A Fine-grained Analysis of Fitted Q-evaluation: Beyond Parametric Models

Jiayi Wang [1] **Zhengling Qi** [2] **Raymond K. W. Wong** [3]

Abstract

In this paper, we delve into the statistical analysis of the fitted Q-evaluation (FQE) method, which focuses on estimating the value of a target policy using offline data generated by some behavior policy. We provide a comprehensive theoretical understanding of FQE estimators under both parameteric and nonparametric models on the Q-function. Specifically, we address three key questions related to FQE that remain largely unexplored in the current literature: (1) Is the optimal convergence rate for estimating the policy value regarding the sample size n ($n^{-1/2}$) achievable for FQE under a non-parametric model with a fixed horizon (T)? (2) How does the error bound depend on the horizon T? (3) What is the role of the probability ratio function in improving the convergence of FQE estimators? Specifically, we show that under the completeness assumption of Q-functions, which is mild in the non-parametric setting, the estimation errors for policy value using both parametric and non-parametric FQE estimators can achieve an optimal rate in terms of n. The corresponding error bounds in terms of both n and T are also established. With an additional realizability assumption on ratio functions, the rate of estimation errors can be improved from $T^{1.5}/\sqrt{n}$ to $T/\sqrt{n}$, which matches the sharpest known bound in the current literature under the tabular setting.

1. Introduction

In reinforcement learning (RL), off-policy evaluation (OPE) is an important topic that focuses on estimating the expected total reward (e.g., the value defined in (1)) of a target policy based on data collected from a potentially different and unknown policy (Sutton & Barto, 2018). OPE is particularly useful in high-stakes domains where the implementation of a new policy can incur significant costs or risks, and has been extensively studied in RL (e.g. Xie et al., 2019; Duan et al., 2020; Yin & Wang, 2020; Chen & Qi, 2022; Ji et al., 2022; Wang et al., 2023). See Uehara et al. (2022) for an overview of the OPE methods. Among various algorithms for OPE, fitted Q-evaluation (FQE) is arguably one of the most popular algorithms. In FQE, Q-functions (defined in (3)) are estimated in a backward manner using supervised learning methods, and their estimates are then used to construct an estimated policy value (see (6)). FQE has demonstrated significant empirical success in many applications (Fu et al., 2021; Voloshin et al., 2019). Its popularity and success have also led to significant theoretical interest in FQE. Several recent studies aim to provide theoretical justifications for its effectiveness (as discussed in detail in Section 1.1). In this work, we delve deeply into the analysis of FQE estimators within the framework of a finite-horizon, time-inhomogeneous Markov Decision Process (MDP). Compared to existing analysis in FQE, our objective is to provide a more comprehensive understanding of the convergence rate on the error bound of estimating the value of a target policy under both parametric and nonparametric models in terms of both the number of episodes n and the horizon T. Specifically, we seek to address the following three fundamental questions related to FQE:

- **Q1**: For the fixed horizon T, how does the convergence rate depend on the number of episodes n given the completeness assumption for Q-functions? Is the optimal convergence rate ($n^{-1/2}$) still achievable under nonparametric models of Q-functions?

- **Q2**: How does the convergence rate depend on the growing horizon T?

- **Q3**: What is the role of the probability ratio functions w_t^π (defined in (2)) in improving the convergence rate for FQE estimators?

We will comment on these questions and the existing progress towards addressing them in the following subsection.

[1]Department of Mathematical Sciences, University of Texas at Dallas, Richardson, USA [2]School of Business, The George Washington University, Washington, D.C., USA [3]Department of Statistics, Texas A&M University, College Station, USA. Correspondence to: Zhengling Qi <qizhengling@email.gwu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1.1. Related Literature

In recent years, many works have studied FQE from a theoretical perspective with the goal to address **Q1** and **Q2**, with varying degrees of success under different modeling assumptions. Below, we survey these efforts and indicate that some important understanding is still lacking.

The first line of research focuses on studying FQE under some parametric model on Q-functions. For example, Duan et al. (2020) assumed linear MDP and established a $T^2/\sqrt{n}$ order for the estimation error of the policy value. See Theorem 2 in Duan et al. (2020). Furthermore, Zhang et al. (2022) studied beyond linear MDP and their analysis allows that Q-functions lie in an almost arbitrarily parametrized function class with some differentiability condition. They obtained the same order for the estimation error as Duan et al. (2020). In those aforementioned works, they showed that the optimal convergence rate with respect to n can be achieved by using FQE under a parametric model of Q-functions. Regarding **Q2**, they obtain a quadratic dependence with respect to T. However, as shown in Yin & Wang (2020), a linear and thus better dependence of T can be achieved under the tabular setting. *This leads to an interesting question that whether linear horizon dependence can be obtained beyond tabular setting.*

The second line of research considers FQE under some nonparametric models of Q-function. Specifically, Nguyen-Tang et al. (2021) provided an error bound for the estimation error of nonparametric FQE using feed-forward ReLU network. They showed that the estimation error is of an order $T^{2-\alpha/(2\alpha+2D)}n^{-\alpha/(2\alpha+2D)}$, where α is a smoothness parameter and D is the dimension of state and action. Compared to Nguyen-Tang et al. (2021), Ji et al. (2022) further assumed a low-dimensional manifold structure in convolutional neural networks and improved the error bound to an order of $T^2 n^{-\alpha/(2\alpha+d)}$, where d is the intrinsic dimension of the state-action space. Regarding **Q1**, these two works are only able to show a slower convergence rate of estimating the policy value than the optimal $n^{-1/2}$ with fixed T. *Hence, it is unclear if $n^{-1/2}$ rate is achievable for nonparametric FQE based on the current literature.* We will provide a positive result to this question in this paper. For **Q2**, Nguyen-Tang et al. (2021) showed the horizon dependence is of an order that is larger than $T^{1.5}$ and Ji et al. (2022) showed a quadratic dependence. *Similarly to the parametric methods, it is unclear if the linear dependence of the horizon can be obtained under non-parametric models.*

Besides FQE, another line of research uses marginal importance sampling (MIS) or probability ratio functions to address the distributional mismatch due to the difference between the target policy and the behavior policy and develop an MIS estimator for OPE. In particular, for time-inhomogeneous settings and for tabular MDP, Xie et al. (2019) and Yin & Wang (2020) showed that the estimation error of the MIS estimator has an order of $T/\sqrt{n}$ under some proper assumptions. To the best of our knowledge, this is the sharpest bound with respect to both n and T in the existing literature. Building on the insights discussed in Section 3.3 of Duan et al. (2020), a connection between FQE estimators and marginalized importance sampling (MIS) estimators (refer to the forms in (8) and (9)) can be established. Given this connection, one naturally wonders if the successful analysis of MIS estimator (under tabular setting) can be leveraged to understand the horizon dependence for the convergence rate of FQE *beyond tabular setting*. Specifically, we ask the following question: is the linear dependence of T in the convergence rate of FQE estimators achievable due to the connection with MIS estimator? Moreover, in the context of a continuous state space with some nonparametric model, we further ask: whether any conditions for the marginalized sampling weights (or the probability ratio function) need to be imposed in order to achieve such sharp dependence of T and how such conditions contribute to addressing the convergence of FQE estimators (**Q3**). We will address all these three intriguing but rarely studied questions, contributing to improving the understanding of FQE.

1.2. Our Contributions

In this work, we focus on addressing the three aforementioned questions in the context of FQE where Q-functions are estimated under either a parametric linear model, or a non-parametric model via linear sieves (Ai & Chen, 2003). In particular, we successfully establish the following results:

1. For fixed T, FQE estimators with Q-functions modeled parametrically is shown to achieve the optimal convergence rate with respect to n ($n^{-1/2}$). When Q-functions are modeled nonparametrically, $n^{-1/2}$ rate can be still obtained when Q-functions satisfy certain smoothness conditions. In contrast, under these conditions, the estimation for Q-functions only has a convergence rate slower than $n^{-1/2}$.

2. For asymptotically growing T, the first-order term of the error bound is shown to have a dependence of $T^{3/2}$ while the higher order term has a dependence of T^3 for both parametric and nonparametric FQE. Our bound has a milder dependence with respect to T for the first-order term compared to current literature, where their bounds have a quadratic dependence of T.

3. When probability ratio functions lie in a space of smooth functions, *without additionally estimating these probability ratio functions (like those double robust estimators)*, the first-order term of the error bound for the vanilla FQE estimators is shown to converge with an order of $T/\sqrt{n}$

Table 1. Comparison on the error bound for the first-order term in existing works. κ is defined in (11). $\tilde{\kappa}$ is the upper bound for the probability ratio functions; D is the dimension of space and action. d is the intrinsic dimension of the state-action space. Some logarithmic orders are omitted in the error bounds.

WORK	PARAMETRIC?	REGULARITY ON Q	ERROR BOUND (W.H.P)
YIN & WANG (2020)	$\surd$	TABULAR	$\mathcal{O}(T\tilde{\kappa}\sqrt{1/n})$
DUAN ET AL. (2020)	$\surd$	LINEAR	$\mathcal{O}(T^2\sqrt{\kappa/n})$
ZHANG ET AL. (2022)	$\surd$	DIFFERENTIABLE	$\mathcal{O}(T^2\sqrt{\kappa/n})$
NGUYEN-TANG ET AL. (2021)	$\times$	BESOV	$\mathcal{O}(T^{2-\alpha/(2\alpha+2D)}\tilde{\kappa}n^{-\alpha/(2\alpha+2D)})$
JI ET AL. (2022)	$\times$	BESOV	$\mathcal{O}(T^2\kappa n^{-\alpha/(2\alpha+d)})$
OUR WORK IN SECTION 3.1	$\surd$	LINEAR	$\mathcal{O}(T^{1.5}\sqrt{\kappa/n})$ $\mathcal{O}(T\tilde{\kappa}\sqrt{1/n})$ WHEN w_t^π ARE LINEAR
OUR WORK IN SECTION 3.2	$\times$	HÖLDER	$\mathcal{O}(T^{1.5}\sqrt{\kappa/n})$ WHEN Q_t^π ARE SMOOTH ENOUGH $\mathcal{O}(T\tilde{\kappa}\sqrt{1/n})$ WHEN w_t^π ARE HÖLDER

in both parametric and nonparametric settings. This bound matches with the sharpest rate of convergence in the tabular setting.

2. Set Up

To set the stage for our theoretical discussion, we review the framework of discrete-time inhomogeneous Markov Decision Processes (MDP) and the fitted Q-evaluation (FQE) for estimating policy value in this section.

2.1. Preliminary

Denote by $\mathcal{M} = (T, \tilde{\mathcal{S}}, \tilde{\mathcal{A}}, \tilde{\Pr}, \tilde{\mathcal{R}})$ a finite-horizon episodic Markov Decision Process (MDP), where the integer T is the length of horizon, $\tilde{\mathcal{S}} = \{\mathcal{S}_t\}_{t=1}^T$ and $\tilde{\mathcal{A}} = \{\mathcal{A}_t\}_{t=1}^T$ are the state spaces and the action spaces across T decision points respectively, $\tilde{\Pr} = \{\Pr_t\}_{t=1}^T$ with $\Pr_t(\bullet \mid s, a)$ representing the transition kernel (probability) given the state $s \in \mathcal{S}_t$ and the action $a \in \mathcal{A}_t$, and $\mathcal{R} = \{R_t\}_{t=1}^T$ are immediate rewards such that $R_t|S_t = s, A_t = a \sim R_t(s, a)$. We take $r_t(s, a)$ as the conditional mean of $R_t(s, a)$. A trajectory generated from $\mathcal{M}$ can be written as $\{S_t, A_t, R_t\}_{t=1}^T$, where $S_t \in \mathcal{S}_t$, $A_t \in \mathcal{A}_t$ and $R_t \in \mathbb{R}$ denote the state, the action and the reward at time t respectively. Without loss of generality, we assume $|R_t| \leq 1$ for $t = 1, \ldots, T$. In the following discussion, for the sake of simplicity, we assume the same state spaces and action spaces across all decision points, denoted by $\mathcal{S} \subset \mathbb{R}^d$ and $\mathcal{A}$, respectively. Additionally, we assume the action space $\mathcal{A}$ is finite.

A policy is defined as a way of choosing actions at each decision time point t. More specifically, denote a target policy as $\pi = \{\pi_t\}_{t=1}^T$, where π_t is a function mapping from the state space $\mathcal{S}$ to a probability mass function over the action space $\mathcal{A}$. Then OPE aims to estimate the value of π defined as

$$\nu(\pi) = \mathbb{E}^\pi \left[\sum_{t=1}^T R_t \right], \tag{1}$$

using the pre-collected data generated by a fixed stationary policy $\pi^b = \{\pi_t^b\}_{t=1}^T$, which is called behavior policy. Here $\mathbb{E}^\pi$ denotes the expectation with respect to the distribution whose actions are generated by the target policy π. We further assume that the pre-collected training data consist of n independent and identically distributed trajectories as $\mathcal{D}_n = \left\{ \{(S_{i,t}, A_{i,t}, R_{i,t})\}_{1 \leq t < T} \right\}_{1 \leq i \leq n}$. For convenience, we omit π^b in the superscript in the notation of the expectation and probability under the distribution induced by π^b.

Next we define notation for several important probability distributions. For any $t = 1, \ldots, T$, we take $\rho_t^\pi(s, a)$ and $\rho_t^b(s, a)$ as the marginal density of (S_t, A_t) at $(s, a) \in \mathcal{S} \times \mathcal{A}$ under the target policy π and behavior policy π^b respectively. And we define the probability ratio function w_t^π as

$$w_t^\pi(s, a) = \rho_t^\pi(s, a)/\rho_t^b(s, a), \tag{2}$$

for $t = 1, \ldots, T$.

2.2. Fitted Q-evaluation

One can find $\nu(\pi)$ by computing the state-action value functions (also known as the Q-functions) defined as

$$Q_t^\pi(s, a) = \mathbb{E}^\pi \left[\sum_{t'=t}^T R_t | S_t = s, A_t = a \right], \tag{3}$$

for $t = 1, \ldots, T$. Then $\nu(\pi) = \int_{(s,a)\in\mathcal{S}\times\mathcal{A}} Q_1^\pi(s,a)\rho_1^\pi(s,a)d(s,a)$. For simplicity, we assume the initial state distribution ρ_1^π is known. To compute the Q-function, it is well-known that $\{Q_t^\pi\}_{t=1}^T$ satisfy the following Bellman equation

$$Q_t^\pi(s,a) = \mathbb{E}\left[R_t + V_{t+1}^\pi(S_{t+1})|S_t = s, A_t = a\right], \quad (4)$$

where $V_t^\pi(s) = \sum_{a\in\mathcal{A}} \pi_t(a'|s)Q_t^\pi(s,a')$.

Motivated by (4), in fitted Q-evaluation (FQE), one can utilize the offline data $\mathcal{D}_n$ and recursively apply a regression technique to learn $Q_T^\pi, Q_{T-1}^\pi, \ldots, Q_1^\pi$ in a sequential and backward order. More specifically, let $\hat{Q}_{T+1}^\pi = 0$, and for $t = T, T-1, \ldots, 1$, one can compute

$$\hat{Q}_t^\pi = \operatorname*{argmin}_{Q\in\mathcal{Q}^{(t)}} \frac{1}{n}\sum_{i=1}^n \left\{ Q(S_{i,t}, A_{i,t}) - \left[R_{i,t} + \sum_{a'\in\mathcal{A}_{t+1}} \pi_t(a \mid S_{i,t+1})\hat{Q}_{t+1}^\pi(S_{i,t+1}, a')\right]\right\}^2 \quad (5)$$

for Q_t^π, where $\mathcal{Q}^{(t)}$ is a hypothesis class for Q_t^π. Then the policy value is estimated via a plug-in estimator:

$$\hat{\nu}(\pi) = \int_{(s,a)\in\mathcal{S}\times\mathcal{A}} \hat{Q}_1^\pi(s,a)\rho_1^\pi(s,a)d(s,a). \quad (6)$$

When considering parametric models, one can use a linear model to approximate Q_t^π for every t (e.g. Duan et al., 2020; Min et al., 2021). For example, take $\Psi_K(s) = [\psi_1(s), \cdots, \psi_K(s)]^\intercal$ as a vector consisting of K features for $s \in \mathcal{S}$. Let $\phi_K(s,a) = [\mathbb{1}(a=1)\Psi_K(s)^\intercal, \ldots, \mathbb{1}(a = |\mathcal{A}|)\Psi_K(s)^\intercal]^\intercal \in \mathbb{R}^{K|A|}$. Then one can let $\mathcal{Q}^{(t)} = \{\phi_K(\cdot)^\intercal\beta : \beta \in \mathbb{R}^{K|\mathcal{A}|}\}$.

Many existing work study the theoretical property of FQE under the linear model assumption such as $Q_t^\pi \in \mathcal{Q}^{(t)}$ (realizability), and K is fixed *independent of* n *and* T. We call this setup the *parametric* setting, as each Q_t^π is modeled via a finite number of parameters. The realizability of such linear modeling hedges strongly on the careful selection of the features, which is often non-trivial. To relax this assumption, an infinite-dimensional modeling for Q_t^π can be used, which we refer to as a *nonparametric* setting. In this work, we assume that for every $a \in \mathcal{A}$, $Q_t^\pi(\cdot, a)$ lies in a Hölder space which, roughly speaking, consists of functions $g : \mathcal{S} \subset \mathbb{R}^d \to \mathbb{R}$ with Hölder continuous derivatives of certain order. Specifically, a Hölder space is defined as

$$\Lambda_\infty(p, L) = \left\{ g : \sup_{0\le\|\alpha\|_1\le\lfloor p\rfloor} \|\partial^\alpha g\|_\infty \le L, \sup_{\alpha:\|\alpha\|_1\le\lfloor p\rfloor} \sup_{x,y\in\mathcal{S}, x\neq y} \frac{|\partial^\alpha g(x) - \partial^\alpha g(y)|}{\|x-y\|_2^{p-\lfloor p\rfloor}} \le L \right\}, \quad (7)$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ are the ℓ_1 and ℓ_2 norm of a vector respectively, $\lfloor p\rfloor$ denotes the integer less or equal to p for $p > 0$, $\alpha = (\alpha_1, \cdots, \alpha_d)$ is a non-negative vector,

$$\partial^\alpha g(x) = \frac{\partial^\alpha g(x)}{\partial^{\alpha_1}\cdots\partial^{\alpha_d}}.$$

A standard approach for nonparametric regression is sieve estimation. We study a linear-sieve estimator in our theoretical study of nonparametric FQE, where an increasing number of basis functions are allowed to approximate Q_t^π in the Hölder space, as sample size increases. To be specific, with slight abuse of notation, take $\Psi_K(\cdot) = [\psi_1(\cdot), \cdots, \psi_K(\cdot)]^\intercal$ as a vector consisting of K sieve basis functions at time point t, where the number of basis functions K *is allowed to depend on* n *and* T. One can take splines or wavelet bases (see for example, Huang (1998) and Chen & Christensen (2018)) for choices of basis functions. Take $\phi_K(s,a) = [\Psi_K(s)^\intercal\mathbb{1}(a=1), \ldots, \Psi_K(s)^\intercal\mathbb{1}(a=|\mathcal{A}|)]^\intercal$. We approximate $\mathcal{Q}^{(t)}$ with space $\tilde{\mathcal{Q}}^{(t)} := \text{span}\{(\phi_K(\cdot,\cdot)\}$ and solve for (5) with $\mathcal{Q}^{(t)}$ replaced by $\tilde{\mathcal{Q}}^{(t)}$. In practice, one can choose different types and numbers of basis functions at different time points. To simplify the notation, we use a universal set of basis functions and a universal number of basis functions.

Sieve estimations have been extensively studied in statistics and econometric communities with their appealing empirical performance and ease of computation (e.g. Geman & Hwang, 1982; Huang, 1998; Ai & Chen, 2003; Chen & Christensen, 2018). Some recent works in OPE also utilize the linear sieves as a tool to study the nonparametric estimation of Q-functions (e.g. Shi et al., 2022; Chen & Qi, 2022; Wang et al., 2023). Many sieve bases can effectively approximate infinite-dimensional spaces that contain a wide range of smooth functions. For example, (tensor-product) B-spline basis and wavelet basis can approximate Hölder space well, with the approximation error decreasing as the number of basis functions increases. See Section 2.2 of Huang (1998) for detailed discussions on the approximation power of these bases.

2.3. Connection with Marginal Importance Sampling Estimator

In this section, we connect FQE with a marginal importance sampling (MIS) estimator, which will help us understand the role of ratio functions in our theoretical analysis in the later sections. To proceed, we introduce some notation. Unless specified otherwise, the following notations apply to both parametric and nonparametric cases. Let $\Sigma_t = \mathbb{E}[\phi_K(S_t, A_t)\phi_K(S_t, A_t)^\intercal] \in \mathbb{R}^{K|\mathcal{A}|\times K|\mathcal{A}|}$, $\hat{\Sigma}_t = \frac{1}{n}\sum_{i=1}^n [\phi_K(S_{i,t}, A_{i,t})\phi_K(S_{i,t}, A_{i,t})^\intercal] \in \mathbb{R}^{K|\mathcal{A}|\times K|\mathcal{A}|}$ and $\Sigma_{t,a} = \mathbb{E}[\psi_K(S_t)\psi_K(S_t)^\intercal \mid A_t = a] \in \mathbb{R}^{K\times K}$.

Define $\mathcal{P}_t^\pi$ and $\hat{\mathcal{P}}_t^\pi$ as the population and estimated condi-

tional expectation operators respectively, such that

$$(\mathcal{P}_t^\pi f)(s,a) = \mathbb{E}\left\{\sum_{a'} \pi_t(a' \mid S_{t+1}) f(S_{t+1}, a') \mid S_t = s, A_t = a\right\},$$
$$(\hat{\mathcal{P}}_t^\pi f)(s,a) = \phi_K(s,a)^\intercal (\hat{\Sigma}_t)^{-1} \left(\tfrac{1}{n}\sum_{i=1}^n \phi_K(S_{i,t}, A_{i,t}) \left[\sum_{a'} \pi_t(a' \mid S_{i,h+1}) f(S_{i,t+1}, a')\right]\right),$$

for $f \in \mathcal{Q}^{(t+1)}$, $t = 1, \ldots, T-1$.

Define $\mathcal{E}_t f = \mathbb{E} f(S_t, A_t) = \int_{(s,a)} f(s,a)\rho_t^b(s,a)d(s,a)$, $\mathcal{E}_t^\pi f = \mathbb{E}^\pi f(S_t, A_t) = \int_{(s,a)} f(s,a)\rho_t^\pi(s,a)d(s,a)$ for any function f. One can verify that the value estimator (6) based FQE can be represented in the following form:

$$\hat{\nu}(\pi) = \sum_{t=1}^T \frac{1}{n} \sum_{i=1}^n \hat{w}_{i,t} R_{i,t} \tag{8}$$

where

$$\hat{w}_{i,t} = \mathcal{E}_1^\pi \left\{ \left(\prod_{t'=0}^{t-1} \hat{\mathcal{P}}_{t'}^\pi \right) [\phi_K^\intercal(\cdot)]^\intercal (\hat{\Sigma}_t)^{-1} \right\} \phi_K(S_{i,t}, A_{i,t}). \tag{9}$$

Note that (8) can be considered as a MIS estimator where $\hat{w}_{i,t}$ is used to estimate the ratio function $w_t^\pi(S_{i,t}, A_{i,t})$ defined in (2)). As discussed in Section 3.3 in Duan et al. (2020), under the tabular case, (8) matches the estimator proposed in Yin & Wang (2020).

Additional notations. We provide some additional notation that will be used later in the paper. Denote by $\|\cdot\|_{\mathcal{L}_2}$ and $\|\cdot\|_\infty$ the $\mathcal{L}_2$-norm and the infinity norm respectively. More specifically, $\|f\|_{\mathcal{L}_2} = \sqrt{\mathbb{E} f^2(S_t, A_t)}$ and $\|f\|_\infty = \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |f(s,a)|$ for some function f defined on $\mathcal{S}\times\mathcal{A}$. Define the projection Π_t such that $\Pi_t g(s,a) = \phi_K(s,a)^\intercal (\Sigma_t)^{-1} \mathbb{E}\left[\phi_K(S_t, A_t) g(S_t, A_t)\right]$ for $g \in \mathcal{Q}^{(t)}$. We use the notation $\lesssim$ ($\gtrsim$) to denote less (greater) than up to an absolute constant. We write $a \asymp b$ if $a \lesssim b$ and $b \gtrsim a$. Lastly, if a non-negative random variable X satisfies $P(X \le c\varrho(n,T)) \to 0$ as $c \to \infty$ for any n, T, we write $X = \mathcal{O}(\varrho(n,T))$ with high probability (W.H.P).

3. Error Bounds

In this section, we analyze the finite-sample upper bound for $|\nu(\pi) - \hat{\nu}(\pi)|$. Note that the transitions are not required to be homogeneous, i.e., $\Pr_t$ can vary across t. Also for the sake of clarity, we present key results in the main text and leave the most general error bounds in Section B of the appendix.

In the following, we impose the following assumptions for Q_t^π, $t = 1, \ldots, T$ and the basis functions.

Assumption 3.1. $r_t \in \mathcal{Q}^{(t)}$, for $t = 1, \ldots, T$. For every $q \in \mathcal{Q}^{(t+1)}$, we have $\mathcal{P}_t^\pi q \in \mathcal{Q}^{(t)}$.

Assumption 3.2. (i) $\mathcal{S} \subset \mathbb{R}^d$ is compact. The densities $\rho_t^b, t = 1, \ldots, T$ are uniformly bounded away from 0 and ∞ on $\mathcal{S}\times\mathcal{A}$, i.e., there exist constants $\underline{M}, \overline{M} > 0$ (independent of T) such that $\underline{M} \le \rho_t^b(s,a) \le \overline{M}$ for all s, a, t. (ii) The minimal eigenvalue of $\Sigma_{t,a}$ is uniformly lower bounded for all t and a. In addition, $\zeta_{K,t} := \sup_{s,a} \|\Sigma_t^{-1/2}\phi_K(s,a)\|_2 = \mathcal{O}(K)$ hold uniformly for all t. (iii) The sieve basis (features) in Ψ_K is Hölder continuous, i.e., there exist finite constants $\omega \ge 0$ and $\omega' > 0$ (independent of K and T) such that the following inequality holds for every t, a and s_t, s_t'.

$$\left\|\Sigma_{t,a}^{-1/2}\{\Psi_K(s_t) - \Psi_K(s_t')\}\right\|_2 \lesssim K^\omega \|s_t - s_t'\|_2^{\omega'}. \tag{10}$$

Assumption 3.1 states the realizability for function spaces $\mathcal{Q}_t$, $t = 1, \ldots, T$ and the completeness for the Bellman operator, which are widely adopted in RL literature (e.g. Ji et al., 2022; Duan et al., 2020; Chen & Jiang, 2019). *When $\mathcal{Q}_t$ is an infinite-dimensional space such as Hölder class, Assumption 3.1 is mild.* Assumption 3.2 states the conditions for the basis functions. Assumption 3.2(i) is standard. The requirement for bounded support can be relaxed to unbounded support with a slight modification, see, e.g., Blundell et al. (2007); Chen & Christensen (2018); Chen & Pouzo (2012). Assumption 3.2(ii) and (iii) are very mild and are satisfied by many commonly used sieve bases, such as splines and wavelets (Chen & Christensen, 2018) under Assumption 3.2(i).

Define

$$\kappa := \frac{1}{T}\sum_{t=1}^T \sup_{f\in\mathcal{Q}^{(t)}} \frac{[\mathcal{E}_t^\pi f]^2}{\|f\|_{\mathcal{L}_2}^2}, \tag{11}$$

The constant κ quantifies the distribution shift between data induced by the behavior policy and the target policy. It shows up in the later error bounds in our analysis and also appears in many prior theoretical studies of OPE (e.g. Duan et al., 2020; Ji et al., 2022). Under Assumption 3.2(i), κ is always upper bounded.

3.1. Parametric Setting: A Preliminary Result

In this section, we provide error bounds under the settings where $\mathcal{Q}^{(t)} = \left\{\phi_K(\cdot,\cdot)^\intercal \beta : \beta \in \mathbb{R}^{K|\mathcal{A}|}\right\}$ for $t = 1, \ldots, T$ and K is a fixed constant. In this case, we have $Q_t^\pi = \Pi_t Q_t^\pi$ for $t = 1, \ldots, T$. which is the same as the in-homogeneous setting with linear function approximation discussed in Duan et al. (2020).

Theorem 3.3. *Under Assumptions 3.1-3.2, and further assume that $T = o([n/(\log n \log T)]^{1/2})$, we have W.H.P,*

$$|\hat{\nu}(\pi) - \nu(\pi)| = \mathcal{O}\left(\sqrt{\frac{T^3\kappa}{n}} + T^3 \frac{\log n \log T}{n}\right). \tag{12}$$

The first term in our bound (12) exhibits an order of $T^{1.5}/\sqrt{n}$. If we focus on the convergent cases, the first

term of (12) dominates (up to logarithmic orders of n and T), and hence is usually of interest. In (12), we refer to the first term as the first-order term and the second term as the higher-order term. In the later discussion, following this convention, we refer the term that has a slowest dependence on n as the first-order term and the remaining terms as the higher-order terms.

Extending the theoretical results from Duan et al. (2020) in the time-homogeneous setting to the in-homogeneous one, their bound will also comprise two major terms. The first-order term will have an order of $T^2/\sqrt{n}$. Compared with their bound, our first order term has an order of $T^{1.5}/\sqrt{n}$. We achieve a sharper horizon dependence by exploiting the fact that the variance of the first order term can be decomposed as a sum of T individual expectations of the conditional variance. See Lemma D.1 for more details. It is important to note that we use the exact same conditions as Duan et al. (2020) to achieve this rate of convergence. As for the higher-order term, our bound exhibits an order of T^3/n and their bound is $T^{3.5}/n$. We have a stronger requirement for basis functions (Assumption 3.2(iii)) and derive the uniform convergence to achieve this. If we drop this assumption, by adopting their proof, we can show the same bound as theirs for the higher-order term.

Next, Theorem 3.5 shows that with an additional realizability assumption (Assumption 3.4) on the probability ratio functions, the convergence rate of the error will depend *linearly* with respect to the horizon T in the first-order term. This is a significant improvement in horizon dependence over the existing literature on the setting of using linear function approximation. Our result also aligns with the sharpest known dependence of the horizon in the first-order term under the tabular setting, as established in Yin & Wang (2020) for their MIS estimator. Note that the MIS estimator in the tabular setting can be considered as a special case of (8) by taking basis functions as indicator functions, due to the equivalence discussed in Section 2.2. In this case, it is also remarked that Assumption 3.4 holds automatically. Therefore, Theorem 3.5 bridges the gap in the current literature by showing linear horizon dependence for more general linear modeling (with potentially continuous state space).

Assumption 3.4. $w_t^\pi \in \{\phi_K(\cdot,\cdot)^\intercal \beta : \beta \in \mathbb{R}^{K|\mathcal{A}|}\}$, $t = 1,\ldots,T$.

Theorem 3.5. *Under Assumptions 3.1-3.2, and 3.4 with the condition that $T = o([n/(\log n \log T)]^{1/2})$, we have W.H.P,*

$$|\hat{\nu}(\pi) - \nu(\pi)| = \mathcal{O}\left(T\sqrt{\frac{\kappa}{n}} + T^3\frac{\log n \log T}{n}\right). \quad (13)$$

3.2. Nonparametric Setting: Key Results

In this subsection, we generalize the parametric setting to the case where Q functions are modeled nonparametrically. As opposed to the previous subsection, one fundamental difference is that linear function approximation with finite K could incur non-negligible approximation errors. In the non-parametric setting, the number basis functions K will grow with the sample size so that the approximation error diminishes asymptotically. However, one needs to control the estimation error due to the increasing model complexity. In addition to the assumptions listed in Section 3.1, we will need the following assumptions for our theoretical study.

Assumption 3.6. For every $a \in \mathcal{A}$ and $t = 1,\ldots,T$, $\{q(\cdot,a) : q \in \mathcal{Q}^{(t)}, \|q\|_\infty \le 1\}$ is a subset of $\Lambda_\infty(p,L)$ with constants $p > d/2$ and $L > 0$.

Assumption 3.7. There exists a constant $\beta_Q > 1/2$ (independent of T) such that $\sup_{q\in\mathcal{Q}^{(t)}(1)} \|q - \Pi_t q\|_\infty \lesssim K^{-\beta_Q}$ for $t = 1,\ldots,T$.

Again, we emphasize that Assumption 3.6 together with Assumption 3.1 are very mild as the Hölder space is very broad. Indeed, we can show that Assumption 3.1 is not hard to satisfy under Assumption 3.6. By using similiar proof arguments of Lemma 1 in Shi et al. (2022), one can show that if the transition kernel $p_t(s'|\cdot,a) \in \Lambda(p,L)$, $t = 1,\ldots,T$, for any $a \in \mathcal{A}$ and $s' \in \mathcal{S}$, then we have $\sup_{q\in\mathcal{Q}^{(t+1)}:\|q\|_\infty\le 1} ||(\mathcal{P}_t^\pi)q||_{\Lambda(p)} \le L$ for any policy π, where $\|\cdot\|_{\Lambda(p)}$ is the Hölder norm defined in the space $\Lambda(p,L)$ for $t = 1,\ldots,T$. Therefore, the completeness assumption (Assumption 3.1) is satisfied. Assumption 3.7 specifies the uniform, projection error of Π_t in sup-norm. which is satisfied when taking the basis functions such as B-spline basis or wavelet. In these cases, one can take $\beta_Q = p/d$ so that Assumption 3.7 holds. See Section 3.1 of Chen & Christensen (2018).

3.2.1. A Slower Convergence Rate: No Realizability on Ratio Functions

Theorem 3.8. *Under Assumptions 3.1-3.2, 3.6-3.7, if we further assume that $K = o(\min\{\sqrt{n/(\log n \log T)}, n/(T^2 \log n \log T)\})$, $T = o(K^{\beta_Q})$, then we have W.H.P, $|\hat{\nu}(\pi) - \nu(\pi)| =$*

$$\mathcal{O}(T^2 K^{-\beta_Q} + \sqrt{\frac{T^3\kappa}{n}} + \frac{T^3 K \log n \log T}{n}). \quad (14)$$

To ensure the existence of K that satisfies the conditions listed in Theorem 3.8, we require

$$T \log T = o\left((n/\log n)^{\frac{\beta_Q}{1+2\beta_Q}}\right). \quad (15)$$

The bound (14) consists of three terms. The first term results from the bias of the approximation. The larger p in Assumption 3.6 is (i.e., the smoother the Q-functions are), the smaller this bias term would be. In the following corollaries, we discuss the dominant term under different scenarios. First, we consider the case where T is bounded.

Corollary 3.9. *Suppose Assumptions 3.1-3.2, 3.6-3.7 hold. We further assume that T is bounded.*

(i) If $1/2 < \beta_Q \leq 1$, then by taking $K \asymp \sqrt{n}/(\log n \log T)$, we have

$$|\hat{\nu}(\pi) - \nu(\pi)| = \mathcal{O}\left(n^{-\beta_Q/2} \log n\right), \text{ W.H.P.} \quad (16)$$

(ii) If $\beta_Q > 1$, then by taking $K \asymp (n/(\log n))^{1/(1+\beta_Q)}$, we have

$$|\hat{\nu}(\pi) - \nu(\pi)| = \mathcal{O}\left(n^{-1/2}\right), \text{ W.H.P.} \quad (17)$$

As shown in case (ii) of Corollary 3.9, when β_Q is large enough, i.e., Q functions are smooth enough, we can achieve the optimal convergence rate $n^{-1/2}$. This addresses the fundamental question **Q1**. When $1/2 < \beta_Q \leq 1$, by choosing K appropriately, our bound is faster than the optimal convergence rate $n^{-\beta_Q/(1+2\beta_Q)}$ for nonparametrically estimating the Q-functions (Chen & Qi, 2022). In other words, even without stricter smoothness assumption, our results indicate that the FQE estimator $\hat{\nu}(\pi)$ can still achieve a non-trivial fast rate, despite that $\hat{\nu}(\pi)$ is a simple plug-in estimator based on estimation of the nonparametric functions $\{Q_t^\pi\}$. In comparison to the bounds derived in Nguyen-Tang et al. (2021) and Ji et al. (2022), we have faster convergence rates with respect to n in both cases of Corollary 3.9. Unlike their analysis, we decompose the error term differently (see the decomposition in Section C), where we separate the first order term (25) and the bias-induced term (27). By choosing K appropriately, we can render the bias asymptotically ignorable to achieve optimal dependence on sample size n.

Next, we discuss the scenario when the horizon T can grow with n. For a neat presentation, we omit the log factors appearing in (18) and (19) in Corollary 3.10.

Corollary 3.10. *Under Assumptions 3.1, 3.2, 3.6 and 3.7, if $T \log T = \mathcal{O}\left((n/\log n)^{\beta/(1+2\beta)}\right)$, then we can take*

$$K \asymp \min\left\{(n/T)^{\frac{1}{1+\beta_Q}}, \sqrt{n}, n/T^2\right\}, \quad (18)$$

and we have W.H.P, $|\hat{\nu}(\pi) - \nu(\pi)| =$

$$\mathcal{O}\left(n^{-\frac{\beta_Q}{1+\beta_Q}} T^{\frac{2+3\beta_Q}{1+\beta_Q}} + T^{2+\beta_Q} n^{-\beta_Q} + T^2 n^{\frac{-\beta_Q}{2}} + \sqrt{\frac{T^3\kappa}{n}}\right). \quad (19)$$

Which of the four terms in (19) dominates depends on the choice of β_Q, n and T. When $\beta_Q > 1$, $\sqrt{T^3/n}$ is the first-order term and we obtain a polynomial dependence of order 1.5 for T. The remaining higher-order terms have a stronger dependence on T but still converge faster due to a better dependence on n. When $\beta_Q \leq 1$, the slowest dependence concerning n becomes $n^{-\beta_Q/2}$ and the corresponding term is quadratic in T (up to logarithmic orders). These results respond to **Q2** and provide some key understanding of the horizon dependence. In the next subsection, we show that FQE estimator $\hat{\nu}(\pi)$ can achieve better horizon dependence with an additional realizability condition of the ratio function, adding another layer of understanding to **Q2** by addressing **Q3**.

3.2.2. A Faster Convergence Rate: Realizability on Ratio Function

In this section, we show a better convergence guarantee for FQE estimator by adopting the following realizability condition on the probability ratio function $w_t^\pi, t = 1, \ldots, T$.

Assumption 3.11. There exists a constant $\beta_w > 1/2$ such that $\sup_t \|w_t^\pi - \Pi_t w_t^\pi\|_\infty \lesssim K^{-\beta_w}$ for $t = 1, \ldots, T$.

Assumption 3.11 imposes the smoothness conditions for the probability ratio functions and assumes that they can be approximated well by the basis functions that are used to model Q functions, as the number of basis functions K increases. Similar to the discussion in Section 3.2.1, this assumption can be fulfilled when $w_t^\pi(\cdot, a)$ belongs to the Hölder spaces for every $a \in \mathcal{A}$ and the basis functions are taken as B-spline basis or wavelet. *This implies that Assumption 3.11 is also very mild.*

Theorem 3.12. *Under Assumptions 3.1, 3.2, 3.6, 3.7 and 3.11, If we further assume $K = o(\min\{\sqrt{n/(\log n \log T)}, n/(T^2 \log n \log T)\})$, $T = o(K^{\beta_Q})$, we have W.H.P,*

$$|\hat{\nu}(\pi) - \nu(\pi)| = \mathcal{O}\left(\frac{T}{\sqrt{n}} + T^2 K^{-\beta_Q - \beta_w} + T^3 K^{-2\beta_Q} + T^3 K^{-\beta_Q}\sqrt{\frac{K \log n \log T}{n}} + \frac{T^3 K \log n \log T}{n}\right). \quad (20)$$

Similar to Theorem 3.8, we require T to satisfy (24) in order to ensure the existence of K. The first term (20) corresponds to the convergence error with respect to the first-order term E1 defined in (25) in Section C of the Appendix. The second term in (20) is the error term that takes into account of the projection error of the probability ratio function w_t^π. As illustrated in the bound (20) in Theorem 3.12, even though we do not explicitly model the probability ratio function w_t^π, there exists some "double-rate robustness" property in FQE that helps to obtain a faster rate convergence as long as the probability ratio function can be approximated well by sieve bases that are used to model Q functions. This responds to **Q3**. Next, we take a deeper look for the bound (20) under different scenarios. To simplify the discussion, we take $\beta_Q = \beta_w$, where we assume w_t^π and Q_t^π have the same degree of smoothness.

Corollary 3.13. *Under Assumptions 3.1, 3.2, 3.6, 3.7 and 3.11, , and further assume* $\beta_Q = \beta_w = \beta > 1/2$, $T \log T = \mathcal{O}\left((n/\log n)^{\beta/(1+2\beta)}\right)$, *by taking the optimal order of* K *such that*

$$K \asymp \{n/(\log n \log T)\}^{\frac{1}{1+2\beta}}, \tag{21}$$

we have W.H.P, $|\hat{\nu}(\pi) - \nu(\pi)| =$

$$\begin{cases} \mathcal{O}\left(\frac{T}{\sqrt{n}}\right), \text{if } T = \mathcal{O}\left(n^{\frac{2\beta-1}{4(1+2\beta)}} (\log n)^{\frac{-\beta}{1+2\beta}}\right), \\ \mathcal{O}\left(T^3 \left(\frac{n}{\log n \log T}\right)^{\frac{-2\beta}{1+2\beta}}\right), \text{otherwise.} \end{cases} \tag{22}$$

As we can see, (21) is equivalent to the optimal order for number of basis functions in common nonparametric regression, up to a logarithmic term (Chen & Christensen, 2018). If the number of horizon T is bounded, we can achieve the optimal convergence rate ($n^{-1/2}$) for $|\hat{\nu}(\pi) - \nu(\pi)|$ even though we estimate Q functions nonparametricly. Compared to Corollary 3.9, we do not require $\beta_Q > 1$ to achieve such optimal convergence rate, adding extra understanding to **Q1**. In Corollary 3.9, the number of basis functions needs to be chosen of order larger than (21) in order to remove the effect of approximation bias. Next, we focus on the horizon dependence (**Q2**). When the horizon T is allowed to grow with n, different regimes for upper bounds emerge depending on the relative order of T to n.

In the scenario where T grows relatively slowly compared to n (case 1 in (22)), the convergence exhibits a $n^{-1/2}$ dependence with respect to n, with a linear dependence on the horizon. To the best of our knowledge, this convergence rate aligns with the best-known rate for FQE in tabular settings (Yin & Wang, 2020) (necessarily parametric), despite our analysis is conducted under a much more challenging nonparametric setting. Conversely, when T grows faster, the error bound exhibits a cubic horizon dependence. In this case, it exhibits a $n^{-2\beta/(1+2\beta)}$ dependence on n, which is better than $n^{-1/2}$ dependence that we found in the aforementioned case. However, it is important to note that the overall rate is slower than the case of slowly growing T due to the stronger dependence on T. In Yin & Wang (2020), their higher order term exhibits a n^{-1} dependence with respect to n and a $T^{3/2}$ dependence with respect to T. In comparison to their tabular setting, our model introduces additional bias terms due to the continuous setting and nonparametric modeling. It remains unclear whether this order can be further improved to match theirs. We leave this as a subject for future research.

4. Simulation Study

We conduct a simulation study to illustrate the behavior of the error $|\hat{\nu}(\pi) - \nu(\pi)|$ with respect to n and T. The goal here is to provide empirical evidence of our theoretical results, and so we use a relatively simple simulation setup for the purpose of clear demonstration. Specifically, the data generative model is given as follows. The state variable is a one-dimensional continuous variable and the action is a binary variable, i.e., $\mathcal{A}_t = \{0, 1\}$ for all t. The initial state follows the uniform distribution within $[-2, 2]$. The transition dynamics are given by $S_{i,t+1} = (2A_{i,t} - 1)f(S_{i,t})$, where $f(x)$ is a function constructed from cubic B-spline basis functions as depicted in Figure 2 in Appendix. The behavior policy independently follows a Bernoulli distribution with mean $1/2$. The immediate reward $R_{i,t}$ is defined as $R_{i,t} = 2S_{i,t+1}$. Though the data follows the homogeneous MDP, we treat it as an inhomogeneous setting and use FQE discussed in Section 2.2 to estimate the value for different target policies. We evaluate the following two different target policies.

(a) $\pi_t(a = 1 \mid s) = 0.5$, for $s \in \mathcal{S}, t = 1, \ldots, T$. This policy is the same as the behavior one.

(b) $\pi_t(a = 1 \mid s) = \exp\{f(s)\}/\{1 + \exp(f(s))\}$, for $s \in \mathcal{S}, t = 1, \ldots, T$. This policy is smooth with respect to the state variable.

For every target policy, we evaluate values with $n = 200, 400, \ldots, 2000$, and $T = 20, 40, \ldots, 200$. We use cubic B-spline to construct basis functions at every step t. The knots are placed at evenly distributed percentiles of samples. The choice of K is tricky to decide in practice. We consider two different approaches to specify K. First, following (21) in Corollary 3.13, we consider taking $\beta = 2$, as the transition function has a continuous second derivative. And we try several different constants c with $K = cn^{1/5}$. We found that $c = 3$ yields an appropriate result and we fix $K = 3n^{1/5}$. For the second approach, we use leave-one-out cross-validation to decide K at every step.

Figure 1 summarize the simulation results for target policies (a) over 200 simulation replicates when K is selected by LOOCV. As we can see, we observe a roughly linear dependence between $|\hat{\nu}(\pi) - \nu(\pi)|$ and T, especially when T is relatively small compared to n. This is reflected in (22) in Corollary 3.13. The simulation results when K is selected using the criterion $K = 3n^{1/5}$ and the results for policy (b) can be found in Section A in Appendix.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

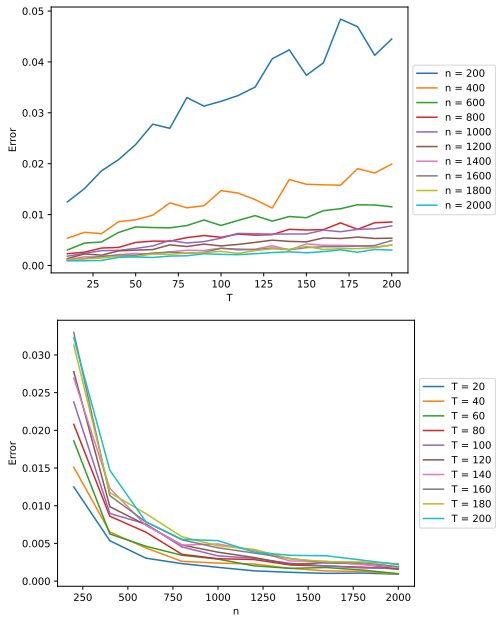

Figure 1. Simulation results for $|\hat{\nu}(\pi) - \nu(\pi)|$ when the target policy π is (a) and K is selected by LOOCV. The upper plot demonstrates the change of error along with the change of T, different curves represent different n. The bottom plot demonstrates the change of error along with the change of n, different curves represent different T.

References

Ai, C. and Chen, X. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

Blundell, R., Chen, X., and Kristensen, D. Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75(6):1613–1669, 2007.

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 Christensen, T. M. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics*, 9(1): 39–84, 2018.

Chen, X. and Pouzo, D. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1):277–321, 2012.

Chen, X. and Qi, Z. On well-posedness and minimax optimal rates of nonparametric q-function estimation in off-policy evaluation. In *International Conference on Machine Learning*, pp. 3558–3582. PMLR, 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.

Fu, J., Norouzi, M., Nachum, O., Tucker, G., Wang, Z., Novikov, A., Yang, M., Zhang, M. R., Chen, Y., Kumar, A., et al. Benchmarks for deep off-policy evaluation. *arXiv preprint arXiv:2103.16596*, 2021.

Geman, S. and Hwang, C.-R. Nonparametric maximum likelihood estimation by the method of sieves. *The annals of Statistics*, pp. 401–414, 1982.

Huang, J. Z. Projection estimation in multiple regression with application to functional anova models. *The annals of statistics*, 26(1):242–272, 1998.

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.

Min, Y., Wang, T., Zhou, D., and Gu, Q. Variance-aware off-policy evaluation with linear function approximation. *Advances in neural information processing systems*, 34: 7598–7610, 2021.

Nguyen-Tang, T., Gupta, S., Tran-The, H., and Venkatesh, S. Sample complexity of offline reinforcement learning with deep relu networks. *arXiv preprint arXiv:2103.06671*, 2021.

Shi, C., Zhu, J., Ye, S., Luo, S., Zhu, H., and Song, R. Off-policy confidence interval estimation with confounded markov decision process. *Journal of the American Statistical Association*, pp. 1–12, 2022.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Uehara, M., Shi, C., and Kallus, N. A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*, 2022.

Van Der Vaart, A. W. and Wellner, J. A. Weak convergence and empirical processes: with applications to statistics, 1996.

Voloshin, C., Le, H. M., Jiang, N., and Yue, Y. Empirical study of off-policy policy evaluation for reinforcement learning. *arXiv preprint arXiv:1911.06854*, 2019.

Wang, J., Qi, Z., and Wong, R. K. W. Projected state-action balancing weights for offline reinforcement learning. *The Annals of Statistics*, 51(4):1639–1665, 2023.

Xie, T., Ma, Y., and Wang, Y.-X. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *Advances in Neural Information Processing Systems*, 32, 2019.

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.

Zhang, R., Zhang, X., Ni, C., and Wang, M. Off-policy fitted q-evaluation with differentiable function approximators: Z-estimation and inference theory. In *International Conference on Machine Learning*, pp. 26713–26749. PMLR, 2022.

A. Additional Figures for the Simulation Results

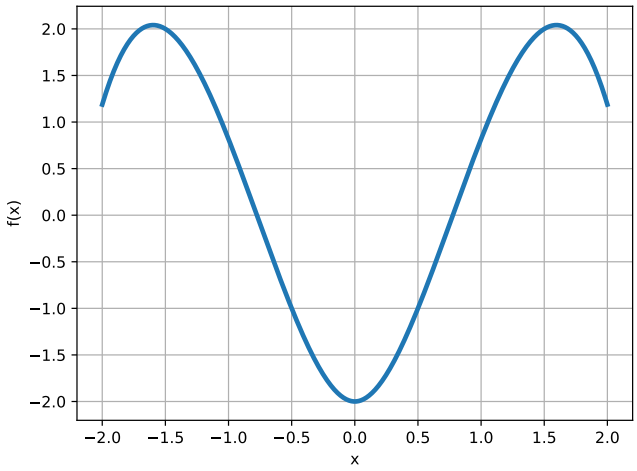

Figure 2. Illustration of the function f in the transition mechanism.

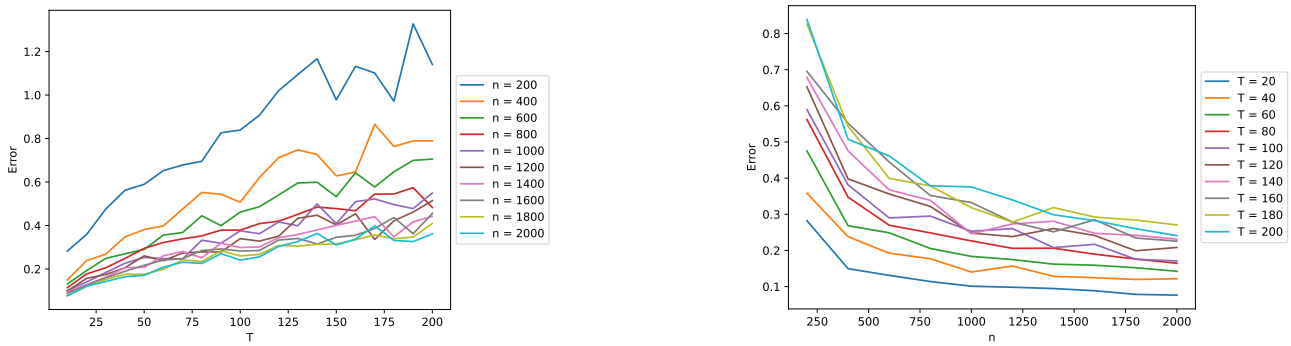

Figure 3. Simulation results for $|\hat{\nu}(\pi) - \nu(\pi)|$ when the target policy π is (b). See detailed description in Figure 1.

In addition to two policies evaluated in Section 4, we also evaluate the following policy

(c)

$$\pi_2(a = 1 \mid s) = \begin{cases} 1 & \text{if } f(s) > 0 \\ 0 & \text{otherwise} \end{cases}, \quad \text{for} \quad s \in \mathcal{S}, t = 1, \ldots, T.$$

This policy is a discontinuous function with respect to the state variable.

B. Detailed Theorem Statements

B.1. Parametric Setting

Theorem B.1. *Under Assumptions 3.1-3.2, we have*

$$|\hat{\nu}(\pi) - \nu(\pi)| = \mathcal{O}\left\{\sqrt{\frac{T^3}{n}\kappa} + T\sum_{t=1}^{T}\left[\left(1 + \sqrt{\frac{K \log n \log T}{n}}\right)^t - 1\right]\sqrt{\frac{\log n \log T}{n}}\right\}, \text{ W.H.P.} \tag{23}$$

If we further assume that

$$T = o([n/(\log n \log T)]^{1/2}),$$

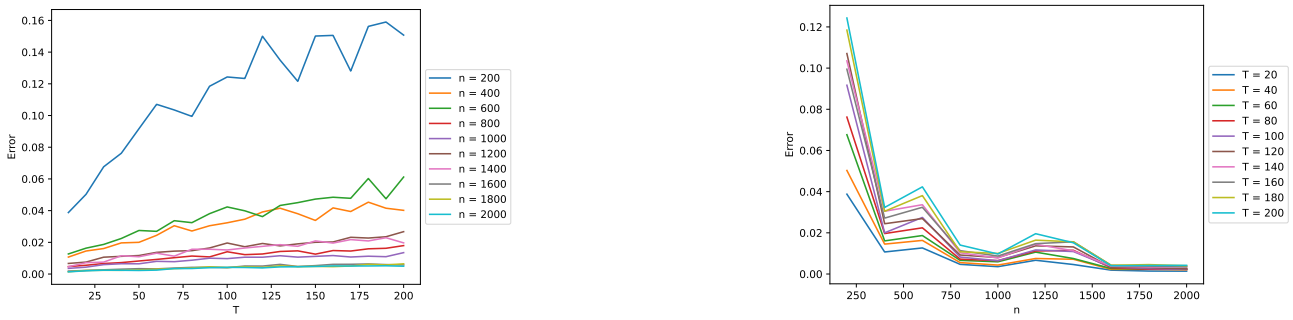

Figure 4. Simulation results for $|\hat{\nu}(\pi) - \nu(\pi)|$ when the target policy π is (a) and K is selected as $K = 3n^{1/5}$. The left plot demonstrate the change of error along with the change of T, different curves represent different n. The right plot demonstrates the change of error along with the change of n, different curves represent different T.

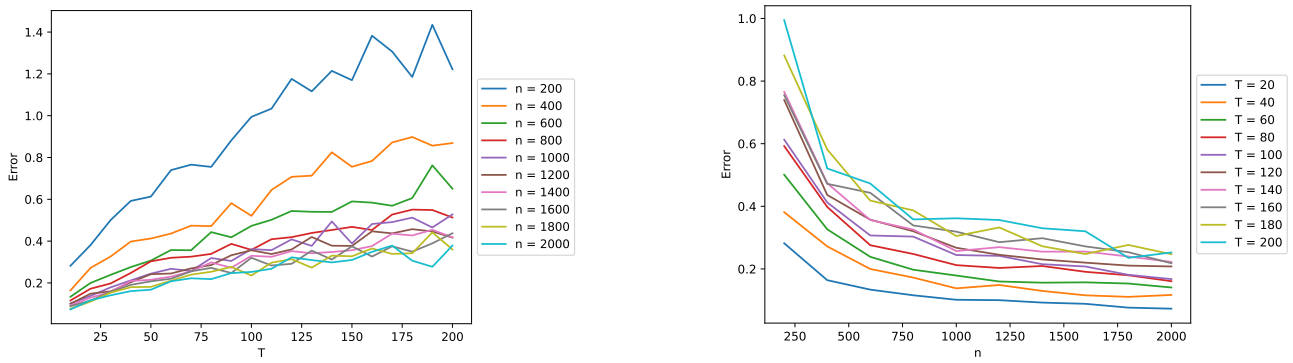

Figure 5. Simulation results for $|\hat{\nu}(\pi) - \nu(\pi)|$ when the target policy π is (b). See detailed description in Figure 4.

we have

$$|\hat{\nu}(\pi) - \nu(\pi)| = \mathcal{O}\left(\sqrt{\frac{T^3}{n}}\kappa + T^3\frac{\log n \log T}{n}\right), \text{ W.H.P.}$$

B.2. Nonparametric Setting

Theorem B.2. *Under Assumptions 3.1-3.2, 3.6-3.7, if $K = o(\sqrt{n/(\log n \log T)})$, then we have*

$$|\hat{\nu}(\pi) - \nu(\pi)| = \mathcal{O}_{\mathrm{p}}\Bigg\{T^2K^{-\beta_Q} + \sqrt{\frac{T^3}{n}}\kappa + T\sum_{t=1}^{T}\left[\left(1 + K^{-\beta_Q} + \sqrt{\frac{K\log n \log T}{n}}\right)^t - 1\right]\left(K^{-\beta_Q} + \sqrt{\frac{K\log n\log T}{n}}\right)\Bigg\}.$$

If we further assume that

$$K = o\{n/(T^2\log n \log T)\},\ T = o(K^{\beta_Q}) \tag{24}$$

we have

$$|\hat{\nu}(\pi) - \nu(\pi)| = \mathcal{O}_{\mathrm{p}}\left(T^2K^{-\beta_Q} + \sqrt{\frac{T^3}{n}}\kappa + \frac{T^3K\log n\log T}{n}\right).$$

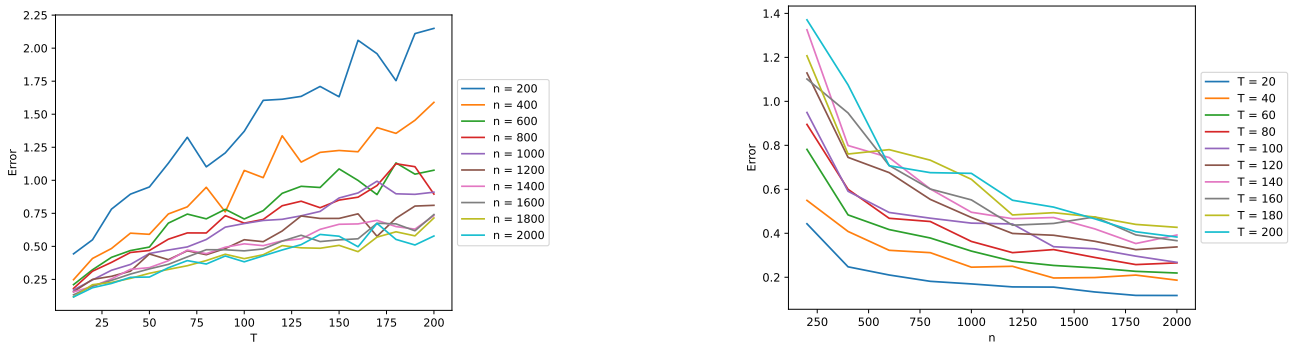

Figure 6. Simulation results for $|\hat{\nu}(\pi) - \nu(\pi)|$ when the target policy π is (c). See detailed description in Figure 1.

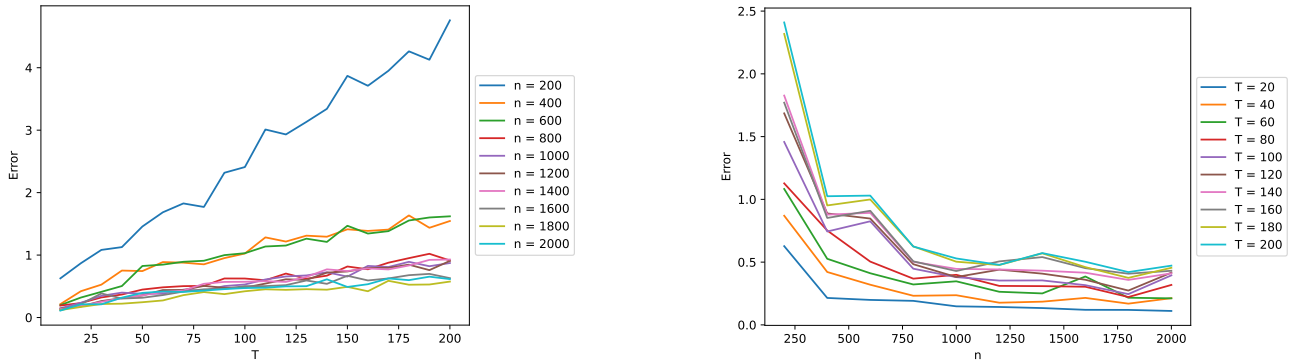

Figure 7. Simulation results for $|\hat{\nu}(\pi) - \nu(\pi)|$ when the target policy π is (c). See detailed description in Figure 4.

Theorem B.3. *Under Assumptions 3.1, 3.2, 3.6, 3.7 and 3.11, $K = \mathcal{O}(\sqrt{n/(\log n \log T)})$, we have*

$$|\hat{\nu}(\pi) - \nu(\pi)| = \mathcal{O}\Bigg\{ \frac{T}{\sqrt{n}} + K^{-\beta_w} \left[T^2 K^{-\beta_Q} + \sqrt{\frac{T^3}{n}} \right] + T \sum_{t=1}^{T} \left[\left(1 + K^{-\beta_Q} + \sqrt{\frac{K \log n \log T}{n}} \right)^t - 1 \right] \left(K^{-\beta_Q} + \sqrt{\frac{K \log n \log T}{n}} \right) \Bigg\}, \text{ W.H.P.}$$

If we further assume (24), we have

$$|\hat{\nu}(\pi) - \nu(\pi)| = \mathcal{O}\left(\frac{T}{\sqrt{n}} + T^2 K^{-\beta_Q - \beta_w} + T^3 K^{-2\beta_Q} + T^3 K^{-\beta_Q} \sqrt{\frac{K \log n \log T}{n}} + T^3 \frac{K \log n \log T}{n} \right), \text{ W.H.P.}$$

C. Proof of Main Theorems

In this section, we provide the proof for Theorem 3.8 and 3.12. The proof for theoretical results in the parametric setting can be derived as a special case by taking β_Q and β_w to be infinity.

First of all, we recall and introduce some notations. Let $\Sigma_t = \mathbb{E}[\phi_K(S_t, A_t)\phi_K(S_t, A_t)^\intercal] \in \mathbb{R}^K$, $\hat{\Sigma}_t = \frac{1}{n}\sum_{i=1}^n [\phi_K(S_{i,t}, A_{i,t})\phi_K(S_{i,t}, A_{i,t})^\intercal] \in \mathbb{R}^K$ and $\Sigma_{t,a} = \mathbb{E}[\psi_K(S_t)\psi_K(S_t)^\intercal \mid A_t = a]$. And we denote $\mathcal{D}_t$ as the collection of historical data up to time step t, i.e., $\mathcal{D}_t = \{S_1, A_1, R_1, \ldots, S_{t-1}, A_{t-1}, R_{t-1}, S_t, A_t\}$. Write $\langle \pi, Q \rangle(\cdot) = \sum_{a \in \mathcal{A}} \pi(a \mid \cdot) Q(\cdot, a)$.

Define $\mathcal{P}_t$ and $\hat{\mathcal{P}}_t^\pi$ as the population and estimated conditional expectation operators respectively, such that

$$(\mathcal{P}_t^\pi f)(s,a) = \mathbb{E}\left\{\sum_{a'} \pi_t(a' \mid S_{t+1}) f(S_{t+1}, a') \mid S_t = s, A_t = a\right\},$$
$$(\hat{\mathcal{P}}_t^\pi f)(s,a) = \phi_K(s,a)^\intercal (\hat{\Sigma}_t)^{-1} \left(\frac{1}{n}\sum_{i=1}^n \phi_K(S_{i,t}, A_{i,t}) \left[\sum_{a'} \pi_t(a' \mid S_{i,h+1}) f(S_{i,t+1}, a')\right]\right),$$

for $f \in \mathcal{Q}^{(t+1)}$, $t = 1, \ldots, T$. In addition, we define Π_t and $\hat{\Pi}_t^\pi$ as the population and estimated projection operators respectively, such that

$$\Pi_t g(s,a) = \phi_K(s,a)^\intercal (\Sigma_t)^{-1} \mathbb{E}\left[\phi_K(S_{i,t}, A_{i,t}) g(S_{i,t}, A_{i,t})\right],$$
$$(\hat{\Pi}_t g)(s,a) = \phi_K(s,a)^\intercal (\hat{\Sigma}_t)^{-1} \left(\frac{1}{n}\sum_{i=1}^n \phi_K(S_{i,t}, A_{i,t}) g(S_{i,t}, A_{i,t})\right),$$
$$(\tilde{\Pi}_t g)(s,a) = \phi_K(s,a)^\intercal (\Sigma_t)^{-1} \left(\frac{1}{n}\sum_{i=1}^n \phi_K(S_{i,h}, A_{i,h}) g(S_{i,h}, A_{i,h})\right)$$

for $g \in \mathcal{Q}^{(t)}$, $t = 1, \ldots, T$.

Then we have the following decomposition

$$\begin{aligned}
\nu(\pi) - \hat{\nu}(\pi) =& \mathcal{E}_1\{Q_1^\pi - \hat{Q}_1^\pi\} \\
=& \mathcal{E}_1\{Q_1^\pi - (\hat{\Pi}_1 R_1 + \hat{\mathcal{P}}_1^\pi Q_2^\pi) + \hat{\mathcal{P}}_1^\pi (Q_2^\pi - \hat{Q}_2^\pi)\} \\
=& \mathcal{E}_1\{[Q_1^\pi - (\hat{\Pi}_1 R_1 + \hat{\mathcal{P}}_1^\pi Q_2^\pi)] + \hat{\mathcal{P}}_1^\pi [Q_2^\pi - (\hat{\Pi}_t R_2 + \hat{\mathcal{P}}_2^\pi Q_3^\pi)] + \hat{\mathcal{P}}_1^\pi \hat{\mathcal{P}}_2^\pi (Q_3^\pi - \hat{Q}_3^\pi)\} \\
=& \cdots \\
=& \mathcal{E}_1\{[Q_1^\pi - (\hat{\Pi}_1 R_1 + \hat{\mathcal{P}}_1^\pi Q_2^\pi)] + \hat{\mathcal{P}}_1^\pi [Q_2^\pi - (\hat{\Pi}_t R_2 + \hat{\mathcal{P}}_2^\pi Q_3^\pi)] + \cdots + \hat{\mathcal{P}}_1^\pi \cdots \hat{\mathcal{P}}_{T-1}^\pi [Q_T^\pi - \hat{\mathcal{P}}_T^\pi R_T]\} \\
=& \mathcal{E}_1\{E_1\} + \mathcal{E}_1\{E_2\} + \mathcal{E}_1\{E_3\},
\end{aligned}$$

where

$$E_1 = \sum_{t=1}^T \left(\prod_{t'=0}^{t-1} \mathcal{P}_{t'}^\pi\right) \tilde{\Pi}_t \left[Q_t^\pi - (R_t + \langle \pi_t, Q_{t+1}^\pi \rangle)\right], \tag{25}$$
$$E_2 = \sum_{t=1}^T \left(\left[\prod_{t'=0}^{t-1} \hat{\mathcal{P}}_{t'}^\pi\right] \hat{\Pi}_t - \left[\prod_{t'=0}^{t-1} \mathcal{P}_{t'}^\pi\right] \tilde{\Pi}_t\right) \left[Q_t^\pi - (R_t + \langle \pi_t, Q_{t+1}^\pi \rangle)\right], \tag{26}$$
$$E_3 = \sum_{t=1}^T \left(\prod_{t'=0}^{t-1} \hat{\mathcal{P}}_{t'}^\pi\right) \left[Q_t - \hat{\Pi}_t Q_t\right], \tag{27}$$

In the following, we focus on these three terms one by one.

C.1. Bounding $\mathcal{E}_1\{E_1\}$

Note that

$$\begin{aligned}
\mathbb{E}(\mathcal{E}_1^\pi[E_1]) =& \sum_{t=1}^T \mathbb{E}\left(\mathcal{E}_1\left\{\left(\prod_{t'=0}^{t-1} \mathcal{P}_{t'}^\pi\right) \tilde{\Pi}_t \left[Q_t^\pi - (R_t + \langle \pi_t, Q_{t+1}^\pi \rangle)\right]\right\}\right) \\
=& \sum_{t=1}^T \mathbb{E}\left[\mathbb{E}\left(\mathcal{E}_1\left\{\left(\prod_{t'=0}^{t-1} \mathcal{P}_{t'}^\pi\right) \tilde{\Pi}_t \left[Q_t^\pi - (R_t + \langle \pi_t, Q_{t+1}^\pi \rangle)\right]\right\} \mid \mathcal{D}_t\right)\right] = 0.
\end{aligned}$$

Then it suffices to derive the bound for the variance of $\mathcal{E}_1\{E_1\}$.

$$\begin{aligned}
&\mathrm{Var}(\mathcal{E}_1^\pi[E_1(S_1,A_1)]) = \mathrm{Var}\left(\mathcal{E}_1^\pi\left\{\sum_{t=1}^T\left(\prod_{t'=0}^{t-1}\mathcal{P}_{t'}^\pi\right)\tilde{\Pi}_t\left[Q_t^\pi-(R_t+\langle\pi_t,Q_{t+1}^\pi\rangle)\right]\right\}\right)\\
&=\sum_{t=1}^T\mathbb{E}\left\{\mathrm{Var}\left(\mathcal{E}_1^\pi\left\{\left(\prod_{t'=0}^{t-1}\mathcal{P}_{t'}^\pi\right)\tilde{\Pi}_t\left[Q_t^\pi-(R_t+\langle\pi_t,Q_{t+1}^\pi\rangle)\right]\right\}\mid\mathcal{D}_t\right)\right\}\\
&=\sum_{t=1}^T\mathbb{E}\left\{\mathrm{Var}\left(\mathcal{E}_t^\pi\left\{\tilde{\Pi}_t\left[Q_t^\pi-(R_t+\langle\pi_t,Q_{t+1}^\pi\rangle)\right]\right\}\mid\mathcal{D}_t\right)\right\}\\
&=\sum_{t=1}^T\mathbb{E}\left\{\mathrm{Var}\left([\mathcal{E}_t^\pi\phi_K]^\intercal{\Sigma_t}^{-1}\left(\frac{1}{n}\sum_{i=1}^n\phi_K(S_{i,t},A_{i,t})\left[Q_t^\pi(S_{i,t},A_{i,t})-(R_{i,t}+\langle\pi_t,Q_{t+1}^\pi\rangle(S_{i,t+1}))\right]\right)\mid\mathcal{D}_t\right)\right\}\cdots\cdots(i)
\end{aligned}$$

The first inequality is due to Lemma D.1.

Next we consider bounding (i) under different conditions in Theorem 3.8 and Theorem 3.12.

- Under conditions in Theorem 3.8.

$$\begin{aligned}
(i)&=\sum_{t=1}^T\mathbb{E}\left\{\mathrm{Var}\left([\mathcal{E}_t^\pi\phi_K]^\intercal{\Sigma_t}^{-1}\left(\frac{1}{n}\sum_{i=1}^n\phi_K(S_{i,t},A_{i,t})\left[Q_t^\pi(S_{i,t},A_{i,t})-(R_{i,t}+\langle\pi_t,Q_{t+1}^\pi\rangle(S_{i,t+1}))\right]\right)\mid\mathcal{D}_t\right)\right\}\\
&=\sum_{t=1}^T\frac{1}{n}\mathbb{E}\left\{\mathrm{Var}\left([\mathcal{E}_t^\pi\phi_K]^\intercal{\Sigma_t}^{-1}\phi_K(S_{i,t},A_{i,t})\left(\left[Q_t^\pi(S_{i,t},A_{i,t})-(R_{i,t}+\langle\pi_t,Q_{t+1}^\pi\rangle(S_{i,t+1}))\right]\right)\mid\mathcal{D}_t\right)\right\}\\
&\lesssim\sum_{t=1}^T\frac{1}{n}(T-t+1)^2\mathbb{E}\left\{[\mathcal{E}_t^\pi\phi_K]^\intercal{\Sigma_t}^{-1}\phi_K(S_{i,t},A_{i,t})\phi_K^\intercal(S_{i,t},A_{i,t})\Sigma_t^{-1}[\mathcal{E}_t^\pi\phi_K]\right\}\\
&=\frac{1}{n}(T-t+1)^2\sum_{t=1}^T[\mathcal{E}_t^\pi\phi_K]^\intercal\Sigma_t^{-1}[\mathcal{E}_t^\pi\phi_K]\\
&\le\frac{T^3}{n}\kappa.
\end{aligned}$$

The second equality is due to the independence between different episodes and the last inequality is due to the definition of κ. Therefore, we have

$$\mathcal{E}_1(E_1)=\mathcal{O}_\mathrm{p}\left(\sqrt{\frac{T^3}{n}\kappa}\right).$$

- Under conditions in Theorem 3.12. Take $\Delta_t(s,a,r,s')=Q_t^\pi(s,a)-[r+\langle\pi,Q_{t+1}^\pi\rangle(s')]$.

$$
\begin{aligned}
(i) &= \sum_{t=1}^{T} \mathbb{E}\left\{ \mathrm{Var}\left([\mathcal{E}_t^\pi \phi_K]^\intercal {\Sigma_t}^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \phi_K(S_{i,t}, A_{i,t}) \Delta_t(S_{i,t}, A_{i,t}, S_{i,t+1}, R_{i,t}) \right) \mid \mathcal{D}_t \right) \right\} \\
&= \sum_{t=1}^{T} \frac{1}{n} \mathbb{E}\left\{ \mathrm{Var}\left(\left[\mathbb{E}\phi_K(S_t, A_t) \frac{\rho_t^\pi(S_t, A_t)}{\rho_t^b(S_t, A_t)} \right]^\intercal {\Sigma_t}^{-1} \phi_K(S_{i,t}, A_{i,t}) \left(\Delta_t(S_{i,t}, A_{i,t}, S_{i,t+1}, R_{i,t}) \right) \mid \mathcal{D}_t \right) \right\} \\
&\le 2 \sum_{t=1}^{T} \frac{1}{n} \mathbb{E}\left\{ \mathrm{Var}\left(\frac{\rho_t^\pi(S_{i,t}, A_{i,t})}{\rho_t^b(S_{i,t}, A_{i,t})} \left(\Delta_t(S_{i,t}, A_{i,t}, S_{i,t+1}, R_{i,t}) \right) \mid S_{i,t}, A_{i,t} \right) \right\} \\
&+ \frac{2}{n} \sum_{t=1}^{T} \mathbb{E}\left\{ \mathrm{Var}\left(\left[\phi_K(S_{i,t}, A_{i,t})^\intercal {\Sigma_t}^{-1} \mathbb{E}\phi_K(S_t, A_t) \frac{\rho_t^\pi(S_t, A_t)}{\rho_t^b(S_t, A_t)} - \frac{\rho_t^\pi(S_{i,t}, A_{i,t})}{\rho_t^b(S_{i,t}, A_{i,t})} \right] \Delta_t(S_{i,t}, A_{i,t}, S_{i,t+1}, R_{i,t}) \mid \mathcal{D}_t \right) \right\} \\
&\le 2 \sum_{t=1}^{T} \frac{1}{n} \sup_{s,a} \frac{\rho_t^\pi(s,a)}{\rho_t^b(s,a)} \mathbb{E}\left\{ \frac{\rho_t^\pi(S_{i,t}, A_{i,t})}{\rho_t^b(S_{i,t}, A_{i,t})} \mathrm{Var}\left(\left(\Delta_t(S_{i,t}, A_{i,t}, S_{i,t+1}, R_{i,t}) \right) \mid S_{i,t}, A_{i,t} \right) \right\} \\
&+ \frac{2}{n} \sum_{t=1}^{T} \mathbb{E}\left\{ \mathrm{Var}\left(\left[\phi_K(S_{i,t}, A_{i,t})^\intercal {\Sigma_t}^{-1} \mathbb{E}\phi_K(S_t, A_t) \frac{\rho_t^\pi(S_t, A_t)}{\rho_t^b(S_t, A_t)} - \frac{\rho_t^\pi(S_{i,t}, A_{i,t})}{\rho_t^b(S_{i,t}, A_{i,t})} \right] \Delta_t(S_{i,t}, A_{i,t}, S_{i,t+1}, R_{i,t}) \mid \mathcal{D}_t \right) \right\} \\
&\le 2 \sum_{t=1}^{T} \frac{1}{n} \left[\sup_{s,a} \frac{\rho_t^\pi(s,a)}{\rho_t^b(s,a)} \right] \mathbb{E}^\pi \left\{ \mathrm{Var}\left(\left(\left[Q_t^\pi(S_{i,t}, A_{i,t}) - (R_{i,t} + \langle \pi_t, Q_{t+1}^\pi \rangle (S_{i,t+1})) \right] \right) \mid S_{i,t}, A_{i,t} \right) \right\} \cdot \cdots \cdot (ii) \\
&+ \frac{2}{n} \sum_{t=1}^{T} \mathbb{E}\left\{ \mathrm{Var}\left(\left[\phi_K(S_{i,t}, A_{i,t})^\intercal {\Sigma_t}^{-1} \mathbb{E}\phi_K(S_t, A_t) \frac{\rho_t^\pi(S_t, A_t)}{\rho_t^b(S_t, A_t)} - \frac{\rho_t^\pi(S_{i,t}, A_{i,t})}{\rho_t^b(S_{i,t}, A_{i,t})} \right] \Delta_t(S_{i,t}, A_{i,t}, S_{i,t+1}, R_{i,t}) \mid \mathcal{D}_t \right) \right\} \\
&\qquad \cdots \cdots (iii).
\end{aligned}
$$

The first equality is due to the independence among episodes and the fact that

$$
\mathcal{E}_t^\pi \phi_K = \mathbb{E}\phi_K(S_t, A_t) \frac{\rho_t^\pi(S_t, A_t)}{\rho_t^b(S_t, A_t)}.
$$

By applying Lemma 3.4 in Yin & Wang (2020), we have

$$
\begin{aligned}
(ii) &\le \frac{2}{n} \left[\sup_{s,a,t} \frac{\rho_t^\pi(s,a)}{\rho_t^b(s,a)} \right] \sum_{t=1}^{T} \mathbb{E}^\pi \left\{ \mathrm{Var}\left(\left(\left[Q_t^\pi(S_{i,t}, A_{i,t}) - (R_{i,t} + \langle \pi_t, Q_{t+1}^\pi \rangle (S_{i,t+1})) \right] \right) \mid S_{i,t}, A_{i,t} \right) \right\} \\
&\le \frac{2}{n} \left[\sup_{s,a,t} \frac{\rho_t^\pi(s,a)}{\rho_t^b(s,a)} \right] \mathrm{Var}^\pi \left(\sum_{t=1}^{T} R_t \right) \le \frac{2}{n} \left[\sup_{s,a,t} \frac{\rho_t^\pi(s,a)}{\rho_t^b(s,a)} \right] T^2 \lesssim \frac{T^2}{n}.
\end{aligned}
$$

As for (iii), we have $|\Delta_t(S_{i,t}, A_{i,t}, S_{i,t+1}, R_{i,t})| \lesssim (T - t + 1)$ due to that $|R_t| \le 1$, and we obtain

$$
\begin{aligned}
(iii) \lesssim & \sum_{t=1}^{T} \frac{1}{n} (T - t + 2)^2 \mathbb{E}\left[\frac{\rho_t^\pi}{\rho_t^b}(S_t, A_t) - \Pi\left(\frac{\rho_t^\pi}{\rho_t^b} \right)(S_t, A_t) \right]^2 \\
\le & \frac{T^2}{n} \sum_{t=1}^{T} \left\| \frac{\rho_t^\pi}{\rho_t^b} - \Pi\left(\frac{\rho_t^\pi}{\rho_t^b} \right) \right\|_{\mathcal{L}_2}^2 \le \frac{T^3}{n} K^{-\beta_w}.
\end{aligned}
$$

Overall, we have

$$
\mathcal{E}_1(E_1) = \mathcal{O}_{\mathrm{p}}\left(\frac{T}{\sqrt{n}} + \sqrt{\frac{T^3 K^{-\beta_w}}{n}} \right).
$$

C.2. Bounding $\mathcal{E}_1(E_2)$

$$\begin{aligned}\mathcal{E}_1^\pi(E_2) &= \sum_{t=1}^T \mathcal{E}_1^\pi \left\{ \left(\left[\prod_{t'=0}^{t-1} \hat{\mathcal{P}}_{t'}^\pi\right] \hat{\Pi}_t - \left[\prod_{t'=0}^{t-1} \mathcal{P}_{t'}^\pi\right] \tilde{\Pi}_t \right) \left[Q_t^\pi - (R_t + \langle \pi_t, Q_{t+1}^\pi \rangle)\right] \right\} \\ &= \sum_{t=1}^T \mathcal{E}_1^\pi \left\{ \left(\prod_{t'=0}^{t-1} \hat{\mathcal{P}}_{t'}^\pi - \prod_{t'=0}^{t-1} \mathcal{P}_{t'}^\pi \right) \hat{\Pi}_t \left[Q_t^\pi - (R_t + \langle \pi_t, Q_{t+1}^\pi \rangle)\right] \right\} \cdots\cdots (i) \\ &\quad + \sum_{t=1}^T \mathcal{E}_1^\pi \left\{ \prod_{t'=0}^{t-1} \mathcal{P}_{t'}^\pi \left(\hat{\Pi}_t - \tilde{\Pi}_t\right) \left[Q_t^\pi - (R_t + \langle \pi_t, Q_{t+1}^\pi \rangle)\right] \right\} \cdots\cdots (ii)\end{aligned}$$

Let's first deal with term (ii). For every t, we have

$$\begin{aligned}&\mathbb{E}\left[\mathcal{E}_1^\pi \left\{ \prod_{t'=0}^{t-1} \mathcal{P}_{t'}^\pi \left(\hat{\Pi}_t - \tilde{\Pi}_t\right) \left[Q_t^\pi - (R_t + \langle \pi_t, Q_{t+1}^\pi \rangle)\right] \right\} \mid \mathcal{D}_t \right] = 0 \\ &\mathbb{E}\left[\mathcal{E}_1^\pi \left\{ \prod_{t'=0}^{t-1} \mathcal{P}_{t'}^\pi \left(\hat{\Pi}_t - \tilde{\Pi}_t\right) \left[Q_t^\pi - (R_t + \langle \pi_t, Q_{t+1}^\pi \rangle)\right] \right\} \right] = 0\end{aligned}$$

Therefore, we consider the variance of

$$\sum_{t=1}^T \mathcal{E}_1^\pi \left\{ \prod_{t'=0}^{t-1} \mathcal{P}_{t'}^\pi \left(\hat{\Pi}_t - \tilde{\Pi}_t\right) \left[Q_t^\pi - (R_t + \langle \pi_t, Q_{t+1}^\pi \rangle)\right] \right\}.$$

Using a similar argument as in Lemma D.1, we can decompose the variance as

$$\begin{aligned}&\mathrm{Var}\left[\sum_{t=1}^T \mathcal{E}_1^\pi \left\{ \prod_{t'=0}^{t-1} \mathcal{P}_{t'}^\pi \left(\hat{\Pi}_t - \tilde{\Pi}_t\right) \left[Q_t^\pi - (R_t + \langle \pi_t, Q_{t+1}^\pi \rangle)\right] \right\} \right] \\ =& \sum_{t=1}^T \mathbb{E}\left\{ \mathrm{Var}\left[\mathcal{E}_1^\pi \left\{ \prod_{t'=0}^{t-1} \mathcal{P}_{t'}^\pi \left(\hat{\Pi}_t - \tilde{\Pi}_t\right) \left[Q_t^\pi - (R_t + \langle \pi_t, Q_{t+1}^\pi \rangle)\right] \right\} \mid \mathcal{D}_t \right] \right\} \\ =& \sum_{t=1}^T \mathbb{E}\left\{ \mathrm{Var}\left[\mathcal{E}_t^\pi \left\{ \left(\hat{\Pi}_t - \tilde{\Pi}_t\right) \left[Q_t^\pi - (R_t + \langle \pi_t, Q_{t+1}^\pi \rangle)\right] \right\} \mid \mathcal{D}_t \right] \right\} \\ =& \sum_{t=1}^T \frac{1}{n} \mathbb{E}\left\{ \left[\mathbb{E}\left(\phi_K(S_t, A_t) \frac{\rho_t^\pi(S_t, A_t)}{\rho_t^b(S_t, A_t)} \right)^\intercal \Sigma_t^{-1/2} \left(\Sigma_t^{1/2} \hat{\Sigma}_t^{-1} \Sigma_t^{1/2} - I_K \right) \Sigma_t^{-1/2} \phi_K(S_t, A_t) \right]^2 \right. \\ &\qquad \left. \mathrm{Var}\left(Q_t^\pi(S_t, A_t) - (R_t + \langle \pi_t, Q_{t+1}^\pi(S_{t+1}) \rangle) \mid \mathcal{D}_t \right) \right\} \\ \le& \sum_{t=1}^T \frac{(T-t+2)^2}{n} \mathbb{E}\left\{ \left\| \mathbb{E}\left(\phi_K(S_t, A_t) \frac{\rho_t^\pi(S_t, A_t)}{\rho_t^b(S_t, A_t)} \right) \Sigma_t^{-1/2} \right\|_2^2 \left\| \Sigma_t^{1/2} \hat{\Sigma}_t^{-1} \Sigma_t^{1/2} - I_K \right\|_2^2 \left\| \Sigma_t^{-1/2} \phi_K(S_t, A_t) \right\|_2^2 \right\} \\ \le& \sum_{t=1}^T \frac{(T-t+2)^2}{n} [\mathcal{E}_t^\pi \phi_K]^\intercal \Sigma_t^{-1} [\mathcal{E}_t^\pi \phi_K] \left(\frac{\zeta_K^2 \log n}{n} \right) \zeta_K^2\end{aligned}$$

The last inequality is by applying Lemma D.2 and the definition of ζ_K. Therefore, we obtain

$$(ii) = \mathcal{O}\left(\sqrt{\frac{T^3 \zeta_K^4 \log n}{n^2}} \right), \text{ W.H.P.}$$

Next, we derive the bound for term (i).

First of all, notice that

$$\|\mathcal{P}_t^\pi f\|_\infty = \sup_{s,a} \mathbb{E}[f^\pi(S') \mid S=s, A=a] \le \sup_{s,a} \mathbb{E}[|f^\pi(S')| \mid S=s, A=a] \le \sup_{s'} f^\pi(s') \le \sup_{(s',a')} f(s',a').$$

Therefore

$$\|\mathcal{P}_t^\pi\|_\infty \le 1.$$

$$\begin{aligned}
&\mathcal{E}_1^\pi \left\{ \left(\prod_{t'=0}^{t-1} \hat{\mathcal{P}}_{t'}^\pi - \prod_{t'=0}^{t-1} \mathcal{P}_{t'}^\pi \right) \hat{\Pi}_t \left[Q_t^\pi - (R_t + \langle \pi_t, Q_{t+1}^\pi \rangle) \right] \right\} \\
=&\mathcal{E}_1^\pi \left\{ \left(\prod_{t'=0}^{t-1} (\mathcal{P}_{t'}^\pi + \hat{\mathcal{P}}_{t'}^\pi - \mathcal{P}_{t'}^\pi) - \prod_{t'=0}^{t-1} \mathcal{P}_{t'}^\pi \right) \hat{\Pi}_t \left[Q_t^\pi - (R_t + \langle \pi_t, Q_{t+1}^\pi \rangle) \right] \right\} \\
=&\mathcal{E}_1^\pi \left\{ \left(\sum_{(\delta_{t,0},\ldots,\delta_{t,t-1}) \in \{0,1\}^t \setminus \{0\}^t} (\mathcal{P}_0^\pi)^{1-\delta_{t,0}} (\hat{\mathcal{P}}_0^\pi - \mathcal{P}_{0'}^\pi)^{\delta_{t,0}} \cdots (\mathcal{P}_{t-1}^\pi)^{1-\delta_{t,t-1}} (\hat{\mathcal{P}}_{t-1}^\pi - \mathcal{P}_{t-1}^\pi)^{\delta_{t,t-1}} \right) \right. \\
&\qquad \left. \hat{\Pi}_t \left[Q_t^\pi - (R_t + \langle \pi_t, Q_{t+1}^\pi \rangle) \right] \right\} \\
\le& \left\| \left(\sum_{(\delta_{t,0},\ldots,\delta_{t,t-1}) \in \{0,1\}^t \setminus \{0\}^t} (\mathcal{P}_0^\pi)^{1-\delta_{t,0}} (\hat{\mathcal{P}}_0^\pi - \mathcal{P}_{0'}^\pi)^{\delta_{t,0}} \cdots (\mathcal{P}_{t-1}^\pi)^{1-\delta_{t,t-1}} (\hat{\mathcal{P}}_{t-1}^\pi - \mathcal{P}_{t-1}^\pi)^{\delta_{t,t-1}} \right) \right. \\
&\qquad \left. \hat{\Pi}_t \left[Q_t^\pi - (R_t + \langle \pi_t, Q_{t+1}^\pi \rangle) \right] \right\|_\infty \\
\le& \sum_{(\delta_{t,0},\ldots,\delta_{t,t-1}) \in \{0,1\}^t \setminus \{0\}^t} \|\mathcal{P}_0^\pi\|_\infty^{1-\delta_{t,0}} \|\hat{\mathcal{P}}_0^\pi - \mathcal{P}_0^\pi\|_\infty^{\delta_{t,0}} \cdots \|\mathcal{P}_{t-1}^\pi\|_\infty^{1-\delta_{t,t-1}} \|\hat{\mathcal{P}}_{t-1}^\pi - \mathcal{P}_{t-1}^\pi\|_\infty^{\delta_{t,t-1}} \\
&\qquad \|\hat{\Pi}_t \left[Q_t^\pi - (R_t + \langle \pi_t, Q_{t+1}^\pi \rangle) \right] \|_\infty \\
\le& \left\{ \left(\|\mathcal{P}_0^\pi\|_\infty + \|\hat{\mathcal{P}}_0^\pi - \mathcal{P}_0^\pi\|_\infty \right) \cdots \left(\|\mathcal{P}_{t-1}^\pi\|_\infty + \|\hat{\mathcal{P}}_{t-1}^\pi - \mathcal{P}_{t-1}^\pi\|_\infty \right) - \|\mathcal{P}_0^\pi\|_\infty \cdots \|\mathcal{P}_{t-1}^\pi\|_\infty \right\} \\
&\qquad \|\hat{\Pi}_t \left[Q_t^\pi - (R_t + \langle \pi_t, Q_{t+1}^\pi \rangle) \right] \|_\infty \\
\le& \left\{ \left(1 + \|\hat{\mathcal{P}}_0^\pi - \mathcal{P}_0^\pi\|_\infty \right) \cdots \left(1 + \|\hat{\mathcal{P}}_{t-1}^\pi - \mathcal{P}_{t-1}^\pi\|_\infty \right) - 1 \right\} \|\hat{\Pi}_t \left[Q_t^\pi - (R_t + \langle \pi_t, Q_{t+1}^\pi \rangle) \right] \|_\infty
\end{aligned}$$

From the argument in Section C.4, with probability at least $1 - 4T[(nK)^{-2} - c_1 \exp\left\{\beta \log n - c_2 \zeta_K^{-1} \sqrt{n}\right\}]$, we have that for any $t = 1, \ldots, T$

$$\begin{aligned}
&\|\hat{\mathcal{P}}_t^\pi - \mathcal{P}_t^\pi\|_\infty \\
&\qquad \lesssim \left(1 + \frac{\zeta_K^2 \sqrt{\log n \log K}}{\sqrt{n}} \right) \sup_{h \in \mathcal{Q}^{(t)}(1)} \|h - \Pi h\|_\infty + \frac{\zeta_K}{\sqrt{n}} + K^{\frac{-1}{2}} \frac{\zeta_K^2 \sqrt{\log n \log K}}{\sqrt{n}} + \frac{\zeta_K^3 \log n \log K}{n}.
\end{aligned}$$

In addition, we have

$$\begin{aligned}
&\|\hat{\Pi}_t \left[Q_t^\pi - (R_t + \langle \pi_t, Q_{t+1}^\pi \rangle) \right] \|_\infty \\
&\qquad = \mathcal{O}\left(\|Q_t^\pi - (R_t + \langle \pi_t, Q_{t+1}^\pi \rangle)\|_\infty \left[\frac{\zeta_K}{\sqrt{n}} + K^{\frac{-1}{2}} \frac{\zeta_K^2 \sqrt{\log n \log K}}{\sqrt{n}} + \frac{\zeta_K^3 \log n \log K}{n} \right] \right), \text{ W.H.P.}
\end{aligned}$$

Denote

$$\begin{aligned}
\xi_{1,n,K} &= \left(1 + \frac{\zeta_K^2 \sqrt{\log n \log K}}{\sqrt{n}} \right) \sup_{h \in \mathcal{Q}^{(t)}(1)} \|h - \Pi h\|_\infty \\
\xi_{2,n,K} &= \frac{\zeta_K}{\sqrt{n}} + K^{\frac{-1}{2}} \frac{\zeta_K^2 \sqrt{\log n \log K}}{\sqrt{n}} + \frac{\zeta_K^3 \log n \log K}{n}
\end{aligned}$$

- Without further condition on T, we bound $\mathcal{E}_1(E_2)$ by

$$\mathcal{E}_1(E_2) = \mathcal{O}\left(\sum_{t=1}^{T}\left\{[(\xi_{1,n,K}+\xi_{2,n,K})+1]^t - 1\right\}\xi_{2,n,K}\right), \text{ W.H.P.}$$

- Under the condition (15), we have

$$\xi_{1,n,K}+\xi_{2,n,K} < 1/T,$$

and therefore

$$\begin{aligned}\left\{\left(1+\|\hat{\mathcal{P}}_0^\pi - \mathcal{P}_0^\pi\|_\infty\right)\cdots\left(1+\|\hat{\mathcal{P}}_{t-1}^\pi - \mathcal{P}_{t-1}^\pi\|_\infty\right)-1\right\}\\ \lesssim t(\xi_{1,n,K}+\xi_{2,n,K}).\end{aligned}$$

Then the term E_2 is bounded by

$$\begin{aligned}\mathcal{E}_1^\pi(E_2) \lesssim& \sum_{t=1}^{T}\left\{t(\xi_{1,n,K}+\xi_{2,n,K})\|Q_t^\pi - (R_t+\langle\pi_t, Q_{t+1}^\pi\rangle)\|_\infty\xi_{2,n,K}\right\}\\ &+\sum_{t=1}^{T}\sup_{s,a}\frac{\rho_t^\pi(s,a)}{\rho_t^b(s,a)}\frac{\zeta_K^2\log K\log n}{n}\left\|Q_t^\pi - (R_t+\langle\pi_t,Q_{t+1}^\pi\rangle)\right\|_\infty\\ &\lesssim T^2(\xi_{1,n,K}+\xi_{2,n,K})\xi_{2,n,K}\left[\sup_{t=1,\ldots,T}\|Q_t^\pi - (R_t+\langle\pi_t,Q_{t+1}^\pi\rangle)\|_\infty\right],\end{aligned}$$

with probability at least $1-4T[(nK)^{-2}-c_1\exp\left\{\beta\log n - c_2\zeta_K^{-1}\sqrt{n}\right\}]$. And we have

$$\mathcal{E}_1^\pi(E_2) = \mathcal{O}\big(T^3(\xi_{1,n,K}+\xi_{2,n,K})\xi_{2,n,K}\big), \text{ W.H.P.}$$

.

C.3. Bounding $\mathcal{E}_1(E_3)$

$$\begin{aligned}\mathcal{E}_1(E_3) =&\mathcal{E}_1\left\{\sum_{t=1}^{T}\left(\prod_{t'=0}^{t-1}\hat{\mathcal{P}}_{t'}^\pi\right)\left[Q_t - \hat{\Pi}_t Q_t\right]\right\}\\ =&\mathcal{E}_1\left\{\sum_{t=1}^{T}\left(\prod_{t'=0}^{t-1}\mathcal{P}_{t'}^\pi\right)\left[Q_t - \hat{\Pi}_t Q_t\right]\right\} + \mathcal{E}_1\left\{\sum_{t=1}^{T}\left(\prod_{t'=0}^{t-1}\hat{\mathcal{P}}_{t'}^\pi - \prod_{t'=0}^{t-1}\mathcal{P}_{t'}^\pi\right)\left[Q_t - \hat{\Pi}_t Q_t\right]\right\}\end{aligned} \tag{28}$$

For the first component in (28),

$$\mathcal{E}_1^\pi\left\{\sum_{t=1}^{T}\left(\prod_{t'=0}^{t-1}\mathcal{P}_{t'}^\pi\right)\left[Q_t - \hat{\Pi}_t Q_t\right]\right\} = \sum_{t=1}^{T}\mathcal{E}_t^\pi\left[Q_t - \hat{\Pi}_t Q_t\right]$$

$$\begin{aligned}\left|\sum_{t=1}^{T}\mathcal{E}_t^\pi\left[Q_t - \hat{\Pi}_t Q_t\right]\right| =& \left|\sum_{t=1}^{T}\mathcal{E}_t^\pi\left[\Pi Q_t - \hat{\Pi}_t Q_t + Q_t - \Pi Q_t\right]\right|\\ \le& \left|\sum_{t=1}^{T}\mathcal{E}_t^\pi\left[\Pi Q_t - \hat{\Pi}_t Q_t\right]\right| + \left|\sum_{t=1}^{T}\mathcal{E}_t^\pi\left[Q_t - \Pi Q_t\right]\right|\\ \le& (i)+(ii)\end{aligned}$$

For term (i), note that

$$\Pi_t Q_t(s,a) = \phi_K(s,a)^\intercal C_{t,K}$$

where $C_{t,K} = (\Sigma_t)^{-1}\mathbb{E}\phi_K(S_t,A_t)Q_t^\pi(S_t,A_t)$, and

$$\begin{aligned}
&\mathbb{E}\phi_K(S_{i,t},A_{i,t})\left[Q_t^\pi(S_{i,t},A_{i,t}) - \phi_K^\intercal(S_{i,t},A_{i,t})^\intercal C_{t,K}\right]\\
=&\mathbb{E}\phi_K(S_{i,t},A_{i,t})\left[Q_t^\pi(S_{i,t},A_{i,t}) - \phi_K^\intercal(S_{i,t},A_{i,t})^\intercal(\Sigma_t)^{-1}\mathbb{E}\phi_K(S_t,A_t)Q_t^\pi(S_t,A_t)\right]\\
=&\mathbb{E}\phi_K(S_{i,t},A_{i,t})\left[Q_t^\pi(S_{i,t},A_{i,t})\right] - \phi_K(S_t,A_t)\left[Q_t^\pi(S_t,A_t)\right] = 0.
\end{aligned}$$

we have

$$\begin{aligned}
&\left|\sum_{t=1}^T \mathcal{E}_t^\pi\left[\Pi Q_t - \hat{\Pi}_t Q_t\right]\right|\\
\le&\left|\sum_{t=1}^T \mathcal{E}_t^\pi\left[\phi_K^\intercal(\cdot,\cdot)(\hat{\Sigma}_t)^{-1}\left\{\frac{1}{n}\sum_{i=1}^n \phi_K(S_{i,t},A_{i,t})Q_t^\pi(S_{i,t},A_{i,t})\right\} - \Pi_t Q_t^\pi\right]\right|\\
=&\left|\sum_{t=1}^T \mathcal{E}_t^\pi\left[\phi_K^\intercal(\cdot,\cdot)(\hat{\Sigma}_t)^{-1}\left\{\frac{1}{n}\sum_{i=1}^n \phi_K(S_{i,t},A_{i,t})[Q_t^\pi(S_{i,t},A_{i,t}) - \phi_K^\intercal(S_{i,t},A_{i,t})C_{t,K}]\right.\right.\\
&\qquad\left.\left.-\mathbb{E}\phi_K(S_{i,t},A_{i,t})\left[Q_t^\pi(S_{i,t},A_{i,t}) - \phi_K^\intercal(S_{i,t},A_{i,t})^\intercal C_{t,K}\right]\right\}\right]\Big|\\
=&\left|\sum_{t=1}^T \mathcal{E}_t^\pi\left[\phi_K^\intercal(\cdot,\cdot)(\hat{\Sigma}_t)^{-1}\left\{\frac{1}{n}\sum_{i=1}^n \phi_K(S_{i,t},A_{i,t})[Q_t^\pi(S_{i,t},A_{i,t}) - \Pi_t Q_t^\pi(S_{i,t},A_{i,t})]\right.\right.\\
&\qquad\left.\left.-\mathbb{E}\phi_K(S_{i,t},A_{i,t})\left[Q_t^\pi(S_{i,t},A_{i,t}) - \Pi_t Q_t^\pi(S_{i,t},A_{i,t})\right]\right\}\right]\Big|.
\end{aligned}$$

Using similiar argument in deriving the bound (29), replacing $\mathcal{P}_t^\pi f$ with $Q_t^\pi - \Pi_t Q_t^\pi / \|Q_t^\pi - \Pi_t Q_t^\pi\|_\infty$, we have

$$\begin{aligned}
&\left|\sum_{t=1}^T \mathcal{E}_t^\pi\left[\Pi Q_t - \hat{\Pi}_t Q_t\right]\right|\\
\le&\mathcal{O}\left(\sum_{t=1}^T \mathcal{E}_t^\pi\left\|\mathcal{E}_1\{\phi_K^\intercal\}\Sigma_t^{-1/2}\right\|_2 \zeta_K\sqrt{\frac{\log n\log T}{n}}\|Q_t^\pi - \Pi_t Q_t^\pi\|_\infty\right)\\
=&\mathcal{O}\left(T^2\kappa\frac{\zeta_K\sqrt{\log n\log T}}{\sqrt{n}}K^{-\beta_Q}\right),\ \text{W.H.P.}
\end{aligned}$$

For term (ii), we derive the bound under different conditions.

- Under conditions in Theorem 3.8.

$$(ii) \le \sum_{t=1}^T \|Q_t^\pi - \Pi_t Q_t^\pi\|_\infty \le T^2 K^{-\beta_Q}.$$

- Under conditions in Theorem 3.12.

  First of all, note that

$$\mathbb{E}\left\{\phi_K(S_t,A_t)\left[Q_t(S_t,A_t) - \Pi Q_t(S_t,A_t)\right]\right\} = \mathbf{0}$$

due to the definition of Π_t. Then we have

$$\begin{aligned}
&\sum_{t=1}^{T} \mathcal{E}_t^\pi \left[Q_t^\pi - \Pi Q_t^\pi\right] \\
=&\sum_{t=1}^{T} \mathbb{E}\left\{\frac{\rho_t^\pi(S_t,A_t)}{\rho_t^b(S_t,A_t)}\left[Q_t(S_t,A_t) - \Pi Q_t(S_t,A_t)\right]\right\} \\
=&\sum_{t=1}^{T} \mathbb{E}\left\{\Pi_t\left\{\frac{\rho_t^\pi}{\rho_t^b}\right\}(S_t,A_t)\left[Q_t(S_t,A_t) - \Pi Q_t(S_t,A_t)\right]\right\} \\
&+\sum_{t=1}^{T} \mathbb{E}\left\{\left(\frac{\rho_t^\pi(S_t,A_t)}{\rho_t^b(S_t,A_t)} - \Pi_t\left\{\frac{\rho_t^\pi}{\rho_t^b}\right\}(S_t,A_t)\right)\left[Q_t(S_t,A_t) - \Pi Q_t(S_t,A_t)\right]\right\} \\
=&0+\sum_{t=1}^{T} \mathbb{E}\left\{\left(\frac{\rho_t^\pi(S_t,A_t)}{\rho_t^b(S_t,A_t)} - \Pi_t\left\{\frac{\rho_t^\pi}{\rho_t^b}\right\}(S_t,A_t)\right)\left[Q_t(S_t,A_t) - \Pi Q_t(S_t,A_t)\right]\right\} \\
\le&\sum_{t=1}^{T} \left\|\frac{\rho_t^\pi}{\rho_t^b} - \Pi_t\left[\frac{\rho_t^\pi}{\rho_t^b}\right]\right\|_{\mathcal{L}_2} \|Q_t - \Pi_t Q_t\|_{\mathcal{L}_2} \\
\le&T^2 K^{-\beta_w} K^{-\beta_Q}.
\end{aligned}$$

The last equality is due to the fact that $\Pi_t\{w_t\}(s,a) = \phi_K(s,a)^\intercal\omega$ for some $\omega \in \mathbb{R}^K$.

For the second component in (28), using a similar idea as in bounding E_2, we have

$$\begin{aligned}
\sum_{t=1}^{T}\left(\prod_{t'=0}^{t-1}\hat{\mathcal{P}}_{t'}^\pi - \prod_{t'=0}^{t-1}\mathcal{P}_{t'}^\pi\right)\left[Q_t - \hat{\mathcal{P}}_t Q_t\right] &\le \sum_{t=1}^{T}\mathcal{O}[t(\xi_{1,n,K}+\xi_{2,n,K})]\left\|Q_t - \hat{\mathcal{P}}_t Q_t\right\|_\infty \\
&\le \sum_{t=1}^{T}\mathcal{O}\left(t(\xi_{1,n,K}+\xi_{2,n,K})\left(1+\frac{\zeta_K^2\sqrt{\log n \log K}}{\sqrt{n}}\right)\|Q_t - \Pi Q_t\|_\infty\right) \\
&\le \mathcal{O}\left\{(\xi_{1,n,K}+\xi_{2,n,K})\left(1+\frac{\zeta_K^2\sqrt{\log n \log K}}{\sqrt{n}}\right)\left(\sum_{t=1}^{T} t\|Q_t - \Pi Q_t\|_\infty\right)\right\} \\
&\le \mathcal{O}\big(T^3(\xi_{1,n,K}+\xi_{2,n,K})K^{-\beta_Q}\big), \text{ W.H.P.}
\end{aligned}$$

The last inequality is due to the condition of ζ_K.

By combining results from Section C.1, C.2 and C.3, we obtain the bounds in Theorem 3.8 and 3.12.

C.4. Bounding $\|\hat{\mathcal{P}}_t^\pi - \mathcal{P}_t^\pi\|_\infty$

$$\begin{aligned}
&[\hat{\mathcal{P}}_t^\pi]f(s,a) - [\mathcal{P}_t^\pi]f(s,a) \\
=&\phi_K(s,a)^\intercal(\hat{\Sigma})^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n}\phi_K(S_i,A_i)\mathcal{P}_t^\pi f(S_i,A_i)\right\} - \Pi(\mathcal{P}_t^\pi f)(s,a) + \Pi(\mathcal{P}_t^\pi f)(s,a) - (\mathcal{P}_t^\pi f)(s,a) \\
&+\phi_K(s,a)^\intercal(\hat{\Sigma})^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n}\phi_K(S_i,A_i)\left[f^\pi(S_i') - \mathcal{P}_t^\pi f(S_i,A_i)\right]\right\} \\
=&I + II
\end{aligned}$$

where I indicates the bias term and II indicates the variance term. For the following, we constrain $f \in \mathcal{Q}^{(t+1)}$ with $\|f\|_\infty \le 1$. Here, we omit te subscirpt t for Π_t and Σ_t.

C.4.1. BOUNDING THE BIAS TERM

We first look at the bias term. From the definition of Πh_0, we know that

$$\Pi h_0(s,a) = \phi_K(s,a)^\intercal(\Sigma)^{-1}\mathbb{E}\left\{\phi_K(S,A)h_0(S,A)\right\}$$

Take $C_{K,f}$ as

$$C_{K,f} = (\Sigma)^{-1}\mathbb{E}\left\{\phi_K(S,A)\mathcal{P}_t^\pi f(S,A)\right\}$$

and Define

$$\zeta_K = \sup_{s,a}\|\Sigma^{-1/2}\phi_K(s,a)\|_2.$$

$$\begin{aligned}
&\phi_K(s,a)^\intercal(\hat{\Sigma})^{-1}\left\{\frac{1}{n}\sum_{i=1}^n \phi_K(S_i,A_i)\mathcal{P}_t^\pi f(S_i,A_i)\right\} - \Pi(\mathcal{P}_t^\pi f)(s,a)\\
=&\phi_K(s,a)^\intercal(\Sigma)^{-1}\left\{\frac{1}{n}\sum_{i=1}^n \phi_K(S_i,A_i)\left(\mathcal{P}_t^\pi f(S_i,A_i) - \phi_k(S_i,A_i)^\intercal C_{K,f}\right)\right.\\
&\quad\left.-\mathbb{E}\left[\phi_K(S,A)\left(\mathcal{P}_t^\pi f(S,A) - \phi_k(S,A)^\intercal C_{K,f}\right)\right]\right\}\\
&+\phi_K(s,a)^\intercal\Sigma^{-1/2}(\Sigma^{1/2}\hat{\Sigma}^{-1}\Sigma^{1/2} - I_K)\Sigma^{-1/2}\left\{\frac{1}{n}\sum_{i=1}^n \phi_K(S_i,A_i)\left(\mathcal{P}_t^\pi f(S_i,A_i) - \phi_k(S_i,A_i)^\intercal C_{K,f}\right)\right.\\
&\quad\left.-\mathbb{E}\left[\phi_K(S,A)\left(\mathcal{P}_t^\pi f(S,A) - \phi_k(S,A)^\intercal C_{K,f}\right)\right]\right\}
\end{aligned}$$

Take $\mathcal{Q}^{(t+1)}(1) = \{f\in\mathcal{Q}^{(t+1)} : \|f\|_\infty \le 1\}$, we have

$$\begin{aligned}
&\sup_{f\in\mathcal{Q}^{(t+1)}(1)}\left\|\phi_K(\cdot,\cdot)^\intercal(\hat{\Sigma})^{-1}\left\{\frac{1}{n}\sum_{i=1}^n \phi_K(S_i,A_i)\mathcal{P}_t^\pi f(S_i,A_i)\right\} - \Pi(\mathcal{P}_t^\pi f)\right\|_\infty\\
\le&\left(\sup_{s,a}\left\|\Sigma^{-1/2}\phi_K(s,a)\right\|_2 + \sup_{s,a}\left\|\Sigma^{-1/2}\phi_K(s,a)\right\|_2\|\Sigma^{1/2}\hat{\Sigma}^{-1}\Sigma^{1/2} - I_K\|\right)\\
&\sup_{f\in\mathcal{Q}^{(t+1)}(1)}\left\|\Sigma^{\frac{-1}{2}}\left\{\frac{1}{n}\sum_{i=1}^n \phi_K(S_i,A_i)\left(\mathcal{P}_t^\pi f(S_i,A_i) - \phi_k(S_i,A_i)^\intercal C_{K,f}\right)\right.\right.\\
&\qquad\left.\left.-\mathbb{E}\left[\phi_K(S,A)\left(\mathcal{P}_t^\pi f(S,A) - \phi_k(S,A)^\intercal C_{K,f}\right)\right]\right\}\right\|_2\\
\le&\zeta_K(1 + \|\Sigma^{1/2}\hat{\Sigma}^{-1}\Sigma^{1/2} - I_K\|)\\
&\sup_{h\in\mathcal{Q}^{(t)}:\|h\|_\infty\le\sup_{h\in\mathcal{Q}^{(t)}}\|h-\Pi h\|_\infty}\left\|\Sigma^{\frac{-1}{2}}\left\{\frac{1}{n}\sum_{i=1}^n \phi_K(S_i,A_i)h(S_i,A_i) - \mathbb{E}\left[\phi_K(S,A)h(S,A)\right]\right\}\right\|_2
\end{aligned}$$

And it remains to bound $\|\Sigma^{1/2}\hat{\Sigma}^{-1}\Sigma^{1/2} - I_K\|$ and

$$\sup_{h\in\mathcal{Q}(1)}\left\|\Sigma^{\frac{-1}{2}}\left\{\frac{1}{n}\sum_{i=1}^n \phi_K(S_i,A_i)h(S_i,A_i) - \mathbb{E}\left[\phi_K(S,A)h(S,A)\right]\right\}\right\|_2.$$

Note that under Assumption 3.6, by Theorem 2.7.3 in Van Der Vaart & Wellner (1996), there exists a constant $B>0$ such that $\mathcal{N}(\mathcal{Q}^{(t)}(1),\|\cdot\|_\infty,\epsilon)\le\exp(B\epsilon^{-d/(p)})$ for $t=1,\dots,T$.

Therefore, by Lemma D.3 and D.2, with probability at least $1-2(nK)^{-2}$, we have

$$\begin{aligned}
&\sup_{f\in\mathcal{Q}^{(t+1)}(1)}\left\|\phi_K(\cdot,\cdot)^\intercal(\hat{\Sigma})^{-1}\left\{\frac{1}{n}\sum_{i=1}^n \phi_K(S_i,A_i)\mathcal{P}_t^\pi f(S_i,A_i)\right\} - \Pi(\mathcal{P}_t^\pi f)\right\|_\infty\\
\lesssim&\frac{\zeta_K^2\sqrt{\log n\log K}}{\sqrt{n}}\sup_{h\in\mathcal{Q}^{(t)}}\|h-\Pi h\|_\infty. \qquad (29)
\end{aligned}$$

And therefore with probability at least $1-2(nK)^{-2}$,

$$\sup_{f\in\mathcal{Q}^{(t+1)}}\sup_{s,a}\left|\phi_K(s,a)^\intercal(\hat{\Sigma})^{-1}\left\{\frac{1}{n}\sum_{i=1}^n\phi_K(S_i,A_i)\mathcal{P}_t^\pi f(S_i,A_i)\right\}-\Pi(\mathcal{P}_t^\pi f)(s,a)+\Pi(\mathcal{P}_t^\pi f)(s,a)-(\mathcal{P}_t^\pi f)(s,a)\right| \lesssim\left(1+\frac{\zeta_K^2\sqrt{\log n\log K}}{\sqrt{n}}\right)\sup_{h\in\mathcal{Q}^{(t)}(1)}\|h-\Pi h\|_\infty.$$

C.4.2. Bounding the variance term

$$\begin{aligned}II(s,a)=&\phi_K(s,a)^\intercal(\hat{\Sigma})^{-1}\left\{\frac{1}{n}\sum_{i=1}^n\phi_K(S_i,A_i)\left[f^\pi(S_i')-\mathcal{P}_t^\pi f(S_i,A_i)\right]\right\}\\=&\phi_K(s,a)^\intercal\Sigma^{-1}\left\{\frac{1}{n}\sum_{i=1}^n\phi_K(S_i,A_i)\left[f^\pi(S_i')-\mathcal{P}_t^\pi f(S_i,A_i)\right]\right\}\\&+\phi_K(s,a)^\intercal\Sigma^{-1/2}\left(\Sigma^{1/2}\Sigma^{-1}\Sigma^{1/2}-I_K\right)\Sigma^{-1/2}\left\{\frac{1}{n}\sum_{i=1}^n\phi_K(S_i,A_i)\left[f^\pi(S_i')-\mathcal{P}_t^\pi f(S_i,A_i)\right]\right\}.\end{aligned}\tag{30}$$

By using the same argument in Lemma D.3, we have with probability at least $1-(nK)^{-2}$,

$$\sup_{f\in\mathcal{Q}^{(t+1)}(1)}\left\|\Sigma^{-1/2}\left\{\frac{1}{n}\sum_{i=1}^n\phi_K(S_i,A_i)\left[f^\pi(S_i')-\mathcal{P}_t^\pi f(S_i,A_i)\right]\right\}\right\|\lesssim\frac{\zeta_K\sqrt{\log n\log K}}{\sqrt{n}}.$$

Then the second term in (30) can be bounded by

$$\begin{aligned}&\sup_{s,a}\sup_{f\in\mathcal{Q}^{(t+1)}}\left|\phi_K(s,a)^\intercal\Sigma^{-1/2}\left(\Sigma^{1/2}\Sigma^{-1}\Sigma^{1/2}-I_K\right)\Sigma^{-1/2}\left\{\frac{1}{n}\sum_{i=1}^n\phi_K(S_i,A_i)\left[f^\pi(S_i')-\mathcal{P}_t^\pi f(S_i,A_i)\right]\right\}\right|\\\le&\sup_{s,a}\|\phi_K^\intercal(s,a)\Sigma^{-1/2}\|_2\left\|\Sigma^{1/2}\Sigma^{-1}\Sigma^{1/2}-I_K\right\|\sup_{f\in\mathcal{Q}^{(t+1)}(1)}\left\|\Sigma^{-1/2}\left\{\frac{1}{n}\sum_{i=1}^n\phi_K(S_i,A_i)\left[f^\pi(S_i')-\mathcal{P}_t^\pi f(S_i,A_i)\right]\right\}\right\|\\\le&\zeta_K\frac{\zeta_K\sqrt{\log n\log K}}{\sqrt{n}}\frac{\zeta_K\sqrt{\log n\log K}}{\sqrt{n}}\end{aligned}$$

with probability at least $1-2(nK)^{-2}$. For the following, we focus on bounding the first term in (30).

Let $\mathcal{X}_n\subset\mathcal{S}\times\mathcal{A}$ be a grid of finitely many points such that for each $(s,a)\in\mathcal{S}\times\mathcal{A}$ there exists a $\overline{(s,a)}_n(s,a)\in\mathcal{X}_n$ such that $\|(s,a)-\overline{(s,a)}_n(s,a)\|\lesssim(\zeta_K K^{-(\omega+1/2)})^{1/\omega'}$, where ω and ω' are the constants defined in Assumption 3.2. By compactness and convexity of the support of (S,A), we may choose $\mathcal{X}_n$ to have cardinality $\#(\mathcal{X}_n)\lesssim n^\beta$ for some constant $0<\beta<\infty$.

$$\begin{aligned}&\sup_{s,a}\left|\phi_K(s,a)^\intercal\Sigma^{-1}\left\{\frac{1}{n}\sum_{i=1}^n\phi_K(S_i,A_i)\left[f^\pi(S_i')-\mathcal{P}_t^\pi f(S_i,A_i)\right]\right\}\right|\\\le&\max_{(s_n,a_n)\in\mathcal{X}_n}\left|\phi_K(s_n,a_n)^\intercal\Sigma^{-1}\left\{\frac{1}{n}\sum_{i=1}^n\phi_K(S_i,A_i)\left[f^\pi(S_i')-\mathcal{P}_t^\pi f(S_i,A_i)\right]\right\}\right|\\&+\sup_{s,a}\left|[\phi_K(s,a)-\phi_K(s_n,a_n)]^\intercal\Sigma^{-1}\left\{\frac{1}{n}\sum_{i=1}^n\phi_K(S_i,A_i)\left[f^\pi(S_i')-\mathcal{P}_t^\pi f(S_i,A_i)\right]\right\}\right|\\\le&\max_{(s_n,a_n)\in\mathcal{X}_n}\left|\phi_K(s_n,a_n)^\intercal\Sigma^{-1}\left\{\frac{1}{n}\sum_{i=1}^n\phi_K(S_i,A_i)\left[f^\pi(S_i')-\mathcal{P}_t^\pi f(S_i,A_i)\right]\right\}\right|\\&+C_\omega K^\omega(\zeta_K K^{-(\omega+1/2)})\left\|\Sigma^{-1/2}\frac{1}{n}\sum_{i=1}^n\phi_K(S_i,A_i)\left[f^\pi(S_i')-\mathcal{P}_t^\pi f(S_i,A_i)\right]\right\|_2.\end{aligned}$$

The second term can be bounded by using the same argument as that in Lemma D.3. We focus on the first term. For any fixed (s_n, a_n), from Lemma D.4, we have

$$\sup_{f\in\mathcal{Q}^{(t+1)}(1)}\left|\phi_K(s,a)^\intercal\Sigma^{-1}\left\{\frac{1}{n}\sum_{i=1}^n\phi_K(S_i,A_i)\left[f^\pi(S_i')-\mathcal{P}_t^\pi f(S_i,A_i)\right]\right\}\right|\lesssim\frac{\zeta_K}{\sqrt{n}}$$

with probability at least $1-c\exp\{-c^{-1}\zeta_K^{-1}\sqrt{n}\}$ for some universal constant $c>0$. It follows by the union bound that

$$\begin{aligned}
&\Pr\left(\max_{(s_n,a_n)\in\mathcal{X}_n}\left|\phi_K(s_n,a_n)^\intercal\Sigma^{-1}\left\{\frac{1}{n}\sum_{i=1}^n\phi_K(S_i,A_i)\left[f^\pi(S_i')-\mathcal{P}_t^\pi f(S_i,A_i)\right]\right\}\right|>\frac{\zeta_K}{\sqrt{n}}\right)\\
\leq&\mathrm{card}(\mathcal{X}_n)\max_{(s_n,a_n)\in\mathcal{X}_n}\Pr\left(\left|\phi_K(s_n,a_n)^\intercal\Sigma^{-1}\left\{\frac{1}{n}\sum_{i=1}^n\phi_K(S_i,A_i)\left[f^\pi(S_i')-\mathcal{P}_t^\pi f(S_i,A_i)\right]\right\}\right|>\frac{\zeta_K}{\sqrt{n}}\right)\\
\lesssim&c_1\exp\left\{\beta\log n-c_2\zeta_K^{-1}\sqrt{n}\right\},
\end{aligned}$$

for some universal constants $c_1, c_2>0$. Therefore,

$$\begin{aligned}
&\sup_{f\in\mathcal{Q}^{(t+1)}}\sup_{s,a}\left|\phi_K(s,a)^\intercal(\hat{\Sigma})^{-1}\left\{\frac{1}{n}\sum_{i=1}^n\phi_K(S_i,A_i)\left[f^\pi(S_i')-\mathcal{P}_t^\pi f(S_i,A_i)\right]\right\}\right|\\
&\qquad\lesssim\frac{\zeta_K}{\sqrt{n}}+K^{\frac{-1}{2}}\frac{\zeta_K^2\sqrt{\log n\log K}}{\sqrt{n}}+\frac{\zeta_K^3\log n\log K}{n}
\end{aligned}$$

with probability at least $1-2(nK)^{-1}-c_1\exp\left\{\beta\log n-c_2\zeta_K^{-1}\sqrt{n}\right\}$.

D. Additional Lemmas

Lemma D.1. *The variance of $\mathcal{E}_1(E_1)$ can be decomposed as*

$$\begin{aligned}
&\mathrm{Var}\left(\mathcal{E}_1^\pi\left\{\sum_{t=1}^T\left(\prod_{t'=0}^{t-1}\mathcal{P}_{t'}^\pi\right)\tilde{\Pi}_t\left[Q_t^\pi-(R_t+\langle\pi_t,Q_{t+1}^\pi\rangle)\right]\right\}\right)\\
=&\sum_{t=1}^T\mathbb{E}\left\{\mathrm{Var}\left(\mathcal{E}_1^\pi\left\{\left(\prod_{t'=0}^{t-1}\mathcal{P}_{t'}^\pi\right)\tilde{\Pi}_t\left[Q_t^\pi-(R_t+\langle\pi_t,Q_{t+1}^\pi\rangle)\right]\right\}\mid\mathcal{D}_t\right)\right\}
\end{aligned}$$

Proof. By iteratively applying law of total variance, we have

$$
\begin{aligned}
&\mathrm{Var}\left(\mathcal{E}_1^\pi\left\{\sum_{t=1}^{T}\left(\prod_{t'=0}^{t-1}\mathcal{P}_{t'}^\pi\right)\tilde{\Pi}_t\left[Q_t^\pi-(R_t+\langle\pi_t,Q_{t+1}^\pi\rangle)\right]\right\}\right)\\
=&\mathbb{E}\left(\mathrm{Var}\left(\mathcal{E}_1^\pi\left\{\sum_{t=1}^{T}\left(\prod_{t'=0}^{t-1}\mathcal{P}_{t'}^\pi\right)\tilde{\Pi}_t\left[Q_t^\pi-(R_t+\langle\pi_t,Q_{t+1}^\pi\rangle)\right]\right\}\mid\mathcal{D}_T\right)\right)\\
&\quad+\mathrm{Var}\left[\mathbb{E}\left(\mathcal{E}_1^\pi\left\{\sum_{t=1}^{T}\left(\prod_{t'=0}^{t-1}\mathcal{P}_{t'}^\pi\right)\tilde{\Pi}_t\left[Q_t^\pi-(R_t+\langle\pi_t,Q_{t+1}^\pi\rangle)\right]\right\}\mid\mathcal{D}_T\right)\right]\\
=&\mathbb{E}\left(\mathrm{Var}\left(\mathcal{E}_1^\pi\left\{\left(\prod_{t'=0}^{T-1}\mathcal{P}_{t'}^\pi\right)\tilde{\Pi}_T\left[Q_T^\pi-(R_T+\langle\pi_t,Q_{T+1}^\pi\rangle)\right]\right\}\mid\mathcal{D}_T\right)\right)\\
&\quad+\mathrm{Var}\left[\mathcal{E}_1^\pi\left\{\sum_{t=1}^{T-1}\left(\prod_{t'=0}^{t-1}\mathcal{P}_{t'}^\pi\right)\tilde{\Pi}_t\left[Q_t^\pi-(R_t+\langle\pi_t,Q_{t+1}^\pi\rangle)\right]\right\}\right]\\
=&\mathbb{E}\left(\mathrm{Var}\left(\mathcal{E}_1^\pi\left\{\left(\prod_{t'=0}^{T-1}\mathcal{P}_{t'}^\pi\right)\tilde{\Pi}_T\left[Q_T^\pi-(R_T+\langle\pi_t,Q_{T+1}^\pi\rangle)\right]\right\}\mid\mathcal{D}_T\right)\right)\\
&\quad+\mathbb{E}\left(\mathrm{Var}\left(\mathcal{E}_1^\pi\left\{\sum_{t=1}^{T-1}\left(\prod_{t'=0}^{t-1}\mathcal{P}_{t'}^\pi\right)\tilde{\Pi}_t\left[Q_t^\pi-(R_t+\langle\pi_t,Q_{t+1}^\pi\rangle)\right]\right\}\mid\mathcal{D}_{T-1}\right)\right)\\
&\quad+\mathrm{Var}\left[\mathbb{E}\left(\mathcal{E}_1^\pi\left\{\sum_{t=1}^{T-1}\left(\prod_{t'=0}^{t-1}\mathcal{P}_{t'}^\pi\right)\tilde{\Pi}_t\left[Q_t^\pi-(R_t+\langle\pi_t,Q_{t+1}^\pi\rangle)\right]\right\}\mid\mathcal{D}_{T-1}\right)\right]\\
=&\mathbb{E}\left(\mathrm{Var}\left(\mathcal{E}_1^\pi\left\{\left(\prod_{t'=0}^{T-1}\mathcal{P}_{t'}^\pi\right)\tilde{\Pi}_T\left[Q_T^\pi-(R_T+\langle\pi_t,Q_{T+1}^\pi\rangle)\right]\right\}\mid\mathcal{D}_T\right)\right)\\
&\quad+\mathbb{E}\left(\mathrm{Var}\left(\mathcal{E}_1^\pi\left\{\left(\prod_{t'=0}^{T-2}\mathcal{P}_{t'}^\pi\right)\tilde{\Pi}_{T-1}\left[Q_{T-1}^\pi-(R_{T-1}+\langle\pi_t,Q_T^\pi\rangle)\right]\right\}\mid\mathcal{D}_{T-1}\right)\right)\\
&\quad+\mathrm{Var}\left[\mathbb{E}\left(\mathcal{E}_1^\pi\left\{\sum_{t=1}^{T-2}\left(\prod_{t'=0}^{t-1}\mathcal{P}_{t'}^\pi\right)\tilde{\Pi}_t\left[Q_t^\pi-(R_t+\langle\pi_t,Q_{t+1}^\pi\rangle)\right]\right\}\mid\mathcal{D}_{T-2}\right)\right]\\
=&\cdots\\
=&\sum_{t=1}^{T}\mathbb{E}\left\{\mathrm{Var}\left(\mathcal{E}_1^\pi\left\{\left(\prod_{t'=0}^{t-1}\mathcal{P}_{t'}^\pi\right)\tilde{\Pi}_t\left[Q_t^\pi-(R_t+\langle\pi_t,Q_{t+1}^\pi\rangle)\right]\right\}\mid\mathcal{D}_t\right)\right\}.
\end{aligned}
$$

□

Lemma D.2. *With probability at least* $1-(nK)^{-2}$*, we have*

$$
\left\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2}-I_K\right\|\lesssim\frac{\zeta_K\sqrt{\log n\log K\log\zeta_K}}{\sqrt{n}}+\frac{\zeta_K^2\log n\log K\log\zeta_K}{n}. \tag{31}
$$

If we further assume that $\zeta_K^2\log n\log K\log\zeta_K=o(n)$*,*

$$
\left\|\Sigma^{1/2}\hat{\Sigma}^{-1}\Sigma^{1/2}-I_K\right\|\lesssim\frac{\zeta_K\sqrt{\log n\log K\log\zeta_K}}{\sqrt{n}} \tag{32}
$$

Proof. By applying Lemma B.5 in (Duan et al., 2020), we obtain (31). Next, we condition on the event that (31) holds, under the condition that $\zeta_K^2\log n\log K\log\zeta_K=o(n)$, we have

$$
\left\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2}-I_K\right\|\le\frac{1}{2}=\frac{1}{2}\lambda_{\min}(I_K).
$$

Then we apply Lemma F.4 in (Chen & Christensen, 2018) and obtain

$$\left\|\Sigma^{1/2}\hat{\Sigma}^{-1}\Sigma^{1/2} - I_K\right\| \leq 2(1+\sqrt{5})[\lambda_{\min}(I_K)]^{-2}\left\|\Sigma^{1/2}\hat{\Sigma}^{-1}\Sigma^{1/2} - I_K\right\| \lesssim \frac{\zeta_K\sqrt{\log n \log K \log \zeta_K}}{\sqrt{n}}.$$

□

Lemma D.3. *Suppose the covering number of space $\mathcal{Q}(1)$ satisfies that $N(\mathcal{Q}(1), \|\cdot\|_\infty, \epsilon) \leq \exp(A\epsilon^{-\alpha})$ for some constant $A > 0$ and $\alpha < 2$. Then with probability at least $1 - (nK)^{-2}$, we have*

$$\sup_{h\in\mathcal{Q}(1)}\left\|\Sigma^{\frac{-1}{2}}\left\{\frac{1}{n}\sum_{i=1}^n \phi_K(S_i,A_i)h(S_i,A_i) - \mathbb{E}\left[\phi_K(S,A)h(S,A)\right]\right\}\right\|_2 \leq C\frac{\zeta_K\sqrt{\log N \log K}}{\sqrt{n}},$$

where the constant C depends on A and α.

Proof. Take r_i, $i = 1, \ldots, n$ as independent Rademacher random variables. Then by symmetrization inequality, we have

$$\begin{aligned}
&\mathbb{E}\sup_{h\in\mathcal{Q}(1)}\left\|\Sigma^{\frac{-1}{2}}\left\{\frac{1}{n}\sum_{i=1}^n \phi_K(S_i,A_i)h(S_i,A_i) - \mathbb{E}\left[\phi_K(S,A)h(S,A)\right]\right\}\right\|_2^2 \\
&\qquad\qquad \lesssim \mathbb{E}\sup_{h\in\mathcal{Q}(1)}\left\|\Sigma^{\frac{-1}{2}}\left\{\frac{1}{n}\sum_{i=1}^n \phi_K(S_i,A_i)h(S_i,A_i)r_i\right\}\right\|_2^2.
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}\sup_{h\in\mathcal{Q}(1)}\left\|\Sigma^{\frac{-1}{2}}\left\{\frac{1}{n}\sum_{i=1}^n \phi_K(S_i,A_i)h(S_i,A_i)r_i\right\}\right\|_2^2 \\
\leq&\frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\sup_{h\in\mathcal{Q}^{(1)}}\|\Sigma^{-1/2}\phi_K(S_i,A_i)h(S_i,A_i)r_i\|_2^2 \\
&+\frac{1}{n^2}\sum_{i\neq j}\mathbb{E}\sup_{h\in\mathcal{Q}(1)} r_ir_jh(S_i,A_i)h(S_j,A_j)\phi_K(S_i,A_i)^\intercal\Sigma^{-1}\phi_K(S_j,A_j) \\
\leq&\frac{1}{n}\zeta_K^2 + \frac{1}{n^2}\sum_{i\neq j}\mathbb{E}\sup_{h\in\mathcal{Q}^{(1)}} r_ir_jh(S_i,A_i)h(S_j,A_j)\phi_K(S_i,A_i)^\intercal\Sigma^{-1}\phi_K(S_j,A_j)
\end{aligned}$$

Next, we focus on bounding

$$\frac{1}{n^2}\sum_{i\neq j}\mathbb{E}\sup_{h\in\mathcal{Q}^{(1)}} r_ir_jh(S_i,A_i)h(S_j,A_j)\phi_K(S_i,A_i)^\intercal\Sigma^{-1}\phi_K(S_j,A_j).$$

As r_i, $i = 1, \ldots, n$ are Rademacher random variables, then there exists a constant σ^2 such that $\mathbb{E}\exp(\lambda r_ir_j) \leq \exp(\lambda^2\sigma^2/2)$ for any $\lambda > 0$. Conditioned on (S_i, A_i), $i = 1, \ldots, n$, then

$$r_ir_j[h_1(S_i,A_i)h_1(S_j,A_j) - h_2(S_i,A_i)h_2(S_j,A_j)]\phi_K(S_i,A_i)^\intercal\Sigma^{-1}\phi_K(S_j,A_j)$$

is a subgaussian random variable with parameter

$$[h_1(S_i,A_i)h_1(S_j,A_j) - h_2(S_i,A_i)h_2(S_j,A_j)]^2\sigma^2.$$

Define

$$\mathcal{S}(h) = \frac{1}{n}\sum_{i\neq j} r_ir_jh(S_i,A_i)h(S_j,A_j)\phi_K(S_i,A_i)^\intercal\Sigma^{-1}\phi_K(S_j,A_j).$$

We know that $r_i r_j$ is independent of $r_{i'} r_{j'}$ as long as either $i \neq i'$ or $j \neq j'$, then conditioned on (S_i, A_i), $i = 1, \ldots, n$, $S(h_1) - S(h_2)$ is a subgaussian with parameter

$$d^2(h_1, h_2) = \frac{1}{n^2} \sum_{i \neq j} [h_1(S_i, A_i) h_1(S_j, A_j) - h_2(S_i, A_i) h_2(S_j, A_j)]^2 \sigma^2.$$

Next, we derive $H^{1/2}(\mathcal{Q}(1), d, \epsilon)$. We know that the covering number $N(\mathcal{Q}, \|\cdot\|_\infty, \epsilon) \leq \exp(A\epsilon^{-\alpha})$. Consider $\mathcal{N} \subset \mathcal{Q}(1)$ as the ϵ-net of $\mathcal{Q}(1)$ with respect to $\|\cdot\|_\infty$. By definition, for any $h \in \mathcal{Q}(1)$, there exists a $u_0 \in \mathcal{N}$, such that

$$\sup_{(s,a)} |h(s,a) - h_0(s,a)| \leq \epsilon. \tag{33}$$

Then

$$\begin{aligned}
d^2(h, h_0) =& \frac{1}{n^2} \sum_{i \neq j} [h(S_i, A_i) h(S_j, A_j) - h_0(S_i, A_i) h_0(S_j, A_j) \phi_K(S_i, A_i)^\intercal \Sigma^{-1} \phi_K(S_j, A_j)]^2 \sigma^2 \\
\leq & \frac{1}{n^2} \sigma^2 \sup_{s,a} \|\Sigma^{-1/2} \phi_K(s,a)\|_2^4 \sum_{i \neq j} [h(S_i, A_i) h(S_j, A_j) - h_0(S_i, A_i) h_0(S_j, A_j)]^2 \\
\leq & \frac{1}{n^2} \sigma^2 \zeta_K^4 \sum_{i \neq j} \{h(S_i, A_i)(h(S_j, A_j) - h_0(S_j, A_j)) + h_0(S_j, A_j)(h(S_i, A_i) - h_0(S_i, A_i)\}^2 \\
\leq & \frac{4}{n^2} \sigma^2 \zeta_K^4 \sum_{i \neq j} (2\epsilon)^2 \leq 4\sigma^2 \zeta_K^4 \epsilon^2.
\end{aligned}$$

Therefore we have

$$\begin{aligned}
N(\mathcal{Q}(1), d, 2\sigma\zeta_K^2 \epsilon) &\leq N(\mathcal{Q}(1), \|\cdot\|_\infty, \epsilon) \\
N(\mathcal{Q}(1), d, \epsilon) \leq N(\mathcal{Q}(1), \|\cdot\|_\infty, \epsilon/(2\sigma\zeta_K^2)) &\leq \exp\left\{ A \left(\frac{\epsilon}{2\sigma\zeta_K^2} \right)^{-\alpha} \right\}.
\end{aligned}$$

Following Dudley's entropy bounds, we have

$$\begin{aligned}
\mathbb{E} \sup_{h \in \mathcal{Q}(1)} \mathcal{S}(h) &\lesssim \mathbb{E} \int_0^D H^{1/2}(\mathcal{Q}(1), d, \epsilon) d\epsilon \\
&\leq \int_0^{\zeta_K^2} \left\{ A \left(\frac{\epsilon}{2\sigma\zeta_K^2} \right)^{-\alpha} \right\}^{1/2} d\epsilon \leq C\zeta_K^2,
\end{aligned}$$

where $D = \sup_{h \in \mathcal{Q}^{(1)}} \sqrt{\frac{1}{n^2} \sum_{i \neq j} [h(S_i, A_i) h(S_j, A_j) \phi_K(S_i, A_i)^\intercal \Sigma^{-1} \phi_K(S_j, A_j)]^2} \leq \sup_{s,a} \|\Sigma^{-1/2} \phi_K(s,a)\|_2^2 \leq \zeta_K^2$, C is a constant depending on A and α.

Therefore we have

$$\begin{aligned}
&\frac{1}{n^2} \sum_{i \neq j} \mathbb{E} \sup_{h \in \mathcal{Q}^{(1)}} r_i r_j h(S_i, A_i) h(S_j, A_j) \phi_K(S_i, A_i)^\intercal \Sigma^{-1} \phi_K(S_j, A_j) \lesssim \frac{1}{n} \zeta_K^2 \\
&\mathbb{E} \sup_{h \in \mathcal{Q}(1)} \left\| \Sigma^{\frac{-1}{2}} \left\{ \frac{1}{n} \sum_{i=1}^n \phi_K(S_i, A_i) h(S_i, A_i) - \mathbb{E}\left[\phi_K(S, A) h(S, A) \right] \right\} \right\|_2 \\
\leq & \left[\mathbb{E} \sup_{h \in \mathcal{Q}(1)} \left\| \Sigma^{\frac{-1}{2}} \left\{ \frac{1}{n} \sum_{i=1}^n \phi_K(S_i, A_i) h(S_i, A_i) - \mathbb{E}\left[\phi_K(S, A) h(S, A) \right] \right\} \right\|_2^2 \right]^{1/2} \\
\lesssim & \left[\mathbb{E} \sup_{h \in \mathcal{Q}(1)} \left\| \Sigma^{\frac{-1}{2}} \left\{ \frac{1}{n} \sum_{i=1}^n \phi_K(S_i, A_i) h(S_i, A_i) r_i \right\} \right\|_2^2 \right]^{1/2} \lesssim \frac{1}{\sqrt{n}} \zeta_K.
\end{aligned}$$

Next, take $X_i = (S_i, A_i)$ and define

$$g(X_1, \ldots, X_n) = \sup_{h \in \mathcal{Q}(1)} \left\| \Sigma^{\frac{-1}{2}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \phi_K(S_i, A_i) h(S_i, A_i) - \mathbb{E}\left[\phi_K(S, A) h(S, A)\right] \right\} \right\|_2$$

We have

$$\begin{aligned}
&|g(X_1, \ldots, X_j, \ldots, X_n) - g(X_1, \ldots, X_j', \ldots, X_n)| \\
\leq &\sup_{h \in \mathcal{Q}(1)} \left| \left\| \Sigma^{\frac{-1}{2}} \left\{ \frac{1}{n} \sum_{i \neq j} \phi_K(S_i, A_i) h(S_i, A_i) - \mathbb{E}\left[\phi_K(S, A) h(S, A)\right] + \frac{1}{n} \phi_K(S_j, A_j) h(S_j, A_j) - \mathbb{E}\left[\phi_K(S, A) h(S, A)\right] \right\} \right\|_2 \right. \\
&\quad \left. - \left\| \Sigma^{\frac{-1}{2}} \left\{ \frac{1}{n} \sum_{i \neq j} \phi_K(S_i, A_i) h(S_i, A_i) - \mathbb{E}\left[\phi_K(S, A) h(S, A)\right] + \frac{1}{n} \phi_K(S_j', A_j') h(S_j', A_j') - \mathbb{E}\left[\phi_K(S, A) h(S, A)\right] \right\} \right\|_2 \right| \\
\leq &\sup_{h \in \mathcal{Q}(1)} \left\| \Sigma^{\frac{-1}{2}} \frac{1}{n} \left[\phi_K(S_j, A_j) h(S_j, A_j) - \phi_K(S_j', A_j') h(S_j', A_j') \right] \right\|_2 \\
\leq &\frac{1}{n} \sup_{s,a} \|\Sigma^{-1/2} \phi_K(s, a)\|_2 \leq \frac{1}{n} \zeta_K,
\end{aligned}$$

where the first and the second inequality is due to the triangle inequality for $\sup$ and $\|\cdot\|_2$. Then we are able to use the bounded difference inequality and we have

$$\begin{aligned}
&\Pr\left(\sup_{h \in \mathcal{Q}(1)} \left\| \Sigma^{\frac{-1}{2}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \phi_K(S_i, A_i) h(S_i, A_i) - \mathbb{E}\left[\phi_K(S, A) h(S, A)\right] \right\} \right\|_2 \right. \\
&\qquad \left. - \mathbb{E} \sup_{h \in \mathcal{Q}(1)} \left\| \Sigma^{\frac{-1}{2}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \phi_K(S_i, A_i) h(S_i, A_i) - \mathbb{E}\left[\phi_K(S, A) h(S, A)\right] \right\} \right\|_2 \geq t \right) \\
\leq &\exp\left\{ -\frac{2t^2}{n(\zeta_K/n)^2} \right\} = \exp\left\{ -\frac{2nt^2}{\zeta_K^2} \right\}.
\end{aligned}$$

By taking $t = \frac{\zeta_K \sqrt{\log n \log K}}{\sqrt{n}}$, we obtain the result. □

Lemma D.4. *Suppose the covering number of space $\mathcal{Q}^{(t+1)}(1)$ satisfies that $N(\mathcal{Q}^{(t+1)}(1), \|\cdot\|_\infty, \epsilon) \leq \exp(A\epsilon^{-\alpha})$ for some constant $A > 0$ and $\alpha \leq 2$. And we assume that $\zeta_K = \mathcal{O}(\sqrt{n})$, then for any (s, a), we have*

$$\left| \sup_{f \in \mathcal{Q}^{(t+1)}(1)} \phi_K(s, a)^\intercal \Sigma^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} \phi_K(S_i, A_i) \left[f^\pi(S_i') - \mathcal{P}_t^\pi f(S_i, A_i) \right] \right\} \right| = \mathcal{O}\left(\frac{\|\Sigma^{-1/2} \phi_K(s, a)\|_2 \sqrt{\log n \log T}}{\sqrt{n}} \right), \text{ W.H.P.}$$

uniformly holds for all $t = 1, \ldots, T$.

Proof. Take r_i, $i = 1, \ldots, n$ as independent Rademacher random variables. Then by symmetrization inequality, we have

$$\begin{aligned}
&\mathbb{E} \sup_{f \in \mathcal{Q}^{(t+1)}(1)} \left| \phi_K(s, a)^\intercal \Sigma^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} \phi_K(S_i, A_i) \left[f^\pi(S_i') - \mathcal{P}_t^\pi f(S_i, A_i) \right] \right\} \right| \\
\lesssim &\mathbb{E} \sup_{f \in \mathcal{Q}^{(t+1)}(1)} \left| \phi_K(s, a)^\intercal \Sigma^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} \phi_K(S_i, A_i) r_i f^\pi(S_i') \right\} \right|.
\end{aligned}$$

Define

$$\mathcal{S}(f) = \sqrt{n} \phi_K(s, a)^\intercal \Sigma^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} \phi_K(S_i, A_i) r_i f^\pi(S_i') \right\}.$$

Then conditioned on (S_i, A_i, S_i'), $i = 1, \ldots, n$, $S(f_1 - f_2)$ is a subgaussian process with parameter

$$d^2(f_1, f_2) = \frac{1}{n} \sum_{i=1}^{n} [\phi_K(s, a)^\intercal \Sigma^{-1} \phi_K(S_i, A_i)]^2 [f_1^\pi(S_i') - f_2^\pi(S_i')]^2.$$

Next, we derive $H^{1/2}(\mathcal{Q}^{(t+1)}(1), d, \epsilon)$. We know that the covering number $N(\mathcal{Q}^{(t+1)}(1), \|\cdot\|_\infty, \epsilon) \le \exp(A\epsilon^{-\alpha})$. Consider $\mathcal{N} \subset \mathcal{Q}^{(t+1)}(1)$ as the ϵ-net of $\mathcal{Q}^{(t+1)}(1)$ with respect to $\|\cdot\|_\infty$. By definition, for any $f \in \mathcal{Q}^{(t+1)}(1)$, there exists a $f_0 \in \mathcal{N}$, such that

$$\sup_{(s,a)} |f(s,a) - f_0(s,a)| \le \epsilon. \tag{34}$$

Then

$$\begin{aligned} d^2(f, f_0) =& \frac{1}{n}\sum_{i=1}^n [\phi_K(s,a)^\intercal \Sigma^{-1} \phi_K(S_i, A_i)]^2 [f^\pi(S_i') - f_0^\pi(S_i')]^2 \\ \le& \frac{1}{n}\sum_{i=1}^n [\phi_K(s,a)^\intercal \Sigma^{-1} \phi_K(S_i, A_i)]^2 \epsilon^2 \\ =& [\phi_K(s,a)^\intercal \Sigma^{-1} \hat{\Sigma} \Sigma^{-1} \phi_K(s,a)] \epsilon^2. \end{aligned}$$

Therefore we have

$$\begin{aligned} & N(\mathcal{Q}^{(t+1)}(1), d, \sqrt{\phi_K(s,a)^\intercal \Sigma^{-1} \hat{\Sigma} \Sigma^{-1} \phi_K(s,a)}\epsilon) \le N(\mathcal{Q}^{(t+1)}(1), \|\cdot\|_\infty, \epsilon) \\ & N(\mathcal{Q}(1), d, \epsilon) \\ & \le N(\mathcal{Q}(1), \|\cdot\|_\infty, \epsilon / \sqrt{\phi_K(s,a)^\intercal \Sigma^{-1} \hat{\Sigma} \Sigma^{-1} \phi_K(s,a)}) \le \exp\left\{ A \left(\frac{\epsilon}{\sqrt{\phi_K(s,a)^\intercal \Sigma^{-1} \hat{\Sigma} \Sigma^{-1} \phi_K(s,a)}} \right)^{-\alpha} \right\}. \end{aligned}$$

Following Dudley's entropy bounds, we have

$$\begin{aligned} \mathbb{E} \sup_{f \in \mathcal{Q}^{(t+1)}(1)} \mathcal{S}(f) \lesssim& \mathbb{E} \int_0^D H^{1/2}(\mathcal{Q}^{(t+1)}(1), d, \epsilon) d\epsilon \\ \le& \mathbb{E} \int_0^{\sqrt{\phi_K(s,a)^\intercal \Sigma^{-1} \hat{\Sigma} \Sigma^{-1} \phi_K(s,a)}} \left\{ A \left(\frac{\epsilon}{\sqrt{\phi_K(s,a)^\intercal \Sigma^{-1} \hat{\Sigma} \Sigma^{-1} \phi_K(s,a)}} \right)^{-\alpha} \right\}^{1/2} d\epsilon \\ \le& C \mathbb{E} \sqrt{\phi_K(s,a)^\intercal \Sigma^{-1} \hat{\Sigma} \Sigma^{-1} \phi_K(s,a)} \\ \le& C \sqrt{\mathbb{E}[\phi_K(s,a)^\intercal \Sigma^{-1} \hat{\Sigma} \Sigma^{-1} \phi_K(s,a)]} \\ =& C \|\Sigma^{-1/2} \phi_K(s,a)\|_2. \end{aligned}$$

where C is a constant depending on A and α.

Next, we apply the Talagrand concentration inequality.

$$\begin{aligned} & \sup_{f \in \mathcal{Q}^{(t+1)}(1)} \left| \phi_K(s,a)^\intercal \Sigma^{-1} \left\{ \phi_K(S_i, A_i) [f^\pi(S_i') - \mathcal{P}_t^\pi f(S_i, A_i)] \right\} \right| \le 2 \|\Sigma^{-1/2} \phi_K(s,a)\|_2 \zeta_K. \\ & \sup_{f \in \mathcal{Q}^{(t+1)}(1)} \mathbb{E} \left| \phi_K(s,a)^\intercal \Sigma^{-1} \left\{ \phi_K(S_i, A_i) [f^\pi(S_i') - \mathcal{P}_t^\pi f(S_i, A_i)] \right\} \right|^2 \le 2 \mathbb{E} \phi_K(s,a)^\intercal \Sigma^{-1} \hat{\Sigma} \Sigma^{-1} \phi_K(s,a) \\ & \qquad \le 2 \|\Sigma^{-1/2} \phi_K(s,a)\|_2^2 \end{aligned}$$

Then we take

$$\begin{aligned} U =& 2 \|\Sigma^{-1/2} \phi_K(s,a)\|_2 \zeta_K \\ V =& n(2 \|\Sigma^{-1/2} \phi_K(s,a)\|_2^2) + 8U\sqrt{n} \|\Sigma^{-1/2} \phi_K(s,a)\|_2. \end{aligned}$$

There exists a universal constant $c > 0$ such that for every $t > 0$,

$$\begin{aligned}
&\Pr\left(\left|\left|\sup_{f\in\mathcal{Q}^{(t+1)}(1)} \phi_K(s,a)^\intercal\Sigma^{-1}\left\{\frac{1}{n}\sum_{i=1}^n \phi_K(S_i,A_i)\left[f^\pi(S_i') - \mathcal{P}_t^\pi f(S_i,A_i)\right]\right\}\right| \right.\\
&\qquad \left. -\mathbb{E}\left|\sup_{f\in\mathcal{Q}^{(t+1)}(1)} \phi_K(s,a)^\intercal\Sigma^{-1}\left\{\frac{1}{n}\sum_{i=1}^n \phi_K(S_i,A_i)\left[f^\pi(S_i') - \mathcal{P}_t^\pi f(S_i,A_i)\right]\right\}\right|\right| > t\right)\\
\leq & c\exp\left\{-\frac{1}{c}\frac{nt}{2\|\Sigma^{-1/2}\phi_K(s,a)\|_2\zeta_K}\log\left(1+\frac{nt2\|\Sigma^{-1/2}\phi_K(s,a)\|_2\zeta_K}{n(2\|\Sigma^{-1/2}\phi_K(s,a)\|_2^2)+8U\sqrt{n}\|\Sigma^{-1/2}\phi_K(s,a)\|_2}\right)\right\}\\
\lesssim & c\exp\left\{-\frac{1}{c}\frac{nt}{2\|\Sigma^{-1/2}\phi_K(s,a)\|_2\zeta_K}\log\left(1+\frac{2nt\|\Sigma^{-1/2}\phi_K(s,a)\|_2\zeta_K}{4n\|\Sigma^{-1/2}\phi_K(s,a)\|_2^2}\right)\right\}\\
\lesssim & c\exp\left\{-\frac{1}{c}\frac{nt}{2\|\Sigma^{-1/2}\phi_K(s,a)\|_2\zeta_K}\log\left(1+t/2\right)\right\},
\end{aligned}$$

where the second inequality is due to the condition that $\zeta_K = \mathcal{O}(\sqrt{n})$. By taking $t = \frac{\|\Sigma^{-1/2}\phi_K(s,a)\|_2\sqrt{\log n\log T}}{\sqrt{n}}$, we obtain

$$\begin{aligned}
&\Pr\left(\left|\left|\sup_{f\in\mathcal{Q}^{(t+1)}(1)} \phi_K(s,a)^\intercal\Sigma^{-1}\left\{\frac{1}{n}\sum_{i=1}^n \phi_K(S_i,A_i)\left[f^\pi(S_i') - \mathcal{P}_t^\pi f(S_i,A_i)\right]\right\}\right| \right.\right.\\
&\qquad \left.\left. -\mathbb{E}\left|\sup_{f\in\mathcal{Q}^{(t+1)}(1)} \phi_K(s,a)^\intercal\Sigma^{-1}\left\{\frac{1}{n}\sum_{i=1}^n \phi_K(S_i,A_i)\left[f^\pi(S_i') - \mathcal{P}_t^\pi f(S_i,A_i)\right]\right\}\right|\right| > \frac{\|\Sigma^{-1/2}\phi_K(s,a)\|_2\sqrt{\log n\log T}}{\sqrt{n}}\right)\\
\lesssim & c\exp\left\{-\frac{1}{c}\frac{\sqrt{n}\sqrt{\log n\log T}}{\zeta_K}\right\}.
\end{aligned}$$

Take a union bound over T, with the condition that $\zeta_K = \mathcal{O}(n)$, we have

$$\begin{aligned}
&\left|\sup_{f\in\mathcal{Q}^{(t+1)}(1)} \phi_K(s,a)^\intercal\Sigma^{-1}\left\{\frac{1}{n}\sum_{i=1}^n \phi_K(S_i,A_i)\left[f^\pi(S_i') - \mathcal{P}_t^\pi f(S_i,A_i)\right]\right\}\right|\\
&\qquad = \mathcal{O}\left(\frac{\|\Sigma^{-1/2}\phi_K(s,a)\|_2\sqrt{\log n\log T}}{\sqrt{n}}\right), \text{ W.H.P.}
\end{aligned}$$

uniformly holds for all $t = 1,\ldots,T$. □