OFFLINE REINFORCEMENT LEARNING WITH DIFFERENTIABLE FUNCTION APPROXIMATION IS PROVABLY EFFICIENT

Ming Yin*, Mengdi Wang†, Yu-Xiang Wang*

*Department of Computer Science University of California, Santa Barbara {ming_yin, yuxiangw}@cs.ucsb.edu
†Department of Electrical and Computer Engineering Princeton University mengdiw@princeton.edu

ABSTRACT

Offline reinforcement learning, which aims at optimizing sequential decision-making strategies with historical data, has been extensively applied in real-life applications. State-Of-The-Art 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, we take a step by considering offline reinforcement learning with differentiable function class approximation (DFA). This function class naturally incorporates a wide range of models with nonlinear/nonconvex structures. 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. We hope our work could draw interest in studying reinforcement learning with differentiable function approximation beyond the scope of current research.

1 INTRODUCTION

Offline reinforcement learning (Lange et al., 2012; Levine et al., 2020) 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. Silver et al. (2017)) 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 (Silver et al., 2017; Schrittwieser et al., 2020), Robotics (Gu et al., 2017; Levine et al., 2018), energy control (Degrave et al., 2022) and Biology (Mahmud et al., 2018; Popova et al., 2018). Nevertheless, practitioners also noticed that algorithms with general function approximators can be quite data 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 (Munos, 2003; Chen and Jiang, 2019; Yang and Wang, 2019; Du et al., 2019; Sun et al., 2019; Modi et al., 2020; Jin et al., 2020b; Ayoub et al., 2020; Zanette et al., 2020; Jin et al., 2021a; Du et al., 2021; Jin et al., 2021b; Zhou et al., 2021; Xie et al., 2021a; Min et al., 2021; Nguyen-Tang et al., 2022; Li et al., 2021; Zanette et al., 2021; Yin et al., 2022; Uehara et al., 2022; Cai et al., 2022). However, most works consider linear model approximators (e.g. linear (mixture) MDPs) or its variants. While the explicit linear

\(^1\)Check Arulkumaran et al. (2017) and the references therein for an overview.
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. (Chen and Jiang, 2019; Xie et al., 2021a; Zhan et al., 2022)). Due to the generic function approximation class considered, those complexity bounds are usually expressed in the standard worst-case fashion \( O(V_{\max} \sqrt{\frac{d}{T}}) \) which lack the characterizations of individual instance behaviors. However, as mentioned in Zanette and Brunskill (2019), 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 1.1** (Parametric Differentiable Function Class). Let \( \mathcal{S}, \mathcal{A} \) be arbitrary state, action spaces and a feature map \( \phi(\cdot, \cdot) : \mathcal{S} \times \mathcal{A} \rightarrow \Psi \subset \mathbb{R}^m \). The parameter space \( \Theta \subset \mathbb{R}^d \). Both \( \Theta \) and \( \Psi \) are compact spaces. Then the parametric function class (for a model \( f : \mathbb{R}^d \times \mathbb{R}^m \rightarrow \mathbb{R} \)) is defined as

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

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

**Remark 1.2.** Differentiable Function Class was recently proposed for studying Off-Policy Evaluation (OPE) Problem (Zhang et al., 2022a) and we adopt it here for the policy learning task. Note by the compactness of \( \Theta, \Psi \) and continuity, there exists constants \( C_{\Theta}, B_{\mathcal{F}}, \kappa_1, \kappa_2, \kappa_3 > 0 \) that bounds: \( \|\theta\|_2 \leq C_{\Theta} \), \( f(\theta, \phi(s, a)) \leq B_{\mathcal{F}} \), \( \|\nabla_\theta f(\theta, \phi(s, a))\|_2 \leq \kappa_1 \), \( \|\nabla^2_{\theta \theta} f(\theta, \phi(s, a))\|_2 \leq \kappa_2 \), \( \|\nabla^3_{\theta \theta \theta} f(\theta, \phi(s, a))\|_2 \leq \kappa_3 \) for all \( \theta \in \Theta \), \( s, a \in \mathcal{S} \times \mathcal{A} \).

Why consider differentiable function class (Definition 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 (Azar et al., 2013; 2017; Sidford et al., 2018; Jin et al., 2018; Cui and Yang, 2020; Agarwal et al., 2020; Yin et al., 2021a; Li et al., 2020; Ren et al., 2021; Xie et al., 2021b; Li et al., 2022; Zhang et al., 2022b; Qiao et al., 2022; Cui and Du, 2022)) or linear MDPs (Yang and Wang, 2020; Jin et al., 2020b; Wang et al., 2020; Jin et al., 2021b; Ding et al., 2021; Wang et al., 2021a; Min et al., 2021) / linear Mixture MDPs (Modi et al., 2020; Cai et al., 2020; Zhang et al., 2021a; Zhou et al., 2021b) (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 (Zanette et al., 2020) and Eluder dimension (Russo and Van Roy, 2013; Jin et al., 2021a)). Differentiable function class strictly generalizes over the previous popular choices, i.e. by choosing \( f(\theta, \phi) = \langle \theta, \phi \rangle \) or specifying \( \phi \) 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, \( \theta \) 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 model \( f \) into Theorem 3.2 to obtain the respective bounds (we do not need separate analysis for different tasks).

1.1 Related works

Reinforcement learning with function approximation. RL with function approximation has a long history that can date back to Bradtkie and Barto (1996); Tsitsiklis and Van Roy (1996). Later,
it draws significant interest for the finite sample analysis (Jin et al., 2020b; Yang and Wang, 2019). Since then, people put tremendous efforts towards generalizing over linear function approximations and examples include Linear Bellman complete models (Zanette et al., 2020), Eluder dimension (Russo and Van Roy, 2013; Jin et al., 2021a), linear deterministic Q* (Wen and Van Roy, 2013) or Bilinear class (Du et al., 2021). 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 Wen and Van Roy (2013)). As a comparison, differentiable function class contains a range of functions that are widely used in practical algorithms (Riedmiller, 2005).

**Offline RL with general function approximation (GFA).** Another interesting thread of work considered offline RL with general function approximation (Ernst et al., 2005; Chen and Jiang, 2019; Liu et al., 2020; Xie et al., 2021a) 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. Zhang et al. (2022a) first considers the differentiable function approximation (DFA) for the off-policy evaluation (OPE) task and builds the asymptotic theory. Fan et al. (2020) analyzes the deep Q-learning with the specific ReLu activations, and Kallus and Uehara (2020) considers semi-parametric / nonparametric methods for offline RL (as opposed to our parametric DFA in 1.1). These are nice complementary studies to our work.  

**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 $\sum_{h=1}^{H} \mathbb{E}_{\pi^\theta, h} \left[ \sqrt{g_\theta(s, a)^\top \Sigma_h^{-1} g_\theta(s, a)} \right]$ where $g_\theta(s, a) := \nabla f(\theta_\pi(\phi(s, a)))$ is the gradient w.r.t. the parameter $\theta_\pi$ at feature $\phi$ 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 $\theta$. This is achieved by analyzing the pessimistic fitted Q-learning (PFQL) algorithm (Theorem 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 $F$.

## 2 Preliminaries

**Episodic Markov decision process.** Let $M = (S, A, P, r, H, d_1)$ to denote a finite-horizon Markov Decision Process (MDP), where $S$ is the arbitrary state space and $A$ is the arbitrary action space which can be infinite or continuous. The transition kernel $P_h : S \times A \mapsto \Delta S \times \Delta S$ 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 $0 \leq r \leq 1$. A policy $\pi = (\pi_1, \ldots, \pi_H)$ assigns each state $s_h \in S$ a probability distribution over actions by
mapping \( s_h \mapsto \pi_h (s_h) \) \( \forall h \in [H] \) and induces a random trajectory \( s_1, a_1, r_1, \ldots, s_H, a_H, r_H, s_{H+1} \) with \( s_1 \sim d_1, a_1 \sim \pi (s_1), s_{h+1} \sim P_h (s_h, a_h), \forall h \in [H] \).

Given a policy \( \pi \), the \( V \)-value functions and state-action value function (Q-functions) \( Q^\pi_h (s, \cdot) \in \mathbb{R}^{S \times A} \) are defined as:

\[
V^\pi_h (s) = \mathbb{E}_s \left[ \sum_{t=h}^H r_t | s_t = s \right], \quad Q^\pi_h (s, a) = \mathbb{E}_s \left[ \sum_{t=h}^H r_t | s_t = s, a_t = a \right], \quad \forall s, a, h \in S, A, [H].
\]

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

\[
Q^\pi_h (s, a) = r_h (s, a) + \int V^\pi_{h+1} (s', \cdot) dP_h (s', a, \cdot).
\]

We define Bellman operator \( \mathcal{P}_h \) for any function \( V \in \mathbb{R}^S \) as \( \mathcal{P}_h (V) = r_h + \int V dP_h \), then \( \mathcal{P}_h (V_{h+1}^{\pi}) = Q_h^{\pi} \) and \( \mathcal{P}_h (V_{h+1}^{\pi_v}) = Q_h^{\pi_v} \).

The performance measure is \( \nu^\pi := \mathbb{E}_{d_1} [V^\pi_1] = \mathbb{E}_{d_1, d_2} \left[ \sum_{t=1}^H r_t \right] \). Lastly, the induced state-action marginal occupancy measure for any \( h \in [H] \) is defined to be: for any \( E \subseteq S \times A \),

\[
d_h^\pi (E) := \mathbb{E} (s_h, a_h, r_h \in E | s_1 \sim d_1, a_1 \sim \pi (s_1), s_{i+1} \sim P_{i+1} (s_{i+1}, a_{i+1}), 1 \leq i \leq h) \text{ and } \mathbb{E}_{\pi, h} [f(s, a)] := \int_{S \times A} f(s, a) d_h^\pi (s, a) dsa.
\]

### Offline Reinforcement Learning

The goal of Offline RL is to learn the policy \( \pi^* := \arg \max_{\pi} \nu^\pi \) using only the historical data \( D = \{ (s_h, a_h, r_h, s_{h+1}) \}_{r \in [H]} \). The data generating behavior policy is denoted as \( \mu \). In the offline regime, we have neither the knowledge about \( \mu \) nor the access to further exploration for a different policy. The agent is asked to find a policy \( \hat{\pi} \) such that \( \nu^\pi - \nu^{\hat{\pi}} \leq \epsilon \) for the given batch data \( D \) and a specified accuracy \( \epsilon > 0 \).

### 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 (Foster et al., 2021). Therefore, under the differentiable function setting (Definition 1.1), we make the following assumptions.

**Assumption 2.1 (Realizability+Bellman Completeness).** The differentiable function class \( \mathcal{F} \) in Definition 1.1 satisfies:

- **Realizability**: for optimal Q^\pi_h, there exists \( \theta_h^\pi \in \Theta \) such that \( Q_h^\pi (s, \cdot) = f (\theta_h^\pi, \phi (\cdot)) \) \( \forall h \);

- **Bellman Completeness**: Let \( G := \{ V (\cdot) \in \mathbb{R}^S : \text{such that } \| V \|_\infty \leq H \} \). Then in this case \( \sup_{V \in G} \inf_{f \in \mathcal{F}} \| f - \mathcal{P}_h (V) \|_\infty \leq \epsilon_F \) for some \( \epsilon_F \geq 0 \).

Realizability and Bellman Completeness are widely adopted in the offline RL analysis with general function approximations (Chen and Jiang, 2019; Xie et al., 2021a) and Assumption 2.1 states its differentiable function approximation version. There are other forms of completeness, e.g. optimistic closure defined in Wang et al. (2021b).

**Data coverage assumption.** Furthermore, in the offline regime, it is known that function approximation cannot be sample efficient for learning a \( \epsilon \)-optimal policy without data-coverage assumptions when \( \epsilon \) is small (i.e. high accuracy) (Wang et al., 2021a). In particular, we consider two types of coverage assumptions and provide guarantees for them separately.

**Assumption 2.2 (Concentrability Coverage).** For any fixed policy \( \pi \), define the marginal state-action occupancy ratio as \( d_h^\pi (s, a) / d_h^\theta (s, a) \) \( \forall s, a \). Then the concentrability coefficient is defined as \( C_{\text{eff}} := \sup_{\pi} \sup_{h \in [H]} \| d_h^\pi (s, a) / d_h^\theta (s, a) \|_2^2 \), where \( \| g(\cdot, \cdot) \|_{2, d^\pi} := \sqrt{\mathbb{E}_{d^\pi} [g(\cdot, \cdot)^2]} \) and \( C_{\text{eff}} < \infty \).

This is the standard coverage assumption that has been widely used in (Ernst et al., 2005; Szepesvári and Munos, 2005; Chen and Jiang, 2019; Xie and Jiang, 2020a), and 2.2 is fully characterized by the MDPs. In addition, we can make an alternative assumption 2.3 that depends on both the MDPs and the function approximation class \( \mathcal{F} \). It assumes a curvature condition for \( \mathcal{F} \).

**Assumption 2.3 (Uniform Coverage).** We have \( \forall h \in [H], \text{there exists } \kappa > 0, \)

- \( E_{\mu, h} \left[ (f(\theta_1, \phi (\cdot), \cdot)) - f(\theta_2, \phi (\cdot), \cdot) \right]^2 \geq \kappa \| \theta_1 - \theta_2 \|_2^2 \), \( \forall \theta_1, \theta_2 \in \Theta ; (\star) \)

- \( E_{\mu, h} \left[ \nabla f(\theta, \phi (s, a)) \cdot \nabla f(\theta, \phi (s, a)) \right] \geq \kappa I, \forall \theta \in \Theta ; (\star \star) \)

\(^3\)Generally speaking, 2.2 and 2.3 are not directly comparable. However, for the specific function class \( f = (\theta, \phi) \) with \( \phi = 1(s, a) \) and tabular MDPs, it is easy to check 2.3 is stronger than 2.2.
In the linear function approximation regime, Assumption 2.3 reduces to 2.4 since (∗) and (**) are identical assumptions. Concretely, if \( f(\theta, \phi) = (\theta, \phi) \), then (∗) \( \mathbb{E}_{\mu,h}[(f(\theta_1, \phi(\cdot)) - f(\theta_2, \phi(\cdot)))]^2 = \mathbb{E}_{\mu,h}[(\phi(s,a) - \theta_1)\theta_2] \geq \kappa \) \( |\theta_1 - \theta_2|^2 \) \( \forall \theta_1, \theta_2 \in \Theta \) \( \supseteq \) 2.4 \( \Leftrightarrow \) (∗∗) \( \mathbb{E}_{\mu,h}[\nabla f(\theta, \phi(s,a)) \cdot \nabla f(\theta, \phi(s,a))] \). Therefore, 2.3 can be considered as a natural extension of 2.4 for differentiable class. We do point that 2.3 can be violated for function class \( \mathcal{F} \) that is “not identifiable” by the data distribution \( \mu \) (i.e., there exists \( f(\theta_1), f(\theta_2) \in \mathcal{F}, \theta_1 \neq \theta_2 \) s.t. \( \mathbb{E}_{\mu,h}[(f(\theta_1, \phi(\cdot)) - f(\theta_2, \phi(\cdot)))]^2 = 0 \). Nevertheless, there are representative non-linear differentiable classes (e.g. generalized linear model (GLM)) satisfying 2.3.

Example 2.4 (Linear function coverage assumption (Wang et al., 2021a; Min et al., 2021; Yin et al., 2022; Xiong et al., 2022)). \( \Sigma_{h,\text{feature}} := \mathbb{E}_{\mu,h}[\phi(s,a)\phi(s,a)] \geq \kappa \) \( \forall h \in [H] \) with some \( \kappa > 0 \).

Example 2.5 (offline generalized linear model (Li et al., 2017; Wang et al., 2021b)). For a known feature map \( \phi : S \times A \rightarrow \mathcal{B}_d \) and link function \( f : [-1, 1] \rightarrow [-1, 1] \) the class of GLM is \( \mathcal{F}_{\text{GLM}} := \{(s, a) \mapsto f((\phi(s,a), \theta)) : \theta \in \Theta \} \) satisfying \( \mathbb{E}_{\mu,h}[\phi(s,a)\phi(s,a)] \geq \kappa I \). Furthermore, \( f(\cdot) \) is either monotonically increasing or decreasing and 0 \( \leq \kappa \leq |f'(z)| \leq K < \infty \) \( |f''(z)| \leq M < \infty \) for all \( |z| \leq 1 \) and some \( \kappa, K, M \). Then \( \mathcal{F}_{\text{GLM}} \) satisfies 2.3, see Appendix B.

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 2), which only requires the concentrability Assumption 2.2. The algorithm is presented in Appendix I.

Theorem 3.1. Choose \( 0 < \lambda \leq 1/2C_\mathcal{F}^2 \) in Algorithm 2. Suppose Assumption 2.1,2.2. Then if \( K \geq \max \left\{ 512 \frac{\kappa^2}{\epsilon_x^2} \left( \log \left( \frac{2Hd}{\delta} \right) + d \log(1 + 4\frac{\lambda^2\kappa^2C_{\text{Gas}}K^3}{\kappa^2}) \right) + \frac{4\lambda^2}{\kappa^2} \right\}, \) with probability \( 1 - \delta \), the output \( \hat{\pi} \) of VFQL guarantees: \( v^* - \hat{v} \leq \sqrt{C_{\text{eff}}H\bar{O}} \left( \frac{H^2d + 4\lambda^2C_{\text{Gas}}K^3}{K} \right) + O(\sqrt{C_{\text{eff}}H^2c_{\mathcal{F}} + c_{\mathcal{F}}} + \epsilon_{\mathcal{F}}) \)

If the model capacity is insufficient, 3.1 will induce extra error due to the large \( \epsilon_{\mathcal{F}} \). If \( \epsilon_{\mathcal{F}} \rightarrow 0 \), the parametric rate \( \frac{1}{\sqrt{N}} \) can be recovered and similar results are derived with general function approximation (GFA) (Chen and Jiang, 2019). 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 2.3 and analyze the pessimistic fitted Q-learning (PFQL, Algorithm 1).

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 (Gordon, 1999; Ernst et al., 2005) 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 maximin objective in Xie et al. (2021a), eqn (3.2)). In particular, FQI with differentiable function approximation resembles the theoretical prototype of neural FQI algorithm (Riedmiller, 2005) and DQN algorithm (Mnih et al., 2015; Fan et al., 2020) 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. (Schulman et al., 2017; Haarnoja et al., 2018)) 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 \( D \)) may witness severe overestimation error due to the extrapolation of model \( f \) (Levine et al., 2020). 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. (Buckman et al., 2020; Kidambi et al., 2020; Jin et al., 2021b)).

---

We mention Xie et al. (2021a) has a nice practical version PSPI, but the convergence is slower (the rate \( O(n^{-\frac{3}{2}})) \).
Algorithm 1 description. Inside the backward iteration of PFQL, Fitted Q-update is performed to optimize the parameter (Line 4). \(\hat{\theta}_h\) is the root of the first-order stationarity equation \(\sum_{k=1}^{K} \left( f(\theta, \phi_h, k) - r_h, k - \hat{V}_{h+1}(s^{k}_{h+1}) \right) \cdot \nabla_\theta^T f(\theta, \phi_h, k) + \lambda \theta = 0 \) and \(\Sigma_h\) is the Gram matrix with respect to \(\nabla_\theta f_{|\theta=\hat{\theta}_h}\). Note for any \(s, a \in X \times A\), \(m(s, a) := (\nabla_\theta f(\hat{\theta}_h, \phi(s, a)))^T \Sigma^{-1} \nabla_\theta f(\hat{\theta}_h, \phi(s, a))\)^{-1} measures the effective sample size that explored \(s, a\) along the gradient direction \(\nabla_\theta f_{|\theta=\hat{\theta}_h}\) and \(\beta / \sqrt{m(s, a)}\) is the estimated uncertainty at \((s, a)\). However, the quantity \(m(s, a)\) depends on \(\hat{\theta}_h\), and \(\hat{\theta}_h\) needs to be close to the true \(\theta^*\) (i.e. \(\hat{Q}_h \approx f(\hat{\theta}_h, \phi)\)) needs to be close to \(Q^*_h\) for the uncertainty estimation \(\Gamma_h\) to be valid, since putting a random \(\theta\) into \((s, a)\) can cause an arbitrary \(\Gamma_h\) that is useless (or might even deteriorate the algorithm). Such an “implicit” constraint over \(\hat{\theta}_h\) imposes the extra difficulty for the theoretical analysis due to that general differentiable functions encode nonlinear structures. Besides, the choice of \(\beta\) is set to be \(\hat{O}(dH)\) in Theorem 3.2 and the extra term \(\hat{O}(1/K)\) in \(\Gamma_h\) is for theoretical reason only.

Algorithm 1 Pessimistic Fitted Q-Learning (PFQL)

1: Input: Offline Dataset \(\mathcal{D} = \{ (s^{k}_{h}, a^{k}_{h}, r^{k}_{h}, s^{k+1}_{h}) \}_{k,h=1}^{K,H}\) Require \(\beta\). Denote \(\phi(s, a).\)
2: Initialization: Set \(\hat{V}_{h+1}() \leftarrow 0 \) and \(\lambda > 0\).
3: for \(h = H, H-1, \ldots, 1\) do
4: \(\hat{\theta}_h \leftarrow \arg \min_{\theta \in \Theta} \left\{ \sum_{k=1}^{K} \left[ f(\theta, \phi_h, k) - r_h, k - \hat{V}_{h+1}(s^{k}_{h+1}) \right]^2 + \lambda \cdot ||\theta||^2 \right\}\)
5: \(\Sigma_h \leftarrow \sum_{k=1}^{K} \nabla_\theta f(\hat{\theta}_h, \phi_h, k) \nabla_\theta^T f(\hat{\theta}_h, \phi_h, k) + \lambda I_d.\)
6: \(\Gamma_h(\cdot, \cdot) \leftarrow \beta \left( \nabla_\theta f(\hat{\theta}_h, \phi(\cdot, \cdot)) \Sigma^{-1} \nabla_\theta f(\hat{\theta}_h, \phi(\cdot, \cdot)) \right) (\hat{\Gamma}_h(\cdot, \cdot) + \hat{O}(1/K)).\)
7: \(\hat{Q}_h(\cdot, \cdot) \leftarrow f(\hat{\theta}_h, \phi(\cdot, \cdot)) - \Gamma_h(\cdot, \cdot).\)
8: \(\pi_h(\cdot, \cdot) \leftarrow \min \left\{ \hat{Q}_h(\cdot, \cdot), H - h + 1 \right\}^+\)
9: \(\min_{\pi} \left\{ \hat{V}_h(\cdot, \cdot), \pi \in \pi_h(\cdot, \cdot) \right\}_A, \hat{V}_h(\cdot) \leftarrow \max_{s, \pi_h} \left\{ \hat{Q}_h(\cdot, \cdot), \pi_h(\cdot, \cdot) \right\}_A\)
10: end for
11: Output: \(\pi_h(\cdot, \cdot)\)

Model-Based vs. Model-Free. PFQL can be viewed as the strict generalization over the previous value iteration based algorithms, e.g. PEVI algorithm (Jin et al. (2021b)), linear MDPs and the VPVI algorithm (Yin and Wang (2021)), tabular MDPs. On one hand, approximate value iteration (AVI) algorithms (Munos, 2005) are usually model-based algorithms (for instance the tabular algorithm VPVI uses empirical model \(\hat{P}\) 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 (Munos, 2007), 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 D.E.F.

Theorem 3.2. Let \(\beta = 8dH\) and choose \(0 < \lambda \leq 1/2C_\phi^2\) in Algorithm 1. Suppose Assumption 2.1.2.3 with \(\epsilon_x = 0.5\). Then if \(K \geq \max \left\{ \frac{512d^4}{\epsilon_x^2}, \frac{\log(\frac{2Hd}{\epsilon})}{d \log(1 + \frac{4\epsilon_x^2 C_\phi K^3}{d})} \right\} \cdot \frac{d}{\epsilon_x^2}\) with probability \(1 - \delta\), for all policy \(\pi\) simultaneously, the output of PFQL guarantees

\[
v^* - v^\theta \leq \sum_{h=1}^{H} 8dH \cdot E_{\pi} \left[ \sqrt{\nabla_\theta^T \nabla_\theta f(\hat{\theta}_h, \phi(s, a_h))) \Sigma_h^{-1} \nabla_\theta f(\hat{\theta}_h, \phi(s, a_h))} \right] \cdot \epsilon + \hat{O}\left( \frac{C_{hot}}{K} \right),
\]

where \(\epsilon\) is a Polylog term and the expectation of \(\pi\) is taken over \(s_h, a_h\). In particular, if further \(K \geq \max \left\{ \hat{O}(\frac{(s^2 + \lambda^2)s^2\log^2(Hd^2)}{\epsilon_x^2}, \frac{128\epsilon_x^4 \log(2d/\delta)}{\epsilon^2} \right\} \) we have

\[
0 \leq v^\pi - v^\theta \leq \sum_{h=1}^{H} 16dH \cdot E_{\pi} \left[ \sqrt{\nabla_\theta^T \nabla_\theta f(\hat{\theta}_h^*, \phi(s, a_h))) \Sigma_h^{-1} \nabla_\theta f(\hat{\theta}_h^*, \phi(s, a_h))} \right] \cdot \epsilon + \hat{O}\left( \frac{C_{hot}}{K} \right).
\]

Here we assume model capacity is sufficient to make the presentation concise. If \(\epsilon_x > 0\), the complexity bound will include the term \(\epsilon_x\). We include more discussion in Appendix H.
Here $\Sigma_h = \sum_{k=1}^K \nabla_{\theta} f(\theta_h^*, \phi(s_h^k, a_h^k)) \nabla_{\theta}^T f(\theta_h^*, \phi(s_h^k, a_h^k)) + \lambda I_d$ and the definition of higher order parameter $C_{\text{hot}}$ can be found in List A.

**Corollary 3.3** (Offline Generalized Linear Models (GLM)). Consider the GLM function class defined in 2.5. Suppose $\beta, \lambda, K$ are defined the same as Theorem 3.2. $\epsilon_F = 0$. Then with probability $1 - \delta$, for all policy $\pi$ simultaneously, PFQL guarantees

$$v^* - v^\pi \leq \sum_{h=1}^H 8dH \cdot E_{\pi} \left[ f((\bar{\theta}_h, \phi(s_h, a_h)))^2 \cdot \phi^T(s_h, a_h) \Sigma_h^{-1} \phi(s_h, a_h) \right] + O\left(\frac{C_{\text{hot}}}{K}\right).$$

**PFQL is provably efficient.** Theorem 3.2 verifies PFQL is statistically efficient. In particular, by Lemma L.5 we have $\|\nabla_{\theta} f(\bar{\theta}_h^*, \phi)\|_{\Sigma_h^{-1}} \leq \frac{2\delta_1}{\sqrt{\kappa R}}$, resulting the main term to be bounded by $\frac{32dH^2\delta_1}{\sqrt{\kappa R}}$ that recovers the standard statistical learning convergence rate $\frac{1}{\sqrt{d}}$.

**Comparing to Jin et al. (2021b).** Theorem 3.2 strictly subsumes the linear MDP learning bound in Jin et al. (2021b). In fact, in the linear case $\nabla_{\theta} f(\theta, \phi) = \nabla_{\theta} (\theta, \phi) = \phi$ and 3.2 reduces to $O(dH \sum_{h=1}^H E_{\pi} \left[ \phi(s_h, a_h)^T \Sigma_h^{-1} \phi(s_h, a_h) \right])$.

**Instance-dependent learning.** Previous studies for offline RL with general function approximation (GFA) (Chen and Jiang, 2019; Xie and Jiang, 2020b) are more of worst-case flavors as they usually rely on the concentrability coefficient $C$. The resulting learning bounds are expressed in the form $O(V_{\max} \sqrt{E_{\pi} / n})$ 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 $\sum_{h=1}^H E_{\pi} \left[ \nabla_{\theta} f(\theta_h^*, \phi) \Sigma_h^{-1} \nabla_{\theta} f(\theta_h^*, \phi) \right]$. This Fisher-Information style quantity characterizes the learning hardness of separate problems explicitly as different for different MDP instances $M_1, M_2$, the coupled $\theta^*_h$, $\theta^*_h$ will generate different performances via the measure $\sum_{h=1}^H E_{\pi} \left[ \nabla_{\theta} f(\theta_h^*, \phi) \Sigma_h^{-1} \nabla_{\theta} f(\theta_h^*, \phi) \right]$ $(i = 1, 2)$. Standard worst-case bounds (e.g. from GFA approximation) cannot explicitly differentiate between problem instances.

**Feature representation vs. Parameters.** One interesting observation from Theorem 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 $F$ satisfies differentiability definition 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 (Uehara et al., 2022)). 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 $|f(\hat{\theta}_h, \phi) - f(\theta_h^*, \phi)|$. This can be estimated by the first order approximation $\nabla f(\hat{\theta}_h, \phi) \cdot (\hat{\theta}_h - \theta_h^*)$. However, different from the least-square value iteration (LSVI) objective (Jin et al., 2020b; 2021b), the fitted $Q$-update (Line 4, Algorithm 1) no longer admits a closed-form solution for $\hat{\theta}_h$. Instead, we can only leverage $\hat{\theta}_h$ is a stationary point of $Z_h(\theta) := \sum_{k=1}^K \left[ f(\theta, \phi, s_h^k, a_h^k) - r_h - \tilde{V}_{h+1}(s_{h+1}^k) \right] \nabla f(\theta, \phi, s_h^k, a_h^k) + \lambda \cdot \theta$ (since $Z_h(\bar{\theta}_h) = 0$). To measure the difference $\hat{\theta}_h - \theta_h^*$ for any $\theta \in \Theta$, we do the Vector Taylor expansion $Z_h(\theta) - Z_h(\hat{\theta}_h) = \Sigma_h (\theta - \hat{\theta}_h) + R_K(\theta)$ (where $R_K(\theta)$ is the higher-order residuals) at the point $\hat{\theta}_h$ with

$$\Sigma_h := \left. \frac{\partial}{\partial \theta} Z_h(\theta) \right|_{\theta = \hat{\theta}_h} = \left. \frac{\partial}{\partial \theta} \left( \sum_{k=1}^K \left[ f(\theta, \phi, s_h^k, a_h^k) - r_h - \tilde{V}_{h+1}(s_{h+1}^k) \right] \nabla f(\theta, \phi, s_h^k, a_h^k) + \lambda \cdot \theta \right) \right|_{\theta = \hat{\theta}_h}
$$

$$= \sum_{k=1}^K \left( f(\hat{\theta}_h, \phi, s_h^k, a_h^k) - r_h - \tilde{V}_{h+1}(s_{h+1}^k) \right) \nabla_{\theta} f(\hat{\theta}_h, \phi, s_h^k, a_h^k) + \sum_{k=1}^K \nabla_{\phi} f(\hat{\theta}_h, \phi, s_h^k, a_h^k) \nabla_{\theta} f(\hat{\theta}_h, \phi, s_h^k, a_h^k) + \lambda I_d \cdot \Sigma_h$$

---

Here $n$ is the number of samples used in the infinite horizon discounted setting and is similar to $K$ in the episodic setting.
The perturbation term $\Delta_{\Sigma_h^{-1}_b}$ encodes one key challenge for solving $\hat{\theta}_h - \theta^*_h$ since it breaks the positive definiteness of $\Sigma_h^{-1}_b$, and, as a result, we cannot invert the $\Sigma_h^{-1}$ in the Taylor expansion of $Z_h$. This is due to DFA (Definition 1.1) is a rich class that incorporates nonlinear curvatures. In the linear function approximation regime, this hurdle will not show up since $\nabla^2_{\theta_h} f \equiv 0$ and $\Delta_{\Sigma_h^{-1}_b}$ 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 (Zhang et al., 2022a). However, for the policy learning task, we cannot use either population quantity or the true parameter $\theta^*_h$ since we need a computable/data-based pessimism $\Gamma_h$ to make the algorithm practical.

### 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 Jin et al. (2021b), the suboptimality gap can be bounded by (Appendix D) $\sum_{h=1}^{H} 2E_\pi \left[ \Gamma_h(s_h, a_h) \right]$ if $|\langle \mathcal{P}_h \hat{V}_{h+1} - f(\hat{\theta}_h, \phi) \rangle(s, a)| \leq \Gamma_h(s, a)$. To deal with $\mathcal{P}_h \hat{V}_{h+1}$, by Assumption 2.1 we can leverage the parameter Bellman operator $T$ (Definition D.1) so that $\mathcal{P}_h \hat{V}_{h+1} = f(\theta^*_T \hat{V}_{h+1}, \phi)$. Next, we apply the second-order approximation to obtain

$$\mathcal{P}_h \hat{V}_{h+1} - f(\hat{\theta}_h, \phi) \approx \nabla^2 f(\hat{\theta}_h, \phi) \hat{\theta}_h + \frac{1}{2} \langle \nabla^2 f(\hat{\theta}_h, \phi) \nabla^2 f(\hat{\theta}_h, \phi) \rangle \hat{\theta}_h.$$  

Later, we use (1) to represent $Z_h(\theta^*_T \hat{V}_{h+1}) - Z_h(\hat{\theta}_h) = \sum_b(\theta^*_T \hat{V}_{h+1} - \hat{\theta}_b) + R_K(\theta^*_T \hat{V}_{h+1}) = \sum_b(\theta^*_T \hat{V}_{h+1} - \hat{\theta}_b) + R_K(\theta^*_T \hat{V}_{h+1})$ by denoting $R_K(\theta^*_T \hat{V}_{h+1}) = \Delta_{\Sigma_h^{-1}_b}(\hat{\theta}_b - \theta^*_T \hat{V}_{h+1}) + R_K(\theta^*_T \hat{V}_{h+1})$.

Now that $\Sigma_h^{-1}$ is invertible thus provides the estimation (note $Z_h(\hat{\theta}_h) = 0$)

$$\theta^*_T \hat{V}_{h+1} - \hat{\theta}_h = \Sigma_h^{-1} \cdot Z_h(\theta^*_T \hat{V}_{h+1}) - \Sigma_h^{-1} R_K(\theta^*_T \hat{V}_{h+1}).$$

However, to handle the higher order terms, we need the explicit finite sample bound for $||\theta^*_T \hat{V}_{h+1} - \hat{\theta}_b||_2$ (or $|\theta^*_T - \theta^*_h||$). In the OPE literature, Zhang et al. (2022a) uses asymptotic theory (Prohorovs Theorem) to show the existence of $B(\delta)$ such that $||\theta^*_T - \theta^*_h|| \leq B(\delta).$ 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 $e^{H} \log(\frac{1}{\delta})$ is an inefficient bound since $K$ needs to be $e^{H}/\delta^2$ large to guarantee $\epsilon$ accuracy.

To address this, we use a novel reduction to general function approximation (GFA) learning proposed in Chen and Jiang (2019). We first bound the loss objective $E_\mu[\ell_h(\theta^*_h)] - E_\mu[\ell_h(\theta^*_T \hat{V}_{h+1})]$ via a “orthogonal” decomposition and by solving a quadratic equation. The resulting bound can be directly used to further bound $||\theta^*_T \hat{V}_{h+1} - \hat{\theta}_h||_2$ for obtaining efficient guarantee $\tilde{O}(\frac{dT}{\sqrt{\kappa K}})$. During the course, the covering technique is applied to extend the finite function hypothesis to all the differentiable functions in 1.1. See Appendix G and Appendix D.E,F for the complete proofs.

### 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 (Talebi and Maillard, 2018; Zhou et al., 2021a) to achieve the near-minimax sample complexity and it uses estimated conditional variances $\text{Var}_{P_(\cdot)}(V_{h+1}^*)$ to reweight each training sample in the LSVI objective. Later, such a technique is leveraged by Min et al. (2021); Yin et al. (2022) to obtained the instance-dependent results. Intuitively, conditional variances $\sigma^2(s, a) := \text{Var}_{P_(\cdot)}(V_{h+1}^*)$ 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 $(s, a, r, s')$. 

\footnote{\textit{i.e.} expanding over $Z_h^*(\theta) := E_{a, s', r}(f(\theta, \phi(s, a)) - r - V_{h+1}^*(s')) \nabla f(\theta, \phi(s, a))$, and the corresponding $\Delta_{\Sigma_h^{-1}_b}$ in $\frac{\partial}{\partial \theta} Z_h(\theta)$ is zero by Bellman equation. \textit{We mention Zhang et al. (2021b) uses variance-aware confidence sets in a slightly different way.}}
rather than weighting all the samples equally. In the differentiable function approximation regime, the update is modified to
\[
\hat{\theta}_h \leftarrow \arg\min_{\theta \in \Theta} \left\{ \sum_{k=1}^{K} \left[ f(\theta, \phi, h, k) - r_{h, k} - \tilde{V}_{h+1}(s_{h+1}^k) \right] \sigma_h^2(s_{h+1}^k, a_{h+1}^k) + \lambda \cdot \| \theta \|^2 \right\}
\]
with \(\sigma_h^2(\cdot, \cdot)\) estimated by the offline data. Notably, empirical algorithms have also shown uncertainty reweighting can improve the performances for both online RL (Mai et al., 2022) and offline RL (Wu et al., 2021). These motivates our variance-aware fitted Q-learning (VAFQL) algorithm 3.

**Theorem 4.1.** Suppose Assumption 2.1.2.3 with \(c_{x, \Phi} = 0\). Let \(\beta = 8d\) and choose \(0 < \lambda \leq 1/(2C_0^2)\) in Algorithm 3. Then if \(K \geq K_0\) and \(\sqrt{d} \geq \tilde{O}(\zeta)\), with probability \(1 - \delta\), for all policy \(\pi\) simultaneously, the output of VAFQL guarantees
\[
v^* - v^\hat{\theta} \leq \sum_{h=1}^{H} 8d \cdot \frac{4d \cdot \tilde{O}(\tilde{C}_h \cdot \zeta)}{K} \cdot \| \hat{\theta} - \theta^* \|_2 + \tilde{O}(\zeta).
\]
where \(\zeta = \mathbb{E}_x f(\theta^*, \phi(\cdot)) \mathbb{E}_s f(\theta^*, \phi(s)) / (\lambda + d)\) and the \(\sigma_h^2(\cdot, \cdot)\) is a Polylog term and the expectation of \(\| \theta - \theta^* \|_2\) is bounded by \(\tilde{O}(\zeta)\).

In particular, to bound the error for \(|u_h, v_h|\), we need to define an operator \(\mathbb{D}\) that is similar to the parameter Bellman operator D.1. The full proof of Theorem 4.1 can be found in Appendix J. Comparing to Theorem 3.2, VAFQL enjoys a net improvement of the horizon dependence since \(\mathbb{V}_\mathcal{A} (V^*_h) \leq H^2\). 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), \(\sigma_h^2 \approx \mathbb{V}_\mathcal{A} V_{h+1}^*(\cdot, \cdot) \equiv 0\) (if ignore the truncation) and \(\Lambda^{-1} \rightarrow 0\). This yields a faster convergence with rate \(\tilde{O}(\frac{1}{K})\). Lastly, when reduced to linear MDPs, 4.1 recovers the results of Yin et al. (2022) except an extra \(\sqrt{d}\).

**On the statistical limits.** To complement the study, we incorporate a minimax lower bound via a reduction to Zanette et al. (2021). The following theorem reveals we cannot improve over Theorem 4.1 by more than a factor of \(\sqrt{d}\) in the most general cases. The full discussion is in K.

**Theorem 4.2** (Minimax lower bound). Specifying the model to have linear representation \(f = (\theta, \phi)\). There exist a pair of universal constants \(c, \tilde{c} > 0\) such that given dimension \(d\), horizon \(H\) and sample size \(K > c(d)\), one can always find a family of MDP instances such that for any algorithm \(\hat{\tau}\) (where \(\Lambda_h^p = \mathbb{E}\left[ \sum_{k=1}^{K} \sum_{a \in \mathcal{A}} \| \nabla f(\hat{\theta}_h^k, \phi(\cdot)) \nabla f(\hat{\theta}_h^k, \phi(s)) \|_2^2 \right] \mid \mathbb{V}_\mathcal{A} (V_{h+1}^*) \equiv 0\) ),
\[
\inf_{\hat{\tau}} \sup_{M \in \mathcal{M}} \mathbb{E}_M [v^* - v^\hat{\tau}] \geq c \sqrt{d} \cdot \sum_{h=1}^{H} \mathbb{E}_{\mathcal{Z}_h^s} \left[ \sqrt{\nabla f(\theta^*, \phi)} (\Lambda_h^p)^{-1} \nabla f(\theta^*, \phi, \cdot) \right].
\]

5 Conclusion, Limitation and Future Directions

In this work, we study offline RL with differentiable function approximation and show the sample efficient learning. We further improve the horizon dependence via a variance aware variant. However, the dependence of the parameter space still scales with \(d\) (whereas for the linear case it is \(\sqrt{d}\)), and this is due to applying covering argument for the rich class of differentiable functions. For large deep models, the parameter dimension is 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 2.3 is unknown under the current analysis. In addition, understanding the connections between the differentiable function approximation and overparameterized neural networks approximation Nguyen-Tang and Arora (2023); Xu and Liang (2022) is important. We leave these open problems as future work. 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 (Jin et al., 2020b; Wang et al., 2021b), reward-free learning (Jin et al., 2020a; Wang et al., 2020) and representation learning (Uehara et al., 2022) might provide new and unified views over the existing studies.
ACKNOWLEDGMENTS

The authors would like to thank anonymous reviewers for their helpful suggestions. Ming Yin would like to thank Chi Jin for the helpful suggestions regarding the assumption for differentiable function class and Andrea Zanette, Xuezhou Zhang for the friendly discussion. Mengdi Wang gratefully acknowledges funding from Office of Naval Research (ONR) N00014-21-1-2288, Air Force Office of Scientific Research (AFOSR) FA9550-19-1-0203, and NSF 19-589, CMMI-1653435. Ming Yin and Yu-Xiang Wang are gratefully supported by National Science Foundation (NSF) Awards #2007117 and #2003257.

REFERENCES


Appendix

A Notation List

\[ \Sigma^p_\theta (\theta) = \mathbb{E}_{\mu, h} \left[ \nabla f(\theta, \phi(s, a)) \cdot \nabla f(\theta, \phi(s, a))^T \right] \]
\[ \kappa = \min_{\theta, \phi} \lambda_{\min}(\Sigma^p_\theta (\theta)) \]
\[ \sigma^2_\theta (s, a) = \max\{1, \text{Var}_{\theta, \phi} (V) (s, a)\} \text{ for any } V \]
\[ \delta = \text{Failure probability} \]
\[ K_0 = \max \left\{ \frac{512}{3\kappa^4} \left( \log(\frac{2Hd}{\delta}) + d \log(1 + \frac{4\kappa_1^2 \kappa_2 C_\theta K^3}{\lambda^2}) \right), \frac{4\lambda}{\kappa} \right\} \]
\[ \zeta = 2 \max_{s' \sim P(\cdot|s, a), h \in [H]} \left( \frac{\lambda_{\min}(\Sigma^p_\theta (\theta))}{\sigma^2_{\theta, \phi}(s, a)} \right) \]
\[ C_{\text{hot}} = \bar{C}_{\text{hot}} \]
\[ C_{\text{hot}}' = \bar{C}_{\text{hot}}' \]

B Further Illustration that Generalized Linear Model Example satisfies 2.3

Recall the definition in 2.5, then:
For (\ast\ast),
\[ \mathbb{E}_{\mu, h} \left[ \nabla f(\theta, \phi(s, a)) \cdot \nabla f(\theta, \phi(s, a))^T \right] = \mathbb{E}_{\mu, h} \left[ f'(\theta, \phi(s, a))^2 \phi(\cdot, \cdot) \cdot \phi(\cdot, \cdot)^T \right] \]
\[ \succ \kappa^2 \mathbb{E}_{\mu, h} \left[ \phi(\cdot, \cdot) \cdot \phi(\cdot, \cdot)^T \right] \succ \kappa^3 I, \quad \forall \theta \in \Theta \]
For (\ast), by Taylor’s Theorem,
\[ \mathbb{E}_{\mu, h} \left[ (f(\theta_1, \phi(\cdot, \cdot)) - f(\theta_2, \phi(\cdot, \cdot)))^2 \right] = \mathbb{E}_{\mu, h} \left[ f'(\theta_{s, a}) \phi(\cdot, \cdot)^2 (\theta_1 - \theta_2)^T \phi(\cdot, \cdot) \phi(\cdot, \cdot)^T (\theta_1 - \theta_2) \right] \]
\[ \geq \kappa^2 \mathbb{E}_{\mu, h} \left[ (\theta_1 - \theta_2)^T \phi(\cdot, \cdot) \phi(\cdot, \cdot)^T (\theta_1 - \theta_2) \right] = \kappa^2 (\theta_1 - \theta_2)^T \mathbb{E}_{\mu, h} \left[ \phi(\cdot, \cdot) \phi(\cdot, \cdot)^T \right] (\theta_1 - \theta_2) \geq \kappa^3 \|\theta_1 - \theta_2\|_2^2 \]
and choose \( \kappa^3 \) as \( \kappa \) in 2.3.

C On the computational complexity

For storage of Pessimistic Fitted Q-learning, at each time step \( h \in [H] \) in Algorithm 1, we need to store \( \hat{\theta}_h, \Sigma_h \) and \( \nabla f(\hat{\theta}_h, \phi, h) \). Therefore, the total space complexity is \( O(dH + d^2H + dKH) \). For computation, assuming \( \hat{\theta}_h \) is solved via SGD and let \( M \) denote the number of gradient steps, then the complexity is dominated by computing \( \hat{\theta}_h, \Sigma_h \) and \( \Sigma_h^{-1} \), which results in \( O(MH + Kh + d^2H) \) 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.

D Some basic constructions

First of all, Recall in the first-order condition, we have
\[ \nabla_{\theta} \left\{ \sum_{k=1}^K \left[ f(\theta, \phi, h, k) - \hat{r}_{h, k} - \hat{V}_{h+1}(s^k_{h+1}) \right] + \lambda \cdot \|\theta\|_2^2 \right\} \bigg|_{\theta = \hat{\theta}_h} = 0, \quad \forall h \in [H]. \]
Therefore, if we define the quantity $Z_h(\cdot, \cdot) \in \mathbb{R}^d$ as

$$Z_h(\theta|V) = \sum_{k=1}^{K} \left[ f(\theta, \phi_{h,k}) - r_{h,k} - V(s_{h+1}^k) \right] \nabla f(\theta, \phi_{h,k}) + \lambda \cdot \theta, \quad \forall \theta \in \Theta, \|V\|_2 \leq H,$$

then we have (recall $\widehat{\theta}_h \in \text{Int}(\Theta)$)

$$Z_h(\widehat{\theta}_h|\hat{V}_{h+1}) = 0.$$

In addition, according to Bellman completeness Assumption 2.1, for any bounded $V(\cdot) \in \mathbb{R}^S$ with $\|V\|_\infty \leq H$, $\inf f \in \mathcal{F} \|f - \mathcal{P}_h(V)\|_\infty \leq \epsilon_f$, $\forall h$ (recall $\mathcal{P}_h(V) = r_h + \int_S VdP_h$). Therefore, we can define the parameter Bellman operator $\mathbb{T}$ as follows.

**Definition D.1** (parameter Bellman operator). By the Bellman completeness Assumption 2.1, for any $\|V\|_\infty \leq H$, we can define the parameter Bellman operator $\mathbb{T} : V \to \theta_{TV} \in \Theta$ such that

$$\theta_{TV} = \arg \min_{\theta \in \Theta} \|f(\theta, \phi) - \mathcal{P}_h(V)\|_\infty.$$ 

Denote $\delta_V := f(\theta_{TV}, \phi) - \mathcal{P}_h(V)$, then we have $\|f(\theta_{TV}, \phi) - \mathcal{P}_h(V)\|_\infty = \|\delta_V\|_\infty \leq \epsilon_f$. In particular, by realizability Assumption 2.1 it holds $\theta_{TV}^{h+1} = \theta_h^*$ and this is due to $f(\theta_{TV}^{h+1}, \phi) = \mathcal{P}_h(V_h^{*+1}) = V_h^* = f(\theta_h^*, \phi)$.

### D.1 Suboptimality decomposition

Denote $\nu_h(s, a) := \mathcal{P}_h \hat{V}_{h+1}(s, a) - \hat{Q}_h(s, a)$, by Jin et al. (2021b) we have the following decomposition.

**Lemma D.2** (Lemma 3.1 of Jin et al. (2021b)). Let $\tilde{\pi} = \{\tilde{\pi}_h\}_{h=1}^{H}$ a policy and $\hat{Q}_h$ be any estimates with $\hat{V}_h = (\hat{Q}_h(s, \cdot), \tilde{\pi}_h(\cdot | s))_A$. Then for any policy $\pi$, we have

$$v^\pi - v^{\tilde{\pi}} = - \sum_{h=1}^{H} E_\pi [\nu_h(s_h, a_h)] + \sum_{h=1}^{H} E_\pi [\nu_h(s_h, a_h)] + \sum_{h=1}^{H} E_\pi [(\hat{Q}_h(s_h, \cdot), \pi_h(\cdot | s_h) - \tilde{\pi}_h(\cdot | s_h))_A].$$

In particular, if we choose $\tilde{\pi}_h(\cdot|s) := \arg \max_{\pi}(\hat{Q}_h(s, \cdot), \pi(\cdot | s))_A$, then

$$v^\pi - v^{\tilde{\pi}} = - \sum_{h=1}^{H} E_\pi [\nu_h(s_h, a_h)] + \sum_{h=1}^{H} E_\pi [\nu_h(s_h, a_h)].$$

**Lemma D.3.** Let $\hat{P}_h$ be the general estimated Bellman operator. Suppose with probability $1 - \delta$, it holds for all $h, s, a \in [H] \times S \times A$ that $|(\mathcal{P}_h \hat{V}_{h+1} - \hat{P}_h \hat{V}_{h+1})(s, a)| \leq \Gamma_h(s, a)$, then it implies $\forall s, a, h \in S \times A \times [H]$, $0 \leq \zeta_h(s, a) \leq 2\Gamma_h(s, a)$. Furthermore, it holds for any policy $\pi$ simultaneously, with probability $1 - \delta$,

$$V^\pi_1(s) - V^{\tilde{\pi}}_1(s) \leq \sum_{h=1}^{H} 2 \cdot \mathbb{E}_\pi [\Gamma_h(s_h, a_h) \mid s_1 = s].$$

**Proof of Lemma D.3.** This is a generic result that holds true for the general MDPs and was first raised by Theorem 4.2 of Jin et al. (2021b). Later, it is summarized in Lemma C.1 of Yin et al. (2022).

With Lemma D.3, we need to bound the term $|\mathcal{P}_h \hat{V}_{h+1}(s, a) - \hat{P}_h \hat{V}_{h+1}(s, a)|$.

---

*Here without loss of generality we assume $Q_h^*$ can be uniquely identified, i.e. there is a unique $\theta^*$ such that $f(\theta^*_h, \phi) = Q_h^*$.*
E  ANALYZING $|\mathcal{P}_h \hat{V}_{h+1}(s, a) - \mathcal{P}_h \hat{V}_{h+1}(s, a)|$ FOR PFQL.

Throughout this section, we suppose $\epsilon_F = 0$, i.e. $f(\theta_{TV}, \phi) = \mathcal{P}_h(V)$. According to the regression oracle (Line 4 of Algorithm 1), the estimated Bellman operator $\mathcal{P}_h$ maps $\hat{V}_{h+1}$ to $\hat{\theta}_h$, i.e. $\mathcal{P}_h \hat{V}_{h+1} = f(\hat{\theta}_h, \phi)$. Therefore (recall Definition D1)

$$\mathcal{P}_h \hat{V}_{h+1}(s, a) - \mathcal{P}_h \hat{V}_{h+1}(s, a) = \mathcal{P}_h \hat{V}_{h+1}(s, a) - f(\hat{\theta}_h, \phi(s, a))$$

(3)

where we apply the first-order Taylor expansion for the differentiable function $f$ at point $\hat{\theta}_h$ and $\text{Hot}_{h, 1}$ is a higher-order term. Indeed, the following Lemma E.1 bounds the $\text{Hot}_{h, 1}$ term with $O(1/K)$.

**Lemma E.1.** Recall the definition (from the above decomposition) $\text{Hot}_{h, 1} := f(\theta_{TV_{h+1}}, \phi(s, a)) - f(\hat{\theta}_h, \phi(s, a)) - \nabla f(\hat{\theta}_h, \phi(s, a)) \left( \theta_{TV_{h+1}} - \hat{\theta}_h \right)$, then with probability $1 - \delta$,

$$|\text{Hot}_{h, 1}| \leq \frac{18H^2\kappa_2(\log(H/\delta) + C_d \log K) + \kappa_2 C_\phi^2}{\kappa K}, \forall h \in [H].$$

**Proof of Lemma E.1.** By second-order Taylor’s Theorem, there exists a point $\xi$ (lies in the line segment of $\hat{\theta}_h$ and $\theta_{TV_{h+1}}$) such that

$$f(\theta_{TV_{h+1}}, \phi(s, a)) - f(\hat{\theta}_h, \phi(s, a)) = \nabla f(\hat{\theta}_h, \phi(s, a))^\top \left( \theta_{TV_{h+1}} - \hat{\theta}_h \right) + \frac{1}{2} \left( \theta_{TV_{h+1}} - \hat{\theta}_h \right)^\top \nabla^2 f(\xi, \phi(s, a)) \left( \theta_{TV_{h+1}} - \hat{\theta}_h \right)$$

Therefore, by directly applying Theorem G.2, with probability $1 - \delta$, for all $h \in [H],$

$$|\text{Hot}_{h, 1}| \leq \frac{1}{2} \kappa_2 \left\| \left( \theta_{TV_{h+1}} - \hat{\theta}_h \right) \right\|_2^2 \leq \frac{18H^2\kappa_2(\log(H/\delta) + C_d \log K) + \kappa_2 C_\phi^2}{\kappa K}$$

$\square$

## E.1  ANALYZING $\nabla f(\hat{\theta}_h, \phi(s, a)) \left( \theta_{TV_{h+1}} - \hat{\theta}_h \right)$ VIA $Z_h$.

From (3) and Lemma E.1, the problem further reduces to bounding $\nabla f(\hat{\theta}_h, \phi(s, a)) \left( \theta_{TV_{h+1}} - \hat{\theta}_h \right)$.

To begin with, we first provide a characterization of $\theta_{TV_{h+1}} - \hat{\theta}_h$. Indeed, by first-order Vector Taylor expansion (Lemma L.1), we have (note $Z_h(\hat{\theta}_h|\hat{V}_{h+1}) = 0$) for any $\theta \in \Theta$,

$$Z_h(\theta|\hat{V}_{h+1}) - Z_h(\hat{\theta}_h|\hat{V}_{h+1}) = \Sigma^*_h(\theta - \hat{\theta}_h) + R_K(\theta),$$

(4)

where $R_K(\theta)$ is the higher-order residuals and $\Sigma^*_h := \frac{\partial}{\partial \theta} Z_h(\theta|\hat{V}_{h+1}) |_{\theta = \hat{\theta}_h}$ with

$$\Sigma^*_h := \frac{\partial}{\partial \theta} Z_h(\theta|\hat{V}_{h+1}) \bigg|_{\theta = \hat{\theta}_h} = \frac{\partial}{\partial \theta} \left( \sum_{k=1}^K \left[ f(\theta, \phi_{h,k}) - r_{h,k} - \hat{V}_{h+1}(s_{h+1}^k) \right] \nabla f(\theta, \phi_{h,k}) + \lambda \cdot \theta \right) \bigg|_{\theta = \hat{\theta}_h}$$

$$= \sum_{k=1}^K \left( f(\hat{\theta}_h, \phi_{h,k}) - r_{h,k} - \hat{V}_{h+1}(s_{h+1}^k) \right) \cdot \nabla^2 f(\hat{\theta}_h, \phi_{h,k}) =: \Delta^*_h$$

and

$$\sum_{k=1}^K \nabla^2 f(\hat{\theta}_h, \phi_{h,k}) \nabla f(\hat{\theta}_h, \phi_{h,k}) + \lambda I_d, =: \Sigma_h$$

(5)
Here $\nabla^2 = \nabla \otimes \nabla$ denotes outer product of gradients.

Note $\Delta \Sigma^*_h$ is not desirable since it could prevent $\Sigma^*_h$ from being positive-definite (and it could cause $\Sigma^*_h$ to be singular). Therefore, we first deal with $\Delta \Sigma^*_h$ in below.

**Lemma E.2.** With probability $1 - \delta$, for all $h \in [H]$,

$$
\frac{1}{K} \| \Delta \Sigma^*_h \|_2 = \frac{1}{K} \sum_{k=1}^{K} \left( f(\hat{\theta}_h, \phi_h, k) - r_h, k - \hat{V}_{h+1}(s^k_{h+1}) \right) \cdot \nabla^2_{\theta \theta} f(\hat{\theta}_h, \phi_h) \\
\leq 9\kappa_2 \max \left( \frac{\kappa_1}{\sqrt{K}}, 1 \right) \sqrt{\frac{dH^2(\log(2H/\delta) + d\log(1 + 2C_{d,H}K_\kappa_3K) + C_{d,\log K})}{K}} + \frac{1}{K}.
$$

**Proof of Lemma E.2.** **Step1:** We prove for fixed $\bar{\theta} \in \Theta$, with probability $1 - \delta$, for all $h \in [H]$,

$$\left\| \frac{1}{K} \sum_{k=1}^{K} \left( f(\hat{\theta}_h, \phi_h, k) - r_h, k - \hat{V}_{h+1}(s^k_{h+1}) \right) \cdot \nabla^2_{\theta \theta} f(\bar{\theta}, \phi_h) \right\|_2 \leq 9\kappa_2 \max \left( \frac{\kappa_1}{\sqrt{K}}, 1 \right) \sqrt{\frac{H^2(\log(2H/\delta) + C_{d,\log K})}{K}}.$$

Indeed, we have

$$
\left\| \frac{1}{K} \sum_{k=1}^{K} \left( f(\hat{\theta}_h, \phi_h, k) - r_h, k - \hat{V}_{h+1}(s^k_{h+1}) \right) \cdot \nabla^2_{\theta \theta} f(\bar{\theta}, \phi_h) \right\|_2 \\
\leq \left\| \frac{1}{K} \sum_{k=1}^{K} \left( f(\bar{\theta}, \phi_h, k) - f(\bar{\theta}, \phi_h, k) \right) \cdot \nabla^2_{\theta \theta} f(\bar{\theta}, \phi_h) \right\|_2 \\
+ \left\| \frac{1}{K} \sum_{k=1}^{K} \left( f(\bar{\theta}, \phi_h, k) - r_h, k - \hat{V}_{h+1}(s^k_{h+1}) \right) \cdot \nabla^2_{\theta \theta} f(\bar{\theta}, \phi_h) \right\|_2.
$$

On one hand, by Theorem G.2 with probability $1 - \delta/2$ for all $h \in [H]$,

$$
\left\| \frac{1}{K} \sum_{k=1}^{K} \left( f(\hat{\theta}_h, \phi_h, k) - f(\bar{\theta}, \phi_h, k) \right) \cdot \nabla^2_{\theta \theta} f(\bar{\theta}, \phi_h) \right\|_2 \leq \kappa_2 \max_{\theta, \phi, a} \| \nabla f(\theta, \phi, a) \|_2 \| \hat{\theta}_h - \bar{\theta} \|_2 \\
\leq \kappa_2 \kappa_1 \| \hat{\theta}_h - \bar{\theta} \|_2 \leq \kappa_2 \kappa_1 \left( \sqrt{\frac{36H^2(\log(H/\delta) + C_{d,\log K}) + 2\lambda C_{\lambda}^2}{\kappa K}} + \sqrt{\frac{b_{d,K,T,F}}{K}} + \sqrt{\frac{2H \epsilon_F}{K}} \right).
$$

On other hand, recall the definition of $\bar{\tau}$, we have

$$
E \left[ \left( f(\bar{\theta}, \phi_h, k) - r_h, k - \hat{V}_{h+1}(s^k_{h+1}) \right) \cdot \nabla^2_{\theta \theta} f(\bar{\theta}, \phi_h) \right] = \left( P_h \bar{V}_{h+1}(s^k_{h+1}) \right) (s^k_{h+1}, a^k_h) \cdot \nabla^2_{\theta \theta} f(\bar{\theta}, \phi_h) = 0.
$$

Also, since $\left\| \left( f(\bar{\theta}, \phi_h, k) - r_h, k - \hat{V}_{h+1}(s^k_{h+1}) \right) \cdot \nabla^2_{\theta \theta} f(\bar{\theta}, \phi_h) \right\|_2 \leq H \kappa_2$, denote $\sigma^2 := K \cdot H^2 \kappa_2$, then by Vector Hoeffding’s inequality (Lemma L.2),

$$
P \left( \left\| \frac{1}{K} \sum_{k=1}^{K} \left( f(\bar{\theta}, \phi_h, k) - r_h, k - \hat{V}_{h+1}(s^k_{h+1}) \right) \cdot \nabla^2_{\theta \theta} f(\bar{\theta}, \phi_h) \right\|_2 \geq \frac{t}{K} \| \left( s^k_{h+1}, a^k_h \right) \|_1 \right) \leq d e^{-t^2/8dKH^2 \kappa_2^2} := \delta
$$

which is equivalent to

$$
P \left( \left\| \frac{1}{K} \sum_{k=1}^{K} \left( f(\bar{\theta}, \phi_h, k) - r_h, k - \hat{V}_{h+1}(s^k_{h+1}) \right) \cdot \nabla^2_{\theta \theta} f(\bar{\theta}, \phi_h) \right\|_2 \leq \sqrt{\frac{8dH^2 \kappa_2^2 \log(4/d)}{K}} \| \left( s^k_{h+1}, a^k_h \right) \|_1 \right) \geq 1 - \delta
$$
Define \( A = \left\{ \frac{1}{K} \sum_{k=1}^{K} \left( f(\theta_{\hat{V}_{h+1}}(k), \phi_h, k) - r_{\hat{V}_{h+1}}(s_{h+1}^k) \right) \cdot \nabla^2_{\theta\theta} f(\theta, \phi)_h \right\} \leq \sqrt{\frac{8dH^2\kappa_2^2 \log(d/\delta)}{K}}, \) then by law of total expectation \( \mathbb{P}(A) = \mathbb{E}[1_A] = \mathbb{E}[\mathbb{E}[1_A|\{s_h^k, a_h^k\}_{k=1}^{K}]] = \mathbb{E}[\mathbb{P}[A|\{s_h^k, a_h^k\}_{k=1}^{K}]] \geq \mathbb{E}[1 - \delta] = 1 - \delta, \) i.e. with probability at least \( 1 - \delta/2 \) (and a union bound),

\[
\left\| \frac{1}{K} \sum_{k=1}^{K} \left( f(\theta_{\hat{V}_{h+1}}(k), \phi_h, k) - r_{\hat{V}_{h+1}}(s_{h+1}^k) \right) \cdot \nabla^2_{\theta\theta} f(\theta, \phi)_h \right\|_2 \leq \sqrt{\frac{8dH^2\kappa_2^2 \log(2dH/\delta)}{K}}, \forall h \in [H].
\]

Using above and equation 6, equation 7 and a union bound, w.p. \( 1 - \delta \), for all \( h \in [H] \),

\[
\left\| \frac{1}{K} \sum_{k=1}^{K} \left( f(\theta_h, \phi_h, k) - r_{\hat{V}_{h+1}}(s_{h+1}^k) \right) \cdot \nabla^2_{\theta\theta} f(\theta, \phi)_h \right\|_2 \leq 6\kappa_2 \kappa_1 \sqrt{\frac{H^2(\log(2H/\delta) + C_{d,log} K)}{K}} + \sqrt{\frac{8dH^2\kappa_2^2 \log(2dH/\delta)}{K}} \leq 9\kappa_2 \max(\frac{\kappa_1}{\sqrt{K}}, 1) \sqrt{\frac{dH^2(\log(2H/\delta) + C_{d,log} K)}{K}}.
\]

**Step2:** we finish the proof of the lemma.

Consider the function class \( \left\{ f(\bar{\theta}) := \left\| \frac{1}{K} \sum_{k=1}^{K} \left( f(\hat{\theta}_h, \phi_h, k) - r_{\hat{V}_{h+1}}(s_{h+1}^k) \right) \cdot \nabla^2_{\theta\theta} f(\hat{\theta}, \phi)_h \right\|_2, \bar{\theta} \in \Theta \right\} \),

then by triangular inequality

\[
|f(\bar{\theta}_1) - f(\bar{\theta}_2)| \leq \left\| \frac{1}{K} \sum_{k=1}^{K} \left( f(\hat{\theta}_h, \phi_h, k) - r_{\hat{V}_{h+1}}(s_{h+1}^k) \right) \cdot \nabla^2_{\theta\theta} f(\hat{\theta}_1, \phi_h) - \nabla^2_{\theta\theta} f(\hat{\theta}_2, \phi_h) \right\|_2 \leq H \cdot \sup_{\theta, \phi} \left\| \nabla^2_{\theta\theta} f(\hat{\theta}_1, \phi_h) - \nabla^2_{\theta\theta} f(\hat{\theta}_2, \phi_h) \right\|_2 \leq H \kappa_3 \left\| \hat{\theta}_1 - \hat{\theta}_2 \right\|_2.
\]

By Lemma L.8, the covering number \( C \) of the \( \epsilon \)-net of the above function class satisfies \( \log C \leq d \log(1 + \frac{2C_{\epsilon} d H^2}{\epsilon}) \). By choosing \( \epsilon = 1/K \), by a union bound over \( C \) cases we obtain for all \( h \in [H] \)

\[
\left\| \frac{1}{K} \sum_{k=1}^{K} \left( f(\hat{\theta}_h, \phi_h, k) - r_{\hat{V}_{h+1}}(s_{h+1}^k) \right) \cdot \nabla^2_{\theta\theta} f(\hat{\theta}_h, \phi_h)_h \right\|_2 \leq 9\kappa_2 \max(\frac{\kappa_1}{\sqrt{K}}, 1) \sqrt{\frac{dH^2(\log(2H/\delta) + d \log(1 + 2C_{\epsilon} H \kappa_3 K) + C_{d,log} K)}{K}} + \frac{1}{K}.
\]

\( \square \)

Combining Lemma E.2 and Theorem G.2 (and a union bound), we directly have

**Corollary E.3.** With probability \( 1 - \delta \),

\[
\left\| \frac{1}{K} \Delta_{\hat{V}_{h+1}}(\hat{\theta}_h - \theta_{\hat{V}_{h+1}}) \right\|_2 \leq \left\| \frac{1}{K} \Delta_{\hat{V}_{h}}(\hat{\theta}_h - \theta_{\hat{V}_{h+1}}) \right\|_2 \leq O \left( \frac{\kappa_2 \max(\frac{\kappa_1}{\kappa}, \frac{1}{\sqrt{K}}) d^2 H^2}{K} \right).
\]

Here \( \tilde{O} \) absorbs all the constants and Polylog terms.

Now we select \( \hat{\theta} = \theta_{\hat{V}_{h+1}} \) in equation 4, and denote \( \bar{R}_{\hat{V}_{h+1}}(\theta_{\hat{V}_{h+1}}) = \Delta_{\hat{V}_{h+1}}(\hat{\theta}_h - \theta_{\hat{V}_{h+1}}) + R_{\hat{V}_{h+1}}(\theta_{\hat{V}_{h+1}}) \), then equation 4 is equivalent to

\[
Z_h(\theta_{\hat{V}_{h+1}}|\hat{V}_{h+1}) \equiv Z_h(\theta_{\hat{V}_{h+1}}|\hat{V}_{h+1}) = \Sigma_h(\theta_{\hat{V}_{h+1}} - \hat{\theta}_h) + R_{\hat{V}_{h+1}}(\theta_{\hat{V}_{h+1}}) = \Sigma_h(\theta_{\hat{V}_{h+1}} - \hat{\theta}_h) + \bar{R}_{\hat{V}_{h+1}}(\theta_{\hat{V}_{h+1}}).
\]

Note \( \lambda > 0 \) implies \( \Sigma_h \) is invertible, then we have (recall \( Z_h(\hat{\theta}_h|\hat{V}_{h+1}) = 0 \))

\[
\theta_{\hat{V}_{h+1}} - \hat{\theta}_h = \Sigma^{-1}_h \left[Z_h(\theta_{\hat{V}_{h+1}}|\hat{V}_{h+1}) - Z_h(\hat{\theta}_h|\hat{V}_{h+1})\right] - \Sigma^{-1}_h \bar{R}_{\hat{V}_{h+1}}(\theta_{\hat{V}_{h+1}})
\]

\[
= \Sigma^{-1}_h \left[Z_h(\theta_{\hat{V}_{h+1}}|\hat{V}_{h+1}) - Z_h(\theta_{\hat{V}_{h+1}}|\hat{V}_{h+1})\right] - \Sigma^{-1}_h \bar{R}_{\hat{V}_{h+1}}(\theta_{\hat{V}_{h+1}}).
\]
Plug it back to equation 3 to get
\[ \nabla f(\hat{\theta}, \phi(s, a)) \left( \tilde{\Theta}_{\hat{\mathcal{T}}(\phi)} \right) = \nabla f(\hat{\theta}, \phi(s, a)) \sum_{k=1}^{K} \left( f(\theta_{TV}, \phi, k) - r_{h, k} - \hat{V}_{h+1}(s_{h+1}) \right) \cdot \nabla_{\hat{\theta}} f(\theta_{TV}, \phi, k) \]
\[ = \nabla f(\hat{\theta}, \phi(s, a)) \sum_{k=1}^{K} \left( f(\theta_{TV}, \phi, k) - r_{h, k} - \hat{V}_{h+1}(s_{h+1}) \right) \cdot \nabla_{\hat{\theta}} f(\theta_{TV}, \phi, k) \]
\[ = \nabla f(\hat{\theta}, \phi(s, a)) \sum_{h=1}^{H} \left( \tilde{R}_K(\theta_{TV}, \phi) + \lambda \theta_{TV} \right) \]
\[ := I \]

We will bound second term Hot 2 to have higher order \( O\left( \frac{1}{K} \right) \) in Section E.5 and focus on the first term. By direct decomposition,

\[ I := \nabla f(\hat{\theta}, \phi(s, a)) \sum_{h=1}^{H} \left( \tilde{R}_K(\theta_{TV}, \phi) + \lambda \theta_{TV} \right) \]
\[ := I_1 \]
\[ + \nabla f(\hat{\theta}, \phi(s, a)) \sum_{h=1}^{H} \left( \tilde{R}_K(\theta_{TV}, \phi) + \lambda \theta_{TV} \right) \]
\[ := I_2 \]
\[ + \nabla f(\hat{\theta}, \phi(s, a)) \sum_{h=1}^{H} \left( \tilde{R}_K(\theta_{TV}, \phi) + \lambda \theta_{TV} \right) \]
\[ := I_3 \]

E.2 Bounding the term I3

We first bound the term I3. We have the following Lemma.

**Lemma E.4.** For any fixed \( V(\cdot) \in \mathbb{R}^S \) with \( \|V\|_{\infty} \leq H \) and any fixed \( \theta \) such that \( \|\theta_{TV} - \theta\|_2 \leq \sqrt{\frac{36H^2(log(H/\delta) + C_{d, log} \cdot K + 2C_{d, 2}^2)}{\kappa K}} \). Let

\[ I_3 := \nabla f(\hat{\theta}, \phi(s, a)) \sum_{h=1}^{H} \left( f(\theta_{TV}, \phi, k) - r_{h, k} - V(s_{h+1}) \right) \cdot \nabla_{\hat{\theta}} f(\theta_{TV}, \phi, k) \]

and if \( K \geq \max \left\{ 512 \frac{d}{\kappa^2} \left( \log \left( \frac{22}{\delta} \right) + d \log(1 + \frac{4x_1D^2K^3}{\kappa^2}) \right), \frac{4M}{\kappa} \right\} \), then with probability \( 1 - \delta \),

where \( D = \max\{\kappa_1, \sqrt{\frac{144dH^2H^2}{\kappa^4} + C_{d, log} \cdot K + 8dH^2 \kappa^2 \lambda C_{d, 2}^2 \log(d/\delta)} \} \)

\[ |I_3| \leq 4\kappa_1 \left( \frac{144dH^2H^2}{\kappa^4} + C_{d, log} \cdot K + 8dH^2 \kappa^2 \lambda C_{d, 2}^2 \log(d/\delta) \right) \left( \frac{1}{K} \right) + O\left( \frac{1}{K^{3/2}} \right). \]
Proof of Lemma E.4. Indeed, with probability $1 - \delta/2$,

$$|\tilde{I}_3| = \left\| \nabla f(\tilde{\theta}_h, \phi(s,a)) - \sum_{k=1}^{K} \left( f(\theta^T V, \phi_{h,k}) - r_{h,k} - V(s_{h+1}^k) \right) \cdot \nabla \theta f(\theta^T V, \phi_{h,k}) - \nabla \theta f(\phi_{h,k}) \right\|$$

$$\leq \left\| \nabla f(\tilde{\theta}_h, \phi(s,a)) \right\|_{\Sigma_h^{-1}} \sum_{k=1}^{K} \left( f(\theta^T V, \phi_{h,k}) - r_{h,k} - V(s_{h+1}^k) \right) \cdot \nabla \theta f(\theta^T V, \phi_{h,k}) - \nabla \theta f(\phi_{h,k}) \right\|_{\Sigma_h^{-1}}$$

$$\leq \left( \frac{2\kappa_1}{\sqrt{\kappa K}} + O\left( \frac{1}{K} \right) \right) \left\| \sum_{k=1}^{K} \left( f(\theta^T V, \phi_{h,k}) - r_{h,k} - V(s_{h+1}^k) \right) \cdot \nabla \theta f(\theta^T V, \phi_{h,k}) - \nabla \theta f(\phi_{h,k}) \right\|_{\Sigma_h^{-1}}$$

where, under the condition $K \geq \max \left\{ \frac{512 \kappa^2}{\delta^2} \left( \log(\frac{d}{\gamma}) + d \log(1 + \frac{4\kappa_1^2 \kappa_2 \kappa K}{\Delta^2}) \right), \frac{4\lambda}{\pi} \right\}$, we applied Lemma L.5.

Next, on one hand, $\left\| \nabla \phi f(\theta^T V, \phi_{h,k}) - \nabla \theta f(\theta, \phi_{h,k}) \right\|_2 \leq \kappa_2 \cdot \left\| \theta^T V - \theta \right\|_2 \leq \kappa_2 \sqrt{\frac{36H^2(\log(H/\delta) + C_{d, \log K}) + 2\kappa C_2^2}{\kappa K}}$. On the other hand,

$$E \left[ \left( f(\theta^T V, \phi_{h,k}) - r_{h,k} - V(s_{h+1}^k) \right) \cdot \nabla \phi f(\theta^T V, \phi_{h,k}) - \nabla \theta f(\theta, \phi_{h,k}) \right]$$

$$= E \left[ \left( f(\theta^T V, \phi_{h,k}) - r_{h,k} - V(s_{h+1}^k) \right) \cdot \nabla \phi f(\theta^T V, \phi_{h,k}) - \nabla \theta f(\theta, \phi_{h,k}) \right]$$

$$= (P_h V)(s_{h+1}^k, a_{h+1}^k) \cdot \nabla \phi f(\theta^T V, \phi_{h,k}) - \nabla \theta f(\theta, \phi_{h,k}) = 0$$

Therefore by Vector Hoeffding’s inequality (Lemma L.2) (also note the condition for boundedness $\left\| \left( f(\theta^T V, \phi_{h,k}) - r_{h,k} - V(s_{h+1}^k) \right) \cdot \nabla \phi f(\theta^T V, \phi_{h,k}) - \nabla \theta f(\theta, \phi_{h,k}) \right\|_2 \leq H_{K^2} \cdot \left\| \theta^T V - \theta \right\|_2 \leq \frac{36H^2(\log(H/\delta) + C_{d, \log K}) + 2\kappa C_2^2}{\kappa K}$) with probability $1 - \delta/2$,

$$\left\| \frac{1}{K} \sum_{k=1}^{K} \left( f(\theta^T V, \phi_{h,k}) - r_{h,k} - V(s_{h+1}^k) \right) \cdot \nabla \phi f(\theta^T V, \phi_{h,k}) - \nabla \theta f(\theta, \phi_{h,k}) \right\|_2 \leq \sqrt{\frac{4d}{K} \left( \frac{36H^2(\log(H/\delta) + C_{d, \log K}) + 2\kappa C_2^2}{\kappa K} \right)^2 \log(d/\gamma)}$$

and this implies with probability $1 - \delta/2$,

$$\left\| \sum_{k=1}^{K} \left( f(\theta^T V, \phi_{h,k}) - r_{h,k} - V(s_{h+1}^k) \right) \cdot \nabla \phi f(\theta^T V, \phi_{h,k}) - \nabla \theta f(\theta, \phi_{h,k}) \right\|_2 \leq \sqrt{\frac{144dH^2 \kappa_2^2 (H^2 \log(H/\delta) + C_{d, \log K}) + 8dH^2 \kappa_2^2 \kappa C^2_2 \log(d/\gamma)}{K}}$$

choose $u = \sum_{k=1}^{K} \left( f(\theta^T V, \phi_{h,k}) - r_{h,k} - V(s_{h+1}^k) \right) \cdot \nabla \phi f(\theta^T V, \phi_{h,k}) - \nabla \theta f(\theta, \phi_{h,k})$ in Lemma L.5, by a union bound we obtain with probability $1 - \delta$,

$$|\tilde{I}_3| = \left\| \nabla f(\tilde{\theta}_h, \phi(s,a)) \right\|_{\Sigma_h^{-1}} \sum_{k=1}^{K} \left( f(\theta^T V, \phi_{h,k}) - r_{h,k} - V(s_{h+1}^k) \right) \cdot \nabla \theta f(\theta^T V, \phi_{h,k}) - \nabla \theta f(\phi_{h,k}) \right\|$$

$$\leq \left( \frac{2\kappa_1}{\sqrt{\kappa K}} + O\left( \frac{1}{K} \right) \right) \left\| \sum_{k=1}^{K} \left( f(\theta^T V, \phi_{h,k}) - r_{h,k} - V(s_{h+1}^k) \right) \cdot \nabla \theta f(\theta^T V, \phi_{h,k}) - \nabla \theta f(\phi_{h,k}) \right\|_{\Sigma_h^{-1}}$$

$$\leq \left( \frac{2\kappa_1}{\sqrt{\kappa K}} + O\left( \frac{1}{K} \right) \right) \left( 2 \sqrt{\frac{144dH^2 \kappa_2^2 (H^2 \log(H/\delta) + C_{d, \log K}) + 8dH^2 \kappa_2^2 \kappa C^2_2 \log(d/\gamma)}{K^3 K}} + O\left( \frac{1}{K^{3/2}} \right) \right)$$

$$= 4\kappa_1 \sqrt{\frac{144dH^2 \kappa_2^2 (H^2 \log(H/\delta) + C_{d, \log K}) + 8dH^2 \kappa_2^2 \kappa C^2_2 \log(d/\gamma)}{K^3}} + O\left( \frac{1}{K^{3/2}} \right).$$
Lemma E.5. Under the same condition as Lemma E.4. With probability $1 - \delta$, 
\[ |I_3| \leq 4\kappa_1 \sqrt{\left(144dH^2\kappa_2^4 (H^2 \log(H/\delta) + D_{d,\log K} + C_{d,\log K}) + 8dH^2\kappa_2^2 \Lambda C_{\Theta}^2 \right) \log(d/\delta) + D_{d,\log K}} \frac{1}{K} + O\left(\frac{1}{K^{3/2}}\right). \]

Here \( D_{d,\log K} := \frac{d \cdot \log(1 + 6C_\Theta (2\kappa_1^2 + H\kappa_2) K)}{K} + d \log(1 + 6C_\Theta H\kappa_2 K) + d \log(1 + 288C_\Theta \kappa_1^2 (\kappa_1 \sqrt{C_{\Theta} + 2\sqrt{B\kappa_1 \kappa_2}}) K^2) + d^2 \log \left(1 + 288\sqrt{dB\kappa_1^4 K^2}\right) = \widetilde{O}(d^2) \) with \( \widetilde{O} \) absorbs Polylog terms.

Proof of Lemma E.5. Define 
\[ h(V, \tilde{\theta}, \theta) = \sum_{k=1}^{K} \left( f(\tilde{\theta}, \phi_{h,k}) - r_{h,k} - V(s_{h+1}^k) \right) \cdot \left( \nabla_{\theta} f(\tilde{\theta}, \phi_{h,k}) - \nabla_{\theta} f(\theta, \phi_{h,k}) \right), \]
then
\[ |h(V_1, \tilde{\theta}_1, \theta_1) - h(V_2, \tilde{\theta}_2, \theta_2)| \]
\[ \leq \sum_{k=1}^{K} \left| \left( f(\tilde{\theta}_1, \phi_{h,k}) - V_1(s_{h+1}) \right) - \left( f(\tilde{\theta}_2, \phi_{h,k}) - V_2(s_{h+1}) \right) \right| \cdot \left| \left( \nabla_{\theta} f(\tilde{\theta}_1, \phi_{h,k}) - \nabla_{\theta} f(\theta_1, \phi_{h,k}) \right) \right| \\
+ \sum_{k=1}^{K} \left| f(\tilde{\theta}_2, \phi_{h,k}) - r_{h,k} - V_2(s_{h+1}) \right| \cdot \left| \left( \nabla_{\theta} f(\tilde{\theta}_2, \phi_{h,k}) - \nabla_{\theta} f(\theta_2, \phi_{h,k}) \right) \right| \\
\leq K \sup_{s,a,a'} \left| f(\tilde{\theta}_1, \phi(s, a)) - f(\tilde{\theta}_2, \phi(s, a)) \right| - \left| V_1(s') - V_2(s') \right| \cdot 2\kappa_1 \\
+ KH \cdot \sup_{s,a,a'} \left| \nabla_{\theta} f(\tilde{\theta}_1, \phi(s, a)) - \nabla_{\theta} f(\theta_1, \phi(s, a)) \right| - \left| \nabla_{\theta} f(\tilde{\theta}_2, \phi(s, a)) - \nabla_{\theta} f(\theta_2, \phi(s, a)) \right| \left| V_1(s') - V_2(s') \right| \\
\leq K \kappa_1^2 \left| \tilde{\theta}_1 - \tilde{\theta}_2 \right| + 2K \kappa_1 \left| V_1 - V_2 \right|_\infty + KH \kappa_2 \left| \tilde{\theta}_1 - \tilde{\theta}_2 \right|_2 + KH \kappa_2 \left| \theta_1 - \theta_2 \right|_2 \\
=(2\kappa_1^4 + H\kappa_2) K \left| \tilde{\theta}_1 - \tilde{\theta}_2 \right|_2 + 2K \kappa_1 \left| V_1 - V_2 \right|_\infty + KH \kappa_2 \left| \theta_1 - \theta_2 \right|_2. 
\]

Let \( C_\Theta \) be the \( \frac{\epsilon/3}{(2\kappa_1^2 + H\kappa_2) K} \) covering net of \( \{ \theta : \| \theta \|_2 \leq C_\Theta \} \), \( C_V \) be the \( \frac{\epsilon}{6\kappa_1 K} \) covering net of \( V \) defined in Lemma L.9 and \( C_\Theta \) be the \( \frac{\epsilon}{3\kappa_2 K} \) covering net of \( \{ \theta : \| \theta \|_2 \leq C_\Theta \} \), then by Lemma L.8 and Lemma L.9,
\[ \log |C_\Theta| \leq d \cdot \log(1 + \frac{6C_\Theta (2\kappa_1^2 + H\kappa_2) K}{\epsilon}), \quad \log |C_V| \leq d \cdot \log(1 + \frac{6C_\Theta H\kappa_2 K}{\epsilon}) \]
\[ \log C_V \leq d \cdot \log \left(1 + \frac{288C_\Theta \kappa_1^2 (\kappa_1 \sqrt{C_{\Theta} + 2\sqrt{B\kappa_1 \kappa_2}}) K^2}{\epsilon^2} \right) + d^2 \log \left(1 + \frac{288\sqrt{dB\kappa_1^4 K^2}}{\epsilon^2} \right). \]

Further notice with probability \( 1 - \delta/2 \) (by Lemma L.5), for all fixed sets of parameters \( \theta, V \) satisfies 
\[ \| \theta_{TV} - \theta \|_2 \leq \sqrt{\frac{36dH^2 (\log(2H/\delta) + C_{d,\log K} + 2\Lambda C_{\Theta}^2)}{K}} \]
simultaneously,
\[ |I_3 - \tilde{I}_3| \leq \left| \nabla_{\theta} f(\tilde{\theta}_h, \phi(s, a)) \right|_{\Sigma_{h+1}^1} \cdot \left| h(\tilde{V}_{h+1}, \theta_{TV_{h+1}}, \tilde{\theta}_h) - h(V, \theta_{TV}, \theta) \right|_{\Sigma_{h+1}^1} \]
\[ \leq \left( \frac{2\kappa_1^2}{\sqrt{K} \kappa} + O\left(\frac{1}{K}\right) \right) \cdot \left| h(\tilde{V}_{h+1}, \theta_{TV_{h+1}}, \tilde{\theta}_h) - h(V, \theta_{TV}, \theta) \right|_{\Sigma_{h+1}^1} \]
and 
\[ \| \theta_{TV_{h+1}} - \tilde{\theta}_h \|_2 \leq \sqrt{\frac{36dH^2 (\log(2H/\delta) + C_{d,\log K} + 2\Lambda C_{\Theta}^2)}{K}} \]
with probability \( 1 - \delta/2 \) by Theorem G.2.

Now, choosing \( \epsilon = O(1/K^2) \) and by Lemma E.4 and union bound over covering instances, we obtain with probability \( 1 - \delta \)
\[ |I_3| \leq 4\kappa_1 \sqrt{\left(144dH^2\kappa_2^4 (H^2 \log(H/\delta) + D_{d,\log K} + C_{d,\log K}) + 8dH^2\kappa_2^2 \Lambda C_{\Theta}^2 \right) \log(d/\delta) + D_{d,\log K}} \frac{1}{K} + O\left(\frac{1}{K^{3/2}}\right). \]
E.3  BOUNDING THE SECOND TERM $I_2$

In this section, we bound the term

$$I_2 := \nabla f(\hat{\theta}, \phi(s, a)) \Sigma_0^{-1} \sum_{k=1}^{K} \left( f(\theta_{TV_h^*, \phi, h, k}) - f(\theta_{TV_{h+1}^*, \phi, h, k}) - V_h^*(s_{h+1}^k) + V_{h+1}^*(s_{h+1}^k) \right) \cdot \nabla^\top f(\hat{\theta}, \phi, h, k).$$

The following Lemma shows that $I_2$ is a higher-order error term with rate $O(1/K)$.

Lemma E.6 (Bounding $I_2$). If $K$ satisfies $K \geq 512 \frac{\kappa^2}{\kappa^4} \left( \log\left( \frac{2d}{\delta} \right) + d \log(1 + \frac{4c_1 c_2 C a K}{\kappa^2}) \right)$, and $K \geq 4\lambda/K$, then with probability $1 - \delta$

$$|I_2| \leq \tilde{O}\left( \frac{\kappa^2 H^2 d^2}{K^2} + \frac{1}{K^{3/2}} \right).$$

Here $O$ absorbs constants and Polylog terms.

Proof of Lemma E.6. **Step 1.** Define $\eta_k(V) := f(\theta_{TV}, \phi, h, k) - f(\theta_{TV_{h+1}^*, \phi, h, k}) - V(s_{h+1}^k) + V_{h+1}^*(s_{h+1}^k)$ and let $\|V(\cdot)\|_{\infty} \leq H$ be any fixed function such that $\sup_{V_{h+1}} \eta_k(V) \leq \tilde{O}(\sqrt{\kappa H^2 \frac{d^2}{K\kappa}})$, i.e. arbitrary fixed $V$ function in the neighborhood (measured by $\eta_k$) of $V_{h+1}^*$. Then by definition of $T$ it holds $E[\eta_k(V, \theta)]|_{s_{h+1}^k, \phi_{h+1}^k} = 0$. Let the fixed $\theta \in \Theta$ be arbitrary and define $x_k(\theta) = \nabla f(\theta, \phi_{h+1}^k)$. Next, define $G_h(\theta) = \sum_{k=1}^{K} \nabla f(\theta, \phi(s_{h+1}^k)) \cdot \nabla f(\theta, \phi(s_{h+1}^k))^\top + \lambda I_d$, since $\|x_k\|_{2} \leq \kappa_1$ and $|\eta_k| \leq \tilde{O}(\kappa_1 H^2 \frac{d^2}{K\kappa})$, by self-normalized Hoeffding’s inequality (Lemma L.3), with probability $1 - \delta$ (recall $t := K$ in Lemma L.3),

$$\left\| \sum_{k=1}^{K} x_k(\theta) \eta_k(V) \right\|_{\mathcal{G}_h(\theta)^{-1}} \leq \tilde{O}(\kappa_1 H^2 \sqrt{\frac{d^2}{K\kappa}}) \sqrt{d \log \left( \frac{\lambda + \kappa_1}{\lambda \kappa_1} \right)}.$$

**Step 2.** Define $h(V, \theta) := \sum_{k=1}^{K} x_k(\theta) \eta_k(V)$ and $H(V, \theta) := \left\| \sum_{k=1}^{K} x_k(\theta) \eta_k(V) \right\|_{\mathcal{G}_h(\theta)^{-1}}$, then note by definition $|\eta_k(V)| \leq 2H$, which implies $h(V, \theta) \leq 2KH_1$ and

$$\|\eta_k(V_1) - \eta_k(V_2)\| \leq \|A h_1 - A h_2\| + \|V_1 - V_2\|_{\infty} \leq 2 \|V_1 - V_2\|_{\infty}$$

and

$$\|h(V_1, \theta_1) - h(V_2, \theta_2)\| \leq \sum_{k=1}^{K} \|x_k(\theta_1) - x_k(\theta_2)\| = K \max_{k} (2H \|x_k(\theta_1) - x_k(\theta_2)\|_{2} + K_1 \|\eta_k(V_1) - \eta_k(V_2)\|_{2})$$

$$\leq K(2H \kappa_2 \|\theta_1 - \theta_2\|_{2} + 2K_1 \|V_1 - V_2\|_{\infty}).$$

Furthermore,

$$\|G_h(\theta_1)^{-1} - G_h(\theta_2)^{-1}\|_{2} \leq \|G_h(\theta_1)^{-1}\|_{2} \|G_h(\theta_1) - G_h(\theta_2)\|_{2} \|G_h(\theta_2)^{-1}\|_{2}$$

$$\leq \frac{1}{\lambda^2} K \sup_{k} \left\| \nabla f(\theta_{1}, \phi_{h+1}^k) \cdot \nabla f(\theta_{1}, \phi_{h+1}^k)^\top - \nabla f(\theta_{2}, \phi_{h+1}^k) \cdot \nabla f(\theta_{2}, \phi_{h+1}^k)^\top \right\|_{2}$$

$$\leq \frac{1}{\lambda^2} K \sup_{k} \left[ \left\| \nabla f(\theta_{1}, \phi_{h+1}^k) \cdot \nabla f(\theta_{1}, \phi_{h+1}^k)^\top \right\|_{2} \cdot \left\| \nabla f(\theta_{2}, \phi_{h+1}^k) \cdot \nabla f(\theta_{2}, \phi_{h+1}^k)^\top \right\|_{2} \right]$$

$$\leq \frac{\kappa_2 K}{\lambda^2} \|\theta_1 - \theta_2\|_{2} \cdot \frac{2K_2}{\lambda^2} \|\theta_1 - \theta_2\|_{2}.$$

All the above imply

$$\|H(V_1, \theta_1) - H(V_2, \theta_2)\| \leq \sqrt{\|h(V_1, \theta_1)^\top G_h(\theta_1)^{-1}h(V_1, \theta_1) - h(V_2, \theta_2)^\top G_h(\theta_2)^{-1}h(V_2, \theta_2)\|}$$

$$\leq \sqrt{\|h(V_1, \theta_1) - h(V_2, \theta_2)\|_2 \cdot \frac{1}{\lambda} \cdot 2KH_1} + \sqrt{2KH_1 \cdot \|G_h(\theta_1)^{-1} - G_h(\theta_2)^{-1}\|_2 \cdot 2KH_1}$$

$$+ \sqrt{2KH_1 \cdot \frac{1}{\lambda} \cdot \|h(V_1, \theta_1) - h(V_2, \theta_2)\|_2}$$

$$\leq \sqrt{2K(2H_2 \kappa \|\theta_1 - \theta_2\|_{2} + 2K_1 \|V_1 - V_2\|_{\infty}) \cdot \frac{1}{\lambda} \cdot 2KH_1} + \sqrt{2KH_1 \cdot \frac{2\kappa_2 K}{\lambda^2} \|\theta_1 - \theta_2\|_{2} \cdot 2KH_1}$$

$$\leq \sqrt{2K^3 H^2 \frac{\kappa_1 \kappa_2}{\lambda} + 8K^3 H^2 \frac{\kappa_1 \kappa_2}{\lambda^2} \cdot \frac{1}{\lambda} \cdot 2KH_1} + \sqrt{2KH_1 \cdot \frac{2\kappa_2 K}{\lambda^2} \|\theta_1 - \theta_2\|_{2} \cdot 2KH_1}$$

$$\leq \sqrt{2K^3 H^2 \frac{\kappa_1 \kappa_2}{\lambda} + 8K^3 H^2 \frac{\kappa_1 \kappa_2}{\lambda^2} \cdot \frac{1}{\lambda} \cdot 2KH_1} + \sqrt{2KH_1 \cdot \frac{2\kappa_2 K}{\lambda^2} \|\theta_1 - \theta_2\|_{2} \cdot 2KH_1}$$

$$\leq \sqrt{2K^3 H^2 \frac{\kappa_1 \kappa_2}{\lambda} + 8K^3 H^2 \frac{\kappa_1 \kappa_2}{\lambda^2} \cdot \frac{1}{\lambda} \cdot 2KH_1} + \sqrt{2KH_1 \cdot \frac{2\kappa_2 K}{\lambda^2} \|\theta_1 - \theta_2\|_{2} \cdot 2KH_1}$$

$$\leq \sqrt{2K^3 H^2 \frac{\kappa_1 \kappa_2}{\lambda} + 8K^3 H^2 \frac{\kappa_1 \kappa_2}{\lambda^2} \cdot \frac{1}{\lambda} \cdot 2KH_1} + \sqrt{2KH_1 \cdot \frac{2\kappa_2 K}{\lambda^2} \|\theta_1 - \theta_2\|_{2} \cdot 2KH_1}$$
Then a \( c \)-covering net of \( \{H(V, \theta)\} \) can be constructed by the union of 
\[
\frac{4(4\sqrt{K^3H^2\kappa_1\kappa_2} + \sqrt{8K^3H^2\kappa_1\kappa_2} + \epsilon^2)}{4(4\sqrt{K^3H^2\kappa_1\kappa_2} + \sqrt{8K^3H^2\kappa_1\kappa_2} + \epsilon^2)}
\]
covering net of \( \{\theta \in \Theta\} \) and 
\[
\frac{5K}{4(4\sqrt{K^3H^2\kappa_1\kappa_2} + \sqrt{8K^3H^2\kappa_1\kappa_2} + \epsilon^2)}
\]
covering net of \( V \) in Lemma L.9. The covering number \( N_c \) satisfies
\[
\log N_c \leq d\log \left( 1 + \frac{8C\theta(4\sqrt{K^3H^2\kappa_1\kappa_2} + \sqrt{8K^3H^2\kappa_1\kappa_2} + \epsilon^2)}{\epsilon^2} \right) + d\log \left( 1 + \frac{8\sqrt{dK}\epsilon}{16(4\sqrt{K^3H^2\kappa_1\kappa_2} + \epsilon^2)} \right).
\]

Step3. First note by definition in Step2
\[
\left\| \sum_{k=1}^{K} \left( f(\theta_V\Phi_k) - f(\theta_{V^*}\Phi_k) \right) - V_{h+1}(s_{h+1}^k) + V_{h+1}^*(s_{h+1}^k) \right\|_{\Sigma_h^{-1}} \leq \sum_{k=1}^{K} \left\| \frac{\partial}{\partial \theta} f(\tilde{\theta}_h, \Phi_k) \right\|_{\Sigma_h^{-1}} \leq O\left( \frac{K\lambda^2}{\sqrt{\kappa K}} + O\left( \frac{1}{K} \right) \right).
\]

where the second inequality uses \( \theta_{TV^*} = \theta^* \) and the third inequality uses Theorem G.2 and Theorem G.3. The last equal sign is due to \( C_{d,log}K \leq O(d^2) \) (recall Lemma G.1).

Now choosing \( \epsilon = O(1/K) \) in Step2 and union bound over both equation 9 and covering number in Step2, we obtain with probability \( 1 - \delta \),
\[
H(\tilde{V}_{h+1}, \tilde{\theta}_h) = \sum_{k=1}^{K} x_k(\tilde{\theta}_h)\eta_k(\tilde{V}_{h+1}) \leq O(\kappa_1 H^2 \sqrt{\frac{d^2}{\kappa K}} + \frac{1}{K} \sqrt{\kappa K})
\]
where we absorb all the Polylog terms. Meanwhile, by Lemma L.5 with probability \( 1 - \delta \),
\[
\left\| \nabla f(\tilde{\theta}_h, \Phi_s) \right\|_{\Sigma_h^{-1}} \leq \frac{2\kappa_1}{\sqrt{\kappa K}} + O\left( \frac{1}{K} \right).
\]

Finally, by equation 10 and equation 11 and a union bound, we have with probability \( 1 - \delta \),
\[
|I_2| := \left\| \frac{\partial}{\partial \theta} f(\tilde{\theta}_h, \Phi_s) \right\|_{\Sigma_h^{-1}} \sum_{k=1}^{K} \left( f(\theta_{V^*}\Phi_k) - f(\theta_{TV^*}\Phi_k) \right) \cdot \left\| \nabla f(\tilde{\theta}_h, \Phi_k) \right\|_{\Sigma_h^{-1}} \leq \frac{2\kappa_1}{\sqrt{\kappa K}} + O\left( \frac{1}{K} \right) \cdot \frac{\sqrt{\kappa K}}{\sqrt{\kappa K}} = O\left( \frac{K^2\lambda^2d^2}{\kappa K} \right)
\]
where the first inequality is Cauchy-Schwarz inequality.

E.4 Bounding the main term \( I_1 \)

In this section, we bound the dominate term
\[
I_1 := \left\| \nabla f(\tilde{\theta}_h, \Phi_s) \right\|_{\Sigma_h^{-1}} \sum_{k=1}^{K} \left( f(\theta_{TV^*}\Phi_k) - r_{h,k} - V_{h+1}^*(s_{h+1}^k) \right) \cdot \left\| \nabla f(\tilde{\theta}_h, \Phi_k) \right\|_{\Sigma_h^{-1}}.
\]
First of all, by Cauchy-Schwarz inequality, we have
\[
|I_1| \leq \left\| \nabla f(\hat{\theta}_h, \phi(s, a)) \right\|_{\Sigma_h^{-1}} \cdot \sum_{k=1}^{K} \left( f(\theta^*_T V_{k+1}^*, \phi_h, k) - r_{h,k} - V_{h+1}^*(s_{h+1}^k) \right) \cdot \nabla^\top f(\hat{\theta}_h, \phi_h, k) \right\|_{\Sigma_h^{-1}}. 
\]
(12)

Then we have the following Lemma to bound $I_1$.

**Lemma E.7.** With probability $1 - \delta$, 
\[
|I_1| \leq 4HD \left\| \nabla f(\hat{\theta}_h, \phi(s, a)) \right\|_{\Sigma_h^{-1}} \cdot C_{\delta, \log K} + \tilde{O}\left(\frac{K_1}{\sqrt{\kappa K}}\right),
\]
where $C_{\delta, \log K}$ only contains Polylog terms.

**Proof of Lemma E.7.** **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 $\|x_k\|_2 \leq \kappa_1$. Also denote $\eta_k := f(\theta^*_T V_{k+1}^*, \phi_h, k) - r_{h,k} - V_{h+1}^*(s_{h+1}^k)$, then $\mathbb{E}[\eta_k | s_{h}^{k}, a_{h}^{k}] = 0$ and $|\eta_k| \leq H$. Now by self-normalized Hoeffding’s inequality (Lemma L.3), with probability $1 - \delta$ (recall $t := K$ in Lemma L.3),
\[
\left\| \sum_{k=1}^{K} x_k(\theta) \eta_k \right\|_{G_h(\theta)^{-1}} \leq 2H \sqrt{d \log \left( \frac{\lambda + KK_1}{\lambda \delta} \right)}.
\]

**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 definition $|\eta_k| \leq H$, which implies $\|h(\theta)\|_2 \leq KH\kappa_1$ and by $x_k(\theta_1) - x_k(\theta_2) = \nabla \theta f(\theta, \phi_h, k) \cdot (\theta_1 - \theta_2)$,
\[
\|h(\theta_1) - h(\theta_2)\|_2 \leq K \max_k (H \|x_k(\theta_1) - x_k(\theta_2)\|_2) \leq HK\kappa_2 \|\theta_1 - \theta_2\|_2.
\]

Furthermore,
\[
\left\| G_h(\theta_1)^{-1} - G_h(\theta_2)^{-1} \right\|_2 \leq \left\| G_h(\theta_1) - G_h(\theta_2) \right\|_2 \left\| G_h(\theta_1)^{-1} \right\|_2 \left\| G_h(\theta_2) \right\|_2 \left\| G_h(\theta_2)^{-1} \right\|_2 \leq \frac{1}{\lambda^2} \sup_k \| \nabla f(\theta_1, \phi_h, k) \cdot \nabla f(\theta_1, \phi_h, k)^\top - \nabla f(\theta_2, \phi_h, k) \cdot \nabla f(\theta_2, \phi_h, k)^\top \|_2 \leq \frac{2\kappa_1 \kappa_2 K}{\lambda^2} \|\theta_1 - \theta_2\|_2.
\]

All the above imply
\[
\|H(\theta_1) - H(\theta_2)\| \leq \sqrt{H(\theta_1)^\top G_h(\theta_1)^{-1} h(\theta_1) - h(\theta_2)^\top G_h(\theta_2)^{-1} h(\theta_2)} \leq \sqrt{\|h(\theta_1) - h(\theta_2)\|_2 \cdot \frac{1}{\lambda} \cdot KH\kappa_1 \cdot \|G_h(\theta_1)^{-1} - G_h(\theta_2)^{-1}\|_2 \cdot KH\kappa_1 + \sqrt{KH\kappa_1 \cdot \|G_h(\theta_1)^{-1} - G_h(\theta_2)^{-1}\|_2 \cdot KH\kappa_1} \leq \sqrt{\|H(\theta_1) - H(\theta_2)\|_2 \cdot \frac{1}{\lambda} \cdot KH\kappa_1 + \sqrt{KH\kappa_1 \cdot \frac{1}{\lambda} \cdot H(\theta_1) - H(\theta_2)\|_2 \cdot KH\kappa_1} \leq \sqrt{\frac{1}{\lambda} \cdot KH\kappa_1 \cdot \|\theta_1 - \theta_2\|_2 \cdot KH\kappa_1 + \sqrt{KH\kappa_1 \cdot \frac{1}{\lambda} \cdot KH\kappa_1 \cdot \|\theta_1 - \theta_2\|_2 \cdot KH\kappa_1}} \leq 2KH\kappa_2 \frac{1}{\lambda} \cdot KH\kappa_1 + \sqrt{KH\kappa_1 \cdot \frac{2\kappa_1 \kappa_2 K}{\lambda^2} \cdot H(\theta_1) - H(\theta_2)\|_2 \cdot KH\kappa_1} \leq \sqrt{\frac{1}{\lambda} \cdot KH\kappa_1 \cdot \frac{2\kappa_1 \kappa_2 K}{\lambda^2} \cdot H(\theta_1) - H(\theta_2)\|_2 \cdot KH\kappa_1} \leq \left( \frac{4K^2\kappa_1^2 \kappa_2^2}{\lambda^2} + \sqrt{\frac{2K^3\kappa_1^3 \kappa_2^3}{\lambda^2}} \right) \sqrt{\|\theta_1 - \theta_2\|_2}
\]

Then a $\epsilon$-covering net of $\{H(\theta)\}$ can be constructed by the union of
\[
\left\{ \frac{\epsilon^2}{\sqrt{4K^2\kappa_1^2 \kappa_2^2 / \lambda^3 + \sqrt{2K^3\kappa_1^3 \kappa_2^3 / \lambda^3}}} \right\} \quad \text{covering net of } \{\theta \in \Theta\}. \quad \text{By Lemma L.8, the covering number } \mathcal{N}_c \text{ satisfies}
\]
\[
\log \mathcal{N}_c \leq d \log \left( 1 + \frac{2C_\Theta \left( \sqrt{4K^2\kappa_1^2 \kappa_2^2 / \lambda^3 + \sqrt{2K^3\kappa_1^3 \kappa_2^3 / \lambda^3}} \right)^2}{\epsilon^2} \right) = \tilde{O}(d)
\]
Step 3. First note by definition in Step 2
\[ \left\| \sum_{k=1}^{K} \left( f(\theta_{TV_{h+1}}, \phi_{h,k}) - r_{h,k} - V_{h+1}^*(s_{h+1}^k) \right) \cdot \nabla_\theta^T f(\hat{\theta}_h, \phi_{h,k}) \right\|_{\Sigma_h^{-1}} = H(\hat{\theta}_h) \]

Now choosing \( \epsilon = O(1/K) \) in Step 2 and union bound over the covering number in Step 2, we obtain with probability \( 1 - \delta \),
\[
H(\hat{\theta}_h) = \left\| \sum_{k=1}^{K} x_k(\hat{\theta}_h) \eta_k \right\|_{G_\epsilon(\hat{\theta}_h)^{-1}} \leq 2H \sqrt{d \left[ \log \left( \frac{\lambda + K \kappa_1}{\lambda \delta} \right) + O(d) \right] + O\left( \frac{1}{K} \right)}.
\]
where we absorb all the Polylog terms. Combining above with equation 12, we obtain with probability \( 1 - \delta \),
\[
|I_1| \leq \left\| \nabla f(\hat{\theta}_h, \phi(s, a)) \right\|_{\Sigma_h^{-1}} \cdot \left| \sum_{k=1}^{K} \left( f(\theta_{TV_{h+1}}, \phi_{h,k}) - r_{h,k} - V_{h+1}^*(s_{h+1}^k) \right) \cdot \nabla_\theta^T f(\hat{\theta}_h, \phi_{h,k}) \right|_{\Sigma_h^{-1}}
\leq \left\| \nabla f(\hat{\theta}_h, \phi(s, a)) \right\|_{\Sigma_h^{-1}} \cdot \left( 2H \sqrt{d \left[ \log \left( \frac{\lambda + K \kappa_1}{\lambda \delta} \right) + O(d) \right] + O\left( \frac{1}{K} \right)} \right)
\leq 4Hd \left\| \nabla f(\hat{\theta}_h, \phi(s, a)) \right\|_{\Sigma_h^{-1}} \cdot C_{\delta, \log \kappa} + O\left( \frac{\kappa_1}{\sqrt{\kappa} K} \right),
\]
where \( C_{\delta, \log \kappa} \) only contains Polylog terms.

E.5 Analyzing Hot \( \_2 \) in Equation 8

Lemma E.8. Recall \( \text{Hot}_2 := \nabla f(\hat{\theta}_h, \phi(s, a)) \Sigma_h^{-1} \left[ \hat{R}_K(\theta_{TV_{h+1}}) + \lambda \theta_{TV_{h+1}} \right] \). If the number of episode \( K \) satisfies \( K \geq \max \left\{ 512 \frac{\kappa_1}{\kappa^2} \left( \log \left( \frac{2d}{\delta} \right) + d \log(1 + \frac{4\kappa_1 \kappa_2 C_0 K^3}{\delta}) \right), \frac{4\lambda}{\kappa} \right\} \), then with probability \( 1 - \delta \),
\[
\left| \nabla f(\hat{\theta}_h, \phi(s, a)) \Sigma_h^{-1} \left[ \hat{R}_K(\theta_{TV_{h+1}}) + \lambda \theta_{TV_{h+1}} \right] \right| \leq \tilde{O}\left( \frac{\kappa_2 \max\left( \frac{\kappa_1}{\kappa}, \frac{1}{\sqrt{\kappa}} \right) d^2 H^2 + \frac{d^2 H^3 \kappa_3 + \kappa_3 C_0}{\kappa}}{K} \right),
\]
where \( \tilde{O} \) absorbs all the constants and Polylog terms.

Proof of Lemma E.8. Step 1: we first show with probability \( 1 - \delta \)
\[
\left| \nabla f(\hat{\theta}_h, \phi(s, a)) \Sigma_h^{-1} \hat{R}_K(\theta_{TV_{h+1}}) \right| \leq \tilde{O}\left( \frac{1}{K} \right).
\]
Recall by plug in \( \theta_{TV_{h+1}} \) in equation 4, we have
\[
Z_h(\theta_{TV_{h+1}}, \hat{V}_{h+1}) = \frac{\partial}{\partial \theta} \left[ Z_h(\hat{\theta}_h, \hat{V}_{h+1})(\theta_{TV_{h+1}} - \hat{\theta}_h) + R_K(\theta_{TV_{h+1}}) \right],
\]
and by second-order Taylor’s Theorem we have
\[
\left\| R_K(\theta_{TV_{h+1}}) \right\|_2 = \left\| Z_h(\theta_{TV_{h+1}}, \hat{V}_{h+1}) - Z_h(\hat{\theta}_h, \hat{V}_{h+1}) - \frac{\partial}{\partial \theta} \left[ Z_h(\theta_{TV_{h+1}}, \hat{V}_{h+1})(\theta_{TV_{h+1}} - \hat{\theta}_h) \right] \right\|_2
\leq \frac{1}{2} \kappa_{z2} \left\| \theta_{TV_{h+1}} - \hat{\theta}_h \right\|^2_2.
\]
Note
\[
\frac{\partial^2}{\partial \theta \partial \theta} Z_h(\theta | \hat{V}_{h+1}) \bigg|_{\theta = \xi} = \frac{\partial}{\partial \theta} \sum_{k=1}^{K} \left( f(\xi, \phi_{h,k}) - r_{h,k} - \hat{V}_{h+1}(s_{h+1}^{k}) \right) \cdot \nabla_{\theta}^2 f(\xi, \phi_{h,k}) \bigg]
\]
\[
+ \sum_{k=1}^{K} \frac{\partial}{\partial \theta} \left( \nabla_{\theta} f(\xi, \phi_{h,k})^\top \nabla_{\theta} f(\xi, \phi_{h,k}) + \lambda I_d \right)
\]

Therefore, we can bound \(\kappa_{z_2}\) with \(\kappa_{z_2} \leq (H\kappa_3 + 3\kappa_1\kappa_2) K\) and this implies with probability \(1 - \delta/2\),
\[
\left\| R_K(\theta_{\hat{T}V_{h+1}}) \right\|_2 \leq \frac{1}{2} \kappa_{z_2} \left\| \theta_{\hat{T}V_{h+1}} - \hat{\theta}_h \right\|_2 \leq \frac{1}{2} \frac{1}{2} (H\kappa_3 + 3\kappa_1\kappa_2) K \cdot \left\| \theta_{\hat{T}V_{h+1}} - \hat{\theta}_h \right\|_2
\]
\[
\leq \frac{1}{2} \left( H\kappa_3 + 3\kappa_1\kappa_2 \right) K \cdot \frac{36H^2(\log(H/\delta) + C_{d,\log K} + 2\lambda C_{\phi}^2)}{\kappa K}
\]
\[
\leq O((H\kappa_3 + 3\kappa_1\kappa_2)H^2d^2/\kappa).
\]

And by Corollary E.3 with probability \(1 - \delta/2\),
\[
\left\| \Delta_{\Sigma_{2}}(\hat{\theta}_{h} - \theta_{\hat{T}V_{h+1}}) \right\|_2 \leq \tilde{O}(1),
\]

Therefore, by Lemma L.5 and a union bound with probability \(1 - \delta\),
\[
|\nabla f(\hat{\theta}_{h}, \phi(s, a))^\top \Sigma_{2}^{-1} R_K(\theta_{\hat{T}V_{h+1}}) | = \left| \nabla f(\hat{\theta}_{h}, \phi(s, a))^\top \Sigma_{2}^{-1} \left( \Delta_{\Sigma_{2}}(\hat{\theta}_{h} - \theta_{\hat{T}V_{h+1}}) + R_K(\theta_{\hat{T}V_{h+1}}) \right) \right|
\]
\[
\leq \left| \nabla f(\hat{\theta}_{h}, \phi(s, a)) \right|_{\Sigma_{2}^{-1}} \left| \Delta_{\Sigma_{2}}(\hat{\theta}_{h} - \theta_{\hat{T}V_{h+1}}) + R_K(\theta_{\hat{T}V_{h+1}}) \right|_{\Sigma_{2}^{-1}}
\]
\[
\leq \left( \frac{2\kappa_1}{\sqrt{\kappa K}} + O\left( \frac{1}{K} \right) \right) \left( \frac{C}{\sqrt{\kappa}} + O\left( \frac{1}{K} \right) \right) = O \left( \kappa_2 \max\left( \frac{\sqrt{\kappa}}{\kappa K}, \frac{1}{\sqrt{\kappa}} \right) d^2H^2 + \frac{\kappa^2 d^2H^2}{\kappa K} \right)
\]

where \(\tilde{O}\) absorbs all the constants and Polylog terms. Here the last inequality uses bound for
\[
\left\| R_K(\theta_{\hat{T}V_{h+1}}) \right\|_2 \text{ and } \left\| \Delta_{\Sigma_{2}}(\hat{\theta}_{h} - \theta_{\hat{T}V_{h+1}}) \right\|_2.
\]

**Step 2:** By Lemma L.5, with probability \(1 - \delta\),
\[
|\nabla f(\hat{\theta}_{h}, \phi(s, a)) \Sigma_{h}^{-1} \lambda_{\alpha_{\hat{T}V_{h+1}}} | \leq \lambda \left| \nabla f(\hat{\theta}_{h}, \phi(s, a)) \right|_{\Sigma_{h}^{-1}} \left( \theta_{\hat{T}V_{h+1}} \right)_{\Sigma_{h}^{-1}}
\]
\[
\leq \lambda \left( \frac{2\kappa_1}{\sqrt{\kappa K}} + O\left( \frac{1}{K} \right) \right) \cdot \left( \frac{2C_{\phi}}{\sqrt{\kappa K}} + O\left( \frac{1}{K} \right) \right) = 4\lambda \frac{\kappa_1 C_{\phi}}{\kappa K} + O\left( \frac{1}{K^2} \right)
\]

\(\square\)

### F Proof of Theorem 3.2

Now we are ready to prove Theorem 3.2. In particular, we prove the first part. Also, recall that we consider the exact Bellman completeness (\(\epsilon_F = 0\)).

#### F.1 The first part

**Proof of Theorem 3.2 (first part).** First of all, from the previous calculation (3), (8), we have
\[
|P_h \hat{V}_{h+1}(s, a) - \hat{P}_h \hat{V}_{h+1}(s, a)| \leq |\nabla f(\hat{\theta}_{h}, \phi(s, a)) \left( \theta_{\hat{T}V_{h+1}} - \hat{\theta}_{h} \right)| + |\text{Hot}_{h, 1}|
\]
\[
\leq |I_1| + |J_2| + |J_3| + |\text{Hot}_{h, 2}| + |\text{Hot}_{h, 1}|
\]
Now by Lemma E.5, Lemma E.6, Lemma E.7, Lemma E.8 and Lemma E.1 (and a union bound), with probability $1 - \delta$,

$$
|I_1| \leq \widetilde{O}(\sqrt{\frac{d^3H^2\kappa^2}{\kappa^3}}) \frac{1}{K},
$$

$$
|I_2| \leq \widetilde{O}\left(\frac{\kappa^2 H^2d^2}{\kappa K}\right) + \widetilde{O}\left(\frac{1}{K^{3/2}}\right),
$$

$$
|I_1| \leq 4Hd \left\| \nabla f(\hat{\theta}_h, \phi(s, a)) \right\|_{\Sigma_h^{-1}} \cdot C_{\delta, log} K + \widetilde{O}\left(\frac{\kappa_1}{\sqrt{K}}\right),
$$

$$
|\text{Hot}_{h,2}| \leq \widetilde{O}\left(\frac{\kappa_2 \max(\frac{\kappa_1}{\sqrt{K}}, 1) d^2 H^2 + \frac{d^2 H^3 + 3K\kappa_1 C_\theta}{K}}{\kappa}\right),
$$

$$
|\text{Hot}_{h,1}| \leq \widetilde{O}\left(\frac{H^2 \kappa_d d^2}{\kappa} \right) \frac{1}{K}.
$$

Finally, Plug the above into Lemma D.3, by a union bound over all $h \in [H]$, we have with probability $1 - \delta$, for any policy $\pi$,

$$
v^{*} - \overline{v} \leq \sum_{h = 1}^{H} 2 \cdot \mathbb{E}_{\pi} \left[ |I_1| + |I_2| + |I_3| + |\text{Hot}_{h,2}| + |\text{Hot}_{h,1}| \right]
$$

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

where $\iota = C_{\delta, log} K$ only contains Polylog terms and

$$
C_{\text{hot}} = \frac{\kappa_1 H}{\sqrt{K}} + \frac{\kappa_2 H^2 d^2}{\kappa} + \sqrt{\frac{d^3 H^2 \kappa^2}{\kappa^3}} \cdot \max(\frac{\kappa_1}{\sqrt{K}}, 1) d^2 H^2 + \frac{d^3 H^3 + 3K\kappa_1 C_\theta + H^2 \kappa^2 d^2}{K}.
$$

\[\Box\]

F.2 THE SECOND PART

Next we prove the second part of Theorem 3.2.

Proof of Theorem 3.2 (second part). Step 1. Choose $\pi = \pi^*$ in the first part, we have

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

Next, by the triangular inequality of the norm to obtain

$$
\left\| \nabla f(\hat{\theta}_h, \phi(s, a_h)) \right\|_{\Sigma_h^{-1}} - \left\| \nabla f(\theta^*_h, \phi(s, a_h)) \right\|_{\Sigma_h^{-1}} \leq \left\| \nabla f(\hat{\theta}_h, \phi(s, a_h)) - \nabla f(\theta^*_h, \phi(s, a_h)) \right\|_{\Sigma_h^{-1}}
$$

$$
= \left\| \nabla^2 f(\xi, \phi(s, a_h)) \cdot (\hat{\theta}_h - \theta^*_h) \right\|_{\Sigma_h^{-1}},
$$

since with probability $1 - \delta$,

$$
\left\| \nabla^2 f(\xi, \phi(s, a_h)) \cdot (\hat{\theta}_h - \theta^*_h) \right\|_{\Sigma_h^{-1}} \leq \kappa_2 \left\| \hat{\theta}_h - \theta^*_h \right\|_{2} \leq \widetilde{O}\left(\frac{\kappa_1 \kappa_2 H^2 d}{\kappa} \sqrt{\frac{1}{K}}\right),
$$

where the last inequality uses part three of Theorem G.3. Then by a union bound and Lemma L.5,

$$
\left\| \nabla^2 f(\xi, \phi(s, a_h)) \cdot (\hat{\theta}_h - \theta^*_h) \right\|_{\Sigma_h^{-1}} \leq \widetilde{O}\left(\frac{\kappa_1 \kappa_2 H^2 d}{\kappa^{3/2}} \cdot \frac{1}{K}\right).
$$
Step 2. Next, we show with probability $1-\delta$, 
\[
\|\nabla_\theta f(\theta^*_h, \phi(s_h, a_h))\|_{\Sigma_h^{-1}} \leq 2 \|\nabla_\theta f(\theta^*_h, \phi(s_h, a_h))\|_{\Sigma_h^{-1}}.
\]
First of all, 
\[
\left\| \frac{1}{K} \Sigma_h - \frac{1}{K} \Sigma_h^* \right\|_2 \leq \left\| \frac{1}{K} \left( \sum_{k=1}^K \nabla f(\theta_h, \phi(s, a))\nabla f(\theta_h, \phi(s, a))^\top - \nabla f(\theta^*_h, \phi(s, a))\nabla f(\theta^*_h, \phi(s, a))^\top \right) \right\|_2 \leq \sup_{s,a} \left( \left\| (\nabla f(\theta_h, \phi(s, a)) - \nabla f(\theta^*_h, \phi(s, a)) ) \nabla f(\theta_h, \phi(s, a)) \right\|_2 + \left\| (\nabla f(\theta_h, \phi(s, a)) - \nabla f(\theta^*_h, \phi(s, a)) ) \nabla f(\theta_h, \phi(s, a)) \right\|_2 \right) \leq 2\kappa_2 \kappa_1 \|\theta_h - \theta^*_h\|_2 \leq O\left( \frac{\kappa_2 \kappa_1^2 H^2 d}{\sqrt{1/K}} \right)
\]
Second, by Lemma L.6 with probability $1-\delta$ 
\[
\left\| \frac{\Sigma_h}{K} - \mathbb{E}_\mu [\nabla_\theta f(\theta^*_h, \phi)\nabla_\theta f(\theta^*_h, \phi)^\top] - \frac{\lambda}{K} \right\| \leq \frac{4\sqrt{2}\kappa_1^2}{\sqrt{K}} \left( \log \frac{2d}{\delta} \right)^{1/2}
\]
This implies 
\[
\left\| \frac{\Sigma_h^*}{K} \right\| \leq \left\| \mathbb{E}_\mu [\nabla_\theta f(\theta^*_h, \phi)\nabla_\theta f(\theta^*_h, \phi)^\top] \right\| + \frac{\lambda}{K} + \frac{4\sqrt{2}\kappa_1^2}{\sqrt{K}} \left( \log \frac{2d}{\delta} \right)^{1/2} \leq \kappa_1^2 + \lambda + 4\sqrt{2}\kappa_1^2 \left( \log \frac{2d}{\delta} \right)^{1/2}
\]
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 $1-\delta$ 
\[
\lambda_{\min}(\frac{\Sigma_h}{K}) \geq \lambda_{\min}(\mathbb{E}_\mu [\nabla_\theta f(\theta^*_h, \phi)\nabla_\theta f(\theta^*_h, \phi)^\top]) + \frac{\lambda}{K} - \frac{4\sqrt{2}\kappa_1^2}{\sqrt{K}} \left( \log \frac{2d}{\delta} \right)^{1/2} \geq \kappa + \frac{\lambda}{K} - \frac{4\sqrt{2}\kappa_1^2}{\sqrt{K}} \left( \log \frac{2d}{\delta} \right)^{1/2} \geq \frac{\kappa}{2}
\]
then 
\[
\left\| \left( \frac{\Sigma_h}{K} \right)^{-1} \right\| \leq \frac{2}{\kappa}. \text{ Similarly, with probability } 1-\delta, \left\| \left( \frac{\Sigma_h}{K} \right)^{-1} \right\| \leq \frac{2}{\kappa}. \text{ Then by Lemma L.7,}
\]
\[
\|\nabla_\theta f(\theta^*_h, \phi(s, a))\|_{K\Sigma_h^{-1}} \leq 1 + \sqrt{\|K\Sigma_h^{-1}\| \cdot \|K\Sigma_h^{-1}\| \cdot \|K\Sigma_h^{-1}\| \cdot \|K\Sigma_h^{-1}\| \cdot \|K\Sigma_h^{-1}\| \cdot \|\nabla_\theta f(\theta^*_h, \phi(s, a))\|_{K\Sigma_h^{-1}} \leq 1 + \frac{4}{\kappa^2} O(\kappa_1^2 + \lambda) \hat{O}\left( \frac{\kappa_2 \kappa_1^2 H^2 d}{\sqrt{1/K}} \right) \cdot \|\nabla_\theta f(\theta^*_h, \phi(s, a))\|_{K\Sigma_h^{-1}} \leq 2 \|\nabla_\theta f(\theta^*_h, \phi(s, a))\|_{K\Sigma_h^{-1}} \].
\]
as long as $K \geq \hat{O}\left( \frac{(\kappa_1^2 + \lambda)^2 \kappa_2 \kappa_1^2 H^2 d^2}{\kappa^3} \right)$. The above is equivalently to 
\[
\|\nabla_\theta f(\theta^*_h, \phi(s_h, a_h))\|_{\Sigma_h^{-1}} \leq 2 \|\nabla_\theta f(\theta^*_h, \phi(s_h, a_h))\|_{\Sigma_h^{-1}} .
\]
Combining Step1, Step2 and a union bound, we have with probability $1 - \delta$,
\[
0 \leq v^* - v^g \leq \sum_{h=1}^{H} 8dH \cdot \mathbb{E}_{\theta^*} \left[ \sqrt{\nabla_{\theta}^T \tilde{f}(\theta^*_h, \phi(s_h, a_h)) \Sigma_{\theta}^{-1} \nabla_{\theta} \tilde{f}(\theta^*_h, \phi(s_h, a_h))} \right] + \tilde{O}(\frac{C_{\text{hot}}}{K})
\]
\[
\leq \sum_{h=1}^{H} 8dH \cdot \mathbb{E}_{\theta^*} \left[ \sqrt{\nabla_{\theta}^T \tilde{f}(\theta^*_h, \phi(s_h, a_h)) \Sigma_{\theta}^{-1} \nabla_{\theta} \tilde{f}(\theta^*_h, \phi(s_h, a_h))} \right] + \tilde{O}(\frac{C_{\text{hot}}}{K}) + \tilde{O}(\frac{\kappa_{1}\kappa_{2}H^4d^2}{K^{3/2}} - \frac{1}{K})
\]
\[
\leq \sum_{h=1}^{H} 16dH \cdot \mathbb{E}_{\theta^*} \left[ \sqrt{\nabla_{\theta}^T \tilde{f}(\theta^*_h, \phi(s_h, a_h)) \Sigma_{\theta}^{-1} \nabla_{\theta} \tilde{f}(\theta^*_h, \phi(s_h, a_h))} \right] + \tilde{O}(\frac{C_{\text{hot}}}{K})
\]
where $C'_{\text{hot}} = C_{\text{hot}} + \frac{\kappa_{1}\kappa_{2}H^4d^2}{K^{3/2}}$.

\[\square\]

### G. PROVABLE EFFICIENCY BY REDUCTION TO GENERAL FUNCTION APPROXIMATION

In this section, we bound the accuracy of the parameter difference $\left\| \hat{\theta}_h - \theta_{TV_{h+1}} \right\|_2$ via a reduction to General Function Approximation scheme in Chen and Jiang (2019).

Recall the objective
\[
\ell_h(\theta) := \frac{1}{K} \sum_{k=1}^{K} \left[ f(\theta, \phi(s^k_h, a^k_h)) - r(s^k_h, a^k_h) - \hat{V}_{h+1}(s^k_{h+1}) \right]^2 + \frac{\lambda}{K} \cdot \left\| \theta \right\|_2^2
\]
(17)

Then by definition, $\hat{\theta}_h := \arg \min_{\theta \in \Theta} \ell_h(\theta)$ and $\theta_{TV_{h+1}}$ satisfies $f(\theta_{TV_{h+1}}, \phi) = \mathcal{P}_h \hat{V}_{h+1} + \delta_{\hat{V}_{h+1}}$.

Therefore, in this case, we have the following lemma:

**Lemma G.1.** Fix $h \in [H]$. With probability $1 - \delta$,
\[
\mathbb{E}_{\mu}[\ell_h(\hat{\theta}_h)] - \mathbb{E}_{\mu}[\ell_h(\theta_{TV_{h+1}})] \leq \frac{36H^2(\log(1/\delta) + C_{d, \log K}) + \lambda C_{\Theta}^2}{K} + \sqrt{\frac{16H^3\epsilon_f(\log(1/\delta) + C_{d, \log K})}{K}} + 4H\epsilon_f.
\]

where the expectation over $\mu$ is taken w.r.t. $(s^k_h, a^k_h, s^k_{h+1}) = 1, \ldots, K$ only (i.e., first compute $\mathbb{E}_{\mu}[\ell_h(\theta)]$ for a fixed $\theta$, then plug-in either $\hat{\theta}_{h+1}$ or $\theta_{TV_{h+1}}$). Here
\[
C_{d, \log K} := d \log(1 + 24C_\Theta (H + 1)\kappa_1 K) + d \log \left( 1 + 288H^2C_\Theta (\kappa_1 \sqrt{C_\Theta} + 2\sqrt{\kappa_1\kappa_2/\lambda}K^2) \right) + d^2 \log \left( 1 + 288H^2dK^2/\lambda \right).
\]

**Proof of Lemma G.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 $f_V(s, a) := f(\theta_{TV}, \phi) = \mathcal{P}_h V + \delta_V$. For any fixed $\theta \in \Theta$, denote $g(s, a) = f(\theta, \phi(s, a))$. Then define\(^{10}\)
\[
X(g, V, f_V) := (g(s, a) - r - V(s'))^2 - (f_V(s, a) - r - V(s'))^2.
\]

Since all episodes are independent of each other, $X_k(g, V, f_V) := X(g(s^k_h, a^k_h), V(s^k_{h+1}), f_V(s^k_h, a^k_h))$ are independent r.v.s and it holds
\[
\frac{1}{K} \sum_{k=1}^{K} X_k(g, V, f_V) = \ell(g) - \ell(f_V).
\]
(18)

\(^{10}\)We abuse the notation here to use either $X(g, V, f_V)$ or $X(\theta, V, f_V)$. They mean the same quantity.
Next, the variance of $X$ is bounded by:

$$\text{Var}[X(g, f, f')] \leq \mathbb{E}_\mu[X(g, f, f')^2]$$

$$= \mathbb{E}_\mu\left[\left((g(s, ah) - rh - V(s_{h+1}))^2 - (f_V(s, ah) - rh - V(s_{h+1}))^2\right)^2\right]$$

$$= \mathbb{E}_\mu\left[(g(s, ah) - f_V(s, ah))^2(g(s, ah) + f_V(s, ah) - 2rh - 2V(s_{h+1}))^2\right]$$

$$\leq 4H^2 \cdot \mathbb{E}_\mu[(g(s, ah) - f_V(s, ah))^2]$$

$$\leq 4H^2 \cdot \mathbb{E}_\mu\left[(g(s, ah) - rh - V(s_{h+1}))^2 - (f_V(s, ah) - rh - V(s_{h+1}))^2\right] + 8H^3\epsilon_F \ (*)$$

$$= 4H^2 \cdot \mathbb{E}_\mu[X(g, f, f')^2] + 8H^3\epsilon_F$$

where the step (*) comes from

$$\mathbb{E}_\mu\left[(g(s, ah) - rh - V(s_{h+1}))^2 - (f_V(s, ah) - rh - V(s_{h+1}))^2\right]$$

$$= \mathbb{E}_\mu\left[(g(s, ah) - f_V(s, ah))^2\right] + \mathbb{E}_\mu\left[2(g(s, ah) - f_V(s, ah))\mathbb{E}_{\rho}[f_V(s, ah) - rh - V(s_{h+1}) | s_h, a_h]\right]$$

$$\geq \mathbb{E}_\mu\left[(g(s, ah) - f_V(s, ah))^2\right] - 2H |\delta_V|_\infty \geq \mathbb{E}_\mu\left[(g(s, ah) - f_V(s, ah))^2\right] - 2H\epsilon_F \ (19)$$

where the last step uses law of total expectation and the definition of $f_V$.

Therefore, by Bernstein inequality, with probability $1 - \delta$,

$$\mathbb{E}_\mu[X(g, f, f')] - \frac{1}{K} \sum_{k=1}^K X_k(g, f, f')$$

$$\leq \sqrt{\frac{2\text{Var}[X(g, f, f')]|\log(1/\delta)|}{K} + \frac{4H^2 \log(1/\delta)}{3K}$$

$$\leq \sqrt{\frac{8H^2\mathbb{E}_\mu[X(g, f, f')]|\log(1/\delta)|}{K} + \frac{16H^3\epsilon_F |\log(1/\delta)|}{K} + \frac{4H^2 \log(1/\delta)}{3K}}.$$

Now, if we choose $g(s, a) := f(\hat{\theta}_h, \phi(s, a))$, then $\hat{\theta}_h$ minimizes $\ell_h(\theta)$, therefore, it also minimizes

$$\frac{1}{K} \sum_{k=1}^K X_k(\theta, \hat{V}_{h+1}, f_{\hat{V}_{h+1}})$$

and this implies

$$\frac{1}{K} \sum_{k=1}^K X_k(\hat{\theta}_h, \hat{V}_{h+1}, f_{\hat{V}_{h+1}}) \leq \frac{1}{K} \sum_{k=1}^K X_k(\theta, \hat{V}_{h+1}, f_{\hat{V}_{h+1}}) = 0.$$

Therefore, we obtain

$$\mathbb{E}_\mu[X(\hat{\theta}_h, \hat{V}_{h+1}, f_{\hat{V}_{h+1}})] \leq \sqrt{\frac{8H^2 \cdot \mathbb{E}_\mu[X(\hat{\theta}_h, \hat{V}_{h+1}, f_{\hat{V}_{h+1}})]|\log(1/\delta)|}{K} + \frac{16H^3\epsilon_F |\log(1/\delta)|}{K} + \frac{4H^2 \log(1/\delta)}{3K}}.$$

However, the above does not hold with probability $1 - \delta$ since $\hat{\theta}_h$ and $\hat{V}_{h+1} := \min\{\max_a f(\hat{\theta}_{h+1}, \phi(\cdot, a)) - \nabla f(\hat{\theta}_{h+1}, \phi(\cdot, a))^T A \cdot \nabla f(\theta, \phi(\cdot, a)), H\}$ (where $A$ is certain symmetric matrix with bounded norm) depend on $\theta_h$ and $\hat{\theta}_{h+1}$ which are data-dependent. Therefore, we need to further apply covering Lemma L.10 and choose $\epsilon = O(1/K)$ and a union bound to obtain with probability $1 - \delta$,

$$\mathbb{E}_\mu[X(\hat{\theta}_h, \hat{V}_{h+1}, f_{\hat{V}_{h+1}})] \leq \sqrt{\frac{8H^2 \cdot \mathbb{E}_\mu[X(\hat{\theta}_h, \hat{V}_{h+1}, f_{\hat{V}_{h+1}})](|\log(1/\delta)| + C_d\log K)}{K} + \frac{7H^2(1/\delta) + C_d\log K}{3K} + \frac{16H^3\epsilon_F(1/\delta) + C_d\log K}{K} + 4H\epsilon_F \ (20)$$
where $C_{d,\log(K)} := \log(1+24C_\Theta(H+1)\kappa_1 K) + d \log \left(1 + 288H^2C_\Theta(\kappa_1\sqrt{C_\Theta} + 2\kappa_1\kappa_2/\lambda)^2K^2\right) + d^2 \log \left(1 + 288H^2\sqrt{d}\kappa_1^2 K^2/\lambda\right)$.

Solving this quadratic equation to obtain with probability $1 - \delta$,

$$
\mathbb{E}_\mu[X(\hat{\theta}_h, \hat{V}_{h+1}, f_{\hat{V}_{h+1}})] \leq \frac{36H^2(\log(1/\delta) + C_{d,\log(K)}) + \sqrt{16H^3\epsilon_F (\log(1/\delta) + C_{d,\log(K)})}}{K} + 4H\epsilon_F.
$$

Now according to equation 18, by definition we finally have with probability $1 - \delta$ (recall the expectation over $\mu$ is taken w.r.t. $(s^k_h, a^k_h, \hat{V}^k_{h+1})$ $k = 1, ..., K$ only)

$$
\mathbb{E}_\mu[\ell_h(\hat{\theta}_{\hat{V}_{h+1}})] - \mathbb{E}_\mu[\ell_h(\theta_{\hat{V}_{h+1}})] = \mathbb{E}_\mu[X(\hat{\theta}_h, \hat{V}_{h+1}, f_{\hat{V}_{h+1}})] \\
\leq \frac{36H^2(\log(1/\delta) + C_{d,\log(K)}) + \sqrt{16H^3\epsilon_F (\log(1/\delta) + C_{d,\log(K)})}}{K} + 4H\epsilon_F. \quad (20)
$$

**Step 2.** If $\lambda > 0$, there is only extra term $\frac{\lambda}{K} \left(\left\Vert \hat{\theta}_h \right\Vert_2 - \left\Vert \theta_{\hat{V}_{h+1}} \right\Vert_2\right) \leq \frac{\lambda}{K} \left\Vert \hat{\theta}_h \right\Vert_2 \leq \frac{\lambda C^2}{K}$ in addition to above. This finishes the proof.

**Theorem G.2 (Provable efficiency (Part I)).** Let $C_{d,\log(K)}$ be the same as Lemma G.1. Then denote $b_{d,K,\epsilon_F} := \frac{\sqrt{16H^3\epsilon_F (\log(1/\delta) + C_{d,\log(K)})}}{K} + 4H\epsilon_F$, with probability $1 - \delta$

$$
\left\Vert \hat{\theta}_h - \theta_{\hat{V}_{h+1}} \right\Vert_2 \leq \sqrt{\frac{36H^2(\log(H/\delta) + C_{d,\log(K)}) + 2\lambda C^2}{\kappa K}} + \sqrt{\frac{b_{d,K,\epsilon_F}}{\kappa}} + \sqrt{\frac{2\kappa H\epsilon_F}{\kappa}}, \forall h \in [H].
$$

**Proof of Theorem G.2.** Apply a union bound in Lemma G.1, we have with probability $1 - \delta$,

$$
\mathbb{E}_\mu[\ell_h(\hat{\theta}_h)] - \mathbb{E}_\mu[\ell_h(\theta_{\hat{V}_{h+1}})] \leq \frac{36H^2(\log(H/\delta) + C_{d,\log(K)}) + 2\lambda C^2}{K} + \frac{b_{d,K,\epsilon_F}}{K} + \frac{2\kappa H\epsilon_F}{K}, \forall h \in [H]
$$

(21)

Now we prove for all $h \in [H]$,

$$
\mathbb{E}_\mu \left[ \left( f(\hat{\theta}_h, \phi(\cdot)) - f(\theta_{\hat{V}_{h+1}}, \phi(\cdot)) \right)^2 \right] \leq \mathbb{E}_\mu \left[ \ell_h(\hat{\theta}_h) - \frac{\lambda}{K} \left\Vert \hat{\theta}_h \right\Vert_2 \right] - \mathbb{E}_\mu \left[ \ell_h(\theta_{\hat{V}_{h+1}}) - \frac{\lambda}{K} \left\Vert \theta_{\hat{V}_{h+1}} \right\Vert_2 \right] + 2H\epsilon_F. \quad (22)
$$

Indeed, similar to equation 20, by definition we have

$$
\mathbb{E}_\mu \left[ \ell_h(\hat{\theta}_h) - \frac{\lambda}{K} \left\Vert \hat{\theta}_h \right\Vert_2 \right] - \mathbb{E}_\mu \left[ \ell_h(\theta_{\hat{V}_{h+1}}) - \frac{\lambda}{K} \left\Vert \theta_{\hat{V}_{h+1}} \right\Vert_2 \right] = \mathbb{E}_\mu[X(\hat{\theta}_h, \hat{V}_{h+1}, f_{\hat{V}_{h+1}})]
$$

$$
= \mathbb{E}_\mu \left[ \left( f(\hat{\theta}_h, \phi(s_h, a_h)) - r_h - \hat{V}_{h+1}(s_h+1) \right)^2 - \left( f(\theta_{\hat{V}_{h+1}}, \phi(s_h, a_h)) - r_h - \hat{V}_{h+1}(s_h+1) \right)^2 \right]
$$

$$
= \mathbb{E}_\mu \left[ \left( f(\hat{\theta}_h, \phi(\cdot)) - f(\theta_{\hat{V}_{h+1}}, \phi(\cdot)) \right)^2 \right]
$$

$$
+ \mathbb{E}_\mu \left[ \left( f(\hat{\theta}_h, \phi(s_h, a_h)) - f(\theta_{\hat{V}_{h+1}}, \phi(s_h, a_h)) \right) \cdot \left( f(\theta_{\hat{V}_{h+1}}, \phi(s_h, a_h)) - r_h - \hat{V}_{h+1}(s_h+1) \right) \right]
$$

$$
\geq \mathbb{E}_\mu \left[ \left( f(\hat{\theta}_h, \phi(\cdot)) - f(\theta_{\hat{V}_{h+1}}, \phi(\cdot)) \right)^2 \right] - 2H\epsilon_F
$$

**Note:** Here in our realization of Lemma L.9, we set $B = 1/\lambda$ (since $\left\Vert \Sigma_h^{-1} \right\Vert_2 \leq 1/\lambda$).
where the third identity uses $\mu$ is taken w.r.t. $s_h, a_h, s_{h+1}$ (recall Lemma G.1) and law of total expectation. The first inequality uses the definition of $\tilde{T}_{T_h+1}$.

Now apply Assumption 2.3, we have

$$
E_\mu \left[ \left( f(\hat{\theta}_h, \phi(\cdot, \cdot)) - f(\tilde{\theta}_{T_{T_h+1}}, \phi(\cdot, \cdot)) \right)^2 \right] \geq \kappa \left\| \hat{\theta}_h - \tilde{\theta}_{T_{T_h+1}} \right\|_2^2,
$$

Combine the above with equation 21 and equation 22, we obtain the stated result. \hfill \Box

**Theorem G.3** (Provable efficiency (Part II)). Let $C_{d, \log K}$ be the same as Lemma G.1 and suppose $\epsilon_F = 0$. Furthermore, suppose $\lambda \leq 1/2C^2_\Theta$ and $K \geq \max \left\{ 512 \frac{\kappa^3}{C^2_\Theta}, \log(\frac{2d}{\delta}) + d \log(1 + \frac{4\kappa^2\kappa^2_2 C_\Theta K^2}{\kappa^2}), \frac{4\lambda}{\kappa} \right\}$. Then, with probability $1 - \delta$, $\forall h \in [H]$, $f(\hat{\theta}_h, \phi(s, a)) - f(\theta^*_h, \phi(s, a)) \leq \left( \kappa_1 H \sqrt{\frac{36H^2(\log(H^2/\delta) + C_{d, \log K}) + 2\lambda C_\Theta^2}{\kappa}} + \frac{2H^2\kappa^1}{\sqrt{\kappa}} \right) \sqrt{\frac{1}{K}} + O\left( \frac{1}{K} \right)$.

Furthermore, we have with probability $1 - \delta$,

$$
\sup \left\| \tilde{V}_h - V_h^* \right\|_\infty \leq \left( \kappa_1 H \sqrt{\frac{36H^2(\log(H^2/\delta) + C_{d, \log K}) + 2\lambda C_\Theta^2}{\kappa}} + \frac{2H^2\kappa^1}{\sqrt{\kappa}} \right) \sqrt{\frac{1}{K}} + O\left( \frac{1}{K} \right)
$$

where $\tilde{O}$ absorbs Polylog terms and higher order terms. Lastly, it also holds for all $h \in [H]$, w.p. $1 - \delta$

$$
\sup \left\| \hat{\theta}_h - \theta^*_h \right\|_2 \leq \left( \kappa_1 H \sqrt{\frac{36H^2(\log(H^2/\delta) + C_{d, \log K}) + 2\lambda C_\Theta^2}{\kappa}} + \frac{2H^2\kappa^1}{\sqrt{\kappa}} \right) \sqrt{\frac{1}{K}} + O\left( \frac{1}{K} \right)
$$

Proof of Theorem G.3. **Step 1:** we show the first result.

We prove this by backward induction. When $h = H + 1$, by convention $f(\hat{\theta}_h, \phi(s, a)) = f(\theta^*_h, \phi(s, a)) = 0$ so the base case holds. Suppose for $h + 1$, with probability $1 - (H - h)\delta$, it holds true that $\sup_{s, a} \left| f(\hat{\theta}_{h+1}, \phi(s, a)) - f(\theta^*_{h+1}, \phi(s, a)) \right| \leq C_{h+1} \sqrt{\frac{1}{K}} + a(h + 1)$, we next consider the case for $t = h$.

On one hand, by Theorem G.2, we have with probability $1 - \delta/2$,

$$
\sup_{s, a} \left| f(\hat{\theta}_h, \phi(s, a)) - f(\theta^*_h, \phi(s, a)) \right|
\leq \sup_{s, a} \left| f(\hat{\theta}_h, \phi(s, a)) - f(\tilde{T}_{T_h+1}, \phi(s, a)) \right| + \sup_{s, a} \left| f(\tilde{T}_{T_h+1}, \phi(s, a)) - f(\theta^*_h, \phi(s, a)) \right|
\leq \kappa_1 \cdot \left\| \tilde{\theta}_h - \theta^*_h \right\|_2 + \sup_{s, a} \left| \nabla f(\xi, \phi(s, a)) \right| + \sup_{s, a} \left| f(\theta^*_h, \phi(s, a)) - f(\theta_{T_{T_h+1}}, \phi(s, a)) \right|
\leq \kappa_1 \cdot \left\| \tilde{T}_{T_h+1} - \theta^*_h \right\|_2 + \frac{\kappa K}{\kappa^2} \left( \frac{36H^2(\log(H/\delta) + C_{d, \log K}) + 2\lambda C_\Theta^2}{\kappa^2} \right) + \left\| \tilde{V}_{h+1} - V_{h+1}^* \right\|_\infty.
$$

Recall $\tilde{V}_{h+1}(\cdot) := \min_{s, a} \{ f(\hat{\theta}_{h+1}, \phi(s, a)) - \Gamma_h(\cdot, a), H \}$ and $V_{h+1}(\cdot) = \max_{s, a} \{ f(\theta^*_{h+1}, \phi(s, a)) - \Gamma_h(\cdot, a), H \}$, we obtain

$$
\left\| \tilde{V}_{h+1} - V_{h+1}^* \right\|_\infty \leq \sup_{s, a} \left| f(\hat{\theta}_{h+1}, \phi(s, a)) - f(\theta^*_{h+1}, \phi(s, a)) \right| + \sup_{h, s, a} \Gamma_h(s, a)
\quad (23)
$$
Note the above holds true for any generic \( \Gamma_h(s, a) \). In particular, according to Algorithm 1, we specify
\[
\Gamma_h(\cdot, \cdot) = dH + \sqrt{\nabla \theta f(\theta_h, \phi(\cdot, \cdot))}^\top \Sigma_{\theta_h}^{-1} \nabla \theta f(\theta_h, \phi(\cdot, \cdot)) \left( + \tilde{O}(\frac{1}{K}) \right)
\]
and by Lemma L.5, with probability \( 1 - \delta \),
\[
\Gamma_h \leq \frac{2dH \kappa_1}{\sqrt{\kappa K}} + \tilde{O}(\frac{1}{K})
\]
and by a union bound this implies with probability \( 1 - (H - h + 1)\delta \),
\[
\sup_{s, a} \left| f(\hat{\theta}_h, \phi(s, a)) - f(\theta_h^*, \phi(s, a)) \right| \leq C_{h+1} \left( \frac{\kappa_1 \sqrt{36H^2 + C_{d, log K}} + 2\kappa C_{\phi}^2}{\kappa K} \right) + \frac{2dH \kappa_1}{\sqrt{\kappa K}} + \tilde{O}(\frac{1}{K}) := C_h \left( \frac{1}{K} \right) + \tilde{O}(\frac{1}{K})
\]
Solving for \( C_h \), we obtain \( C_h \leq \kappa_1 H \sqrt{\frac{36H^2 + C_{d, log K} + 2\kappa C_{\phi}^2}{\kappa K}} + \frac{2dH \kappa_1}{\sqrt{\kappa K}} + \tilde{O}(\frac{1}{K}) \) for all \( H \). By a union bound (replacing \( \delta \) by \( \delta/H \)), we obtain the stated result.

**Step2:** Utilizing the intermediate result equation 23, we directly have with probability \( 1 - \delta \),
\[
\sup_{h} \left| \tilde{V}_h - V_h^* \right| \leq \sup_{s, a} \left| f(\hat{\theta}_h, \phi(s, a)) - f(\theta_h^*, \phi(s, a)) \right| + \frac{2dH \kappa_1}{\sqrt{\kappa K}} + O(\frac{1}{K}),
\]
where \( \sup_{s, a} \left| f(\hat{\theta}_h, \phi(s, a)) - f(\theta_h^*, \phi(s, a)) \right| \) can be bounded using Step1.

**Step3:** Denote \( M := \left( \kappa_1 H \sqrt{\frac{36H^2 + C_{d, log K} + 2\kappa C_{\phi}^2}{\kappa K}} + \frac{2dH \kappa_1}{\sqrt{\kappa K}} \right) \tilde{O}(\frac{1}{K}) \), then by Step1 we have with probability \( 1 - \delta \) (here \( \xi \) is some point between \( \hat{\theta}_h \) and \( \theta_h^* \) for all \( h \in [H] \)
\[
M^2 \geq \sup_{s, a} \left| f(\hat{\theta}_h, \phi(s, a)) - f(\theta_h^*, \phi(s, a)) \right|^2 \\
\geq \mathbb{E}_{\mu, h}[(f(\hat{\theta}_h, \phi(s, a)) - f(\theta_h^*, \phi(s, a))^2] \geq \kappa \left\| \hat{\theta}_h - \theta_h^* \right\|_2^2
\]
where the last inequality is by Assumption 2.3. Solve this to obtain the stated result.

\[
\square
\]

**H WITH POSITIVE BELLMAN COMPLETENESS COEFFICIENT \( \epsilon_F > 0 \)**

In Theorem 3.2, we consider the case where \( \epsilon_F = 0 \). If \( \epsilon_F > 0 \), similar guarantee can be achieved with the measurement of model misspecification. For instance, the additional error \( \sqrt{\frac{36H^2 \epsilon_F + C_{d, log K}}{\kappa K}} + 4H \epsilon_F \) will show up in Lemma G.1 (as stated in the current version), \( \sqrt{\frac{b_d, K, \epsilon_F}{\kappa}} \) will show up in Lemma G.2. Then the decomposition in equation 3 will incur the extra \( \delta_{\tilde{V}_{h+1}} \) term with \( \delta_{\tilde{V}_{h+1}} \) might not be 0. The analysis with positive \( \epsilon_F > 0 \) 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 \( \epsilon_F = 0 \) so the readers can focus on the more critical components that characterize the hardness of differentiable function class.

**I VFQL AND ITS ANALYSIS**

We present the *vanilla fitted Q-learning* (VFQL) Algorithm 2 as follows. For VFQL, no pessimism is used and we assume \( \hat{\theta}_h \in \Theta \) without loss of generality.
Algorithm 2 Vanilla Fitted Q-Learning (VFQL)

1: **Input:** Offline Dataset $\mathcal{D} = \{(s^k_h, a^k_h, r^k_h, s^k_{h+1})\}_{k, h=1}^{K, H}$. Denote $\phi_{h,k} := \phi(s^k_h, a^k_h)$.
2: **Initialization:** Set $\hat{V}_{h+1}(\cdot) \leftarrow 0$ and $\lambda > 0$.
3: for $h = H, H - 1, \ldots, 1$ do
4:   Set $\hat{\theta}_h \leftarrow \arg\min_{\theta \in \Theta} \left\{ \sum_{k=1}^{K} \left[ f(\theta, \phi_{h,k}) - r_h, k - \hat{V}_{h+1}(s^k_{h+1}) \right]^2 + \lambda \cdot \|\theta\|^2 \right\}$
5:   Set $\hat{Q}_h(\cdot, \cdot) \leftarrow \min \left\{ f(\hat{\theta}_h, \phi(\cdot, \cdot)), H - h + 1 \right\}$
6:   Set $\hat{\pi}_h(\cdot | \cdot) \leftarrow \arg\max_{\pi_h} \langle \hat{Q}_h(\cdot, \cdot), \pi_h(\cdot | \cdot) \rangle_A$, $\hat{V}_h(\cdot) \leftarrow \max_{\pi_h} \langle \hat{Q}_h(\cdot, \cdot), \pi_h(\cdot | \cdot) \rangle_A$
7: end for
8: **Output:** $\{\hat{\pi}_h\}_{h=1}^{H}$.

I.1 Analysis for VFQL (Theorem 3.1)

Recall $\ell_h(s, a) := \mathbb{P}_h \hat{V}_{h+1}(s, a) - \hat{Q}_h(s, a)$ and the definition of Bellman operator $D.1$. Note $\min\{\cdot, H - h + 1\}^+$ is a non-expansive operator, therefore we have

$$|\ell_h(s, a)| = |\mathbb{P}_h \hat{V}_{h+1}(s, a) - \hat{Q}_h(s, a)| = \left| \min \left\{ \mathbb{P}_h \hat{V}_{h+1}(s, a), H - h + 1 \right\}^+ - \min \left\{ f(\hat{\theta}_h, \phi(\cdot, \cdot)), H - h + 1 \right\}^+ \right|$$

$$\leq \left| \mathbb{P}_h \hat{V}_{h+1}(s, a) - f(\hat{\theta}_h, \phi(\cdot, \cdot)) \right| \leq f(\theta_{\hat{V}_h+1}) - f(\hat{\theta}_h, \phi(\cdot, \cdot)) + \epsilon_F.$$

By Lemma $D.2$, we have for any $\pi$,

$$v^\pi - v^\pi = - \sum_{h=1}^{H} \mathbb{E}_h[\ell_h(s_h, a_h)] + \sum_{h=1}^{H} \mathbb{E}_h[\ell_h(s_h, a_h)] \leq - \sum_{h=1}^{H} \mathbb{E}_h[\ell_h(s_h, a_h)] + \sum_{h=1}^{H} \mathbb{E}_h[\ell_h(s_h, a_h)]$$

$$\leq \sum_{h=1}^{H} \mathbb{E}_h[|f(\theta_{\hat{V}_h+1}, \phi(\cdot, \cdot)) - f(\hat{\theta}_h, \phi(\cdot, \cdot))|] + \sum_{h=1}^{H} \mathbb{E}_h[|f(\theta_{\hat{V}_h+1}, \phi(\cdot, \cdot)) - f(\hat{\theta}_h, \phi(\cdot, \cdot))|] + 2H \epsilon_F$$

$$\leq \sum_{h=1}^{H} \sqrt{\mathbb{E}_h[|f(\theta_{\hat{V}_h+1}, \phi(\cdot, \cdot)) - f(\hat{\theta}_h, \phi(\cdot, \cdot))|^2]} + \sum_{h=1}^{H} \sqrt{\mathbb{E}_h[|f(\theta_{\hat{V}_h+1}, \phi(\cdot, \cdot)) - f(\hat{\theta}_h, \phi(\cdot, \cdot))|^2]} + 2H \epsilon_F$$

$$\leq 2\sqrt{C_{\epsilon_F}} \sum_{h=1}^{H} \sqrt{\mathbb{E}_\mu, h[|f(\theta_{\hat{V}_h+1}, \phi(\cdot, \cdot)) - f(\hat{\theta}_h, \phi(\cdot, \cdot))|^2]} + 2H \epsilon_F$$

where the second inequality uses Cauchy inequality and the third one uses the definition of concentrability coefficient $2.2$.

Next, for VFQL, there is no pessimism therefore the quantity $\hat{B}$ in Lemma L.10 is zero, hence the covering number applied in Lemma G.1 is bounded by $C_{d, \log(K)} \leq \overline{O}(d)$ and

$$\mathbb{E}_\mu[\ell_h(\hat{\theta}_h)] - \mathbb{E}_\mu[\ell_h(\theta_{\hat{V}_h+1})] \leq \frac{36H^2(\log(1/\delta) + C_{d, \log K}) + \lambda C_\Theta^2}{K} + \frac{16H^3 \epsilon_F (\log(1/\delta) + C_{d, \log K})}{K} + 4H \epsilon_F.$$

Now leveraging equation 21 and equation 22 in Theorem G.2 to obtain

$$\mathbb{E}_\mu \left[ \left( f(\theta_{\hat{V}_h+1}, \phi(\cdot, \cdot)) - f(\theta_{\hat{V}_h+1}, \phi(\cdot, \cdot)) \right)^2 \right] \leq \mathbb{E}_\mu \left[ \ell_h(\hat{\theta}_h) - \frac{\lambda \|\hat{\theta}_h\|^2}{K} \right] - \mathbb{E}_\mu \left[ \ell_h(\theta_{\hat{V}_h+1}) - \frac{\lambda \|\theta_{\hat{V}_h+1}\|^2}{K} \right] + 2H \epsilon_F$$

$$\leq \frac{36H^2(\log(H/\delta) + C_{d, \log K}) + 2\lambda C_\Theta^2}{K} + b_{d, K, \epsilon_F} + 2H \epsilon_F.$$
Plug the above into equation 24, we obtain with probability $1 - \delta$, for all policy $\pi$,

$$v^\pi - v^\tilde{\pi} \leq 2\sqrt{c_{\text{eff}}H\sqrt{\frac{36H^2(\log(H/\delta) + C_{d,\log K}) + 2\lambda C_{\Theta}^2}{K}} + b_{d,\epsilon,\pi} + 2H\epsilon_F + 2H\epsilon_F}$$

$$= 2\sqrt{c_{\text{eff}}H\sqrt{\frac{36H^2(\log(H/\delta) + C_{d,\log K}) + 2\lambda C_{\Theta}^2}{K}} + \sqrt{\frac{16H^3\epsilon_F(1/\delta) + C_{d,\log K}}{K}} + 6H\epsilon_F + 2H\epsilon_F}$$

$$= \sqrt{c_{\text{eff}}H \cdot O \left( \frac{H^2d + \lambda C_{\Theta}^2}{K} + \sqrt{\frac{H^3d\epsilon_F}{K}} \right)} + O(\sqrt{c_{\text{eff}}H^3\epsilon_F} + H\epsilon_F)$$

This finishes the proof of Theorem 3.1.

### J Proofs for VAFQL

In this section, we present the analysis for variance-aware fitted Q learning (VAFQL). Throughout the whole section, we assume $\epsilon_F = 0$, i.e., the exact Bellman-Completeness holds. The algorithm is presented in the following. Before giving the proofs of Theorem 3, we first prove some useful lemmas.

**Algorithm 3 Variance-Aware Fitted Q Learning (VAFQL)**

1. **Input:** Split dataset $D = \{(s_h^k, a_h^k, r_h^k)\}_{k,h=1}^{K,H}$, $D' = \{(s_h^k, a_h^k, r_h^k)\}_{k,h=1}^{K,H}$. Require $\beta$.
2. **Initialization:** Set $\tilde{V}_{h+1}(\cdot) \leftarrow 0$. Denote $\phi_{h,k} := \phi(s_h^k, a_h^k)$, $\phi_{h,k} := \phi(s_h^k, a_h^k)$
3. **for** $h = H, H - 1, \ldots, 1$ **do**
   4. **Set** $u_h \leftarrow \arg\min_{\theta \in \Theta_h} \left\{ \sum_{k=1}^{K} \left[ f(\theta, \phi_{h,k}) - \tilde{V}_{h+1}(s_h^k) \right]^2 + \lambda \cdot ||\theta||_2^2 \right\}$
   5. **Set** $v_h \leftarrow \arg\min_{\theta \in \Theta_h} \left\{ \sum_{k=1}^{K} \left[ f(\theta, \phi_{h,k}) - \tilde{V}_{h+1}(s_h^k) \right]^2 + \lambda \cdot ||\theta||_2^2 \right\}$
   6. **Set** $[\text{Var}_{\tilde{V}_{h+1}}(\cdot, \cdot)](\cdot) \leftarrow [f(u_h, \phi(\cdot, \cdot))][H-h+1]^2 - [f(u_h, \phi(\cdot, \cdot))][H-h+1]^2$
   7. **Set** $[\tilde{\sigma}_h(\cdot, \cdot)]^2 \leftarrow \max(1, \text{Var}_{\tilde{V}_{h+1}}(\cdot, \cdot))$
   8. **Set** $\hat{\theta}_h \leftarrow \arg\min_{\theta \in \Theta_h} \left\{ \sum_{k=1}^{K} \left[ f(\theta, \phi_{h,k}) - \sigma_{h,k} - \tilde{V}_{h+1}(s_h^k) \right]^2 / \tilde{\sigma}_h^2(s_h^k, a_h^k) + \lambda \cdot ||\theta||_2^2 \right\}$
   9. **Set** $\Lambda_h \leftarrow \sum_{k=1}^{K} \nabla f(\hat{\theta}_h, \phi_{h,k})^T / \tilde{\sigma}_h^2(s_h^k, a_h^k) + \lambda \cdot \tilde{I}$
   10. **Set** $\Gamma_h(\cdot, \cdot) \leftarrow \beta \nabla \tilde{\sigma}_h(\cdot, \cdot) / \tilde{\sigma}_h(\cdot, \cdot) + \Lambda_h^{-1} \nabla \tilde{\sigma}_h(\cdot, \cdot) ( - \hat{\theta}_h(\cdot, \cdot) )$
   11. **Set** $\bar{Q}_h(\cdot, \cdot) \leftarrow f(\hat{\theta}_h, \phi(\cdot, \cdot)) - \Gamma_h(\cdot, \cdot), \bar{Q}_h(\cdot, \cdot) \leftarrow \min \{ \bar{Q}_h(\cdot, \cdot), H = h + 1 \}$
   12. **Set** $\bar{\pi}_h(\cdot | \cdot, \cdot) \leftarrow \arg\max_{\pi} \bar{Q}_h(\cdot, \cdot, \pi(\cdot, \cdot))$
13. **end for**
14. **Output:** $\{\bar{\pi}_h\}_{h=1}^H$.

#### J.1 Provable Efficiency for Variance-Aware Fitted Q Learning

Recall the objective

$$\ell_h(\theta) := \frac{1}{K} \sum_{k=1}^{K} \left[ f(\theta, \phi(s_h^k, a_h^k)) - r_h^k - \tilde{V}_{h+1}(s_h^k) \right]^2 / \tilde{\sigma}_h^2(s_h^k, a_h^k) + \lambda \cdot ||\theta||_2^2$$

Then by definition, $\hat{\theta}_h := \arg\min_{\theta \in \Theta_h} \ell_h(\theta)$ and $\theta_{\pi^{\tilde{V}_{h+1}}}$ satisfies $f(\theta_{\pi^{\tilde{V}_{h+1}}}, \phi) = \mathcal{P}_h(\tilde{V}_{h+1}(s_h^k))$ (recall $\epsilon_F = 0$). Therefore, in this case, we have the following lemma:

**Lemma J.1.** Fix $h \in [H]$. With probability $1 - \delta$,

$$\mathbb{E}_\mu[\ell_h(\hat{\theta}_h)] - \mathbb{E}_\mu[\ell_h(\theta_{\pi^{\tilde{V}_{h+1}}})] \leq \frac{36H^2(\log(1/\delta) + C_{d,\log K}) + \lambda C_{\Theta}^2}{K}$$

where the expectation over $\mu$ is taken w.r.t. $(s_h^k, a_h^k, s_{h+1}^k)^{k=1,\ldots,K}$ only (i.e., first compute $\mathbb{E}_\mu[\ell_h(\theta)]$ for a fixed $\theta$, then plug-in either $\hat{\theta}_{h+1}$ or $\theta_{\pi^{\tilde{V}_{h+1}}}$). Here

36
\[C_{d,\log(K)} := d \log(1+24C_\Theta(H+1)\kappa K) + d \log \left( 1 + 288H^2C_\Theta(\kappa_1 \sqrt{C_\Theta} + 2\sqrt{\kappa_1 \kappa_2 / \lambda})^2K^2 \right) + d^2 \log \left( 1 + 288H^2\sqrt{d\kappa_1^2K^2 / \lambda} \right) + d \log(1 + 16C_\Theta H^2\kappa_1 K) + d \log(1 + 32C_\Theta H^3\kappa_1 K).\]

**Proof of Lemma 11.** **Step 1:** Consider the case where \(\lambda = 0\). Indeed, fix \(h \in [H]\) and any function \(V(\cdot) \in \mathbb{R}_+^S\). Similarly, define \(f_V(s,a) := f(\theta V, \phi) = P_h V\). For any fixed \(\theta \in \Theta\), denote \(g(s,a) = f(\theta, \phi(s,a))\). Moreover, for any \(u, v \in \Theta\), define
\[
\sigma^2_{u,v}(\cdot, \cdot) := \max\{1, f(v, \phi(\cdot))\}_{[0, (H-1)2]} - \left[ f(u, \phi(\cdot))\right]_{[0, (H-1)2]}^2
\]
Then define (we omit the subscript \(u, v\) 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, \(X_k(g, V, f_V) := X(g(s_h^k, a_h^k), V(s_h^k+1), f_V(s_h^k, a_h^k), \sigma^2(s_h^k, a_h^k))\) are independent r.v.s and it holds
\[
\frac{1}{K} \sum_{k=1}^{K} X_k(g, V, f_V, \sigma^2) = \ell(g) - \ell(f_V). \tag{25}
\]
Next, the variance of \(X\) is bounded by
\[
\text{Var}[X(g, V, f_V, \sigma^2)] \leq \mathbb{E}_\mu [X(g, f, f_V, \sigma^2)]
\]
\[
=E_\mu \left[ \frac{(g(s_h,a_h) - r_h - V(s_h+1))^2 - (f_V(s_h,a_h) - r_h - V(s_h+1))^2}{\sigma^2(s_h,a_h)^2} \right]
\]
\[
\leq 4H^2 \cdot \mathbb{E}_\mu \left[ \frac{g(s_h,a_h) - f_V(s_h,a_h)^2}{\sigma^2(s_h,a_h)^2} \right]
\]
\[
= 4H^2 \cdot \mathbb{E}_\mu [X(g, f, f_V, \sigma^2)]
\]

\((*)\) follows from that
\[
\mathbb{E}_\mu \left[ \frac{f(\hat{\theta}_h, \phi(s_h,a_h)) - f(\theta V_{h+1}, \phi(s_h,a_h))}{\sigma^2(s_h,a_h)} \right] = \mathbb{E} \left[ f\left(\theta V_{h+1}, \phi(s_h,a_h)\right) - r_h - \hat{V}_{h+1}(s_h+1)\right] = 0.
\]
Therefore, by Bernstein inequality, with probability \(1 - \delta\),
\[
\mathbb{E}_\mu [X(g, f, f_V, \sigma^2)] - \frac{1}{K} \sum_{k=1}^{K} X_k(g, f, f_V, \sigma^2) \leq \sqrt{\frac{2 \text{Var}[X(g, f, f_V, \sigma^2)] \log(1/\delta)}{K}} + \frac{4H^2 \log(1/\delta)}{3K}
\]
\[
\leq \sqrt{\frac{8H^2 \mathbb{E}_\mu [X(g, f, f_V, \sigma^2)] \log(1/\delta)}{K}} + \frac{4H^2 \log(1/\delta)}{3K}.
\]
Now, if we choose \(g(s,a) := f(\hat{\theta}_h, \phi(s,a))\) and \(u = u_h, v = v_h\) from Algorithm 3, then \(\hat{\theta}_h\) minimizes \(f_h(\theta)\), therefore, it also minimizes \(\frac{1}{K} \sum_{k=1}^{K} X_i(\theta, \hat{V}_{h+1}, f_{\hat{V}_{h+1}}, \hat{\sigma}^2_h)\) and this implies
\[
\frac{1}{K} \sum_{k=1}^{K} X_k(\hat{\theta}_h, \hat{V}_{h+1}, f_{\hat{V}_{h+1}}, \hat{\sigma}^2_h) \leq \frac{1}{K} \sum_{k=1}^{K} X_k(\theta V_{h+1}, \hat{V}_{h+1}, f_{\hat{V}_{h+1}}, \hat{\sigma}^2_h) = 0.
\]
Thus, we obtain
\[
\mathbb{E}_\mu[X(\hat{\theta}_h, \hat{V}_{h+1}, f_{\hat{V}_{h+1}}, \sigma^2_h)] \leq \sqrt{\frac{8H^2 \cdot \mathbb{E}_\mu[X(\hat{\theta}_h, \hat{V}_{h+1}, f_{\hat{V}_{h+1}}, \sigma^2_h)] \log(1/\delta)}{K}} + \frac{4H^2 \log(1/\delta)}{3K}.
\]

However, the above does not hold with probability $1 - \delta$ since $\hat{\theta}_h$, $\sigma^2_h$ and $\hat{V}_{h+1} := \min\{\max_a f(\hat{\theta}_h, \phi(\cdot, a)) - \nabla f(\hat{\theta}_h, \phi(\cdot, a))^\top A : \nabla f(\phi(\cdot), H)\}$ (where $A$ is certain symmetric matrix with bounded norm) depend on $\theta_h$, $\hat{\theta}_{h+1}$ which are data-dependent. Therefore, we need to further apply covering Lemma L.11 and choose $\epsilon = O(1/K)$ and a union bound to obtain with probability $1 - \delta$,
\[
\mathbb{E}_\mu[X(\hat{\theta}_h, \hat{V}_{h+1}, f_{\hat{V}_{h+1}}, \sigma^2_h)] \leq \sqrt{\frac{8H^2 \cdot \mathbb{E}_\mu[X(\hat{\theta}_h, \hat{V}_{h+1}, f_{\hat{V}_{h+1}}, \sigma^2_h)](\log(1/\delta) + C_{d,\log K})}{K}} + \frac{4H^2(\log(1/\delta) + C_{d,\log K})}{3K}.
\]

where $C_{d,\log K} := d\log(1 + 24C_\phi(H + 1)\kappa_1 K) + d\log \left(1 + 288H^2C_\phi(\kappa_1 \sqrt{C_\phi} + 2\sqrt{\kappa_1\kappa_2}K^2)\right) + d^2\log \left(1 + 288H^2\sqrt{d\kappa_1^2K^2}/\lambda\right) + d\log(1 + 16C_\phi H^2\kappa_1 K) + d\log(1 + 32C_\phi H^3\kappa_1 K)$ (where we let $B = 1/\lambda$ since $\|\Lambda_h^{-1}\|_2 \leq 1/\lambda$). Solving this quadratic equation to obtain with probability $1 - \delta$,
\[
\mathbb{E}_\mu[X(\theta_h, \hat{V}_{h+1}, f_{\hat{V}_{h+1}})] \leq \frac{36H^2(\log(1/\delta) + C_{d,\log K})}{K}.
\]

Now according to equation 25, by definition we finally have with probability $1 - \delta$ (recall the expectation over $\mu$ is taken w.r.t. $(s_k^h, a_k^h, s_k^h) \in \mathcal{K}$)
\[
\mathbb{E}_\mu[\ell_h(\hat{\theta}_h) - \mathbb{E}_\mu[\ell_h(\theta_{\hat{V}_{h+1}})]] = \mathbb{E}_\mu[X(\theta_h, \hat{V}_{h+1}, f_{\hat{V}_{h+1}})] \leq \frac{36H^2(\log(1/\delta) + C_{d,\log K})}{K}
\]
where we used $f(\theta_{\hat{V}_{h+1}}, \phi) = P_h \hat{V}_{h+1} = f_{\hat{V}_{h+1}}$.

**Step 2.** If $\lambda > 0$, there is only extra term $\frac{\lambda}{K} \left(\|\hat{\theta}_h\|_2 - \|\theta_{\hat{V}_{h+1}}\|_2\right) \leq \frac{\lambda}{K} \|\hat{\theta}_h\|_2 \leq \frac{\lambda C_{\phi}^2}{K}$ in addition to above. This finishes the proof.

- **Theorem J.2 (Provable efficiency for VAFQL).** Let $C_{d,\log K}$ be the same as Lemma J.1. Then, with probability $1 - \delta$

\[
\|\hat{\theta}_h - \theta_{\hat{V}_{h+1}}\|_2 \leq \sqrt{\frac{36H^4(\log(H/\delta) + C_{d,\log K} + 2\lambda C_{\phi}^2)}{K}} \quad \forall h \in [H].
\]

**Proof of Theorem J.2.** Apply a union bound in Lemma J.1, we have with probability $1 - \delta$,
\[
\mathbb{E}_\mu[\ell_h(\hat{\theta}_h)] - \mathbb{E}_\mu[\ell_h(\theta_{\hat{V}_{h+1}})] \leq \frac{36H^2(\log(H/\delta) + C_{d,\log K}) + \lambda C_{\phi}^2}{K} \quad \forall h \in [H]
\]

\[
\Rightarrow \mathbb{E}_\mu[\ell_h(\hat{\theta}_h) - \frac{\lambda}{K} \|\hat{\theta}_h\|_2^2 - \mathbb{E}_\mu[\ell_h(\theta_{\hat{V}_{h+1}}) - \frac{\lambda}{K} \|\theta_{\hat{V}_{h+1}}\|_2^2] \leq \frac{36H^2(\log(H/\delta) + C_{d,\log K} + 2\lambda C_{\phi}^2)}{K}\]

Now we prove for all $h \in [H]$,
\[
\mathbb{E}_\mu \left[ (f(\hat{\theta}_h, \phi(\cdot, \cdot)) - f(\theta_{\hat{V}_{h+1}}, \phi(\cdot, \cdot))^2 \right] = \mathbb{E}_\mu \left[ \ell_h(\hat{\theta}_h) - \frac{\lambda}{K} \|\hat{\theta}_h\|_2^2 \right] - \mathbb{E}_\mu \left[ \ell_h(\theta_{\hat{V}_{h+1}}) - \frac{\lambda}{K} \|\theta_{\hat{V}_{h+1}}\|_2^2 \right].
\]
Indeed, identical to equation 26,

\[
\begin{align*}
\mathbb{E}_\mu \left[ \ell_h(\tilde{\theta}_h) - \frac{\lambda \|\tilde{\theta}_h\|^2}{K} \right] &- \mathbb{E}_\mu \left[ \ell_h(\theta_{\hat{V}h+1}) - \frac{\lambda \|\theta_{\hat{V}h+1}\|^2}{K} \right] = \mathbb{E}_\mu[X(\tilde{\theta}_h, \hat{V}_{h+1}, f_{\hat{V}_{h+1}})] \\
= &\mathbb{E}_\mu \left( \frac{f(\tilde{\theta}_h, \phi(s_h, a_h)) - r_h - \hat{V}_{h+1}(s_h, a_h)}{\sigma^2_h(s_h, a_h)} \right)^2 \frac{1}{\sigma^2_h(s_h, a_h)} \\
+ &\mathbb{E}_\mu \left( \frac{f(\tilde{\theta}_h, \phi(s_h, a_h)) - f(\theta_{\hat{V}h+1}, \phi(s_h, a_h)) \cdot \left( \theta_{\hat{V}h+1} \phi(s_h, a_h) - r_h - \hat{V}_{h+1}(s_h, a_h) \right)}{\sigma^2_h(s_h, a_h)} \right)^2 \frac{1}{\sigma^2_h(s_h, a_h)} \\
= &\mathbb{E}_\mu \left( \frac{f(\tilde{\theta}_h, \phi(s_h, a_h)) - f(\theta_{\hat{V}h+1}, \phi(s_h, a_h)) \cdot \left( \theta_{\hat{V}h+1} \phi(s_h, a_h) - r_h - \hat{V}_{h+1}(s_h, a_h) \right)}{\sigma^2_h(s_h, a_h)} \right)^2 \frac{1}{\sigma^2_h(s_h, a_h)} \\
= &\mathbb{E}_\mu \left( \frac{f(\tilde{\theta}_h, \phi(s_h, a_h)) - f(\theta_{\hat{V}h+1}, \phi(s_h, a_h)) \cdot \left( \theta_{\hat{V}h+1} \phi(s_h, a_h) - r_h - \hat{V}_{h+1}(s_h, a_h) \right)}{\sigma^2_h(s_h, a_h)} \right)^2 \frac{1}{\sigma^2_h(s_h, a_h)} \\
&\mathbb{E}_\mu \left( \frac{f(\tilde{\theta}_h, \phi(s_h, a_h)) - f(\theta_{\hat{V}h+1}, \phi(s_h, a_h)) \cdot \left( \theta_{\hat{V}h+1} \phi(s_h, a_h) - r_h - \hat{V}_{h+1}(s_h, a_h) \right)}{\sigma^2_h(s_h, a_h)} \right)^2 \frac{1}{\sigma^2_h(s_h, a_h)} \\
\end{align*}
\]

where the third identity uses law of total expectation and that \( \mu \) is taken w.r.t. \( s_h, \theta_h, s_{h+1} \) only (recall Lemma J.1) so the \( \sigma^2_h \) can be move outside of the conditional expectation. \( \Box \) The fourth identity uses the definition of \( \theta_{\hat{V}h+1} \) since \( f(\theta_{\hat{V}h+1}, \phi(s, a)) = \mathcal{P}_{h,s,a} \hat{V}_{h+1} \).

Then we have

\[
\begin{align*}
\mathbb{E}_\mu \left[ \frac{f(\tilde{\theta}_h, \phi(s, a)) - f(\theta_{\hat{V}h+1}, \phi(s, a))}{\sigma^2_h(s_h, a_h)} \right]^2 \frac{1}{\sigma^2_h(s_h, a_h)} &\geq \mathbb{E}_\mu \left[ \frac{f(\tilde{\theta}_h, \phi(s, a)) - f(\theta_{\hat{V}h+1}, \phi(s, a))}{\sigma^2_h(s_h, a_h)} \right]^2 \frac{1}{\sigma^2_h(s_h, a_h)} \\
&\geq \kappa^2 \frac{1}{\kappa^2} \left( \frac{\lambda}{K} \right)^2 \frac{1}{\kappa^2} \frac{1}{K}
\end{align*}
\]

where the third identity uses \( \mu \) is over \( s_h, a_h \) only and the last one uses \( \sigma^2_h(s, a) \leq H^2 \). Combine the above with equation 27 and equation 28, we obtain the stated result.

**Theorem J.3** (Provable efficiency of VAFQL (Part II)). Let \( C_{d, \log \kappa} \) be the same as Lemma J.1. Furthermore, suppose \( \lambda \leq 1/2C_{d, \log \kappa} \) and \( K \geq \max \left\{ 512 \frac{\log(\frac{H}{2\delta})}{\kappa} + d \log(1 + 2\kappa \lambda_{d, \log \kappa} K^2), \frac{K}{\kappa} \right\} \).

Then, with probability \( 1 - \delta \), \( h \in [H] \)

\[
\sup_h \left| f(\tilde{\theta}_h, \phi(s, a)) - f(\theta_{\hat{V}h+1}, \phi(s, a)) \right| \leq \left( \kappa_1 H \sqrt{36H^4(\log(H/\delta) + C_{d, \log \kappa}) + 2\kappa^2 C_{d, \log \kappa}^2} + \frac{2dH^3\kappa_1}{\sqrt{\kappa}} \right) \sqrt{\frac{1}{K}} + O\left( \frac{1}{K} \right).
\]

Furthermore, we have with probability \( 1 - \delta \),

\[
\sup_h \| \hat{V}_h - V^*_h \|_\infty \leq \left( \kappa_1 H \sqrt{36H^4(\log(H/\delta) + C_{d, \log \kappa}) + 2\kappa^2 C_{d, \log \kappa}^2} + \frac{2dH^3\kappa_1}{\sqrt{\kappa}} \right) \sqrt{\frac{1}{K}} + O\left( \frac{1}{K} \right)
\]

where \( \tilde{O} \) absorbs Polylog terms and higher order terms. Lastly, it also holds for all \( h \in [H] \), w.p. \( 1 - \delta \)

\[
\left\| \tilde{\theta}_h - \theta^*_h \right\|_2 \leq \left( \kappa_1 H \sqrt{72H^4(\log(H^2/\delta) + C_{d, \log \kappa}) + 4\kappa^2 C_{d, \log \kappa}^2} + \frac{4H^3d\kappa_1}{\kappa} \right) \sqrt{\frac{1}{K}} + O\left( \frac{1}{K} \right)
\]

\[\text{Recall } \sigma^2_h \text{ computed in Algorithm 3 uses an independent copy } D'.\]
Proof of Theorem J.3. Step1: we show the first result.

We prove this by backward induction. When \( h = H + 1 \), by convention \( f(\hat{\theta}_h, \phi(s, a)) = f(\theta^*_h, \phi(s, a)) = 0 \) so the base case holds. Suppose for \( h + 1 \), with probability \( 1 - (H - h)\delta \),
\[
\sup_{s, a} \left| f(\hat{\theta}_h, \phi(s, a)) - f(\theta^*_h, \phi(s, a)) \right| \leq C_{h+1} \sqrt{\frac{1}{K}},
\]
we next consider the case for \( t = h \).

On one hand, by Theorem J.2, we have with probability \( 1 - \delta/2 \),
\[
\sup_{s, a} \left| f(\hat{\theta}_h, \phi(s, a)) - f(\theta^*_h, \phi(s, a)) \right| 
\leq \sup_{s, a} \left| f(\hat{\theta}_h, \phi(s, a)) - f(\theta_TV_{h+1}, \phi(s, a)) \right| + \sup_{s, a} \left| f(\theta_TV_{h+1}, \phi(s, a)) - f(\theta^*_h, \phi(s, a)) \right|
\leq \sup_{s, a} \left| \nabla f(\xi, \phi(s, a))^\top (\hat{\theta}_h - \theta_TV_{h+1}) \right| + \sup_{s, a} \left| f(\theta_TV_{h+1}, \phi(s, a)) - f(\theta^*_h, \phi(s, a)) \right|
\leq \kappa_1 \sup_{s, a} \left| \nabla f(\xi, \phi(s, a)) \right| + \sup_{s, a} \left| f(\theta_TV_{h+1}, \phi(s, a)) - f(\theta^*_h, \phi(s, a)) \right|
\leq \kappa_1 \sqrt{\frac{36H^4(\log(H/\delta) + C_d \log K)}{\kappa K}} + \sqrt{\frac{2\lambda^2 \delta}{\kappa K}} \leq \sqrt{\frac{36H^4(\log(H/\delta) + C_d \log K)}{\kappa K}} + \sqrt{\frac{2\lambda^2 \delta}{\kappa K}} \leq \kappa_1 \sqrt{\frac{H^2 \lambda^{-1}}{\kappa K}} + \sqrt{\frac{2\lambda^2 \delta}{\kappa K}}.
\]

Recall we have the form \( \hat{V}_{h+1}(\cdot) : = \min \{ \max_a f(\theta_{h+1}, \phi(\cdot, a)) - \Gamma_H(\cdot, a), H \} \) and \( V_{h+1}^*(\cdot) = \max_a f(\theta_{h+1}^*, \phi(\cdot, a)) = \min \{ \max_a f(\theta_{h+1}^*, \phi(\cdot, a)) \} \),
\[
\left| \hat{V}_{h+1} - V_{h+1}^* \right| \leq \sup_{s, a} \left| f(\theta_{h+1}, \phi(s, a)) - f(\theta_{h+1}^*, \phi(s, a)) \right| + \sup_{h, s, a} \Gamma_H(s, a).
\]

Note the above holds true for any generic \( \Gamma_H(s, a) \). In particular, according to Algorithm 3, we specify
\[
\Gamma_h(\cdot, \cdot) = d_v \nabla \phi f(\hat{\theta}_h, \phi(\cdot, \cdot)) \Lambda_h^{-1} \nabla \phi f(\hat{\theta}_h, \phi(\cdot, \cdot)) + \tilde{O}(\frac{1}{K})
\]
and by Lemma L.5, with probability \( 1 - \delta \) (note here \( \kappa_1 \lambda^{-1} \) is replaced by \( \Lambda_h^{-1} \) and \( \frac{\|A_h^{-1}\|}{2} \leq H^2/\kappa \),
\[
\Gamma_h \leq \frac{2dH^2 \kappa_1}{\sqrt{\kappa K}} + O(\frac{1}{K})
\]
and by a union bound this implies with probability \( 1 - (H - h + 1)\delta \),
\[
\sup_{s, a} \left| f(\hat{\theta}_h, \phi(s, a)) - f(\theta^*_h, \phi(s, a)) \right| 
\leq C_{h+1} \sqrt{\frac{1}{K}} + \kappa_1 \sqrt{\frac{36H^4(\log(H/\delta) + C_d \log K)}{\kappa K}} + \sqrt{\frac{2\lambda^2 \delta}{\kappa K}} + O(\frac{1}{K}) =: C_h \sqrt{\frac{1}{K}}.
\]

Solving for \( C_h \), we obtain \( C_h \leq \kappa_1 H \sqrt{\frac{36H^4(\log(H/\delta) + C_d \log K)}{\kappa K}} + \frac{2dH^2 \kappa_1}{\sqrt{\kappa K}} + O(\frac{1}{K}) \) for all \( H \). By a union bound (replacing \( \delta \) by \( \delta/H \)), we obtain the stated result.

Step2: Utilizing the intermediate result equation 29, we directly have with probability \( 1 - \delta \),
\[
\sup_{h} \left| \hat{V}_{h} - V_{h}^* \right| \leq \sup_{s, a} \left| f(\hat{\theta}_h, \phi(s, a)) - f(\theta^*_h, \phi(s, a)) \right| + \frac{2dH^2 \kappa_1}{\sqrt{\kappa K}} + O(\frac{1}{K}),
\]
where \( \sup_{s, a} \left| f(\hat{\theta}_h, \phi(s, a)) - f(\theta^*_h, \phi(s, a)) \right| \) can be bounded using Step1.

Step3: Denote \( M := \left( \kappa_1 H \sqrt{\frac{36H^4(\log(H^2/\delta) + C_d \log K)}{\kappa K}} + \frac{2dH^2 \kappa_1}{\sqrt{\kappa K}} \right) \sqrt{\frac{1}{K}} + O(\frac{1}{K}) \), then by
Step1 we have with probability \( 1 - \delta \) (here \( \xi \) is some point between \( \hat{\theta}_h \) and \( \theta^*_h \) for all \( h \in [H] \)
\[
M^2 \geq \sup_{s, a} \left| f(\hat{\theta}_h, \phi(s, a)) - f(\theta^*_h, \phi(s, a)) \right| \geq \mathbb{E}_\mu \left[ \left| f(\hat{\theta}_h, \phi(s, a)) - f(\theta^*_h, \phi(s, a)) \right| \right] \geq \kappa \left\| \hat{\theta}_h - \theta^*_h \right\|_2^2
\]
where the last step is by Assumption 2.3. Solving this to obtain the stated result. □

J.2 Bounding \( \hat{\sigma}_h^2 - \sigma_h^2 \)

Recall the definition \( \sigma_h^2(\cdot, \cdot) := \max\{1, [\text{Var}_{P_h}V^2_{h+1}(\cdot, \cdot)]\} \). In this section, we bound the term 
\[ |\hat{\sigma}_h^2(\cdot, \cdot) - \sigma_h^2(\cdot, \cdot)| \] 
and 
\[ \hat{\sigma}_h^2(\cdot, \cdot) := \max\{1, [f(u_h, \phi(\cdot, \cdot))_{[0,H-h-1]}] - [f(u_h, \phi(\cdot, \cdot))_{[0,H-h-1]}] \}
\]
where 
\[ \hat{\sigma}_h^2(\cdot, \cdot) := \max\{1, [f(h, \phi(\cdot, \cdot))]_{[0,H-h+1]} - [f(h, \phi(\cdot, \cdot))]_{[0,H-h+1]} \}
\]

and true parameters \( u_h^*, v_h^* \) satisfy 
\[ f(u_h^*, \phi(\cdot, \cdot)) = E_{P(s', \cdot)}[V^2_h(s')], f(v_h^*, \phi) = E_{P(s', \cdot)}[V^2_h(s')] \]. Furthermore, we define 
\[ \sigma_{h+1}^2(\cdot, \cdot) := \max\{1, [\text{Var}_{P_{h+1}}V_{h+1}^2(\cdot, \cdot)]\} \]
and the parameter Expectation operator \( \mathbb{E} \): \( V \in \mathbb{R}^S \to \theta_{1V} \in \Theta \) such that: 
\[ f(\theta_{1V}, \phi) = \mathbb{E}_{P_{\cdot}}[V(s')], \forall \|V\| \leq B_X. \]

Note \( \theta_{1V} \in \Theta \) by Bellman completeness, reward \( r \) is constant and differentiability (Definition 1.1) is an additive closed property. By definition, 
\[ |\hat{\sigma}_h^2 - \sigma_h^2| \leq |f(h, \phi(\cdot, \cdot))_{[0,H-h-1]} - [f(h, \phi(\cdot, \cdot))]_{[0,H-h-1]}| 
\]
and 
\[ |\sigma_h^2(\cdot, \cdot) - \sigma_h^2(\cdot, \cdot)| \leq |f(h, \phi(\cdot, \cdot))_{[0,H-h-1]} - [f(h, \phi(\cdot, \cdot))]_{[0,H-h-1]}| 
\]

We first give the following result.

**Lemma J.4.** Suppose \( \lambda \leq 1/2C_0^2 \) and \( K \geq \max \left\{ 512\kappa^2 \left( \log(\frac{2d}{\delta}) + d \log(1 + \frac{4\kappa^3C_0K^3}{\lambda^2}) \right) \right\} \).

Then, with probability \( 1 - \delta \), \( \forall h \in [H], \)
\[ \|u_h - \theta_{\mathbb{E}V_h}(\cdot, \cdot)\|_2 \leq \sqrt{\frac{36H^2(\log(H/\delta) + \widetilde{O}(d^2)) + 2\lambda C_0^2}{K}} \kappa, \forall h \in [H], \]
\[ \|v_h - \theta_{\mathbb{E}V_h^2}(\cdot, \cdot)\|_2 \leq \sqrt{\frac{36H^4(\log(H/\delta) + \widetilde{O}(d^2)) + 2\lambda C_0^2}{K}} \kappa, \forall h \in [H], \]
and
\[ \sup_{s,a} |f(u_h, \phi(s, a)) - f(u_h^*, \phi(s, a))| \leq \left( \kappa_2 \sqrt{\frac{36H^2(\log(H^2/\delta) + \widetilde{O}(d^2)) + 2\lambda C_0^2}{K}} + \frac{2H^2 d_{\mathcal{S}_1}}{\sqrt{K}} \right) \sqrt{\frac{1}{K}} + O(\frac{1}{K}), \]
\[ \sup_{s,a} |f(v_h, \phi(s, a)) - f(v_h^*, \phi(s, a))| \leq \left( \kappa_2 \sqrt{\frac{36H^4(\log(H^2/\delta) + \widetilde{O}(d^2)) + 2\lambda C_0^2}{K}} + \frac{2H^4 d_{\mathcal{S}_1}}{\sqrt{K}} \right) \sqrt{\frac{1}{K}} + O(\frac{1}{K}). \]

The above directly implies for all \( h \in [H] \), with probability \( 1 - \delta \), 
\[ |\sigma_h^2(\cdot, \cdot) - \hat{\sigma}_h^2(\cdot, \cdot)| \leq \left( 3\kappa_1 H^2 \left( \frac{36H^2(\log(H^2/\delta) + \widetilde{O}(d^2)) + 2\lambda C_0^2}{K} + \frac{6H^2 d_{\mathcal{S}_1}}{\sqrt{K}} \right) \right) \sqrt{\frac{1}{K}} + O(\frac{1}{K}), \]
\[ |\hat{\sigma}_h^2 - \sigma_{h+1}^2| \leq 3H \kappa_1 \sqrt{\frac{36H^4(\log(H/\delta) + \widetilde{O}(d^2)) + 2\lambda C_0^2}{K}}. \]
Proof of Lemma J.4. In fact, the proof follows a reduction from the provable efficiency procedure conducted in Section G. This is due to the regression procedure in equation 30 is the same as the procedure equation 17 except the parameter Bellman operator $T$ is replaced by the parameter expectation operator $\mathcal{J}$ (recall here $\hat{\phi}_{h,k}$ uses the independent copy $D'$ and $O(d^2)$ comes from the covering argument.). Concretely, the $X(g, V, f_V)$ used in Lemma G.1 will be modified to $X(g, V, f_V) = (g(s, a) - V(s'))^2 - (f(\theta_{jV}, \phi(s, a)) - V(s'))^2$ by removing reward information and the decomposition
\[
\mathbb{E}_\mu [(g(s_h, a_h) - V(s_{h+1}))^2 - (f(\theta_{jV}, \phi(s_h, a_h)) - V(s_{h+1}))^2] = \mathbb{E}_\mu [(g(s_h, a_h) - f(\theta_{jV}, \phi(s_h, a_h)))^2]
\]
holds true. Then with probability $1 - \delta$,
\[
|\sigma_h^2 - \bar{\sigma}_{h+1}^2| \leq |f(v_h, \phi) - f(v_h, \phi)| + 2H \cdot |f(u_h, \phi) - f(v_h, \phi)|
\leq \left( 3\kappa_1 H^2 \sqrt{\frac{36H^4(\log(H/\delta) + O(d^2)) + 2\lambda C_\phi^2}{\kappa} + \frac{6H^4 d_2 \kappa_1}{\sqrt{\kappa}}} \right) \sqrt{\frac{1}{K}} + O\left( \frac{1}{K} \right),
\]
and
\[
|\bar{\sigma}_{h+1}^2 - \sigma_{h+1}^2| \leq |f(v_h, \phi) - f(\theta_{j\hat{\phi}_{h+1}}, \phi)| + 2H \cdot |f(u_h, \phi) - f(\theta_{j\hat{\phi}_{h+1}}, \phi)|
\leq \kappa_1 \left\| v_h - \theta_{j\hat{\phi}_{h+1}} \right\|_2 + 2H \kappa_1 \left\| u_h - \theta_{j\hat{\phi}_{h+1}} \right\|_2
\leq 3H \kappa_1 \frac{\sqrt{36H^4(\log(H/\delta) + O(d^2)) + 2\lambda C_\phi^2}}{\kappa K}.
\]

\[\square\]

J.3 Proof of Theorem 4.1

In this section, we sketch the proof of Theorem 4.1 since the most components are identical to Theorem 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
\[
\nabla_{\theta} \left\{ \sum_{k=1}^{K} \left[ \frac{f(\theta, \phi_{h,k}) - r_{h,k} - \hat{V}_{h+1}(s_{h+1}^{k})}{\sigma_h^2(s_{h+1}^{k}, a_{h}^{k})} \right] + \lambda \cdot \| \theta \|_2^2 \right\}_{\theta = \hat{\theta}_h} = 0, \ \forall h \in [H].
\]

Therefore, if we define the quantity $Z_h(\cdot, \cdot) \in \mathbb{R}^d$ as
\[
Z_h(\theta) = \nabla_{\theta} \left\{ \sum_{k=1}^{K} \frac{f(\theta, \phi_{h,k}) - r_{h,k} - V(s_{h+1}^{k})}{\sigma(s_{h+1}^{k}, a_{h}^{k})} \right\} + \lambda \cdot \| \theta \|_2^2, \ \forall \theta \in \Theta, \| V \|_2 \leq H,
\]
then we have
\[
Z_h(\theta) = 0.
\]

According to the regression oracle (Line 8 of Algorithm 3), the estimated Bellman operator $\hat{P}_h$ maps $\hat{V}_{h+1}$ to $\hat{\theta}_h$, i.e. $\hat{P}_h \hat{V}_{h+1} = f(\hat{\theta}_h, \phi)$. Therefore (recall Definition D.1)
\[
P_h \hat{V}_{h+1}(s, a) - \hat{P}_h \hat{V}_{h+1}(s, a) = P_h \hat{V}_{h+1}(s, a) - f(\hat{\theta}_h, \phi(s, a))
= f(\theta_{j\hat{\phi}_{h+1}}, \phi(s, a)) - f(\theta_{j\hat{\phi}_{h+1}}, \phi(s, a))
= f(\theta_{j\hat{\phi}_{h+1}}, \phi(s, a)) \left( \theta_{j\hat{\phi}_{h+1}} - \hat{\theta}_h \right) + \text{Hot}_{h,1}.
\]

where we apply the first-order Taylor expansion for the differentiable function $f$ at point $\hat{\theta}_h$ and $\text{Hot}_{h,1}$ is a higher-order term. Indeed, the following Lemma E.1 bounds the $\text{Hot}_{h,1}$ term with $O\left( \frac{1}{K} \right)$.  

42
Lemma J.5. Recall the definition (from the above decomposition) \( \text{Hot}_{h, 1} := f(\hat{\theta}_{\hat{V}_{h+1}}, \phi(s, a)) - f(\hat{\theta}_h, \phi(s, a)) - \nabla f(\hat{\theta}_h, \phi(s, a)) \left( \hat{\theta}_{\hat{V}_{h+1}} - \hat{\theta}_h \right) \), then with probability \( 1 - \delta \),

\[
|\text{Hot}_{h, 1}| \leq \tilde{O}(\frac{1}{K}), \quad \forall h \in [H].
\]

**Proof.** The proof is identical to that of Lemma E.1 but with the help of Lemma J.2. \( \square \)

Next, according to the expansion of \( Z_h(\theta|V_{h+1}, \sigma_{h}^2) \), we have

\[
\nabla f(\hat{\theta}_h, \phi(s, a)) \left( \hat{\theta}_{\hat{V}_{h+1}} - \hat{\theta}_h \right) = I_1 + I_2 + I_3 + \text{Hot}_2,
\]

where

\[
\text{Hot}_2 := \nabla f(\hat{\theta}_h, \phi(s, a)) \Lambda_h^{-1} \left[ \tilde{R}_K(\hat{\theta}_{\hat{V}_{h+1}}) + \lambda \hat{\theta}_{\hat{V}_{h+1}} \right]
\]

\[
\Delta \Lambda_h = \sum_{k=1}^{K} \left( f(\hat{\theta}_h, \phi, h, k) - r_{h,k} - \hat{V}_{h+1}(s_{h+1}) \right) \cdot \frac{\nabla \phi(\hat{\theta}_h, \phi, h,k)}{\sigma^2(h, a_{h,k})} \]

\[
\Lambda_h = \sum_{k=1}^{K} \frac{\nabla \phi(\hat{\theta}_h, \phi, h,k) \cdot \nabla \phi(\hat{\theta}_h, \phi, h,k)}{\sigma^2(h, a_{h,k})} + \lambda I_d
\]

\[
\tilde{R}_K(\theta|\hat{V}_{h+1}) = \Delta \Lambda_h \theta - \lambda \hat{\theta}_{\hat{V}_{h+1}} + R_K(\theta|\hat{V}_{h+1})
\]

where \( R_K(\theta|\hat{V}_{h+1}) \) is the second order residual that is bounded by \( \tilde{O}(1/K) \) and

\[
I_1 = \nabla f(\hat{\theta}_h, \phi(s, a)) \Lambda_h^{-1} \left( f(\theta_{TV^{+}_{h+1}}, \phi, h, k) - r_{h,k} - V_{h+1}(s_{h+1}) \right) \cdot \frac{\nabla \phi(\hat{\theta}_h, \phi, h,k)}{\sigma^2(h, a_{h,k})}
\]

\[
I_2 = \nabla f(\hat{\theta}_h, \phi(s, a)) \Lambda_h^{-1} \sum_{k=1}^{K} \left( f(\theta_{TV^{+}_{h+1}}, \phi, h, k) - f(\theta_{TV^{+}_{h+1}}, \phi, h, k) - \hat{V}_{h+1}(s_{h+1}) \right) \cdot \frac{\nabla \phi(\hat{\theta}_h, \phi, h,k)}{\sigma^2(h, a_{h,k})}
\]

\[
I_3 = \nabla f(\hat{\theta}_h, \phi(s, a)) \Lambda_h^{-1} \sum_{k=1}^{K} \left( f(\theta_{TV^{+}_{h+1}}, \phi, h, k) - r_{h,k} - \hat{V}_{h+1}(s_{h+1}) \right) \cdot \frac{\nabla \phi(\hat{\theta}_h, \phi, h,k) - \nabla \phi(\hat{\theta}_h, \phi, h,k)}{\sigma^2(h, a_{h,k})}
\]

Similar to the PFQL case, \( I_2, I_3, \) \( \text{Hot}_2 \) can be bounded to have order \( \tilde{O}(1/K) \) via provably efficiency theorems in Section J.1 and in particular, the inclusion of \( z_h^{2} \) will not cause additional order in \( d \).

Now we prove the dominate term \( I_1 \).

**Lemma J.6.** With probability \( 1 - \delta \),

\[
|I_1| \leq 4Hd \left\| \nabla f(\hat{\theta}_h, \phi(s, a)) \right\|_{\Sigma_h} \cdot C_{\delta, \log K} + \tilde{O}(\frac{K_1}{\sqrt{K}}),
\]

where \( C_{\delta, \log K} \) only contains Polylog terms.

**Proof of Lemma J.6.** First of all, by Cauchy-Schwarz inequality, we have

\[
|I_1| \leq \left\| \nabla f(\hat{\theta}_h, \phi(s, a)) \right\|_{\Lambda_h^{-1}} \sum_{k=1}^{K} \left( f(\theta_{TV^{+}_{h+1}}, \phi, h, k) - r_{h,k} - V_{h+1}(s_{h+1}) \right) \cdot \frac{\nabla \phi(\hat{\theta}_h, \phi, h,k)}{\sigma^2(h, a_{h,k})}
\]

(33)

Recall that \( z_{u,v}^2(\cdot, \cdot) := \max\{1, f(u, \phi(\cdot, \cdot))|_{[0,H-H+1]} - [f(u, \phi(\cdot, \cdot))|_{[0,H-H+1]} \}^2 \).

\footnote{Note in Lemma L.11, we only have additive terms that has the same order has Lemma L.10.}
Step1. Let the fixed $\theta \in \Theta$ be arbitrary and fixed $u, v$ such that $\sigma_{uv}^2(\phi, \cdot) \geq \frac{1}{4} \sigma_{uv}^2(\phi, \cdot)$ and define $x_k(\theta, u, v) = \nabla \theta f(\theta, \phi_{h,k})/\sigma_{uv}(s_h^k, a_h^k)$. Next, define $G_{uv}(\theta) = \sum_{k=1}^{K} \nabla f(\theta, \phi(s_h^k, a_h^k)) \cdot \nabla f(\theta, \phi(s_h^k, a_h^k))$ and define $\eta_k := \{ f(\theta, \phi(s_h^k, a_h^k)) \}$ and define $\sigma_{uv}(s_h^k, a_h^k)$, then $\|x_k\|_2 \leq \kappa_1$. Also denote $\eta_k := \{ f(\theta, \phi(s_h^k, a_h^k)) - r_{h,k} - V_{h+1}(s_h^k, a_h^k) \}$ and define $\sigma_{uv}(s_h^k, a_h^k)$, then $\E[\eta_k | s_h^k, a_h^k] = 0$ and

$$ \text{Var}[\eta_k | s_h^k, a_h^k] = \frac{\text{Var}[f(\theta, \phi(s_h^k, a_h^k)) - r_{h,k} - V_{h+1}(s_h^k, a_h^k)]}{\sigma_{uv}(s_h^k, a_h^k)} \leq 2\text{Var}[f(\theta, \phi(s_h^k, a_h^k)) - r_{h,k} - V_{h+1}(s_h^k, a_h^k)]/\sigma_{uv}(s_h^k, a_h^k) \leq \frac{2}{\sigma_{uv}(s_h^k, a_h^k)} \leq 2,$$

then by Self-normalized Bernstein’s inequality (Lemma L.4), with probability $1 - \delta$,

$$ \left\| \sum_{k=1}^{K} x_k(\theta, u, v) \eta_k \right\|_{G(\theta, u, v)^{-1}} \leq 16 \left\{ \log \left( 1 + \frac{K \eta_k^2}{2 \lambda d} \right) \cdot \log \left( \frac{4K^2}{\delta} \right) + \zeta \log \left( \frac{4K^2}{\delta} \right) \right\} \leq \mathcal{O}(\sqrt{d})$$

where $|\eta_k| \leq \zeta$ with $\zeta = 2\max_{s, a, s'} |f(\theta, \phi(s, a)) - r_{h,k} - V_{h+1}(s')|/\sigma_{uv}(s, a)$ and the last inequality uses $\sqrt{d} \geq \mathcal{O}(\zeta)$.

Step2. Define $h(\theta, u, v) := \sum_{k=1}^{K} x_k(\theta, u, v) \eta_k(\theta, u, v)$ and $H(\theta, u, v) := \|h(\theta, u, v)\|_{G_{uv}(\theta)^{-1}}$, then $\|h(\theta_1, u_1, v_1) - h(\theta_2, u_2, v_2)\|_2 \leq K \max \|x_k \cdot \eta_k\|_2 = \sum_{k=1}^{K} x_k(\theta_1, u_1, v_1) - x_k(\theta_2, u_2, v_2)\|_2 \leq K \max \\|x_k \cdot \eta_k\|_2 \leq K \kappa_1 \|h_1 - h_2\|_2 + K H \kappa_1 \|\sigma_{uv}^2 - \sigma_{uv}^2\|_2$

Furthermore,

$$ \|G_{uv}(\theta_1, u_1, v_1)^{-1} - G_{uv}(\theta_2, u_2, v_2)^{-1}\|_2 \leq \|G_{uv}(\theta_1, u_1, v_1)^{-1}\|_2 \|G_{uv}(\theta_2, u_2, v_2)^{-1}\|_2 \leq \left\| \sum_{k=1}^{K} x_k(\theta_1, u_1, v_1) - x_k(\theta_2, u_2, v_2) \right\|_2 \leq K \kappa_1 \|h_1 - h_2\|_2 + K H \kappa_1 \|\sigma_{uv}^2 - \sigma_{uv}^2\|_2$$

All the above imply

$$ \|H(\theta_1, u_1, v_1) - H(\theta_2, u_2, v_2)\| \leq \sqrt{\|h(\theta_1, u_1, v_1) - h(\theta_2, u_2, v_2)\|_2} \cdot \frac{1}{\lambda} \cdot \kappa H \kappa_1 + \sqrt{K H \kappa_1 \cdot \|G_{uv}^{-1}(\theta_1) - G_{uv}^{-1}(\theta_2)\|_2} \cdot \kappa H \kappa_1$$

$$ + \sqrt{K \kappa_1 \cdot \|h(\theta_1, u_1, v_1) - h(\theta_2, u_2, v_2)\|_2} \leq 2 \sqrt{K H \kappa_1 \cdot \|\theta_1 - \theta_2\|_2 + \|\sigma_{uv}^2 - \sigma_{uv}^2\|_2} \cdot \frac{1}{\lambda} \cdot \kappa H \kappa_1 + \sqrt{K H \kappa_1 \cdot \|h(\theta_1, u_1, v_1) - h(\theta_2, u_2, v_2)\|_2} \cdot \kappa H \kappa_1$$

$$ \leq \sqrt{2K K \kappa_1 \cdot \|\theta_1 - \theta_2\|_2 + \|\sigma_{uv}^2 - \sigma_{uv}^2\|_2} \cdot \frac{1}{\lambda} \cdot \kappa H \kappa_1 + \sqrt{K H \kappa_1 \cdot \|h(\theta_1, u_1, v_1) - h(\theta_2, u_2, v_2)\|_2} \cdot \kappa H \kappa_1$$

$$ \leq 2 \left( \sqrt{K H \kappa_1 \cdot \|\theta_1 - \theta_2\|_2 + \|\sigma_{uv}^2 - \sigma_{uv}^2\|_2} \cdot \frac{1}{\lambda} \cdot \kappa H \kappa_1 + \sqrt{K H \kappa_1 \cdot \|h(\theta_1, u_1, v_1) - h(\theta_2, u_2, v_2)\|_2} \right) \cdot \kappa H \kappa_1$$

note

$$ |\sigma_{uv}^2 - \sigma_{uv}^2| \leq |f(v_1, \phi(s, a)) - f(v_2, \phi(s, a))| + 2H |f(u_1, \phi(s, a)) - f(u_2, \phi(s, a))| \leq \kappa_1 \|v_1 - v_2\|_2 + 2H \kappa_1 \|u_1 - u_2\|_2,$$

Then a $\varepsilon$-covering net of $\{H(\theta, u, v)\}$ can be constructed by the union of covering net for $\theta, u, v$ and by Lemma L.8, the covering number $N_{\varepsilon}$ satisfies (where $\bar{O}$ absorbs Polylog terms)

$$ \log N_{\varepsilon} \leq \bar{O}(d)$$
Step 3. First note by definition in Step 2
\[
\left\| \sum_{k=1}^{K} \frac{f(\theta_{TV_{k+1}}, \phi_{h,k}) - r_{h,k} - V_{h+1}^*(s_{h+1})}{\sigma_h^2(s_h, a_h)} \cdot \nabla_\theta f(\hat{\theta}_h, \phi_{h,k}) \right\|_{\Lambda_h^{-1}} = H(\hat{\theta}_h, \mathbf{u}_h, \mathbf{v}_h)
\]
Now choosing \( \epsilon = O(1/K) \) in Step 2 and union bound over the covering number in Step 2, we obtain with probability \( 1 - \delta \) (recall \( \sqrt{d} \geq \tilde{O}(\zeta) \)),
\[
H(\hat{\theta}_h, \mathbf{u}_h, \mathbf{v}_h) \leq 16 \sqrt{d \log \left( 1 + \frac{K \kappa_1^2}{\lambda d} \right) \cdot \log \left( \frac{4K^2}{\delta} \right)} + \tilde{O}(d) + 4\zeta \log \left( \frac{4K^2}{\delta} \right) + \tilde{O}(d) + O\left( \frac{1}{K} \right)
\]
where we absorb all the Polylog terms. Combing above with equation 33, we obtain with probability \( 1 - \delta \),
\[
|I_1| \leq \left\| \nabla f(\hat{\theta}_h, \phi(s, a)) \right\|_{\Lambda_h^{-1}} \cdot H(\hat{\theta}_h, \mathbf{u}_h, \mathbf{v}_h)
\leq \left\| \nabla f(\hat{\theta}_h, \phi(s, a)) \right\|_{\Lambda_h^{-1}} \cdot \left[ \tilde{O}(d) + O\left( \frac{1}{K} \right) \right]
\leq \tilde{O} \left( d \left\| \nabla f(\hat{\theta}_h, \phi(s, a)) \right\|_{\Lambda_h^{-1}} \right) + \tilde{O}\left( \frac{\kappa_1}{\sqrt{dK}} \right),
\]
\( \square \)

Combing dominate term \( I_1 \) (via Lemma J.6) and all other higher order terms we can obtain the first result together with Lemma D.3.

The proof of the second result is also very similar to the proofs in Section F.2. Concretely, when picking \( \pi = \pi^* \), we can convert the quantity
\[
\sqrt{\nabla_\theta f(\hat{\theta}_h, \phi(s_h, a_h))\Lambda_h^{-1}\nabla_\theta f(\hat{\theta}_h, \phi(s_h, a_h))}
\]
to
\[
\sqrt{\nabla_\theta f(\hat{\theta}_h^*, \phi(s_h, a_h))\Lambda_h^{-1}\nabla_\theta f(\hat{\theta}_h^*, \phi(s_h, a_h))}
\]
using Theorem J.3, and convert
\[
\sqrt{\nabla_\theta f(\hat{\theta}_h^*, \phi(s_h, a_h))\Lambda_h^{-1}\nabla_\theta f(\hat{\theta}_h^*, \phi(s_h, a_h))}
\]
to
\[
\sqrt{\nabla_\theta f(\theta_h^*, \phi(s_h, a_h))\Lambda_h^{-1}\nabla_\theta f(\theta_h^*, \phi(s_h, a_h))}
\]
using Lemma J.4.

K THE LOWER BOUND

Theorem K.1 (Restatement of Theorem 4.2). Specifying the model to have linear representation \( f = (\theta, \phi) \). 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 \( \tilde{\pi} \)
\[
\inf_{\tilde{\pi}} \sup_{M \in \mathcal{M}} \mathbb{E}_M [v^* - v_{\tilde{\pi}}] \geq c \sqrt{d} \cdot \sum_{h=1}^{H} \mathbb{E}_{\pi_h} \left[ \sqrt{\nabla_\theta^T f(\theta_h^*, \phi(\cdot, \cdot))} \Lambda_h^{-1} \nabla_\theta f(\theta_h^*, \phi(\cdot, \cdot)) \right],
\]
(34)
where \( \Lambda_h^{-1} = \mathbb{E} \left[ \sum_{k=1}^{K} \nabla x(\theta_h^*, \phi(s_h^*, a_h^*)) \nabla x(\theta_h^*, \phi(s_h^*, a_h^*))^T \right] \).

Remark K.2. Note Theorem 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 \( F \) satisfies Assumption 2.1, 2.3. Therefore, for the lower bound construction it suffices to find one model \( f \) such that the lower bound equation 34 holds. Here we simply choose the linear function approximation.
K.1 Regarding the proof of lower bound

The proof of Theorem 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 Yin et al. (2022), and the original proof comes from Theorem 2 of Zanette et al. (2021).

Concretely, all the proofs in Theorem 3.5 of Yin et al. (2022) follows and the only modification is to replace

\[
\sqrt{\mathbb{E}_n^*[\phi] (A_h^*)^{-1}} \mathbb{E}_n^*[\phi] \leq \frac{1}{2} \| \phi (1, u_h) \| (A_h^*)^{-1} + \frac{1}{2} \| \phi (-1, u_h) \| (A_h^*)^{-1}
\]

in Section E.5 by

\[
\mathbb{E}_n^* \left[ \phi (\cdot, \cdot) (A_h^*)^{-1} \phi (\cdot, \cdot) \right] = \frac{1}{2} \| \phi (+1, u_h) \| (A_h^*)^{-1} + \frac{1}{2} \| \phi (-1, u_h) \| (A_h^*)^{-1},
\]

and the final result holds with \( \phi (\cdot, \cdot) = \nabla_\theta f (\theta^*_h, \phi (\cdot, \cdot)) \) by the reduction \( f = (\theta, \phi) \).

L Auxiliary lemmas

Lemma L.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 \( \theta, \theta' \in \Theta \), there exists \( \xi \in \text{the line segment of } \theta \) and \( \theta' \), such that

\[
f(\theta) - f(\theta') = \nabla f(\theta) (x - \theta) + \frac{1}{2} (x - \theta) \nabla^2 f(\theta)(x - \theta) + \ldots + \frac{1}{(k-1)!} \nabla^{k-1} f(\theta) \left( \bigotimes (x - \theta) \right)^{k-1}
\]

Here \( \nabla^k f(\theta) \) denotes \( k \)-dimensional tensor and \( \bigotimes \) denotes tensor product.

Lemma L.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 \( 1 - \delta \),

\[
\left\| \frac{1}{n} \sum_{i=1}^{n} X^{(i)} \right\|_2 \leq \sqrt{\frac{4dR^2}{n} \log \left( \frac{d}{\delta} \right)}.
\]

Proof of Lemma L.2. Since \( \|X\|_2 \leq R \) implies \( |X_j| \leq R \), by the univariate Hoeffding’s inequality, for a fixed \( j \in \{1, \ldots, d\} \), denote \( Y_j := \frac{1}{n} \sum_{i=1}^{n} X^{(i)}_j \). Then with probability \( 1 - \delta \) (note \( |X^{(i)}_j| \leq R \),

\[
\Pr \left( |Y_j| \geq 2 \sqrt{\frac{R^2}{n} \log \left( \frac{1}{\delta} \right)} \right) \leq \delta.
\]

By a union bound,

\[
\exists i \text{ s.t. } |Y_j| \geq 2 \sqrt{\frac{R^2}{n} \log \left( \frac{1}{\delta} \right)} \iff \forall i \nabla |Y_j| \leq 2 \sqrt{\frac{R^2}{n} \log \left( \frac{1}{\delta} \right)} \geq 1 - \delta\)
\]

\[
\iff \forall i \ n \leq \frac{4R^2}{n} \log \left( \frac{1}{\delta} \right) \geq 1 - \delta \iff \forall Y \nabla |Y| \geq \frac{4R^2}{n} \log \left( \frac{1}{\delta} \right) \geq 1 - \delta \)
\]

\[
\iff \forall |Y| \leq \frac{4dR^2}{n} \log \left( \frac{1}{\delta} \right) \geq 1 - \delta.
\]

Lemma L.3 (Hoeffding inequality for self-normalized martingales (Abbasi-Yadkori et al., 2011)). Let \( \{\eta_t\}_{t=1}^{\infty} \) be a real-valued stochastic process. Let \( \{\mathcal{F}_t\}_{t=0}^{\infty} \) be a filtration, such that \( \eta_t \) is \( \mathcal{F}_t \)-measurable. Assume \( \eta_t \) also satisfies \( \eta_t \) 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 \( t > 0 \),

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

**Lemma L.4** (Bernstein inequality for self-normalized martingales (Zhou et al., 2021a)). Let \( \{\eta_t\}_{t=1}^{\infty} \) be a real-valued stochastic process. Let \( \{\mathcal{F}_t\}_{t=0}^{\infty} \) be a filtration, such that \( \eta_t \) is \( \mathcal{F}_t \)-measurable. Assume \( \eta_t \) also satisfies

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

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 \( t > 0 \),

\[
\left\| \sum_{s=1}^{t} x_s \eta_s \right\|^2_{\Lambda_t^{-1}} \leq 8 \sigma^2 \sqrt{d \log \left( 1 + \frac{tL^2}{\lambda \delta} \right) \log \left( \frac{4t^2}{\delta} \right)} + 4R \log \left( \frac{4t^2}{\delta} \right).
\]

**Lemma L.5.** Let \( \nabla f(\theta, \phi(\cdot, \cdot)) : \mathcal{S} \times A \rightarrow \mathbb{R}^d \) be a bounded function s.t. \( \sup_{\theta \in \Theta} \|\nabla f(\theta, \phi(\cdot, \cdot))\|_2 \leq \kappa_1 \). If \( K \) satisfies

\[
K \geq \max \left\{ 512 \frac{\kappa_1^4}{\kappa_2^2} \left( \log \left( \frac{2d}{\delta} \right) + d \log \left( 1 + \frac{4\kappa_1 B^2 \kappa_2 C \kappa_3}{\lambda^2} \right) \right), 4\lambda \right\}
\]

Then with probability at least \( 1 - \delta \), for all \( \|u\|_2 \leq B \) simultaneously, it holds that

\[
\|u\|_{\Sigma_h^{-1}} \leq \frac{2B}{\sqrt{K}} + O\left( \frac{1}{K} \right)
\]

where \( \Sigma_h = \sum_{k=1}^{K} \nabla f(\tilde{\theta}_h, \phi(s_h^k, \eta_h^k)) \cdot \nabla f(\tilde{\theta}_h, \phi(s_h^k, \eta_h^k))^\top + \lambda I_d \).

**Proof of Lemma L.5.** For a fixed \( \theta \), define \( \tilde{G} = \sum_{k=1}^{K} \nabla f(\theta, \phi(s_h^k, \eta_h^k)) \cdot \nabla f(\theta, \phi(s_h^k, \eta_h^k))^\top + \lambda I_d \), and \( G = \mathbb{E}_{\mu}[\nabla f(\theta, \phi(s_h, \eta_h)) \cdot \nabla f(\theta, \phi(s_h, \eta_h))^\top] \), then by Lemma H.5. of Min et al. (2021), as long as

\[
K \geq \max\left\{ 512 \kappa_1^4 \|G^{-1}\|^2 \log \left( \frac{2d}{\delta} \right), 4\lambda \|G^{-1}\|_2 \right\},
\]

then with probability \( 1 - \delta \), for all \( u \in \mathbb{R}^d \) simultaneously, \( \|u\|_{G^{-1}} \leq \frac{\sqrt{2}}{\sqrt{K}} \|u\|_{G^{-1}} \). As a corollary, if we constrain \( u \) 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,

\[
\|u\|_{\tilde{G}^{-1}} \leq \frac{2}{\sqrt{K}} \|u\|_{G^{-1}} = \frac{2}{\sqrt{K}} \sqrt{u^\top G^{-1}u} \leq \frac{2B \sqrt{\|G^{-1}\|_2}}{\sqrt{K}}.
\]

Next, for any \( \theta \), define

\[
h_u(\theta) := \|u\|_{\tilde{G}^{-1}} = \sqrt{u^\top \tilde{G}^{-1}u} = \sqrt{u^\top \left( \sum_{k=1}^{K} \nabla f(\theta, \phi(s_h^k, \eta_h^k)) \cdot \nabla f(\theta, \phi(s_h^k, \eta_h^k))^\top + \lambda I_d \right) u^{-1}}
\]

and \( \tilde{G}(\theta) = \sum_{k=1}^{K} \nabla f(\theta, \phi(s_h^k, \eta_h^k)) \cdot \nabla f(\theta, \phi(s_h^k, \eta_h^k))^\top + \lambda I_d \), we have for any \( \theta_1, \theta_2 \)

\[
\|\tilde{G}(\theta_1) - \tilde{G}(\theta_2)\|_2 \leq \sum_{k=1}^{K} \left| \nabla f(\theta_1, \phi(s_h^k, \eta_h^k)) - \nabla f(\theta_2, \phi(s_h^k, \eta_h^k)) \right| \cdot \left| \nabla f(\theta_1, \phi(s_h^k, \eta_h^k))^\top - \nabla f(\theta_2, \phi(s_h^k, \eta_h^k))^\top \right|
\]

\[
+ \sum_{k=1}^{K} \nabla f(\theta_2, \phi(s_h^k, \eta_h^k)) \left( \nabla f(\theta_1, \phi(s_h^k, \eta_h^k)) - \nabla f(\theta_2, \phi(s_h^k, \eta_h^k)) \right)^\top
\]

\[
\leq K \kappa_2 \kappa_1 \|\theta_1 - \theta_2\|_2 + K \kappa_2 \kappa_1 \|\theta_1 - \theta_2\|_2 \leq 2K \kappa_2 \kappa_1 \|\theta_1 - \theta_2\|_2.
\]
Use the basic inequality for $a, b > 0 \Rightarrow |\sqrt{a} - \sqrt{b}| \leq \sqrt{|a - b|}$,
\[
\sup_u \left| h_u(\theta_1) - h_u(\theta_2) \right| \leq \sup_u \sqrt{u^T (\bar{G}(\theta_1)^{-1} - \bar{G}(\theta_2)^{-1}) u} \leq \sqrt{B^2 \cdot \| \bar{G}(\theta_1)^{-1} - \bar{G}(\theta_2)^{-1} \|_2}
\]
\[
\leq \sqrt{B^2 \cdot \| \bar{G}(\theta_1) \|_2 \| \bar{G}(\theta_2) \|_2} \frac{1}{\lambda} = \sqrt{\frac{2B^2 K_\kappa \kappa_2 \| \theta_1 - \theta_2 \|_2}{\lambda^2}}
\]

Therefore, the $\epsilon$-covering net of $\{ h(\theta) : \theta \in \Theta \}$ is implies by the $\frac{\epsilon^2}{2K_\kappa \kappa^2}$-covering net of $\{ \theta : \theta \in \Theta \}$, so by Lemma L.8, the covering number $N_\epsilon$ satisfies
\[
\log N_\epsilon \leq d \log \left( 1 + \frac{4B^2 K_\kappa \kappa_2 C_\Theta}{\lambda^2} \right).
\]

Select $\theta = \bar{\theta}_h$. Choose $\epsilon = O(1/K)$ and by a union bound over equation 36 to get with probability $1 - \delta$, for all $\|u\|_2 \leq B$ (note By Assumption 2.3 $\|G^{-1}\|_2 \leq 1/\kappa$),
\[
\|u\|_{\Sigma_h^{-1}} \leq \frac{2B}{\sqrt{\kappa K}} + O\left( \frac{1}{K} \right)
\]

if (union bound over the condition equation 35)
\[
K \geq \max \left\{ 512 \frac{\kappa^4}{\kappa^2} \left( \log \left( \frac{2d}{\delta} \right) + d \log \left( 1 + \frac{4\kappa_1 B^2 \kappa_2 C_\Theta K^3}{\lambda^2} \right) \lambda \right) \right\}
\]

where this condition is satisfied by the Lemma statement.

\[\square\]

**Lemma L.6.** Let $\phi : S \times A \rightarrow \mathbb{R}^d$ satisfies $\|\phi(s, a)\| \leq C$ for all $s, a \in S \times A$. 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 $\nu$. Then with probability $1 - \delta$,
\[
\| \bar{G}_K - \mathbb{E}_\nu \left[ \bar{G}_K \right] \| \leq 4\sqrt{2C^2} \sqrt{\frac{\log \left( \frac{2d}{\delta} \right)}{\kappa}}^{1/2}.
\]

**Proof of Lemma L.6.** See Lemma H.5 of Yin et al. (2022) or Lemma H.4 of Lemma Min et al. (2021) for details.

**Lemma L.7 (Lemma H.4 in Yin et al. (2022)).** Let $\Lambda_1$ and $\Lambda_2 \in \mathbb{R}^{d \times d}$ are two positive semi-definite matrices. Then:
\[
\| \Lambda_1^{-1} \| \leq \| \Lambda_2^{-1} \| + \| \Lambda_1^{-1} \| \cdot \| \Lambda_2^{-1} \| \cdot \| \Lambda_1 - \Lambda_2 \|
\]

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

for all $\phi \in \mathbb{R}^d$.

L.1 COVERING ARGUMENTS

**Lemma L.8.** (Covering Number of Euclidean Ball) For any $\epsilon > 0$, the $\epsilon$-covering number of the Euclidean ball in $\mathbb{R}^d$ with radius $R > 0$ is upper bounded by $(1 + 2R/\epsilon)^d$.

**Lemma L.9.** Define $\mathcal{V}$ to be the class mapping $S$ to $\mathbb{R}$ with the parametric form
\[
V(\cdot) := \min_a \{ \max f(\theta, \phi(\cdot, a)) - \nabla f(\theta, \phi(\cdot, a))^\top A \cdot \nabla f(\theta, \phi(\cdot, a)) : H \}
\]

where the parameter spaces are $\{ \theta : \| \theta \|_2 \leq C_\Theta \}$ and $\{ A : \| A \|_2 \leq B \}$. Let $\mathcal{N}_\epsilon^\mathcal{V}$ be the covering number of $\epsilon$-net with respect to $l_\infty$ distance, then we have
\[
\log \mathcal{N}_\epsilon^\mathcal{V} \leq d \log \left( 1 + \frac{8C_\Theta (K_1 \sqrt{C_\Theta} + 2\sqrt{B^2 K_1 K_2})^2}{\epsilon^2} \right) + d^2 \log \left( 1 + \frac{8\sqrt{d}B \kappa^2}{\epsilon^2} \right).
\]

48
Proof of Lemma L.9.

\[ \sup_{s,a} |V_1(s) - V_2(s)| \]
\[ \leq \sup_{s,a} \left| f(\theta_1, \phi(s, a)) - \sqrt{\nabla f(\theta_1, \phi(s, a))}^\top A_1 \cdot \nabla f(\theta_1, \phi(s, a)) - f(\theta_2, \phi(s, a)) + \sqrt{\nabla f(\theta_2, \phi(s, a))}^\top A_2 \cdot \nabla f(\theta_2, \phi(s, a)) \right| \]
\[ = \sup_{s,a} \left| \nabla f(\xi, \phi(s, a)) \cdot (\theta_1 - \theta_2) - \sqrt{\nabla f(\theta_1, \phi(s, a))}^\top A_1 \cdot \nabla f(\theta_1, \phi(s, a)) + \sqrt{\nabla f(\theta_2, \phi(s, a))}^\top A_2 \cdot \nabla f(\theta_2, \phi(s, a)) \right| \]
\[ \leq \kappa_1 \cdot \|\theta_1 - \theta_2\|_2 + \sup_{s,a} \left| \nabla f(\theta_1, \phi(s, a)) \right| \cdot \|\theta_1 - \theta_2\|_2 \cdot \|A_1 - A_2\|_2 \]
\[ \leq \kappa_1 \cdot \|\theta_1 - \theta_2\|_2 + \sup_{s,a} \left| \nabla f(\theta_1, \phi(s, a)) \right| \cdot \|\theta_1 - \theta_2\|_2 \cdot \|A_1 - A_2\|_2 \]
\[ \leq \kappa_1 \cdot \|\theta_1 - \theta_2\|_2 + \frac{\kappa_1}{2} \cdot \|A_1 - A_2\|_2 \]
\[ \leq \kappa_1 \cdot \|\theta_1 - \theta_2\|_2 + \frac{\kappa_1}{2} \cdot \|A_1 - A_2\|_2 \]
\[ \leq \kappa_1 \cdot \|\theta_1 - \theta_2\|_2 + \frac{\kappa_1}{2} \cdot \|A_1 - A_2\|_2 \]
\[ \leq \frac{\kappa_1}{2} \cdot \|A_1 - A_2\|_2 \]
\[ \leq \frac{\kappa_1}{2} \cdot \|A_1 - A_2\|_2 \]

Here \( \|\cdot\|_F \) is Frobenius norm. Let \( C_\theta \) be the \( \frac{\epsilon^2}{4(\kappa_1 \sqrt{C_{\theta \theta}} + 2\sqrt{B\kappa_1\kappa_2})} \)-net of space \( \{\theta : \|\theta\|_2 \leq C_\theta\} \) and \( C_w \) be the \( \frac{\epsilon^2}{4\kappa_1} \)-net of the space \( \{A : \|A\|_F \leq \sqrt{dB}\} \), then by Lemma L.8,

\[ |C_{w}\| \leq \left( 1 + \frac{8C_{\theta}(\kappa_1 \sqrt{C_{\theta \theta}} + 2\sqrt{B\kappa_1\kappa_2})^2}{\epsilon^2} \right)^d, \quad |C_{A}\| \leq \left( 1 + \frac{8\sqrt{dB}\kappa_1^2}{\epsilon^2} \right)^d \]

Therefore, the covering number of space \( V \) satisfies

\[ \log \mathcal{N}_\epsilon^V \leq \log(|C_{w}| \cdot |C_{A}|) \leq d\log \left( 1 + \frac{8C_{\theta}(\kappa_1 \sqrt{C_{\theta \theta}} + 2\sqrt{B\kappa_1\kappa_2})^2}{\epsilon^2} \right) + d^2 \log \left( 1 + \frac{8\sqrt{dB}\kappa_1^2}{\epsilon^2} \right) \]

Lemma L.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 \( f_{V_{\theta'}} := f_{\frac{\delta V}{V}} + V_{\theta'} \) and \( V(s) \) has form \( V_{\theta}(s) \) that belongs to \( V \) (as defined in Lemma L.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 \( \mathcal{H} = \{h(\theta, \theta') := \mathbb{E}_{\mu}[X(\theta, \theta')] : \|\theta\|_2 \leq C_\theta, V_\theta \in V \} \) has the covering number of \( (\epsilon + 4H_c)^d \)-net bounded by

\[ d\log(1 + \frac{24C_{\theta}(H_c + 1)\kappa_1}{\epsilon}) + d\log \left( 1 + \frac{88H^2 C_{\theta}(\kappa_1 \sqrt{C_{\theta \theta}} + 2\sqrt{B\kappa_1\kappa_2})^2}{\epsilon^2} \right) + d^2 \log \left( 1 + \frac{88H^2\sqrt{dB}\kappa_1^2}{\epsilon^2} \right). \]

Proof of Lemma L.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 \((\theta_1, \theta_1'), (\theta_2, \theta_2')\),
\[
|X(\theta_1, \theta_1') - X(\theta_2, \theta_2')| \leq |f(\theta_1, \phi(s, a))^2 - f(\theta_2, \phi(s, a))^2|
\]
\[
+ |f_{\nu_1}(s, a)^2 - f_{\nu_2}(s, a)^2| + 2|f_{\nu_1'}(s, a) - f_{\nu_2'}(s, a)| \cdot (r + V_{\theta_1}'(s'))
\]
\[
+ 2f_{\nu_2}(s, a) \cdot |V_{\theta_1}'(s') - V_{\theta_2}'(s')| + 2|f(\theta_1, \phi(s, a)) - f(\theta_2, \phi(s, a))| \cdot (r + V_{\theta_1}'(s'))
\]
\[
+ 2|f(\theta_2, \phi(s, a))| \cdot |V_{\theta_1}'(s') - V_{\theta_2}'(s')|
\]
\[
\leq 2H \cdot |f(\theta_1, \phi(s, a)) - f(\theta_2, \phi(s, a))| + 2H \cdot |f_{\nu_1'}(s, a) - f_{\nu_2'}(s, a)|
\]
\[
+ 4H \cdot |V_{\theta_2}(s') - V_{\theta_2}'(s')| + 4(H + 1) \cdot |f(\theta_1, \phi(s, a)) - f(\theta_2, \phi(s, a))|
\]
\[
\leq (6H + 1) \cdot |f(\theta_1, \phi(s, a)) - f(\theta_2, \phi(s, a))| + 2H \max_{s'} |V_{\theta_1}'(s') - V_{\theta_2}'(s')| + 4H \epsilon \mathcal{F}
\]
\[
+ 4H \cdot |V_{\theta_2}(s') - V_{\theta_2}'(s')|
\]
\[
\leq (6H + 1) \|\nabla f(\xi, \phi(s, a))\|_2 \cdot \|\theta_1 - \theta_2\|_2 + 6H \|V_{\theta_1}' - V_{\theta_2}'\|_{\infty} + 4H \epsilon \mathcal{F}
\]
\[
\leq (6H + 1) \kappa_1 \cdot \|\theta_1 - \theta_2\|_2 + 6H \|V_{\theta_1}' - V_{\theta_2}'\|_{\infty} + 4H \epsilon \mathcal{F}
\]

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
\[
|\mathbb{E}_\mu[X(\theta_1, \theta_1')] - \mathbb{E}_\mu[X(\theta_2, \theta_2')]| \leq \sup_{s, a, s'} |X(\theta_1, \theta_1') - X(\theta_2, \theta_2')|
\]
\[
\leq (6H + 1) \kappa_1 \cdot \|\theta_1 - \theta_2\|_2 + 6H \|V_{\theta_1}' - V_{\theta_2}'\|_{\infty} + 4H \epsilon \mathcal{F}
\]

Now let \(C_1\) be the \(\frac{24C_0(H + 1)\kappa_1}{\epsilon}\)-net of \(\{\theta : \|\theta\|_2 \leq C_0\}\) and \(C_2\) be the \(\epsilon/6H\)-net of \(\mathcal{V}\), applying Lemma L.8 and Lemma L.9 to obtain
\[
\log |C_1| \leq d \log(1 + \frac{24C_0(H + 1)\kappa_1}{\epsilon}) \quad \log |C_2| \leq d \log \left(1 + \frac{288H^2C_0(\kappa_1 \sqrt{C_0} + 2\sqrt{B\kappa_1}k_2)^2}{\epsilon^2}\right) + d^2 \log \left(1 + \frac{288H^2\sqrt{B\kappa_1}k_2^2}{\epsilon^2}\right)
\]
which implies the covering number of \(\mathcal{H}\) to be bounded by
\[
\log |C_1| \cdot |C_2| \leq d \log(1 + \frac{24C_0(H + 1)\kappa_1}{\epsilon}) + d \log \left(1 + \frac{288H^2C_0(\kappa_1 \sqrt{C_0} + 2\sqrt{B\kappa_1}k_2)^2}{\epsilon^2}\right) + d^2 \log \left(1 + \frac{288H^2\sqrt{B\kappa_1}k_2^2}{\epsilon^2}\right)
\]
\(\blacksquare\)

**Lemma L.11.** Denote \(\sigma^2_{u,v}(\cdot, \cdot) := \max\{1, f(v, \phi(\cdot, \cdot))|_{[0,(H-h+1)^2]} - [f(u, \phi(\cdot, \cdot))|_{[0,(H-h+1)^2]}\}^2)\)
and define
\[
X(\theta, \theta', u, v) := \frac{(f(\theta, \phi(s, a)) - r - V_{\theta'}(s'))^2 - (f_{\nu'}(s, a) - r - V_{\theta'}(s'))^2}{\sigma^2_{u,v}(s, a)}
\]
where \(f_V := \mathcal{P}_h V\) and \(V(s)\) has form \(V_\theta(s)\) that belongs to \(\mathcal{V}\) (as defined in Lemma L.9). Here \(X(\theta, \theta', u, v)\) is a function of \(s, a, r, s'\) as well, and we suppress the notation for conciseness only. Then the function class \(\mathcal{H} = \{h(\theta, \theta', u, v) := \mathbb{E}_\mu[X(\theta, \theta', u, v)] ||\|\theta\|_2 \leq C_0, \theta, \theta' \in \mathcal{V}\}\) has the covering number of \(\epsilon\)-net bounded by
\[
d \log(1 + \frac{24C_0(H + 1)\kappa_1}{\epsilon}) + d \log \left(1 + \frac{288H^2C_0(\kappa_1 \sqrt{C_0} + 2\sqrt{B\kappa_1}k_2)^2}{\epsilon^2}\right) + d^2 \log \left(1 + \frac{288H^2\sqrt{B\kappa_1}k_2^2}{\epsilon^2}\right)
\]
\[
+ d \log(1 + \frac{16C_0H^2\kappa_1}{\epsilon}) + d \log(1 + \frac{32C_0H^3\kappa_1}{\epsilon})
\]

**Proof of Lemma L.11.** Recall \(\sigma^2_{u,v}(\cdot, \cdot) := \max\{1, f(v, \phi(\cdot, \cdot))|_{[0,(H-h+1)^2]} - [f(u, \phi(\cdot, \cdot))|_{[0,(H-h+1)^2]}\}^2\},\) and since \(\max\) truncation are non-expansive operations, then we can achieve for any \(s, a\)
\[
|\sigma^2_{u_1,v_1}(s, a) - \sigma^2_{u_2,v_2}(s, a)| \leq |f(v_1, \phi(s, a)) - f(v_2, \phi(s, a))| + 2H |f(u_1, \phi(s, a)) - f(u_2, \phi(s, a))| + 2H \kappa_1 \|u_1 - u_2\|_2
\]
Hence
\[
\left| \bar{X}(\theta_1, \theta_1', u_1, v_1) - \bar{X}(\theta_2, \theta_2', u_2, v_2) \right| = \left| \frac{X(\theta_1, \theta_1') - X(\theta_2, \theta_2')}{\sigma_{u_1,v_1}^2} \right|
\]
\[
\leq \left| \frac{X(\theta_1, \theta_1') - X(\theta_2, \theta_2')}{\sigma_{u_1,v_1}^2} \right| + \left| \frac{X(\theta_2, \theta_2')}{\sigma_{u_2,v_2}^2} \left( \sigma_{u_1,v_1}^2 - \sigma_{u_2,v_2}^2 \right) \right|
\]
\[
\leq |X(\theta_1, \theta_1') - X(\theta_2, \theta_2')| + 2H^2 |\sigma_{u_1,v_1}^2 - \sigma_{u_2,v_2}^2|
\]
\[
\leq |X(\theta_1, \theta_1') - X(\theta_2, \theta_2')| + 2H^2 \kappa_1 \| v_1 - v_2 \|_2 + 4H^3 \kappa_1 \| u_1 - u_2 \|_2
\]
\[
\leq (6H + 1) \kappa_1 \| \theta_1 - \theta_2 \|_2 + 6H \| V_{\theta_1} - V_{\theta_2} \|_{\infty} + 2H^2 \kappa_1 \| v_1 - v_2 \|_2 + 4H^3 \kappa_1 \| u_1 - u_2 \|_2
\]

Note the above holds true for all \( s, a, r, s' \), therefore it implies
\[
|E_{\mu}[\bar{X}(\theta_1, \theta_1', u_1, v_1)] - E_{\mu}[\bar{X}(\theta_2, \theta_2', u_2, v_2)]|
\leq (6H + 1) \kappa_1 \| \theta_1 - \theta_2 \|_2 + 6H \| V_{\theta_1} - V_{\theta_2} \|_{\infty} + 2H^2 \kappa_1 \| v_1 - v_2 \|_2 + 4H^3 \kappa_1 \| u_1 - u_2 \|_2
\]

and similar to Lemma L.10, the covering number of \( \epsilon \)-net will be bounded by
\[
d \log(1 + \frac{24C_\omega(H + 1) \kappa_1}{\epsilon}) + d \log \left( 1 + \frac{288H^2 C_\omega (\kappa_1 \sqrt{C_\omega} + 2\sqrt{DB\kappa_2})}{\epsilon^2} \right) + d^2 \log \left( 1 + \frac{288H^2 \sqrt{d} B \kappa_3}{\epsilon^2} \right)
\]
\[
+ d \log(1 + \frac{16C_\omega H^2 \kappa_1}{\epsilon}) + d \log(1 + \frac{32C_\omega H^3 \kappa_1}{\epsilon})
\]

Comparing to Lemma L.10, the last two terms are incurred by covering \( u, v \) arguments.