_h\widehat{V}_{h+1}(s,a) - f(\widehat{\theta}_h, \phi(\cdot,\cdot))\right| \le \left|f(\theta_{\mathbb{T}\widehat{V}_{h+1}}) - f(\widehat{\theta}_h, \phi(\cdot,\cdot))\right| + \epsilon_{\mathcal{F}}.
\end{aligned}$$

By Lemma C.3.1, we have for any π,

$$\begin{aligned}
v^\pi - v^{\widehat{\pi}} =& -\sum_{h=1}^H E_{\widehat{\pi}}[\iota_h(s_h,a_h)] + \sum_{h=1}^H E_\pi[\iota_h(s_h,a_h)] \le \sum_{h=1}^H E_{\widehat{\pi}}[|\iota_h(s_h,a_h)|] + \sum_{h=1}^H E_\pi[|\iota_h(s_h,a_h)|] \\
\le& \sum_{h=1}^H \mathbb{E}_{\widehat{\pi}}[|f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi(\cdot,\cdot)) - f(\widehat{\theta}_h, \phi(\cdot,\cdot))|] + \sum_{h=1}^H \mathbb{E}_\pi[|f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi(\cdot,\cdot)) - f(\widehat{\theta}_h, \phi(\cdot,\cdot))|] + 2H\epsilon_{\mathcal{F}} \\
\le& \sum_{h=1}^H \sqrt{\mathbb{E}_{\widehat{\pi}}[|f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi(\cdot,\cdot)) - f(\widehat{\theta}_h, \phi(\cdot,\cdot))|^2]} + \sum_{h=1}^H \sqrt{\mathbb{E}_\pi[|f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi(\cdot,\cdot)) - f(\widehat{\theta}_h, \phi(\cdot,\cdot))|^2]} + 2H\epsilon_{\mathcal{F}} \\
\le& 2\sqrt{C_{\text{eff}}} \sum_{h=1}^H \sqrt{\mathbb{E}_{\mu,h}[|f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi(\cdot,\cdot)) - f(\widehat{\theta}_h, \phi(\cdot,\cdot))|^2]} + 2H\epsilon_{\mathcal{F}}
\end{aligned} \tag{C.22}$$

where the second inequality uses Cauchy inequality and the third one uses the definition of concentrability coefficient 4.2.2.

Next, for VFQL, there is no pessimism therefore the quantity B in Lemma C.11.10 is zero, hence the covering number applied in Lemma C.6.1 is bounded by $C_{d,\log(K)} \leq \widetilde{O}(d)$ and

$$\mathbb{E}_\mu[\ell_h(\widehat{\theta}_h)] - \mathbb{E}_\mu[\ell_h(\theta_{\mathbb{T}\widehat{V}_{h+1}})] \leq \frac{36H^2(\log(1/\delta) + C_{d,\log K}) + \lambda C_\Theta^2}{K} + \sqrt{\frac{16H^3\epsilon_{\mathcal{F}}(\log(1/\delta) + C_{d,\log K})}{K}} + 4H\epsilon_{\mathcal{F}}.$$

Now leveraging (C.19) and (C.20) in Theorem C.6.1 to obtain

$$\begin{aligned}\mathbb{E}_\mu\left[\left(f(\widehat{\theta}_h, \phi(\cdot,\cdot)) - f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi(\cdot,\cdot))\right)^2\right] \leq & \mathbb{E}_\mu\left[\ell_h(\widehat{\theta}_h) - \frac{\lambda\left\|\widehat{\theta}_h\right\|_2^2}{K}\right] - \mathbb{E}_\mu\left[\ell_h(\theta_{\mathbb{T}\widehat{V}_{h+1}}) - \frac{\lambda\left\|\theta_{\mathbb{T}\widehat{V}_{h+1}}\right\|_2^2}{K}\right] + 2H\epsilon_{\mathcal{F}} \\ \leq & \frac{36H^2(\log(H/\delta) + C_{d,\log K}) + 2\lambda C_\Theta^2}{K} + b_{d,K,\epsilon_{\mathcal{F}}} + 2H\epsilon_{\mathcal{F}}\end{aligned}$$

Plug the above into (C.22), we obtain with probability $1-\delta$, for all policy π,

$$\begin{aligned}v^\pi - v^{\widehat{\pi}} \leq & 2\sqrt{C_{\text{eff}}}H\sqrt{\frac{36H^2(\log(H/\delta) + C_{d,\log K}) + 2\lambda C_\Theta^2}{K} + b_{d,K,\epsilon_{\mathcal{F}}} + 2H\epsilon_{\mathcal{F}}} + 2H\epsilon_{\mathcal{F}} \\ = & 2\sqrt{C_{\text{eff}}}H\sqrt{\frac{36H^2(\log(H/\delta) + C_{d,\log K}) + 2\lambda C_\Theta^2}{K} + \sqrt{\frac{16H^3\epsilon_{\mathcal{F}}(\log(1/\delta) + C_{d,\log K})}{K}} + 6H\epsilon_{\mathcal{F}}} + 2H\epsilon_{\mathcal{F}} \\ = & \sqrt{C_{\text{eff}}}H \cdot \widetilde{O}\left(\sqrt{\frac{H^2 d + \lambda C_\Theta^2}{K}} + \sqrt[4]{\frac{H^3 d\epsilon_{\mathcal{F}}}{K}}\right) + O(\sqrt{C_{\text{eff}}H^3\epsilon_{\mathcal{F}}} + H\epsilon_{\mathcal{F}})\end{aligned}$$

This finishes the proof of Theorem 4.3.1.

C.9 Proofs for VAFQL

In this section, we present the analysis for *variance-aware fitted Q learning* (VAFQL). Throughout the whole section, we assume $\epsilon_{\mathcal{F}} = 0$, *i.e.* the exact Bellman-Completeness holds. The algorithm is presented in the following. Before giving the proofs of Theorem 6, we first prove some useful lemmas.

Algorithm 6 Variance-Aware Fitted Q Learning (VAFQL)

Input: Split dataset $\mathcal{D} = \left\{\left(s_h^k, a_h^k, r_h^k\right)\right\}_{k,h=1}^{K,H}$ $\mathcal{D}' = \left\{\left(\bar{s}_h^k, \bar{a}_h^k, \bar{r}_h^k\right)\right\}_{k,h=1}^{K,H}$. Require β.
Initialization: Set $\widehat{V}_{H+1}(\cdot) \leftarrow 0$. Denote $\phi_{h,k} := \phi(s_h^k, a_h^k), \bar{\phi}_{h,k} := \phi(\bar{s}_h^k, \bar{a}_h^k)$
for $h = H, H-1, \ldots, 1$ **do**
  Set $\boldsymbol{u}_h \leftarrow \operatorname{argmin}_{\theta\in\Theta} \left\{ \sum_{k=1}^K \left[f\left(\theta, \bar{\phi}_{h,k}\right) - \widehat{V}_{h+1}(\bar{s}_{h+1}^k)\right]^2 + \lambda \cdot \|\theta\|_2^2 \right\}$
  Set $\boldsymbol{v}_h \leftarrow \operatorname{argmin}_{\theta\in\Theta} \left\{ \sum_{k=1}^K \left[f\left(\theta, \bar{\phi}_{h,k}\right) - \widehat{V}_{h+1}^2(\bar{s}_{h+1}^k)\right]^2 + \lambda \cdot \|\theta\|_2^2 \right\}$
  Set $\left[\widehat{\mathrm{Var}}_h \widehat{V}_{h+1}\right](\cdot,\cdot) = f(\boldsymbol{v}_h, \phi(\cdot,\cdot))_{[0,(H-h+1)^2]} - \left[f(\boldsymbol{u}_h, \phi(\cdot,\cdot))_{[0,H-h+1]}\right]^2$
  Set $\widehat{\sigma}_h(\cdot,\cdot)^2 \leftarrow \max\{1, \widehat{\mathrm{Var}}_{P_h} \widehat{V}_{h+1}(\cdot,\cdot)\}$
  Set $\widehat{\theta}_h \leftarrow \operatorname{argmin}_{\theta\in\Theta} \left\{ \sum_{k=1}^K \left[f\left(\theta, \phi_{h,k}\right) - r_{h,k} - \widehat{V}_{h+1}(s_{h+1}^k)\right]^2 / \widehat{\sigma}_h^2(s_h^k, a_h^k) + \lambda \cdot \|\theta\|_2^2 \right\}$
  Set $\Lambda_h \leftarrow \sum_{k=1}^K \nabla f(\widehat{\theta}_h, \phi_{h,k}) \nabla f(\widehat{\theta}_h, \phi_{h,k})^\top / \widehat{\sigma}^2(s_h^k, a_h^k) + \lambda \cdot I$,
  Set $\Gamma_h(\cdot,\cdot) \leftarrow \beta \sqrt{\nabla_\theta f(\widehat{\theta}_h, \phi(\cdot,\cdot))^\top \Lambda_h^{-1} \nabla_\theta f(\widehat{\theta}_h, \phi(\cdot,\cdot))} \left(+\widetilde{O}(\frac{1}{K})\right)$
  Set $\bar{Q}_h(\cdot,\cdot) \leftarrow f(\widehat{\theta}_h, \phi(\cdot,\cdot)) - \Gamma_h(\cdot,\cdot)$, $\widehat{Q}_h(\cdot,\cdot) \leftarrow \min\left\{\bar{Q}_h(\cdot,\cdot), H-h+1\right\}^+$
  Set $\widehat{\pi}_h(\cdot \mid \cdot) \leftarrow \arg\max_{\pi_h} \left\langle \widehat{Q}_h(\cdot,\cdot), \pi_h(\cdot \mid \cdot)\right\rangle_{\mathcal{A}}$, $\widehat{V}_h(\cdot) \leftarrow \max_{\pi_h} \left\langle \widehat{Q}_h(\cdot,\cdot), \pi_h(\cdot \mid \cdot)\right\rangle_{\mathcal{A}}$
end for
Output: $\left\{\widehat{\pi}_h\right\}_{h=1}^H$.

C.9.1 Provable Efficiency for Variance-Aware Fitted Q Learning

Recall the objective

$$\ell_h(\theta) := \frac{1}{K} \sum_{k=1}^K \left[f\left(\theta, \phi(s_h^k, a_h^k)\right) - r(s_h^k, a_h^k) - \widehat{V}_{h+1}(s_{h+1}^k)\right]^2 / \widehat{\sigma}_h^2(s_h^k, a_h^k) + \frac{\lambda}{K} \cdot \|\theta\|_2^2$$

Then by definition, $\widehat{\theta}_h := \operatorname{argmin}_{\theta\in\Theta} \ell_h(\theta)$ and $\theta_{\mathbb{T}\widehat{V}_{h+1}}$ satisfies $f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi) = \mathcal{P}_h\widehat{V}_{h+1}(s^k_{h+1})$ (recall $\epsilon_{\mathcal{F}} = 0$). Therefore, in this case, we have the following lemma:

Lemma C.9.1. *Fix $h \in [H]$. With probability $1-\delta$,*

$$\mathbb{E}_\mu[\ell_h(\widehat{\theta}_h)] - \mathbb{E}_\mu[\ell_h(\theta_{\mathbb{T}\widehat{V}_{h+1}})] \le \frac{36H^2(\log(1/\delta) + C_{d,\log K}) + \lambda C_\Theta^2}{K}$$

where the expectation over μ is taken w.r.t. $(s^k_h, a^k_h, s^k_{h+1})$ $k = 1, ..., K$ only (i.e., first compute $\mathbb{E}_\mu[\ell_h(\theta)]$ for a fixed θ, then plug-in either $\widehat{\theta}_{h+1}$ or $\theta_{\mathbb{T}\widehat{V}_{h+1}}$). Here $C_{d,\log(K)} := d\log(1 + 24C_\Theta(H+1)\kappa_1 K) + d\log\left(1 + 288H^2C_\Theta(\kappa_1\sqrt{C_\Theta} + 2\sqrt{\kappa_1\kappa_2/\lambda})^2K^2\right) + d^2\log\left(1 + 288H^2\sqrt{d}\kappa_1^2K^2/\lambda\right) + d\log(1 + 16C_\Theta H^2\kappa_1 K) + d\log(1 + 32C_\Theta H^3\kappa_1 K)$.

Proof: [Proof of Lemma C.9.1]

Step1: Consider the case where $\lambda = 0$. Indeed, fix $h \in [H]$ and any function $V(\cdot) \in \mathbb{R}^S$. Similarly, define $f_V(s,a) := f(\theta_{\mathbb{T}V}, \phi) = \mathcal{P}_h V$. For any fixed $\theta \in \Theta$, denote $g(s,a) = f(\theta, \phi(s,a))$. Moreover, for any $u, v \in \Theta$, define

$$\sigma^2_{u,v}(\cdot,\cdot) := \max\{1, f(v, \phi(\cdot,\cdot))_{[0,(H-h+1)^2]} - \left[f(u, \phi(\cdot,\cdot))_{[0,H-h+1]}\right]^2\}$$

Then define (we omit the subscript u, v of $\sigma^2_{u,v}$ for the illustration purpose when there is no ambiguity)

$$X(g, V, f_V, \sigma^2) := \frac{(g(s,a) - r - V(s'))^2 - (f_V(s,a) - r - V(s'))^2}{\sigma^2_{u,v}(s,a)}.$$

Since all episodes are independent of each other, $X_k(g, V, f_V) := X(g(s^k_h, a^k_h), V(s^k_{h+1}), f_V(s^k_h, a^k_h), \sigma^2(s^k_h, a^k_h))$ are independent r.v.s and it holds

$$\frac{1}{K}\sum_{k=1}^K X_k(g, V, f_V, \sigma^2) = \ell(g) - \ell(f_V). \tag{C.23}$$

Next, the variance of X is bounded by

$$
\begin{aligned}
&\mathrm{Var}[X(g,V,f_V,\sigma^2)] \le \mathbb{E}_\mu[X(g,f,f_V,\sigma^2)^2] \\
=&\mathbb{E}_\mu\left[\left(\left(g(s_h,a_h)-r_h-V(s_{h+1})\right)^2-\left(f_V(s_h,a_h)-r_h-V(s_{h+1})\right)^2\right)^2/\sigma^2(s_h,a_h)^2\right] \\
=&\mathbb{E}_\mu\left[\frac{(g(s_h,a_h)-f_V(s_h,a_h))^2}{\sigma^2(s_h,a_h)}\cdot\frac{(g(s_h,a_h)+f_V(s_h,a_h)-2r_h-2V(s_{h+1}))^2}{\sigma^2(s_h,a_h)}\right] \\
\le&4H^2\cdot\mathbb{E}_\mu[\frac{(g(s_h,a_h)-f_V(s_h,a_h))^2}{\sigma^2(s_h,a_h)}] \\
=&4H^2\cdot\mathbb{E}_\mu\left[\frac{\left(g(s_h,a_h)-r_h-V(s_{h+1})\right)^2-\left(f_V(s_h,a_h)-r_h-V(s_{h+1})\right)^2}{\sigma^2(s_h,a_h)}\right] \quad (*) \\
=&4H^2\cdot\mathbb{E}_\mu[X(g,f,f_V,\sigma^2)]
\end{aligned}
$$

$(*)$ follows from that

$$
\mathbb{E}_\mu\left[\frac{f(\widehat{\theta}_h,\phi(s_h,a_h))-f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(s_h,a_h))}{\sigma^2(s_h,a_h)}\cdot\mathbb{E}\left(f\left(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(s_h,a_h)\right)-r_h-\widehat{V}_{h+1}(s_{h+1})\middle|s_h,a_h\right)\right]=0.
$$

Therefore, by Bernstein inequality, with probability $1-\delta$,

$$
\begin{aligned}
&\mathbb{E}_\mu[X(g,f,f_V,\sigma^2)]-\frac{1}{K}\sum_{k=1}^K X_k(g,f,f_V,\sigma^2) \\
\le&\sqrt{\frac{2\mathrm{Var}[X(g,f,f_V,\sigma^2)]\log(1/\delta)}{K}}+\frac{4H^2\log(1/\delta)}{3K} \\
\le&\sqrt{\frac{8H^2\mathbb{E}_\mu[X(g,f,f_V,\sigma^2)]\log(1/\delta)}{K}}+\frac{4H^2\log(1/\delta)}{3K}.
\end{aligned}
$$

Now, if we choose $g(s,a) := f(\widehat{\theta}_h,\phi(s,a))$ and $u=\boldsymbol{u}_h, v=\boldsymbol{v}_h$ from Algorithm 6, then $\widehat{\theta}_h$

minimizes $\ell_h(\theta)$, therefore, it also minimizes $\frac{1}{K}\sum_{k=1}^{K} X_i(\theta, \widehat{V}_{h+1}, f_{\widehat{V}_{h+1}}, \widehat{\sigma}_h^2)$ and this implies

$$\frac{1}{K}\sum_{k=1}^{K} X_k(\widehat{\theta}_h, \widehat{V}_{h+1}, f_{\widehat{V}_{h+1}}, \widehat{\sigma}_h^2) \leq \frac{1}{K}\sum_{k=1}^{K} X_k(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \widehat{V}_{h+1}, f_{\widehat{V}_{h+1}}, \widehat{\sigma}_h^2) = 0.$$

Thus, we obtain

$$\mathbb{E}_\mu[X(\widehat{\theta}_h, \widehat{V}_{h+1}, f_{\widehat{V}_{h+1}}, \widehat{\sigma}_h^2)] \leq \sqrt{\frac{8H^2 \cdot \mathbb{E}_\mu[X(\widehat{\theta}_h, \widehat{V}_{h+1}, f_{\widehat{V}_{h+1}}, \widehat{\sigma}_h^2)]\log(1/\delta)}{K}} + \frac{4H^2\log(1/\delta)}{3K}.$$

However, the above does not hold with probability $1-\delta$ since $\widehat{\theta}_h, \widehat{\sigma}_h^2$ and $\widehat{V}_{h+1} := \min\{\max_a f(\widehat{\theta}_{h+1}, \phi(\cdot, a)) - \sqrt{\nabla f(\widehat{\theta}_{h+1}, \phi(\cdot, a))^\top A \cdot \nabla f(\theta, \phi(\cdot, a))}, H\}$ (where A is certain symmetric matrix with bounded norm) depend on $\widehat{\theta}_h, \widehat{\theta}_{h+1}$ which are data-dependent. Therefore, we need to further apply covering Lemma C.11.11 and choose $\epsilon = O(1/K)$ and a union bound to obtain with probability $1-\delta$,

$$\mathbb{E}_\mu[X(\widehat{\theta}_h, \widehat{V}_{h+1}, f_{\widehat{V}_{h+1}}, \widehat{\sigma}_h^2)] \leq \sqrt{\frac{8H^2 \cdot \mathbb{E}_\mu[X(\widehat{\theta}_h, \widehat{V}_{h+1}, f_{\widehat{V}_{h+1}}, \widehat{\sigma}_h^2)](\log(1/\delta) + C_{d,\log K})}{K}} + \frac{4H^2(\log(1/\delta) + C_{d,\log K})}{3K}.$$

where $C_{d,\log(K)} := d\log(1+24C_\Theta(H+1)\kappa_1 K) + d\log\left(1 + 288H^2C_\Theta(\kappa_1\sqrt{C_\Theta} + 2\sqrt{\kappa_1\kappa_2/\lambda})^2K^2\right) + d^2\log\left(1 + 288H^2\sqrt{d}\kappa_1^2K^2/\lambda\right) + d\log(1+16C_\Theta H^2\kappa_1 K) + d\log(1+32C_\Theta H^3\kappa_1 K)$ (where we let $B = 1/\lambda$ since $\left\|\Lambda_h^{-1}\right\|_2 \leq 1/\lambda$). Solving this quadratic equation to obtain with probability $1-\delta$,

$$\mathbb{E}_\mu[X(\widehat{\theta}_h, \widehat{V}_{h+1}, f_{\widehat{V}_{h+1}})] \leq \frac{36H^2(\log(1/\delta) + C_{d,\log K})}{K}.$$

Now according to (C.23), by definition we finally have with probability $1-\delta$ (recall the expectation over μ is taken w.r.t. $(s_h^k, a_h^k, s_{h+1}^k)$ $k = 1, ..., K$ only)

$$\mathbb{E}_\mu[\ell_h(\widehat{\theta}_{h+1})] - \mathbb{E}_\mu[\ell_h(\theta_{\mathbb{T}\widehat{V}_{h+1}})] = \mathbb{E}_\mu[X(\widehat{\theta}_h, \widehat{V}_{h+1}, f_{\widehat{V}_{h+1}})] \leq \frac{36H^2(\log(1/\delta) + C_{d,\log K})}{K} \quad \text{(C.24)}$$

where we used $f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi) = \mathcal{P}_h \widehat{V}_{h+1} = f_{\widehat{V}_{h+1}}$.

Step2. If $\lambda > 0$, there is only extra term $\frac{\lambda}{K}\left(\left\|\widehat{\theta}_h\right\|_2 - \left\|\theta_{\mathbb{T}\widehat{V}_{h+1}}\right\|_2\right) \leq \frac{\lambda}{K}\left\|\widehat{\theta}_h\right\|_2 \leq \frac{\lambda C_\Theta^2}{K}$ in addition to above. This finishes the proof.

Theorem C.9.1 (Provable efficiency for VAFQL)**.** *Let $C_{d,\log K}$ be the same as Lemma C.9.1. Then, with probability* $1-\delta$

$$\left\|\widehat{\theta}_h - \theta_{\mathbb{T}\widehat{V}_{h+1}}\right\|_2 \leq \sqrt{\frac{36H^4(\log(H/\delta) + C_{d,\log K}) + 2\lambda C_\Theta^2}{\kappa K}},\ \forall h \in [H].$$

Proof: [Proof of Theorem C.9.1] Apply a union bound in Lemma C.9.1, we have with probability $1-\delta$,

$$\begin{aligned}
&\mathbb{E}_\mu[\ell_h(\widehat{\theta}_h)] - \mathbb{E}_\mu[\ell_h(\theta_{\mathbb{T}\widehat{V}_{h+1}})] \leq \frac{36H^2(\log(H/\delta) + C_{d,\log K}) + \lambda C_\Theta^2}{K},\ \ \forall h \in [H] \\
\Rightarrow &\mathbb{E}_\mu[\ell_h(\widehat{\theta}_h) - \frac{\lambda}{K}\left\|\widehat{\theta}_h\right\|_2^2] - \mathbb{E}_\mu[\ell_h(\theta_{\mathbb{T}\widehat{V}_{h+1}}) - \frac{\lambda}{K}\left\|\theta_{\mathbb{T}\widehat{V}_{h+1}}\right\|_2^2] \leq \frac{36H^2(\log(H/\delta) + C_{d,\log K}) + 2\lambda C_\Theta^2}{K}
\end{aligned} \tag{C.25}$$

Now we prove for all $h \in [H]$,

$$\mathbb{E}_\mu\left[\left(f(\widehat{\theta}_h, \phi(\cdot,\cdot)) - f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi(\cdot,\cdot))\right)^2\right] = \mathbb{E}_\mu\left[\ell_h(\widehat{\theta}_h) - \frac{\lambda\left\|\widehat{\theta}_h\right\|_2^2}{K}\right] - \mathbb{E}_\mu\left[\ell_h(\theta_{\mathbb{T}\widehat{V}_{h+1}}) - \frac{\lambda\left\|\theta_{\mathbb{T}\widehat{V}_{h+1}}\right\|_2^2}{K}\right]. \tag{C.26}$$

Indeed, identical to (C.24),

$$
\begin{aligned}
&\mathbb{E}_\mu\left[\ell_h(\widehat{\theta}_h)-\frac{\lambda\left\|\widehat{\theta}_h\right\|_2^2}{K}\right]-\mathbb{E}_\mu\left[\ell_h(\theta_{\mathbb{T}\widehat{V}_{h+1}})-\frac{\lambda\left\|\theta_{\mathbb{T}\widehat{V}_{h+1}}\right\|_2^2}{K}\right]=\mathbb{E}_\mu[X(\widehat{\theta}_h,\widehat{V}_{h+1},f_{\widehat{V}_{h+1}})]\\
=&\mathbb{E}_\mu\left(\left[f\left(\widehat{\theta}_h,\phi(s_h,a_h)\right)-r_h-\widehat{V}_{h+1}(s_{h+1})\right]^2/\widehat{\sigma}_h^2(s_h,a_h)-\left[f\left(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(s_h,a_h)\right)-r_h-\widehat{V}_{h+1}(s_{h+1})\right]^2/\widehat{\sigma}_h^2(s_h,a_h)\right)\\
=&\mathbb{E}_\mu\left[\left(f(\widehat{\theta}_h,\phi(\cdot,\cdot))-f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(\cdot,\cdot))\right)^2/\widehat{\sigma}_h^2(\cdot,\cdot)\right]\\
+&\mathbb{E}_\mu\left[\left(f(\widehat{\theta}_h,\phi(s_h,a_h))-f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(s_h,a_h))\right)\cdot\left(f\left(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(s_h,a_h)\right)-r_h-\widehat{V}_{h+1}(s_{h+1})\right)/\widehat{\sigma}_h^2(s_h,a_h)\right]\\
=&\mathbb{E}_\mu\left[\left(f(\widehat{\theta}_h,\phi(\cdot,\cdot))-f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(\cdot,\cdot))\right)^2/\widehat{\sigma}_h^2(\cdot,\cdot)\right]\\
+&\mathbb{E}_\mu\left[\left(f(\widehat{\theta}_h,\phi(s_h,a_h))-f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(s_h,a_h))\right)\cdot\mathbb{E}\left(f\left(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(s_h,a_h)\right)-r_h-\widehat{V}_{h+1}(s_{h+1})\middle|s_h,a_h\right)/\widehat{\sigma}_h^2(s_h,a_h)\right]\\
=&\mathbb{E}_\mu\left[\left(f(\widehat{\theta}_h,\phi(\cdot,\cdot))-f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(\cdot,\cdot))\right)^2/\widehat{\sigma}_h^2(\cdot,\cdot)\right]
\end{aligned}
$$

where the third identity uses law of total expectation and that μ is taken w.r.t. s_h, a_h, s_{h+1} only (recall Lemma C.9.1) so the $\widehat{\sigma}_h^2$ can be move outside of the conditional expectation.[4] The fourth identity uses the definition of $\theta_{\mathbb{T}\widehat{V}_{h+1}}$ since $f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(s,a))=\mathcal{P}_{h,s,a}\widehat{V}_{h+1}$.

Then we have

$$
\begin{aligned}
&\mathbb{E}_\mu\left[\left(f(\widehat{\theta}_h,\phi(\cdot,\cdot))-f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(\cdot,\cdot))\right)^2/\widehat{\sigma}_h^2(\cdot,\cdot)\right]\\
\geq&\mathbb{E}_\mu\left[\left(f(\widehat{\theta}_h,\phi(\cdot,\cdot))-f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(\cdot,\cdot))\right)^2\right]/H^2\geq\frac{\kappa}{H^2}\left\|\widehat{\theta}_h-\theta_{\mathbb{T}\widehat{V}_{h+1}}\right\|_2^2,
\end{aligned}
$$

where the third identity uses μ is over s_h, a_h only and the last one uses $\widehat{\sigma}_h^2(\cdot,\cdot)\leq H^2$. Combine the above with (C.25) and (C.26), we obtain the stated result.

Theorem C.9.2 (Provable efficiency of VAFQL (Part II))**.** *Let $C_{d,\log K}$ be the same as Lemma C.9.1. Furthermore, suppose $\lambda\leq 1/2C_\Theta^2$ and $K\geq\max\left\{512\frac{\kappa_1^4}{\kappa^2}\left(\log(\frac{2d}{\delta})+d\log(1+\frac{4\kappa_1^3\kappa_2 C_\Theta K^3}{\lambda^2})\right),\frac{4\lambda}{\kappa}\right\}$.*

[4] Recall $\widehat{\sigma}_h^2$ computed in Algorithm 6 uses an independent copy $\mathcal{D}'$.

Then, with probability $1-\delta$, $\forall h \in [H]$

$$\sup_{s,a}\left|f(\widehat{\theta}_h,\phi(s,a))-f(\theta_h^\star,\phi(s,a))\right| \le \left(\kappa_1 H\sqrt{\frac{36H^4(\log(H/\delta)+C_{d,\log K})+2\lambda C_\Theta^2}{\kappa}}+\frac{2dH^3\kappa_1}{\sqrt{\kappa}}\right)\sqrt{\frac{1}{K}}+O(\frac{1}{K}),$$

Furthermore, we have with probability $1-\delta$,

$$\begin{aligned}\sup_h\left\|\widehat{V}_h-V_h^\star\right\|_\infty \le& \left(\kappa_1 H\sqrt{\frac{36H^4(\log(H/\delta)+C_{d,\log K})+2\lambda C_\Theta^2}{\kappa}}+\frac{2dH^3\kappa_1}{\sqrt{\kappa}}\right)\sqrt{\frac{1}{K}}+O(\frac{1}{K})\\ =&\widetilde{O}\left(\kappa_1 H^3\sqrt{\frac{d^2}{\kappa}}\sqrt{\frac{1}{K}}\right)\end{aligned}$$

where $\widetilde{O}$ *absorbs Polylog terms and higher order terms. Lastly, it also holds for all* $h \in [H]$, *w.p.* $1-\delta$

$$\begin{aligned}\left\|\widehat{\theta}_h-\theta_h^\star\right\|_2 \le& \left(\kappa_1 H\frac{\sqrt{72H^4(\log(H^2/\delta)+C_{d,\log K})+4\lambda C_\Theta^2}}{\kappa}+\frac{4H^3d\kappa_1}{\kappa}\right)\sqrt{\frac{1}{K}}+O(\frac{1}{K})\\ =&\widetilde{O}\left(\frac{\kappa_1 H^3 d}{\kappa}\sqrt{\frac{1}{K}}\right)\end{aligned}$$

Proof: [Proof of Theorem C.9.2] **Step1:** we show the first result.

We prove this by backward induction. When $h = H+1$, by convention $f(\widehat{\theta}_h,\phi(s,a)) = f(\theta_h^\star,\phi(s,a)) = 0$ so the base case holds. Suppose for $h+1$, with probability $1-(H-h)\delta$, $\sup_{s,a}\left|f(\widehat{\theta}_h,\phi(s,a))-f(\theta_h^\star,\phi(s,a))\right| \le C_{h+1}\sqrt{\frac{1}{K}}$, we next consider the case for $t = h$.

On one hand, by Theorem C.9.1, we have with probability $1-\delta/2$,

$$
\begin{aligned}
&\sup_{s,a}\left|f(\widehat{\theta}_h,\phi(s,a))-f(\theta_h^\star,\phi(s,a))\right| \\
\leq&\sup_{s,a}\left|f(\widehat{\theta}_h,\phi(s,a))-f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(s,a))\right|+\sup_{s,a}\left|f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(s,a))-f(\theta_h^\star,\phi(s,a))\right| \\
=&\sup_{s,a}\left|\nabla f(\xi,\phi(s,a))^\top(\widehat{\theta}_h-\theta_{\mathbb{T}\widehat{V}_{h+1}})\right|+\sup_{s,a}\left|f(\theta_{\mathbb{T}\widehat{V}_{h+1}},\phi(s,a))-f(\theta_{\mathbb{T}V_{h+1}^\star},\phi(s,a))\right| \\
\leq&\kappa_1\cdot\left\|\widehat{\theta}_h-\theta_{\mathbb{T}\widehat{V}_{h+1}}\right\|_2+\sup_{s,a}\left|\mathcal{P}_{h,s,a}\widehat{V}_{h+1}-\mathcal{P}_{h,s,a}V_{h+1}^\star\right| \\
\leq&\kappa_1\sqrt{\frac{36H^4(\log(H/\delta)+C_{d,\log K})+2\lambda C_\Theta^2}{\kappa K}}+\left\|\widehat{V}_{h+1}-V_{h+1}^\star\right\|_\infty,
\end{aligned}
$$

Recall we have the form $\widehat{V}_{h+1}(\cdot) := \min\{\max_a f(\widehat{\theta}_{h+1},\phi(\cdot,a))-\Gamma_h(\cdot,a),H\}$ and $V_{h+1}^\star(\cdot)=\max_a f(\theta_{h+1}^\star,\phi(\cdot,a))=\min\{\max_a f(\theta_{h+1}^\star,\phi(\cdot,a)),H\}$, we obtain

$$
\left\|\widehat{V}_{h+1}-V_{h+1}^\star\right\|_\infty\leq\sup_{s,a}\left|f(\widehat{\theta}_{h+1},\phi(s,a))-f(\theta_{h+1}^\star,\phi(s,a))\right|+\sup_{h,s,a}\Gamma_h(s,a) \tag{C.27}
$$

Note the above holds true for any generic $\Gamma_h(s,a)$. In particular, according to Algorithm 6, we specify

$$
\Gamma_h(\cdot,\cdot)=d\sqrt{\nabla_\theta f(\widehat{\theta}_h,\phi(\cdot,\cdot))^\top\Lambda_h^{-1}\nabla_\theta f(\widehat{\theta}_h,\phi(\cdot,\cdot))}\left(+\widetilde{O}(\frac{1}{K})\right)
$$

and by Lemma C.11.5, with probability $1-\delta$ (note here Σ_h^{-1} is replaced by Λ_h^{-1} and $\left\|\Lambda_h^{-1}\right\|_2\leq H^2/\kappa$),

$$
\Gamma_h\leq\frac{2dH^2\kappa_1}{\sqrt{\kappa K}}+O(\frac{1}{K})
$$

and by a union bound this implies with probability $1-(H-h+1)\delta$,

$$\begin{aligned}&\sup_{s,a}\left|f(\widehat{\theta}_h,\phi(s,a))-f(\theta_h^\star,\phi(s,a))\right|\\ \leq& C_{h+1}\sqrt{\frac{1}{K}}+\kappa_1\sqrt{\frac{36H^4(\log(H/\delta)+C_{d,\log K})+2\lambda C_\Theta^2}{\kappa K}}+\frac{2dH^2\kappa_1}{\sqrt{\kappa K}}+O(\frac{1}{K}):=C_h\sqrt{\frac{1}{K}}.\end{aligned}$$

Solving for C_h, we obtain $C_h \leq \kappa_1 H\sqrt{\frac{36H^4(\log(H/\delta)+C_{d,\log K})+2\lambda C_\Theta^2}{\kappa}}+H\frac{2dH^2\kappa_1}{\sqrt{\kappa}}$ for all H. By a union bound (replacing δ by δ/H), we obtain the stated result.

Step2: Utilizing the intermediate result (C.27), we directly have with probability $1-\delta$,

$$\sup_h\left\|\widehat{V}_h-V_h^\star\right\|_\infty\leq\sup_{s,a}\left|f(\widehat{\theta}_h,\phi(s,a))-f(\theta_h^\star,\phi(s,a))\right|+\frac{2dH^2\kappa_1}{\sqrt{\kappa K}}+O(\frac{1}{K}),$$

where $\sup_{s,a}\left|f(\widehat{\theta}_h,\phi(s,a))-f(\theta_h^\star,\phi(s,a))\right|$ can be bounded using Step1.

Step3: Denote $M:=\left(\kappa_1 H\sqrt{\frac{36H^4(\log(H^2/\delta)+C_{d,\log K})+2\lambda C_\Theta^2}{\kappa}}+\frac{2H^3d\kappa_1}{\sqrt{\kappa}}\right)\sqrt{\frac{1}{K}}+O(\frac{1}{K})$, then by Step1 we have with probability $1-\delta$ (here ξ is some point between $\widehat{\theta}_h$ and $\theta_h^\star$) for all $h\in[H]$

$$\begin{aligned}M^2&\geq\sup_{s,a}\left|f(\widehat{\theta}_h,\phi(s,a))-f(\theta_h^\star,\phi(s,a))\right|^2\\ &\geq\mathbb{E}_\mu[\Big(f(\widehat{\theta}_h,\phi(s,a))-f(\theta_h^\star,\phi(s,a))\Big)^2]\geq\kappa\left\|\widehat{\theta}_h-\theta_h^\star\right\|_2^2\end{aligned}$$

where the last step is by Assumption 4.2.3. Solving this to obtain the stated result.

C.9.2 Bounding $|\widehat{\sigma}_h^2 - \sigma_h^{\star 2}|$

Recall the definition $\sigma_h^{\star 2}(\cdot,\cdot) = \max\{1, [\mathrm{Var}_{P_h} V_{h+1}^{\star}](\cdot,\cdot)\}$. In this section, we bound the term $|\widehat{\sigma}_h^2 - \sigma_h^{\star 2}| := \left\|\widehat{\sigma}_h^2(\cdot,\cdot) - \sigma_h^{\star 2}(\cdot,\cdot)\right\|_\infty$ and

$$\begin{aligned} \boldsymbol{u}_h &= \operatorname*{argmin}_{\theta\in\Theta} \left\{ \frac{1}{K}\sum_{k=1}^{K} \left[f\left(\theta, \bar{\phi}_{h,k}\right) - \widehat{V}_{h+1}(\bar{s}_{h+1}^k) \right]^2 + \frac{\lambda}{K}\cdot\|\theta\|_2^2 \right\} \\ \boldsymbol{v}_h &= \operatorname*{argmin}_{\theta\in\Theta} \left\{ \frac{1}{K}\sum_{k=1}^{K} \left[f\left(\theta, \bar{\phi}_{h,k}\right) - \widehat{V}_{h+1}^2(\bar{s}_{h+1}^k) \right]^2 + \frac{\lambda}{K}\cdot\|\theta\|_2^2 \right\} \end{aligned} \tag{C.28}$$

where

$$\widehat{\sigma}_h^2(\cdot,\cdot) := \max\{1, f(\boldsymbol{v}_h, \phi(\cdot,\cdot))_{[0,(H-h+1)^2]} - \left[f(\boldsymbol{u}_h, \phi(\cdot,\cdot))_{[0,H-h+1]}\right]^2\}$$

and true parameters $\boldsymbol{u}_h^\star, \boldsymbol{v}_h^\star$ satisfy $f(\boldsymbol{u}_h^\star, \phi(\cdot,\cdot)) = \mathbb{E}_{P(s'|\cdot,\cdot)}[V_h^\star(s')]$, $f(\boldsymbol{v}_h^\star, \phi) = \mathbb{E}_{P(s'|\cdot,\cdot)}[V_h^{\star 2}(s')]$. Furthermore, we define

$$\sigma^2_{\widehat{V}_{h+1}}(\cdot,\cdot) := \max\{1, [\mathrm{Var}_{P_h} \widehat{V}_{h+1}](\cdot,\cdot)\}$$

and the *parameter Expectation operator* $\mathbb{J} : V \in \mathbb{R}^S \to \theta_{\mathbb{J}V} \in \Theta$ such that:

$$f(\theta_{\mathbb{J}V}, \phi) = \mathbb{E}_{P_h}[V(s')], \quad \forall \|V\|_2 \le \mathcal{B}_{\mathcal{F}}.$$

Note $\theta_{\mathbb{J}V} \in \Theta$ by Bellman completeness, reward r is constant and differentiability (Definition 4.1.1) is an additive closed property. By definition,

$$\begin{aligned} |\widehat{\sigma}_h^2 - \sigma^2_{\widehat{V}_{h+1}}| \le& |f(\boldsymbol{v}_h, \phi) - f(\theta_{\mathbb{J}\widehat{V}^2_{h+1}}, \phi)| + |f(\boldsymbol{u}_h, \phi)^2 - f(\theta_{\mathbb{J}\widehat{V}_{h+1}}, \phi)^2| \\ \le& |f(\boldsymbol{v}_h, \phi) - f(\theta_{\mathbb{J}\widehat{V}^2_{h+1}}, \phi)| + 2H\cdot|f(\boldsymbol{u}_h, \phi) - f(\theta_{\mathbb{J}\widehat{V}_{h+1}}, \phi)| \end{aligned}$$

and

$$\begin{aligned}|\sigma_h^{\star 2} - \widehat{\sigma}_h^2| \leq & |f(\boldsymbol{v}_h^\star, \phi) - f(\boldsymbol{v}_h, \phi)| + |f(\boldsymbol{u}_h^\star, \phi)^2 - f(\boldsymbol{v}_h, \phi)^2| \\ \leq & |f(\boldsymbol{v}_h^\star, \phi) - f(\boldsymbol{v}_h, \phi)| + 2H \cdot |f(\boldsymbol{u}_h^\star, \phi) - f(\boldsymbol{v}_h, \phi)|\end{aligned}$$

We first give the following result.

Lemma C.9.2. *Suppose* $\lambda \leq 1/2C_\Theta^2$ *and* $K \geq \max\left\{512\frac{\kappa_1^4}{\kappa^2}\left(\log(\frac{2d}{\delta}) + d\log(1 + \frac{4\kappa_1^3\kappa_2 C_\Theta K^3}{\lambda^2})\right), \frac{4\lambda}{\kappa}\right\}$. *Then, with probability* $1-\delta$, $\forall h \in [H]$,

$$\begin{aligned}\left\|\boldsymbol{u}_h - \theta_{\mathbb{J}\widehat{V}_{h+1}}\right\|_2 \leq & \sqrt{\frac{36H^2(\log(H/\delta) + \widetilde{O}(d^2)) + 2\lambda C_\Theta^2}{\kappa K}}, \ \forall h \in [H], \\ \left\|\boldsymbol{v}_h - \theta_{\mathbb{J}\widehat{V}_{h+1}^2}\right\|_2 \leq & \sqrt{\frac{36H^4(\log(H/\delta) + \widetilde{O}(d^2)) + 2\lambda C_\Theta^2}{\kappa K}}, \ \forall h \in [H].\end{aligned}$$

and

$$\begin{aligned}\sup_{s,a}\left|f(\boldsymbol{u}_h, \phi(s,a)) - f(\boldsymbol{u}_h^\star, \phi(s,a))\right| \leq & \left(\kappa_1 H\sqrt{\frac{36H^2(\log(H^2/\delta) + \widetilde{O}(d^2)) + 2\lambda C_\Theta^2}{\kappa}} + \frac{2H^2 d\kappa_1}{\sqrt{\kappa}}\right)\sqrt{\frac{1}{K}} + O(\frac{1}{K}), \\ \sup_{s,a}\left|f(\boldsymbol{v}_h, \phi(s,a)) - f(\boldsymbol{v}_h^\star, \phi(s,a))\right| \leq & \left(\kappa_1 H\sqrt{\frac{36H^4(\log(H^2/\delta) + \widetilde{O}(d^2)) + 2\lambda C_\Theta^2}{\kappa}} + \frac{2H^3 d\kappa_1}{\sqrt{\kappa}}\right)\sqrt{\frac{1}{K}} + O(\frac{1}{K}).\end{aligned}$$

The above directly implies for all $h \in [H]$, *with probability* $1-\delta$,

$$\begin{aligned}|\sigma_h^{\star 2} - \widehat{\sigma}_h^2| \leq & \left(3\kappa_1 H^2\sqrt{\frac{36H^4(\log(H^2/\delta) + \widetilde{O}(d^2)) + 2\lambda C_\Theta^2}{\kappa}} + \frac{6H^4 d\kappa_1}{\sqrt{\kappa}}\right)\sqrt{\frac{1}{K}} + O(\frac{1}{K}) \\ |\widehat{\sigma}_h^2 - \sigma_{\widehat{V}_{h+1}}^2| \leq & 3H\kappa_1\sqrt{\frac{36H^4(\log(H/\delta) + \widetilde{O}(d^2)) + 2\lambda C_\Theta^2}{\kappa K}}.\end{aligned}$$

Proof: [Proof of Lemma C.9.2] In fact, the proof follows a reduction from the provable

efficiency procedure conducted in Section C.6. This is due to the regression procedure in (C.28) is the same as the procedure (C.15) except the *parameter Bellman operator* $\mathbb{T}$ is replaced by the *parameter Expectation operator* $\mathbb{J}$ (recall here $\bar{\phi}_{h,k}$ uses the independent copy $\mathcal{D}'$ and $\widetilde{O}(d^2)$ comes from the covering argument.). Concretely, the $X(g, V, f_V)$ used in Lemma C.6.1 will be modified to $X(g, V, f_V) = (g(s,a) - V(s'))^2 - (f(\theta_{\mathbb{J}V}, \phi(s,a)) - V(s'))^2$ by removing reward information and the decomposition

$$\mathbb{E}\mu\left[(g(s_h,a_h) - V(s_{h+1}))^2 - (f(\theta_{\mathbb{J}V}, \phi(s_h,a_h)) - V(s_{h+1}))^2\right] = \mathbb{E}_\mu\left[(g(s_h,a_h) - f(\theta_{\mathbb{J}V}, \phi(s_h,a_h)))^2\right]$$

holds true. Then with probability $1-\delta$,

$$\begin{aligned}
|\sigma_h^{\star 2} - \widehat{\sigma}_h^2| \leq & |f(\boldsymbol{v}_h^\star, \phi) - f(\boldsymbol{v}_h, \phi)| + 2H \cdot |f(\boldsymbol{u}_h^\star, \phi) - f(\boldsymbol{v}_h, \phi)| \\
\leq & \left(3\kappa_1 H^2 \sqrt{\frac{36H^4(\log(H^2/\delta) + \widetilde{O}(d^2)) + 2\lambda C_\Theta^2}{\kappa}} + \frac{6H^4 d\kappa_1}{\sqrt{\kappa}}\right)\sqrt{\frac{1}{K}} + O(\frac{1}{K}).
\end{aligned}$$

and

$$\begin{aligned}
|\widehat{\sigma}_h^2 - \sigma^2_{\widehat{V}_{h+1}}| \leq & |f(\boldsymbol{v}_h, \phi) - f(\theta_{\mathbb{J}\widehat{V}^2_{h+1}}, \phi)| + 2H \cdot |f(\boldsymbol{u}_h, \phi) - f(\theta_{\mathbb{J}\widehat{V}_{h+1}}, \phi)| \\
\leq & \kappa_1 \left\|\boldsymbol{v}_h - \theta_{\mathbb{J}\widehat{V}^2_{h+1}}\right\|_2 + 2H\kappa_1 \left\|\boldsymbol{u}_h - \theta_{\mathbb{J}\widehat{V}_{h+1}}\right\|_2 \\
\leq & 3H\kappa_1 \sqrt{\frac{36H^4(\log(H/\delta) + \widetilde{O}(d^2)) + 2\lambda C_\Theta^2}{\kappa K}}.
\end{aligned}$$

C.9.3 Proof of Theorem 4.4.1

In this section, we sketch the proof of Theorem 4.4.1 since the most components are identical to Theorem 4.3.2. We will focus on highlighting the difference for obtaining the tighter bound.

First of all, Recall in the first-order condition, we have

$$\nabla_\theta \left\{ \sum_{k=1}^{K} \frac{\left[f\left(\theta, \phi_{h,k}\right) - r_{h,k} - \widehat{V}_{h+1}\left(s_{h+1}^k\right) \right]^2}{\widehat{\sigma}_h^2(s_h^k, a_h^k)} + \lambda \cdot \|\theta\|_2^2 \right\} \Bigg|_{\theta=\widehat{\theta}_h} = 0, \quad \forall h \in [H].$$

Therefore, if we define the quantity $Z_h(\cdot,\cdot) \in \mathbb{R}^d$ as

$$Z_h(\theta|V, \sigma^2) = \sum_{k=1}^{K} \frac{\left[f\left(\theta, \phi_{h,k}\right) - r_{h,k} - V\left(s_{h+1}^k\right) \right]}{\sigma(s_h^k, a_h^k)} \frac{\nabla f(\theta, \phi_{h,k})}{\sigma(s_h^k, a_h^k)} + \lambda \cdot \theta, \quad \forall \theta \in \Theta, \|V\|_2 \le H,$$

then we have

$$Z_h(\widehat{\theta}_h|\widehat{V}_{h+1}, \widehat{\sigma}_h^2) = 0.$$

According to the regression oracle (Line 8 of Algorithm 6), the estimated Bellman operator $\widehat{\mathcal{P}}_h$ maps $\widehat{V}_{h+1}$ to $\widehat{\theta}_h$, *i.e.* $\widehat{\mathcal{P}}_h \widehat{V}_{h+1} = f(\widehat{\theta}_h, \phi)$. Therefore (recall Definition C.3.1)

$$\begin{aligned}
&\mathcal{P}_h \widehat{V}_{h+1}(s,a) - \widehat{\mathcal{P}}_h \widehat{V}_{h+1}(s,a) = \mathcal{P}_h \widehat{V}_{h+1}(s,a) - f(\widehat{\theta}_h, \phi(s,a)) \\
=& f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi(s,a)) - f(\widehat{\theta}_h, \phi(s,a)) \\
=& \nabla f(\widehat{\theta}_h, \phi(s,a)) \left(\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h \right) + \mathrm{Hot}_{h,1},
\end{aligned} \tag{C.29}$$

where we apply the first-order Taylor expansion for the differentiable function f at point $\widehat{\theta}_h$ and $\mathrm{Hot}_{h,1}$ is a higher-order term. Indeed, the following Lemma C.4.1 bounds the $\mathrm{Hot}_{h,1}$ term with $\widetilde{O}(\frac{1}{K})$.

Lemma C.9.3. *Recall the definition (from the above decomposition)* $\mathrm{Hot}_{h,1} := f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi(s,a)) - f(\widehat{\theta}_h, \phi(s,a)) - \nabla f(\widehat{\theta}_h, \phi(s,a)) \left(\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h \right)$*, then with probability* $1-\delta$*,*

$$\left|\mathrm{Hot}_{h,1}\right| \le \widetilde{O}(\frac{1}{K}), \quad \forall h \in [H].$$

Proof: The proof is identical to that of Lemma C.4.1 but with the help of Lemma C.9.1. Next, according to the expansion of $Z_h(\theta|\widehat{V}_{h+1}, \widehat{\sigma}_h^2)$, we have

$$\nabla f(\widehat{\theta}_h, \phi(s,a)) \left(\theta_{\mathbb{T}\widehat{V}_{h+1}} - \widehat{\theta}_h\right) = I_1 + I_2 + I_3 + \mathrm{Hot}_2, \tag{C.30}$$

where

$$\begin{aligned}
\mathrm{Hot}_2 :=& \nabla f(\widehat{\theta}_h, \phi(s,a))\Lambda_h^{-1}\left[\widetilde{R}_K(\theta_{\mathbb{T}\widehat{V}_{h+1}}) + \lambda\theta_{\mathbb{T}\widehat{V}_{h+1}}\right] \\
\Delta_{\Lambda_h^s} =& \sum_{k=1}^{K} \frac{\left(f(\widehat{\theta}_h, \phi_{h,k}) - r_{h,k} - \widehat{V}_{h+1}(s_{h+1}^k)\right)\cdot \nabla_{\theta\theta}^2 f(\widehat{\theta}_h, \phi_{h,k})}{\widehat{\sigma}^2(s_h^k, a_h^k)} \\
\Lambda_h =& \sum_{k=1}^{K} \frac{\nabla_\theta f(\widehat{\theta}_h, \phi_{h,k})\nabla_\theta^\top f(\widehat{\theta}_{h,k}, \phi_{h,k})}{\widehat{\sigma}^2(s_h^k, a_h^k)} + \lambda I_d \\
\widetilde{R}_K(\theta_{\mathbb{T}\widehat{V}_{h+1}}) =& \Delta_{\Lambda_h^s}(\widehat{\theta}_h - \theta_{\mathbb{T}\widehat{V}_{h+1}}) + R_K(\theta_{\mathbb{T}\widehat{V}_{h+1}})
\end{aligned}$$

where $R_K(\theta_{\mathbb{T}\widehat{V}_{h+1}})$ is the second order residual that is bounded by $\widetilde{O}(1/K)$ and

$$\begin{aligned}
I_1 =& \nabla f(\widehat{\theta}_h, \phi(s,a))\Lambda_h^{-1} \sum_{k=1}^{K} \frac{\left(f(\theta_{\mathbb{T}V_{h+1}^\star}, \phi_{h,k}) - r_{h,k} - V_{h+1}^\star(s_{h+1}^k)\right)\cdot \nabla_\theta^\top f(\widehat{\theta}_h, \phi_{h,k})}{\widehat{\sigma}_h^2(s_h^k, a_h^k)} \\
I_2 =& \nabla f(\widehat{\theta}_h, \phi(s,a))\Lambda_h^{-1} \sum_{k=1}^{K} \frac{\left(f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi_{h,k}) - f(\theta_{\mathbb{T}V_{h+1}^\star}, \phi_{h,k}) - \widehat{V}_{h+1}(s_{h+1}^k) + V_{h+1}^\star(s_{h+1}^k)\right)\cdot \nabla_\theta^\top f(\widehat{\theta}_h, \phi_{h,k})}{\widehat{\sigma}_h^2(s_h^k, a_h^k)} \\
I_3 =& \nabla f(\widehat{\theta}_h, \phi(s,a))\Lambda_h^{-1} \sum_{k=1}^{K} \frac{\left(f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi_{h,k}) - r_{h,k} - \widehat{V}_{h+1}(s_{h+1}^k)\right)\cdot \left(\nabla_\theta^\top f(\theta_{\mathbb{T}\widehat{V}_{h+1}}, \phi_{h,k}) - \nabla_\theta^\top f(\widehat{\theta}_h, \phi_{h,k})\right)}{\widehat{\sigma}_h^2(s_h^k, a_h^k)}
\end{aligned}$$

Similar to the PFQL case, I_2, I_3, Hot_2 can be bounded to have order $O(1/K)$ via provably efficiency theorems in Section C.9.1 and in particular, the inclusion of $\sigma_{u,v}^2$ will not cause additional order in d.[5] Now we prove the result for the dominate term I_1.

[5]Note in Lemma C.11.11, we only have additive terms that has the same order has Lemma C.11.10.

Lemma C.9.4. *With probability* $1-\delta$,

$$|I_1| \le 4Hd \left\|\nabla f(\widehat{\theta}_h, \phi(s,a))\right\|_{\Sigma_h^{-1}} \cdot C_{\delta,\log K} + \widetilde{O}(\frac{\kappa_1}{\sqrt{\kappa}K}),$$

where $C_{\delta,\log K}$ *only contains Polylog terms.*

Proof: [Proof of Lemma C.9.4] First of all, by Cauchy–Schwarz inequality, we have

$$|I_1| \le \left\|\nabla f(\widehat{\theta}_h, \phi(s,a))\right\|_{\Lambda_h^{-1}} \cdot \left\|\sum_{k=1}^{K} \frac{\left(f(\theta_{\mathbb{T}V_{h+1}^\star}, \phi_{h,k}) - r_{h,k} - V_{h+1}^\star(s_{h+1}^k)\right) \cdot \nabla_\theta^\top f(\widehat{\theta}_h, \phi_{h,k})}{\widehat{\sigma}_h^2(s_h^k, a_h^k)}\right\|_{\Lambda_h^{-1}}. \tag{C.31}$$

Recall that $\sigma_{u,v}^2(\cdot,\cdot) := \max\{1, f(v,\phi(\cdot,\cdot))_{[0,(H-h+1)^2]} - \left[f(u,\phi(\cdot,\cdot))_{[0,H-h+1]}\right]^2\}$.

Step1. Let the fixed $\theta \in \Theta$ be arbitrary and fixed u, v such that $\sigma_{u,v}^2(\cdot,\cdot) \ge \frac{1}{2}\sigma_{\boldsymbol{u}_h^\star, \boldsymbol{v}_h^\star}^2(\cdot,\cdot) = \frac{1}{2}\sigma_h^{\star 2}(\cdot,\cdot)$ and define $x_k(\theta,u,v) = \nabla_\theta f(\theta,\phi_{h,k})/\sigma_{u,v}(s_h^k,a_h^k)$. Next, define $G_{u,v}(\theta) = \sum_{k=1}^K \nabla f(\theta,\phi(s_h^k,a_h^k))\cdot \nabla f(\theta,\phi(s_h^k,a_h^k))^\top/\sigma_{u,v}^2(s_h^k,a_h^k) + \lambda I_d$, then $\|x_k\|_2 \le \kappa_1$. Also denote $\eta_k := [f(\theta_{\mathbb{T}V_{h+1}^\star},\phi_{h,k}) - r_{h,k} - V_{h+1}^\star(s_{h+1}^k)]/\sigma_{u,v}(s_h^k,a_h^k)$, then $\mathbb{E}[\eta_k|s_h^k,a_h^k] = 0$ and

$$\begin{aligned}
\mathrm{Var}[\eta_k|s_h^k,a_h^k] =& \frac{\mathrm{Var}[f(\theta_{\mathbb{T}V_{h+1}^\star},\phi_{h,k}) - r_{h,k} - V_{h+1}^\star(s_{h+1}^k)|s_h^k,a_h^k]}{\sigma_{u,v}^2(s_h^k,a_h^k)} \\
\le& \frac{2\mathrm{Var}[f(\theta_{\mathbb{T}V_{h+1}^\star},\phi_{h,k}) - r_{h,k} - V_{h+1}^\star(s_{h+1}^k)|s_h^k,a_h^k]}{\sigma_h^{\star 2}(s_h^k,a_h^k)} \\
=& \frac{2[\mathrm{Var}_{P_h}V_{h+1}^\star](s_h^k,a_h^k)}{\sigma_h^{\star 2}(s_h^k,a_h^k)} \le 2,
\end{aligned}$$

then by Self-normalized Bernstein's inequality (Lemma C.11.4), with probability $1-\delta$,

$$\left\|\sum_{k=1}^K x_k(\theta,u,v)\eta_k\right\|_{G(\theta,u,v)^{-1}} \le 16\sqrt{d\log\left(1+\frac{K\kappa_1^2}{\lambda d}\right)\cdot\log\left(\frac{4K^2}{\delta}\right)} + 4\zeta\log\left(\frac{4K^2}{\delta}\right) \le \widetilde{O}(\sqrt{d})$$

where $|\eta_k| \le \zeta$ with $\zeta = 2\max_{s,a,s'} \frac{|f(\theta_{\mathbb{T}V_{h+1}^\star},\phi(s,a)) - r - V_{h+1}^\star(s')|}{\sigma_h^\star(s,a)}$ and the last inequality uses $\sqrt{d} \ge$

$\widetilde{O}(\zeta)$.

Step2. Define $h(\theta,u,v) := \sum_{k=1}^{K} x_k(\theta,u,v)\eta_k(u,v)$ and $H(\theta,u,v) := \|h(\theta,u,v)\|_{G_{u,v}(\theta)^{-1}}$,

$$\begin{aligned}
&\left\|h(\theta_1,u_1,v_1)-h(\theta_2,u_2,v_2)\right\|_2 \le K\max_k \left\|(x_k\cdot\eta_k)(\theta_1,u_1,v_1)-(x_k\cdot\eta_k)(\theta_2,u_2,v_2)\right\|_2\\
&\le K\max_k\left\{H\left|\frac{\nabla f(\theta_1,\phi_{h,k})-\nabla f(\theta_2,\phi_{h,k})}{\sigma^2_{u_1,v_1}(s_h^k,a_h^k)}\right| + H\kappa_1\left|\frac{\sigma^2_{u_1,v_1}(s_h^k,a_h^k)-\sigma^2_{u_2,v_2}(s_h^k,a_h^k)}{\sigma^2_{u_1,v_1}(s_h^k,a_h^k)\sigma^2_{u_2,v_2}(s_h^k,a_h^k)}\right|\right\}\\
&\le KH\kappa_1\left\|\theta_1-\theta_2\right\|_2 + KH\kappa_1\left\|\sigma^2_{u_1,v_1}-\sigma^2_{u_2,v_2}\right\|_2
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\left\|G_h(\theta_1,u_1,v_1)^{-1}-G_h(\theta_2,u_2,v_2)^{-1}\right\|_2 \le \left\|G_h(\theta_1,u_1,v_1)^{-1}\right\|_2\left\|G_h(\theta_1,u_1,v_1)-G_h(\theta_2,u_2,v_2)\right\|_2\left\|G_h(\theta_2,u_2,v_2)^{-1}\right\|_2\\
\le&\frac{1}{\lambda^2}K\sup_k\left\|\frac{\nabla f(\theta_1,\phi_{h,k})\cdot\nabla f(\theta_1,\phi_{h,k})^\top}{\sigma^2_{u_1,v_1}(s_h^k,a_h^k)}-\frac{\nabla f(\theta_2,\phi_{h,k})\cdot\nabla f(\theta_2,\phi_{h,k})^\top}{\sigma^2_{u_2,v_2}(s_h^k,a_h^k)}\right\|_2\\
\le&\frac{1}{\lambda^2}\left(K\kappa_2\kappa_1\left\|\theta_1-\theta_2\right\|_2+K\kappa_1^2\left\|\sigma^2_{u_1,v_1}-\sigma^2_{u_2,v_2}\right\|_2\right)
\end{aligned}$$

All the above imply

$$\begin{aligned}
&|H(\theta_1,u_1,v_1)-H(\theta_2,u_2,v_2)| \le \sqrt{\left|h(\theta_1,u_1,v_1)^\top G_{u_1,v_1}(\theta_1)^{-1}h(\theta_1,u_1,v_1)-h(\theta_2,u_2,v_2)^\top G_{u_2,v_2}(\theta_2)^{-1}h(\theta_2,u_2,v_2)\right|}\\
\le&\sqrt{\left\|h(\theta_1,u_1,v_1)-h(\theta_2,u_2,v_2)\right\|_2\cdot\frac{1}{\lambda}\cdot KH\kappa_1}+\sqrt{KH\kappa_1\cdot\left\|G_{u_1,v_1}(\theta_1)^{-1}-G_{u_2,v_2}(\theta_2)^{-1}\right\|_2\cdot KH\kappa_1}\\
+&\sqrt{(KH\kappa_1\cdot\frac{1}{\lambda})\cdot\left\|h(\theta_1,u_1,v_1)-h(\theta_2,u_2,v_2)\right\|_2}\\
\le&2\sqrt{KH\kappa_1(\left\|\theta_1-\theta_2\right\|_2+\left\|\sigma^2_{u_1,v_1}-\sigma^2_{u_2,v_2}\right\|_2)\cdot\frac{1}{\lambda}\cdot KH\kappa_1}+\sqrt{K^2H^2\kappa_1^2\cdot\frac{K\kappa_1}{\lambda^2}\left(\kappa_2\left\|\theta_1-\theta_2\right\|_2+\kappa_1\left\|\sigma^2_{u_1,v_1}-\sigma^2_{u_2,v_2}\right\|_2\right)}\\
\le&\left(\sqrt{4K^2H^2\kappa_1^2/\lambda}+\sqrt{K^3H^2\kappa_1^3\kappa_2/\lambda^2}\right)\sqrt{\left\|\theta_1-\theta_2\right\|_2}+\left(\sqrt{4K^2H^2\kappa_1^2/\lambda}+\sqrt{K^3H^2\kappa_1^4/\lambda^2}\right)\sqrt{\left\|\sigma^2_{u_1,v_1}-\sigma^2_{u_2,v_2}\right\|_2}
\end{aligned}$$

note

$$\begin{aligned}|\sigma^2_{u_1,v_1}(s,a)-\sigma^2_{u_2,v_2}(s,a)| \leq& \left|f(v_1,\phi(s,a))-f(v_2,\phi(s,a))\right| + 2H\left|f(u_1,\phi(s,a))-f(u_2,\phi(s,a))\right| \\ \leq& \kappa_1\left\|v_1-v_2\right\|_2 + 2H\kappa_1\left\|u_1-u_2\right\|_2,\end{aligned}$$

Then a ϵ-covering net of $\{H(\theta,u,v)\}$ can be constructed by the union of covering net for θ, u, v and by Lemma C.11.8, the covering number $\mathcal{N}_\epsilon$ satisfies (where $\widetilde{O}$ absorbs Polylog terms)

$$\log \mathcal{N}_\epsilon \leq \widetilde{O}(d)$$

Step3. First note by definition in Step2

$$\left\|\sum_{k=1}^{K}\frac{\left(f(\theta_{\mathbb{T}V^\star_{h+1}},\phi_{h,k})-r_{h,k}-V^\star_{h+1}(s^k_{h+1})\right)\cdot\nabla^\top_\theta f(\widehat{\theta}_h,\phi_{h,k})}{\widehat{\sigma}^2_h(s^k_h,a^k_h)}\right\|_{\Lambda_h^{-1}} = H(\widehat{\theta}_h,\boldsymbol{u}_h,\boldsymbol{v}_h)$$

Now choosing $\epsilon = O(1/K)$ in Step2 and union bound over the covering number in Step2, we obtain with probability $1-\delta$ (recall $\sqrt{d}\geq\widetilde{O}(\zeta)$),

$$\begin{aligned}H(\widehat{\theta}_h,\boldsymbol{u}_h,\boldsymbol{v}_h) \leq& 16\sqrt{d\log\left(1+\frac{K\kappa_1^2}{\lambda d}\right)\cdot[\log\left(\frac{4K^2}{\delta}\right)+\widetilde{O}(d)]} + 4\zeta[\log\left(\frac{4K^2}{\delta}\right)+\widetilde{O}(d)] + O(\frac{1}{K}) \\ \leq& \widetilde{O}(d)+O(\frac{1}{K})\end{aligned}$$

where we absorb all the Polylog terms. Combing above with (C.31), we obtain with probability

$1-\delta$,

$$
\begin{aligned}
|I_1| \leq & \left\|\nabla f(\widehat{\theta}_h, \phi(s,a))\right\|_{\Lambda_h^{-1}} \cdot H(\widehat{\theta}_h, \boldsymbol{u}_h, \boldsymbol{v}_h) \\
\leq & \left\|\nabla f(\widehat{\theta}_h, \phi(s,a))\right\|_{\Lambda_h^{-1}} \cdot \left[\widetilde{O}(d) + O(\frac{1}{K})\right] \\
\leq & \widetilde{O}\left(d \left\|\nabla f(\widehat{\theta}_h, \phi(s,a))\right\|_{\Lambda_h^{-1}}\right) + \widetilde{O}(\frac{\kappa_1}{\sqrt{\kappa}K}),
\end{aligned}
$$

Combing dominate term I_1 (via Lemma C.9.4) and all other higher order terms we can obtain the first result together with Lemma C.3.2.

The proof of the second result is also very similar to the proofs in Section C.5.2. Concretely, when picking $\pi = \pi^\star$, we can convert the quantity

$$
\sqrt{\nabla_\theta^\top f(\widehat{\theta}_h, \phi(s_h,a_h))\Lambda_h^{-1}\nabla_\theta f(\widehat{\theta}_h, \phi(s_h,a_h))}
$$

to

$$
\sqrt{\nabla_\theta^\top f(\theta_h^\star, \phi(s_h,a_h))\Lambda_h^{-1}\nabla_\theta f(\theta_h^\star, \phi(s_h,a_h))}
$$

using Theorem C.9.2, and convert

$$
\sqrt{\nabla_\theta^\top f(\theta_h^\star, \phi(s_h,a_h))\Lambda_h^{-1}\nabla_\theta f(\theta_h^\star, \phi(s_h,a_h))}
$$

to

$$
\sqrt{\nabla_\theta^\top f(\theta_h^\star, \phi(s_h,a_h))\Lambda_h^{\star-1}\nabla_\theta f(\theta_h^\star, \phi(s_h,a_h))}
$$

using Lemma C.9.2.

C.10 The lower bound

Theorem C.10.1 (Restatement of Theorem 4.4.2)**.** *Specifying the model to have linear representation $f = \langle \theta, \phi \rangle$. There exist a pair of universal constants $c, c' > 0$ such that given dimension d, horizon H and sample size $K > c'd^3$, one can always find a family of MDP instances such that for any algorithm $\widehat{\pi}$*

$$\inf_{\widehat{\pi}} \sup_{M \in \mathcal{M}} \mathbb{E}_M \left[v^\star - v^{\widehat{\pi}} \right] \geq c \sqrt{d} \cdot \sum_{h=1}^{H} \mathbb{E}_{\pi^\star} \left[\sqrt{\nabla_\theta^\top f(\theta_h^\star, \phi(\cdot,\cdot))(\Lambda_h^{\star,p})^{-1} \nabla_\theta f(\theta_h^\star, \phi(\cdot,\cdot))} \right], \tag{C.32}$$

where $\Lambda_h^{\star,p} = \mathbb{E}\Big[\sum_{k=1}^{K} \frac{\nabla_\theta f(\theta_h^\star, \phi(s_h^k, a_h^k)) \cdot \nabla_\theta f(\theta_h^\star, \phi(s_h^k, a_h^k))^\top}{\mathrm{Var}_h(V_{h+1}^\star)(s_h^k, a_h^k)} \Big]$.

Remark 8. *Note Theorem 4.4.2 is a valid lower bound for comparison. This is because the upper bound result holds true for all model f such that the corresponding $\mathcal{F}$ satisfies Assumption 4.2.1, 4.2.3. Therefore, for the lower bound construction it suffices to find one model f such that the lower bound* (C.32) *holds. Here we simply choose the linear function approximation.*

C.10.1 Regarding the proof of lower bound

The proof of Theorem 4.4.2 can be done via a reduction to linear function approximation lower bound. In fact, it can be directly obtained from Theorem 3.5 of [106], and the original proof comes from Theorem 2 of [41].

Concretely, all the proofs in Theorem 3.5 of [106] follows and the only modification is to replace

$$\sqrt{\mathbb{E}_{\pi^\star}[\phi]^\top \left(\Lambda_h^\star\right)^{-1} \mathbb{E}_{\pi^\star}[\phi]} \leq \frac{1}{2} \left\| \phi\left(+1, u_h\right) \right\|_{\left(\Lambda_h^{\star,p}\right)^{-1}} + \frac{1}{2} \left\| \phi\left(-1, u_h\right) \right\|_{\left(\Lambda_h^{\star,p}\right)^{-1}}$$

in Section E.5 by

$$\mathbb{E}_{\pi^\star}\left[\sqrt{\phi(\cdot,\cdot)^\top(\Lambda_h^{\star,p})^{-1}\phi(\cdot,\cdot)}\right] = \frac{1}{2}\left\|\phi(+1,u_h)\right\|_{(\Lambda_h^{\star,p})^{-1}} + \frac{1}{2}\left\|\phi(-1,u_h)\right\|_{(\Lambda_h^{\star,p})^{-1}},$$

and the final result holds with $\phi(\cdot,\cdot) = \nabla_\theta f(\theta_h^\star, \phi(\cdot,\cdot))$ by the reduction $f = \langle\theta,\phi\rangle$.

C.11 Helpful Results

Lemma C.11.1 (k-th Order Mean Value Form of Taylor's Expansion)**.** *Let $k \geq 1$ be an integer and let function $f : \mathbb{R}^d \to \mathbb{R}$ be k times differentiable and continuous over the compact domain $\Theta \subset \mathbb{R}^d$. Then for any $x, \theta \in \Theta$, there exists ξ in the line segment of x and θ, such that*

$$\begin{aligned} f(x) - f(\theta) =& \nabla f(\theta)^\top(x-\theta) + \frac{1}{2!}(x-\theta)^\top\nabla^2_{\theta\theta}f(\theta)(x-\theta) + \ldots + \frac{1}{(k-1)!}\nabla^{k-1}f(\theta)\left(\bigotimes(x-\theta)\right)^{k-1} \\ &+\frac{1}{k!}\nabla^k f(\xi)\left(\bigotimes(x-\theta)\right)^k. \end{aligned}$$

Here $\nabla^k f(\theta)$ denotes k-dimensional tensor and $\bigotimes$ denotes tensor product.

Lemma C.11.2 (Vector Hoeffding's Inequality)**.** *Let $X = (X_1, \ldots, X_d)$ be d-dimensional vector Random Variable with $E[X] = 0$ and $\|X\|_2 \leq R$. $X^{(1)}, \ldots, X^{(n)}$'s are n samples. Then with probability $1-\delta$,*

$$\left\|\frac{1}{n}\sum_{i=1}^n X^{(i)}\right\|_2 \leq \sqrt{\frac{4dR^2}{n}\log(\frac{d}{\delta})}.$$

Proof: [Proof of Lemma C.11.2] Since $\|X\|_2 \leq R$ implies $|X_j| \leq R$, by the univariate Hoeffding's inequality, for a fixed $j \in \{1, ..., d\}$, denote $Y_j := \frac{1}{n}\sum_{i=1}^n X_j^{(i)}$. Then with probability $1-\delta$ (note $|X_j^{(i)}| \leq R$),

$$\mathbb{P}\left(|Y_j| \geq 2\sqrt{\frac{R^2}{n}\log(\frac{1}{\delta})}\right) \leq \delta.$$

By a union bound,

$$\mathbb{P}\left(\exists\ i\ s.t.\ |Y_j| \geq 2\sqrt{\frac{R^2}{n}\log(\frac{1}{\delta})}\right) \leq d\delta \Leftrightarrow \mathbb{P}\left(\forall\ i\ \ |Y_j| \leq 2\sqrt{\frac{R^2}{n}\log(\frac{1}{\delta})}\right) \geq 1 - d\delta$$
$$\Leftrightarrow \mathbb{P}\left(\forall\ i\ \ Y_j^2 \leq \frac{4R^2}{n}\log(\frac{1}{\delta})\right) \geq 1 - d\delta \Rightarrow \mathbb{P}\left(\|Y\|_2 \leq \sqrt{\frac{4dR^2}{n}\log(\frac{1}{\delta})}\right) \geq 1 - d\delta$$
$$\Leftrightarrow \mathbb{P}\left(\|Y\|_2 \leq \sqrt{\frac{4dR^2}{n}\log(\frac{d}{\delta})}\right) \geq 1 - \delta.$$

Lemma C.11.3 (Hoeffding inequality for self-normalized martingales [160])**.** *Let $\{\eta_t\}_{t=1}^{\infty}$ be a real-valued stochastic process. Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration, such that η_t is $\mathcal{F}_t$-measurable. Assume η_t also satisfies η_t given $\mathcal{F}_{t-1}$ is zero-mean and R-subgaussian,* i.e.

$$\forall \lambda \in \mathbb{R}, \quad \mathbb{E}\left[e^{\lambda\eta_t} \mid \mathcal{F}_{t-1}\right] \leq e^{\lambda^2 R^2/2}$$

Let $\{x_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process where x_t is $\mathcal{F}_{t-1}$ measurable and $\|x_t\| \leq L$. Let $\Lambda_t = \lambda I_d + \sum_{s=1}^{t} x_s x_s^{\top}$. Then for any $\delta > 0$, with probability $1 - \delta$, for all $t > 0$,

$$\left\|\sum_{s=1}^{t} x_s\eta_s\right\|_{\Lambda_t^{-1}}^2 \leq 8R^2 \cdot \frac{d}{2}\log\left(\frac{\lambda + tL}{\lambda\delta}\right).$$

Lemma C.11.4 (Bernstein inequality for self-normalized martingales [76])**.** *Let $\{\eta_t\}_{t=1}^{\infty}$ be a real-valued stochastic process. Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration, such that η_t is $\mathcal{F}_t$-measurable. Assume η_t also satisfies*

$$|\eta_t| \leq R, \mathbb{E}\left[\eta_t \mid \mathcal{F}_{t-1}\right] = 0, \mathbb{E}\left[\eta_t^2 \mid \mathcal{F}_{t-1}\right] \leq \sigma^2.$$

Let $\{x_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process where x_t is $\mathcal{F}_{t-1}$ measurable and $\|x_t\| \leq L$.

Let $\Lambda_t = \lambda I_d + \sum_{s=1}^{t} x_s x_s^\top$. *Then for any* $\delta > 0$, *with probability* $1-\delta$, *for all* $t > 0$,

$$\left\| \sum_{s=1}^{t} \mathbf{x}_s \eta_s \right\|_{\mathbf{\Lambda}_t^{-1}} \leq 8\sigma \sqrt{d \log\left(1 + \frac{tL^2}{\lambda d}\right) \cdot \log\left(\frac{4t^2}{\delta}\right)} + 4R \log\left(\frac{4t^2}{\delta}\right)$$

Lemma C.11.5. *Let* $\nabla f(\theta, \phi(\cdot,\cdot)) : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$ *be a bounded function s.t.* $\sup_{\theta\in\Theta} \|\nabla f(\theta, \phi(\cdot,\cdot))\|_2 \leq \kappa_1$. *If* K *satisfies*

$$K \geq \max\left\{ 512 \frac{\kappa_1^4}{\kappa^2} \left(\log(\frac{2d}{\delta}) + d\log(1 + \frac{4\kappa_1 B^2 \kappa_2 C_\Theta K^3}{\lambda^2}) \right), \frac{4\lambda}{\kappa} \right\}$$

Then with probability at least $1-\delta$, *for all* $\|u\|_2 \leq B$ *simultaneously, it holds that*

$$\|u\|_{\Sigma_h^{-1}} \leq \frac{2B}{\sqrt{\kappa K}} + O(\frac{1}{K})$$

where $\Sigma_h = \sum_{k=1}^{K} \nabla f(\widehat{\theta}_h, \phi(s_h^k, a_h^k)) \cdot \nabla f(\widehat{\theta}_h, \phi(s_h^k, a_h^k))^\top + \lambda I_d$.

Proof: [Proof of Lemma C.11.5] For a fixed θ, define $\bar{G} = \sum_{k=1}^{K} \nabla f(\theta, \phi(s_h^k, a_h^k)) \cdot \nabla f(\theta, \phi(s_h^k, a_h^k))^\top + \lambda I_d$, and $G = \mathbb{E}_\mu[\nabla f(\theta, \phi(s_h, a_h)) \cdot \nabla f(\theta, \phi(s_h, a_h))^\top]$, then by Lemma H.5. of [82], as long as

$$K \geq \max\{512\kappa_1^4 \left\|G^{-1}\right\|^2 \log(\frac{2d}{\delta}), 4\lambda \left\|G^{-1}\right\|_2\}, \tag{C.33}$$

then with probability $1-\delta$, for all $u \in \mathbb{R}^d$ simultaneously, $\|u\|_{\bar{G}^{-1}} \leq \frac{2}{\sqrt{K}} \|u\|_{G^{-1}}$. As a corollary, if we constraint u to the subspace $\|u\|_2 \leq B$, then we have: with probability $1-\delta$, for all $\{u \in \mathbb{R}^d : \|u\|_2 \leq B\}$ simultaneously,

$$\|u\|_{\bar{G}^{-1}} \leq \frac{2}{\sqrt{K}} \|u\|_{G^{-1}} = \frac{2}{\sqrt{K}} \sqrt{u^\top G^{-1} u} \leq \frac{2B\sqrt{\|G^{-1}\|_2}}{\sqrt{K}}. \tag{C.34}$$

Next, for any θ, define

$$h_u(\theta) := \|u\|_{\bar{G}^{-1}} = \sqrt{u^\top \bar{G}^{-1} u} = \sqrt{u^\top \left(\sum_{k=1}^{K} \nabla f(\theta, \phi(s_h^k, a_h^k)) \cdot \nabla f(\theta, \phi(s_h^k, a_h^k))^\top + \lambda I_d \right)^{-1} u}$$

and $\bar{G}(\theta) = \sum_{k=1}^{K} \nabla f(\theta, \phi(s_h^k, a_h^k)) \cdot \nabla f(\theta, \phi(s_h^k, a_h^k))^\top + \lambda I_d$, we have for any θ_1, θ_2

$$\begin{aligned}
\left\| \bar{G}(\theta_1) - \bar{G}(\theta_2) \right\|_2 \leq & \left\| \sum_{k=1}^{K} \left(\nabla f(\theta_1, \phi(s_h^k, a_h^k)) - \nabla f(\theta_2, \phi(s_h^k, a_h^k)) \right) \cdot \nabla f(\theta_1, \phi(s_h^k, a_h^k))^\top \right\| \\
& + \left\| \sum_{k=1}^{K} \nabla f(\theta_2, \phi(s_h^k, a_h^k)) \left(\nabla f(\theta_1, \phi(s_h^k, a_h^k)) - \nabla f(\theta_2, \phi(s_h^k, a_h^k)) \right)^\top \right\| \\
\leq & K\kappa_2\kappa_1 \left\| \theta_1 - \theta_2 \right\|_2 + K\kappa_2\kappa_1 \left\| \theta_1 - \theta_2 \right\|_2 \leq 2K\kappa_2\kappa_1 \left\| \theta_1 - \theta_2 \right\|_2 .
\end{aligned}$$

Use the basic inequality for $a, b > 0 \Rightarrow |\sqrt{a} - \sqrt{b}| \leq \sqrt{|a - b|}$,

$$\begin{aligned}
\sup_u |h_u(\theta_1) - h_u(\theta_2)| \leq & \sup_u \sqrt{\left| u^\top \left(\bar{G}(\theta_1)^{-1} - \bar{G}(\theta_2)^{-1} \right) u \right|} \leq \sqrt{B^2 \cdot \left\| \bar{G}(\theta_1)^{-1} - \bar{G}(\theta_2)^{-1} \right\|_2} \\
\leq & \sqrt{B^2 \cdot \left\| \bar{G}(\theta_1)^{-1} \right\|_2 \left\| \bar{G}(\theta_1) - \bar{G}(\theta_2) \right\|_2 \left\| \bar{G}(\theta_2)^{-1} \right\|_2} \\
\leq & \sqrt{B^2 \frac{1}{\lambda} 2K\kappa_2\kappa_1 \left\| \theta_1 - \theta_2 \right\|_2 \frac{1}{\lambda}} = \sqrt{\frac{2B^2 K\kappa_1\kappa_2 \left\| \theta_1 - \theta_2 \right\|_2}{\lambda^2}}
\end{aligned}$$

Therefore, the ϵ-covering net of $\{h(\theta) : \theta \in \Theta\}$ is implies by the $\frac{\lambda^2\epsilon^2}{2KB^2\kappa_1\kappa_2}$-covering net of $\{\theta : \theta \in \Theta\}$, so by Lemma C.11.8, the covering number $\mathcal{N}_\epsilon$ satisfies

$$\log \mathcal{N}_\epsilon \leq d \log(1 + \frac{4B^2 K\kappa_1\kappa_2 C_\Theta}{\lambda^2\epsilon^2}).$$

Select $\theta = \widehat{\theta}_h$. Choose $\epsilon = O(1/K)$ and by a union bound over (C.34) to get with probability

$1-\delta$, for all $\|u\|_2 \leq B$ (note By Assumption 4.2.3 $\left\|G^{-1}\right\|_2 \leq 1/\kappa$),

$$\|u\|_{\Sigma_h^{-1}} \leq \frac{2B}{\sqrt{\kappa K}} + O(\frac{1}{K})$$

if (union bound over the condition (C.33))

$$K \geq \max\left\{ 512\frac{\kappa_1^4}{\kappa^2}\left(\log(\frac{2d}{\delta}) + d\log(1+\frac{4\kappa_1 B^2\kappa_2 C_\Theta K^3}{\lambda^2})\right), \frac{4\lambda}{\kappa}\right\}$$

where this condition is satisfied by the Lemma statement.

Lemma C.11.6. *let $\phi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$ satisfies $\|\phi(s,a)\| \leq C$ for all $s,a \in \mathcal{S}\times\mathcal{A}$. For any $K>0, \lambda>0$, define $\bar{G}_K = \sum_{k=1}^K \phi(s_k,a_k)\phi(s_k,a_k)^\top + \lambda I_d$ where (s_k,a_k)'s are i.i.d samples from some distribution ν. Then with probability $1-\delta$,*

$$\left\|\frac{\bar{G}_K}{K} - \mathbb{E}_\nu\left[\frac{\bar{G}_K}{K}\right]\right\| \leq \frac{4\sqrt{2}C^2}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2}.$$

Proof: [Proof of Lemma C.11.6] See Lemma H.5 of [106] or Lemma H.4 of Lemma [82] for details.

Lemma C.11.7 (Lemma H.4 in [106])**.** *Let Λ_1 and $\Lambda_2 \in \mathbb{R}^{d\times d}$ are two positive semi-definite matrices. Then:*

$$\left\|\Lambda_1^{-1}\right\| \leq \left\|\Lambda_2^{-1}\right\| + \left\|\Lambda_1^{-1}\right\|\cdot\left\|\Lambda_2^{-1}\right\|\cdot\|\Lambda_1-\Lambda_2\|$$

and

$$\|\phi\|_{\Lambda_1^{-1}} \leq \left[1+\sqrt{\left\|\Lambda_2^{-1}\right\|\|\Lambda_2\|\cdot\left\|\Lambda_1^{-1}\right\|\cdot\|\Lambda_1-\Lambda_2\|}\right]\cdot\|\phi\|_{\Lambda_2^{-1}}.$$

for all $\phi \in \mathbb{R}^d$.

C.11.1 Covering Arguments

Lemma C.11.8. *(Covering Number of Euclidean Ball) For any $\epsilon > 0$, the ϵ-covering number of the Euclidean ball in $\mathbb{R}^d$ with radius $R > 0$ is upper bounded by $(1 + 2R/\epsilon)^d$.*

Lemma C.11.9. *Define $\mathcal{V}$ to be the class mapping $\mathcal{S}$ to $\mathbb{R}$ with the parametric form*

$$V(\cdot) := \min\{\max_a f(\theta, \phi(\cdot, a)) - \sqrt{\nabla f(\theta, \phi(\cdot, a))^\top A \cdot \nabla f(\theta, \phi(\cdot, a))}, H\}$$

where the parameter spaces are $\{\theta : \|\theta\|_2 \le C_\Theta\}$ and $\{A : \|A\|_2 \le B\}$. Let $\mathcal{N}_\epsilon^{\mathcal{V}}$ be the covering number of ϵ-net with respect to l_∞ distance, then we have

$$\log \mathcal{N}_\epsilon^{\mathcal{V}} \le d \log\left(1 + \frac{8C_\Theta(\kappa_1\sqrt{C_\Theta} + 2\sqrt{B\kappa_1\kappa_2})^2}{\epsilon^2}\right) + d^2 \log\left(1 + \frac{8\sqrt{d}B\kappa_1^2}{\epsilon^2}\right).$$

Proof: [Proof of Lemma C.11.9]

$$\begin{aligned}
&\sup_s |V_1(s) - V_2(s)| \\
\le &\sup_{s,a} \left| f(\theta_1, \phi(s,a)) - \sqrt{\nabla f(\theta_1, \phi(s,a))^\top A_1 \cdot \nabla f(\theta_1, \phi(s,a))} - f(\theta_2, \phi(s,a)) + \sqrt{\nabla f(\theta_2, \phi(s,a))^\top A_2 \cdot \nabla f(\theta_2, \phi(s,a))} \right| \\
= &\sup_{s,a} \left| \nabla f(\xi, \phi(s,a)) \cdot (\theta_1 - \theta_2) - \sqrt{\nabla f(\theta_1, \phi(s,a))^\top A_1 \cdot \nabla f(\theta_1, \phi(s,a))} + \sqrt{\nabla f(\theta_2, \phi(s,a))^\top A_2 \cdot \nabla f(\theta_2, \phi(s,a))} \right| \\
\le &\kappa_1 \cdot \|\theta_1 - \theta_2\|_2 + \sup_{s,a} \left| \sqrt{\nabla f(\theta_1, \phi(s,a))^\top A_1 \cdot \nabla f(\theta_1, \phi(s,a))} - \sqrt{\nabla f(\theta_2, \phi(s,a))^\top A_2 \cdot \nabla f(\theta_2, \phi(s,a))} \right| \\
\le &\kappa_1 \cdot \|\theta_1 - \theta_2\|_2 + \sup_{s,a} \sqrt{|[\nabla f(\theta_1, \phi(s,a)) - \nabla f(\theta_2, \phi(s,a))]^\top A_1 \cdot \nabla f(\theta_1, \phi(s,a))|} \\
&+ \sup_{s,a} \sqrt{|\nabla f(\theta_2, \phi(s,a))^\top (A_1 - A_2) \cdot \nabla f(\theta_1, \phi(s,a))|} + \sup_{s,a} \sqrt{|\nabla f(\theta_2, \phi(s,a))^\top A_2 \cdot [\nabla f(\theta_1, \phi(s,a)) - \nabla f(\theta_2, \phi(s,a))]|} \\
\le &\kappa_1 \cdot \|\theta_1 - \theta_2\|_2 + 2\sup_{s,a} \sqrt{\|\nabla f(\theta_1, \phi(s,a)) - \nabla f(\theta_2, \phi(s,a))\|_2 \cdot B \cdot \kappa_1} + \sqrt{\kappa_1^2 \|A_1 - A_2\|_2} \\
\le &\kappa_1 \cdot \|\theta_1 - \theta_2\|_2 + 2\sup_{s,a} \sqrt{\|\nabla f(\theta_1, \phi(s,a)) - \nabla f(\theta_2, \phi(s,a))\|_2 \cdot B \cdot \kappa_1} + \sqrt{\kappa_1^2 \|A_1 - A_2\|_2} \\
\le &\kappa_1 \cdot \|\theta_1 - \theta_2\|_2 + 2\sup_{s,a} \sqrt{\|\nabla f(\theta_1, \phi(s,a))\|_2 \cdot \|\theta_1 - \theta_2\|_2 \cdot B \cdot \kappa_1} + \sqrt{\kappa_1^2 \|A_1 - A_2\|_2} \\
\le &\kappa_1 \cdot \|\theta_1 - \theta_2\|_2 + 2\sqrt{\kappa_2 \cdot \|\theta_1 - \theta_2\|_2 \cdot B \cdot \kappa_1} + \sqrt{\kappa_1^2 \|A_1 - A_2\|_2} \\
\le &\left(\kappa_1\sqrt{C_\Theta} + 2\sqrt{B\kappa_1\kappa_2}\right)\sqrt{\|\theta_1 - \theta_2\|_2} + \kappa_1\sqrt{\|A_1 - A_2\|_2} \le \left(\kappa_1\sqrt{C_\Theta} + 2\sqrt{B\kappa_1\kappa_2}\right)\sqrt{\|\theta_1 - \theta_2\|_2} + \kappa_1\sqrt{\|A_1 - A_2\|_F}
\end{aligned}$$

Here $\|\cdot\|_F$ is Frobenius norm. Let C_θ be the $\frac{\epsilon^2}{4(\kappa_1\sqrt{C_\Theta}+2\sqrt{B\kappa_1\kappa_2})^2}$-net of space $\{\theta : \|\theta\|_2 \le C_\Theta\}$

and C_w be the $\frac{\epsilon^2}{4\kappa_1^2}$-net of the space $\{A : \|A\|_F \le \sqrt{d}B\}$, then by Lemma C.11.8,

$$|C_w| \le \left(1 + \frac{8C_\Theta(\kappa_1\sqrt{C_\Theta} + 2\sqrt{B\kappa_1\kappa_2})^2}{\epsilon^2}\right)^d, \quad |C_A| \le \left(1 + \frac{8\sqrt{d}B\kappa_1^2}{\epsilon^2}\right)^{d^2}$$

Therefore, the covering number of space $\mathcal{V}$ satisfies

$$\log \mathcal{N}_\epsilon^{\mathcal{V}} \le \log(|C_w|\cdot|C_A|) \le d\log\left(1 + \frac{8C_\Theta(\kappa_1\sqrt{C_\Theta} + 2\sqrt{B\kappa_1\kappa_2})^2}{\epsilon^2}\right) + d^2\log\left(1 + \frac{8\sqrt{d}B\kappa_1^2}{\epsilon^2}\right)$$

Lemma C.11.10 (Covering of $\mathbb{E}_\mu(X(g, V, f))$)**.** *Define*

$$X(\theta, \theta') := (f(\theta, \phi(s,a)) - r - V_{\theta'}(s'))^2 - (f_{V_{\theta'}}(s,a) - r - V_{\theta'}(s'))^2,$$

where $f_V := \mathcal{P}_h V + \delta_V$ and $V(s)$ has form $V_\theta(s)$ that belongs to $\mathcal{V}$ (as defined in Lemma C.11.9). Here $X(\theta, \theta')$ is a function of s, a, r, s' as well, and we suppress the notation for conciseness only. Then the function class $\mathcal{H} = \{h(\theta, \theta') := \mathbb{E}_\mu[X(\theta, \theta')] \mid \|\theta\|_2 \le C_\Theta, V_\theta \in \mathcal{V}\}$ has the covering number of $(\epsilon + 4H\epsilon_{\mathcal{F}})$-net bounded by

$$d\log(1 + \frac{24C_\Theta(H+1)\kappa_1}{\epsilon}) + d\log\left(1 + \frac{288H^2C_\Theta(\kappa_1\sqrt{C_\Theta} + 2\sqrt{B\kappa_1\kappa_2})^2}{\epsilon^2}\right) + d^2\log\left(1 + \frac{288H^2\sqrt{d}B\kappa_1^2}{\epsilon^2}\right).$$

Proof: [Proof of Lemma C.11.10] First of all,

$$X(\theta, \theta') = f(\theta, \phi(s,a))^2 - f_{V_{\theta'}}(s,a)^2 - 2f(\theta, \phi(s,a))\cdot(r + V_{\theta'}(s')) + 2f_{V_{\theta'}}(s,a)\cdot(r + V_{\theta'}(s')),$$

For any (θ_1, θ_1'), (θ_2, θ_2'),

$$
\begin{aligned}
&|X(\theta_1, \theta_1') - X(\theta_2, \theta_2')| \leq |f(\theta_1, \phi(s,a))^2 - f(\theta_2, \phi(s,a))^2| \\
&+|f_{V_{\theta_1'}}(s,a)^2 - f_{V_{\theta_2'}}(s,a)^2| + 2|f_{V_{\theta_1'}}(s,a) - f_{V_{\theta_2'}}(s,a)| \cdot (r + V_{\theta_1'}(s')) \\
&+2f_{V_{\theta_2'}}(s,a) \cdot |V_{\theta_1'}(s') - V_{\theta_2'}(s')| + 2|f(\theta_1, \phi(s,a)) - f(\theta_2, \phi(s,a))| \cdot (r + V_{\theta_1'}(s')) \\
&+2|f(\theta_2, \phi(s,a))| \cdot |V_{\theta_1'}(s') - V_{\theta_2'}(s')| \\
&\leq 2H \cdot |f(\theta_1, \phi(s,a)) - f(\theta_2, \phi(s,a))| + 2H \cdot |f_{V_{\theta_1'}}(s,a) - f_{V_{\theta_2'}}(s,a)| \\
&+4H \cdot |V_{\theta_1'}(s') - V_{\theta_2'}(s')| + 4(H+1) \cdot |f(\theta_1, \phi(s,a)) - f(\theta_2, \phi(s,a))| \\
&\leq (6H+1) \cdot |f(\theta_1, \phi(s,a)) - f(\theta_2, \phi(s,a))| + 2H \max_{s'} |V_{\theta_1'}(s') - V_{\theta_2'}(s')| + 4H\epsilon_{\mathcal{F}} \\
&+4H \cdot |V_{\theta_1'}(s') - V_{\theta_2'}(s')| \\
&\leq (6H+1) \|\nabla f(\xi, \phi(s,a))\|_2 \cdot \left\|\theta_1 - \theta_2\right\|_2 + 6H \left\|V_{\theta_1'} - V_{\theta_2'}\right\|_\infty + 4H\epsilon_{\mathcal{F}} \\
&\leq (6H+1)\kappa_1 \cdot \left\|\theta_1 - \theta_2\right\|_2 + 6H \left\|V_{\theta_1'} - V_{\theta_2'}\right\|_\infty + 4H\epsilon_{\mathcal{F}}
\end{aligned}
$$

where the second inequality comes from $f_V = \mathcal{P}_h V + \delta_V$. Note the above holds true for all s, a, r, s', therefore it implies

$$
\begin{aligned}
|\mathbb{E}_\mu[X(\theta_1, \theta_1')] - \mathbb{E}_\mu[X(\theta_2, \theta_2')]| &\leq \sup_{s,a,s'} |X(\theta_1, \theta_1') - X(\theta_2, \theta_2')| \\
&\leq (6H+1)\kappa_1 \cdot \left\|\theta_1 - \theta_2\right\|_2 + 6H \left\|V_{\theta_1'} - V_{\theta_2'}\right\|_\infty + 4H\epsilon_{\mathcal{F}}
\end{aligned}
$$

Now let $\mathcal{C}_1$ be the $\frac{\epsilon}{12(H+1)\kappa_1}$-net of $\{\theta : \|\theta\|_2 \leq C_\Theta\}$ and $\mathcal{C}_2$ be the $\epsilon/6H$-net of $\mathcal{V}$, applying

Lemma C.11.8 and Lemma C.11.9 to obtain

$$\log|\mathcal{C}_1| \leq d\log(1+\frac{24C_\Theta(H+1)\kappa_1}{\epsilon}),$$
$$\log|\mathcal{C}_2| \leq d\log\left(1+\frac{288H^2C_\Theta(\kappa_1\sqrt{C_\Theta}+2\sqrt{B\kappa_1\kappa_2})^2}{\epsilon^2}\right)+d^2\log\left(1+\frac{288H^2\sqrt{d}B\kappa_1^2}{\epsilon^2}\right)$$

which implies the covering number of $\mathcal{H}$ to be bounded by

$$\log|\mathcal{C}_1|\cdot|\mathcal{C}_2| \leq d\log(1+\frac{24C_\Theta(H+1)\kappa_1}{\epsilon})+d\log\left(1+\frac{288H^2C_\Theta(\kappa_1\sqrt{C_\Theta}+2\sqrt{B\kappa_1\kappa_2})^2}{\epsilon^2}\right)$$
$$+d^2\log\left(1+\frac{288H^2\sqrt{d}B\kappa_1^2}{\epsilon^2}\right).$$

Lemma C.11.11. *Denote* $\sigma^2_{u,v}(\cdot,\cdot) := \max\{1, f(v,\phi(\cdot,\cdot))_{[0,(H-h+1)^2]} - \left[f(u,\phi(\cdot,\cdot))_{[0,H-h+1]}\right]^2\}$ *and define*

$$\bar{X}(\theta,\theta',u,v) := \frac{(f(\theta,\phi(s,a))-r-V_{\theta'}(s'))^2-(f_{V_{\theta'}}(s,a)-r-V_{\theta'}(s'))^2}{\sigma^2_{u,v}(s,a)},$$

where $f_V := \mathcal{P}_h V$ *and* $V(s)$ *has form* $V_\theta(s)$ *that belongs to* $\mathcal{V}$ *(as defined in Lemma C.11.9). Here* $\bar{X}(\theta,\theta',u,v)$ *is a function of* s,a,r,s' *as well, and we suppress the notation for conciseness only. Then the function class* $\mathcal{H} = \{h(\theta,\theta',u,v) := \mathbb{E}_\mu[\bar{X}(\theta,\theta',u,v)] \mid \|\theta\|_2 \leq C_\Theta, V_\theta \in \mathcal{V}\}$ *has the covering number of* ϵ*-net bounded by*

$$d\log(1+\frac{24C_\Theta(H+1)\kappa_1}{\epsilon})+d\log\left(1+\frac{288H^2C_\Theta(\kappa_1\sqrt{C_\Theta}+2\sqrt{B\kappa_1\kappa_2})^2}{\epsilon^2}\right)+d^2\log\left(1+\frac{288H^2\sqrt{d}B\kappa_1^2}{\epsilon^2}\right)$$
$$+d\log(1+\frac{16C_\Theta H^2\kappa_1}{\epsilon})+d\log(1+\frac{32C_\Theta H^3\kappa_1}{\epsilon})$$

Proof: [Proof of Lemma C.11.11] Recall $\sigma^2_{u,v}(\cdot,\cdot) := \max\{1, f(v,\phi(\cdot,\cdot))_{[0,(H-h+1)^2]} - \left[f(u,\phi(\cdot,\cdot))_{[0,H-h+1]}\right]^2\}$, and since max, truncation are non-expansive operations, then we can

achieve for any s, a

$$|\sigma^2_{u_1,v_1}(s,a) - \sigma^2_{u_2,v_2}(s,a)| \leq \left|f(v_1, \phi(s,a)) - f(v_2, \phi(s,a))\right| + 2H\left|f(u_1, \phi(s,a)) - f(u_2, \phi(s,a))\right|$$
$$\leq \kappa_1 \left\|v_1 - v_2\right\|_2 + 2H\kappa_1 \left\|u_1 - u_2\right\|_2,$$

Hence

$$\begin{aligned}
&\left|\bar{X}(\theta_1, \theta_1', u_1, v_1) - \bar{X}(\theta_2, \theta_2', u_2, v_2)\right| = \left|\frac{X(\theta_1, \theta_1')}{\sigma^2_{u_1,v_1}} - \frac{X(\theta_2, \theta_2')}{\sigma^2_{u_2,v_2}}\right| \\
&\leq \left|\frac{X(\theta_1, \theta_1') - X(\theta_2, \theta_2')}{\sigma^2_{u_1,v_1}}\right| + \left|\frac{X(\theta_2, \theta_2')}{\sigma^2_{u_1,v_1}\sigma^2_{u_2,v_2}}\left(\sigma^2_{u_1,v_1} - \sigma^2_{u_2,v_2}\right)\right| \\
&\leq \left|X(\theta_1, \theta_1') - X(\theta_2, \theta_2')\right| + 2H^2\left|\sigma^2_{u_1,v_1} - \sigma^2_{u_2,v_2}\right| \\
&\leq \left|X(\theta_1, \theta_1') - X(\theta_2, \theta_2')\right| + 2H^2\kappa_1 \left\|v_1 - v_2\right\|_2 + 4H^3\kappa_1 \left\|u_1 - u_2\right\|_2 \\
&\leq (6H+1)\kappa_1 \cdot \left\|\theta_1 - \theta_2\right\|_2 + 6H\left\|V_{\theta_1'} - V_{\theta_2'}\right\|_\infty + 2H^2\kappa_1 \left\|v_1 - v_2\right\|_2 + 4H^3\kappa_1 \left\|u_1 - u_2\right\|_2
\end{aligned}$$

Note the above holds true for all s, a, r, s', therefore it implies

$$\begin{aligned}
&\left|\mathbb{E}_\mu[\bar{X}(\theta_1, \theta_1', u_1, v_1)] - \mathbb{E}_\mu[\bar{X}(\theta_2, \theta_2', u_2, v_2)]\right| \\
\leq &(6H+1)\kappa_1 \cdot \left\|\theta_1 - \theta_2\right\|_2 + 6H\left\|V_{\theta_1'} - V_{\theta_2'}\right\|_\infty + 2H^2\kappa_1 \left\|v_1 - v_2\right\|_2 + 4H^3\kappa_1 \left\|u_1 - u_2\right\|_2
\end{aligned}$$

and similar to Lemma C.11.10, the covering number of ϵ-net will be bounded by

$$d\log(1 + \frac{24C_\Theta(H+1)\kappa_1}{\epsilon}) + d\log\left(1 + \frac{288H^2C_\Theta(\kappa_1\sqrt{C_\Theta} + 2\sqrt{B\kappa_1\kappa_2})^2}{\epsilon^2}\right) + d^2\log\left(1 + \frac{288H^2\sqrt{d}B\kappa_1^2}{\epsilon^2}\right)$$
$$+ d\log(1 + \frac{16C_\Theta H^2\kappa_1}{\epsilon}) + d\log(1 + \frac{32C_\Theta H^3\kappa_1}{\epsilon})$$

Comparing to Lemma C.11.10, the last two terms are incurred by covering u, v arguments.

Appendix D

Assisting lemmas

Lemma D.0.1 (Multiplicative Chernoff bound [161]). *Let X be a Binomial random variable with parameter p, n. For any $1 \geq \theta > 0$, we have that*

$$\mathbb{P}[X < (1-\theta)pn] < e^{-\frac{\theta^2 pn}{2}}. \quad \text{and} \quad \mathbb{P}[X \geq (1+\theta)pn] < e^{-\frac{\theta^2 pn}{3}}$$

Lemma D.0.2 (Hoeffding's Inequality [162]). *Let $x_1, ..., x_n$ be independent bounded random variables such that $\mathbb{E}[x_i] = 0$ and $|x_i| \leq \xi_i$ with probability 1. Then for any $\epsilon > 0$ we have*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} x_i \geq \epsilon\right) \leq e^{-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}\xi_i^2}}.$$

Lemma D.0.3 (Bernstein's Inequality). *Let $x_1, ..., x_n$ be independent bounded random variables such that $\mathbb{E}[x_i] = 0$ and $|x_i| \leq \xi$ with probability 1. Let $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}[x_i]$, then with probability $1-\delta$ we have*

$$\frac{1}{n}\sum_{i=1}^{n} x_i \leq \sqrt{\frac{2\sigma^2 \cdot \log(1/\delta)}{n}} + \frac{2\xi}{3n}\log(1/\delta)$$

Lemma D.0.4 (Empirical Bernstein's Inequality [163]). *Let $x_1, ..., x_n$ be i.i.d random variables*

such that $|x_i| \leq \xi$ *with probability* 1. *Let* $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ *and* $\widehat{V}_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, *then with probability* $1-\delta$ *we have*

$$\left|\frac{1}{n}\sum_{i=1}^{n} x_i - \mathbb{E}[x]\right| \leq \sqrt{\frac{2\widehat{V}_n \cdot \log(2/\delta)}{n}} + \frac{7\xi}{3n}\log(2/\delta).$$

Lemma D.0.5 (Freedman's inequality [164])**.** *Let* X *be the martingale associated with a filter* $\mathcal{F}$ *(i.e.* $X_i = \mathbb{E}[X|\mathcal{F}_i]$*) satisfying* $|X_i - X_{i-1}| \leq M$ *for* $i = 1, ..., n$. *Denote* $W := \sum_{i=1}^{n} \mathrm{Var}(X_i|\mathcal{F}_{i-1})$ *then we have*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon, W \leq \sigma^2) \leq 2e^{-\frac{\epsilon^2}{2(\sigma^2 + M\epsilon/3)}}.$$

Or in other words, with probability $1-\delta$,

$$|X - \mathbb{E}[X]| \leq \sqrt{8\sigma^2 \cdot \log(1/\delta)} + \frac{2M}{3} \cdot \log(1/\delta), \quad Or \quad W \geq \sigma^2.$$

Lemma D.0.6 (Empirical Bernstein Inequality)**.** *Let* $n \geq 2$ *and* $V \in \mathbb{R}^S$ *be any function with* $||V||_\infty \leq H$, P *be any* S*-dimensional distribution and* $\widehat{P}$ *be its empirical version using* n *samples. Then with probability* $1-\delta$,

$$\left|\sqrt{\mathrm{Var}_{\widehat{P}}(V)} - \sqrt{\frac{n-1}{n}\mathrm{Var}_P(V)}\right| \leq 2H\sqrt{\frac{\log(2/\delta)}{n-1}}.$$

Proof: This is a directly application of Theorem 10 in [163]. Indeed, by direct translating Theorem 10 of [163],

$$V_n(V) = \frac{1}{n(n-1)}\sum_{1\leq i<j\leq n}\left(V(s_i) - V(s_j)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(V(s_i) - \overline{V})^2 = \mathrm{Var}_{\widehat{P}}(V).$$

where $s_i \sim P$ are i.i.d random variables and

$$\begin{aligned}
\mathbb{E}[V_n] =& \mathbb{E}\left[\mathrm{Var}_{\widehat{P}}(V)\right] = \mathbb{E}\left[\mathbb{E}_{\widehat{P}}[V^2] - \left(\mathbb{E}_{\widehat{P}}[V]\right)^2\right] \\
=& \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} V^2(s_i)\right] - \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} V(s_i)\right)^2\right] \\
=& \mathbb{E}\left[V^2\right] - \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n} V^2(s_i) + 2\sum_{1\le i<j\le n} V(s_i)V(s_j)\right] \\
=& \mathbb{E}\left[V^2\right] - \frac{1}{n}\mathbb{E}\left[V^2\right] - 2\frac{n(n-1)/2}{n^2}(\mathbb{E}[V])^2 \\
=& \frac{n-1}{n}\mathrm{Var}_P(V).
\end{aligned}$$

Therefore by Theorem 10 of [163] we directly have the result.

D.0.1 Extend Value Difference

The extended value difference lemma helps characterize the difference between the estimated value $\widehat{V}_1$ and the true value V_1^{π}, which was first summarized in [56] and also used in [39].

Lemma D.0.7 (Extended Value Difference (Section B.1 in [56])). *Let $\pi = \{\pi_h\}_{h=1}^H$ and $\pi' = \{\pi'_h\}_{h=1}^H$ be two arbitrary policies and let $\{\widehat{Q}_h\}_{h=1}^H$ be any given Q-functions. Then define $\widehat{V}_h(s) := \langle \widehat{Q}_h(s,\cdot), \pi_h(\cdot \mid s)\rangle$ for all $s \in \mathcal{S}$. Then for all $s \in \mathcal{S}$,*

$$\begin{aligned}
\widehat{V}_1(s) - V_1^{\pi'}(s) = &\sum_{h=1}^{H} \mathbb{E}_{\pi'}\left[\left\langle \widehat{Q}_h\left(s_h,\cdot\right), \pi_h\left(\cdot \mid s_h\right) - \pi'_h\left(\cdot \mid s_h\right)\right\rangle \mid s_1 = s\right] \\
&+ \sum_{h=1}^{H} \mathbb{E}_{\pi'}\left[\widehat{Q}_h\left(s_h,a_h\right) - \left(\mathcal{T}_h\widehat{V}_{h+1}\right)\left(s_h,a_h\right) \mid s_1 = s\right]
\end{aligned} \tag{D.1}$$

where $(\mathcal{T}_h V)(\cdot,\cdot) := r_h(\cdot,\cdot) + (P_h V)(\cdot,\cdot)$ for any $V \in \mathbb{R}^S$.

Proof:

Denote $\xi_h = \widehat{Q}_h - \mathcal{T}_h \widehat{V}_{h+1}$. For any $h \in [H]$, we have

$$\begin{aligned}
\widehat{V}_h - V_h^{\pi'} &= \langle \widehat{Q}_h, \pi_h \rangle - \langle Q_h^{\pi'}, \pi'_h \rangle \\
&= \langle \widehat{Q}_h, \pi_h - \pi'_h \rangle + \langle \widehat{Q}_h - Q_h^{\pi'}, \pi'_h \rangle \\
&= \langle \widehat{Q}_h, \pi_h - \pi'_h \rangle + \langle P_h(\widehat{V}_{h+1} - V_{h+1}^{\pi'}) + \xi_h, \pi'_h \rangle \\
&= \langle \widehat{Q}_h, \pi_h - \pi'_h \rangle + \langle P_h(\widehat{V}_{h+1} - V_{h+1}^{\pi'}), \pi'_h \rangle + \langle \xi_h, \pi'_h \rangle
\end{aligned}$$

recursively apply the above for $\widehat{V}_{h+1} - V_{h+1}^{\pi'}$ and use the $\mathbb{E}_{\pi'}$ notation (instead of the inner product of P_h, π'_h) we can finish the prove of this lemma.

The following lemma helps to characterize the gap between any two policies.

Lemma D.0.8. *Let $\widehat{\pi} = \left\{\widehat{\pi}_h\right\}_{h=1}^H$ and $\widehat{Q}_h(\cdot,\cdot)$ be the arbitrary policy and Q-function and also $\widehat{V}_h(s) = \langle \widehat{Q}_h(s,\cdot), \widehat{\pi}_h(\cdot|s) \rangle \ \forall s \in \mathcal{S}$. and $\xi_h(s,a) = (\mathcal{T}_h \widehat{V}_{h+1})(s,a) - \widehat{Q}_h(s,a)$ element-wisely. Then for any arbitrary π, we have*

$$\begin{aligned}
V_1^{\pi}(s) - V_1^{\widehat{\pi}}(s) = &\sum_{h=1}^{H} \mathbb{E}_{\pi}\left[\xi_h(s_h,a_h) \mid s_1 = s\right] - \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}\left[\xi_h(s_h,a_h) \mid s_1 = s\right] \\
&+ \sum_{h=1}^{H} \mathbb{E}_{\pi}\left[\langle \widehat{Q}_h\left(s_h,\cdot\right), \pi_h\left(\cdot|s_h\right) - \widehat{\pi}_h\left(\cdot|s_h\right)\rangle \mid s_1 = x\right]
\end{aligned}$$

where the expectation are taken over s_h, a_h.

Proof: Note the gap can be rewritten as

$$V_1^{\pi}(s) - V_1^{\widehat{\pi}}(s) = V_1^{\pi}(s) - \widehat{V}_1(s) + \widehat{V}_1(s) - V_1^{\widehat{\pi}}(s).$$

By Lemma D.0.7 with $\pi = \widehat{\pi}$, $\pi' = \pi$, we directly have

$$V_1^{\pi}(s) - \widehat{V}_1(s) = \sum_{h=1}^{H} \mathbb{E}_{\pi}\left[\xi_h(s_h, a_h) \mid s_1 = s\right] + \sum_{h=1}^{H} \mathbb{E}_{\pi}\left[\langle \widehat{Q}_h\left(s_h, \cdot\right), \pi_h\left(\cdot|s_h\right) - \widehat{\pi}_h\left(\cdot|s_h\right)\rangle \mid s_1 = s\right] \tag{D.2}$$

Next apply Lemma D.0.7 again with $\pi = \pi' = \widehat{\pi}$, we directly have

$$\widehat{V}_1(s) - V_1^{\widehat{\pi}}(s) = -\sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}\left[\xi_h(s_h, a_h) \mid s_1 = s\right]. \tag{D.3}$$

Combine the above two results we prove the stated result.

Bibliography

[1] M. Yin, *On the Statistical Complexity of Offline Policy Evaluation for Tabular Reinforcement Learning*. PhD thesis, UC Santa Barbara, 2023.

[2] M. Yin and Y.-X. Wang, *Asymptotically efficient off-policy evaluation for tabular reinforcement learning*, in *International Conference on Artificial Intelligence and Statistics*, pp. 3948–3958, PMLR, 2020.

[3] M. Yin, Y. Bai, and Y.-X. Wang, *Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning*, in *International Conference on Artificial Intelligence and Statistics*, pp. 1567–1575, PMLR, 2021.

[4] M. Yin and Y.-X. Wang, *Optimal uniform ope and model-based offline reinforcement learning in time-homogeneous, reward-free and task-agnostic settings*, *Advances in neural information processing systems* (2021).

[5] M. Yin and Y.-X. Wang, *Towards instance-optimal offline reinforcement learning with pessimism*, *Advances in neural information processing systems* (2021).

[6] M. Yin, Y. Duan, M. Wang, and Y.-X. Wang, *Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism*, *International Conference on Learning Representations,* (2022).

[7] M. Yin, M. Wang, and Y.-X. Wang, *Offline reinforcement learning with differentiable function approximation is provably efficient*, *International Conference on Learning Representations,* (2023).

[8] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, *et. al.*, *Mastering the game of go without human knowledge*, *nature* **550** (2017), no. 7676 354–359.

[9] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, *et. al.*, *Grandmaster level in starcraft ii using multi-agent reinforcement learning*, *Nature* **575** (2019), no. 7782 350–354.

[10] S. Levine, A. Kumar, G. Tucker, and J. Fu, *Offline reinforcement learning: Tutorial, review, and perspectives on open problems*, *arXiv preprint arXiv:2005.01643* (2020).

[11] S. Lange, T. Gabel, and M. Riedmiller, *Batch reinforcement learning*, in *Reinforcement learning*, pp. 45–73. Springer, 2012.

[12] H. Le, C. Voloshin, and Y. Yue, *Batch policy learning under constraints*, in *International Conference on Machine Learning*, pp. 3703–3712, 2019.

[13] J. Chen and N. Jiang, *Information-theoretic considerations in batch reinforcement learning*, in *International Conference on Machine Learning*, pp. 1042–1051, 2019.

[14] T. Xie and N. Jiang, *Q* approximation schemes for batch reinforcement learning: A theoretical comparison*, in *Uncertainty in Artificial Intelligence*, pp. 550–559, 2020.

[15] T. Xie and N. Jiang, *Batch value-function approximation with only realizability*, *arXiv preprint arXiv:2008.04990* (2020).

[16] M. Yin, Y. Bai, and Y.-X. Wang, *Near-optimal offline reinforcement learning via double variance reduction*, *Advances in Neural Information Processing Systems* (2021).

[17] T. Ren, J. Li, B. Dai, S. S. Du, and S. Sanghavi, *Nearly horizon-free offline reinforcement learning*, *Advances in neural information processing systems* (2021).

[18] P. Rashidinejad, B. Zhu, C. Ma, J. Jiao, and S. Russell, *Bridging offline reinforcement learning and imitation learning: A tale of pessimism*, *arXiv preprint arXiv:2103.12021* (2021).

[19] T. Xie, N. Jiang, H. Wang, C. Xiong, and Y. Bai, *Policy finetuning: Bridging sample-efficient offline and online reinforcement learning*, *Advances in neural information processing systems* (2021).

[20] C. Gulcehre, Z. Wang, A. Novikov, T. L. Paine, S. G. Colmenarejo, K. Zolna, R. Agarwal, J. Merel, D. Mankowitz, C. Paduraru, *et. al.*, *Rl unplugged: Benchmarks for offline reinforcement learning*, *Advances in neural information processing systems* (2020).

[21] J. Fu, A. Kumar, O. Nachum, G. Tucker, and S. Levine, *D4rl: Datasets for deep data-driven reinforcement learning*, *arXiv preprint arXiv:2004.07219* (2020).

[22] J. Fu, M. Norouzi, O. Nachum, G. Tucker, Z. Wang, A. Novikov, M. Yang, M. R. Zhang, Y. Chen, A. Kumar, *et. al.*, *Benchmarks for deep off-policy evaluation*, *International Conference on Learning Representations* (2021).

[23] M. Janner, Q. Li, and S. Levine, *Reinforcement learning as one big sequence modeling problem*, *arXiv preprint arXiv:2106.02039* (2021).

[24] C. Diehl, T. S. Sievernich, M. Krüger, F. Hoffmann, and T. Bertram, *Uncertainty-aware model-based offline reinforcement learning for automated driving*, *IEEE Robotics and Automation Letters* **8** (2023), no. 2 1167–1174.

[25] C. M. Hruschka, M. Schmidt, D. Töpfer, and S. Zug, *Uncertainty-adaptive, risk based motion planning in automated driving*, in *2019 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*, pp. 1–7, IEEE, 2019.

[26] M. Yin, W. Chen, M. Wang, and Y.-X. Wang, *Offline stochastic shortest path: Learning, evaluation and towards optimality*, *Uncertainty in Artificial Intelligence,* (2022).

[27] R. Liu, J. L. Greenstein, J. C. Fackler, J. Bergmann, M. M. Bembea, and R. L. Winslow, *Offline reinforcement learning with uncertainty for treatment strategies in sepsis*, *arXiv preprint arXiv:2107.04491* (2021).

[28] G. Gao, S. Ju, M. S. Ausin, and M. Chi, *Hope: Human-centric off-policy evaluation for e-learning and healthcare*, *arXiv preprint arXiv:2302.09212* (2023).

[29] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[30] A. Agarwal, S. Kakade, and L. F. Yang, *Model-based reinforcement learning with a generative model is minimax optimal*, in *Conference on Learning Theory*, pp. 67–83, 2020.

[31] C. Szepesvári and R. Munos, *Finite time bounds for sampling based fitted value iteration*, in *Proceedings of the 22nd international conference on Machine learning*, pp. 880–887, 2005.

[32] Y. Liu, A. Swaminathan, A. Agarwal, and E. Brunskill, *Off-policy policy gradient with state distribution correction*, in *Uncertainty in Artificial Intelligence*, 2019.

[33] A. Antos, R. Munos, and C. Szepesvari, *Fitted q-iteration in continuous action-space mdps*, in *Advances in Neural Information Processing Systems*, pp. 9–16, 2008.

[34] A. Antos, C. Szepesvári, and R. Munos, *Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path*, *Machine Learning* **71** (2008), no. 1 89–129.

[35] Y. Liu, A. Swaminathan, A. Agarwal, and E. Brunskill, *Provably good batch reinforcement learning without great exploration*, *arXiv preprint arXiv:2007.08202* (2020).

[36] J. D. Chang, M. Uehara, D. Sreenivas, R. Kidambi, and W. Sun, *Mitigating covariate shift in imitation learning via offline data without great coverage*, *Advances in Neural Information Processing Systems* (2021).

[37] R. Kidambi, A. Rajeswaran, P. Netrapalli, and T. Joachims, *Morel: Model-based offline reinforcement learning*, *Advances in Neural Information Processing Systems* (2020).

[38] M. Uehara and W. Sun, *Pessimistic model-based offline rl: Pac bounds and posterior sampling under partial coverage*, *arXiv preprint arXiv:2107.06226* (2021).

[39] Y. Jin, Z. Yang, and Z. Wang, *Is pessimism provably efficient for offline rl?*, *International Conference on Machine Learning* (2020).

[40] T. Xie, C.-A. Cheng, N. Jiang, P. Mineiro, and A. Agarwal, *Bellman-consistent pessimism for offline reinforcement learning*, *Advances in neural information processing systems* (2021).

[41] A. Zanette, M. J. Wainwright, and E. Brunskill, *Provable benefits of actor-critic methods for offline reinforcement learning*, 2021.

[42] R. Wang, D. P. Foster, and S. M. Kakade, *What are the statistical limits of offline rl with linear function approximation?*, *International Conference on Machine Learning* (2021).

[43] A. Zanette, *Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online rl*, *International Conference on Machine Learning* (2021).

[44] M. G. Azar, I. Osband, and R. Munos, *Minimax regret bounds for reinforcement learning*, in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 263–272, JMLR. org, 2017.

[45] A. Krishnamurthy, A. Agarwal, and J. Langford, *Pac reinforcement learning with rich observations*, *Advances in neural information processing systems* (2016).

[46] N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire, *Contextual decision processes with low bellman rank are pac-learnable*, in *International Conference on Machine Learning-Volume 70*, pp. 1704–1713, 2017.

[47] Z. Zhang, X. Ji, and S. S. Du, *Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon*, *arXiv preprint arXiv:2009.13503* (2020).

[48] N. Jiang and A. Agarwal, *Open problem: The dependence of sample complexity lower bounds on planning horizon*, in *Conference On Learning Theory*, pp. 3395–3398, 2018.

[49] Y. Bai, T. Xie, N. Jiang, and Y.-X. Wang, *Provably efficient q-learning with low switching cost*, in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[50] W. C. Cheung, D. Simchi-Levi, and R. Zhu, *Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism*, *arXiv preprint arXiv:2006.14389* (2020).

[51] O.-A. Maillard, T. A. Mann, and S. Mannor, *How hard is my mdp?" the distribution-norm to the rescue"*, *Advances in Neural Information Processing Systems* **27** (2014) 1835–1843.

[52] A. Zanette and E. Brunskill, *Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds*, in *International Conference on Machine Learning*, pp. 7304–7312, PMLR, 2019.

[53] Z. Wen and B. Van Roy, *Efficient exploration and value function generalization in deterministic systems*, *Advances in Neural Information Processing Systems* **26** (2013).

[54] A. Zanette and E. Brunskill, *Problem dependent reinforcement learning bounds which can identify bandit structure in mdps*, in *International Conference on Machine Learning*, pp. 5747–5755, PMLR, 2018.

[55] S. Bubeck and N. Cesa-Bianchi, *Regret analysis of stochastic and nonstochastic multi-armed bandit problems*, *Foundations and Trends in Machine Learning* (2012).

[56] Q. Cai, Z. Yang, C. Jin, and Z. Wang, *Provably efficient exploration in policy optimization*, in *International Conference on Machine Learning*, pp. 1283–1294, PMLR, 2020.

[57] C. Xiao, Y. Wu, J. Mei, B. Dai, T. Lattimore, L. Li, C. Szepesvari, and D. Schuurmans, *On the optimality of batch policy optimization algorithms*, in *International Conference on Machine Learning*, pp. 11362–11371, PMLR, 2021.

[58] G. J. Gordon, *Approximate solutions to Markov decision processes*. Carnegie Mellon University, 1999.

[59] D. Ernst, P. Geurts, and L. Wehenkel, *Tree-based batch mode reinforcement learning*, *Journal of Machine Learning Research* **6** (2005) 503–556.

[60] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et. al.*, *Human-level control through deep reinforcement learning*, *nature* **518** (2015), no. 7540 529–533.

[61] S. Fujimoto, D. Meger, and D. Precup, *Off-policy deep reinforcement learning without exploration*, in *International Conference on Machine Learning*, pp. 2052–2062, PMLR, 2019.

[62] A. Kumar, J. Fu, G. Tucker, and S. Levine, *Stabilizing off-policy q-learning via bootstrapping error reduction*, *Advances in Neural Information Processing Systems* (2019).

[63] Y. Wu, G. Tucker, and O. Nachum, *Behavior regularized offline reinforcement learning*, *arXiv preprint arXiv:1911.11361* (2019).

[64] T. Yu, G. Thomas, L. Yu, S. Ermon, J. Zou, S. Levine, C. Finn, and T. Ma, *Mopo: Model-based offline policy optimization*, *arXiv preprint arXiv:2005.13239* (2020).

[65] A. Kumar, A. Zhou, G. Tucker, and S. Levine, *Conservative q-learning for offline reinforcement learning*, *Advances in Neural Information Processing Systems* (2020).

[66] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, *Decision transformer: Reinforcement learning via sequence modeling*, *arXiv preprint arXiv:2106.01345* (2021).

[67] I. Kostrikov, A. Nair, and S. Levine, *Offline reinforcement learning with in-sample q-learning*, in *International Conference on Learning Representations*, 2022.

[68] L. Yang and M. Wang, *Sample-optimal parametric q-learning using linearly additive features*, in *International Conference on Machine Learning*, pp. 6995–7004, PMLR, 2019.

[69] L. Yang and M. Wang, *Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound*, in *International Conference on Machine Learning*, pp. 10746–10756, PMLR, 2020.

[70] A. Modi, N. Jiang, A. Tewari, and S. Singh, *Sample complexity of reinforcement learning using linearly combined model ensembles*, in *International Conference on Artificial Intelligence and Statistics*, pp. 2010–2020, PMLR, 2020.

[71] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan, *Provably efficient reinforcement learning with linear function approximation*, in *Conference on Learning Theory*, pp. 2137–2143, PMLR, 2020.

[72] A. Ayoub, Z. Jia, C. Szepesvari, M. Wang, and L. Yang, *Model-based reinforcement learning with value-targeted regression*, in *International Conference on Machine Learning*, pp. 463–474, PMLR, 2020.

[73] S. Du, A. Krishnamurthy, N. Jiang, A. Agarwal, M. Dudik, and J. Langford, *Provably efficient rl with rich observations via latent state decoding*, in *International Conference on Machine Learning*, pp. 1665–1674, PMLR, 2019.

[74] W. Sun, N. Jiang, A. Krishnamurthy, A. Agarwal, and J. Langford, *Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches*, in *Conference on learning theory*, pp. 2898–2933, PMLR, 2019.

[75] A. Zanette, A. Lazaric, M. Kochenderfer, and E. Brunskill, *Learning near optimal policies with low inherent bellman error*, in *International Conference on Machine Learning*, pp. 10978–10989, PMLR, 2020.

[76] D. Zhou, Q. Gu, and C. Szepesvari, *Nearly minimax optimal reinforcement learning for linear mixture markov decision processes*, in *Conference on Learning Theory*, pp. 4532–4576, PMLR, 2021.

[77] C. Jin, Q. Liu, and S. Miryoosefi, *Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms*, *arXiv preprint arXiv:2102.00815* (2021).

[78] S. S. Du, S. M. Kakade, J. D. Lee, S. Lovett, G. Mahajan, W. Sun, and R. Wang, *Bilinear classes: A structural framework for provable generalization in rl*, *International Conference on Machine Learning* (2021).

[79] R. Munos, *Error bounds for approximate policy iteration*, in *ICML*, vol. 3, pp. 560–567, 2003.

[80] Y. Jin, Z. Yang, and Z. Wang, *Is pessimism provably efficient for offline rl?*, in *International Conference on Machine Learning*, pp. 5084–5096, PMLR, 2021.

[81] Y. Duan, C. Jin, and Z. Li, *Risk bounds and rademacher complexity in batch reinforcement learning*, *International Conference on Machine Learning* (2021).

[82] Y. Min, T. Wang, D. Zhou, and Q. Gu, *Variance-aware off-policy evaluation with linear function approximation*, *Advances in neural information processing systems* (2021).

[83] T. Nguyen-Tang, S. Gupta, H. Tran-The, and S. Venkatesh, *On finite-sample analysis of offline reinforcement learning with deep relu networks*, *arXiv preprint arXiv:2103.06671* (2021).

[84] Y. Wang, R. Wang, S. S. Du, and A. Krishnamurthy, *Optimism in reinforcement learning with generalized linear function approximation*, in *International Conference on Learning Representations*, 2021.

[85] R. Wang, S. S. Du, L. F. Yang, and R. Salakhutdinov, *On reward-free reinforcement learning with linear function approximation*, *Advances in neural information processing systems* (2020).

[86] J. He, D. Zhou, and Q. Gu, *Logarithmic regret for reinforcement learning with linear function approximation*, in *International Conference on Machine Learning*, pp. 4171–4180, PMLR, 2021.

[87] Z. Liu, Y. Zhang, Z. Fu, Z. Yang, and Z. Wang, *Provably efficient generative adversarial imitation learning for online and offline setting with linear function approximation*, *arXiv preprint arXiv:2108.08765* (2021).

[88] L. Shi, G. Li, Y. Wei, Y. Chen, and Y. Chi, *Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity*, *arXiv preprint arXiv:2202.13890* (2022).

[89] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, *On the theory of policy gradient methods: Optimality, approximation, and distribution shift*, *Journal of Machine Learning Research* **22** (2021), no. 98 1–76.

[90] R. Wang, D. P. Foster, and S. M. Kakade, *What are the statistical limits of offline rl with linear function approximation?*, *International Conference on Learning Representations* (2021).

[91] A. Zanette, *Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online rl*, *International Conference on Machine Learning* (2021).

[92] D. J. Foster, A. Krishnamurthy, D. Simchi-Levi, and Y. Xu, *Offline reinforcement learning: Fundamental barriers for value function approximation*, *arXiv preprint arXiv:2111.10919* (2021).

[93] M. S. Talebi and O.-A. Maillard, *Variance-aware regret bounds for undiscounted reinforcement learning in mdps*, in *Algorithmic Learning Theory*, pp. 770–805, PMLR, 2018.

[94] Z. Zhang, J. Yang, X. Ji, and S. S. Du, *Variance-aware confidence set: Variance-dependent bound for linear bandits and horizon-free bound for linear mixture mdp*, *arXiv preprint arXiv:2101.12745* (2021).

[95] Y. Duan, Z. Jia, and M. Wang, *Minimax-optimal off-policy evaluation with linear function approximation*, in *International Conference on Machine Learning*, pp. 8334–8342, 2020.

[96] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[97] A. Wagenmaker, Y. Chen, M. Simchowitz, S. S. Du, and K. Jamieson, *First-order regret in reinforcement learning with linear function approximation: A robust estimation approach*, *arXiv preprint arXiv:2112.03432* (2021).

[98] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, *et. al.*, *Mastering atari, go, chess and shogi by planning with a learned model*, *Nature* **588** (2020), no. 7839 604–609.

[99] S. Gu, E. Holly, T. Lillicrap, and S. Levine, *Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates*, in *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 3389–3396, IEEE, 2017.

[100] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, *Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection*, *The International journal of robotics research* **37** (2018), no. 4-5 421–436.

[101] J. Degrave, F. Felici, J. Buchli, M. Neunert, B. Tracey, F. Carpanese, T. Ewalds, R. Hafner, A. Abdolmaleki, D. de Las Casas, *et. al.*, *Magnetic control of tokamak plasmas through deep reinforcement learning*, *Nature* **602** (2022), no. 7897 414–419.

[102] M. Mahmud, M. S. Kaiser, A. Hussain, and S. Vassanelli, *Applications of deep learning and reinforcement learning to biological data*, *IEEE transactions on neural networks and learning systems* **29** (2018), no. 6 2063–2079.

[103] M. Popova, O. Isayev, and A. Tropsha, *Deep reinforcement learning for de novo drug design*, *Science advances* **4** (2018), no. 7 eaap7885.

[104] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, *Deep reinforcement learning: A brief survey*, *IEEE Signal Processing Magazine* **34** (2017), no. 6 26–38.

[105] G. Li, Y. Chen, Y. Chi, Y. Gu, and Y. Wei, *Sample-efficient reinforcement learning is feasible for linearly realizable mdps with limited revisiting*, *Advances in Neural Information Processing Systems* **34** (2021) 16671–16685.

[106] M. Yin, Y. Duan, M. Wang, and Y.-X. Wang, *Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism*, *International Conference on Learning Representations* (2022).

[107] M. Uehara, X. Zhang, and W. Sun, *Representation learning for online and offline rl in low-rank mdps*, in *International Conference on Learning Representations*, 2022.

[108] Q. Cai, Z. Yang, and Z. Wang, *Reinforcement learning from partial observation: Linear function approximation with provable sample efficiency*, in *International Conference on Machine Learning*, pp. 2485–2522, PMLR, 2022.

[109] W. Zhan, B. Huang, A. Huang, N. Jiang, and J. D. Lee, *Offline reinforcement learning with realizability and single-policy concentrability*, *arXiv preprint arXiv:2202.04634* (2022).

[110] R. Zhang, X. Zhang, C. Ni, and M. Wang, *Off-policy fitted q-evaluation with differentiable function approximators: Z-estimation and inference theory*, *International Conference on Machine Learning* (2022).

[111] M. G. Azar, R. Munos, and H. J. Kappen, *Minimax pac bounds on the sample complexity of reinforcement learning with a generative model*, *Machine learning* **91** (2013), no. 3 325–349.

[112] A. Sidford, M. Wang, X. Wu, L. Yang, and Y. Ye, *Near-optimal time and sample complexities for solving markov decision processes with a generative model*, in *Advances in Neural Information Processing Systems*, pp. 5186–5196, 2018.

[113] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan, *Is q-learning provably efficient?*, in *Advances in Neural Information Processing Systems*, pp. 4863–4873, 2018.

[114] Q. Cui and L. F. Yang, *Is plug-in solver sample-efficient for feature-based reinforcement learning?*, in *Advances in neural information processing systems*, 2020.

[115] G. Li, Y. Wei, Y. Chi, Y. Gu, and Y. Chen, *Sample complexity of asynchronous q-learning: Sharper analysis and variance reduction*, *Advances in neural information processing systems* **33** (2020) 7031–7043.

[116] G. Li, L. Shi, Y. Chen, Y. Chi, and Y. Wei, *Settling the sample complexity of model-based offline reinforcement learning*, *arXiv preprint arXiv:2204.05275* (2022).

[117] Z. Zhang, X. Ji, and S. Du, *Horizon-free reinforcement learning in polynomial time: the power of stationary policies*, in *Conference on Learning Theory*, pp. 3858–3904, PMLR, 2022.

[118] D. Qiao, M. Yin, M. Min, and Y.-X. Wang, *Sample-efficient reinforcement learning with loglog (t) switching cost*, *International Conference on Machine Learning* (2022).

[119] Q. Cui and S. S. Du, *When is offline two-player zero-sum markov game solvable?*, *arXiv preprint arXiv:2201.03522* (2022).

[120] D. Ding, X. Wei, Z. Yang, Z. Wang, and M. Jovanovic, *Provably efficient safe exploration via primal-dual policy optimization*, in *International Conference on Artificial Intelligence and Statistics*, pp. 3304–3312, PMLR, 2021.

[121] W. Zhang, D. Zhou, and Q. Gu, *Reward-free model-based reinforcement learning with linear function approximation*, *Advances in Neural Information Processing Systems* **34** (2021) 1582–1593.

[122] D. Zhou, J. He, and Q. Gu, *Provably efficient reinforcement learning for discounted mdps with feature mapping*, in *International Conference on Machine Learning*, pp. 12793–12802, PMLR, 2021.

[123] D. Russo and B. Van Roy, *Eluder dimension and the sample complexity of optimistic exploration*, *Advances in Neural Information Processing Systems* **26** (2013).

[124] S. J. Bradtke and A. G. Barto, *Linear least-squares algorithms for temporal difference learning*, *Machine learning* **22** (1996), no. 1 33–57.

[125] J. Tsitsiklis and B. Van Roy, *Analysis of temporal-diffference learning with function approximation*, *Advances in neural information processing systems* **9** (1996).

[126] M. Riedmiller, *Neural fitted q iteration–first experiences with a data efficient neural reinforcement learning method*, in *European conference on machine learning*, pp. 317–328, Springer, 2005.

[127] J. Fan, Z. Wang, Y. Xie, and Z. Yang, *A theoretical analysis of deep q-learning*, in *Learning for Dynamics and Control*, pp. 486–489, PMLR, 2020.

[128] N. Kallus and M. Uehara, *Double reinforcement learning for efficient off-policy evaluation in markov decision processes.*, *J. Mach. Learn. Res.* **21** (2020), no. 167 1–63.

[129] Y. Wang, R. Wang, S. S. Du, and A. Krishnamurthy, *Optimism in reinforcement learning with generalized linear function approximation*, *International Conference on Learning Representations* (2021).

[130] W. Xiong, H. Zhong, C. Shi, C. Shen, L. Wang, and T. Zhang, *Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent mdp and markov game*, *arXiv preprint arXiv:2205.15512* (2022).

[131] L. Li, Y. Lu, and D. Zhou, *Provably optimal algorithms for generalized linear contextual bandits*, in *International Conference on Machine Learning*, pp. 2071–2080, PMLR, 2017.

[132] J. Schulman, X. Chen, and P. Abbeel, *Equivalence between policy gradients and soft q-learning*, *arXiv preprint arXiv:1704.06440* (2017).

[133] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, *Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor*, in *International conference on machine learning*, pp. 1861–1870, PMLR, 2018.

[134] J. Buckman, C. Gelada, and M. G. Bellemare, *The importance of pessimism in fixed-dataset policy optimization*, *arXiv preprint arXiv:2009.06799* (2020).

[135] R. Munos, *Error bounds for approximate value iteration*, in *Proceedings of the National Conference on Artificial Intelligence*, vol. 20, p. 1006, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

[136] R. Munos, *Performance bounds in l_p-norm for approximate value iteration*, *SIAM journal on control and optimization* **46** (2007), no. 2 541–561.

[137] Z. Zhang, J. Yang, X. Ji, and S. S. Du, *Improved variance-aware confidence sets for linear bandits and linear mixture mdp*, *Advances in Neural Information Processing Systems* **34** (2021).

[138] V. Mai, K. Mani, and L. Paull, *Sample efficient deep reinforcement learning via uncertainty estimation*, *International Conference on Learning Representations* (2022).

[139] Y. Wu, S. Zhai, N. Srivastava, J. Susskind, J. Zhang, R. Salakhutdinov, and H. Goh, *Uncertainty weighted actor-critic for offline reinforcement learning*, *International Conference on Machine Learning* (2021).

[140] T. Nguyen-Tang and R. Arora, *Provably efficient neural offline reinforcement learning via perturbed rewards*, .

[141] T. Xu and Y. Liang, *Provably efficient offline reinforcement learning with trajectory-wise reward*, *arXiv preprint arXiv:2206.06426* (2022).

[142] C. Jin, A. Krishnamurthy, M. Simchowitz, and T. Yu, *Reward-free exploration for reinforcement learning*, in *International Conference on Machine Learning*, pp. 4870–4879, PMLR, 2020.

[143] N. Jiang and L. Li, *Doubly robust off-policy value evaluation for reinforcement learning*, in *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pp. 652–661, JMLR. org, 2016.

[144] S. Madhow, D. Xiao, M. Yin, and Y.-X. Wang, *Offline policy evaluation for reinforcement learning with adaptively collected data*, *arXiv preprint arXiv:2306.14063* (2023).

[145] T. Nguyen-Tan, M. Yin, S. Gupta, S. Venkates, and R. Arora, *On instance-dependent bounds for offline reinforcement learning with linear function approximation*, *Association for the Advancement of Artificial Intelligence,* (2023).

[146] N. L. Kuang, M. Yin, M. Wang, Y.-X. Wang, and Y.-A. Ma, *Posterior sampling with delayed feedback for reinforcement learning with linear function approximation*, *arXiv preprint arXiv:2310.18919* (2023).

[147] D. Qiao, M. Yin, M. Min, and Y.-X. Wang, *Sample-efficient reinforcement learning with loglog(t) switching cost*, *International Conference on Machine Learning,* (2022).

[148] D. Qiao, M. Yin, and Y.-X. Wang, *Logarithmic switching cost in reinforcement learning beyond linear mdps*, *arXiv preprint arXiv:2302.12456* (2023).

[149] C. Liu, M. Yin, and Y.-X. Wang, *No-regret linear bandits beyond realizability*, *arXiv preprint arXiv:2302.13252* (2023).

[150] S. Feng, M. Yin, R. Huang, Y.-X. Wang, J. Yang, and Y. Liang, *Non-stationary reinforcement learning under general function approximation*, *arXiv preprint arXiv:2306.00861* (2023).

[151] S. Feng, M. Yin, Y.-X. Wang, J. Yang, and Y. Liang, *Model-free algorithm with improved sample efficiency for zero-sum markov games*, *arXiv preprint arXiv:2308.08858* (2023).

[152] J. Li, E. Zhang, M. Yin, Q. Bai, Y.-X. Wang, and W. Y. Wang, *Offline reinforcement learning with closed-form policy improvement operators*, *NeurIPS workshop in Offline RL,* (2022).

[153] W. Chen, M. Yin, M. Ku, E. Wan, X. Ma, J. Xu, T. Xia, X. Wang, and P. Lu, *Theoremqa: A theorem-driven question answering dataset*, *arXiv preprint arXiv:2305.12524* (2023).

[154] K. Zhang, M. Yin, and Y.-X. Wang, *Why quantization improves generalization: Ntk of binary weight neural networks*, *arXiv preprint arXiv:2206.05916* (2022).

[155] R. I. Brafman and M. Tennenholtz, *R-max-a general polynomial time algorithm for near-optimal reinforcement learning*, *Journal of Machine Learning Research* **3** (2002), no. Oct 213–231.

[156] T. Jung and P. Stone, *Gaussian processes for sample efficient reinforcement learning with rmax-like exploration*, in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 601–616, Springer, 2010.

[157] F. Chung and L. Lu, *Concentration inequalities and martingale inequalities: a survey*, *Internet Mathematics* **3** (2006), no. 1 79–127.

[158] P. D. Sampson and P. Guttorp, *Nonparametric estimation of nonstationary spatial covariance structure*, *Journal of the American Statistical Association* **87** (1992), no. 417 108–119.

[159] J. A. Tropp, *User-friendly tail bounds for sums of random matrices*, *Foundations of computational mathematics* **12** (2012), no. 4 389–434.

[160] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, *Improved algorithms for linear stochastic bandits*, in *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

[161] H. Chernoff *et. al.*, *A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations*, *The Annals of Mathematical Statistics* **23** (1952), no. 4 493–507.

[162] K. Sridharan, *A gentle introduction to concentration inequalities*, *Dept. Comput. Sci., Cornell Univ., Tech. Rep* (2002).

[163] A. Maurer and M. Pontil, *Empirical bernstein bounds and sample variance penalization*, *Conference on Learning Theory* (2009).

[164] J. Tropp *et. al.*, *Freedman's inequality for matrix martingales*, *Electronic Communications in Probability* **16** (2011) 262–270.