Off-Policy Fitted Q-Evaluation with Differentiable Function Approximators: Z-Estimation and Inference Theory

Ruiqi Zhang [1] **Xuezhou Zhang** [2] **Chengzhuo Ni** [2] **Mengdi Wang** [2 3]

Abstract

Off-Policy Evaluation (OPE) serves as one of the cornerstones in Reinforcement Learning (RL). Fitted Q Evaluation (FQE) with various function approximators, especially deep neural networks, has gained practical success. While statistical analysis has proved FQE to be minimax-optimal with tabular, linear and several nonparametric function families, its practical performance with more general function approximator is less theoretically understood. We focus on FQE with general *differentiable function approximators*, making our theory applicable to neural function approximations. We approach this problem using the Z-estimation theory and establish the following results: The FQE estimation error is asymptotically normal with explicit variance determined jointly by the tangent space of the function class at the ground truth, the reward structure, and the distribution shift due to off-policy learning; The finite-sample FQE error bound is dominated by the same variance term, and it can also be bounded by function class-dependent divergence, which measures how the off-policy distribution shift intertwines with the function approximator. In addition, we study bootstrapping FQE estimators for error distribution inference and estimating confidence intervals, accompanied by a Cramer-Rao lower bound that matches our upper bounds. The Z-estimation analysis provides a generalizable theoretical framework for studying off-policy estimation in RL and provides sharp statistical theory for FQE with differentiable function approximators.

[*]Equal contribution [1]School of Mathematical Sciences, Peking University [2]Department of Electrical Computer Engineering, Princeton University [3]Deepmind. Correspondence to: Mengdi Wang <mengdiw@princeton.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

1. Introduction

Off-Policy Evaluation (OPE) studies the problem where one aims to estimate the expected cumulative rewards of a target policy in a Markov decision process, by using only off-policy data, e.g. data generated from some possibly unknown behavior policy different from the one to be evaluated. OPE plays a central role in sample-efficient reinforcement learning, both in online RL where it helps off-policy algorithms achieve superior sample efficiency through data reuse, and in offline RL such as medical applications where online experimentation becomes prohibited.

Among various approaches to OPE (more discussed in Section 2), we focus on a specific family of algorithms known as Fitted Q-Evaluation (FQE), that performs iterative regressions to estimate the Q function of the target policy (Munos & Szepesvári, 2008; Le et al., 2019). Q or value function approximation is widely used in RL practice in conjunction with neural networks, and has been a key component in many empirically successful RL algorithms, such as the Deep-Q-Network (DQN) and its variants (Mnih et al., 2013), and Actor-Critic and its variants (Mnih et al., 2016). For neural networks, their differentiability is arguably a most important feature that contributes to its generalizability, computation efficiency, and versatility. While most statistical analyses for OPE or FQE focus on specific tabular, parametric (linear, finite) or non-parametric (kernel, Hölder) function classes, we wish to establish theories that are flexible enough to capture most commonly used differential function approximators, including neural networks.

Statistical theory of OPE is nontrivial due to the distribution shift from off-policy data to the target policy's unknown state-action occupancy measures. When a function approximator is employed, the effect of distribution shift becomes more subtle.

In this paper we adopt a substantially more powerful theoretical tool, i.e. the Z-estimation theory, to analyze statistical properties of FQE with general differentiable function class. Roughly speaking, we can view the FQE estimator as the root to a system of equations (i.e. Karush–Kuhn–Tucker conditions for least-square regression), whose parameters come from an empirical process. We study the statistical properties of such root and for more information about the background of Z-estimation theory please see Section 6 and subsequently in Appendix A.1.1.

Contributions of this paper are three folds:

1. We show that the FQE estimation error is asymptotically normal and give its variance σ^2 in a closed form. The variance σ^2 is determined jointly by the tangent plane of the function class $\mathcal{F}$ at the ground truth, the reward and the distribution shift.
2. We established an $O(\sigma K^{-1/2})$ high-probability finite-sample error bound for FQE. This σ can be further relaxed to a restricted $\chi^2_{\mathcal{F}}$-divergence between the behavior policy's and the target policy's occupancy measures. For any two probability measure p_1 and p_2, we define

$$\chi^2_{\mathcal{F}}(p_1, p_2) := \sup_{g \in \mathcal{F}} \frac{[\mathbb{E}_{p_1} g(x)]^2}{\mathbb{E}_{p_2}[g^2(x)]} - 1. \quad (1)$$

   It is a function class-dependent divergence between two distributions, and it measures the *partial distribution shift with respect to $\mathcal{F}$*.
3. For statistical inference, we consider bootstrapping the FQE estimator and show that the bootstrap estimators are distributionally consistent. Lastly we provide a Cramer-Rao lower bound that matches the error upper bounds, showing that differentiable FQE is statistical-optimal.

These results generalize prior minimax-optimal OPE theory for tabular and linear MDP. They apply to a substantially broader family of common function approximators. See Section 6.4 for a detailed discussion.

2. Related Work

Off Policy Evaluation Off policy evaluation(OPE) is a fundamental problem in batch RL. One classic family of OPE methods estimates the value function in an iterative fashion, including Fitted Q Evaluation(FQE) (Munos & Szepesvári, 2008; Le et al., 2019) and its variant Lasso FQE (Hao et al., 2021a). Another family of methods uses importance sampling to address the distribution mismatch and get unbiased estimation for policy value(Precup, 2000). Vanilla Importance sampling is known to suffer from large variance that in the worst case can scale exponentially with horizon length(Yin & Wang, 2020; Jiang & Li, 2016). This drawback was improved by several variance reduction techniques, including doubly robust estimation (Jiang & Li, 2016) and marginal importance ratio estimation (Xie et al., 2019). Numerous empirical studies have shown their effectiveness for variance reduction and compare their strength and weakness(Thomas & Brunskill, 2016; Li et al., 2015). For tabular MDP with full data coverage, (Yin & Wang, 2020) showed that marginal importance sampling(MIS) estimator is asymptotic optimal with mean square error bound matching the Cramer Rao lower bound in (Jiang & Li, 2016). Other methods include Hybrid Value Estimation and so on (Jin et al., 2022).

OPE with Linear Function Approximation A variety of literature focused on linear function approximator under Bellman completeness, i.e, the Bellman operator maps to state-action value functions that are linear combination of given features (Hao et al., 2021b; Duan et al., 2020; Wang et al., 2019). (Duan et al., 2020) provided a minimax lower bound for FQE with linear function approximation, and showed that it matches the upper bound. They also revealed that the dominating term of this minimax lower bound mainly depends on a χ^2-divergence term, which measures the distribution mismatch, *in the feature space*, between the data distribution and the occupancy measure of the policy to be evaluated. (Hao et al., 2021b) further combines bootstrapping with linear FQE and show that the bootstrap estimator is asymptotically efficient. (Min et al., 2021) further provided a tighter upper bound with smaller instance-dependent constants. Recently, (Perdomo et al., 2022) provided an instant-dependent sharper characterization for intrinsic complexity of FQE and LSTD with linear function approximation.

OPE with Nonparametric Approximation Many efforts have studied the use of nonparametric or seminonparametric function approximators. (Kallus & Uehara, 2020) proposed a Double Reinforcement Learning(DRL) estimator using properly estimated Q-functions and marginal density ratios, and proved DRL estimator matches a semiparametric efficiency limit for OPE. (Uehara et al., 2020) proved another two estimators based on MIS achieve this efficiency limit. Moreover, (Shi et al., 2021; Uehara et al., 2021) proposed their estimators for Q-functions derived the L^2-norm convergence rate. (Chen & Qi, 2022) provided minimax optimal convergence rate under weaker condition via reformulating Q-function estimation as non-parametric instrumental variables estimation problem. (Ji et al., 2022) explored deep FQE with Holder function class via low-dimension manifold structure and provided a sharp finite sample upper bound.

Other Works with Function Approximation Function approximation has many applications in different methods of RL, such as Fitted Q Iteration(FQI) and Policy Optimization(Munos & Szepesvári, 2008; Farahmand et al., 2016). (Duan et al., 2021) studied on-policy evaluation, and provided non-asymptotic error bound of kernel Least Square Temporal Difference(LSTD) estimator. Apart from linear function approximation(Cai et al., 2020; Jin et al., 2019; Zhou et al., 2021), (Chen & Jiang, 2019) provided analysis for FQI with general but finite function class, while (Le et al., 2019) focused on function class with finite pseudo-

Work	Method	Para?	Function Class	Key Assumption	Result
(Xie et al., 2019)(Yin & Wang, 2020)	MIS	Yes	Tabular	Concentrability	$\lvert v_\pi - \widehat{v}_\pi\rvert \le \frac{C_{\mu,\bar{\mu}}}{\sqrt{N}} + o(\frac{1}{\sqrt{N}})$, Cramer-Rao Lower Bound and Locally Minimax
(Uehara et al., 2020)	MWL / MQL	No	General	Concentrability and Stronger Realizability	$\lvert\widehat{v}_\pi - v_\pi\rvert \le \frac{C}{\sqrt{N}} + \varepsilon_{\text{approx}}$, Semi-parametric Asymptotic Lower Bound
(Duan et al., 2020)	FQE	Yes	Linear	Completeness	Minimax Optimal, $\lvert\widehat{v}^\pi - v^\pi\rvert \le \frac{CH^2}{\sqrt{N}}\left[\sum_{h=1}^H \sqrt{1+\chi^2_{\mathcal{Q}}(\mu_h,\bar{\mu})}\right] + O(\frac{1}{N})$.
(Hao et al., 2021b)	FQE	Yes	Linear	Completeness	Asymptotic Normality, Cramer-Rao Lower Bound and Distributional Consistency
(Kallus & Uehara, 2020)	DRL	No	General	Concentratability and Proper rate of Nuisance Estimator	Semi-Efficiency, $\lvert\widehat{v}_\pi - v_\pi\rvert \le \sqrt{\frac{2\log(14/\delta)\,\mathrm{Effbd}(\mathcal{M}_2)}{K}} + O(\frac{1}{K})$.
(Ji et al., 2022)	FQE	No	Holder	Completeness	$\mathbb{E}\lvert v^\pi - \widehat{v}^\pi\rvert \le CH^2\kappa\left(K^{-\frac{\alpha}{2\alpha+d}} + \sqrt{D/K}\right)\log^{\frac{3}{2}} K$.
Our Work	FQE	Yes	General and Differentiable	Completeness	$\lvert\widehat{v}^\pi - v^\pi\rvert \le \frac{CH^2}{\sqrt{N}}\left[\sum_{h=1}^H \sqrt{1+\chi^2_{\mathcal{G}_h}(\mu,\bar{\mu})}\right] + O(\frac{1}{N})$, Asymptotic Normality, Cramer-Rao Lower Bound and Distributional Consistency

Table 1. **Comparison on Different Function Approximators**: Para: Parametric; MIS: Marginal Importance Sampling; MWL: Minimax Weight Learning; MQL: Minimax Q-Function Learning; FQE: Fitted Q-Evaluation; DRL: Double Reinforcement Learning.

dimension. Recently, FQI with more general function class, such as two-layer neural networks(Long et al., 2021) and deep ReLU networks (Fan et al., 2020; Nguyen-Tang et al., 2021) have been studied.

Comparison of OPE theories Due to space limit, we postpone more discussions to Appendix F. See Table 1 for a list of directly comparable results including this paper.

3. Preliminaries

Markov Decision Process We consider the finite-horizon time-homogeneous Markov Decision Process denoted by $\mathcal{MDP}(\mathcal{S},\mathcal{A},\mathcal{P},\pi,r,\xi,H)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state space and the action space and their joint space is denoted by $\mathcal{X}=\mathcal{S}\times\mathcal{A}$. Denote $r\in\mathbb{R}^{\mathcal{X}}$ as the reward function, where $r(s,a)\in[0,1]$ for all $s\in\mathcal{S}$ and $a\in\mathcal{A}$. The transition probability is determined by $p(\cdot\mid s,a)$, and policy $\pi(\cdot\mid s)$ gives the probability distribution on $\mathcal{A}$ conditional on current state s. Throughout our paper we use π to denote the target policy to evaluate. ξ is the initial distribution and H is horizon length. We denote $\boldsymbol{\tau}=(s_1,a_1,s_2,a_2,...,s_H,a_H,s_{H+1})$ as a random trajectory in the data. The state-action value function (Q function) is $Q_h(s,a):=\mathbb{E}^\pi\left[\sum_{h'=h}^H r(s_{h'},a_{h'})\mid s_h=s,a_h=a\right]$, for $h\in[H]$, where $\mathbb{E}^\pi$ denotes expectation over random trajectories generated by the target policy π. The target value function is $V_h(s):=\mathbb{E}^\pi\left[\sum_{h'=h}^H r(s_{h'},a_{h'})\mid s_h=s\right]$.

Off-Policy Policy Evaluation (OPE) Our goal is to estimate the cumulative reward of target policy π from an initial distribution $\xi(\cdot)$, which is a scalar value defined as $v_\pi:=\mathbb{E}^\pi\left[\sum_{h=1}^H r(s_h,a_h)\mid s_1\sim\xi(\cdot)\right]$. We focus on the off-policy learning problem, where logged experiences were pre-collected according to some (possibly unknown) behavior policy and no more interaction with the MDP environment is allowed. Specifically, suppose we have K episodes of data $\{\boldsymbol{\tau}_k\}_{k=1}^K$, which are i.i.d sampled using the behavior policy $\bar{\pi}$. We sometimes use $\mathcal{D}=\{(s_n,a_n,r_n)\}_{n\in[N]}=\left\{\left(s_h^k,a_h^k,r_h^k\right)\right\}_{h\in[H],k\in[K]}$ to denote our batch samples where $N=KH$ is the total number of state transitions.

We denote μ as the state-action occupation measure generated by policy π from the initial distribution. Also we denote $\bar{\mu}$ as the population distribution, i.e., the state-action measure generated by behaviour policy $\bar{\pi}$.

4. Fitted Q Evaluation (FQE) with Differential Function Approximation

While there exist a variety of OPE algorithms, fitted Q evaluation is most popular in practice. Popularity of such value-based methods is related to their compatibility with deep learning, where training and function fitting is most convenient when minimizing squared losses.

4.1. FQE Algorithm

Fitted Q Evaluation exploits iterative regression to approximate Q functions and eventually estimates target policy value by integrating estimator of Q function (Le et al., 2019; Fonteneau et al., 2013).

Let $\mathcal{F}$ be a class of function approximators. FQE framework can be summarized as follows: We let $\widehat{Q}_{H+1} = 0$ and for $h = H, H-1, ..., 1$, we iteratively solve

$$\widehat{Q}_h = \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{N} \sum_{n=1}^{N} \left[f(s_n, a_n) - y_n \right]^2 + \lambda \rho(f) \right\}. \tag{2}$$

where $y_n = r(s_n, a_n) + \int_{\mathcal{A}} \widehat{Q}_{h+1}(s_{n+1}, a)\pi(a \mid s_{n+1}) da$, $\lambda > 0$ and $\rho(\cdot)$ is a proper regularizer. The full algorithm is given below.

Algorithm 1 Framework for Fitted Q Evaluation

1: **Input:** Target policy π, initial distribution ξ, dataset $\mathcal{D} = \{(s_h^k, a_h^k, s_{h+1}^k, r_h^k)\}_{h\in[H],k\in[K]}$, function class $\mathcal{F}$
2: **Initialize** $\widehat{Q}_{H+1}(s,a) = 0, \forall (s,a) \in \mathcal{X}$.
3: **for** $h = H, H-1, \ldots, 1$ **do**
4: Solve (2)
5: **end for**
6: **Return** $\widehat{v}_\pi = \int_{\mathcal{S}\times\mathcal{A}} \widehat{Q}_1(s,a)\pi(a|s)\xi(s) \mathrm{d}a \mathrm{d}s$.

4.2. FQE with a Differentiable Function Class

Let there be a feature map $\phi(\cdot) : \mathcal{S} \times \mathcal{A} \to \Psi \subset \mathbb{R}^m$, where Ψ is the space of state-action feature vectors. We consider function approximators that take $\phi(s,a)$ as input and output an estimated value. For simplicity, previous theory often confined $\mathcal{F}$ to some linear function class (Hao et al., 2021b; Duan et al., 2020). Note this contains the tabular case by letting $\phi(s,a)$'s be one-hot features.

We consider *almost arbitrarily* parametrized $\mathcal{F}$ with mild smoothness condition, i.e.,

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

where Θ is the parameter space. We make no distinction between $f_\theta(\phi)$ and $f(\theta, \phi)$. Without loss of generosity, we assume $\mathbf{0} \in \Theta$ and $f(\mathbf{0}, \phi(s,a)) = 0$ for any $(s,a) \in \mathcal{S}\times\mathcal{A}$. We use $\nabla_\theta f(\theta,\phi)$ to denote its partial derivative with respect to θ, and use $\nabla_\theta^2 f(\theta,\phi)$ to denote Hessian matrix. By parametrization the recursive minimization in FQE can be turned into least square optimization in θ. We assume the regularization on function f is actually on parameter $\theta \in \Theta$, hence we make no distinction between $\rho(f)$ and $\rho(\theta)$. We denote $\widehat{Q}_h = f(\widehat{\theta}_h, \phi)$ and we use ϕ_n for short of $\phi(s_n, a_n)$, then (2) is equivalent to

$$\widehat{\theta}_h = \arg\min_{\theta\in\Theta} \left\{ \frac{1}{2N} \sum_{n=1}^{N} \left[f(\theta, \phi_n) - y_n(\widehat{\theta}_{h+1}) \right]^2 + \lambda\rho(\theta) \right\}, \tag{3}$$

where $y_n(\theta') = r(s_n, a_n) + \int_{\mathcal{A}} f(\theta', \phi(s_{n+1}, a))\pi(a \mid s_{n+1}) \mathrm{d}a$. Then, the final FQE estimator becomes $\widehat{v}_\pi = \int_{\mathcal{S}\times\mathcal{A}} f(\widehat{\theta}_1, \phi(s,a))\pi(a \mid s)\xi(s) da ds$. We denote by $\theta^* = (\theta_1^*, \ldots, \theta_H^*)$ the ground truth parameter, which are solutions to (3) when there is no regularizer and the empirical sum is replaced with expectation over the population distribution.

5. Assumptions

In this section, we summarize main assumptions for statistical theory of differentiable FQE.

Assumption 5.1 (Compactness)**.** *$\Theta \subset \mathbb{R}^d$ and $\Psi \subset \mathbb{R}^m$ are compact. We denote $\mathrm{int}(\Theta)$ as its interior. We assume $\widehat{\theta}_h \in \mathrm{int}(\Theta)$ for $h \in [H]$.*

Compactness of state-action space is a natural assumption and was also assumed in (Hao et al., 2021b; Duan et al., 2020; Hao et al., 2021a) and (Yang et al., 2020). Without loss of generality, we assume that the ground truth belongs to the interior of Θ, i.e., $\{\theta : \|\theta - \theta_h^*\|_2 \leq 1\} \subset \mathrm{int}(\Theta)$ for all h.

Assumption 5.2 (Differentiability)**.** *For any $f \in \mathcal{F}$, $f(\theta,\phi)$ is third-time continuously differentiable in $\Theta \times \Psi$ with respect to θ and ϕ. The regularizer function $\rho(\theta)$ is differentiable with bounded gradient in Θ.*

This assumption requires only sufficient smoothness of f. It does not require f to take any specific parametric form or belong to certain kernel space. This smooth function class is actually much more expressive than linear function approximators. For example, most differentiable neural networks satisfy our assumptions. Other examples include spline function class widely used in statistical learning, or Gaussian processes. While linear predictors are sometimes enough in the rare cases where high quality hand-crafted features are available, in most modern machine learning applications, especially in those where deep learning thrives, high quality features are often not available and high nonlinear functions of raw features need to be learned.

Assumption 5.3 (Policy Completeness)**.** *For any function f, we define the operator $\mathcal{P} : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$, such that for any $(s,a) \in \mathcal{X}$,*

$$(\mathcal{P}f)(s,a) = \mathbb{E}^\pi \left[r(s,a) + \int_{\mathcal{A}} f(s',a')\pi(a' \mid s')\mathrm{d}a' \middle| s, a \right]. \tag{4}$$

We assume $r \in \mathcal{F}$ and for any $f \in \mathcal{F}$, we have $\mathcal{P}f \in \mathcal{F}$.

Remark on policy completeness assumption This assumption requires sufficient expressiveness of $\mathcal{F}$. It implies the realizability for Q functions, i.e., the groundtruth θ^* fully recovers the Q functions $f(\theta_h^*, \phi) = Q_h(s, a)$. It is a rather crucial assumption for RL with function approximation, and commonly used in OPE literatures (Hao et al., 2021b; Duan et al., 2020; Le et al., 2019; Hao et al., 2021a; Fan et al., 2020). (Chen & Jiang, 2019) conjectured the realizability alone is not enough for sample-efficient offline RL. Later (Wang et al., 2020) verified that only assuming realizability without policy completeness can lead to exponential sample complexity, unless one assumes a strong concentrability condition such that the distribution shift be uniformly bounded across state-action space ($\sup_{s,a} \frac{\mu(s,a)}{\bar{\mu}(s,a)} < C$) (Munos & Szepesvári, 2008; Farahmand et al., 2016; Le et al., 2019).

The concentrability condition is rather restrictive and requires that the target and behavior policies be extremely close. However, for a simple linear Gaussian system, even if $\mu, \bar{\mu}$ are two close Gaussian distributions, a small mean difference easily leads to $\sup_{s,a} \frac{\mu(s,a)}{\bar{\mu}(s,a)} = \infty$. An exception is the work of Uehara et al. (2020), which proposed a minimax approach for OPE requiring only realizability, albeit on both the Q function and the density function μ using two function classes. However, minimax optimization is computationally harder to implement than least-square regression.

Simply put, concentrability and completeness are about the different aspect of the problem. Concentrability describes the level of distribution shift, while completeness describes the expressivity of function class. Neither implies the other. Our goal is to handle the case with relatively large distribution shift. When huge distribution shift breaks the concentrability, policy completeness is the least requirement that had been found to guarantee sample efficiency. Whether there exists weaker assumption than completeness that allows for sample efficient OPE is still an open problem. Note that even if the completeness fails to hold and there exists a nonzero approximation error, our results still apply to bounding the statistical error of $\hat{\theta}$ and the approximation error can be handled by classic results on approximate value iteration (Szepesvári & Munos, 2005).

In what follows, we will explicate the dependency of error upper bound on the distribution shift and $\mathcal{F}$. We will show that FQE can be efficient without strong concentrability.

Notation Let $\theta_1, \theta_2, ..., \theta_H \in \mathbb{R}^d$ be variables and define $\theta^\top = (\theta_1^\top, \theta_2^\top, ..., \theta_H^\top) \in \mathbb{R}^{Hd}$ as the variable stacked by θ_h $h \in [H]$. Denote $\theta^{*\top} = (\theta_1^{*\top}, \theta_2^{*\top}, ..., \theta_H^{*\top}) \in \mathbb{R}^{Hd}$, and denote the FQE estimator as $\widehat{\theta}^\top = (\widehat{\theta}_1^\top, \widehat{\theta}_2^\top, ..., \widehat{\theta}_H^\top)$. For any matrix $E \in \mathbb{R}^{d_1 \times d_2}$ (including scalars and vectors as special cases), we define $\frac{\partial}{\partial \theta} E_\theta = \nabla_\theta E_\theta = (\nabla_\theta^1 E_\theta, \nabla_\theta^2 E_\theta, \ldots, \nabla_\theta^m E_\theta) \in \mathbb{R}^{d_1 \times m d_2}$. We denote $\mathbb{E}_{\boldsymbol{\tau}}$ or $\mathbb{E}$ as the expectation over the population distribution generated by behavior policy, and $\mathbb{E}^\pi$ as the expectation over target policy. For any matrix E, we denote $\|E\|$ as its operator norm, i.e. its maximal singular value. We use ϕ_h and r_h for short of $\phi(s_h, a_h)$ and $r(s_h, a_h)$.

6. Z-Estimation Theory for Differentiable FQE Estimators

In this section, we study statistical properties of the differentiable FQE estimator. The challenge with such general function approximation is the lack of analytical expressions for ground true and estimated parameters. We adopt the Z-Estimator theory as a central tool and all the proof will be deferred to Appendix.

6.1. FQE as a Z Estimator

First we show that the FQE can be written as a Z-Estimator, which means the estimator takes the form of the root of some systems. From the optimality condition of (3) for interior solutions, we know

$$\nabla_\theta \left\{ \frac{1}{2N} \sum_{n=1}^{N} \left[f(\widehat{\theta}_h, \phi_n) - y_n(\widehat{\theta}_{h+1}) \right]^2 + \lambda \rho(\widehat{\theta}) \right\} = 0.$$

For any sample path $\boldsymbol{\tau} = (s_1, a_1, ..., s_H, a_H)$ and any $\theta = (\theta_1^\top, \theta_2^\top, ..., \theta_H^\top)^\top \in \Theta^H$ and $\theta_{H+1} = \mathbf{0}$, we define the $z_{(h)}$ function as

$$z_{(h)}(\theta, \boldsymbol{\tau}) = \sum_{j=1}^{H} \Big(f(\theta_h, \phi_j) - r_j - y_j(\theta_{h+1}) \Big) \cdot \nabla_\theta^\top f(\theta_h, \phi_j),$$

where $y_j(\theta) := \int_{\mathcal{A}} f(\theta, \phi(s_{j+1}, a')) \pi(a' \mid s_{j+1}) \mathrm{d}a'$. Denote

$$z(\theta, \boldsymbol{\tau})^\top = (z_{(1)}^\top(\theta, \boldsymbol{\tau}), z_{(2)}^\top(\theta, \boldsymbol{\tau}), ..., z_{(H)}^\top(\theta, \boldsymbol{\tau})) \in \mathbb{R}^{Hd}$$

and its expectation $Z(\theta) := \mathbb{E}_{\boldsymbol{\tau}} \{ z(\theta, \boldsymbol{\tau}) \}$. The groundtruth θ^* is the root of the expected Z function, i.e.,

$$Z(\theta^*) = 0. \tag{5}$$

Define the empirical Z function as

$$Z_K(\theta) := \frac{1}{K} \sum_{k=1}^{K} z(\theta, \boldsymbol{\tau}_k)$$

and denote $R(\theta) := (\nabla_\theta \rho(\theta_1), \nabla_\theta \rho(\theta_2), \ldots, \nabla_\theta \rho(\theta_H))^\top$. Then we have

$$Z_K(\widehat{\theta}) + \lambda R(\widehat{\theta}) = \mathbf{0}. \tag{6}$$

Therefore $\widehat{\theta}$ is the root of above equation and thus is a Z-Estimator. We use $\widehat{\theta}$ interchangeably with $\widehat{\theta}_K = (\widehat{\theta}_{K,1}^\top, ..., \widehat{\theta}_{K,H}^\top)^\top$ to explicate its dependence on K.

Define the Z function class as

$$\mathcal{Z} := \bigcup_{h=1}^{H} \bigcup_{i=1}^{d} \left\{ z_{h,i}(\theta, \cdot) : \theta \in \Theta^H \subset \mathbb{R}^{Hd} \right\},$$

where $z_{h,i}$ denotes the i-th entey of $z_{(h)}(\theta, \boldsymbol{\tau})$. The complexity of $\mathcal{Z}$, which comes from the complexity of $\mathcal{F}$, determines the statistical efficiency of FQE. We analyze the estimation error via Tarlor expansions, concentration inequalities and bounding the complexity of $\mathcal{Z}$ using its bracketing integral (Kosorok, 2008).

6.2. Asymptotic Normality and Variance

Theorem 6.1 (Asymptotic Normality)**.** *Under assumption 5.1, 5.2 and 5.3, if θ^* is the unique root of $Z(\theta)$ in Θ^H and the Jacobian matrix of $Z(\theta)$ at θ^* is non-singular. Then, when $K \to \infty$ and $\lambda = o(K^{-1/2})$, we have*

$$\sqrt{K}\left(\widehat{v}_\pi - v_\pi\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

The asymptotic variance is given by

$$\sigma^2 = \frac{1}{H} \sum_{h_1, h_2 = 1}^{H} \nu_{h_1}^\top \Sigma_{h_1}^{-1} \Omega_{h_1, h_2} \Sigma_{h_2}^{-1} \nu_{h_2}. \tag{7}$$

where for $h, i, j \in [H]$,

$$\Sigma_h = \mathbb{E}\left[\frac{1}{H}\sum_{j=1}^{H} \left(\nabla_{\theta_h} f\left(\theta_h^*, \phi_j\right)\right)^\top \left(\nabla_{\theta_h} f\left(\theta_h^*, \phi_j\right)\right)\right];$$

$$\nu_h^\top = \mathbb{E}^\pi\left[\nabla_{\theta_h} f\left(\theta_h^*, \phi\left(s_h, a_h\right)\right) \mid s_1 \sim \xi(\cdot)\right];$$

$$\Omega_{i,j} = \mathbb{E}\left[\frac{1}{H}\sum_{h=1}^{H} \left(\nabla_{\theta_i}^\top f\left(\theta_i^*, \phi_h\right)\right)\left(\nabla_{\theta_j} f\left(\theta_j^*, \phi_h\right)\right)\varepsilon_{i,h}\varepsilon_{j,h}\right];$$

$$\varepsilon_{j,h} = f\left(\theta_j^*, \phi_h\right) - r_h - \mathbb{E}^\pi\left[f(\theta_{j+1}^*, \phi_{h+1}) \middle| s_{h+1}\right].$$

Remark 6.2. In linear case, this expression for variance is exactly the same as in (Hao et al., 2021b). In that case, all Σ_h become to the dataset's covariance matrix, ν_h becomes the state feature expectation $\phi(s_h, a_h)$ under target policy. They become both independent of θ_h^*. In tabular case with time-inhomogeneous MDP and one-hot feature, all Σ_h become diagonal and the asymptotic variance matches the result in Remark 3.2 in (Yin & Wang, 2020).

Proof of Theorem 6.1 Our proof technique is highly nontrivial since with non-linear function class we can no longer obtain closed-form solution as in linear case. The central tool for the proof is Z-Estimator Master Theorem (Kosorok, 2008). To use this theorem, we need to verify the function class comprising all entries of $z(\xi, \boldsymbol{\tau})$ indexed by $\theta \in \Theta^H$ is both Glivenko-Cantelli and Donsker. Glivenko-Cantelli and Donsker are properties of a function class that measure its complexity. Z-Estimator theory mainly tells us that, with a function class not too complex, asymptotic normality holds. For our differentiable $\mathcal{F}$, we will construct ε-brackets to bound the bracketing integral, which is a relatively complex approach in high-dimension statistics. A gentle introduction to these tools can be found in Appendix Section A.1.1. Asymptotic normality implies the following corollary, which implies that the convergence rate of $|\widehat{v}_\pi - v_\pi|$ is $O(\frac{1}{\sqrt{K}})$.

Corollary 6.3. *For any $\delta > 0$, there exists a constant $B(\delta) > 0$ such that*

$$\sup_{K \in \mathbb{N}} \mathbb{P}\left(\left\|\sqrt{K}\left(\widehat{\theta}_K - \theta^*\right)\right\|_2 > B(\delta)\right) \leq \delta.$$

6.3. Finite Sample Error Upper Bound

Next we will show finite-sample error bound for FQE. In addition to the $O(\frac{1}{\sqrt{K}})$ rate, we will show that the leading order term in FQE error largely depends on the variance σ^2. We also provide a reward-free error bound that depends on a function class-dependent divergence, which measures the partial mismatch between μ and $\bar{\mu}$ with respect to the $\mathcal{F}$ space. Our results strictly generalize the minmax-optimal error bounds for linear FQE (Hao et al., 2021b; Duan et al., 2020; Agarwal et al., 2019).

Assumption 6.4 (Data Coverage)**.** Let C_2 be a positive constant. We assume for any $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $h \in [H]$,

$$\nabla_\theta f(\theta_h^*, \phi(s, a))\Sigma_h^{-1}\nabla_\theta^\top f(\theta_h^*, \phi(s, a)) \leq C_2 d.$$

To interpret this assumption, consider the case where f is linear. In this case, the assumption reduces to $\phi(s, a)^\top \Sigma^{-1} \phi(s, a) \leq C_2 d$. It holds as long as the data has a non-singular covariance Σ. Even if $\sup_{s,a} \frac{\mu(s,a)}{\bar{\mu}(s,a)} = \infty$, data can still be well-conditioned and cover every dimension of $\nabla_\theta f(\theta^*, \phi)$, thus satisfying Assumption 6.4. Such coverage can be achieved with as small as $O(Hd)$ sample transitions, while full data coverage requires $\Omega(SA)$ samples. It is a much weaker assumption than assuming $\sup_{s,a} \frac{\mu(s,a)}{\bar{\mu}(s,a)} < C$ or assuming $\min_{s,a} \mu(s, a) > c$.

For $h \in [H]$, we define

$$\mathcal{G}_h := \left\{\left(\nabla_\theta f\left(\theta_h^*, \phi(\cdot)\right)\right)\mu : \mu \in \mathbb{R}^d\right\}. \tag{8}$$

We denote $\alpha = (\alpha_1, \alpha_2, ..., \alpha_d)$ as a multi-index, and the order of α is $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_n$. We denote $\partial_\theta^\alpha f(\theta, \phi) = \partial_{\theta_1}^{\alpha_1}\partial_{\theta_2}^{\alpha_2}\cdots\partial_{\theta_d}^{\alpha_d} f = \frac{\partial^{|\alpha|} f}{\partial\theta_1^{\alpha_1}\partial\theta_2^{\alpha_2}\cdots\partial\theta_d^{\alpha_d}}$, and for $l = 1, 2, 3$, we define

$$\kappa_l := \sup_{\theta \in \Theta, \phi \in \Psi} \sup_{|\alpha| = l} |\partial_\theta^\alpha f(\theta, \phi)| < \infty.$$

Our main result on finite sample upper bound is given below.

Theorem 6.5 (Finite Sample Upper Bound). *Under the assumptions of Theorem 6.1 and Assumption 6.4, let $\lambda = 0$, for every fixed $K \geq \max\left\{\frac{2(C_2 d+1)^2}{C_2 d}\log\left(\frac{12dH}{\delta}\right), B(\delta/3)^2\right\}$, and every $\delta > 0$, with probability at least $1-\delta$, we have*

(i) Variance-aware error bound:

$$|\widehat{v}_\pi - v_\pi| \leq \sqrt{\frac{2\log(6/\delta)\sigma^2}{K}} + \frac{1}{K}\left[\frac{2}{3}\ln\left(\frac{6}{\delta}\right)\sqrt{C_2 d}\sum_{h=1}^{H}(H-h+1)\sqrt{\nu_h^\top \Sigma_h^{-1}\nu_h} + C\right], \tag{9}$$

(ii) Reward-free error bound:

$$|\widehat{v}_\pi - v_\pi| \leq \left[\sum_{h=1}^{H}(H-h+1)\sqrt{1+\chi^2_{\mathcal{G}_h}(\mu,\bar{\mu})}\right] \cdot \left[\sqrt{\frac{\log(12/\delta)}{2KH}} + \frac{4\ln(12dH/\delta)}{3K}\sqrt{C_2 dH}\right] + \frac{C}{K}. \tag{10}$$

where

$$C = B(\delta/3)\left[H^{2\frac{1}{2}}d\kappa_1^2 B_0 + 2H^{\frac{3}{2}}d\kappa_2 B_0 + 4d^{\frac{3}{2}}\kappa_1\kappa_2 B_0\right] + B(\delta/3)^2\left[B_0 H^3 d\left[\kappa_3 H + 3\kappa_1\kappa_2 + 2\sqrt{d}\kappa_1\kappa_3\right]\right]$$

is a constant, where $B_0 := \max_{h\in[H]}\sqrt{\nu_h^\top\Sigma_h^{-2}\nu_h}$, D is a constant dependent on H and d only and $\chi^2_{\mathcal{G}_h}$ is restricted χ^2-divergence defined as (1).

Proof pf Theorem 6.5 We decompose $\widehat{v}_\pi - v_\pi$ into a sum of a first order term and higher order terms by using Taylor expansion. We bound the first order term with Freedman's Inequality (lemma E.2). To give a reward-free characterization for distribution shift, we used concentration inequalities to do more precise analysis. The higher order term is much more complex than that of linear case. We use upper bound for supremum norm of empirical process to bound them conditional on the event $\left\{\left\|\widehat{\theta}_K - \theta^*\right\| \leq \frac{B(\delta)}{\sqrt{K}}\right\}$ for any fixed K instead of directly bounding it via Bernstein Inequality in linear case. Exact characterization of $B(\delta)$ is not possible without more specific assumptions on $\mathcal{F}$, which is an open problem in empirical process theory and beyond our scope. Historically, we are the first to apply Z-estimation theory to RL and proved it to be useful, since they have long been thought not to be strongly related.

Remark 6.6. The error bound (i) is variance-aware and the tightest in the sense that σ^2 equals to the asymptotic variance given by Theorem 6.1. The error bound (ii) is reward-free in the sense that it does not involve the reward function r at all. It is a worst-case error bound that determined solely by the distribution shift of off-policy learning, measured by $\sqrt{1+\chi^2_{\mathcal{G}_h}(\mu,\bar{\mu})}$. In this linear function case, $1+\chi^2_{\mathcal{G}_h}(\mu,\bar{\mu})$ can be bounded by the relative condition number between the two distributions' covariance matrices. Even without function approximation, if $\mu, \bar{\mu}$ are two Gaussians with the same variance and different means, we have $\sup_{s,a}\frac{\mu(s,a)}{\bar{\mu}(s,a)} = \infty$, but their chi-square divergence is finite.

Remark 6.7. Analysis of Z estimator is in essence analysis of empirical process and Taylor expansion and is much more complex than classical Central Limit Theorem(CLT). Classical CLT applies to the i.i.d. random variables, but state transitions from a single trajectory are highly dependent, so vanilla CLT analysis does not directly apply. In addition, combining CLT and Taylor expansion can only give asymptotic rates and cannot give finite sample bound as we obtained using Z estimator.

6.4. Special Cases

Next we show that our results match the best known error bounds for FQE in the tabular case and in the case of linear function approximation.

Tabular Case In tabular case, there are finite states and finite actions in the MDP we consider, and we can always represent all Q-functions in tables. We denote

$$\tilde{\mu} := \frac{2}{H(H+1)}\mathbb{E}^\pi\left[\sum_{h=1}^{H}(H-h+1)\mathbf{1}(s_h = s, a_h = a)\right], \tag{11}$$

Then Corollary 1 in (Duan et al., 2020) provided the upper bound in tabular case as

$$|\widehat{v}_\pi - v| \leq CH^2\sqrt{1+\chi^2(\tilde{\mu},\bar{\mu})}\sqrt{\frac{\log(12/\delta)}{2KH}} + O(\frac{1}{K}),$$

where C is a constant and χ^2 is standard χ^2 divergence. We give a outline sketch in the Appendix. Further (Duan et al., 2020) proved the upper bound in time inhomogeneous MDP matches the upper bound in Theorem 3.1 in (Yin & Wang, 2020). The same result can be derived as a special case of our theorem.

Linear Case In case of linear function class, since all Σ_h are the same, we can bound the dominant error term in another sharper way. In linear case, θ_h^* and $\phi(s,a)$ will have the same dimension, and all $\mathcal{G}_h$ will be $\mathcal{G} = \left\{\mu^\top\phi(\cdot) \mid \mu \in \mathbb{R}^d\right\}$. Under the assumption $\phi^\top(s,a)\Sigma^{-1}\phi(s',a') \geq 0$, thee upper bound can be improved to

$$|\widehat{v}_\pi - v| \leq \frac{H(H+1)}{2}\sqrt{\frac{\log(12/\delta)}{2KH}}\sqrt{1+\chi^2_{\mathcal{G}}(\tilde{\mu},\bar{\mu})} + O(\frac{1}{K}).$$

This matches the minimax lower bound in (Duan et al., 2020) and shows that in linear case this upper bound is nearly optimal. Another special case with stronger assumption for data coverage is discussed in Appendix C.1.

7. Bootstrapping FQE and Distributional Consistency

FQE provides a point estimator. In order to quantify its uncertainty in practice, there is need to infer the estimator's distribution and to compute confidence intervals(Feng et al., 2020; Hao et al., 2021b).

7.1. Bootstrapping General FQE Estimator

We consider using bootstrap for statstical inference of FQE (Efron, 1982). These bootstrap estimators allow us to estimate the error distribution and perform downstream inference tasks such as confidence intervals, variance estimation (Hao et al., 2021b). We will bootstrap by episodes, instead of by transitions as in some previous research (Kostrikov & Nachum, 2020). It was shown in (Hao et al., 2021b) that bootstrapping by transitions, which are dependent, for FQE might lead to inconsistency.

Denote $W^\circ = (W_1, W_2, ..., W_K)^\top$ as the bootstrap weights sampled according to certain distribution. In bootstrapping general FQE, we use $\widehat{Q}_h^\circ$ to denote bootstrapping estimator of Q functions. We let $\widehat{Q}_{H+1}^\circ(s,a) = 0$ and (2) is turned into

$$\widehat{Q}_h^\circ = \underset{f\in\mathcal{F}}{\arg\min}\left\{\frac{1}{2N}\sum_{k=1}^{K} W_k \sum_{h=1}^{H}\left[f(s_h^k, a_h^k) - y_h^{k\circ}\right]^2 + \lambda\rho(f)\right\}. \tag{12}$$

where $y_h^{k\circ} = r\left(s_h^k, a_h^k\right) + \int_{\mathcal{A}} \widehat{Q}_{h+1}^\circ\left(s_{h+1}^k, a\right)\pi\left(a \mid s_{h+1}^k\right) da$. Similar to Section 4, under assumption of parameteric function space, we turn this problem into a minimization in parameter space and we denote $\widehat{Q}_h^\circ = f(\widehat{\theta}_h^\circ, \phi)$. Then (12) can be turned into

$$\widehat{\theta}_h^\circ = \underset{\theta\in\Theta}{\arg\min}\left\{\sum_{k=1}^{K}\frac{W_k}{2N}\sum_{h=1}^{H}\left[f\left(\theta, \phi_h^k\right) - y_h^{k\circ}\left(\widehat{\theta}_{h+1}^{k\circ}\right)\right]^2 + \lambda\rho(\theta)\right\}. \tag{13}$$

where $y_h^k\left(\widehat{\theta}_{h+1}^\circ\right) := \int_{\mathcal{A}} f\left(\widehat{\theta}_{h+1}^\circ, \phi\left(s_{h+1}^k, a'\right)\right)\pi\left(a' \mid s_{h+1}^k\right) \mathrm{d}a'$. Then the general bootstrapping FQE estimator can be written as

$$\widehat{v}_\pi^\circ = \int_{\mathcal{S}\times\mathcal{A}} f(\widehat{\theta}_1^\circ, \phi(s,a))\pi(a \mid s)\xi(s)\mathrm{d}a\mathrm{d}s. \tag{14}$$

For $\theta \in \Theta^H$, we denote $Z_K^\circ(\theta) := \frac{1}{K}\sum_{k=1}^{K} W_k z\left(\theta, \boldsymbol{\tau}_k\right)$. From the perspective of Z-Estimator, if we assume $\widehat{\theta}_h^\circ \in \operatorname{int}(\Theta)$ and write $\widehat{\theta}_{K,h}^\circ = \widehat{\theta}_h^\circ$ to explicate its dependency on K, then we have when $\widehat{\theta}_K^\circ = (\widehat{\theta}_{K,1}^{\circ\top}, ..., \widehat{\theta}_{K,H}^{\circ\top})^\top$,

$$Z_K^\circ\left(\widehat{\theta}_K^\circ\right) + \lambda R(\widehat{\theta}_K^\circ) = 0,$$

which implies $\widehat{\theta}_K^\circ$ is a bootstrapping Z-estimator. Next we consider standard bootstrap and one of its simple alternatives. Both bootstrapping FQE estimator have nice asymptotic properties and can be exploited to estimate variance and confidence interval.

Vanilla Bootstrap When W° follows multinomial distribution with probability $(1/K, 1/K, ..., 1/K)$, (14) leads to vanilla bootstrapping FQE estimator. In other words, it amounts to resampling K episodes $\boldsymbol{\tau}_1^\circ, \boldsymbol{\tau}_2^\circ, ..., \boldsymbol{\tau}_K^\circ$ independently with replacement from the data $\{\boldsymbol{\tau}_k\}_{k=1}^K$, and solving (2) iteratively using resampled data. (Hao et al., 2021b) used vanilla bootstrapping strategy to construct asymptotic confidence interval of policy value, and proved that bootstrapping error has the same asymptotic distribution with standard FQE error.

Multiplier Bootstrap Suppose U is a non-constant nonnegative random variable with finite mean $m > 0$ and variance $0 < \eta^2 < \infty$, and U satisfies

$$\int_0^\infty \sqrt{P(|U| > x)}dx < \infty. \tag{15}$$

This is slightly stronger than assuming a bounded second order moment, but is weaker than boundedness of $2+\varepsilon$ order moment for any $\varepsilon > 0$. We let $u_1, u_2, ..., u_K, ..$ are an infinite series of i.i.d. samples from the distribution of U and independent of all trajectories and $\bar{u}_K = \frac{1}{K}\sum_{k=1}^K u_k$. Setting $W_k = u_k/\bar{u}_K$ in (12) and (14) gives multiplier bootstrapping FQE estimator. For example, when U follows standard exponential distribution, W° is called Dirichlet weights and the resulting multiplier bootstrap is Bayesian bootstrap(Rubin, 1981).

Algorithm 2 Bootstrapping General FQE

Input: Target policy π, initial state distribution ξ, dataset $\mathcal{D} = \{(s_h^k, a_h^k, s_{h+1}^k, r_h^k)\}_{h\in[H],k\in[K]}$. If multiplier bootstrap used, input random vector U which satisfied (15).
if Vanilla bootstrap **then**
  Sample $(W_1, ..., W_K) \sim$ multinomial distribution with probability $(1/K, ..., 1/K)$.
else if Multiplier Bootstrap **then**
  Sample $u_k \sim U, W_k = \frac{K u_k}{\sum_{k=1}^K u_k}$.
end if
Initialize $\widehat{\theta}_{H+1}^\circ = 0$.
for $h = H, H-1, \ldots, 1$ **do**
  Solve (13)
end for
Return $\widehat{v}_\pi^\circ = \int_{\mathcal{S}\times\mathcal{A}} f\left(\widehat{\theta}_1^\circ, \phi(s,a)\right)\pi(a \mid s)\xi(s)\mathrm{d}a\mathrm{d}s$.

7.2. Distributional Consistency

We show that bootstrapping FQE with differential function approximator is distributionally consistent, i.e. the limit distribution of bootstrapping error $\widehat{\theta}_K^\circ - \widehat{\theta}_K$ conditional on batch data is that of standard FQE error $\widehat{\theta}_K - \theta^*$ established in Theorem 6.1 or a multiplication of it.

Theorem 7.1 (Distributional Consistency)**.** *Under the assumptions of Theorem 6.1, when $K \to \infty$ and $\lambda = o(K^{-1/2})$, conditional on $\{\boldsymbol{\tau}_k\}_{k=1}^K$, we have*

$$\sqrt{K}\left(\widehat{v}_\pi^\circ - \widehat{v}_\pi\right) \xrightarrow{d} \mathcal{N}(0, k_0\sigma^2).$$

for the same σ as defined in Eq. (7). For vanilla bootstrapping FQE estimator, $k_0 = 1$. For multiplier bootstrapping FQE estimator, $k_0 = \frac{\eta^2}{m^2}$.

This distribution consistency certifies that we may use the bootstrapping FQE to approximate the estimation error distributions and inference tasks. For example, we may estimate the confidence interval (CI) of the FQE by taking quantiles; see Algorithm 3. Suppose our batch data $\mathcal{D}$ is generated from a probability space $\left(\mathcal{X}^H, \Sigma\left(\mathcal{X}^H\right), \mathbb{P}_{\mathcal{D}}\right)$, and the bootstrap weight W° is from an independent probability space $(\mathcal{W}, \Omega, \mathbb{P}_W)$. Their joint probability measure is $\mathbb{P}_{\mathcal{D}W^\circ}$. Let $\mathbb{P}_{W^\circ|\mathcal{D}}$ denote the conditional distribution once the data $\mathcal{D}$ is given. Next corollary shows consistency of the CI estimate.

Corollary 7.2. *Denote the lower δ-th quantile of bootstrapping error distribution*

$$q_\delta^\pi = \inf\left\{t : \mathbb{P}_{W^\circ|\mathcal{D}}\left(\widehat{v}_\pi^\circ - \widehat{v}_\pi \le t\right) \ge \delta\right\}.$$

We construct the $1-\delta$ confidence interval of the policy value by: $\mathrm{CI}(\delta) = \left[\widehat{v}_\pi - \frac{1}{\sqrt{k_0}} q_{1-\delta/2}^\pi, \widehat{v}_\pi - \frac{1}{\sqrt{k_0}} q_{\delta/2}^\pi\right]$. *Then we have when $K \to \infty$,*

$$\mathbb{P}_{\mathcal{D}W^\circ}\left(v_\pi \in \mathrm{CI}(\delta)\right) \to 1 - \delta.$$

This gives a convenient and provable way to construct confidence interval for OPE.

Algorithm 3 Bootstrapping Confidence Interval

1: **Input:** Target policy π, initial state distribution ξ, dataset $\mathcal{D} = \{(s_h^k, a_h^k, s_{h+1}^k, r_h^k)\}_{h\in[H],k\in[K]}$, confidence level δ, number of bootstrap B. If multiplier bootstrap used, input random vector U which satisfied (15).
2: Compute standard FQE estimator $\widehat{v}(\mathcal{D})$.
3: **for** $b = 1, 2, \ldots, B$ **do**
4: Sample a bootstrapping weight W_b° according to multinomial distribution or random variable U.
5: Compute $\widehat{v}\left(\mathcal{D}; W_b^\circ\right)$.
6: **end for**
7: Compute errors $\varepsilon_b := \widehat{v}\left(\mathcal{D}; W_b^\circ\right) - \widehat{v}(\mathcal{D})$ for $b \in [B]$.
8: Compute $\delta/2$ and $1 - \delta/2$ empirical quantile of $\left\{\varepsilon_1, ..., \varepsilon_B\right\}$, denoted as $\widehat{q}_{\delta/2}^\pi$ and $\widehat{q}_{1-\delta/2}^\pi$.
9: **Return** $\left[\widehat{v}\left(\mathcal{D}\right) - \widehat{q}_{1-\delta/2}^\pi, \widehat{v}\left(\mathcal{D}\right) - \widehat{q}_{\delta/2}^\pi\right]$.

8. Information-Theoretic Lower Bound

An estimator is called asymptotic efficient if its variance is minimal among all unbiased estimators. Cramer-Rao lower bound is a theoretical lower bound for variance of all unbiased estimators. We will show that the general FQE estimator can achieve Cramer-Rao lower bound for variance and is therefore asymptotically efficient.

Theorem 8.1 (Cramer Rao Lower Bound)**.** *Under assumptions of Theorem 6.1, the variance of any unbiased estimator of v_π is lower bounded by σ^2 in (7).*

Proof of Theorem 8.1 We prove the Cramer-Rao lower bound via the standard Influence Function method, which is a common technique in proving Cramer-Rao lower bounds. But our proof is much more technical than that in (Hao et al., 2021b), since we handle non-linear smooth f.

To our knowledge, (Jiang & Li, 2015) established the first Cramer-Rao lower bound for off policy evaluation in tabular MDP. (Hao et al., 2021b) established the lower bound and proved asymptotic efficiency for using linear function approximation. Our lower bound generalize these previous results.

9. Summary and Discussion

This paper studies statistical properties of fitted Q evaluation using compact, differentiable function approximators. It establishes a set of statistical results including asymptotic normality, finite-sample error bounds, characterization of distribution shift and Cramer-Rao lower bound.

Apart from Z estimator, M estimator is another powerful tool to analysis both asymptotic and non-asymptotic properties of FQE estimators. However, directly applying M estimation theory is not as straightforward as expected. Specifically, FQE is not exactly an M estimator due to the iterative minimization structure, which is not equivalent to minimizing a single loss function. In contrast, FQE is naturally a single Z estimator.

In this paper, we focus on differentiable FQE. Analysis via M estimator is more applied in non-smooth function clas, which is beyond our scope and requires a completely different analysis framework. In particular, when f is non-smooth, there will be no Taylor expansion and hence, no asymptotic normality and no finite sample upper bound. In conclusion, although M estimator will probably not yield results of equal strength as what we have obtained, we can expect its potential application on analysis for non-smooth estimators.

Acknowledgement We are grateful to the support by NSF grants IIS-2107304, CMMI-1653435, AFOSR grant and ONR grant 1006977.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in markov decision processes. *CoRR*, abs/1908.00261, 2019. URL http://arxiv.org/abs/1908.00261.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051. PMLR, 2019.

Chen, X. and Qi, Z. On well-posedness and minimax optimal rates of nonparametric q-function estimation in off-policy evaluation, 2022.

Duan, Y., Jia, Z., and Wang, M. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pp. 2701–2709. PMLR, 2020.

Duan, Y., Wang, M., and Wainwright, M. J. Optimal policy evaluation using kernel-based temporal difference methods. *arXiv preprint arXiv:2109.12002*, 2021.

Efron, B. *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.

Fan, J., Wang, Z., Xie, Y., and Yang, Z. A theoretical analysis of deep q-learning, 2020.

Farahmand, A.-m., Ghavamzadeh, M., Szepesvári, C., and Mannor, S. Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research*, 17(1):4809–4874, 2016.

Feng, Y., Ren, T., Tang, Z., and Liu, Q. Accountable off-policy evaluation with kernel bellman statistics. In *International Conference on Machine Learning*, pp. 3102–3111. PMLR, 2020.

Fonteneau, R., Murphy, S. A., Wehenkel, L., and Ernst, D. Batch mode reinforcement learning based on the synthesis of artificial trajectories. *Annals of operations research*, 208(1):383–416, 2013.

Hao, B., Duan, Y., Lattimore, T., Szepesvári, C., and Wang, M. Sparse feature selection makes batch reinforcement learning more sample efficient. In *International Conference on Machine Learning*, pp. 4063–4073. PMLR, 2021a.

Hao, B., Ji, X., Duan, Y., Lu, H., Szepesvári, C., and Wang, M. Bootstrapping statistical inference for off-policy evaluation, 2021b.

Ji, X., Chen, M., Wang, M., and Zhao, T. Sample complexity of nonparametric off-policy evaluation on low-dimensional manifolds using deep networks. *arXiv preprint arXiv:2206.02887*, 2022.

Jiang, N. and Li, L. Doubly robust off-policy evaluation for reinforcement learning. *CoRR*, abs/1511.03722, 2015. URL http://arxiv.org/abs/1511.03722.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661. PMLR, 2016.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation, 2019.

Jin, X.-K., Liu, X.-H., Jiang, S., and Yu, Y. Hybrid value estimation for off-policy evaluation and offline reinforcement learning. *arXiv preprint arXiv:2206.02000*, 2022.

Kallus, N. and Uehara, M. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *J. Mach. Learn. Res.*, 21:167–1, 2020.

Kosorok, M. R. *Introduction to empirical processes and semiparametric inference.* Springer, 2008.

Kostrikov, I. and Nachum, O. Statistical bootstrapping for uncertainty estimation in off-policy evaluation. *arXiv preprint arXiv:2007.13609*, 2020.

Le, H., Voloshin, C., and Yue, Y. Batch policy learning under constraints. In *International Conference on Machine Learning*, pp. 3703–3712. PMLR, 2019.

Li, L., Munos, R., and Szepesvári, C. Toward minimax off-policy value estimation. In *Artificial Intelligence and Statistics*, pp. 608–616. PMLR, 2015.

Long, J., Han, J., and E, W. An l^2 analysis of reinforcement learning in high dimensions with kernel and neural network approximation. *CoRR*, abs/2104.07794, 2021. URL https://arxiv.org/abs/2104.07794.

Min, Y., Wang, T., Zhou, D., and Gu, Q. Variance-aware off-policy evaluation with linear function approximation. *arXiv preprint arXiv:2106.11960*, 2021.

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

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937. PMLR, 2016.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9 (5), 2008.

Nguyen-Tang, T., Gupta, S., Tran-The, H., and Venkatesh, S. Sample complexity of offline reinforcement learning with deep relu networks, 2021.

Perdomo, J. C., Krishnamurthy, A., Bartlett, P., and Kakade, S. A sharp characterization of linear estimators for offline policy evaluation. *arXiv preprint arXiv:2203.04236*, 2022.

Precup, D. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, pp. 80, 2000.

Rubin, D. B. The bayesian bootstrap. *The annals of statistics*, pp. 130–134, 1981.

Shi, C., Zhang, S., Lu, W., and Song, R. Statistical inference of the value function for reinforcement learning in infinite horizon settings, 2021.

Szepesvári, C. and Munos, R. Finite time bounds for sampling based fitted value iteration. In *Proceedings of the 22nd international conference on Machine learning*, pp. 880–887, 2005.

Thomas, P. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 2139–2148. PMLR, 2016.

Uehara, M., Huang, J., and Jiang, N. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pp. 9659–9668. PMLR, 2020.

Uehara, M., Imaizumi, M., Jiang, N., Kallus, N., Sun, W., and Xie, T. Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency, 2021.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Van Der Vaart, A. W., van der Vaart, A., van der Vaart, A. W., and Wellner, J. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 1996.

Wang, R., Foster, D. P., and Kakade, S. M. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020.

Wang, Y., Wang, R., Du, S. S., and Krishnamurthy, A. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

Xie, T., Ma, Y., and Wang, Y.-X. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *arXiv preprint arXiv:1906.03393*, 2019.

Yang, Z., Jin, C., Wang, Z., Wang, M., and Jordan, M. I. Bridging exploration and general function approximation in reinforcement learning: Provably efficient kernel and neural value iterations. *arXiv e-prints*, pp. arXiv–2011, 2020.

Yin, M. and Wang, Y.-X. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3948–3958. PMLR, 2020.

Zhou, D., He, J., and Gu, Q. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pp. 12793–12802. PMLR, 2021.

A. Proof of Asymptotic Normality and Distributional Consistency

A.1. Proof of Theorem 6.1 and Theorem 7.1

Since Theorem 6.1 and Theorem 7.1 are proven by the same technique, we will prove them simultaneously.

A.1.1. Preliminaries on Glivenko Cantelli and Donsker Function Class

Before we present our main result, we prove some basic properties of the following function class. Below $z_{h,i}(\theta,\cdot)$ is the $i-$th entry of $z_{(h)}(\theta,\cdot)$ for $i\in[d]$.

$$\mathcal{Z} := \bigcup_{h=1}^{H}\bigcup_{i=1}^{d}\left\{z_{h,i}\left(\theta,\cdot\right) : \theta\in\Theta^H\subset\mathbb{R}^{Hd}\right\}. \tag{16}$$

We remark that each function in $\mathcal{Z}$ is a map from $(\mathcal{S}\times\mathcal{A})^H$ to $\mathbb{R}$ and is indexed by $\theta\in\Theta^H$. We present the standard definition of $\mathbb{P}$-Glivenko Cantelli and $\mathbb{P}$-Donsker in asymptotic statistics as follows. In the definition and following proof, we denote $\mathbb{P}$ as a random probability measure and $\mathbb{P}_n$ as its empirical measure. We further denote $\mathbb{G}_n(\cdot) := \sqrt{n}\left(\mathbb{P}_n(\cdot)-\mathbb{P}(\cdot)\right)$ as empirical process.

Definition A.1 (P-Glivenko-Cantelli and P-Donsker(Van der Vaart, 2000))**.** Suppose $\mathcal{U}$ is a measurable function class. We denote $l^\infty(\mathcal{U})$ as the set of all bounded functions on $\mathcal{U}$. We call $\mathcal{U}$ is ($\mathbb{P}$-)Glivenko-Cantelli (or equivalently, strong Glivenko-Cantelli), if

$$\|\mathbb{P}_n u-\mathbb{P}u\|_{\mathcal{U}} := \sup_{u\in\mathcal{U}}|\mathbb{P}_n u-\mathbb{P}u| \xrightarrow{a.s.} 0; \tag{17}$$

We call a $\mathcal{U}$ of measurable functions is ($\mathbb{P}$-)Donsker, if the sequence of processes $\{\mathbb{G}_n u : u\in\mathcal{U}\}$ converges in distribution to a tight limit process G in $l^\infty(\mathcal{U})$. By converging in distribution in $l^\infty(\mathcal{U})$, we mean for every bounded, continuous function $g: l^\infty(\mathcal{U})\to\mathbb{R}$, it always holds that $\mathbb{E}[g(\mathbb{G}_n u)]\to\mathbb{E}[g(G)]$ when n tends to infinity. When there is only one probability measure $\mathbb{P}$, we omit the prefix $\mathbb{P}$- in front of Glivenko-Cantelli or Donsker. We define a class $\mathcal{U}$ of vector-valued functions $u: x\to\mathbb{R}^k$ to be Glivenko-Cantelli or Donsker if the union of the k coordinate classes is Glivenko-Cantelli or Donsker.

Glivenko-Cantelli and Donsker function classes are difficult to verify by definition. A usual sufficient condition involves the concept of bracketing number and bracketing integral.

Definition A.2 (Bracketing Number and Bracketng Integral)**.** For a function $u\in\mathcal{U}$, we define its $L_r(\mathbb{P})$-norm as $\|u\|_{\mathbb{P},r} = (\mathbb{P}|u|^r)^{1/r}$. Given two functions l and h with finite $L_r(\mathbb{P})$-norm (need not to be in $\mathcal{U}$), we define the bracket $[l,h]$ as all functions f with $l\le f\le h$ in the whole domain. An ε-bracket in $L_r(\mathbb{P})$ is a bracket $[l,h]$ with $\mathbb{P}|h-l|^r<\varepsilon^r$. We define the bracketing number $N_{[]}(\varepsilon,\mathcal{U},L_r(\mathbb{P}))$ as the minimal number of ε- brackets needed to cover $\mathcal{U}$, i.e. each function $u\in\mathcal{U}$ is contained in a ε- bracket. The speed of growth of bracketing number are described using the bracketing integral.

$$J_{[]}\left(\delta,\mathcal{U},L_r(\mathbb{P})\right) = \int_0^\delta\sqrt{\log N_{[]}\left(\varepsilon,\mathcal{U},L_r(\mathbb{P})\right)}d\varepsilon. \tag{18}$$

Next two propositions give us sufficient conditions for a measurable function class to be Glivenko-Cantelli or Donsker.

Proposition A.3 (Glivenko-Cantelli (Van Der Vaart et al., 1996))**.** *Every function class $\mathcal{U}$ of measurable functions such that for any $\varepsilon>0$,*

$$N_{[]}(\varepsilon,\mathcal{U},L_1(\mathbb{P}))<\infty$$

is $\mathbb{P}$-Glivenko-Cantelli.

Proposition A.4 (Donsker(Van Der Vaart et al., 1996))**.** *Every function class $\mathcal{U}$ of measurable functions with*

$$J_{[]}\left(1,\mathcal{U},L_2(\mathbb{P})\right)<\infty$$

is $\mathbb{P}$-Donsker.

It is obvious that Glivenko Cantellli and Donsker function class has the following relation.

Proposition A.5. *Every Donsker function class is Glivenko Cantelli.*

A.1.2. Z-Estimator Master Theorem

Our central tool will be the following Z-Estimator Master Theorem. It provides sufficient condition for asymptotic normality of Z-estimators and bootstrapping Z-estimators. We use this lemma to prove both asymptotic normality and distributional consistency.

Lemma A.6 (Z-Estimator Master Theorem(Kosorok, 2008))**.** *Let $u(\theta, \boldsymbol{\tau}) : \mathcal{X}^H \to \mathbb{R}^{Hd}$ be a Borel-measurable function with $\boldsymbol{\tau} \in \mathcal{X}^H$. Let*

$$U(\theta) = \mathbb{E}_{\boldsymbol{\tau}} u(\theta, \boldsymbol{\tau}),\ U_K(\theta) = \frac{1}{K}\sum_{k=1}^{K} u(\theta, \boldsymbol{\tau}_k),\ U_K^{\circ}(\theta) = \frac{1}{K}\sum_{k=1}^{K} W_k u(\theta, \boldsymbol{\tau}_k),$$

where $W^{\circ} = (W_1, W_2, ..., W_K)^{\top}$ is either vanilla bootstrapping weights or multiplier bootstrapping weights. When W° is vanilla bootstrapping weights, W° follows multinomial distribution with parameter $(1/K, 1/K, ..., 1/K)$. When W° is multiplier bootstrapping weights, we have $W_k = \frac{K u_k}{\bar{u}_K}$, where u_k are i.i.d samples from distribution of random vector U which satisfies (15).

Let $\Theta \subset \mathbb{R}^d$ be compact, and assume $\theta^* \in \operatorname{int}(\Theta^H)$ satisfies $U(\theta^*) = 0$. Assume the following:

- *(i) For any sequence $\{\theta^l\} \in \operatorname{int}\Theta^H, U(\theta^l) \to 0$ implies $\|\theta^l - \theta^*\| \to 0$;*
- *(ii) The class $\mathcal{U} = \{u(\theta, \boldsymbol{\tau}) : \theta \in \Theta^H\}$ is P-Glivenko-Cantelli;*
- *(iii) For some $\delta > 0$, the class $\mathcal{U}_\delta := \{u(\theta, \boldsymbol{\tau}) : \theta \in \Theta^H, \|\theta - \theta^*\| \le \delta\}$ is P-Donsker*
- *(iv) $\mathbb{E}\|u(\theta, \boldsymbol{\tau}) - u(\theta^*, \boldsymbol{\tau})\|^2 \to 0$ as $\|\theta - \theta^*\| \to 0$.*
- *(v) $\mathbb{E}\|u(\theta^*, \boldsymbol{\tau})\|^2 < \infty$, and $U(\theta)$ is differentiable at θ^* with non-singular Jacobian matrix $V(\theta^*)$.*
- *(vi) $U_K(\widehat{\theta}_K) = o(K^{-1/2})$ and $U_K^{\circ}(\widehat{\theta}_K^{\circ}) = o(K^{-1/2})$.*

Then, we have

$$\sqrt{K}\left(\widehat{\theta}_K - \theta^*\right) \xrightarrow{d} N\left(0, V(\theta^*)^{-1}\mathbb{E}\left[u(\theta^*, \boldsymbol{\tau})u(\theta^*, \boldsymbol{\tau})^{\top}\right]\left[V(\theta^*)^{-1}\right]^{\top}\right)$$

and

$$\sqrt{K}\left(\widehat{\theta}_K^{\circ} - \widehat{\theta}_K\right) \xrightarrow{d} N\left(0, k_0 V(\theta^*)^{-1}\mathbb{E}\left[u(\theta^*, \boldsymbol{\tau})u(\theta^*, \boldsymbol{\tau})^{\top}\right]\left[V(\theta^*)^{-1}\right]^{\top}\right).$$

conditionally on $\boldsymbol{\tau}_1, ..., \boldsymbol{\tau}_K$. When we use vanilla bootstrap, $k_0 = 1$; when we use multiplier bootstrap, $k_0 = \frac{\eta^2}{m^2}$, where m and η^2 are population mean and variance of distribution of bootstrapping weights.

A.1.3. Completion of the Proof of Asymptotic Normality

All we need to do is to verify that function class $\mathcal{Z}$ satisfies all conditions in Z-Estimator Master Theorem. From compactness of Θ and the uniqueness of root of $Z(\theta)$ in the assumption of Theorem 6.1, (i) and (v) is verified. By continuity of $z(\theta, \boldsymbol{\tau})$ with respect to θ, (iv) holds. (vi) holds by letting $\lambda = o(K^{-1/2})$.

Next we will prove (ii) and (iii). Since every Donsker function class is Glivenko-Cantelli, we only need to prove (iii) for any $\delta > 0$. Then (ii) can be proven by letting $\delta = \operatorname{diam}(\Theta)$. Since $f(\theta, \phi)$ is third times continuously differentiable in a compact set $\Theta \times \Psi$, $f(\theta, \phi)$ has continuous derivatives up to third order. We denote $\alpha = (\alpha_1, \alpha_2, ..., \alpha_d)$ as a multi-index. We define the order of α or the degree of α as

$$|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_n$$

and

$$\partial_\theta^\alpha f(\theta, \phi) = \partial_{\theta_1}^{\alpha_1}\partial_{\theta_2}^{\alpha_2}\cdots\partial_{\theta_d}^{\alpha_d} f = \frac{\partial^{|\alpha|} f}{\partial\theta_1^{\alpha_1}\partial\theta_2^{\alpha_2}\cdots\partial\theta_d^{\alpha_d}} \tag{19}$$

We denote for $l = 1, 2, 3$,

$$\kappa_l := \sup_{\theta\in\Theta, \phi\in[0,1]^m}\ \sup_{|\alpha|=l} |\partial_\theta^\alpha f(\theta, \phi)| < \infty. \tag{20}$$

Then

$$\sup_{\theta\in\Theta,\phi\in[0,1]^m} \|\nabla_\theta f(\theta,\phi)\|_2 \le \sqrt{d}\kappa_1,$$

and

$$\sup_{\theta\in\Theta,\phi\in[0,1]^m} \left\|\nabla_\theta^2 f(\theta,\phi)\right\|_2 \le \sup_{\theta\in\Theta,\phi\in[0,1]^m} \left\|\nabla_\theta^2 f(\theta,\phi)\right\|_{\mathrm{F}} \le d\kappa_2.$$

where $\|\cdot\|_{\mathrm{F}}$ is Frobenius norm. This implies $\|f(\theta,\phi)-f(\theta',\phi)\|_2 \le \sqrt{d}\kappa_1 \|\theta-\theta'\|_2$ and $\|\nabla_\theta f(\theta,\phi)-\nabla_\theta f(\theta',\phi)\|_2 \le d\kappa_2 \|\theta-\theta'\|_2$ for every $\theta,\theta'\in\Theta$ and $\phi\in\Psi$. In the following proof, we denote $\theta^i=(\theta_1^{i\top},\theta_2^{i\top},...,\theta_H^{i\top})^\top\in\Theta^H$ for $i=1,2$ and $\theta_{H+1}^i=0$. We use $\boldsymbol{\tau}$ to represent a random trajectory. We define

$$\varepsilon_{j,h} := f\left(\theta_j^*, \phi(s_h,a_h)\right) - r\left(s_h,a_h\right) - \int_{\mathcal{A}} f\left(\theta_{j+1}^*, \phi(s_{h+1},a')\right)\pi\left(a' \mid s_{h+1}\right) da',$$

then since for any $(s,a)\in\mathcal{S}\times\mathcal{A}$, $r(s,a)\in[0,1]$, we have $Q_h(s,a)=f(\theta_h^*,\phi(s,a))\in[0,H-h+1]$ and $|\varepsilon_{j,h}|\le H-h+1$ for $h,j\in[H]$. Further for $i=1,2$ we define

$$Y_{j,h}^i := f\left(\theta_j^i, \phi(s_h,a_h)\right) - r\left(s_h,a_h\right) - \int_{\mathcal{A}} f\left(\theta_{j+1}^i, \phi(s_{h+1},a)\right)\pi\left(a \mid s_{h+1}\right) da$$

Therefore, we have

$$\begin{aligned}
&\left|z\left(\theta^1,\boldsymbol{\tau}\right)-z\left(\theta^2,\boldsymbol{\tau}\right)\right|^2 \\
\le&\sum_{j=1}^H \left|\sum_{h=1}^H Y_{j,h}^1\cdot\nabla_\theta f\left(\theta_j^1,\phi(s_h,a_h)\right) - Y_{j,h}^2\cdot\nabla_\theta f\left(\theta_j^2,\phi(s_h,a_h)\right)\right|^2 \\
\le&2\sum_{j=1}^H \left|\sum_{h=1}^H Y_{j,h}^1\left[\nabla_\theta f\left(\theta_j^1,\phi(s_h,a_h)\right)-\nabla_\theta f\left(\theta_j^2,\phi(s_h,a_h)\right)\right]\right|^2 + 2\sum_{j=1}^H\left|\sum_{h=1}^H\left(Y_{j,h}^1-Y_{j,h}^2\right)\nabla_\theta f\left(\theta_j^2,\phi(s_h,a_h)\right)\right|^2 \\
\le&2d^2\kappa_2^2\sum_{j=1}^H\left(\sum_{h=1}^H\left|Y_{j,h}^1\right|\cdot\left|\theta_j^1-\theta_j^2\right|\right)^2 + 2d\kappa_1^2\sum_{j=1}^H\left(\sum_{h=1}^H\left|Y_{j,h}^1-Y_{j,h}^2\right|\right)^2.
\end{aligned}$$

Next we bound $Y_{j,h}^i$. Since

$$\begin{aligned}
\left|Y_{j,h}^1\right| &\le \left|\varepsilon_{j,h}\right| + \left|\varepsilon_{j,h}-Y_{j,h}^1\right| \\
&\le (H-h+1)+\kappa_1\sqrt{d}\left[\left|\theta_j^*-\theta_j^1\right|+\left|\theta_{j+1}^*-\theta_{j+1}^1\right|\right] \le H+2\kappa_1\sqrt{d}\delta.
\end{aligned}$$

and

$$\left|Y_{j,h}^1-Y_{j,h}^2\right| \le \kappa_1\sqrt{d}\left[\left|\theta_j^1-\theta_j^2\right|+\left|\theta_{j+1}^1-\theta_{j+1}^2\right|\right],$$

we have

$$\begin{aligned}
&\left|z\left(\theta^1,\boldsymbol{\tau}\right)-z\left(\theta^2,\boldsymbol{\tau}\right)\right|^2 \\
\le&2d^2\kappa_2^2\sum_{j=1}^H\left[\left(H^2+2\kappa_1\delta H\right)^2\left(\sum_{h=1}^H\left|\theta_j^1-\theta_j^2\right|^2\right)\right] + 2\kappa_1^4H^2\left(\sum_{j=1}^H\left|\theta_j^1-\theta_j^2\right|+\left|\theta_{j+1}^1-\theta_{j+1}^2\right|\right)^2 \\
\le&\left[2\kappa_2^2\left(H^2+2\kappa_1\sqrt{d}\delta H\right)^2H+8d^2\kappa_1^4H^3\right]\left\|\theta^1-\theta^2\right\|_2^2.
\end{aligned}$$

We denote $C_1>0$ as where

$$C_1^2 = 2d^2\kappa_2^2\left(H^2+2\kappa_1\sqrt{d}\delta H\right)^2H+8d^2\kappa_1^4H^3. \tag{21}$$

Then $\left|z\left(\theta^1,\boldsymbol{\tau}\right)-z\left(\theta^2,\boldsymbol{\tau}\right)\right| \le C_1\left\|\theta^1-\theta^2\right\|_2$. Therefore, for every $\|\theta-\theta^*\|\le\delta$, we consider the high-dimensional bracket

$$[z(\theta,\boldsymbol{\tau})-C_1\sqrt{Hd}\boldsymbol{\varepsilon}, z(\theta,\boldsymbol{\tau})+C_1\sqrt{Hd}\boldsymbol{\varepsilon}],$$

where $\boldsymbol{\varepsilon}$ is the Hd-dimension vector with every entry being ε. Since the brackets we consider are one-dimensional, these brackets have $L_2(\mathbb{P})$-size of $2C_1\sqrt{Hd}\varepsilon$. If $\left\|\theta^1-\theta^2\right\|_\infty \leq \varepsilon$, then $\left|z\left(\theta^1, \boldsymbol{\tau}\right)-z\left(\theta^2, \boldsymbol{\tau}\right)\right| \leq \sqrt{Hd}C_1\varepsilon$, which implies $z(\theta^2,\boldsymbol{\tau}) \in [z(\theta^1,\boldsymbol{\tau})-C_1\sqrt{Hd}\boldsymbol{\varepsilon}, z(\theta^1,\boldsymbol{\tau})+C_1\sqrt{Hd}\boldsymbol{\varepsilon}]$, and every dimension of $z(\theta^2,\boldsymbol{\tau})$ is contained in the corresponding one-dimensional bracket. And since every dimension of $\theta^1-\theta^2$ spans within a distance no more than 2δ, we can partition every dimension of $\|\theta-\theta^*\| \leq \delta$ into grids with meshwidth being 2ε. By dividing like this, we will get no more than $(\frac{\delta}{\varepsilon})^{Hd}$ hypercubes. For every $\|\theta-\theta^*\| \leq \delta$, there exists one hypercube whose center θ^c satisfies $\|\theta-\theta^c\|_\infty \leq \varepsilon$. Then we have $z(\theta,\boldsymbol{\tau}) \in [z(\theta^c,\boldsymbol{\tau})-C_1\sqrt{Hd}\boldsymbol{\varepsilon}, z(\theta^c,\boldsymbol{\tau})+C_1\sqrt{Hd}\boldsymbol{\varepsilon}]$, and every component of $z(\theta,\boldsymbol{\tau})$ is contained in the corresponding one-dimensional bracket. Therefore, every function of $\mathcal{Z}_\delta$ is contained in a bracket. Summing over all brackets, we have

$$N_{[]}\left(2C_1\sqrt{Hd}\varepsilon, \mathcal{Z}_\delta, L_r(\mathbb{P})\right) \leq Hd\left(\frac{\delta}{\varepsilon}\right)^{Hd}.$$

This implies

$$N_{[]}\left(\varepsilon, \mathcal{Z}_\delta, L_r(\mathbb{P})\right) \leq Hd\left(\frac{2C_1\sqrt{Hd}\delta}{\varepsilon}\right)^{Hd}. \tag{22}$$

Because that the bracketing number is a decreasing function of ε, whether the bracketing integral converges or not depends on the asymptotic behavior of bracketing number for $\varepsilon \to 0$. From the estimation above, $\sqrt{\log N_{[]}(\varepsilon, \mathcal{Z}_\delta, L_2(\mathbb{P}))}$ is of order $O(\sqrt{\log(\frac{1}{\varepsilon})})$, the integral of which converges. By Glibenko-Cantelli's Theorem and Donsker's Theorem, we know $\mathcal{Z}$ is both Glivenko Cantelli and Donsker. In this way, we verify all conditions of Z-Estimator Master Theorem and we have

$$\sqrt{K}\left(\widehat{\theta}_K-\theta^*\right) \xrightarrow{d} N\left(0, \frac{1}{H}\Sigma_*^{-1}\Omega\Sigma_*^{-\top}\right); \tag{23}$$

and conditioned on $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, ..., \boldsymbol{\tau}_K$,

$$\sqrt{K}\left(\widehat{\theta}_K^\circ-\widehat{\theta}_K\right) \xrightarrow{d} N\left(0, \frac{k_0}{H}\Sigma_*^{-1}\Omega\Sigma_*^{-\top}\right); \tag{24}$$

where

$$\Sigma_* = \frac{1}{H}\left.\frac{\partial}{\partial\xi}Z(\xi)\right|_{\xi=\theta^*} \text{ and } \Omega = \frac{1}{H}\,\mathbb{E}_{\boldsymbol{\tau}}\left[z(\xi,\boldsymbol{\tau})z(\xi,\boldsymbol{\tau})^\top\right]\big|_{\xi=\theta^*}. \tag{25}$$

When we adopt vanilla bootstrap strategy, $k_0=1$, when we adopt multiplier bootstrap, $k_0=\frac{\eta^2}{m^2}$. Finally we compute the covariance matrix in the asymptotic variance. Below, we write out Σ_* as a partitioned matrix:

$$\Sigma_* = \begin{pmatrix} \Sigma_1 & -A_1 & 0 & \dots & 0 \\ 0 & \Sigma_2 & -A_2 & \dots & 0 \\ 0 & 0 & \Sigma_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \Sigma_H \end{pmatrix}$$

where for $h \in [H]$,

$$\begin{aligned}\Sigma_h = \frac{1}{H}\mathbb{E}_{\boldsymbol{\tau}}\Bigg\{&\sum_{h'=1}^{H}\left(f(\theta_h^*, \phi(s_{h'}, a_{h'})) - r(s_{h'}, a_{h'}) - \int_{\mathcal{A}} f(\theta_{h+1}^*, \phi(s_{h'+1}, a))\pi(a \mid s_{h'+1})da\right)\nabla_{\theta\theta}^2 f(\theta_h^*, \phi(s_{h'}, a_{h'})) \\ &+ \sum_{h'=1}^{H}\left(\nabla_\theta f(\theta_h^*, \phi(s_{h'}, a_{h'}))\right)^\top\left(\nabla_\theta f(\theta_h^*, \phi(s_{h'}, a_{h'}))\right)\Bigg\}.\end{aligned}$$

Since

$$\mathbb{E}_{\boldsymbol{\tau}}\left\{f(\theta_h^*, \phi(s_{h'}, a_{h'})) - r(s_{h'}, a_{h'}) - \int_{\mathcal{A}} f(\theta_{h+1}^*, \phi(s_{h'+1}, a))\pi(a \mid s_{h'+1})da \,\middle|\, s_{h'}, a_{h'}\right\} = 0,$$

by taking conditional expectation we find the first term in the expectation vanishes and

$$\Sigma_h = \frac{1}{H}\mathbb{E}_{\boldsymbol{\tau}}\left\{\sum_{h'=1}^{H}\left(\nabla_\theta f(\theta_h^*, \phi(s_{h'}, a_{h'}))\right)^\top\left(\nabla_\theta f(\theta_h^*, \phi(s_{h'}, a_{h'}))\right)\right\}. \tag{26}$$

For $h \in [H-1]$,

$$A_h = -\frac{1}{H}\nabla_\beta \mathbb{E}_{\boldsymbol{\tau}}\left\{\sum_{h'=1}^{H}\left(f(\alpha,\phi(s_{h'},a_{h'})) - r(s_{h'},a_{h'}) - \int_{\mathcal{A}} f(\beta,\phi(s_{h'+1},a))\pi(a \mid s_{h'+1})da\right)\nabla_\theta f(\alpha,\phi(s_{h'},a_{h'}))\right\}\Big|_{\alpha=\theta_h^*,\beta=\theta_{h+1}^*}.$$

Here the gradient means partial derivatives with respect to β and take $\alpha=\theta_h^*, \beta=\theta_{h+1}^*$. Therefore, the inverse of Σ_* can be expressed as

$$\Sigma_*^{-1} = \begin{pmatrix} \Sigma_1^{-1} & \Sigma_1^{-1}A_1\Sigma_2^{-1} & \Sigma_1^{-1}A_1\Sigma_2^{-1}A_2\Sigma_3^{-1} & \dots & \Sigma_1^{-1}A_1\Sigma_2^{-1}...A_{H-1}\Sigma_H^{-1} \\ 0 & \Sigma_2^{-1} & \Sigma_2^{-1}A_2\Sigma_3^{-1} & \dots & \Sigma_2^{-1}A_2\Sigma_3^{-1}...A_{H-1}\Sigma_H^{-1} \\ 0 & 0 & \Sigma_3^{-1} & \dots & \Sigma_3^{-1}A_3\Sigma_4^{-1}...A_{H-1}\Sigma_H^{-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \Sigma_H^{-1} \end{pmatrix}.$$

Moreover we express Ω as $(\Omega_{i,j})_{i,j\in[H]}$ where $\Omega_{i,j} \in \mathbb{R}^{d\times d}$ is

$$\Omega_{i,j} = \frac{1}{H}\sum_{h_1=1}^{H}\sum_{h_2=1}^{H}\mathbb{E}\left[\left(\nabla_\theta f\left(\theta_i^*,\phi(s_{h_1},a_{h_1})\right)\right)^\top\left(\nabla_\theta f\left(\theta_j^*,\phi(s_{h_2},a_{h_2})\right)\right)\varepsilon_{i,h_1}\varepsilon_{j,h_2}\right] \in \mathbb{R}^{d\times d}$$

where $\varepsilon_{j,h}$ is defined in Theorem 6.1. We take the conditional expectation to find that the cross terms vanish. Therefore the expression of Ω_{ij} can be simplified as

$$\Omega_{i,j} = \mathbb{E}\left[\frac{1}{H}\sum_{h=1}^{H}\left(\nabla_\theta f\left(\theta_i^*,\phi(s_h,a_h)\right)\right)^\top\left(\nabla_\theta f\left(\theta_j^*,\phi(s_h,a_h)\right)\right)\varepsilon_{i,h}\varepsilon_{j,h}\right] \in \mathbb{R}^{d\times d}. \tag{27}$$

Therefore, we have the following convergence(the second one is conditional on the original data).

$$\sqrt{K}\left(\widehat{\theta}_1 - \theta_1\right), \quad \sqrt{K}\left(\widehat{\theta}_1^\circ - \widehat{\theta}_1\right) \xrightarrow{d} N\left(0, \frac{1}{H}\sum_{h_1=1}^{H}\sum_{h_2=1}^{H}\left(\prod_{i=1}^{h_1-1}\Sigma_i^{-1}A_i\right)\Sigma_{h_1}^{-1}\Omega_{h_1,h_2}\Sigma_{h_2}^{-1}\left(\prod_{j=1}^{h_2-1}\Sigma_j^{-1}A_j\right)^\top\right); \tag{28}$$

Finally we use multivariate delta method and eventually get

$$\sqrt{K}\left(\widehat{v}_\pi - v_\pi\right), \quad \sqrt{K}\left(\widehat{v}_\pi^\circ - \widehat{v}_\pi\right) \xrightarrow{d} N(0,\sigma^2), \tag{29}$$

where

$$\sigma^2 := \sum_{h_1=1}^{H}\sum_{h_2=1}^{H}\left(\int_{\mathcal{S}\times\mathcal{A}}\nabla_\theta f(\theta_1^*,\phi(s,a))\pi(a \mid s)\xi(s)dads\right)$$
$$\cdot\left(\prod_{i=1}^{h_1-1}\Sigma_i^{-1}A_i\right)\Sigma_{h_1}^{-1}\Omega_{h_1,h_2}\Sigma_{h_2}^{-1}\left(\prod_{j=1}^{h_2-1}\Sigma_j^{-1}A_j\right)^\top\left(\int_{\mathcal{S}\times\mathcal{A}}\nabla_\theta f(\theta_1^*,\phi(s,a))\pi(a \mid s)\xi(s)dads\right)^\top.$$

To simplify expressions for Σ_h and A_h we define

$$F(\alpha,\beta) := \frac{1}{2}\nabla_\alpha\mathbb{E}_{\bar{\mu}}\left\{\left[f(\alpha,\phi(s,a)) - r(s,a) + \int_{\mathcal{A}} f(\beta,\phi(s',a'))\pi(a' \mid s')da'\right]^2\right\}$$

for $\alpha,\beta \in \Theta$. We notice that

$$\Sigma_h = \nabla_\alpha^\top F\big|_{\alpha=\theta_h^*,\beta=\theta_{h+1}^*} \text{ and } A_h = \nabla_\beta^\top F\big|_{\alpha=\theta_h^*,\beta=\theta_{h+1}^*}.$$

Here we define another operator $\mathbb{T}$ to simplify the expression for σ^2. Policy completeness under parameterized $\mathcal{F}$ can induce a mapping on Θ, which is defined as

$$f_{\mathbb{T}\theta}(s,a) = (\mathcal{P} f_\theta)(s,a)? \tag{30}$$

That is, $\mathbb{T}\theta$ is the corresponding parameter of $\mathcal{P}f$. Note that $F(\theta_h^*, \theta_{h+1}^*) = 0$ and $F(\cdot,\cdot)$ is continuously differentiable in $\Theta \times \Theta$. Since the Jacobian matrix of $Z(\theta)$ at θ^* is non-singular and is partitioned diagonal matrix with Σ_h on its diagonal, we have Σ_h is non-singular for $h \in [H]$. Therefore, we apply the theorem of implicit function and get that, for any $h \in [H]$, there exists unique vector-valued function G_h defined on a neighbourhood of θ_{h+1}^* inside Θ such that $F(G_h(\theta_{h+1}^*), \theta_{h+1}^*) = 0$. By definition of the operator $\mathbb{T}$, without loss of generality we let each G_h be a confinement of $\mathbb{T}$. That is, $\mathbb{T}$ is actually the operator defined on every neighbourhood of θ_h^* which satisfies (30). We denote the Jacobian matrix of $\mathbb{T}$ at θ as $D\mathbb{T}(\theta)$. From theorem of implicit functions we know $\mathbb{T}$ has continuous Jacobian matrix on neighborhood of any θ_h^*. If we use $D\mathbb{T}(\cdot)$ to denote its Jacobian matrix, then

$$D\mathbb{T}(\theta_{h+1}^*) = \Sigma_h^{-1} A_h.$$

Therefore,

$$\begin{aligned}
&\left(\int_{\mathcal{S}\times\mathcal{A}} \nabla_{\theta_1^*} f(\theta_1^*, \phi(s,a)) \pi(a \mid s) \xi(s) da ds\right) \left(\prod_{i=1}^{h-1} \Sigma_i^{-1} A_i\right) \\
=&\left(\int_{\mathcal{S}\times\mathcal{A}} \nabla_{\theta_1^*} f(\theta_1^*, \phi(s,a)) \pi(a \mid s) \xi(s) da ds\right) \left(\prod_{i=1}^{h-1} D\mathbb{T}(\theta_{i+1})\right) \\
=&\mathbb{E}\left[\nabla_{\theta_h^*} f\left(\mathbb{T}^{h-1}\left(\theta_h^*\right), \phi(s_1,a_1)\right) \middle| s_1 \sim \xi(\cdot), a_1 \sim \pi(\cdot \mid s_1)\right].
\end{aligned}$$

This expectation is actually dependent on transition probability and independent of policy, hence we can transform it to expectation over population generated by target policy π.

$$\begin{aligned}
&\left(\int_{\mathcal{S}\times\mathcal{A}} \nabla_{\theta_1^*} f(\theta_1^*, \phi(s,a)) \pi(a \mid s) \xi(s) da ds\right) \left(\prod_{i=1}^{h-1} \Sigma_i^{-1} A_i\right) \\
=&\mathbb{E}^\pi\left[\nabla_{\theta_h^*} f\left(\mathbb{T}^{h-1}\left(\theta_h^*\right), \phi(s_1,a_1)\right) \middle| s_1 \sim \xi(\cdot), a_1 \sim \pi(\cdot \mid s_1)\right] \\
=&\mathbb{E}^\pi\left[\nabla_{\theta_h^*} \int_{\mathcal{S}\times\mathcal{A}} f(\mathbb{T}^{h-2}\left(\theta_h^*\right), \phi(s_2,a_2)) \pi(a_2 \mid s_2) p(s_2 \mid s_1, a_1) da_2 ds_1 \middle| s_1 \sim \xi(\cdot), a_1 \sim \pi(\cdot \mid s_1)\right] \\
=&\mathbb{E}^\pi\left[\nabla_{\theta_h^*} f(\mathbb{T}^{h-2}\left(\theta_h^*\right), \phi(s_2,a_2)) \middle| s_1 \sim \xi(\cdot), a_1 \sim \pi(\cdot \mid s_1)\right] \\
=&\ldots \\
=&\mathbb{E}^\pi\left[\nabla_{\theta_h^*} f\left(\theta_h^*, \phi(s_h,a_h)\right) \middle| s_1 \sim \xi(\cdot), a_1 \sim \pi(\cdot \mid s_1)\right] = \nu_h^\top.
\end{aligned} \tag{31}$$

The last equation is due to the definition of $\nu_h^\top$ in Theorem 6.1. Then we have

$$\sigma^2 = \frac{1}{H} \sum_{h_1,h_2=1}^{H} \nu_{h_1}^\top \Sigma_{h_1}^{-1} \Omega_{h_1,h_2} \Sigma_{h_2}^{-1} \nu_{h_2}. \tag{32}$$

This finishes our proof for asymptotic normality.

A.2. Proof of Corollary 6.3

This corollary can be derived directly from the following lemma.

Lemma A.7 (Prohorov's Theorem, Theorem 2.4 in (Van der Vaart, 2000))**.** *Let X_n be random variables in $\mathbb{R}^k$ and $X_n \xrightarrow{d} X$ for some X, then $\{X_n : n \in \mathbb{N}\}$ is uniformly tight, i.e. for every $\delta > 0$, there exists a constant B, such that*

$$\sup_{n\in\mathbb{N}} \mathbb{P}\left(\|X_n\|_2 > B\right) < \delta.$$

A.3. Proof of Corollary 7.2

We define $\Phi(t) := \mathbb{P}\left(N(0, k_0\sigma^2) \le t\right)$, where σ^2 is defined as (7) and $N(0,\sigma^2)$ represents a random variable which follows this distribution. From the asymptotic normality and distributional consistency, we have when $K \to \infty$,

$$\mathbb{P}_{\mathcal{D}}\left(\sqrt{K}\left(\widehat{v}_\pi - v_\pi\right) \le t\right) \to \mathbb{P}\left(N(0, k_0\sigma^2) \le \sqrt{k_0}t\right) = \Phi(\sqrt{k_0}t), \quad \mathbb{P}_{W^\circ|\mathcal{D}}\left(\sqrt{K}\left(\widehat{v}^\circ_\pi - \widehat{v}_\pi\right) \le t\right) \to \Phi\left(t\right).$$

According to lemma 21.2 in (Van der Vaart, 2000), denote the cumulative distribution functions of $\sqrt{K}\left(\widehat{v}^\circ_\pi - \widehat{v}_\pi\right)$ as Φ_K, then $\Phi_K \xrightarrow{d} \Phi$ implies $\Phi_K^{-1} \xrightarrow{d} \Phi^{-1}$. Therefore, $\sqrt{K}q_\delta^\pi = \Phi_K^{-1}\left(\delta\right) \to \Phi^{-1}\left(\delta\right)$. Therefore,

$$\mathbb{P}_{\mathcal{D}W^\circ}\left(\widehat{v}_\pi - v_\pi \ge \frac{1}{\sqrt{k_0}}q^\pi_{\frac{\delta}{2}}\right) = \mathbb{P}_{\mathcal{D}W^\circ}\left(\sqrt{K}\left(\widehat{v}_\pi - v_\pi\right) \ge \sqrt{\frac{K}{k_0}}q^\pi_{\frac{\delta}{2}}\right) \to \mathbb{P}_{\mathcal{D}W^\circ}\left(N\left(0,\sigma^2\right) \ge \frac{1}{\sqrt{k_0}}\Phi^{-1}\left(\frac{\delta}{2}\right)\right) = 1-\frac{\delta}{2}.$$

We can bound $\mathbb{P}_{\mathcal{D}W^\circ}\left(\widehat{v}_\pi - v_\pi \le -\frac{1}{\sqrt{k_0}}q^\pi_{1-\frac{\delta}{2}}\right)$ similarly, and by an argument of union bound we can get the result for asymptotic confidence interval.

B. Proof of Theorem 6.5

We are going to decompose FQE error $\widehat{v}_\pi - v_\pi$ into first order term and higher order term. Since we let $\lambda = 0$, we do not have bias term containing λ. Since we have

$$\widehat{v}_\pi - v_\pi = \int_{\mathcal{S}\times\mathcal{A}}\left[f(\widehat{\theta}_1, \phi(s,a)) - f(\theta_1^*, \phi(s,a))\right]\pi(a \mid s)\xi(s)dads,$$

the first order term will be

$$\int_{\mathcal{S}\times\mathcal{A}}\left[\nabla_\theta f(\theta_1^*, \phi(s,a)) \cdot \left(\widehat{\theta}_1 - \theta_1^*\right)\right]\pi(a \mid s)\xi(s)dads.$$

Next we expand $\widehat{\theta}_1 - \theta_1$. From multivariate Taylor expansion we have

$$\sqrt{K}\left[Z(\widehat{\theta}_K) - Z(\theta^*)\right] = \sqrt{K}H\Sigma_*(\widehat{\theta}_K - \theta^*) + \sqrt{K}R_K. \tag{33}$$

where R_K is Taylor remainder and Σ_* is defined in (A.1.3). Notice that $Z(\theta^*) = Z_K(\widehat{\theta}_K) = 0$, we have

$$\begin{aligned}\sqrt{K}\left(\widehat{\theta}_K - \theta^*\right) &= \frac{\sqrt{K}}{H}\Sigma_*^{-1}\left(Z(\widehat{\theta}_K) - Z(\theta^*)\right) - \frac{\sqrt{K}}{H}\Sigma_*^{-1}R_K \\ &= -\frac{1}{\sqrt{K}H}\Sigma_*^{-1}\sum_{k=1}^K z\left(\theta^*, \boldsymbol{\tau}_k\right) + \frac{\Sigma_*^{-1}}{H}\left(\sqrt{K}\mathbb{E}_{\boldsymbol{\tau}} z(\widehat{\theta}_K, \boldsymbol{\tau}) + \frac{1}{\sqrt{K}}\sum_{k=1}^K z(\theta^*, \boldsymbol{\tau}_k)\right) - \frac{\sqrt{K}}{H}\Sigma_*^{-1}R_K.\end{aligned} \tag{34}$$

Since the true value function is explicitly dependent only on θ_1^*, we take the first d entries of this decomposition. We denote $R_{K,h}$ as the vector comprising $(h-1)d+i$-th to hd-th entry of R_K, hence

$$\begin{aligned}\widehat{\theta}_1 - \theta_1^* = &-\frac{1}{KH}\sum_{h=1}^H\left(\prod_{i=1}^{h-1}\Sigma_i^{-1}A_i\right)\Sigma_h^{-1}\left(\sum_{k=1}^K z_{(h)}(\theta^*, \boldsymbol{\tau}_k)\right) \\ &+ \frac{1}{H}\sum_{h=1}^H\left(\prod_{i=1}^{h-1}\Sigma_i^{-1}A_i\right)\Sigma_h^{-1}\left(\mathbb{E}_{\boldsymbol{\tau}} z_{(h)}(\widehat{\theta}_K, \boldsymbol{\tau}) + \frac{1}{K}\sum_{k=1}^K z_{(h)}(\theta^*, \boldsymbol{\tau}_k)\right) - \frac{1}{H}\sum_{h=1}^H\left(\prod_{i=1}^{h-1}\Sigma_i^{-1}A_i\right)\Sigma_h^{-1}R_{K,h}.\end{aligned}$$

Therefore, the first order term of total error is

$$I_1 := -\frac{1}{KH}\int_{\mathcal{S}\times\mathcal{A}}\left[\nabla_\theta f(\theta_1^*, \phi(s,a)) \cdot \sum_{h=1}^H\left(\prod_{i=1}^{h-1}\Sigma_i^{-1}A_i\right)\Sigma_h^{-1}\left(\sum_{k=1}^K z_{(h)}(\theta^*, \boldsymbol{\tau}_k)\right)\right]\pi(a \mid s)\xi(s)dads, \tag{35}$$

and the remaining part will be the higher order term.

$$I_2 := \widehat{v}_\pi - v_\pi - I_1. \tag{36}$$

B.1. First Order Term

The only difference between Variance aware error bound (9) and Reward-free error bound (10) is the contraction technique for the first order term. We deal with them separately.

B.1.1. Variance Aware Error Bound

We first bound I_1 using Bernstein's Inequality (lemma E.1). From this upper bound we see the dependency of its dominant term on asymptotic variance σ^2 in Theorem 6.1. By the same argument of (31), we have

$$\left(\int_{\mathcal{S}\times\mathcal{A}} \nabla_{\theta_1^*} f\left(\theta_1^*, \phi(s,a)\right) \pi(a \mid s)\xi(s) da ds\right)\left(\prod_{i=1}^{h-1} \Sigma_i^{-1} A_i\right) = \nu_h^\top.$$

We define

$$e_k := -\frac{1}{H}\int_{\mathcal{S}\times\mathcal{A}}\left[\nabla_\theta f(\theta_1, \phi(s,a)) \cdot \sum_{h=1}^{H}\left(\prod_{i=1}^{h-1}\Sigma_i^{-1}A_i\right)\Sigma_h^{-1}\cdot z_{(h)}(\theta^*, \boldsymbol{\tau}_k)\right]\pi(a \mid s)\xi(s) da ds.$$

Then by the derivation in (31), we have

$$\begin{aligned}
e_k &= -\frac{1}{H}\sum_{h=1}^{H}\nu_h^\top \Sigma_h^{-1} z_{(h)}(\theta^*, \boldsymbol{\tau}_k) \\
&= -\frac{1}{H}\sum_{h=1}^{H}\sum_{j=1}^{H}\nu_h^\top \Sigma_h^{-1}\left[\nabla_\theta^\top f(\theta_h^*, \phi(s_j^k, a_j^k))\right]\varepsilon_{h,j}^k;
\end{aligned}$$

The first order term will be independent sum of e_k,

$$I_1 = \frac{1}{K}\sum_{k=1}^{K} e_k.$$

From assumption 6.4 and because of $\left|\varepsilon_{h,j}^k\right| \le H-h+1$, we can bound e_k by

$$\begin{aligned}
|e_k| &\le \frac{1}{H}\sum_{h=1}^{H}\sum_{j=1}^{H}\left\|\nu_h^\top \Sigma_h^{-\frac{1}{2}}\right\|_2 \left\|\Sigma_h^{-\frac{1}{2}}\cdot\nabla_\theta^\top f(\theta_h, \phi(s_j^k, a_j^k))\right\|_2 \left|\varepsilon_{h,j}^k\right| \\
&\le \sqrt{C_2 d}\sum_{h=1}^{H}(H-h+1)\sqrt{\nu_h^\top \Sigma_h^{-1}\nu_h}.
\end{aligned}$$

Then we calculate the variance for e_k. By definition of θ_h^*, we have

$$\mathbb{E}\left[\left[\nabla_\theta^\top f(\theta_h^*, \phi(s_j^k, a_j^k))\right]\varepsilon_{h,j}^k\right] = 0,$$

hence $\mathbb{E}\left[e_k\right] = 0$ and $\mathrm{Var}\left[e_k\right] = \mathbb{E}\left[e_k^2\right]$ for $k \in [K]$. We have

$$\mathbb{E}\left[e_k^2\right] = \frac{1}{H^2}\sum_{h_1=1}^{H}\sum_{h_2=1}^{H}\sum_{j=1}^{H}\sum_{l=1}^{H}\left[\nu_{h_1}^\top \Sigma_{h_1}^{-1}\mathbb{E}\left[\left(\nabla_\theta^\top f(\theta_{h_1}^*, \phi(s_j^k, a_j^k))\right)\left(\nabla_\theta f(\theta_{h_2}^*, \phi(s_l^k, a_l^k))\right)\varepsilon_{h_1,j}^k\varepsilon_{h_2,l}^k\right]\Sigma_{h_2}^{-1}\nu_{h_2}\right].$$

By conditional expectation, when $j < l$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\nabla_\theta^\top f(\theta_{h_1}^*, \phi(s_j^k, a_j^k))\right)\left(\nabla_\theta f(\theta_{h_2}^*, \phi(s_l^k, a_l^k))\right)\varepsilon_{h_1,j}^k\varepsilon_{h_2,l}^k\right] \\
=&\mathbb{E}\left\{\mathbb{E}\left[\left(\nabla_\theta^\top f(\theta_{h_1}^*, \phi(s_j^k, a_j^k))\right)\left(\nabla_\theta f(\theta_{h_2}^*, \phi(s_l^k, a_l^k))\right)\varepsilon_{h_1,j}^k\varepsilon_{h_2,l}^k\middle| s_j^k, a_j^k\right]\right\} = 0,
\end{aligned}$$

hence all cross term vanish and

$$\mathbb{E}\left[e_k^2\right]=\frac{1}{H^2}\sum_{h_1=1}^H\sum_{h_2=1}^H\sum_{j=1}^H\left[\nu_{h_1}^\top\Sigma_{h_1}^{-1}\mathbb{E}\left[\left(\nabla_\theta^\top f(\theta_{h_1}^*,\phi(s_j^k,a_j^k))\right)\left(\nabla_\theta f(\theta_{h_2}^*,\phi(s_j^k,a_j^k))\right)\varepsilon_{h_1,j}^k\varepsilon_{h_2,j}^k\right]\Sigma_{h_2}^{-1}\nu_{h_2}\right]$$
$$=\frac{1}{H}\sum_{h_1=1}^H\sum_{h_2=1}^H\sum_{j=1}^H\left[\nu_{h_1}^\top\Sigma_{h_1}^{-1}\Omega_{h_1,h_2}\Sigma_{h_2}^{-1}\nu_{h_2}\right]=\sigma^2,$$

where Ω_{h_1,h_2} and σ^2 are defined as in Theorem 6.1. We use lemma E.1, then we have for any $\varepsilon>0$,

$$\mathbb{P}\left(\left\|\sum_{k=1}^K e_k\right\|\geq\varepsilon\right)\leq 2\exp\left(-\frac{\varepsilon^2/2}{K\sigma^2+\varepsilon\sqrt{C_2 d}\sum_{h=1}^H(H-h+1)\sqrt{\nu_h^\top\Sigma_h^{-1}\nu_h}/3}\right)$$

We choose

$$\varepsilon:=\ln\left(\frac{2}{\delta}\right)\left[\sqrt{2K}\sigma+\frac{2}{3}\sqrt{C_2 d}\sum_{h=1}^H(H-h+1)\sqrt{\nu_h^\top\Sigma_h^{-1}\nu_h}\right],$$

then RHS is bounded by $\delta>0$. Therefore, with probability at least $1-\delta$,

$$|I_1|\leq\ln\left(\frac{2}{\delta}\right)\left[\sqrt{\frac{2\sigma^2}{K}}+\frac{2}{3K}\sqrt{C_2 d}\sum_{h=1}^H(H-h+1)\sqrt{\nu_h^\top\Sigma_h^{-1}\nu_h}\right].$$

This gives one upper bound for the first order term. Notice that this bound depends on asymptotic variance σ^2. When we apply this bound into final upper bound, we get variance-aware error bound (9).

B.1.2. Reward-Free Error Bound

Next, we are going to replace this dependence on σ^2 with a dependence on a reward-free χ^2 divergence, with which we give an upper bound for the worst instance of reward. The main difference is that we use another decomposition of the first order term and use Freedman's Inequality as our central tool. We can simplify the expression for I_1 as

$$I_1=-\frac{1}{KH}\sum_{k=1}^K\sum_{h=1}^H\sum_{j=1}^H\nu_h^\top\Sigma_h^{-1}\left[\nabla_\theta^\top f\left(\theta_h^*,\phi\left(s_j^k,a_j^k\right)\right)\right]\varepsilon_{h,j}^k,$$

where

$$\varepsilon_{h,j}^k=f\left(\theta_h^*,\phi\left(s_j^k,a_j^k\right)\right)-r\left(s_j^k,a_j^k\right)-\int_{\mathcal{A}}f(\theta_{h+1}^*,\phi(s_{j+1}^k,a'))\pi(a'\mid s_{j+1}^k)da'.$$

Below we can decompose I_1 into $N=KH$ items.We denote $\varepsilon_{h,n}=\varepsilon_{h,j}^k$ and $(s_n,a_n)=(s_j^k,a_j^k)$ if $n=(h-1)K+k$. Then if we define

$$u_n:=-\sum_{h=1}^H\nu_h^\top\Sigma_h^{-1}\left[\nabla_\theta^\top f\left(\theta_h^*,\phi\left(s_n,a_n\right)\right)\right]\varepsilon_{h,n},$$

then

$$I_1=\frac{1}{N}\sum_{n=1}^N u_n.$$

Define $\mathcal{F}_n$ as the σ field generated by $s_1,a_1,....,s_n,a_n$, then $\{\mathcal{F}_n\}_{n=1}^N$ is a filtration. Since

$$\mathbb{E}\left[\nabla_\theta^\top f\left(\theta_h^*,\phi\left(s_n,a_n\right)\right)\left[f\left(\theta_h^*,\phi\left(s_n,a_n\right)\right)-r\left(s_n,a_n\right)-\int_{\mathcal{A}}f\left(\theta_{h+1}^*,\phi\left(s_{n+1},a'\right)\right)\pi\left(a'\mid s_{n+1}\right)da'\right]\middle|s_n,a_n\right]$$
$$=\nabla_\theta^\top\mathbb{E}\left[\left[f\left(\theta,\phi\left(s_n,a_n\right)\right)-r\left(s_n,a_n\right)-\int_{\mathcal{A}}f\left(\theta_{h+1}^*,\phi\left(s_{n+1},a'\right)\right)\pi\left(a'\mid s_{n+1}\right)da'\right]^2\middle|s_n,a_n\right]\Bigg|_{\theta=\theta_h^*}=0,$$

we have $\mathbb{E}\left[u_n \mid \mathcal{F}_n\right]=0$, and $\{u_n\}_{n=1}^N$ is a martingale difference sequence. We use Freedman's Inequality (lemma E.2) to analyze I_1. Under data coverage assumption (Assumption 6.4), we have

$$\begin{aligned}
|u_n| &\leq \sum_{h=1}^H\left\|\nu_h^{\top} \Sigma_h^{-\frac{1}{2}}\right\|\left\|\Sigma_h^{-\frac{1}{2}} \nabla_\theta^{\top} f\left(\theta_h^*, \phi\left(s_j^k, a_j^k\right)\right)\right\|\left|\varepsilon_{h, n}\right| \\
&\leq \sqrt{C_2 d} \sum_{h=1}^H(H-h+1) \sqrt{\nu_h^{\top} \Sigma_h^{-1} \nu_h} .
\end{aligned}$$

Next we estimate the conditional variance $\operatorname{Var}\left[u_n \mid \mathcal{F}_n\right]$.

$$\begin{aligned}
\operatorname{Var}\left[u_n \mid \mathcal{F}_n\right] &= \mathbb{E}\left[\left(\sum_{h=1}^H \nu_h^{\top} \Sigma_h^{-1}\left[\nabla_\theta^{\top} f\left(\theta_h^*, \phi\left(s_n, a_n\right)\right)\right] \varepsilon_{h, n}\right)^2 \middle| \mathcal{F}_n\right] \\
&\leq \left(\sum_{h=1}^H \frac{\sqrt{\nu_h^{\top} \Sigma_h^{-1} \nu_h}}{H-h+1} \operatorname{Var}\left(r\left(s_n, a_n\right)+\int_{\mathcal{A}} f\left(\theta_h^*, \phi\left(s_{n+1}, a^{\prime}\right)\right) \pi\left(a^{\prime} \mid s_{n+1}\right) d a^{\prime} \middle| \mathcal{F}_n\right)\right) \\
&\quad \cdot\left(\sum_{h=1}^H \frac{H-h+1}{\sqrt{\nu_h^{\top} \Sigma_h^{-1} \nu_h}}\left(\nu_h^{\top} \Sigma_h^{-1}\left[\nabla_\theta^{\top} f\left(\theta_h^*, \phi\left(s_n, a_n\right)\right)\right]\right)^2\right) \qquad \text{(Cauchy-Schwarz)}
\end{aligned}$$

Since $r\left(s_n, a_n\right)+\int_{\mathcal{A}} f\left(\theta_h^*, \phi\left(s_{n+1}, a^{\prime}\right)\right) \pi\left(a^{\prime} \mid s_{n+1}\right) d a^{\prime} \in[0, H-h+1]$, we have

$$\operatorname{Var}\left(r\left(s_n, a_n\right)+\int_{\mathcal{A}} f\left(\theta_h^*, \phi\left(s_{n+1}, a^{\prime}\right)\right) \pi\left(a^{\prime} \mid s_{n+1}\right) d a^{\prime}\right) \leq \frac{1}{4}(H-h+1)^2,$$

then

$$\operatorname{Var}\left[u_n \mid \mathcal{F}_n\right] \leq \frac{1}{4}\left(\sum_{h=1}^H(H-h+1) \sqrt{\nu_h^{\top} \Sigma_h^{-1} \nu_h}\right) \cdot\left(\sum_{h=1}^H \frac{H-h+1}{\sqrt{\nu_h^{\top} \Sigma_h^{-1} \nu_h}}\left(\nu_h^{\top} \Sigma_h^{-1}\left[\nabla_\theta^{\top} f\left(\theta_h^*, \phi\left(s_n, a_n\right)\right)\right]\right)^2\right),$$

and

$$\begin{aligned}
&\sum_{n=1}^N \operatorname{Var}\left[u_n \mid \mathcal{F}_n\right] \\
\leq& \frac{N}{4}\left(\sum_{h=1}^H(H-h+1) \sqrt{\nu_h^{\top} \Sigma_h^{-1} \nu_h}\right)\left(\sum_{h=1}^H \frac{H-h+1}{\sqrt{\nu_h^{\top} \Sigma_h^{-1} \nu_h}} \nu_h^{\top} \Sigma_h^{-1}\left[\frac{1}{N} \sum_{n=1}^N \nabla_\theta^{\top} f\left(\theta_h^*, \phi\left(s_n, a_n\right)\right) \nabla_\theta f\left(\theta_h^*, \phi\left(s_n, a_n\right)\right)\right] \Sigma_h^{-1} \nu_h\right) \\
\leq& \frac{N}{4}\left(\sum_{h=1}^H(H-h+1) \sqrt{\nu_h^{\top} \Sigma_h^{-1} \nu_h}\right)\left(\sum_{h=1}^H(H-h+1) \sqrt{\nu_h^{\top} \Sigma_h^{-1} \nu_h}\left\|\Sigma_h^{-\frac{1}{2}}\left[\frac{1}{N} \sum_{n=1}^N \nabla_\theta^{\top} f\left(\theta_h^*, \phi\left(s_n, a_n\right)\right) \nabla_\theta f\left(\theta_h^*, \phi\left(s_n, a_n\right)\right)\right] \Sigma_h^{-\frac{1}{2}}\right\|\right)
\end{aligned}$$

where we use $\mu^{\top} \Sigma^{-1} X \Sigma^{-1} \mu \leq \mu^{\top} \Sigma^{-1} \mu \cdot\left\|\Sigma^{-1 / 2} X \Sigma^{-1 / 2}\right\|_2$ for any $\mu \in \mathbb{R}^d, X \in \mathbb{R}^{d \times d}$. We use a special case of lemma C.1 when $h_1=h_2=h$ and we have with probability at least $1-\frac{\delta}{2 H}$,

$$\left\|\Sigma_h^{-\frac{1}{2}}\left[\frac{1}{N} \sum_{n=1}^N \nabla_\theta^{\top} f\left(\theta_h^*, \phi\left(s_n, a_n\right)\right) \nabla_\theta f\left(\theta_h^*, \phi\left(s_n, a_n\right)\right)\right] \Sigma_h^{-\frac{1}{2}}\right\| \leq 1+\sqrt{\frac{2 C_2 d}{K} \log \left(\frac{4 d H}{\delta}\right)}+\frac{2\left(C_2 d+1\right)}{3 K} \log \left(\frac{4 d H}{\delta}\right).$$

By union bound, we have with probability at least $1-\frac{\delta}{2}$, for all $h \in[H]$, the bounds above hold simultaneously and hence

$$\sum_{n=1}^N \operatorname{Var}\left[u_n \mid \mathcal{F}_n\right] \leq \frac{N}{4}\left(\sum_{h=1}^H(H-h+1) \sqrt{\nu_h^{\top} \Sigma_h^{-1} \nu_h}\right)^2\left(1+\sqrt{\frac{2 C_2 d}{K} \log \left(\frac{4 d H}{\delta}\right)}+\frac{2\left(C_2 d+1\right)}{3 K} \log \left(\frac{4 d H}{\delta}\right)\right).$$

We take σ_0^2 equals the right hand side of inequality above, then

$$\mathbb{P}\left(\sum_{n=1}^{N} \operatorname{Var}\left[u_n \mid \mathcal{F}_n\right] \geq \sigma_0^2\right) \leq \frac{\delta}{2}.$$

Freedman's Inequality implies that for any $\varepsilon > 0$, we have

$$\mathbb{P}\left(\left|\sum_{n=1}^{N} u_n\right| \geq \varepsilon, \sum_{n=1}^{N} \operatorname{Var}\left[u_n \mid \mathcal{F}_n\right] \leq \sigma_0^2\right) \leq 2\exp\left(-\frac{\varepsilon^2/2}{\sigma_0^2 + \varepsilon\sqrt{C_2 d}\sum_{h=1}^{H}(H-h+1)\sqrt{\nu_h^\top \Sigma_h^{-1}\nu_h/3}}\right).$$

We take

$$\varepsilon := \sqrt{2\log\left(\frac{4}{\delta}\right)}\sigma_0 + \log\left(\frac{4}{\delta}\right)\frac{2\sqrt{C_2 d}\sum_{h=1}^{H}(H-h+1)\sqrt{\nu_h^\top \Sigma_h^{-1}\nu_h}}{3},$$

then

$$\mathbb{P}\left(\left|\sum_{n=1}^{N} u_n\right| \geq \varepsilon\right) \leq \mathbb{P}\left(\sum_{n=1}^{N} \operatorname{Var}\left[u_n \mid \mathcal{F}_n\right] \geq \sigma_0^2\right) + \mathbb{P}\left(\left|\sum_{n=1}^{N} u_n\right| \geq \varepsilon, \sum_{n=1}^{N} \operatorname{Var}\left[u_n \mid \mathcal{F}_n\right] \leq \sigma_0^2\right) \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta$$

Eventually we use $\sqrt{1+x} \leq 1 + \frac{x}{2}$ for any $x \geq 0$ to get with probability at least $1-\delta$,

$$|I_1| \leq \sqrt{\frac{\log(4/\delta)}{2KH}}\left(\sum_{h=1}^{H}(H-h+1)\sqrt{\nu_h^\top \Sigma_h^{-1}\nu_h}\right) + \Delta I_1,$$

where

$$\Delta I_1 = \left(\sum_{h=1}^{H}(H-h+1)\sqrt{\nu_h^\top \Sigma_h^{-1}\nu_h}\right)\left[\frac{7}{6K}\sqrt{\frac{C_2 d}{H}}\log\left(\frac{4dH}{\delta}\right) + \frac{C_2 d + 1}{3\sqrt{2H}K^{\frac{3}{2}}}\log\left(\frac{4dH}{\delta}\right)^{\frac{3}{2}}\right]$$

We define the function classes

$$\mathcal{G}_h = \left\{\left(\nabla_\theta f\left(\theta_h^*, \phi(s,a)\right)\right) \cdot \mu : \mu \in \mathbb{R}^d\right\} \tag{37}$$

Notice that by Cauchy-Schwarz Inequality, we have

$$\sqrt{\nu_h^\top \Sigma_h^{-1}\nu_h} = \sup_{\mu\in\mathbb{R}^d}\frac{\mu^\top \nu_h}{\sqrt{\mu^\top \Sigma_h \mu}}.$$

For $\mu \in \mathbb{R}^d$, we take $g \in \mathcal{G}_h$ such that $g(s,a) = \left(\nabla_\theta f\left(\theta_h^*, \phi(s,a)\right)\right) \cdot \mu$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. Then

$$\mu^\top \nu_h = \mathbb{E}^\pi\left[g(s_h, a_h) \mid s_1 \sim \xi(\cdot)\right].$$

Additionally, we have

$$\mu^\top \Sigma_h \mu = \mathbb{E}_{\boldsymbol{\tau}}\left\{\frac{1}{H}\sum_{j=1}^{H}\left[\left(\nabla_\theta f\left(\theta_h^*, \phi(s_j, a_j)\right)\right) \cdot \mu\right]^2\right\} = \mathbb{E}_{\boldsymbol{\tau}}\left\{\frac{1}{H}\sum_{h=1}^{H} g^2\left(s_h, a_h\right)\right\}$$

Then,

$$\sqrt{\nu_h^\top \Sigma_h^{-1}\nu_h} = \sup_{\mu\in\mathbb{R}^d}\frac{\mu^\top \nu_h}{\sqrt{\mu^\top \Sigma_h \mu}} = \sup_{g\in\mathcal{G}_h}\frac{\mathbb{E}^\pi\left[g\left(s_h, a_h\right) \mid s_1 \sim \xi(\cdot)\right]}{\sqrt{\mathbb{E}\left[\frac{1}{H}\sum_{h=1}^{H} g^2\left(s_h, a_h\right)\right]}}.$$

We substitute this supremum into upper bound and this concludes the proof for the first order term.

B.2. Higher Order Term

The higheer order term comprises three parts: one part comes from higher order term of $f(\widehat{\theta}_1, \phi) - f(\theta_1^*, \phi)$, the other two parts come from higher order term and Taylor remainder of $\widehat{\theta}_1 - \theta_1^*$. We denote respectively as

$$I_{21} := \int_{\mathcal{S}\times\mathcal{A}} \left[f\left(\widehat{\theta}_1, \phi(s,a)\right) - f\left(\theta_1^*, \phi(s,a)\right) - \nabla_\theta f\left(\theta_1^*, \phi(s,a)\right) \cdot \left(\widehat{\theta}_1 - \theta_1\right) \right] \pi(a \mid s)\xi(s) dads;$$

$$I_{22} := \frac{1}{H} \int_{\mathcal{S}\times\mathcal{A}} \left[\nabla_\theta f\left(\theta_1^*, \phi(s,a)\right) \sum_{h=1}^{H} \left(\prod_{i=1}^{h-1} \Sigma_i^{-1} A_i \right) \Sigma_h^{-1} \left(\mathbb{E}_\tau z_{(h)}\left(\widehat{\theta}_K, \boldsymbol{\tau}\right) + \frac{1}{K} \sum_{k=1}^{K} z_{(h)}\left(\theta^*, \boldsymbol{\tau}_k\right) \right) \right] \pi(a \mid s)\xi(s) dads;$$

$$I_{23} := -\frac{1}{H} \int_{\mathcal{S}\times\mathcal{A}} \left[\nabla_\theta f\left(\theta_1^*, \phi(s,a)\right) \sum_{h=1}^{H} \left(\prod_{i=1}^{h-1} \Sigma_i^{-1} A_i \right) \Sigma_h^{-1} R_{K,h} \right] \pi(a \mid s)\xi(s) dads.$$

We bound them separately. We keep the notation in (19) and use $\alpha = (\alpha_1, \alpha_2, ..., \alpha_d)$ as multi-index. If $x = (x_1, x_2, ..., x_d)$ is a vector of same dimension, we denote $x^\alpha := x_1^{\alpha_1} x_2^{\alpha_2} ... x_d^{\alpha_d}$. Denote $\Delta\theta_1^* := \widehat{\theta}_1 - \theta_1^*$, from Taylor's Theorem, we have

$$f\left(\widehat{\theta}_1, \phi(s,a)\right) - f\left(\theta_1^*, \phi(s,a)\right) - \nabla_\theta f\left(\theta_1^*, \phi(s,a)\right) \cdot \left(\widehat{\theta}_1 - \theta_1^*\right) = \sum_{|\alpha|=2} \frac{\left(\Delta\theta_1^*\right)^\alpha}{2} \partial^\alpha f(\theta_1^* + c\Delta\theta_1^*)$$

for some $c \in (0,1)$. $\sum_{|\alpha|=2}$ denotes summation over all second order derivatives. We recall the definition of $\kappa_l, l = 1,2,3$,

$$\kappa_l := \sup_{\theta\in\Theta, \phi\in\Psi |\alpha|=l} \sup |\partial_\theta^\alpha f(\theta, \phi)| < \infty,$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ is a multi-index with order $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_n$, and $\partial_\theta^\alpha f(\theta, \phi) = \partial_{\theta_1}^{\alpha_1} \partial_{\theta_2}^{\alpha_2} \cdots \partial_{\theta_d}^{\alpha_d} f = \frac{\partial^{|\alpha|} f}{\partial\theta_1^{\alpha_1} \partial\theta_2^{\alpha_2} \cdots \partial\theta_d^{\alpha_d}}$ .s Hence we have

$$|I_{21}| \le \frac{\kappa_2}{2} \int_{\mathcal{S}\times\mathcal{A}} \left| \sum_{|\alpha|=2} \left(\Delta\theta_1^*\right)^\alpha \right| \pi(a \mid s)\xi(s) dads \le \frac{\kappa_2}{2} \left\|\Delta\theta_1^*\right\|_1^2 \le \frac{\kappa_2 d}{2} \left\|\Delta\theta_1^*\right\|_2^2$$

From Corrollary 6.3, we have for every fixed K, with probability at least $1-\delta$, the event $\mathcal{E}_K$ happens, where

$$\mathcal{E}_K := \left\{ \left\| \widehat{\theta}_K - \theta^* \right\|_2 \le \frac{B(\delta)}{\sqrt{K}} \right\}.$$

Under $\mathcal{E}_K$, we have $\|\Delta\theta_1^*\|_2^2 \le \frac{B(\delta)^2}{K}$, hence

$$|I_{21}| \le \frac{\kappa_2 d B(\delta)^2}{2K}.$$

Next we bound I_{22}. From (31) and definition of $\widehat{\theta}_K$, we have

$$I_{22} = \frac{1}{H} \sum_{h=1}^{H} \nu_h^\top \Sigma_h^{-1} \left(\mathbb{E}_{\boldsymbol{\tau}} z_{(h)}\left(\widehat{\theta}_K, \boldsymbol{\tau}\right) - \mathbb{E}_{\boldsymbol{\tau}} z_{(h)}\left(\theta^*, \boldsymbol{\tau}\right) + \frac{1}{K} \sum_{k=1}^{K} z_{(h)}\left(\theta^*, \boldsymbol{\tau}_k\right) - \frac{1}{K} \sum_{k=1}^{K} z_{(h)}\left(\widehat{\theta}_K, \boldsymbol{\tau}_k\right) \right).$$

We use $\mathbb{G}_K := \sqrt{K}\left(\mathbb{P}_K - \mathbb{P}\right)$ to denote empirical measure and we have

$$I_{22} = \frac{1}{H\sqrt{K}} \mathbb{G}_K \left[\sum_{h=1}^{H} \nu_h^\top \Sigma_h^{-1} \left(z_{(h)}\left(\widehat{\theta}_K, \boldsymbol{\tau}\right) - z_{(h)}\left(\theta^*, \boldsymbol{\tau}\right) \right) \right].$$

To bound this empirical process, we use lemma E.4. We define the following function class.

$$\mathcal{M} := \left\{ m(\xi, \cdot) = C_3 \sum_{h=1}^{H} \nu_h^\top \Sigma_h^{-1} \left(z_{(h)}\left(\theta, \boldsymbol{\tau}\right) - z_{(h)}\left(\theta^*, \boldsymbol{\tau}\right) \right) + \frac{1}{2} \middle| \, \|\theta - \theta^*\| \le \frac{B(\delta)}{\sqrt{K}} \right\} \tag{38}$$

Since lemma E.4 requires all function in a certain class take values in $[0,1]$, the constant C_3 and $\frac{1}{2}$ in the expression make each function in $\mathcal{M}$ satisfy this requirement and $\frac{1}{2}$ will not influence bracketing number of this function class. Each function in $\mathcal{M}$ is indexed by $\theta \in \Theta^H$ and is from $(\mathcal{S}\times\mathcal{A})^H$ to $[0,1]$. We denote

$$B_0 := \sup_{h\in[H]} \sqrt{\nu_h^\top \Sigma_h^{-2} \nu_h}. \tag{39}$$

Notice that here the K is fixed, hence without loss of generosity we let $K \geq B(\delta)^2$. Then for every $\theta^1, \theta^2 \in \Theta^H$ such that $\|\theta^i - \theta^*\| \leq \frac{B(\delta)}{\sqrt{K}} \leq 1$, we have

$$\begin{aligned}
&\left|\nu_h^\top \Sigma_h^{-1}\left(z_{(h)}\left(\theta^1, \boldsymbol{\tau}\right) - z_{(h)}\left(\theta^2, \boldsymbol{\tau}\right)\right)\right| \\
\leq &\sum_{j=1}^H \left|\nu_h^\top \Sigma_h^{-1}\left(f(\theta_h^1, \phi_j) - f(\theta_h^2, \phi_j) + \int_{\mathcal{A}} f\left(\theta_{h+1}^2, \phi_{j+1}\right)\pi\left(a', s_{j+1}\right) da' - \int_{\mathcal{A}} f\left(\theta_{h+1}^1, \phi_{j+1}\right)\pi\left(a', s_{j+1}\right) da'\right)\cdot \nabla_\theta^\top f(\theta_h^1, \phi_j)\right| \\
&+ \sum_{j=1}^H \left|\nu_h^\top \Sigma_h^{-1}\left(f(\theta_h^2, \phi_j) - r_j - \int_{\mathcal{A}} f\left(\theta_{h+1}^2, \phi_{j+1}\right)\pi\left(a', s_{j+1}\right) da'\right)\left(\nabla_\theta^\top f(\theta_h^1, \phi_j) - \nabla_\theta^\top f(\theta_h^2, \phi_j)\right)\right| \\
\leq &\sum_{j=1}^H \sqrt{\nu_h^\top \Sigma_h^{-2}\nu_h}\sqrt{d}\kappa_1\left[\left\|\theta_h^1 - \theta_h^2\right\| + \left\|\theta_{h+1}^1 - \theta_{h+1}^2\right\|\right]\left\|\nabla_\theta^\top f(\theta_h^1, \phi_j)\right\| \\
&+ \sum_{j=1}^H \sqrt{\nu_h^\top \Sigma_h^{-2}\nu_h} d\kappa_2 \left\|\theta_h^1 - \theta_h^2\right\| \left|\left(f(\theta_h^2, \phi_j) - r_j - \int_{\mathcal{A}} f\left(\theta_{h+1}^2, \phi_{j+1}\right)\pi\left(a', s_{j+1}\right) da'\right)\right| \\
\leq &H d\kappa_1^2 B_0\left[\left\|\theta_h^1 - \theta_h^2\right\| + \left\|\theta_{h+1}^1 - \theta_{h+1}^2\right\|\right] \\
&+ H d\kappa_2 B_0 \left\|\theta_h^1 - \theta_h^2\right\|\left[\left(f(\theta_h^*, \phi_j) - r_j - \int_{\mathcal{A}} f\left(\theta_{h+1}^*, \phi_{j+1}\right)\pi\left(a', s_{j+1}\right) da'\right) + \sqrt{d}\kappa_1\left[\left\|\theta_h^* - \theta_h^2\right\| + \left\|\theta_{h+1}^* - \theta_{h+1}^2\right\|\right]\right].
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left|\sum_{h=1}^H \nu_h^\top \Sigma_h^{-1}\left(z_{(h)}\left(\theta^1, \boldsymbol{\tau}\right) - z_{(h)}\left(\theta^2, \boldsymbol{\tau}\right)\right)\right| \\
\leq &H^{\frac{3}{2}} d\kappa_1^2 B_0 \left\|\theta^1 - \theta^2\right\| + H d\kappa_2 B_0 \sum_{h=1}^H (H-h+1)\left\|\theta_h^1 - \theta_h^2\right\| + H d^{\frac{3}{2}}\kappa_1\kappa_2 B_0 \sum_{h=1}^H \left\|\theta_h^1 - \theta_h^2\right\|\left[\left\|\theta_h^* - \theta_h^2\right\| + \left\|\theta_{h+1}^* - \theta_{h+1}^2\right\|\right] \\
\leq &\left[H^{\frac{3}{2}} d\kappa_1^2 B_0 + H^{\frac{5}{2}} d\kappa_2 B_0 + 2H d^{\frac{3}{2}}\kappa_1\kappa_2 B_0 \left\|\theta^* - \theta^2\right\|\right]\left\|\theta^1 - \theta^2\right\| \\
\leq &\left[H^{\frac{3}{2}} d\kappa_1^2 B_0 + H^{\frac{5}{2}} d\kappa_2 B_0 + 2H d^{\frac{3}{2}}\kappa_1\kappa_2 B_0\right]\left\|\theta^1 - \theta^2\right\|
\end{aligned}$$

A special case to the bound above is when $\theta^1 = \theta$ and $\theta^2 = \theta^*$, this implies for arbitrary trajectory $\boldsymbol{\tau}$ and θ such that $\|\theta - \theta^*\| \leq \frac{B(\delta)}{\sqrt{K}} \leq 1$, we have

$$\left|\sum_{h=1}^H \nu_h^\top \Sigma_h^{-1}\left(z_{(h)}\left(\theta, \boldsymbol{\tau}\right) - z_{(h)}\left(\theta^*, \boldsymbol{\tau}\right)\right)\right| \leq \left[H^{\frac{3}{2}} d\kappa_1^2 B_0 + H^{\frac{5}{2}} d\kappa_2 B_0 + 2H d^{\frac{3}{2}}\kappa_1\kappa_2 B_0\right]\frac{B(\delta)}{\sqrt{K}}.$$

Therefore, we take

$$C_3 := \left[H^{\frac{3}{2}} d\kappa_1^2 B_0 + H^{\frac{5}{2}} d\kappa_2 B_0 + 2H d^{\frac{3}{2}}\kappa_1\kappa_2 B_0\right]^{-1}\frac{\sqrt{K}}{2B(\delta)},$$

then each function in $\mathcal{M}$ takes value in $[0,1]$. Notice that all functions in $\mathcal{M}$ are Lipschitz continuous, we can give an upper bound for its Lipschitz norm and similar to proof of Theorem 6.1 we can bound the bracketing number. The definition of C_3

and the bound for Lipschitz norm of functions in $\mathcal{M}$ imply that the difference of any two functions in $\mathcal{M}$ can be bounded by

$$\left|m(\theta^1, \boldsymbol{\tau}) - m(\theta^2, \boldsymbol{\tau})\right| \leq \frac{\sqrt{K}}{2B(\delta)} \left\|\theta^1 - \theta^2\right\|.$$

Similar to the calculation in the proof of Theorem 6.1, we have

$$N_{[]}\left(\varepsilon, \mathcal{M}, L_2(\mathbb{P})\right) \leq \left(\frac{2\sqrt{Hd}\frac{\sqrt{K}}{2B(\delta)}\frac{B(\delta)}{\sqrt{K}}}{\varepsilon}\right)^{Hd} = \left(\frac{\sqrt{Hd}}{\varepsilon}\right)^{Hd}$$

We take $V = Hd$ and $U = \sqrt{Hd}$, then using lemma E.4, we have

$$\mathbb{P}\left(\|\mathbb{G}_K\|_{\mathcal{M}} > t\right) \leq \left(\frac{Dt}{\sqrt{Hd}}\right)^{Hd} \exp(-2t^2) \leq \left(\frac{D}{\sqrt{Hd}}\right)^{Hd} \exp(-2t^2 + Hdt).$$

where $D > 0$ is a constant dependent on U only. If we take

$$t = \ln\left(\frac{1}{\delta}\right) + Hd\ln\left(e + \frac{D}{\sqrt{Hd}}\right),$$

then conditional on $\mathcal{E}_K$, with probability at least $1-\delta$, we have

$$\|\mathbb{G}_K\|_{\mathcal{M}} \leq \ln\left(\frac{1}{\delta}\right) + Hd\ln\left(e + \frac{D}{\sqrt{Hd}}\right).$$

Therefore when $K \geq B(\delta)^2$, under $\mathcal{E}_K$, with probability at least $1-\delta$,

$$\begin{aligned}
|I_{22}| &\leq \frac{1}{C_3 H\sqrt{K}} \|\mathbb{G}_K\|_{\mathcal{M}} \\
&\leq \frac{2B(\delta)}{HK}\left[H^{\frac{3}{2}} d\kappa_1^2 B_0 + H^{\frac{5}{2}} d\kappa_2 B_0 + 2Hd^{\frac{3}{2}}\kappa_1\kappa_2 B_0\right]\left[\ln\left(\frac{1}{\delta}\right) + Hd\ln\left(e + \frac{D}{\sqrt{Hd}}\right)\right].
\end{aligned}$$

Finally we give an high-probability upper bound for I_{23}. We have

$$I_{23} = -\frac{1}{H}\sum_{h=1}^{H} \nu_h^\top \Sigma_h^{-1} R_{K,h}.$$

where $R_{K,h} = (R^1_{K,h}, R^2_{K,h}, ...R^d_{K.h})^\top$ is the Taylor remainder of $\mathbb{E}z_{(h)}(\theta, \boldsymbol{\tau})$ at θ^*. Note that $\mathbb{E}z_{(h)}(\theta, \boldsymbol{\tau})$ is only dependent on θ_h, θ_{h+1}. If we denote $\Delta\theta_h := ((\widehat{\theta}_{K,h} - \theta^*_h)^\top, (\widehat{\theta}_{K,h+1} - \theta^*_{h+1})^\top)^\top$, By Taylor Theorem, we have for some $c \in (0,1)$,

$$R^i_{K,h} = \sum_{|\alpha|=2} \frac{\left(\Delta\theta_h\right)^2}{\alpha} \mathbb{E}\left[\partial^\alpha_{\theta_h,\theta_{h+1}} z^i_{(h)}\left(\theta^* + c\left(\widehat{\theta}_K - \theta^*\right), \boldsymbol{\tau}\right)\right].$$

where the notation keeps the same as (19) and $\partial^\alpha_{\theta_h,\theta_{h+1}}$ means the α order derivatives with respect to θ_h, θ_{h+1} only. Remember for $\theta = (\theta^1_1, \theta^2_1, ..., \theta^d_1, \theta^1_2, ..., \theta^d_H)^\top$ and $\theta_h = (\theta^1_h, ..., \theta^d_h)^\top$ we have

$$z^i_{(h)}(\theta, \boldsymbol{\tau}) = \sum_{j=1}^{H}\left(f\left(\theta_h, \phi_j\right) - r_j - \int_{\mathcal{A}} f\left(\theta_{h+1}, \phi\left(s_{j+1}, a'\right)\right)\pi\left(a' \mid s_{j+1}\right) da'\right) \cdot \nabla^i_{\theta_h} f\left(\theta_h, \phi_j\right).$$

We can compute the upper bound for partial derivatives of $z^i_{(h)}$. We have under $\mathcal{E}_K$, for any trajectory $\boldsymbol{\tau}$,

$$\left|\left(f\left(\theta_h, \phi_n\right) - r_n - \int_{\mathcal{A}} f\left(\theta_{h+1}, \phi\left(s_{n+1}, a'\right)\right)\pi\left(a' \mid s_{n+1}\right) da'\right)\right| \leq D_0 := (H-h+1) + 2\sqrt{d}\kappa_1\frac{B(\delta)}{\sqrt{K}}.$$

Therefore we have for $i, \in [d]$,

$$\left|\frac{\partial^2}{\partial\theta_h^i\partial\theta_h^j}z_{(h)}^i(\theta,\boldsymbol{\tau})\right| \le H\left(\kappa_3 D_0 + 3\kappa_1\kappa_2\right), \quad \left|\frac{\partial^2}{\partial\theta_h^i\partial\theta_{h+1}^j}z_{(h)}^i(\theta,\boldsymbol{\tau})\right| \le H\kappa_1\kappa_2, \quad \left|\frac{\partial^2}{\partial\theta_{h+1}^i\partial\theta_{h+1}^j}z_{(h)}^i(\theta,\boldsymbol{\tau})\right| \le H\kappa_1\kappa_2,$$

Therefore, we have

$$\begin{aligned}\left|R_{K,h}^i\right| &\le \frac{H(\kappa_3 D_0 + 3\kappa_1\kappa_2)}{2}\sum_{|\alpha|=2}|(\Delta\theta_h)^\alpha| \le \frac{H(\kappa_3 D_0 + 3\kappa_1\kappa_2)}{2}\left[\left\|\widehat{\theta}_{K,h} - \theta_h^*\right\|_1 + \left\|\widehat{\theta}_{K,h+1} - \theta_{h+1}^*\right\|_1\right]^2 \\ &\le H^2(\kappa_3 D_0 + 3\kappa_1\kappa_2)\left[\left\|\widehat{\theta}_{K,h} - \theta_h^*\right\|_2^2 + \left\|\widehat{\theta}_{K,h+1} - \theta_{h+1}^*\right\|_2^2\right].\end{aligned}$$

Hence under $\mathcal{E}_K$,

$$\|R_{K,h}\|_2 \le \|R_{K,h}\|_1 \le H^2 d(\kappa_3 D_0 + 3\kappa_1\kappa_2)\frac{B(\delta)^2}{K}$$

Therefore, we have

$$|I_{23}| \le B_0 H^3 d\left[\kappa_3 H + 3\kappa_1\kappa_2 + 2\sqrt{d}\kappa_1\kappa_3\frac{B(\delta)}{\sqrt{K}}\right]\frac{B(\delta)^2}{K}.$$

B.3. Completion of the proof

We combine the bound for first order and higher order term and formalize the final upper bound. We only prove for reward-free error bound (10). For variance-aware error bound, the proof is almost the same, except that we use the other bound for first order term. For now, we fix a K such that

$$K \ge \max\left\{\frac{2(C_2 d + 1)^2}{C_2 d}\log\left(\frac{12dH}{\delta}\right), B(\delta/3)^2\right\},$$

then with probability at least $1 - \frac{\delta}{3}$,

$$|I_1| \le \sqrt{\frac{\log(12/\delta)}{2KH}}\left(\sum_{h=1}^H (H-h+1)\sqrt{\nu_h^\top\Sigma_h^{-1}\nu_h}\right) + \left(\sum_{h=1}^H (H-h+1)\sqrt{\nu_h^\top\Sigma_h^{-1}\nu_h}\right)\left[\frac{4}{3K}\sqrt{C_2 dH}\log\left(\frac{12dH}{\delta}\right)\right].$$

At the same time, with probability at least $1 - \frac{\delta}{3}$,

$$\left\|\widehat{\theta}_K - \theta^*\right\|_2 \le \frac{B(\delta/3)}{\sqrt{K}}. \tag{40}$$

Under (40), we have

$$\begin{aligned}|I_{21}| &\le \frac{\kappa_2 d B(\delta/3)^2}{2K}; \\ |I_{23}| &\le B_0 H^3 d\left[\kappa_3 H + 3\kappa_1\kappa_2 + 2\sqrt{d}\kappa_1\kappa_3\frac{B(\delta/3)}{\sqrt{K}}\right]\frac{B(\delta/3)^2}{K}.\end{aligned}$$

and with probability at least $1 - \frac{\delta}{3}$,

$$|I_{22}| \le \frac{2B(\delta)}{HK}\left[H^{\frac{3}{2}}d\kappa_1^2 B_0 + H^{\frac{5}{2}}d\kappa_2 B_0 + 2Hd^{\frac{3}{2}}\kappa_1\kappa_2 B_0\right]\left[\ln\left(\frac{3}{\delta}\right) + Hd\ln\left(e + \frac{D}{\sqrt{Hd}}\right)\right]$$

By union bound, we have when $K \ge B(\delta/3)^2$, with probability at least $1 - \delta$,

$$\begin{aligned}|\widehat{v}_\pi - v_\pi| &\le \sqrt{\frac{\log(12/\delta)}{2KH}}\left(\sum_{h=1}^H (H-h+1)\sqrt{\nu_h^\top\Sigma_h^{-1}\nu_h}\right) + \left(\sum_{h=1}^H (H-h+1)\sqrt{\nu_h^\top\Sigma_h^{-1}\nu_h}\right)\left[\frac{4}{3K}\sqrt{C_2 dH}\log\left(\frac{12dH}{\delta}\right)\right] \\ &+ \frac{B(\delta/3)}{K}\left[H^{2\frac{1}{2}}d\kappa_1^2 B_0 + 2H^{\frac{3}{2}}d\kappa_2 B_0 + 4d^{\frac{3}{2}}\kappa_1\kappa_2 B_0\right] + \frac{B(\delta/3)^2}{K}\left[B_0 H^3 d\left[\kappa_3 H + 3\kappa_1\kappa_2 + 2\sqrt{d}\kappa_1\kappa_3\right]\right],\end{aligned}$$

which implies the upper bound in Theorem 6.5. Here D is a constant dependent only on H and d.

C. Proof and Discussion for Special Cases

C.1. Special Case: With Positivity Condition

We first provide a stronger assumption for data coverage, under which we can get a sharper error bound. Other than assumption 6.4, we further assume for any $(s,a),(s',a') \in \mathcal{S}\times\mathcal{A}$ and $h \in [H]$,

$$\nabla_\theta f(\theta_h^*, \phi(s,a))\Sigma_h^{-1}\nabla_\theta^\top f(\theta_h^*, \phi(s',a')) \geq 0,$$

then with probability at least $1-\delta$,

$$|\widehat{v}_\pi - v_\pi| \leq \left[\sum_{h_1=1}^{H}\sum_{h_2=1}^{H}(H-h_1+1)(H-h_2+1)\right.$$
$$\left.\cdot\sqrt{\nu_{h_1}^\top\Sigma_{h_1}^{-1}\nu_{h_1}}\sqrt{\nu_{h_2}^\top\Sigma_{h_2}^{-1}\nu_{h_2}}\sigma_{h_1,h_2}\right]^{\frac{1}{2}}\sqrt{\frac{\log(12/\delta)}{2HK}} + O(\frac{1}{K}).$$

where $\sigma_{h_1,h_2} := \left\|\Sigma_{h_1}^{-\frac{1}{2}}\Sigma_{h_1,h_2}\Sigma_{h_2}^{-\frac{1}{2}}\right\|_2$ and

$$\Sigma_{h_1,h_2} := \mathbb{E}\left[\frac{1}{H}\sum_{j=1}^{H}\nabla_\theta^\top f\left(\theta_{h_1},\phi_j\right)\nabla_\theta f\left(\theta_{h_2},\phi_j\right)\right].$$

Since $\sigma_{h_1,h_2} \leq 1$, this bound is sharper than (10). When all covariance matrices are zero, all cross terms vanish and this bound is much better than (10).

The only difference brought about by additional positivity condition will be on the bound for the first order term. Since

$$I_1 = -\frac{1}{KH}\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{j=1}^{H}\nu_h^\top\Sigma_h^{-1}\left[\nabla_\theta^\top f\left(\theta_h^*, \phi\left(s_j^k, a_j^k\right)\right)\right]\varepsilon_{h,j}^k,$$

Similarly we denote $\varepsilon_{h,n} = \varepsilon_{h,j}^k$ and $(s_n, a_n) = (s_j^k, a_j^k)$ if $n = (j-1)K + k$, and we define

$$u_n := -\sum_{h=1}^{H}\nu_h^\top\Sigma_h^{-1}\left[\nabla_\theta^\top f\left(\theta_h^*, \phi\left(s_n, a_n\right)\right)\right]\varepsilon_{h,n},$$

then

$$I_1 = \frac{1}{N}\sum_{n=1}^{N}u_n.$$

Define $\mathcal{F}_n$ is the σ field generated by $s_1, a_1, ..., s_n, a_n$, then $\{\mathcal{F}_n\}_{n=1}^N$ is a filtration. Since $\mathbb{E}\left[u_n \mid \mathcal{F}_n\right] = 0$, we have $\{u_n\}_{n=1}^N$ is a martingale difference sequence. We use the following E.2 to analyze I_1. Parallel to the previus analysis, it is easy to see that

$$|u_n| \leq \sqrt{C_2 d}\sum_{h=1}^{H}(H-h+1)\sqrt{\nu_h^\top\Sigma_h^{-1}\nu_h}$$

The main difference with Theorem 6.5 is the estimate of variance $\mathrm{Var}\left[u_n \mid \mathcal{F}_n\right]$. We expand this conditional variance into

$$\mathrm{Var}\left[u_n \mid \mathcal{F}_n\right] = \mathbb{E}\left[\left(\sum_{h=1}^{H}\nu_h^\top\Sigma_h^{-1}\left[\nabla_\theta^\top f\left(\theta_h^*, \phi\left(s_n, a_n\right)\right)\right]\varepsilon_{h,n}\right)^2\middle|\mathcal{F}_n\right]$$
$$= \sum_{h_1=1}^{H}\sum_{h_2=1}^{H}\left(\nu_{h_1}^\top\Sigma_{h_1}^{-1}\nabla_\theta^\top f\left(\theta_{h_1}^*, \phi\left(s_n, a_n\right)\right)\right)\left(\nu_{h_2}^\top\Sigma_{h_2}^{-1}\nabla_\theta^\top f\left(\theta_{h_2}^*, \phi\left(s_n, a_n\right)\right)\right)\mathbb{E}\left[\varepsilon_{h_1,n}\varepsilon_{h_2,n}\middle|\mathcal{F}_n\right].$$

By definition of $\varepsilon_{h,n}$, we have

$$\begin{aligned}\mathbb{E}\left[\varepsilon_{h_1,n}\varepsilon_{h_2,n}\middle|\mathcal{F}_n\right] &= \mathrm{Cov}\left[r\left(s_n,a_n\right)+\int_{\mathcal{A}} f\left(\theta^*_{h_1+1},\phi\left(s_{n+1},a'\right)\right)\pi\left(a'\mid s_{n+1}\right)da',\right.\\ &\qquad \left. r\left(s_n,a_n\right)+\int_{\mathcal{A}} f\left(\theta^*_{h_2+1},\phi\left(s_{n+1},a'\right)\right)\pi\left(a'\mid s_{n+1}\right)da'\middle|\mathcal{F}_n\right]\\ &\le \left[\mathrm{Var}\left(r\left(s_n,a_n\right)+\int_{\mathcal{A}} f\left(\theta^*_{h_1+1},\phi\left(s_{n+1},a'\right)\right)\pi\left(a'\mid s_{n+1}\right)da'\middle|\mathcal{F}_n\right)\right.\\ &\qquad \left.\mathrm{Var}\left(r\left(s_n,a_n\right)+\int_{\mathcal{A}} f\left(\theta^*_{h_2+1},\phi\left(s_{n+1},a'\right)\right)\pi\left(a'\mid s_{n+1}\right)da'\middle|\mathcal{F}_n\right)\right]^{\frac{1}{2}}\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Cauchy-Schwarz)}\\ &\le \frac{1}{4}\left(H-h_1+1\right)\left(H-h_2+1\right).\end{aligned}$$

The last inequality comes from $r\left(s_n,a_n\right)+\int_{\mathcal{A}} f\left(\theta^*_{h+1},\phi\left(s_{n+1},a'\right)\right)\pi\left(a'\mid s_{n+1}\right)da' \in [0,H-h+1]$. Under the condition of $\nabla_\theta f(\theta^*_h,\phi(s,a))\Sigma_h^{-1}\nabla_\theta^\top f(\theta^*_h,\phi(s',a'))\ge 0$ for any $(s,a),(s',a')\in\mathcal{S}\times\mathcal{A}$, we have

$$\begin{aligned}\mathrm{Var}\left[u_n\mid\mathcal{F}_n\right] &\le \frac{1}{4}\sum_{h_1=1}^{H}\sum_{h_2=1}^{H}\left(\nu_{h_1}^\top\Sigma_{h_1}^{-1}\nabla_\theta^\top f\left(\theta^*_{h_1},\phi\left(s_n,a_n\right)\right)\right)\left(\nu_{h_2}^\top\Sigma_{h_2}^{-1}\nabla_\theta^\top f\left(\theta^*_{h_2},\phi\left(s_n,a_n\right)\right)\right)(H-h_1+1)(H-h_2+1)\\ &= \left(\frac{1}{2}\sum_{h=1}^{H}(H-h+1)\nu_h^\top\Sigma_h^{-1}\nabla_\theta^\top f\left(\theta^*_h,\phi\left(s_n,a_n\right)\right)\right)^2.\end{aligned}$$

Therefore,

$$\sum_{n=1}^{N}\mathrm{Var}\left[u_n\mid\mathcal{F}_n\right]\le\frac{1}{4}\sum_{h_1=1}^{H}\sum_{h_2=1}^{H}(H-h_1+1)(H-h_2+1)\nu_{h_1}^\top\Sigma_{h_1}^{-1}\left(\sum_{n=1}^{N}\nabla_\theta^\top f\left(\theta^*_{h_1},\phi\left(s_n,a_n\right)\right)\nabla_\theta f\left(\theta^*_{h_2},\phi\left(s_n,a_n\right)\right)\right)\Sigma_{h_2}^{-1}\nu_{h_2} \tag{41}$$

We use lemma C.1 to bound

$$\left\|\Sigma_{h_1}^{-\frac{1}{2}}\left(\sum_{n=1}^{N}\nabla_\theta^\top f\left(\theta^*_{h_1},\phi\left(s_n,a_n\right)\right)\nabla_\theta f\left(\theta^*_{h_2},\phi\left(s_n,a_n\right)\right)\right)\Sigma_{h_2}^{-\frac{1}{2}}\right\|.$$

Lemma C.1. *Under the assumption that for any $h\in[H]$,*

$$\nabla_\theta f(\theta^*_h,\phi(s,a))\Sigma_h^{-1}\nabla_\theta^\top f(\theta^*_h,\phi(s,a))\le C_2 d,$$

with probability at least $1-\delta$,

$$\begin{aligned}&\left\|\Sigma_{h_1}^{-\frac{1}{2}}\left(\sum_{n=1}^{N}\nabla_\theta^\top f\left(\theta^*_{h_1},\phi\left(s_n,a_n\right)\right)\nabla_\theta f\left(\theta^*_{h_2},\phi\left(s_n,a_n\right)\right)\right)\Sigma_{h_2}^{-\frac{1}{2}}\right\|\\ \le& N\left(\sigma_{h_1,h_2}+\sqrt{\frac{2C_2 d}{K}\log\left(\frac{2d}{\delta}\right)}+\frac{2\left(C_2 d+\sigma_{h_1,h_2}\right)}{3K}\log\left(\frac{2d}{\delta}\right)\right),\end{aligned}$$

where $\sigma_{h_1,h_2}:=\left\|\Sigma_{h_1}^{-\frac{1}{2}}\Sigma_{h_1,h_2}\Sigma_{h_2}^{-\frac{1}{2}}\right\|_2$ and

$$\Sigma_{h_1,h_2}:=\mathbb{E}\left[\frac{1}{H}\sum_{j=1}^{H}\nabla_\theta^\top f\left(\theta^*_{h_1},\phi\left(s_j,a_j\right)\right)\cdot\nabla_\theta f\left(\theta^*_{h_2},\phi\left(s_j,a_j\right)\right)\right].$$

The proof of lemma C.1 is deferred to C.4. By this lemma we have with fixed h_1, h_2 and with probability at least $1 - \frac{\delta}{H^2}$,

$$\begin{aligned}
&\left\| \Sigma_{h_1}^{-\frac{1}{2}} \left(\sum_{n=1}^{N} \nabla_\theta^\top f\left(\theta_{h_1}, \phi\left(s_n, a_n\right)\right) \nabla_\theta f\left(\theta_{h_2}, \phi\left(s_n, a_n\right)\right) \right) \Sigma_{h_2}^{-\frac{1}{2}} \right\| \\
\leq & N \left(\sigma_{h_1,h_2} + \sqrt{\frac{2C_2 d}{K} \log\left(\frac{2dH^2}{\delta}\right)} + \frac{2\left(C_2 d + \sigma_{h_1,h_2}\right)}{3K} \log\left(\frac{2dH^2}{\delta}\right) \right).
\end{aligned} \tag{42}$$

By union bound we have with probability at least $1-\delta$, (42) holds for every $h_1, h_2 \in [H]$. This gives with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{n=1}^{N} \mathrm{Var}\left[u_n \mid \mathcal{F}_n\right] \leq & \frac{1}{4} \sum_{h_1=1}^{H} \sum_{h_2=1}^{H} \left(H - h_1 + 1\right)\left(H - h_2 + 1\right) \\
\leq & \frac{KH}{4} \sum_{h_1=1}^{H} \sum_{h_2=1}^{H} \left(H - h_1 + 1\right)\left(H - h_2 + 1\right) \sqrt{\nu_{h_1}^\top \Sigma_{h_1}^{-1} \nu_{h_1}} \sqrt{\nu_{h_2}^\top \Sigma_{h_2}^{-1} \nu_{h_2}} \\
& \cdot \left(\sigma_{h_1,h_2} + \sqrt{\frac{2C_2 d}{K} \log\left(\frac{2dH^2}{\delta}\right)} + \frac{2\left(C_2 d + \sigma_{h_1,h_2}\right)}{3K} \log\left(\frac{2dH^2}{\delta}\right) \right).
\end{aligned}$$

We take

$$\begin{aligned}
\sigma_0^2 := & \frac{KH}{4} \sum_{h_1=1}^{H} \sum_{h_2=1}^{H} \left(H - h_1 + 1\right)\left(H - h_2 + 1\right) \sqrt{\nu_{h_1}^\top \Sigma_{h_1}^{-1} \nu_{h_1}} \sqrt{\nu_{h_2}^\top \Sigma_{h_2}^{-1} \nu_{h_2}} \\
& \cdot \left(\sigma_{h_1,h_2} + \sqrt{\frac{2C_2 d}{K} \log\left(\frac{4dH^2}{\delta}\right)} + \frac{2\left(C_2 d + \sigma_{h_1,h_2}\right)}{3K} \log\left(\frac{4dH^2}{\delta}\right) \right),
\end{aligned}$$

then

$$\mathbb{P}\left(\sum_{n=1}^{N} \mathrm{Var}\left[u_n \mid \mathcal{F}_n\right] \geq \sigma_0^2 \right) \leq \frac{\delta}{2}.$$

Freedman's Inequality implies that for any $\varepsilon > 0$, we have

$$\mathbb{P}\left(\left| \sum_{n=1}^{N} u_n \right| \geq \varepsilon, \sum_{n=1}^{N} \mathrm{Var}\left[u_n \mid \mathcal{F}_n\right] \leq \sigma_0^2 \right) \leq 2 \exp\left(- \frac{\varepsilon^2/2}{\sigma_0^2 + \varepsilon \sqrt{C_2 d} \sum_{h=1}^{H} (H - h + 1) \sqrt{\nu_h^\top \Sigma_h^{-1} \nu_h} / 3} \right).$$

We take

$$\varepsilon := \sqrt{2 \log\left(\frac{4}{\delta}\right)} \sigma_0 + \log\left(\frac{4}{\delta}\right) \frac{2\sqrt{C_2 d} \sum_{h=1}^{H} (H - h + 1) \sqrt{\nu_h^\top \Sigma_h^{-1} \nu_h}}{3},$$

then

$$\mathbb{P}\left(\left| \sum_{n=1}^{N} u_n \right| \geq \varepsilon \right) \leq \mathbb{P}\left(\sum_{n=1}^{N} \mathrm{Var}\left[u_n \mid \mathcal{F}_n\right] \geq \sigma_0^2 \right) + \mathbb{P}\left(\left| \sum_{n=1}^{N} u_n \right| \geq \varepsilon, \sum_{n=1}^{N} \mathrm{Var}\left[u_n \mid \mathcal{F}_n\right] \leq \sigma_0^2 \right) \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta.$$

Notice that

$$\frac{\varepsilon}{N} = \sqrt{\frac{\log(4/\delta)}{2HK}} \cdot \sqrt{\sum_{h_1=1}^{H} \sum_{h_2=1}^{H} \left(H - h_1 + 1\right)\left(H - h_2 + 1\right) \sqrt{\nu_{h_1}^\top \Sigma_{h_1}^{-1} \nu_{h_1}} \sqrt{\nu_{h_2}^\top \Sigma_{h_2}^{-1} \nu_{h_2}} \sigma_{h_1,h_2}} + O\left(\frac{1}{K}\right),$$

we insert this into the first order term and hence finish the proof.

C.2. Special Case: Linear Function Class

Before we provide the result in linear function case, we briefly describe the simplest case with linear function approximation $f(\theta,\phi) := \theta^\top \phi(s,a)$ (Hao et al., 2021b). Assumption 5.3 for policy completeness here is equivalent to that there exists $M \in \mathbb{R}^{d\times d}$ such that

$$\phi(s,a)^\top M = \mathbb{E}\left[\left(\int_{\mathcal{A}} \phi(s',a')\pi(a' \mid s')\right)^\top \Big| s,a\right].$$

This implies that for $h \in [H]$, we have $\theta_h^* = M\theta_{h+1}^* + \theta_r^*$ and $\theta_H^* = \theta_r^*$ for some $\theta_r^* \in \Theta$. θ_r^* is the parameter of reward function $r(s,a)$. In linear setting, (2) has following explicit expression.

$$\widehat{\theta}_h = \widehat{M}\widehat{\theta}_{h+1} + \widehat{R}. \tag{43}$$

where

$$\widehat{M} = \widehat{\Sigma}^{-1}\sum_{n=1}^{N} \phi(s_n,a_n)\left(\int_{\mathcal{A}} \phi(s_{n+1},a')\pi(a' \mid s_{n+1})\right)^\top, \widehat{R} = \widehat{\Sigma}^{-1}\sum_{n=1}^{N} r_n\phi(s_n,a_n), \widehat{\Sigma} = \sum_{n=1}^{N} \phi(s_n,a_n)\phi(s_n,a_n)^\top + \lambda I_d$$

(Hao et al., 2021b) showed that, the limit distribution of $\sqrt{N}(\widehat{v}_\pi - v_\pi)$ is $N(0,\sigma^2)$, where

$$\sigma^2 = \sum_{h_1=1}^{H}\sum_{h_2=1}^{H} (\nu_{h_1})^\top \Sigma^{-1}\Omega_{h_1,h_2}\Sigma^{-1}\nu_{h_2}, \tag{44}$$

where

$$\nu_h = \mathbb{E}^\pi[\phi(s_h,a_h) \mid s_1 \sim \xi(\cdot)]; \quad \Sigma = \mathbb{E}\left[\frac{1}{H}\sum_{h=1}^{H} \phi(s_h,a_h)\phi(s_h,a_h)^\top\right];$$

$$\Omega_{h_1,h_2} = \mathbb{E}\left[\frac{1}{H}\sum_{h'=1}^{H} \phi(s_{h'},a_{h'})\phi(s_{h'},a_{h'})^\top \varepsilon_{h_1,h'}\varepsilon_{h_2,h'}\right]; \quad \varepsilon_{h_1,h'} = Q_{h_1}(s_{h'},a_{h'}) - (r_{h'} + V_{h_1+1}(s_{h'+1})). \tag{45}$$

Further, they showed that FQE with linear function approximation achieves Cramer Rao lower bound for variance, and bootstrapping error with linear function approximation has the same limit distribution as standard FQE error. Actually, this expression can be immediately derived using our results. In linear case, we can provide a sharper bound for the dominant term in $\widehat{v}_\pi - v_\pi$, which comes from a different variance calculation in Freeman's Inequality. Here, our data coverage assumption will be $\phi^\top(s,a)\Sigma^{-1}\phi(s,a) \le C_2 d$ and the positivity condition will be $\phi^\top(s,a)\Sigma^{-1}\phi(s',a') \ge 0$. Similar to the derivation before, we have $I_1 = \frac{1}{N}\sum_{n=1}^{N} u_n$, where $u_n := -\sum_{h=1}^{H} \nu_h^\top \Sigma^{-1}[\phi(s_n,a_n)]\varepsilon_{h,n}$. Define $\mathcal{F}_n$ is the σ field generated by $s_1,a_1,\ldots,s_n,a_n$, then $\{u_n\}_{n=1}^N$ is a martingale difference sequence and $|u_n| \le \sqrt{C_2 d}\sum_{h=1}^{H}(H-h+1)\sqrt{\nu_h^\top \Sigma_h^{-1}\nu_h}$. In this case, (41) becomes

$$\begin{aligned}
\sum_{n=1}^{N} \operatorname{Var}[u_n \mid \mathcal{F}_n] &\le \frac{1}{4}\sum_{h_1=1}^{H}\sum_{h_2=1}^{H}(H-h_1+1)(H-h_2+1)\nu_{h_1}^\top\Sigma^{-1}\left(\sum_{n=1}^{N}\phi(s_n,a_n)\phi^\top(s_n,a_n)\right)\Sigma^{-1}\nu_{h_2} \\
&\le \frac{1}{4}\left(\sum_{h=1}^{H}(H-h+1)\nu_h\right)^\top \Sigma^{-1}\left(\sum_{n=1}^{N}\phi(s_n,a_n)\phi(s_n,a_n)^\top\right)\Sigma^{-1}\left(\sum_{h=1}^{H}(H-h+1)\nu_h\right) \\
&\le \frac{1}{4}\left(\sum_{h=1}^{H}(H-h+1)\nu_h\right)^\top \Sigma^{-1}\left(\sum_{h=1}^{H}(H-h+1)\nu_h\right)\left\|\Sigma^{-\frac{1}{2}}\left(\sum_{n=1}^{N}\phi(s_n,a_n)\phi(s_n,a_n)^\top\right)\Sigma^{-\frac{1}{2}}\right\|.
\end{aligned}$$

When we take f as linear, lemma C.1 shows that with probability at least $1-\delta$,

$$\left\|\Sigma^{-\frac{1}{2}}\left(\sum_{n=1}^{N}\phi(s_n,a_n)\phi(s_n,a_n)^\top\right)\Sigma^{-\frac{1}{2}}\right\| \le N\left(1 + \sqrt{\frac{2C_2 d}{K}\log\left(\frac{2d}{\delta}\right)} + \frac{2(C_2 d+1)}{3K}\log\left(\frac{2d}{\delta}\right)\right),$$

We take

$$\sigma^2 := \frac{N}{4}\left(\sum_{h=1}^{H}(H-h+1)\nu_h\right)^\top \Sigma^{-1}\left(\sum_{h=1}^{H}(H-h+1)\nu_h\right)\left(1+\sqrt{\frac{2C_2d}{K}\log\left(\frac{2d}{\delta}\right)}+\frac{2\left(C_2d+1\right)}{3K}\log\left(\frac{2d}{\delta}\right)\right),$$

and

$$\varepsilon := \sqrt{2\log\left(\frac{4}{\delta}\right)}\sigma+\log\left(\frac{4}{\delta}\right)\frac{2\sqrt{C_2d}\sum_{h=1}^{H}(H-h+1)\sqrt{\nu_h^\top\Sigma_h^{-1}\nu_h}}{3}.$$

By Freeman's Inequality, for any $\varepsilon>0$,

$$\mathbb{P}\left(\left|\sum_{n=1}^{N}u_n\right|\geq\varepsilon,\sum_{n=1}^{N}\mathrm{Var}\left[u_n\mid\mathcal{F}_n\right]\leq\sigma^2\right)\leq 2\exp\left(-\frac{\varepsilon^2/2}{\sigma^2+\varepsilon\sqrt{C_2d}\sum_{h=1}^{H}(H-h+1)\sqrt{\nu_h^\top\Sigma_h^{-1}\nu_h/3}}\right),$$

then

$$\mathbb{P}\left(\left|\sum_{n=1}^{N}u_n\right|\geq\varepsilon\right)\leq\mathbb{P}\left(\sum_{n=1}^{N}\mathrm{Var}\left[u_n\mid\mathcal{F}_n\right]\geq\sigma^2\right)+\mathbb{P}\left(\left|\sum_{n=1}^{N}u_n\right|\geq\varepsilon,\sum_{n=1}^{N}\mathrm{Var}\left[u_n\mid\mathcal{F}_n\right]\leq\sigma^2\right)\leq\frac{\delta}{2}+\frac{\delta}{2}=\delta.$$

Hence the upper bound for the first order term can be improved to

$$|I_1|\leq\sqrt{\left(\sum_{h=1}^{H}(H-h+1)\nu_h\right)^\top\Sigma^{-1}\left(\sum_{h=1}^{H}(H-h+1)\nu_h\right)}\cdot\sqrt{\frac{\ln(4/\delta)}{2KH}}+O(\frac{1}{K}).$$

By Cauchy-Schwarz Inequality, we have

$$\sqrt{\left(\sum_{h=1}^{H}(H-h+1)\nu_h\right)^\top\Sigma^{-1}\left(\sum_{h=1}^{H}(H-h+1)\nu_h\right)}=\sup_{\mu\in\mathbb{R}^d}\frac{\mu^\top\left(\sum_{h=1}^{H}(H-h+1)\nu_h\right)}{\sqrt{\mu^\top\Sigma\mu}}.$$

For $\mu\in\mathbb{R}^d$, we take $g\in\mathcal{G}$ such that $g(s,a)=\phi^\top(s,a)\mu$ for any $(s,a)\in\mathcal{S}\times\mathcal{A}$. Then $\mu^\top\left(\sum_{h=1}^{H}(H-h+1)\nu_h\right)=\mathbb{E}^\pi\left[(H-h+1)g(s_h,a_h)\mid s_1\sim\xi(\cdot)\right]$. Additionally, we have

$$\mu^\top\Sigma\mu=\mathbb{E}_{\boldsymbol{\tau}}\left\{\frac{1}{H}\sum_{j=1}^{H}\left[\phi^\top(s_j,a_j)\cdot\mu\right]^2\right\}=\mathbb{E}_{\boldsymbol{\tau}}\left\{\frac{1}{H}\sum_{h=1}^{H}g^2\left(s_h,a_h\right)\right\}$$

Then,

$$\begin{aligned}\sqrt{\left(\sum_{h=1}^{H}(H-h+1)\nu_h\right)^\top\Sigma^{-1}\left(\sum_{h=1}^{H}(H-h+1)\nu_h\right)}&=\sup_{g\in\mathcal{G}}\frac{\mathbb{E}^\pi\left[(H-h+1)g\left(s_h,a_h\right)\mid s_1\sim\xi(\cdot)\right]}{\sqrt{\mathbb{E}\left[\frac{1}{H}\sum_{h=1}^{H}g^2\left(s_h,a_h\right)\right]}}\\&=\frac{H(H+1)}{2}\sqrt{1+\chi^2_{\mathcal{G}}(\tilde{\mu},\bar{\mu})}.\end{aligned}$$

We replace δ with $\delta/3$, bound the higher order term as section B.2, and substitute this supremum into upper bound. This implies the result in linear function case.

C.3. Special Case: Tabular Case

This part is proven in (Duan et al., 2020), but for completeness, we briefly sketch the outline. In tabular case, the MDP we consider has finite state space and action space. We denote their cardinality as $|\mathcal{S}|$ and $|\mathcal{A}|$. Intrinsically, we can always parametrize tabular MDP in a linear way, hence tabular case is a special case of linear case. Our feature map is an $|\mathcal{S}| \times |\mathcal{A}|$- dimensional indicator vector $\phi(s,a) = \mathbf{1}_{s,a}$, where the $(s,a)-$th entry is 1 and other entries are 0. Covariance martrix Σ is diagonal with non-negative entries, hence positivity condition holds naturally. By the definition of ν_h and Σ, we have $\nu_h((s,a)) = \mathbb{P}^\pi\left[s_h = s, a_h = a \mid s_1 \sim \xi(\cdot)\right]$ and $\Sigma((s,a),(s,a)) = \bar{\mu}(s,a)$. Further we have $\sum_{h=1}^H (H-h+1)\nu_h((s,a)) = \frac{1}{2}H(H+1)\tilde{\mu}(s,a)$, where $\tilde{\mu}$ is defined as (11). Therefore, the coefficient in the dominant term will become

$$\sqrt{\left(\sum_{h=1}^H (H-h+1)\nu_h\right)^\top \Sigma^{-1}\left(\sum_{h=1}^H (H-h+1)\nu_h\right)} = \sum_{h=1}^H (H-h+1)\sqrt{\sum_{s\in\mathcal{S},a\in\mathcal{A}} \frac{\tilde{\mu}((s,a))^2}{\bar{\mu}(s,a)}} = \frac{H(H+1)}{2}\sqrt{1+\chi^2\left(\tilde{\mu},\bar{\mu}\right)}.$$

This matches the result optimal result in (Yin & Wang, 2020).

C.4. Proof of Lemma C.1

We fix a pair of h_1, h_2 and we denote

$$\Sigma_{h_1,h_2} := \mathbb{E}\left[\frac{1}{H}\sum_{j=1}^H \nabla_\theta^\top f\left(\theta_{h_1}^*, \phi\left(s_j,a_j\right)\right)\cdot \nabla_\theta f\left(\theta_{h_2}^*, \phi\left(s_j,a_j\right)\right)\right],$$

and

$$X_k := \frac{1}{H}\sum_{h=1}^H \Sigma_{h_1}^{-\frac{1}{2}} \nabla_\theta^\top f\left(\theta_{h_1}^*, \phi\left(s_j^k,a_j^k\right)\right)\nabla_\theta f\left(\theta_{h_2}^*, \phi\left(s_j^k,a_j^k\right)\right)\Sigma_{h_2}^{-\frac{1}{2}},$$

then $X_1, X_2, .., X_K$ are independent with $\mathbb{E}\left[X_k\right] = \Sigma_{h_1}^{-\frac{1}{2}}\Sigma_{h_1,h_2}\Sigma_{h_2}^{-\frac{1}{2}}$. Below we use matrix E.3 to bound $\frac{1}{K}\sum_{k=1}^K X_k$. If we denote

$$\Phi_{h,k} := \left(\nabla_\theta^\top f\left(\theta_h^*, \phi\left(s_1^k,a_1^k\right)\right), \nabla_\theta^\top f\left(\theta_h^*, \phi\left(s_2^k,a_2^k\right)\right), ..., \nabla_\theta^\top f\left(\theta_h^*, \phi\left(s_H^k,a_H^k\right)\right)\right) \in \mathbb{R}^{d\times H},$$

then $X_k = \frac{1}{H}\Sigma_{h_1}^{-\frac{1}{2}}\Phi_{h_1,k}\Phi_{h_2,k}^\top\Sigma_{h_2}^{-\frac{1}{2}}$. For any vector $\mu\in\mathbb{R}^d$, we have

$$\begin{aligned}\mu^\top\mathbb{E}\left[X_k^2\right]\mu = \mathbb{E}\left[\left\|X_k\mu\right\|_2^2\right] &= \frac{1}{H^2}\mathbb{E}\left[\left\|\Sigma_{h_1}^{-\frac{1}{2}}\Phi_{h_1,k}\Phi_{h_2,k}^\top\Sigma_{h_2}^{-\frac{1}{2}}\mu\right\|_2^2\right]\\ &\le \frac{1}{H^2}\mathbb{E}\left[\left\|\Sigma_{h_1}^{-\frac{1}{2}}\Phi_{h_1,k}\right\|_2^2\left\|\Phi_{h_2,k}^\top\Sigma_{h_2}^{-\frac{1}{2}}\mu\right\|_2^2\right]\end{aligned}$$

From assumption 6.4, we have $\left|\left(\Phi_{h_1,k}^\top\Sigma_{h_1}^{-1}\Phi_{h_1,k}\right)_{ij}\right| \le C_2 d$ for any $i,j\in[H]$, we have

$$\left\|\Sigma_{h_1}^{-\frac{1}{2}}\Phi_{h_1,k}\right\|_2^2 = \left\|\Phi_{h_1,k}^\top\Sigma_{h_1}^{-1}\Phi_{h_1,k}\right\|_2 \le \left\|\Phi_{h_1,k}^\top\Sigma_{h_1}^{-1}\Phi_{h_1,k}\right\|_{\mathrm{F}} \le C_2 dH.$$

Therefore,

$$\mu^\top\mathbb{E}\left[X_k^2\right]\mu \le \frac{C_2 d}{H}\mathbb{E}\left[\left\|\Phi_{h_2,k}^\top\Sigma_{h_2}^{-\frac{1}{2}}\mu\right\|_2^2\right] = \frac{C_2 d}{H}\mu^\top\mathbb{E}\left[\Sigma_{h_2}^{-\frac{1}{2}}\Phi_{h_2,k}\Phi_{h_2,k}^\top\Sigma_{h_2}^{-\frac{1}{2}}\right]\mu = C_2 d\left\|\mu\right\|_2^2.$$

because $\frac{1}{H}\mathbb{E}\left[\Sigma_{h_2}^{-\frac{1}{2}}\Phi_{h_2,k}^\top\Phi_{h_2,k}\Sigma_{h_2}^{-\frac{1}{2}}\right] = I_d$ by definition. Therefore,

$$\mathrm{Var}\left[X_k\right] \preceq \mathbb{E}\left[X_k^2\right] \preceq C_2 d I_d.$$

and if we denote $\sigma_{h_1,h_2} := \left\|\Sigma_{h_1}^{-\frac{1}{2}}\Sigma_{h_1,h_2}\Sigma_{h_2}^{-\frac{1}{2}}\right\|_2$, then

$$\left\|X_k - \Sigma_{h_1}^{-\frac{1}{2}}\Sigma_{h_1,h_2}\Sigma_{h_2}^{-\frac{1}{2}}\right\|_2 \le \left\|X_k\right\|_2 + \left\|\Sigma_{h_1}^{-\frac{1}{2}}\Sigma_{h_1,h_2}\Sigma_{h_2}^{-\frac{1}{2}}\right\|_2 \le \frac{1}{H}\left\|\Sigma_{h_1}^{-\frac{1}{2}}\Phi_{h_1,k}\right\|_2^2\left\|\Sigma_{h_2}^{-\frac{1}{2}}\Phi_{h_2,k}\right\|_2^2 + \sigma_{h_1,h_2} \le C_2 d + \sigma_{h_1,h_2}.$$

Therefore, by matrix Bernstein inequality, we have for $\varepsilon > 0$,

$$\mathbb{P}\left(\left\|\sum_{k=1}^{K}\left(X_k - \Sigma_{h_1}^{-\frac{1}{2}}\Sigma_{h_1,h_2}\Sigma_{h_2}^{-\frac{1}{2}}\right)\right\|_2 \geq \varepsilon\right) \leq 2d\exp\left(-\frac{\varepsilon^2/2}{C_2dK + (C_2d + \sigma_{h_1,h_2})\,\varepsilon/3}\right).$$

Therefore, with probability at least $1-\delta$,

$$\left\|\frac{1}{K}\sum_{k=1}^{K}\left(X_k - \Sigma_{h_1}^{-\frac{1}{2}}\Sigma_{h_1,h_2}\Sigma_{h_2}^{-\frac{1}{2}}\right)\right\|_2 \leq \sqrt{\frac{2C_2d}{K}\log\left(\frac{2d}{\delta}\right)} + \frac{2\,(C_2d + \sigma_{h_1,h_2})}{3K}\log\left(\frac{2d}{\delta}\right).$$

D. Proof of Information-Theoretic Lower Bound

We first compute the influence function of v_π and show that the expectation of squared influence function meets the variance term in (7). We denote

$$p_\eta(s' \mid s,a) = p(s' \mid s,a) + \eta\Delta p(s',s,a).$$

where Δp is arbitrary probability shift. When $\eta = 0$, this notation is same as our original transition probability $p_0(s' \mid s,a) = p(s' \mid s,a)$. Suppose Δp satisfies $(\Delta p)\mathcal{F} \subset \mathcal{F}$. We denote $Q_{h,\eta}(s,a)$ and $V_{h,\eta}(s)$ as the Q function and state value function with transition probability being $p_\eta(s' \mid s,a)$, and $Q_{h,0}(s,a) = Q_h(s,a), V_{h,0}(s) = V_h(s)$. θ_h^* will be dependent on η, hence we write $\theta_{h,\eta}^*$ to explicate this dependency. When $\eta = 0$, we let $\theta_{h,0}^* = \theta_h^*$. We emphasize again that $\mathbb{E}^\pi$ denotes expectation over population generated by target policy, and $\mathbb{E}$ or $\mathbb{E}_{\boldsymbol{\tau}}$ denotes that by behavior policy. Let $\bar{\mu}$ be the occupancy measure of state action pair generated by behavior policy. We define the score function as

$$l_\eta(s' \mid s,a) := \frac{\partial}{\partial\eta}\log p_\eta(s' \mid s,a) \;\text{ and }\; l_\eta(\boldsymbol{\tau}) := \sum_{h=1}^{H} l_\eta(s_{h+1} \mid s_h,a_h).$$

When η vanishes, we have $l(s' \mid s,a) := \frac{\partial}{\partial\eta}\log p_\eta(s' \mid s,a)\Big|_{\eta=0}$ and $l(\boldsymbol{\tau}) := \sum_{h=1}^{H} l(s_{h+1} \mid s_h,a_h)$. Our objective function is

$$v_{\pi,\eta} := \mathbb{E}^\pi\left[\sum_{h=1}^{H} r(s_h,a_h)\middle| s_1 \sim \xi(\cdot), p_\eta\right].$$

We take its derivatives and then let $\eta = 0$ to compute the influence function.

$$\begin{aligned}
\frac{\partial}{\partial\eta}v_{\pi,\eta}\Big|_{\eta=0} &= \frac{\partial}{\partial\eta}\left[\sum_{h=1}^{H}\int_{(\mathcal{S}\times\mathcal{A})^h} r(s_h,a_h)\xi(s_1)\prod_{h'=1}^{h-1}p_\eta(s_{h'+1} \mid s_{h'},a_{h'})\prod_{h'=1}^{h}\pi(a_{h'} \mid s_{h'})d\boldsymbol{\tau}_h\right]\Bigg|_{\eta=0} \\
&= \sum_{h=1}^{H}\int_{(\mathcal{S}\times\mathcal{A})^h} r(s_h,a_h)\xi(s_1)\sum_{h'=1}^{h-1}l_\eta(s_{h'+1} \mid s_{h'},a_{h'})\prod_{h'=1}^{h-1}p_\eta(s_{h'+1} \mid s_{h'},a_{h'})\prod_{h'=1}^{h}\pi(a_{h'} \mid s_{h'})d\boldsymbol{\tau}_h\Bigg|_{\eta=0} \\
&= \int_{(\mathcal{S}\times\mathcal{A})^H}\left[\sum_{h=1}^{H-1}l_\eta(s_{h+1} \mid s_h,a_h)\sum_{h'=h+1}^{H}r(s_{h'},a_{h'})\right]\xi(s_1)\prod_{j=1}^{H}p_\eta(s_{j+1} \mid s_j,a_j)\pi(a_j \mid s_j)d\boldsymbol{\tau}\Bigg|_{\eta=0} \\
&= \mathbb{E}^\pi\left[\sum_{h=1}^{H-1}l_\eta(s_{h+1} \mid s_h,a_h)\sum_{h'=h+1}^{H}r(s_{h'},a_{h'})\middle| s_1 \sim \xi, p_\eta\right]\Bigg|_{\eta=0} \\
&= \mathbb{E}^\pi\left[\sum_{h=1}^{H}\mathbb{E}^\pi\Big[\,l_\eta(s_{h+1} \mid s_h,a_h)V_{h+1,\eta}(s_{h+1})\big| s_h,a_h\Big]\middle| s_1 \sim \xi, p_\eta\right]\Bigg|_{\eta=0}.
\end{aligned}$$

Since

$$\mathbb{E}^\pi\left[l_\eta(s' \mid s,a)V_{h+1,\eta}(s') \mid s,a\right] = \int_{\mathcal{S}\times\mathcal{A}}\frac{\partial}{\partial\eta}f(\theta_{h+1,\eta}^*,\phi(s,a))p_\eta(s' \mid s,a)\pi(a' \mid s')ds'da' = \frac{\partial}{\partial\eta}f(\theta_{h,\eta}^*,\phi(s,a)),$$

we have

$$\begin{aligned}
\frac{\partial}{\partial \eta} v_{\pi,\eta}\Big|_{\eta=0} &= \mathbb{E}^{\pi}\left[\sum_{h=1}^{H} \frac{\partial}{\partial \eta} f(\theta^*_{h,\eta}, \phi(s_h, a_h))\Big|_{\eta=0} \Big| s_1 \sim \xi, p\right] \\
&= \mathbb{E}^{\pi}\left[\sum_{h=1}^{H} \left(\nabla_{\theta_h} f(\theta^*_h, \phi(s_h, a_h))\right) \Sigma_h^{-1} \Sigma_h \frac{\partial \theta_{h,\eta}}{\partial \eta}\Big|_{\eta=0} \Big| s_1 \sim \xi, p\right].
\end{aligned}$$

Notice that

$$\begin{aligned}
\Sigma_h &= \frac{1}{H} \mathbb{E}_\tau \left\{\sum_{h'=1}^{H} \left(\nabla_{\theta_h} f\left(\theta^*_h, \phi(s_{h'}, a_{h'})\right)\right)^\top \left(\nabla_{\theta_h} f\left(\theta^*_h, \phi(s_{h'}, a_{h'})\right)\right)\right\} \\
&= \mathbb{E}_{(s,a)\sim\bar{\mu}}\left[\left(\nabla_{\theta_h} f\left(\theta^*_h, \phi(s,a)\right)\right)^\top \left(\nabla_{\theta_h} f\left(\theta^*_h, \phi(s,a)\right)\right)\right],
\end{aligned}$$

then

$$\begin{aligned}
&\frac{\partial}{\partial \eta} v_{\pi,\eta}\Big|_{\eta=0} \\
=&\mathbb{E}^{\pi}\left[\sum_{h=1}^{H} \left(\nabla_{\theta_h} f(\theta^*_h, \phi(s_h, a_h))\right) \Sigma_h^{-1} \mathbb{E}_{(s,a)\sim\bar{\mu}}\left[\left(\nabla_{\theta_h} f\left(\theta^*_h, \phi(s,a)\right)\right)^\top \left(\nabla_{\theta_h} f\left(\theta^*_h, \phi(s,a)\right)\right) \frac{\partial \theta_{h,\eta}}{\partial \eta}\Big|_{\eta=0}\right] \Big| s_1 \sim \xi, p\right] \\
=&\mathbb{E}^{\pi}\left[\sum_{h=1}^{H} \left(\nabla_{\theta_h} f(\theta^*_h, \phi(s_h, a_h))\right) \Sigma_h^{-1} \mathbb{E}_{(s,a)\sim\bar{\mu}}\left[\left(\nabla_{\theta_h} f\left(\theta^*_h, \phi(s,a)\right)\right)^\top \frac{\partial}{\partial \eta} f(\theta^*_{h,\eta}, \phi(s,a))\Big|_{\eta=0}\right] \Big| s_1 \sim \xi, p\right] \\
=&\mathbb{E}_{(s,a)\sim\bar{\mu}}\left\{\sum_{h=1}^{H} \mathbb{E}^{\pi}\left[\nabla_{\theta_h} f(\theta^*_h, \phi(s_h, a_h)) \Big| s_1 \sim \xi, p\right] \Sigma_h^{-1} \left(\nabla_{\theta_h} f\left(\theta^*_h, \phi(s,a)\right)\right)^\top l(s' \mid s,a) V_{h+1}(s')\right\}.
\end{aligned}$$

We define

$$w_h(s,a) := \mathbb{E}^{\pi}\left[\nabla_{\theta_h} f\left(\theta^*_h, \phi(s_h, a_h)\right) \mid s_1 \sim \xi, p\right] \Sigma_h^{-1} \left(\nabla_{\theta_h} f\left(\theta^*_h, \phi(s,a)\right)\right)^\top \in \mathbb{R},$$

and use $\mathbb{E}\left[l(s'|s,a)|s,a\right] = 0$ to get

$$\begin{aligned}
\frac{\partial}{\partial \eta} v_{\pi,\eta}\Big|_{\eta=0} &= \mathbb{E}_{(s,a)\sim\bar{\mu}}\left\{\sum_{h=1}^{H} w_h(s,a) l(s' \mid s,a) V_{h+1}(s')\right\} \\
&= \mathbb{E}_{(s,a)\sim\bar{\mu}}\left\{\sum_{h=1}^{H} w_h(s,a) l(s' \mid s,a) \Big[V_{h+1}(s') - \mathbb{E}\left[V_{h+1}(s')|s,a\right]\Big]\right\} \\
&= \frac{1}{H} \mathbb{E}\left\{\sum_{j=1}^{H} \sum_{h=1}^{H} w_h(s_j, a_j) l(s_{j+1} \mid s_j, a_j) \Big[V_{h+1}(s_{j+1}) - \mathbb{E}\left[V_{h+1}(s_{j+1})|s_j, a_j\right]\Big]\right\} \\
&= \frac{1}{H} \mathbb{E}\left\{l(\boldsymbol{\tau}) \sum_{j=1}^{H} \sum_{h=1}^{H} w_h(s_j, a_j) \Big[V_{h+1}(s_{j+1}) - \mathbb{E}\left[V_{h+1}(s_{j+1})|s_j, a_j\right]\Big]\right\}.
\end{aligned}$$

This give us the influence function of our objective function. If we denote

$$q(s,a,s') := \sum_{h=1}^{H} w_h(s,a) \Big[V_{h+1}(s') - \mathbb{E}\left[V_{h+1}(s')|s,a\right]\Big],$$

then the influence function can be written as

$$\mathcal{I}_p(\boldsymbol{\tau}) := \frac{1}{H} \sum_{j=1}^{H} \sum_{h=1}^{H} w_h(s_j, a_j) \Big[V_{h+1}(s_{j+1}) - \mathbb{E}\left[V_{h+1}(s_{j+1})|s_j, a_j\right]\Big] = \frac{1}{H} \sum_{h=1}^{H} q(s_h, a_h, s_{h+1}). \tag{46}$$

Next, we square the influence function and take expectation with it. Then we use $\mathbb{E}\left[q(s,a,s') \mid s,a\right] = 0$ to make cross terms vanish.

$$\begin{aligned}\mathbb{E}\left\{\mathcal{I}_p(\boldsymbol{\tau})^2\right\} &= \frac{1}{H^2}\mathbb{E}\left\{\sum_{h=1}^{H} q(s_h,a_h,s_{h+1})\right\}^2 \\ &= \frac{1}{H^2}\sum_{h=1}^{H}\mathbb{E}\left\{q(s_h,a_h,s_{h+1})^2\right\} \\ &= \frac{1}{H^2}\sum_{h=1}^{H}\sum_{h_1=1}^{H}\sum_{h_2=1}^{H}\mathbb{E}\Big\{w_{h_1}(s_h,a_h)w_{h_2}(s_h,a_h)\varepsilon_{h_1,h}\varepsilon_{h_2,h}\Big\}.\end{aligned}$$

where for $j,h\in[H]$,

$$\varepsilon_{j,h} := f\left(\theta_j^*,\phi_h\right) - r_h - \int_{\mathcal{A}} f\left(\theta_{j+1}^*,\phi_{h+1}\right)\pi\left(a_{h+1}\mid s_{h+1}\right)da_{h+1}\in\mathbb{R}.$$

Then we have

$$\begin{aligned}&\mathbb{E}\left\{\mathcal{I}_p(\boldsymbol{\tau})^2\right\} \\ =&\frac{1}{H^2}\sum_{h_1=1}^{H}\sum_{h_2=1}^{H}\mathbb{E}^\pi\left[\nabla_{\theta_{h_1}} f\left(\theta_{h_1}^*,\phi(s_{h_1},a_{h_1})\right)\mid s_1\sim\xi,p\right]\Sigma_{h_1}^{-1}\cdot\mathbb{E}\left[\sum_{h=1}^{H}\Big(\nabla_{\theta_{h_1}} f(\theta_{h_1},\phi(s_h,a_h))\Big)^\top\right. \\ &\left.\Big(\nabla_{\theta_{h_2}} f(\theta_{h_2}^*,\phi(s_h,a_h))\Big)\varepsilon_{h_1,h}\varepsilon_{h_2,h}\right]\Sigma_{h_2}^{-1}\mathbb{E}^\pi\left[\nabla_{\theta_{h_2}} f\left(\theta_{h_2}^*,\phi(s_{h_2},a_{h_2})\right)\mid s_1\sim\xi,p\right]^\top \\ =&\frac{1}{H}\sum_{h_1=1}^{H}\sum_{h_2=1}^{H}\nu_{h_1}^\top\Sigma_{h_1}^{-1}\Omega_{h_1,h_2}\Sigma_{h_2}^{-1}\nu_{h_2}\end{aligned}$$

where $\nu_h,\Sigma_h,\Omega_{h_1,h_2}$ are defined in Theorem 6.1. In conclusion, the expression above implies that

$$\mathbb{E}\left\{\mathcal{I}_p(\boldsymbol{\tau})^2\right\} = \sigma^2.$$

where σ^2 is defined as (7), and this proves the Cramer Rao lower bound for variance.

E. Technical Lemmas for Contraction

Lemma E.1 (Bernstein's Inequality)**.** *Let $X_1,\ldots,X_N$ be independent mean-zero random variables such that $|X_i|\le K$ all i. Then, for every $t\ge 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{N}X_i\right|\ge t\right\}\le 2\exp\left(-\frac{t^2/2}{\sigma^2+Kt/3}\right)$$

Here $\sigma^2=\sum_{i=1}^{N}\mathbb{E}X_i^2$ is the variance of the sum.

Lemma E.2 (Freedman's Inequality)**.** *Consider a real-valued martingale $\{Y_k : k=0,1,2,\ldots\}$ with difference sequence $\{X_k : k=1,2,3,\ldots\}$. Assume that the difference sequence is uniformly bounded:*

$$X_k\le R \quad \textit{almost surely} \quad \textit{for } k=1,2,3,\ldots$$

Define the predictable quadratic variation process of the martingale:

$$W_k := \sum_{j=1}^{k}\mathbb{E}_{j-1}\left(X_j^2\right)\quad \textit{for } k=1,2,3,\ldots$$

Then, for all $t\ge 0$ and $\sigma^2>0$,

$$\mathbb{P}\left\{\exists k\ge 0 : Y_k\ge t \textit{ and } W_k\le\sigma^2\right\}\le\exp\left\{-\frac{t^2/2}{\sigma^2+Rt/3}\right\}$$

Lemma E.3 (Matrix Bernstein inequality)**.** *Let $X_1, \ldots, X_N$ be independent mean-zero $n \times n$ symmetric random matrices, such that $\|X_i\| \leq K$ almost surely for all i. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left\{\left\|\sum_{i=1}^{N} X_i\right\| \geq t\right\} \leq 2n \exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right)$$

Here $\sigma^2 = \left\|\sum_{i=1}^{N} \mathbb{E}X_i^2\right\|$ is the norm of the matrix variance of the sum.

Lemma E.4 (Theorem 2.14.9 of (Van Der Vaart et al., 1996))**.** *Let $\mathcal{H}$ be a class of measurable functions $g : \mathcal{X} \mapsto [0,1]$ that satisfies*

$$N_{[]}\left(\varepsilon, \mathcal{H}, L_2(\mathbb{P})\right) \leq \left(\frac{U}{\varepsilon}\right)^V, \quad \textit{for every } 0 < \varepsilon < U.$$

We denote the empirical process:

$$\mathbb{G}_n(\cdot) = \sqrt{n}(\mathbb{P}_n(\cdot) - \mathbb{P}(\cdot))$$

For any function class $\mathcal{H}$, we define the supremum norm

$$\|\mathbb{G}_n\|_{\mathcal{H}} := \sup_{g \in \mathcal{H}} \left\{\sqrt{n}(\mathbb{P}_n g - \mathbb{P} g)\right\}$$

Then, for every $t > 0$,

$$\mathrm{P}\left(\|\mathbb{G}_n\|_{\mathcal{H}} > t\right) \leq \left(\frac{Dt}{\sqrt{V}}\right)^V e^{-2t^2}$$

for a constant D that depends on U only.

F. A Brief Comparison to Related Works

We compare our setting and results with some related works. Here we focus on works which dealt with policy evaluation. Researches on policy optimization(Jin et al., 2019; Cai et al., 2020) or Fitted Q Iteration (Long et al., 2021; Fan et al., 2020) approximated by linear, kernel function, or neural networks are not directly related and not included in this comparison. (Yin & Wang, 2020) used marginal importance sampling to set up asymptotic efficiency in batch tabular RL. Since MIS is not an iterative method, they don't need to approximate Q-functions explicitly. But they assume stronger concentrability(Discussion after Assumption 2.2 and Assumption 2.3). These two assumptions imply $\frac{\mu(s,a)}{\bar{\mu}(s,a)} \leq C$ in (Chen & Jiang, 2019). Their main result (Theorem 3.1) is about $\mathbb{E}_{\bar{\mu}}\left[\widehat{v}_\pi - v_\pi\right]$, but using Chebyshev's Inequality we can soon get a high probability bound for $\left|\widehat{v}_\pi - v_\pi\right|$. Roughly speaking, this high probability upper bound is $\left|\widehat{v}_\pi - v_\pi\right| \leq \frac{C_{\mu,\bar{\mu}}}{\sqrt{N}}$, where $C_{\mu,\bar{\mu}} = C\sqrt{\sum_{h=1}^{H}\sum_{s_h,a_h}\frac{d_h^\pi(s_h)^2}{d_h^{\bar{\pi}}(s_h)}\frac{\pi(a_h|s_h)^2}{\bar{\pi}(a_h|s_h)}\operatorname{Var}\left[\left(V_{h+1}\left(s_{h+1}^1\right) + r_h^1\right) \mid s_h^1 = s_h, a_h^1 = a_h\right]}$ is a constant capturing distribution mismatch. This constant is similar to a χ^2-divergence.

Our work is mostly related to (Hao et al., 2021b) and (Duan et al., 2020). They considered linear function approximation with possibly infinite state-action space. Under policy completeness, (Duan et al., 2020) showed that $\left|\widehat{v}^\pi - v^\pi\right| \leq \frac{C}{\sqrt{N}}\left[\sum_{h=1}^{H}(H-h+1)\sqrt{1+\chi^2_{\mathcal{Q}}(\mu_h,\bar{\mu})}\right] + O(\frac{1}{N})$, where μ_h is marginal distribution of (s_h, a_h) generated under target policy, $\mathcal{Q}$ is linear function class spanned by feature map $\phi(s,a)$, and is also the space where all Q-functions lie in. This χ^2-divergence is a special case of ours. (Hao et al., 2021b) established asymptotic results under the same setting, including asymptotic normality(equivalent to asymptotic upper bound), and asymptotic lower bound. The asymptotic variance meets Remark 3.2 in (Yin & Wang, 2020) in tabular case. Moreover they consider standard bootstrap, while we consider one of its alternatives to get a more general asymptotic confidence interval.

(Uehara et al., 2020) does not assume completeness, but it assumes stronger realizability, not only for Q-functions but for density ratio as well. Although two realizability are not necessary for their finite upper bound, they still need certain assumption stronger than realizability for Q-functions only. Their concentrability condition(Assumption 2) is the same as that in (Chen & Jiang, 2019). They proved the sample complexity of MWL / MQL estimators and provided semi-parametric efficiency. Another example of non-parametric OPE is (Kallus & Uehara, 2020), which matches the semi-parametric lower bound fr variance as well. Assumption 1 in (Kallus & Uehara, 2020) assumes a full data coverage and bounded density ratio, which is stronger than that in (Chen & Jiang, 2019). They further assume proper rates of estimators of Q-functions and density ratios, but they do not focus on estimators of both nuisance. Under these assumptions, they get an $O(\frac{1}{\sqrt{K}})$ error rate, and the dominating term depends on effective bound of MDP $\mathrm{Effbd}\,(\mathcal{M}_2)$ which captures the distribution mismatch(See equation (4) in their paper). (Duan et al., 2021) considered online policy evaluation, hence neither concentratability nor completeness is needed. They consider a uniformly bounded kernel function class, which is intrinsically non-parametric when there are infinite eigenvalues for kernel operator. Their result is based on $L^2(\xi^\pi)$-norm of statistical error of value function approximator, and R in their result captures the structual property of the projected fixed point(In case of no approximation error, this is just policy value). In linear kernel setting, this reduces into linear LSTD estimation and has an error rate of order $O(\frac{1}{\sqrt{N}})$. In another case, where the eigenvalues for kernel functions decay with α-polynomial rate, they derive the dependency of the norm of error on both sample size and effective horizon.