On Well-posedness and Minimax Optimal Rates of Nonparametric Q-function Estimation in Off-policy Evaluation

Xiaohong Chen [1] **Zhengling Qi** [2]

Abstract

We study the off-policy evaluation (OPE) problem in an infinite-horizon Markov decision process with continuous states and actions. We recast the Q-function estimation into a special form of the nonparametric instrumental variables (NPIV) estimation problem. We first show that under one mild condition the NPIV formulation of Q-function estimation is well-posed in the sense of L^2-*measure of ill-posedness* with respect to the data generating distribution, bypassing a strong assumption on the discount factor γ imposed in the recent literature for obtaining the L^2 convergence rates of various Q-function estimators. Thanks to this new well-posed property, we derive the first minimax lower bounds for the convergence rates of nonparametric estimation of Q-function and its derivatives in both sup-norm and L^2-norm, which are shown to be the same as those for the classical nonparametric regression (Stone, 1982). We then propose a sieve two-stage least squares estimator and establish its rate-optimality in both norms under some mild conditions. Our general results on the well-posedness and the minimax lower bounds are of independent interest to study not only other nonparametric estimators for Q-function but also efficient estimation on the value of any target policy in off-policy settings.

1. Introduction

In recent years, there is a surging interest in studying batch reinforcement learning (RL), which utilizes previously collected data to perform sequential decision making (Sutton & Barto, 2018) and does not require interacting with task environment or accessing a simulator. The batch RL techniques are especially attractive in many high-stake real-world application domains where it is too costly or infeasible to access a simulator, such as mobile health (Liao et al., 2018), robotics (Pinto & Gupta, 2016), digital marketing (Thomas et al., 2017) and precision medicine (Kosorok & Laber, 2019), and others. Nevertheless, the batch setting still posits several theoretical challenges that tamper the generalizability of many RL algorithms in practice. Among them, one central challenge is the *distributional mismatch* between the data collecting process and the target distribution for evaluation (Levine et al., 2020).

Motivated by these, we study the off-policy evaluation (OPE) problem, which is considered one of fundamental problems in batch RL. The goal of OPE is to leverage pre-collected data generated by a so-called behavior policy to evaluate the performance (e.g., value) of a new/target policy. In particular, we investigate theoretical property of nonparametric estimation of Q-function in the setting of infinite-horizon Markov decision processes (MDPs) (with discounted rewards, continuous states and actions).

We make several important contributions to the existing literature. Motivated by Bellman equation, we formulate Q-function estimation under the framework of a nonparametric instrumental variable (NPIV) model. We first show that, under mild regularity conditions, the NPIV formulation of Q-function estimation is well-posed in the sense of L^2-*measure of ill-posedness* with respect to the data generating distribution. This essentially justifies the valid use of the L^2-norm of Bellman error/residual to measure the accuracy of Q-function estimation in the batch setting. Next, we derive the minimax lower bounds for the convergence rates in sup-norm and in L^2-norm for the estimation of Q-function and its derivatives. Thanks to the general well-posedness result, the lower bounds are shown to be the same as those for the nonparametric regression estimation in the i.i.d. setting (Stone, 1982; Tsybakov, 2009). Thus the nonparametric Q-function estimation could be as easy as the nonparametric regression in terms of the worst case rate. Using the NPIV formulation, we also propose sieve 2SLS estimators to estimate the Q-function (and its derivatives)

Author names are sorted alphabetically. [1]Cowles Foundation for Research in Economics, Yale University [2]Department of Decision Sciences, George Washington University. Correspondence to: Xiaohong Chen <xiaohong.chen@yale.edu>, Zhengling Qi <qizhengling@gwu.edu>.

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

and establish their convergence rates in both sup-norm and L^2-norm. In particular, B-spline and wavelet 2SLS estimators are shown to achieve the sup-norm lower bound for Hölder class of Q-functions (and the derivatives), and many more linear sieve (such as polynomials, cosines, splines, wavelets) 2SLS estimators are shown to achieve the L^2-norm lower bound for Sobolev class of Q-functions (and the derivatives). Our results on L^2-norm convergence rates under mild conditions are particularly useful for obtaining efficient estimation and optimal inference on the value (i.e., the expectation of the Q function) of a target policy. To the best of our knowledge, ours are the first minimax results for non-parametrically estimating Q-function of continuous states and actions in the off-policy setting. The general results on the well-posedness and the minimax lower bounds (in sup-norm and in L^2-norm) are of independent interest to study properties of other nonparametric estimators for Q-function and the related estimators of the marginal importance weight (see, e.g., (Liu et al., 2018)) in the off-policy setting.

1.1. Closely Related Work

Estimation of Q-function for a fixed policy is a key building block for many RL algorithms. There is a growing literature on nonparametric estimation of Q-function in the infinite-horizon and off-policy setting. See some recent theoretical development in (Farahmand et al., 2016; Shi et al., 2020; Uehara et al., 2021) among many others. Specifically, (Farahmand et al., 2016) established L^2 error bound for Bellman error of their Q-function estimator. (Shi et al., 2020; Uehara et al., 2021) derived that L^2-norm convergence rates and error bounds for their respective nonparametric Q-function estimators under a strong assumption that is essentially equivalent to restricting the discount factor γ to be close to zero. Our well-posedness result implies that their L^2-norm convergence rates of their respective estimators for Q-function remain valid without their strong assumption on the discount rate γ. See Section 3 and Remark 5.6 for more detailed discussions.

The connection of estimating Q-function in Bellman equation to instrumental variables estimation, to the best of our knowledge, has been first pointed out by (Bradtke & Barto, 1996) for their celebrated least-squares temporal difference (LSTD) method for parametric models. Recently, the relation between nonparametric Q-function estimation and nonparametric instrumental variables (NPIV) estimation has also been observed by some applied work (such as (Chen et al., 2021)) and theoretical work (such as (Duan et al., 2021) that focuses on the on-policy setting). The NPIV model has been extensively investigated in econometric literature; see, e.g., (Newey & Powell, 2003; Ai & Chen, 2003; Hall & Horowitz, 2005; Blundell et al., 2007; Darolles et al., 2011; Chen & Reiss, 2011; Chen & Christensen, 2018) for earlier reference. However, there is some subtle difference between the nonparametric Q-function estimation and the NPIV one. It is known that a generic NPIV model with continuous endogenous variables is a difficult ill-posed inverse problem in econometrics, but we show that estimation of a nonparametric Q-function of continuous states and actions can be well-posed under mild regularity conditions that are typically assumed in batch RL literature. Our well-posedness result implies that nonparametric estimation and inference on OPE and related batch RL problems could be much simpler than the difficult ill-posed NPIV problems studied in the existing econometric literature.

The rest of the paper is organized as follows. Section 2 presents the framework of infinite-horizon MDPs and some necessary notations. In Section 3, we show that the nonparametric Q-function estimation in sup-norm and in L^2-norm are both well-posed. Section 4 establishes the minimax lower bounds for the rates of convergence for nonparametric estimation of Q-function in sup-norm and in L^2-norm respectively. In Section 5, we propose sieve 2SLS estimation of the Q-function and its derivatives. Under some mild condition, we establish their rates of convergence in both sup-norm and L^2-norm, which coincide with the lower bounds. Section 6 briefly concludes. All proofs are given in the appendix.

2. Preliminaries and Notation

Consider a single trajectory $\{(S_t, A_t, R_t)\}_{t\geq 0}$ where (S_t, A_t, R_t) denotes the state-action-reward triplet collected at time t. Let $\mathcal{S}$ and $\mathcal{A}$ be the state and action spaces, respectively. We assume both state and action are *continuous* (as the discrete and finite spaces are easier). A policy associated with this trajectory defines an agent's way of choosing the action at each decision time t. In this paper, we focus on using the batch data to evaluate the performance of a stationary policy denoted by π, which is a function mapping from the state space $\mathcal{S}$ to a probability distribution over $\mathcal{A}$. In particular, $\pi(a \,|\, s)$ refers to the probability density function of choosing action $a \in \mathcal{A}$ given the state value $s \in \mathcal{S}$. In addition, let $\mathcal{S} \times \mathcal{A} \subseteq \mathbb{R}^d$ for some $d \geq 2$, and $\mathcal{B}(\mathcal{S})$ be the family of Borel subsets of $\mathcal{S}$.

The main goal of this paper is to estimate the so-called Q-function of a target policy π using the batch data. Specifically, given a stationary policy π and any state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, we define Q-function as

$$Q^\pi(s,a) = \sum_{t=0}^{+\infty} \gamma^t \mathbb{E}^\pi(R_t | S_0 = s, A_0 = a),$$

where $\mathbb{E}^\pi$ denotes the expectation assuming the actions are selected according to π, and $0 \leq \gamma < 1$ denotes some discounted factor that balances the trade-off between immediate and future rewards. We consider the framework of

a time-homogeneous MDP and hence make the following two assumptions, which are foundation of many Q-function estimations.

Assumption 2.1. There exists a transition kernel P such that for every $t \geq 0$, $s \in \mathcal{S}$, $a \in \mathcal{A}$ and any set $B \in \mathcal{B}(\mathcal{S})$,

$$\begin{aligned}&\Pr(S_{t+1} \in B \,|\, S_t = s, A_t = a, \{S_j, A_j, R_j\}_{0 \leq j < t}) \\ =&P(S_{t+1} \in B \,|\, S_t = s, A_t = a),\end{aligned}$$

where $\{S_j, A_j, R_j\}_{0 \leq j < t} = \emptyset$ if $t = 0$. In addition, there exists a probability density function q for the transition kernel P.

Assumption 2.2. For every $t \geq 0$, $R_t = \mathcal{R}(S_t, A_t, S_{t+1})$, i.e., a measurable function of (S_t, A_t, S_{t+1}). In addition, there exists a finite constant $R_{\max}$ such that $|R_t| \leq R_{\max}$ for all $t \geq 0$.

Let $r(s,a) = \mathbb{E}[R_t \,|\, S_t = s, A_t = a]$ for every $t \geq 0$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Assumption 2.2 implies that $|r(S_t, A_t)| \leq R_{\max}$ for all $t \geq 0$. We note that the uniformly bounded reward assumption is imposed for simplicity only, and can be replaced by assuming existence of higher order conditional moments of R_t given (S_t, A_t); see, e.g., (Chen & Christensen, 2015; 2018).

To estimate Q^π, by Assumptions 2.1 and 2.2, one approach is to solve the following Bellman equation, i.e.,

$$\begin{aligned}&Q^\pi(s,a) = \mathbb{E}[R_t \,|\, S_t = s, A_t = a] \qquad (1)\\ &+\gamma \mathbb{E}[\int_{a' \in \mathcal{A}} \pi(a' \,|\, S_{t+1}) Q^\pi(S_{t+1}, a') \mathrm{d}a' \,|\, S_t = s, A_t = a],\end{aligned}$$

for any $t \geq 0$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Throughout this paper, we assume the integration with respect to π in (1) can be exactly evaluated as long as the integrand is known. In practice, one can use Monte Carlo method to approximate this integration since the target policy π is known.

Now suppose the given batch data consist of N trajectories, which correspond to N independent and identically distributed copies of $\{(S_t, A_t, R_t)\}_{t \geq 0}$. For $1 \leq i \leq N$, data collected from the ith trajectory are represented by $\{(S_{i,t}, A_{i,t}, R_{i,t}, S_{i,t+1})\}_{0 \leq t < T}$. We then aim to leverage this batch data to estimate Q-function of a target policy π. Before presenting our theoretical results and methods, we make one additional assumption on the data generating process. Let π^b be a stationary policy and $\pi^b(a \,|\, s)$ refers to the conditional probability density of choosing the action a given the state value s.

Assumption 2.3. The batch data $\mathcal{D}_N = \{(S_{i,t}, A_{i,t}, R_{i,t}, S_{i,t+1})\}_{0 \leq t < T, 1 \leq i \leq N}$ are generated by the policy π^b.

Assumptions 2.1-2.3 are standard in the literature of batch RL. Note that in the literature the policy π^b is often called the behavior policy and mostly different from the target one π. Next, we introduce the average visitation probability measure. Let $q_t^{\pi^b}(s,a)$ be the marginal probability density of a state-action pair (s,a) at the decision point t induced by the behavior policy π^b. Then the average visitation probability density across T decision points is defined as

$$\bar{d}_T^{\pi^b}(s,a) = \frac{1}{T} \sum_{t=0}^{T-1} q_t^{\pi^b}(s,a).$$

The corresponding expectation with respect to $\bar{d}_T^{\pi^b}$ is denoted by $\bar{\mathbb{E}}$. We further let $q_t^\pi(s', a' \,|\, s, a)$ be the t-step visitation probability density function induced by a policy π at (s', a') given an initial state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Notation: For generic sequences $\{\varpi(N)\}_{N \geq 1}$ and $\{\theta(N)\}_{N \geq 1}$, the notation $\varpi(N) \gtrsim \theta(N)$ (resp. $\varpi(N) \lesssim \theta(N)$) means that there exists a sufficiently large constant (resp. small) constant $c_1 > 0$ (resp. $c_2 > 0$) such that $\varpi(N) \geq c_1 \theta(N)$ (resp. $\varpi(N) \leq c_2 \theta(N)$). We use $\varpi(N) \asymp \theta(N)$ when $\varpi(N) \gtrsim \theta(N)$ and $\varpi(N) \lesssim \theta(N)$. For matrix and vector norms, we use $\| \bullet \|_{\ell_q}$ to denote either the vector ℓ_q-norm or operator norm induced by the vector ℓ_q-norm, for $1 \leq q < \infty$, when there is no confusion. $\lambda_{\min}(\bullet)$ and $\lambda_{\max}(\bullet)$ denote the minimum and maximum eigenvalues of some square matrix, respectively. For any random variable X, we use $L^q(X)$ to denote the class of all measurable functions with finite q-th moments for $1 \leq q \leq \infty$. Then the L^q-norm is denoted by $\| \bullet \|_{L^q(X)}$. When there is no confusion in the underlying distribution, we also write it as $\| \bullet \|_{L^q}$ or $\| \bullet \|_q$. In particular, $\| \bullet \|_\infty$ denotes the sup-norm. In addition, we use Big O_p and small o_p as the convention. We often use (S, A, R, S') or (S, A, S') to represent some generic transition tuples, where the transition probability density is q. Lastly, we introduce the Hölder class of functions $g : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}$ with smoothness $p > 0$ as

$$\begin{aligned}\Lambda_\infty(p, L) \triangleq \Bigg\{ g \quad | \sup_{0 \leq \|\alpha\|_1 \leq \lfloor p \rfloor} \|\partial^\alpha g\|_\infty \leq L, \\ \sup_{\alpha : \|\alpha\|_1 = \lfloor p \rfloor} \sup_{x, y \in \mathcal{X}, x \neq y} \frac{|\partial^\alpha g(x) - \partial^\alpha g(y)|}{\|x - y\|_{\ell_2}^{\alpha - \lfloor p \rfloor}} \leq L \Bigg\},\end{aligned}$$

where $\mathcal{X} = \mathcal{S} \times \mathcal{A} \subset \mathbb{R}^d$ is a compact rectangular support with nonempty interior, $\lfloor p \rfloor$ denotes the integer no larger than p for any $p > 0$, a non-negative vector $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_d)$ and

$$\partial^\alpha g(x) = \frac{\partial^\alpha g(x)}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_d^{\alpha_d}}.$$

We let $\Lambda_2(p, L)$ be the Sobolev space of smoothness p with radius L and support $\mathcal{X}$, where the underlying measure is Lebesque measure.

3. A Special Form of NPIV Models: Well-posedness

In this section, we formulate Q-function estimation under the framework of a nonparametric instrumental variables (NPIV) model, which has been extensively studied in econometrics (e.g., (Ai & Chen, 2003; Newey & Powell, 2003; Blundell et al., 2007)). A generic NPIV model takes the expression as

$$Y = h_0(X) + U, \quad \text{with} \quad \mathbb{E}[U|W] = 0, \tag{2}$$

where h_0 is an unknown function to estimate, X is called endogenous variables, W is called instrumental variables, and U represents some random error. Motivated by Equation (1), we consider the following special form of a NPIV model with Assumptions 2.1-2.3 for Q-function estimation:

$$R_t = h^\pi(S_t, A_t, S_{t+1}; Q^\pi) + U_t, \text{with} \quad \mathbb{E}[U_t|S_t, A_t] = 0 \tag{3}$$

for $0 \le t \le T-1$, where

$$h^\pi(s, a, s'; Q) = Q(s, a) - \gamma \int_{a' \in \mathcal{A}} \pi(a'|s') Q(s', a') \mathrm{d}a'.$$

We also write $h^\pi(s, a, s'; Q)$ as $h^\pi(Q)(s, a, s')$ and $h_0^\pi = h^\pi(Q^\pi)$ when there is no confusion. By requiring $\mathbb{E}[U_t|S_t, A_t] = 0$ for $0 \le t \le T-1$, we recover the Bellman Equation (1). Therefore Model (3) can be used to estimate Q^π nonparametrically, where S_{t+1} can be understood as endogenous variables and (S_t, A_t) as instrumental variables under the framework of the NPIV model. Let $L^2(S, A)$ be the space of square integrable functions against the probability measure with density $\bar{d}_T^{\pi^b}$ and $L^2(S, A, S')$ against the probability measure with density $\bar{d}_T^{\pi^b} \times q$. Denote the conditional expectation operator by $\mathcal{T} : L^2(S, A, S') \to L^2(S, A)$, i.e., for every $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\mathcal{T} f(s, a) = \mathbb{E}[f(S, A, S')|S = s, A = a]$$

and in particular,

$$\mathcal{T} h^\pi(Q)(s, a) = \mathbb{E}[h^\pi(S, A, S'; Q)|S = s, A = a]\,.$$

3.1. Well-posedness in sup-norm

In this subsection, we show that Q-function estimation is in general well-posed in sup-norm given by the following lemma.

Lemma 3.1. *For any discount factor $0 \le \gamma < 1$ and any uniformly bounded function Q defined over $(\mathcal{S}, \mathcal{A})$, the following inequalities hold.*

$$\frac{1}{1+\gamma} \|h^\pi(Q - Q^\pi)\|_\infty \le \|Q - Q^\pi\|_\infty \tag{4}$$
$$\le \frac{1}{1-\gamma} \|\mathcal{T} h^\pi(Q - Q^\pi)\|_\infty \le \frac{1}{1-\gamma} \|h^\pi(Q - Q^\pi)\|_\infty.$$

Lemma 3.1 implies that to obtain the sup-norm rate for $\widehat{Q}^\pi$, it is sufficient to focus on $\|h^\pi(\widehat{Q}^\pi - Q^\pi)\|_\infty$, which is the sup-norm of so-called *temporal difference* error. One key reason of having such an inequality is the fact that Bellman operator is γ-contractive with respect to the sup-norm. However, it is hard to develop an estimator that minimizes the sup-norm of Bellman error in the batch setting so as to directly bound the sup-norm. Instead most existing methods are focused on minimizing the L^2-norm of Bellman error. This motivates us to study the well-posedness in L^2-norm below.

3.2. Well-posedness in L^2-norm

Lemma 3.1 in general may not hold for L^2-norm with respect to the data generating process (e.g., $\bar{d}_T^{\pi^b}$) due to the distributional mismatch between the behavior policy and the target one, which is one fundamental barrier in analyzing OPE problem in the literature as discussed in the introduction. To characterize the difficulty of L^2-estimating Q^π under Model (3), we define a *L^2-measure of ill-posedness* as

$$\overline{\tau} = \sup_{Q \in L^2(S,A)} \frac{\|h^\pi(Q)\|_{L^2(S,A,S')}}{\|\mathcal{T} h^\pi(Q)\|_{L^2(S,A)}}. \tag{5}$$

It can be seen that $\overline{\tau} \ge 1$ and could be arbitrarily large in general, which can be used to quantify the level of ill-posedness in estimating Q^π. We impose the following mild assumption to ensure the well-posedness in L^2-norm in the sense of $\overline{\tau} \lesssim 1$.

Assumption 3.2. (a) There exist positive constants $p_{\min}$ and $p_{1,\max}$ such that the average visitation probability density function $\bar{d}_T^{\pi^b}$ satisfies $p_{\min} \le \bar{d}_T^{\pi^b}(s, a) \le p_{1,\max}$ for every $(s, a) \in \mathcal{S} \times \mathcal{A}$. (b) The target policy π is absolutely continuous with respect to π^b and $q^\pi(s', a' \mid s, a) \le p_{2,\max}$ for some positive constant $p_{2,\max}$.

Let $p_{\max} = \max(p_{1,\max}, p_{2,\max})$. In general, boundedness assumption on the data generating probability density in Assumption 3.2 (a) is standard in the classical non-parametric estimation such as (Huang et al., 1998; Chen & Christensen, 2015). In our setting, that the average visitation probability density is uniformly bounded away from 0 is also called *coverage* assumption frequently used in RL literature such as (Precup et al., 2000; Antos et al., 2008a; Kallus & Uehara, 2019) among many others. This assumption can be relaxed to the so-called *partial coverage* if one is willing to impose some structure assumption on Q^π. See recent studies in (Duan et al., 2020; Xie et al., 2021; Agarwal et al., 2021; Uehara & Sun, 2021). Assumption 3.2 (b) imposes one mild identification condition on the target policy. It essentially states that our batch data are able to identify the value of the target policy. Lastly, we remark that when both $\mathcal{S}$ and $\mathcal{A}$ are discrete and finite, Assumption 3.2 (b) is automatically satisfied because of Assumption 3.2 (a). In the following,

we use $\|\bullet\|_{2,\nu}$ to denote L^2-norm with respect to some probability distribution/density ν.

Now we are ready to present a key theorem in this paper, which can not only be used to establish the minimax-optimal sup-norm and L^2-norm rates for estimating Q^π, but also provide a foundation for many existing OPE estimators.

Theorem 3.3. *For any policy π, discount factor $0 \le \gamma < 1$, and any two square integrable functions Q_1 and Q_2 defined over $(\mathcal{S}, \mathcal{A})$ with respect to $\bar{d}_T^{\pi^b}$, under Assumptions 2.1, 2.3 and 3.2, the following inequalities hold.*

$$\sqrt{\frac{p_{\min}}{p_{\max}}}(1-\gamma)\|Q_1 - Q_2\|_{2,\bar{d}_T^{\pi^b}} \tag{6}$$
$$\le \|\mathcal{T}h^\pi(Q_1 - Q_2)\|_{2,\bar{d}_T^{\pi^b}} \le \|h^\pi(Q_1 - Q_2)\|_{2,\bar{d}_T^{\pi^b}\times q}.$$

In particular, the L^2 measure of ill-posedness

$$\overline{\tau} \lesssim \frac{\sqrt{p_{\max}(1 + \frac{p_{\max}\gamma^2}{p_{\min}})}}{\sqrt{p_{\min}}(1-\gamma)} \lesssim 1.$$

Theorem 3.3 rigorously justifies the validity of using L^2-norm to measure the Bellman error, which has been widely adopted in the existing literature for constructing various estimators for the Q-function. To see this, let $Q_1 = Q^\pi$ and $Q_2 = \widetilde{Q}$ in Theorem 3.3, where $\widetilde{Q}$ denotes some estimator for Q^π. Then the first inequality in (6) with Bellman equation (1) implies that

$$\|\widetilde{Q} - Q^\pi\|_{2,\bar{d}_T^{\pi^b}} \lesssim \|r + (\gamma\mathcal{P}^\pi - \mathcal{I})\widetilde{Q}\|_{2,\bar{d}_T^{\pi^b}},$$

where the right hand side of the above inequality is called Bellman error (or residual) and recall that r is the reward function defined in Assumption 2.2. Therefore L^2-norm of Bellman error of any Q-function estimator provides a valid upper bound for the L^2 error bound of this estimator to the true Q^π. Many existing estimators such as (Antos et al., 2008b; Farahmand et al., 2016; Uehara & Jiang, 2019; Feng et al., 2020) indeed are based on minimizing the L^2-norm of Bellman error. Therefore our Theorem 3.3 provides a theoretical guarantee for their procedures. Notice that Theorem 3.3 is established without imposing any restriction on the structure of Q-function, it can be used to obtain L^2 error bounds for many different non-parametric estimators of Q-function obtained using different models and/or methods such as LSTD, kernel methods or neural networks. For example, combining our Theorem 3.3 with Theorem 11 of (Farahmand et al., 2016) immediately gives L^2-error bound for their estimator to the true Q^π. Applying our Theorem 3.3 to Example 6 of (Uehara et al., 2021) one can obtain L^2-error bound for their neural network estimator to Q^π.

We also remark that the well-posed result in Theorem 3.3 can be extended to other metrics such as L^1-norm, based on which one may develop a new estimator for Q-function by minimizing the empirical approximation of L^1-norm of Bellman error. We conjecture that such an estimator could achieve robustness compared with the existing ones, especially when the reward distribution is heavy tailed. Lastly, there is a very recent work (Wang et al., 2022), which developed a sufficient and necessary condition for establishing the well-posedness of Bellman operator in terms of L^2-norm with respect to the data generating process. Besides they also developed some sufficient conditions that are similar to our Assumption 3.2 in establishing this well-posedness.

4. Minimax Lower Bounds

In this section, we establish minimax lower bounds in both sup-norm and in L^2-norm for estimation of nonparametric Q-function in OPE problem. The well-posedness property essentially indicates that non-parametric Q-function estimation is as easy as the classical non-parametric regression in the i.i.d. setting in terms of the worst case rate.

Recall that by Theorem 3.3, under Assumptions 2.1, 2.3 and 3.2, for any square integrable function Q defined over $\mathcal{S}\times\mathcal{A}$, we have

$$\sqrt{\frac{p_{\min}}{p_{\max}}}(1-\gamma)\|Q\|_{2,\bar{d}_T^{\pi^b}} \le \|\mathcal{T}h^\pi(Q)\|_{2,\bar{d}_T^{\pi^b}} \tag{7}$$
$$\le \|h^\pi(Q)\|_{2,\bar{d}_T^{\pi^b}\times q} \lesssim \|Q\|_{2,\bar{d}_T^{\pi^b}}.$$

Denote a generic transition tuple as $\{S_{i,t}, A_{i,t}, R_{i,t}, S'_{i,t}\}$ indexed by (i,t). Then we have the following lower bound results for estimating Q^π and its derivative in terms of the sup-norm.

Theorem 4.1. *Let d^ν be the average visitation probability density defined over $\mathcal{S}\times\mathcal{A}$ induced by some policy ν such that Assumption 3.2 holds with $\bar{d}_T^{\pi^b}$ and π^b replaced by d^ν and ν respectively. Suppose the data $\mathcal{D}_N = \{S_{i,t}, A_{i,t}, R_{i,t}, S'_{i,t}\}_{1\le i\le N, 0\le t\le T-1}$ are i.i.d. from Model (3), where the probability density of $(S_{i,t}, A_{i,t})$ is d^ν with the transition probability density q and for every $0 \le t \le T-1$ and $1 \le i \le N$, $\mathbb{E}[U_{i,t}^2 \,|\, S_{i,t}, A_{i,t}] \ge \sigma^2$, where σ is some positive constant, then we have for any $0 \le \|\alpha\|_{\ell_1} < p$,*

$$\liminf_{NT\to\infty} \inf_{\widehat{Q}} \sup_{Q\in\Lambda_\infty(p,L)} \Pr^Q\Big(\|\partial^\alpha\widehat{Q} - \partial^\alpha Q\|_\infty \tag{8}$$
$$\ge c(\log(NT)/NT)^{(p-\|\alpha\|_{\ell_1})/(2p+d)}\Big) \ge c' > 0,$$

for some constants c and c', where $\inf_{\widehat{Q}}$ denotes the infimum over all estimators using $\mathcal{D}_N$, and $\Pr^Q$ denotes the joint probability distribution of $\mathcal{D}_N$ with $h^\pi = h^\pi(Q)$ in Model (3).

The following theorem provides lower bound results in terms of L^2-norm.

Theorem 4.2. *Under all conditions in Theorem 4.1, for* $0 \leq \|\alpha\|_{\ell_1} < p$, *we have*

$$\liminf_{NT\to\infty} \inf_{\widehat{Q}} \sup_{Q\in\Lambda_2(p,L)} \Pr^{Q}\left(\|\partial^{\alpha}\widehat{Q} - \partial^{\alpha}Q\|_2 \geq \bar{c}(NT)^{(\|\alpha\|_{\ell_1}-p)/(2p+d)}\right) \geq \bar{c}' > 0, \tag{9}$$

for some constant $\bar{c}$ *and* $\bar{c}'$.

As we can see from Theorems 4.1 and 4.2, the minimax lower bounds for the rates of estimating Q-function and its derivatives are the same as those for nonparametric regression in the i.i.d setting (Stone, 1982). To the best of our knowledge, these are the first lower bound results for nonparametrically estimating Q-function and its derivatives in the infinite-horizon MDP. In the following section, we proposed simple estimators that match these lower bounds.

5. Sieve 2SLS Estimation of Q-function

Given the NPIV Model (3) as a reformulation of Bellman equation, we now adopt the idea from for example (Blundell et al., 2007; Chen & Christensen, 2018) to construct a sieve 2SLS estimator for Q^{π}. Define two sieve basis functions as

$$\psi^J(s,a) = (\psi_{J1}(s,a), \cdots, \psi_{JJ}(s,a))^{\top}, \tag{10}$$

$$b^K(s,a) = (b_{K1}(s,a), \cdots, b_{KK}(s,a))^{\top}, \tag{11}$$

to model Q^{π} and the space of instrumental variables respectively. Let $\Psi_J = \text{closure}\{\Psi_{J1}, \ldots, \Psi_{JJ}\} \subset L^2(S,A)$ and $B_K = \text{closure}\{b_{K1}, \ldots, b_{KK}\} \subset L^2(S,A)$ denote the sieve spaces for Q^{π} and instrumental variables, respectively. Here the underlying probability measure of $L^2(S,A)$ is $\bar{d}_T^{\pi^b}$. Examples of basis functions include splines or wavelet bases (See more examples in (Huang et al., 1998; Chen, 2007)). The construction of wavelet bases can also be found in Appendix C. We remark that the numbers of basis functions J and K are allowed to grow with either N or T, but require that $J \leq K \leq cJ$ for some $c \geq 1$. Due to the special structure of Model (3), it also makes sense to simply let $K = J$ and $\psi^J = b^K$. Additionally, we let $\psi_{\pi}^J(s) = (\int_{a\in\mathcal{A}} \pi(a|s)\psi_{J1}(s,a)\mathrm{d}a, \cdots, \int_{a\in\mathcal{A}} \pi(a|s)\psi_{JJ}(s,a)\mathrm{d}a)^{\top}$. Correspondingly, a sample version of all these functions can be defined as

$$\begin{aligned}
&\Psi = (\psi^J(S_{1,1}, A_{1,1}) \cdots, \\
&\psi^J(S_{N,T-1}, A_{N,T-1}))^{\top} \in \mathbb{R}^{(NT)\times J}, \\
&B = (b^K(S_{1,0}, A_{1,0}), b^K(S_{1,1}, A_{1,1}) \cdots, \\
&b^K(S_{N,T-1}, b_{N,T-1}))^{\top} \in \mathbb{R}^{(NT)\times K}, \\
&G_{\pi} = (\psi_{\pi}^J(S_{1,1}), \psi_{\pi}^J(S_{1,2}) \cdots, \\
&\psi_{\pi}^J(S_{1,T}), \psi_{\pi}^J(S_{2,1}), \cdots, \psi_{\pi}^J(S_{N,T}))^{\top} \in \mathbb{R}^{(NT)\times J}.
\end{aligned}$$

For notational simplicity, let $\kappa_{\pi}^J(s,a,s') = \psi^J(s,a) - \gamma\psi_{\pi}^J(s')$, and correspondingly $\Gamma_{\pi} = \Psi - \gamma G_{\pi}$. We also denote $\kappa_{J_j}^{\pi}(s,a,s') = \psi_{Jj}(s,a) - \gamma\int_{a'\in\mathcal{A}} \pi(a' \mid s')\mathrm{d}a'\psi_{J_j}^{\pi}(s',a')$ for each element of $\kappa_{\pi}^J(s,a,s') \in \mathbb{R}^J$. Then the sieve 2SLS estimator for Q^{π} can be constructed as

$$\widehat{Q}^{\pi}(s,a) = \psi^J(s,a)^{\top}\widehat{c}, \tag{12}$$

$$\text{with} \quad \widehat{c} = \left[\Gamma_{\pi}^{\top}B(B^{\top}B)^{-}B^{\top}\Gamma_{\pi}\right]^{-}\Gamma_{\pi}^{\top}B(B^{\top}B)^{-}B^{\top}\mathbf{R},$$

where $(Z)^{-}$ denotes the generalized inverse of some matrix Z and $\mathbf{R} = (R_{1,0}, R_{1,1} \cdots, R_{N,T-1})^{\top} \in \mathbb{R}^{NT\times 1}$. The corresponding estimator for the derivatives of Q^{π} is denoted by $\partial^{\alpha}\widehat{Q}^{\pi}$ for any vector α. Here $\widehat{c}$ can be understood as a minimizer of the following optimization problem.

$$\underset{c\in\mathbb{R}^J}{minimize} \quad \|B(B^{\top}B)^{-1}B^{\top}(\mathbf{R} - \Gamma_{\pi}c)\|_{\ell_2}^2.$$

Note that the sieve 2SLS estimator given in (12) becomes the solution of the modified Bellman residual minimiazion in (Farahmand et al., 2016) when their function spaces are modeled by sieve ones.

5.1. Sieve measure of ill-posedness in NPIV

An important quantity related to a generic NPIV model (2) is called *sieve* L^2 *measure of ill-posedness*, which characterizes the difficulty of non-parametrically estimating h_0 using the sieve estimation. Here a similar measure of ill-posedness can be defined under Model (3). Let $\Theta_J^{\pi} = \{h^{\pi}(Q) \in L^2(S,A,S') : Q \in \Psi_J\}$. Adapting from the sieve L^2 measure of ill-posedness in (Blundell et al., 2007), we define an average *sieve* L^2 *measure of ill-posedness* across T decision points under Model (3) as

$$\tau_J = \sup_{h\in\Theta_J^{\pi}: h\neq 0} \frac{\|h\|_{L^2(S,A,S')}}{\|\mathcal{T}h\|_{L^2(S,A)}}. \tag{13}$$

It can be seen that $\tau_J \geq 1$. Basically τ_J measures how much information has been smoothed out by the conditional expectation operator $\mathcal{T}$ over the space Θ_J^{π}. For a generic NPIV model (2), τ_J grows to infinity as J goes to infinity; see, e.g. (Blundell et al., 2007; Chen & Christensen, 2018). By definition we have $\tau_J \leq \bar{\tau} \lesssim 1$ for all $J \geq 1$. Thus Theorem 3.3 directly implies that the NPIV Model (3) is also well-posed under the L^2 sieve measure of ill-posedness defined in (13). Based on this result, minimax-optimal sup-norm and L^2-norm rates for the sieve 2SLS estimator of Q^{π} can be established in the following subsections.

5.2. Sup-norm Convergence Rates

In this subsection, we establish the sup-norm convergence rate of $\widehat{Q}^{\pi}$ to Q^{π}. We first introduce an additional assumption on the data generating process.

Assumption 5.1. The stochastic process $\{S_t, A_t\}_{t\geq 0}$ induced by the behavior policy π^b is a stationary, exponentially $\boldsymbol{\beta}$-mixing stochastic process. The $\boldsymbol{\beta}$-mixing coefficient at time lag k satisfies that $\beta_k \leq \beta_0 \exp(-\beta_1 k)$ for

$\beta_0 \geq 0$ and $\beta_1 > 0$. The induced stationary density is denoted by d^{π^b}.

Assumption 5.1 is imposed to characterize the dependency among observations over time because the observed data modeled by MDP are *not* i.i.d. and transition tuples are dependent. Most of previous works in RL assume transition tuples are independent, which is stronger than Assumption 5.1 . The $\boldsymbol{\beta}$-mixing coefficient at time lag k basically means that the dependency between $\{S_t, A_t\}_{t\leq j}$ and $\{S_t, A_t\}_{t\geq (j+k)}$ decays to 0 at an exponential rate with respect to k. See (Bradley, 2005) for the exact definition of the exponentially $\boldsymbol{\beta}$-mixing. A fast mixing rate is imposed here mainly for technical simplicity and our sup-norm and L^2-norm convergence rates are not affected by the mixing coefficients. Indeed, Assumption 5.1 can be relaxed to stationary distribution with certain algebraic β-mixing, and by using the matrix Bernstein inequality for general β-mixing sequences developed by (Chen & Christensen, 2015), one may obtain the same convergence rates as those in Theorems 5.4 and 5.5 below. We can also relax the strictly stationary assumption with some extra notation. Since this is not our focus, we do not impose the weakest possible assumptions on the temporal dependence in this paper. When both Assumptions 3.2 and 5.1 hold, the average visitation probability density $\bar{d}_T^{\pi^b}$ used in Assumption 3.2 becomes the induced stationary density d^{π^b}. We will omit ν in $\| \bullet \|_{2,\nu}$ when $\nu = d^{\pi^b}$. Throughout the remaining of this section, unless otherwise specified, (S, A) has the probability density d^{π^b} and the density of (S, A, S') is $d^{\pi^b} \times q$.

Define $L_{2,h^\pi}(S, A, S') = \{h^\pi(Q) : Q \in L^2(S, A)\}$. Let $\Pi_J : L_{2,h^\pi}(S, A, S') \to \Theta_J^\pi$ denote the $L^2_{h^\pi}(S, A, S')$ mapping onto Θ_J^π, i.e., $\Pi_J h_0^\pi = h_0^\pi(\overline{\Pi}_J Q^\pi)$, where $\overline{\Pi}_J Q^\pi = \arg\min_{Q\in\Psi_J} \|Q^\pi - Q\|_{L^2(S,A)}$, and let $\Pi_K : L^2(S, A) \to B_K$ denote the $L^2(S, A)$ orthogonal projection onto B_K. Let $\widetilde{\Pi}_J h_0^\pi = \arg\min_{h\in\Theta_J^\pi} \|\Pi_K \mathcal{T}(h_0^\pi - h)\|_{L^2(S,A)}$ denote the sieve 2SLS projection of h_0^π onto Θ_J^π. Let $\Theta_{J,1}^\pi = \{h \in \Theta_J^\pi \mid \|h\|_{L^2(S,A,S')} = 1\}$. We make one additional assumption below for controlling the approximation error of using the sieve bases.

Assumption 5.2. (a) $\sup_{h\in\Theta_{J,1}^\pi} \|(\Pi_K\mathcal{T} - \mathcal{T})h\|_{L^2(S,A)} = o_J(1)$, where $o_J(1)$ refers to a quantity that converges to 0 when $J \to \infty$; (b) $\|\widetilde{\Pi}_J(h_0^\pi - \Pi_J h_0^\pi)\|_\infty \leq C_1 \times \|h_0^\pi - \Pi_J h_0^\pi\|_\infty$ for some constants C_1.

Assumption 5.2 (a) is a mild condition on approximating Θ_J^π by a sieve space B_K. For fixed J (and K), $\sup_{h\in\Theta_{J,1}^\pi} \|(\Pi_K\mathcal{T} - \mathcal{T})h\|_{L^2(S,A)}$ can be interpreted as an inherent Bellman error (for a fixed policy π), which is widely used in the literature of RL such as the analysis of fitted-q iteration (See Assumption 4.2 of (Agarwal et al., 2019)). Assumption 5.2 (b) is also mild because that $\|\widetilde{\Pi}_J(h_0^\pi - \Pi_J h_0^\pi)\|_{L^2(S,A)} \leq \|h_0^\pi - \Pi_J h_0^\pi\|_{L^2(S,A)}$ holds automatically by the projection property. Here we strengthen it in terms of the sup-norm.

To derive the sup-norm convergence rate, following the proof of (Chen & Christensen, 2018), we split $\|\widehat{Q}^\pi - Q^\pi\|_\infty$ into two terms. Let $\widetilde{Q}^\pi(s, a) = \psi^J(s, a)^\top \widetilde{c}$ with $\widetilde{c} = [\Gamma_\pi^\top B(B^\top B)^- B^\top \Gamma_\pi]^- \Gamma_\pi^\top B(B^\top B)^- B^\top H_0$, where

$$H_0 = (h_0^\pi(S_{1,0}, A_{1,0}, S_{1,1}), h_0^\pi(S_{1,1}, A_{1,1}, S_{1,2}), \cdots, h_0^\pi(S_{N,T-1}, A_{N,T-1}, S_{N,T}))^\top \in \mathbb{R}^{NT}.$$

Then by triangle inequality, we have $\|\widehat{Q}^\pi - Q^\pi\|_\infty \leq \|\widehat{Q}^\pi - \widetilde{Q}^\pi\|_\infty + \|Q^\pi - \widetilde{Q}^\pi\|_\infty$. The first term $\|\widehat{Q}^\pi - \widetilde{Q}^\pi\|_\infty$ can be interpreted as an estimation error, while the second term $\|Q^\pi - \widetilde{Q}^\pi\|_\infty$ can be understood as the approximation error. Denote $G_{\kappa,J}^\pi = \mathbb{E}[\kappa_\pi^J(S, A, S')\kappa_\pi^J(S, A, S')^\top]$ and $e_J = \lambda_{\min}(G_{\kappa,J}^\pi)$. Let

$$\begin{aligned}
\zeta_\kappa^\pi &= \sup_{s,a,s'} \|[G_{\kappa,J}^\pi]^{-1/2}\kappa_\pi^J(s, a, s')\|_{\ell^2} \\
G_b &= \mathbb{E}[b^K(S, A)b^K(S, A)^\top] \\
\xi_\psi &= \sup_{s,a} \|\psi^J(s, a)\|_{\ell^1} \\
\zeta_b &= \sup_{s,a} \|G_b^{-1/2}b^K(s, a)\|_{\ell^2}
\end{aligned}$$

for each J and K by omitting their dependence on J and K for notation simplicity, and define $\zeta = \max\{\zeta_b, \zeta_\kappa^\pi\}$. In the following lemma, we derive bounds for the aforementioned two terms.

Lemma 5.3. *(1) Let Assumptions 2.1-2.3, 3.2, 5.1, and Assumption 5.2(a) hold. If $\zeta^2\sqrt{\log(NT)/NT} = O(1)$, then we have the following sup-norm bound for the estimation error.*

$$\|\widehat{Q}^\pi - \widetilde{Q}^\pi\|_\infty = O_p\Big(\xi_\psi\sqrt{(\log J)/(NTe_J)}\Big). \quad (14)$$

(2) Let Assumptions 5.1-5.2 hold. If $\zeta^2\sqrt{\log(J)\log(NT)/NT} = O(1)$ then the approximation error can be controlled by

$$\|Q^\pi - \widetilde{Q}^\pi\|_\infty = O_p(\|Q^\pi - \overline{\Pi}_J Q^\pi\|_\infty). \quad (15)$$

By examining the proof of Lemma 5.3, it is possible to derive the finite sample error bounds for both $\|\widehat{Q}^\pi - \widetilde{Q}^\pi\|_\infty$ and $\|Q^\pi - \widetilde{Q}^\pi\|_\infty$. We omit them for brevity.

Theorem 5.4. *Let Assumptions 2.1-2.3, 3.2, 5.1, and 5.2 hold and $Q^\pi \in \Lambda_\infty(p, L)$ for some L. Suppose that the sieve space Ψ_J is spanned by a B-spline or wavelet basis of (Cohen et al., 1993) with regularity larger than p, and B_K is spanned by a wavelet, spline or cosine basis. If $J\sqrt{\log(J)\log(NT)/NT} = O(1)$, then we have:*

$$\|\widehat{Q}^\pi - Q^\pi\|_\infty = O_p\big(J^{-p/d} + \sqrt{J(\log J)/(NT)}\big). \quad (16)$$

Further, by choosing $J \asymp (\frac{NT}{\log(NT)})^{d/(2p+d)}$ *and assuming* $2p > d$*, we have for all* $0 \leq \|\alpha\|_{\ell_1} < p$*,*

$$\|\partial^\alpha \widehat{Q}^\pi - \partial^\alpha Q^\pi\|_\infty = O_p((\frac{\log(NT)}{NT})^{\frac{p-\|\alpha\|_{\ell_1}}{2p+d}}). \quad (17)$$

The smoothness parameter p in Theorem 5.4 represents the smoothness level of the true Q-function and characterizes the size of the functional class that the true Q-function belongs to. We require $2p > d$ in Theorem 5.4 mainly for deriving the sup-norm rate in order to achieve the optimality when considering Hölder class of functions. Theorem 5.4 shows that in terms of the batch data sample size of NT, the sup-norm rate of our 2SLS estimator $\widehat{Q}^\pi$ to Q^π is the same as the optimal one in the classical non-parametric regression estimation (Stone, 1982) using B-splines. The sup-norm convergence rate of $\widehat{Q}^\pi$ can be useful to develop uniform confidence bands (UCBs) for the Q-function using results in (Chen & Christensen, 2018) for example. Such UCBs may be incorporated into the framework of pessimistic RL algorithms such as (Jin et al., 2021; Xie et al., 2021). In addition, we can also show that the sup-norm bounds on the constant factor γ is of order $(1-\gamma)^{-3}$. Finally, the results on the sup-norm rates for estimating the derivatives of the Q-function may be useful to some actor-critic algorithms such as (Silver et al., 2014; Kallus & Uehara, 2020; Xu et al., 2021).

5.3. L^2-norm Convergence Rates

In this subsection we present the L^2 convergence rates of our 2SLS estimator for the Q-function. We do not require Assumption 5.2 (b) as the L^2-stability condition holds automatically.

Theorem 5.5. *Let Assumptions 2.1-2.3, 3.2, 5.1, and 5.2 (a) hold. If* $\zeta\sqrt{\log(NT)\log(J)/NT} = o(1)$*, then:*

$$\|\widehat{Q}^\pi - Q^\pi\|_2 = O_p(\sqrt{J/(NT)} + \|Q^\pi - \overline{\Pi}_J Q^\pi\|_2). \quad (18)$$

If $Q^\pi \in \Lambda_2(p, L)$ *with* $p > 0$*, and* Ψ_J *and* B_K *are spanned by some commonly used bases such as polynomials, trigonometric polynomials, splines and wavelets with regularity greater than* p*, by choosing* $J \asymp (NT)^{d/(2p+d)}$*, we have: for all* $0 \leq \|\alpha\|_{\ell_1} < p$*,*

$$\|\partial^\alpha \widehat{Q}^\pi - \partial^\alpha Q^\pi\|_2 = O_p\left((NT)^{(\|\alpha\|_{\ell_1}-p)/(2p+d)}\right).$$

According to Theorem 5.5, the sieve 2SLS estimator $\widehat{Q}^\pi$ achieves the minimax optimal L^2-norm convergence rate to Q^π under conditions much weaker than those for the optimal sup-norm convergence rate. Let F be a known marginal distribution of the initial state. It is well-known that one can estimate the value of a target policy π, i.e.,$v^\pi = \int_{s\in\mathcal{S}}[\int_{a\in\mathcal{A}} \pi(a\,|\,s)Q^\pi(s,a)\mathrm{d}a]F(\mathrm{d}s)$ by a simple plug-in sieve 2SLS estimator $\widehat{v}^\pi = \int_{s\in\mathcal{S}}[\int_{a\in\mathcal{A}} \pi(a\,|\,s)\widehat{Q}^\pi(s,a)\mathrm{d}a]F(\mathrm{d}s)$. Theorem 5.5 is particularly useful in establishing the asymptotic normality of $\sqrt{NT}(\widehat{v}^\pi - v^\pi)$.

Remark 5.6. (1) Recently (Shi et al., 2020) presented a sieve LSTD estimator for Q^π and obtained the L^2-norm rate of convergence (See (E.46) in appendix of their paper for more details) under some conditions including their Assumption (A3.) or a small discount factor γ condition. They then apply their L^2-norm convergence rate to establish the $\sqrt{NT}$-asymptotic normality of plug-in sieve LSTD estimator for the value v^π. Note that their sieve LSTD is a special case of our sieve 2SLS with $B_K = \Psi_J$ and $K = J$, and the sieve LSTD automatically satisfies our Assumption 5.2 (a). Our Theorem 5.5 establishes the L^2-norm convergene rate for their sieve LSTD estimator without the need to impose the strong condition of a small discount factor γ (or Assumption (A3.)). Thus we may require weaker conditions for establishing the asymptotic normality of the plug-in sieve 2SLS estimator for the value. We leave details to the longer version of the paper. (2) In this paper, to obtain the optimal rates of convergence in L^2-norm (and sup-norm) of our sieve 2SLS estimator for the Q^π function, we assume strictly stationary data for simplicity. We note that (Shi et al., 2020) did not impose this strict stationarity in their L^2-norm rate and asymptotic normality calculation. However, they need to assume the distribution of the initial state $S_{i,0}$ in the batch data is bounded away from 0 uniformly in i. Indeed it is possible to replace the strict stationary condition in our Assumption 5.1 by imposing the geometric ergodicity and using the truncation argument to obtain the same sup-norm and L^2-norm convergence rates for our sieve 2SLS estimator. We leave it for the future work.

6. Conclusion

In this paper, we consider nonparametric estimation of Q-function of continuous states and actions in the OPE setting. Under some mild conditions, we show that the NPIV model (3) for estimating Q-function nonparametrically is well-posed in the sense of L^2-measure of ill-posedness, bypassing the need of imposing a strong condition on the discount factor γ in the recent literature. The well-posedness property effectively implies that the minimax lower bounds for nonparametric estimation of Q-function coincide with those for a nonparametric regression in sup-norm and in L^2-norm under the i.i.d. setting. Under mild sufficient conditions, we also establish that the sup-norm and the L^2-norm rates of convergence of our proposed sieve 2SLS estimators for Q-function achieve the lower bounds, and hence are minimax-optimal. These rate results are useful for optimal estimation and inference on various functionals, such as the value, of the Q-function by plugging in our sieve 2SLS

estimators. In particular, one can easily develop uniform confidence bands (UCBs) for the Q-function by slightly modifying the UCBs result in (Chen & Christensen, 2018) for a NPIV function estimated via a spline or wavelet sieve 2SLS. We leave this to future work due to the length of the paper.

In this paper we focus on the direct method of using Bellman equation to nonparametrically estimate Q-function in the OPE setting. In the existing literature, there are two additional model-free approaches to perform OPE. One is using the recently proposed marginal importance sampling for the infinite horizon setting such as (Liu et al., 2018; Nachum et al., 2019; Xie et al., 2019; Uehara et al., 2020; Zhang et al., 2020a;b). The other approach combines the direct method and marginal importance sampling to construct the so-called doubly robust estimators for the value of the target policy (see, e.g., (Kallus & Uehara, 2019; Tang et al., 2020; Shi et al., 2021) among many others). Our results on the well-posedness and the minimax lower bounds for Q function estimation should be useful to establish theoretical properties of these alternative approaches under conditions that are weaker than the existing ones. Finally, since OPE serves as the foundation of many RL algorithms, our results on Q-function estimation of a target policy can also be useful to other policy learning methods such as those proposed in (Ernst et al., 2005; Antos et al., 2008b; Le et al., 2019; Liao et al., 2020; Jin et al., 2021; Zanette et al., 2021). We leave details to future work.

References

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 2019.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

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

Antos, A., Szepesvári, C., and Munos, R. Fitted q-iteration in continuous action-space mdps. In *Advances in Neural Information Processing Systems*, volume 20, 2008a.

Antos, A., Szepesvári, C., and Munos, R. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008b. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-007-5038-2. URL http://link.springer.com/10.1007/s10994-007-5038-2.

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

Bradley, R. C. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

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

Chen, X. Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 6: 5549–5632, 2007.

Chen, X. and Christensen, T. M. Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, 188(2):447–465, 2015.

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 Reiss, M. On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory*, 27(3):497–521, 2011.

Chen, Y., Xu, L., Gulcehre, C., Paine, T. L., Gretton, A., de Freitas, N., and Doucet, A. On instrumental variable regression for deep offline policy evaluation. *arXiv preprint arXiv:2105.10148*, 2021.

Cohen, A., Daubechies, I., and Vial, P. Wavelets on the interval and fast wavelet transforms. *Applied and computational harmonic analysis*, 1993.

Darolles, S., Fan, Y., Florens, J.-P., and Renault, E. Nonparametric instrumental regression. *Econometrica*, 79(5): 1541–1565, 2011.

Dedecker, J. and Louhichi, S. Maximal inequalities and empirical central limit theorems. In *Empirical process techniques for dependent data*, pp. 137–159. Springer, 2002.

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

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

Ernst, D., Geurts, P., Wehenkel, L., and Littman, L. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.

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

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

Hall, P. and Horowitz, J. L. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33(6):2904 – 2929, 2005. doi: 10.1214/009053605000000714. URL https://doi.org/10.1214/009053605000000714.

Huang, J. Z. et al. Projection estimation in multiple regression with application to functional anova models. *Annals of Statistics*, 26(1):242–272, 1998.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021.

Kallus, N. and Uehara, M. Efficiently breaking the curse of horizon: Double reinforcement learning in infinite-horizon processes. *arXiv preprint arXiv:1909.05850*, 2019.

Kallus, N. and Uehara, M. Statistically efficient off-policy policy gradients. In *International Conference on Machine Learning*, pp. 5089–5100. PMLR, 2020.

Kosorok, M. R. and Laber, E. B. Precision medicine. *Annual review of statistics and its application*, 6:263–286, 2019.

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

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Liao, P., Dempsey, W., Sarker, H., Hossain, S. M., al'Absi, M., Klasnja, P., and Murphy, S. Just-in-time but not too much: Determining treatment timing in mobile health. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 2(4):179, 2018.

Liao, P., Qi, Z., and Murphy, S. Batch policy learning in average reward markov decision processes. *arXiv preprint arXiv:2007.11771*, 2020.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pp. 5356–5366, 2018.

Nachum, O., Chow, Y., Dai, B., and Li, L. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, pp. 2315–2325, 2019.

Newey, W. K. and Powell, J. L. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5): 1565–1578, 2003.

Pinto, L. and Gupta, A. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In *2016 IEEE international conference on robotics and automation (ICRA)*, pp. 3406–3413. IEEE, 2016.

Precup, D., Sutton, R. S., and Singh, S. P. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 759–766, 2000.

Shi, C., Zhang, S., Lu, W., and Song, R. Statistical inference of the value function for reinforcement learning in infinite horizon settings. *Journal of the Royal Statistical Society: Series B, in press*, 2020.

Shi, C., Wan, R., Chernozhukov, V., and Song, R. Deeply-debiased off-policy interval estimation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 9580–9591. PMLR, 2021.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In *International conference on machine learning*, pp. 387–395. PMLR, 2014.

Stone, C. J. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pp. 1040–1053, 1982.

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

Tang, Z., Feng, Y., Li, L., Zhou, D., and Liu, Q. Doubly robust bias reduction in infinite horizon off-policy estimation. In *International Conference on Learning Representations*, 2020.

Thomas, P. S., Theocharous, G., Ghavamzadeh, M., Durugkar, I., and Brunskill, E. Predictive off-policy policy evaluation for nonstationary decision problems, with applications to digital marketing. In *Twenty-Ninth IAAI Conference*, 2017.

Tropp, J. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.

Tropp, J. A. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer, 2009.

Uehara, M. and Jiang, N. Minimax weight and q-function learning for off-policy evaluation. *arXiv preprint arXiv:1910.12809*, 2019.

Uehara, M. and Sun, W. Pessimistic model-based offline rl: Pac bounds and posterior sampling under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

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

Uehara, M., Imaizumi, M., Jiang, N., Kallus, N., Sun, W., and Xie, T. Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*, 2021.

Wang, J., Qi, Z., and Wong, R. K. Projected state-action balancing weights for offline reinforcement learning. *arXiv preprint arXiv:2109.04640*, 2022.

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

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *arXiv preprint arXiv:2106.06926*, 2021.

Xu, T., Yang, Z., Wang, Z., and Liang, Y. Doubly robust off-policy actor-critic: Convergence and optimality. *arXiv preprint arXiv:2102.11866*, 2021.

Zanette, A., Wainwright, M. J., and Brunskill, E. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34, 2021.

Zhang, R., Dai, B., Li, L., and Schuurmans, D. Gendice: Generalized offline estimation of stationary values. *arXiv preprint arXiv:2002.09072*, 2020a.

Zhang, S., Liu, B., and Whiteson, S. Gradientdice: Rethinking generalized offline estimation of stationary values. In *International Conference on Machine Learning*, pp. 11194–11203. PMLR, 2020b.

A. Proofs in Section 3

A.1. Proof of Lemma 3.1

It is sufficient to show that $\|Q - Q^\pi\|_\infty \leq \frac{1}{1-\gamma}\|\mathcal{T}h^\pi(Q - Q^\pi)\|_\infty$, while other inequalities can be readily seen. It can be observed that

$$\|Q - Q^\pi\|_\infty \leq \|\mathcal{T}h^\pi(Q - Q^\pi)\|_\infty + \gamma\|\mathbb{E}^\pi\left[(Q - Q^\pi)(S', A') \,|\, S = \bullet, A = \bullet\right]\|_\infty \tag{19}$$

$$\leq \|\mathcal{T}h^\pi(Q - Q^\pi)\|_\infty + \gamma\|Q - Q^\pi\|_\infty, \tag{20}$$

where the first line follows the triangle inequality. This immediately implies

$$\|Q - Q^\pi\|_\infty \leq \frac{1}{1-\gamma}\|\mathcal{T}h^\pi(Q - Q^\pi)\|_\infty. \tag{21}$$

A.2. Proof of Theorem 3.3

For the first statement of Theorem 3.3, it is enough to focus on the first inequality, while the second one is given by Jensen's inequality. Let $\mathcal{I}$ be the identity operator and $\mathcal{P}^\pi$ be the operator such that $\mathcal{P}^\pi f(s,a) = \mathbb{E}^\pi\left[f(S_{t+1}, A_{t+1}) \,|\, S_t = s, A_t = a\right]$ for any $t \geq 0$. By induction, we can show that $(\mathcal{P}^\pi)^k f(s,a) = \mathbb{E}^\pi\left[f(S_{t+k}, A_{t+k}) \,|\, S_t = s, A_t = a\right]$. For some integer $\bar{t}$, which will be specified later, we have

$$\|Q_1 - Q_2\|_{2,\bar{d}_T^{\pi^b}} \leq \underbrace{\|\left(\mathcal{I} - \gamma^{\bar{t}}(\mathcal{P}^\pi)^{\bar{t}}\right)(Q_1 - Q_2)\|_{2,\bar{d}_T^{\pi^b}}}_{(I)} + \gamma^{\bar{t}}\underbrace{\|(\mathcal{P}^\pi)^{\bar{t}}(Q_1 - Q_2)\|_{2,\bar{d}_T^{\pi^b}}}_{(II)}.$$

We first focus on deriving an upper bound for (II). By Jensen's inequality, we can show that

$$\begin{aligned}
\{(II)\}^2 &\leq \int_{s\in\mathcal{S},a\in\mathcal{A}} \mathbb{E}^\pi\left[(Q_1 - Q_2)^2(S_{\bar{t}}, A_{\bar{t}}) \,|\, S_0 = s, A_0 = a\right] \bar{d}_T^{\pi^b}(s,a)\mathrm{d}s\mathrm{d}a \\
&= \int_{s\in\mathcal{S},a\in\mathcal{A}} \int_{s'\in\mathcal{S},a'\in\mathcal{A}} (Q_1 - Q_2)^2(s',a') q_{\bar{t}}^\pi(s', a' \,|\, s, a)\mathrm{d}s'\mathrm{d}a' \bar{d}_T^{\pi^b}(s,a)\mathrm{d}s\mathrm{d}a \\
&= \int_{s'\in\mathcal{S},a'\in\mathcal{A}} (Q_1 - Q_2)^2(s',a') \widetilde{q}_{T;\bar{t}}^{\pi^b;\pi}(s',a')\mathrm{d}s'\mathrm{d}a' \\
&= \int_{s'\in\mathcal{S},a'\in\mathcal{A}} (Q_1 - Q_2)^2(s',a') \frac{\widetilde{q}_{T;\bar{t}}^{\pi^b;\pi}(s',a')}{\bar{d}_T^{\pi^b}(s',a')} \bar{d}_T^{\pi^b}(s',a')\mathrm{d}s'\mathrm{d}a' \\
&\leq \frac{p_{\max}}{p_{\min}}\|Q_1 - Q_2\|^2_{2,\bar{d}_T^{\pi^b}},
\end{aligned}$$

where $\widetilde{q}_{T;\bar{t}}^{\pi^b;\pi}(s',a')$ refers to the marginal probability density function by composition between $\bar{d}_T^{\pi^b}$ and $q_{\bar{t}}^\pi$. The last equation holds because $\widetilde{q}_{T;\bar{t}}^{\pi^b;\pi}$ is absolutely continuous with respect to $\bar{d}_T^{\pi^b}(s',a')$ by Assumption 3.2. The last inequality is also given by Assumption 3.2 since $\widetilde{q}_{T;\bar{t}}^{\pi^b;\pi}(s',a') = \overline{\mathbb{E}}\left[q_{\bar{t}}^\pi(s',a' \,|\, S, A)\right] \leq p_{\max}$ for every $(s,a) \in \mathcal{S}\times\mathcal{A}$ (As long as one-step transition density is bounded above, $\bar{t}$-step will also be bounded above.). Now for any $\varepsilon > 0$, we can choose $\bar{t}$ sufficiently large such that

$$\gamma^{\bar{t}}\sqrt{p_{\max}/p_{\min}} \leq \varepsilon,$$

which implies that $\gamma^{\bar{t}} \times (II) \leq \varepsilon\|Q_1 - Q_2\|_{2,\bar{d}_T^{\pi^b}}$. This further shows that

$$\|Q_1 - Q_2\|_{2,\bar{d}_T^{\pi^b}} \leq (1-\varepsilon)^{-1} \times (I).$$

In the following, we derive an upper bound for (I). Let $g = (\mathcal{I} - \gamma\mathcal{P}^\pi)(Q_1 - Q_2)$. By a similar argument as before, we have

$$\begin{aligned}
(I) &= \| \left(\mathcal{I} - \gamma\mathcal{P}^\pi + \gamma\mathcal{P}^\pi - \gamma^2(\mathcal{P}^\pi)^2 + \cdots + \gamma^{\bar{t}-1}(\mathcal{P}^\pi)^{\bar{t}-1} - \gamma^{\bar{t}}(\mathcal{P}^\pi)^{\bar{t}}\right)(Q_1 - Q_2) \|_{2,\bar{d}_T^{\pi_b}} \\
&\leq \sum_{k=0}^{\bar{t}-1} \gamma^k \|(\mathcal{P}^\pi)^k(\mathcal{I} - \gamma\mathcal{P}^\pi)(Q_1 - Q_2)\|_{2,\bar{d}_T^{\pi_b}} \\
&= \sum_{k=0}^{\bar{t}-1} \gamma^k \|(\mathcal{P}^\pi)^k g\|_{2,\bar{d}_T^{\pi_b}} \\
&\leq \sum_{k=0}^{\bar{t}-1} \gamma^k \sqrt{\frac{p_{\max}}{p_{\min}}} \|g\|_{2,\bar{d}_T^{\pi_b}} \\
&\leq \frac{1-\gamma^{\bar{t}}}{1-\gamma} \sqrt{\frac{p_{\max}}{p_{\min}}} \|g\|_{2,\bar{d}_T^{\pi_b}}.
\end{aligned}$$

Summarizing together, we can obtain that

$$\|Q_1 - Q_2\|_{2,\bar{d}_T^{\pi_b}} \leq \frac{(1-\varepsilon)^{-1}(1-\gamma^{\bar{t}})}{1-\gamma} \sqrt{\frac{p_{\max}}{p_{\min}}} \|\mathcal{T}h^\pi(Q_1 - Q_2)\|_{2,\bar{d}_T^{\pi_b}},$$

where we note that $\mathcal{T}h^\pi(Q_1 - Q_2) = g$. Since ε is arbitrary, let ε go to 0, we have

$$\|Q_1 - Q_2\|_{2,\bar{d}_T^{\pi_b}} \leq \frac{1}{1-\gamma} \sqrt{\frac{p_{\max}}{p_{\min}}} \|\mathcal{T}h^\pi(Q_1 - Q_2)\|_{2,\bar{d}_T^{\pi_b}}$$

In the remaining proof, we show $\overline{\tau}$ is bounded above. Note that for any $Q \in L^2(S, A)$,

$$\begin{aligned}
\|h^\pi(Q)\|^2_{L^2(S,A,S')} &= \overline{\mathbb{E}}\left[\left(Q(S,A) - \gamma \int_{a' \in \mathcal{A}} \pi(a' \mid S')Q(S', a')\mathrm{d}a'\right)^2\right] \\
&\lesssim 2\overline{\mathbb{E}}\left[(Q(S,A))^2\right] + \frac{2p_{\max}\gamma^2}{p_{\min}} \int Q^2(s,a)\bar{d}_T^{\pi_b}(s,a)\mathrm{d}s\mathrm{d}a \\
&\lesssim (1 + \frac{p_{\max}\gamma^2}{p_{\min}})\|Q\|^2_{2,\bar{d}_T^{\pi_b}},
\end{aligned}$$

where the first inequality is given by AM-GM, Jensen's inequalities and Assumption 3.2 by noting that $\bar{d}_{T+1}^{\pi_b}(s,a) \lesssim p_{\max}$ for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Then by the first inequality given in (6), we can show that

$$\overline{\tau} \lesssim \frac{\sqrt{p_{\max}(1 + \frac{p_{\max}\gamma^2}{p_{\min}})}}{(1-\gamma)\sqrt{p_{\min}}},$$

which concludes our proof.

B. Notations

In this section, we clarify several notations used in the remaining appendix. Unless specified, for any transition tuple (S, A, S'), the probability density of (S, A) is $\sim d^{\pi_b}$ and the probability density of S' given (S, A) is q. In addition, $\mathbb{E}$ refers to the expectation taken with respect to d^{π_b}. We recall the definition of some quantities below, which will appear in our proof.

$$\begin{array}{rcccl}
G_\kappa^\pi & = & G_{\kappa,J}^\pi & = & \mathbb{E}[\kappa_\pi^J(S,A,S')\kappa_\pi^J(S,A,S')^\top] = \mathbb{E}[\Gamma_\pi^\top \Gamma_\pi/(NT)] \\
G_b & = & G_{b,K} & = & \mathbb{E}[b^K(S,A)b^K(S,A)^\top] = \mathbb{E}[B^\top B/(NT)] \\
G_\psi & = & G_{\psi,J} & = & \mathbb{E}[\psi^J(S,A)\psi^J(S,A)^\top] \\
\Sigma^\pi & = & \Sigma_{K,J}^\pi & = & \mathbb{E}[b^K(S,A)\kappa_\pi^J(S,A,S')^\top] = \mathbb{E}[B^\top \Gamma_\pi/(NT)].
\end{array}$$

We assume that Σ^π has a full column rank J. Denote $e_J = \lambda_{\min}(G^\pi_{\kappa,J})$. Let

$$\zeta^\pi_\kappa = \zeta^\pi_{\kappa,J} = \sup_{s,a,s'} \|[G^\pi_{\kappa,J}]^{-1/2}\kappa^J_\pi(s,a,s')\|_{\ell^2} \qquad \zeta_b = \zeta_{b,K} = \sup_{s,a} \|G_b^{-1/2}b^K(s,a)\|_{\ell^2}$$
$$\xi^\pi_\kappa = \xi^\pi_{\kappa,J} = \sup_{s,a,s'} \|\kappa^J_\pi(s,a,s')\|_{\ell^1} \qquad \xi_\psi = \xi_J = \sup_{s,a} \|\psi^J_\pi(s,a)\|_{\ell^1}$$

for each J and K, and $\zeta = \max\{\zeta_{b,K}, \zeta^\pi_{\kappa,J}\}$. Define

$$(G_b^{-1/2}\Sigma^\pi)^-_l = \left[[\Sigma^\pi]^\top (G_b)^{-1}\Sigma^\pi\right]^{-1}[\Sigma^\pi]^\top (G_b)^{-1/2},$$

and similarly for $(\widehat{G}_b^{-1/2}\widehat{\Sigma}^\pi)^-_l$, where

$$\widehat{\Sigma}^\pi = \frac{B^\top \Gamma_\pi}{NT} \quad \text{and} \quad \widehat{G}_b = \frac{B^\top B}{NT}.$$

C. Lower bounds

In this section, the probability density of $(S_{i,t}, A_{i,t})$ is d^ν and the expectation is with respect to the density d^ν.

C.1. Lower bounds for Sup-norm Rates

The proof mainly follows that of Theorem 3.2 in (Chen & Christensen, 2018). Consider the Gaussian reduced form of NPIV model with known operator $\mathcal{T}$:

$$\begin{aligned} R_{i,t} &= \mathcal{T}h^\pi_0(S_{i,t}, A_{i,t}) + U_{i,t}, \\ U_{i,t} \mid (S_{i,t}, A_{i,t}) &\sim \mathcal{N}(0, \sigma^2(S_{i,t}, A_{i,t})), \end{aligned} \tag{22}$$

for $1 \le i \le N$ and $0 \le t \le T-1$. The known of operator $\mathcal{T}$ is equivalent to knowing the transition density q. By Lemma 1 of (Chen & Reiss, 2011), the minimax lower bound of Model (3) is no smaller than Model (22). In the following, we thus focus on Model (22) and make use of Theorem 2.5 of (Tsybakov, 2009).

We restrict $\mathcal{S} \times \mathcal{A} = [0,1]^d$. Let $\{\widetilde{\phi}_{j,k,G}, \widetilde{\psi}_{j,k,G}\}_{j,k,G}$ be a tensor-product wavelet basis of regularity larger than p for $L^2([0,1]^d)$, where j is the resolution level, $k = (k_1, k_2, \cdots, k_d) \in \{0, 1, \cdots, 2^j - 1\}^d$, and G is a vector indicating which element in a Daubechies pair $\{\phi, \psi\}$ is used. Note that ϕ has support $[-M+1, M]$ for some positive integer M. All these pairs are generated by CDV wavelets (Cohen et al., 1993). Following the proof of (Chen & Christensen, 2018), we consider a class of submodels around Q^π. In particular, for a given j, consider the wavelet space $(\mathcal{S} \times \mathcal{A})_j$, which consists of 2^{jd} functions $\{\widetilde{\psi}_{j,k,G}\}_{k \in \{0,\cdots,2^j-1\}^d}$ with G chosen as all ψ functions. For some constant r, consider $\{\widetilde{\psi}_{j,k,G}\}_{k \in \{r,\cdots,2^j-1-M\}^d}$ as interior wavelets and $\widetilde{\psi}_{j,k,G}(s,a) = \Pi_{m=1}^{d-1}\psi_{j,k_m}(s_m)\psi_{j,k_d}(a)$ for $k = (k_1, \cdots, k_d) \in \{r, \cdots, 2^j - 1 - M\}^d$, where $s = (s_1, \cdots, s_{d-1}) \in \mathcal{S}$, $a \in \mathcal{A}$ and $\psi_{j,k_m}(\bullet) = 2^{j/2}\psi(2^j(\bullet) - k_m)$ for $1 \le m \le d$. Then for sufficiently large j, there exists a set $\mathcal{I} \subseteq \{r, \cdots, (2^j - M - 1)\}^d$ of interior wavelets with $\text{Card}(\mathcal{I}) \gtrsim 2^{dj}$, where $\text{Card}(\bullet)$ refers to the cardinality, such that at least one coordinate of $\text{support}(\widetilde{\psi}_{j,k_1,G})$ and $\text{support}(\widetilde{\psi}_{j,k_2,G})$ is empty for all $k_1 \ne k_2 \in \{r, \cdots, 2^j - 1 - M\}^d$. In addition, we have $\text{Card}(\mathcal{I}) \lesssim 2^{jd}$ by definition.

Then for any $Q^\pi \in \Lambda_\infty(p, L)$ such that $\|Q^\pi\|_{\Lambda^p_\infty} \le L/2$, where $\|\bullet\|_{\Lambda^p_\infty}$ is the Besov norm and for each $i \in \mathcal{I}$, define

$$Q^\pi_i = Q^\pi + c_0 2^{-j(p+d/2)}\widetilde{\psi}_{j,i,G}.$$

Correspondingly, for every (s,a,s'), let

$$h^\pi_i(s,a,s') = h^\pi_0(s,a,s') + c_0 2^{-j(p+d/2)}\left(\widetilde{\psi}_{j,i,G}(s,a) - \gamma\int_{a' \in \mathcal{A}} \pi(a'|s')\widetilde{\psi}_{j,i,G}(s',a')\mathrm{d}a'\right),$$

where c_0 is some positive constant specified later. It can be seen that for all $i \in \mathcal{I}$,

$$\|c_0 2^{-j(p+d/2)}\widetilde{\psi}_{j,i,G}(s,a)\|_{\Lambda^p_\infty} \lesssim c_0.$$

Hence $\|Q^\pi_i\|_{\Lambda^p_\infty} \le L$ for sufficient small c_0.

For $i \in \{0\}^d \cup \mathcal{I}$, consider Model (22) with the true function h_i^π and define the joint probability density of $\{S_{j,t}, A_{j,t}, R_{j,t}, S'_{j,t}\}_{1\le j\le N, 0\le t\le T-1}$ as P_i such that

$$P_i = \Pi_{j=0}^{N}\Pi_{t=0}^{T-1} d^\nu(S_{j,0}, A_{j,0}) p_{h_i^\pi}(R_{j,t}|S_{j,t}, A_{j,t}) q(S'_{j,t} \mid S_{j,t}, A_{j,t}),$$

by recalling that they are $i.i.d.$ samples, where $p_{h_i^\pi}$ denotes the conditional density of reward given a state-action pair. In particular, when $i = \{0\}^d$, $Q_i^\pi = Q^\pi$ and $h_i^\pi = h^\pi$.

First of all, for sufficiently small c_0, we can show

$$\|c_0 2^{-j(p+d/2)} \widetilde{\psi}_{j,i,G}\|_{\Lambda_\infty^p} \lesssim c_0 \le L$$

for every $i \in \mathcal{I}$. In addition, by Equation (7), we have

$$\sqrt{p_{\min}/p_{\max}}(1-\gamma) c_0 2^{-j(p+d/2)} \le \|\mathcal{T} c_0 2^{-j(p+d/2)} \left(\widetilde{\psi}_{j,i,G}(S,A) - \gamma \int_{a'\in\mathcal{A}} \pi(a'|S') \widetilde{\psi}_{j,i,G}(S', a') \mathrm{d}a' \right) \|_2 \tag{23}$$

$$\lesssim c_0 2^{-j(p+d/2)}. \tag{24}$$

Secondly, for every $i \in \mathcal{I}$, the Kullback-Leibler distance $K(P_i, P_0)$ can be bounded as

$$K(P_i, P_0) \le \frac{1}{2} c_0^2 2^{-j(2p+d)} \sum_{m=1}^{N}\sum_{t=0}^{T-1} \mathbb{E}\left[\frac{\left(\mathcal{T}\left(\widetilde{\psi}_{j,i,G}(S_{m,t}, A_{m,t}) - \int_{a'\in\mathcal{A}} \pi(a'|S'_{m,t}) \widetilde{\psi}_{m,i,G}(S'_{m,t}, a') \mathrm{d}a' \right) \right)^2}{\sigma^2(S_{m,t}, A_{m,t})} \right] \tag{25}$$

$$\lesssim NT c_0^2 2^{-j(2p+d)}, \tag{26}$$

by the condition in Theorem 4.1. By choosing $2^j \asymp (NT/\log(NT))^{1/(2p+d)}$, it gives that

$$K(P_i, P_0) \lesssim c_0^2 \log(NT),$$

and $\log(\text{Card}(\mathcal{I})) \gtrsim j \gtrsim \log(NT) - \log\log(NT)$. So for sufficiently small c_0 and large NT, $K(P_i, P_0) \le 1/8 \log(\text{Card}(\mathcal{I}))$ for every $i \in \mathcal{I}$.

Lastly, it can be seen that for $i_1, i_2 \in \mathcal{I}$ and $i_1 \ne i_2$,

$$\begin{aligned}
\|\partial^\alpha Q_{i_1}^\pi - \partial^\alpha Q_{i_2}^\pi\|_\infty &= c_0 2^{-j(p+d/2)} \|\partial^\alpha \widetilde{\psi}_{j,i_1,G} - \partial^\alpha \widetilde{\psi}_{j,i_2,G}\|_\infty \\
&\gtrsim 2 c_0 2^{-j(p+d/2)} 2^{jd/2} 2^{j\|\alpha\|_{\ell_1}} \|\psi^{|\alpha|}\|_\infty \\
&= 2 c_0 2^{-j(p-\|\alpha\|_{\ell_1})} \|\psi^{|\alpha|}\|_\infty,
\end{aligned}$$

where the first inequality is given by recalling that at least one coordinate of support$(\widetilde{\psi}_{j,k_1,G})$ and support$(\widetilde{\psi}_{j,k_2,G})$ is empty for all $k_1 \ne k_2 \in \{r, \cdots, 2^j - 1 - M\}^d$. Here $\psi^{|\alpha|}$ refers to $\Pi_{m=1}^d \partial^{\alpha_m}\psi$.

Note that $2^{-j(p-\|\alpha\|_{\ell_1})} = (\log(NT)/NT)^{(p-\|\alpha\|_{\ell_1})/(2p+d)}$. Then Theorem 2.5 of (Tsybakov, 2009) implies that for any $0 \le \|\alpha\|_{\ell_1} < p$,

$$\liminf_{NT\to\infty} \inf_{\widehat{Q}} \sup_{Q\in\Lambda_\infty(p,L)} \Pr^Q\left(\|\partial^\alpha \widehat{Q} - \partial^\alpha Q\|_\infty \ge c(\log(NT)/NT)^{(p-\|\alpha\|_{\ell_1})/(2p+d)} \right) \ge c' > 0, \tag{27}$$

for some constants c and c'.

C.2. Lower bounds for L^2-norm Rates

The proof mainly follows that of Theorem G.3 in (Chen & Christensen, 2018). Again we focus on Model (22) and apply Theorem 2.5 of (Tsybakov, 2009).

We restrict $\mathcal{S} \times \mathcal{A} = [0,1]^d$. Let $\{\widetilde{\phi}_{j,k,G}, \widetilde{\psi}_{j,k,G}\}_{j,k,G}$ be a tensor-product wavelet basis of regularity larger than p for $L^2([0,1]^d)$, where j is the resolution level, $k = (k_1, k_2, \cdots, k_d) \in \{0, 1, \cdots, 2^j - 1\}^d$, and G is a vector indicating which element in a Daubechies pair $\{\phi, \psi\}$ is used. Note that ϕ has support $[-M+1, M]$ for some positive integer

M. All these pairs are generated by CDV wavelets (Cohen et al., 1993). Following the proof of (Chen & Christensen, 2018), we consider a class of submodels around Q^π. In particular, for a given j, consider the wavelet space $(\mathcal{S}\times\mathcal{A})_j$, which consists of 2^{jd} functions $\{\widetilde{\psi}_{j,k,G}\}_{k\in\{0,\cdots,2^j-1\}^d}$ with G chosen as all ψ functions. For some constant r, consider $\{\widetilde{\psi}_{j,k,G}\}_{k\in\{r,\cdots,2^j-1-M\}^d}$ as interior wavelets and $\widetilde{\psi}_{j,k,G}(s,a)=\Pi_{m=1}^{d-1}\psi_{j,k_m}(s_m)\psi_{j,k_d}(a)$ for $k=(k_1,\cdots,k_d)\in\{r,\cdots,2^j-1-M\}^d$, where $s=(s_1,\cdots,s_{d-1})\in\mathcal{S}$, $a\in\mathcal{A}$ and $\psi_{j,k_m}(\bullet)=2^{j/2}\psi(2^j(\bullet)-k_m)$ for $1\le m\le d$. Then for sufficiently large j, there exists a set $\mathcal{I}\subseteq\{r,\cdots,(2^j-M-1)\}^d$ of interior wavelets with $\text{Card}(\mathcal{I})\gtrsim 2^{dj}$, where $\text{Card}(\bullet)$ refers to the cardinality, such that at least one coordinate of $\text{support}(\widetilde{\psi}_{j,k_1,G})$ and $\text{support}(\widetilde{\psi}_{j,k_2,G})$ is empty for all $k_1\neq k_2\in\{r,\cdots,2^j-1-M\}^d$. In addition, we have $\text{Card}(\mathcal{I})\lesssim 2^{jd}$ by definition.

Then for any $Q^\pi\in\Lambda_2(p,L/2)$, where $\|\bullet\|_{\Lambda^p_{2,2}}$ is the Sobolev norm with smoothness p. For each $\theta=\{\theta_i\}_{i\in\mathcal{I}}$, where $\theta_i\in\{0,1\}$, define

$$Q^\pi_\theta=Q^\pi+c_0 2^{-j(p+d/2)}\sum_{i\in\mathcal{I}}\theta_i\widetilde{\psi}_{j,i,G}(s,a),$$

Correspondingly, let

$$h^\pi_\theta(s,a,s')=h^\pi_0(s,a,s')+c_0 2^{-j(p+d/2)}\left(\sum_{i\in\mathcal{I}}\theta_i\widetilde{\psi}_{j,i,G}(s,a)-\gamma\int_{a'\in\mathcal{A}}\pi(a'|s')\sum_{i\in\mathcal{I}}\theta_i\widetilde{\psi}_{j,i,G}(s',a')\mathrm{d}a'\right),$$

where c_0 is some positive constant specified later. Based on the construction, there are $2^{\text{Card}(\mathcal{I})}$ combinations of θ. It can be seen that for every θ,

$$\begin{aligned}&\|c_0 2^{-j(p+d/2)}\sum_{i\in\mathcal{I}}\theta_i\widetilde{\psi}_{j,i,G}(\bullet,\bullet)\|_{\Lambda^p_{2,2}}\\ \lesssim& c_0 2^{-j(p+d/2)}\sqrt{\sum_{i\in\mathcal{I}}\theta_i^2 2^{2jp}}\\ \le& c_0\end{aligned}$$

Hence $\|Q^\pi_\theta\|_{\Lambda^p_{2,2}}\le L$ for sufficient small c_0.

First of all, it can be seen that for every θ_1 and θ_2,

$$\begin{aligned}\|\partial^\alpha Q^\pi_{\theta_1}-\partial^\alpha Q^\pi_{\theta_2}\|_2&=c_0 2^{-j(p+d/2-\|\alpha\|_{\ell_1})}\|\sum_{i\in\mathcal{I}}(\theta_{1,i}-\theta_{2,i})2^{j/2}\psi^{|\alpha|}(2^j\bullet-i)\|_2\\ &\gtrsim c_0 2^{-j(p+d/2-\|\alpha\|_{\ell_1})}\sqrt{\sum_{i\in\mathcal{I}}(\theta_{1,i}-\theta_{2,i})^2\|2^{j/2}\psi^{|\alpha|}(2^j\bullet-i)\|_2^2}\\ &\asymp 2c_0 2^{-j(p+d/2-\|\alpha\|_{\ell_1})}\sqrt{\sum_{i\in\mathcal{I}}\mathbb{I}(\theta_{1,i}\neq\theta_{2,i})},\end{aligned}$$

where the second inequality is given by recalling that at least one coordinate of $\text{support}(\widetilde{\psi}_{j,k_1,G})$ and $\text{support}(\widetilde{\psi}_{j,k_2,G})$ is empty for all $k_1\neq k_2\in\{r,\cdots,2^j-1-M\}^d$. Here $\psi^{|\alpha|}(2^j\bullet-i)=\Pi_{m=1}^d\partial^{\alpha_m}\psi(2^j\bullet-i_m)$. The last line is based on that $\widetilde{\psi}_{j,i,G}\in C^\omega$ with $\omega>p>\|\alpha\|_{\ell_1}$ and is compactly supported with the bounded above and below density, then $\|2^{j/2}\psi^{(\alpha)}(2^j\bullet-i)\|_2\asymp 1$. Take j large enough. By Varshamov-Gilbert bound, we can show that there exists a subset of $\{\theta^{(0)},\cdots,\theta^{(I^*)}\}$ such that $\theta^{(0)}=\{0\}^{\text{Card}(\mathcal{I})}$, $I^*\asymp 2^{\text{Card}(\mathcal{I})}$, and $\sqrt{\sum_{j\in\mathcal{I}}\mathbb{I}(\theta^{(i)}_j\neq\theta^{(k)}_j)}\gtrsim 2^{jd/2}$, where $0\le i\le k\le I^*$. Therefore $\|Q^\pi_i-Q^\pi_k\|_2\gtrsim c_0 2^{-j(p-\|\alpha\|_{\ell_1})}$ for $0\le i\le k\le I^*$, where we denote $Q_{\theta^{(i)}}=Q_i$. Similarly, we denote $h^\pi_{\theta^{(i)}}=h^\pi_i$.

For $0\le m\le I^*$, consider Model (22) with the true function h^π_m and define the joint probability distribution of $\{S_{j,t},A_{j,t},R_{j,t},S'_{j,t}\}_{1\le j\le N,0\le t\le T-1}$ as P_i such that

$$P_m=\Pi_{j=0}^N\Pi_{t=0}^{T-1}d^\nu(S_{j,0},A_{j,0})p_{h^\pi_m}(R_{j,t}|S_{j,t},A_{j,t})q(S'_{j,t}\,|\,S_{j,t},A_{j,t}),$$

by recalling that they are $i.i.d.$ samples.

Secondly, for sufficiently small c_0, we can show for every $0 \le m \le I^*$

$$\|c_0 2^{-j(p+d/2)} \sum_{i\in\mathcal{I}} \theta_i^{(m)} \widetilde{\psi}_{j,i,G}\|_{\Lambda^p_{2,2}} \lesssim c_0 \le L/2$$

. In addition, by Equation (7), we have

$$\begin{aligned}
\sqrt{p_{\min}/p_{\max}}(1-\gamma)c_0 2^{-j(p+d/2)} \|\sum_{i\in\mathcal{I}} \theta_i^{(m)} \widetilde{\psi}_{j,i,G}\|_2 \lesssim \|\mathcal{T} c_0 2^{-j(p+d/2)} \Big(& \sum_{i\in\mathcal{I}} \theta_i^{(m)} \widetilde{\psi}_{j,i,G}(S,A) \\
& -\gamma \int_{a'\in\mathcal{A}} \pi(a'|S') \sum_{i\in\mathcal{I}} \theta_i^{(m)} \widetilde{\psi}_{j,i,G}(S',a') \mathrm{d}a' \Big) \|_2 \\
\lesssim c_0 2^{-j(p+d/2)} & \|\sum_{i\in\mathcal{I}} \theta_i^{(m)} \widetilde{\psi}_{j,i,G}\|_2 \\
\lesssim c_0 2^{-j(p+d/2)} & \sqrt{\sum_{i\in\mathcal{I}} (\theta^{(m)})^2} \\
\asymp c_0 2^{-jp}. &
\end{aligned}$$

Moreover, for every $0 \le m \le I^*$, the distance $K(P_m, P_0)$ can be bounded as

$$\begin{aligned}
& K(P_m, P_0) \\
\le & \frac{1}{2} c_0^2 2^{-j(2p+d)} \sum_{k=1}^{N} \sum_{t=0}^{T-1} \mathbb{E}[\frac{(\mathcal{T}(\sum_{i\in\mathcal{I}} \theta_i^{(m)} \widetilde{\psi}_{j,i,G}(S_{k,t}, A_{k,t}) - \int_{a'\in\mathcal{A}} \pi(a'|S'_{k,t}) \sum_{i\in\mathcal{I}} \theta_i^{(m)} \widetilde{\psi}_{j,i,G}(S'_{k,t}, a') \mathrm{d}a'))^2}{\sigma^2(S_{k,t}, A_{k,t})}] \\
\lesssim & NT c_0^2 2^{-j(2p)},
\end{aligned}$$

by the condition in Theorem 4.1. By choosing $2^j \asymp (NT)^{1/(2p+d)}$, it gives that

$$K(P_m, P_0) \lesssim c_0^2 (NT)^{d/(2p+d)},$$

and $\log(I^*) \gtrsim 2^{jd} \asymp (NT)^{d/(2p+d)}$ by recalling that $I^* \asymp 2^{\mathrm{Card}(\mathcal{I})}$ and $\mathrm{Card}(\mathcal{I}) \asymp 2^{jd}$. So for sufficiently small c_0 and large NT, $K(P_m, P_0) \le 1/8 \log(\mathcal{I}^*)$ for every $1 \le m \in \mathcal{I}^*$.

Note that $2^{-j(p-\|\alpha\|_{\ell_1})} = (NT)^{(\|\alpha\|_{\ell_1}-p)/(2p+d)}$. Then Theorem 2.5 of (Tsybakov, 2009) implies that

$$\liminf_{NT\to\infty} \inf_{\widehat{Q}} \sup_{Q\in\Lambda_2(p,L)} \Pr^Q \left(\|\widehat{Q} - Q\|_2 \ge \bar{c} (NT)^{(\|\alpha\|_{\ell_1}-p)/(2p+d)} \right) \ge \bar{c}' > 0, \tag{28}$$

for some constants $\bar{c}$ and $\bar{c}'$.

D. Proof of Theorem 5.4

Let $Q^\pi_{0,J}$ solves $\inf_{Q\in\Psi_J} \|Q^\pi - Q\|_\infty$. Under all assumptions in Lemmas 5.3, we have as long as $\zeta^2 \sqrt{\log(J)\log(NT)/NT} = O(1)$,

$$\begin{aligned}
& \|\widehat{Q}^\pi - Q^\pi\|_\infty \\
\le & \|\widehat{Q}^\pi - \widetilde{Q}^\pi\|_\infty + \|\widetilde{Q}^\pi - Q^\pi\|_\infty \\
\le & O_p\Big(\frac{R_{\max}}{1-\gamma} \tau_J \xi_J \sqrt{(\log J)/(NTe_J)} \Big) + O_p(1) \times \|Q^\pi - \overline{\Pi} Q^\pi\|_\infty \\
\le & O_p\Big(\xi_J \sqrt{(\log J)/(NTe_J)} \Big) + O_p(1) \|\overline{\Pi}_J\|_\infty \|Q^\pi - Q^\pi_{0,J}\|_\infty,
\end{aligned}$$

where the first inequality is by triangle inequality and the last inequality is given by Lebesgue's lemma and $\tau_J \lesssim 1$.

To proceed our proof, we only consider the wavelet basis of (Cohen et al., 1993) for B_K and Ψ_J, while results of other bases given in Theorem 5.4 can be derived similarly. Based on the property of wavelet basis, we can show that $\|\overline{\Pi}_J\|_\infty \lesssim 1$, where the proof is given in (Chen & Christensen, 2015). Since $Q^\pi \in \Lambda_\infty(p, L)$, we have $\|Q^\pi - Q^\pi_{0,J}\|_\infty \lesssim O(J^{-p/d})$ by e.g., (Huang et al., 1998). Summarizing together, we have

$$\|\widehat{Q}^\pi - Q^\pi\|_\infty = O_p\Big(\tau_J \xi_J \sqrt{(\log J)/(NTe_J)} + J^{-p/d}\Big).$$

According to Lemma H.1 and by the property of wavelet bases, $e_J \gtrsim (1-\gamma)^2 p_{\min}^2/p_{\max} \gtrsim 1$. Similarly, we can show that $\zeta_b \le \zeta \lesssim \sqrt{J}$, and $\xi_J \lesssim \sqrt{J}$. Hence we have our first statement that

$$\|\widehat{Q}^\pi - Q^\pi\|_\infty = O_p\big(J^{-p/d} + \sqrt{J(\log J)/(NT)}\big), \tag{29}$$

as long as $J\sqrt{\log(J)\log(NT)/NT} = O(1)$. Lastly, by choosing $J \asymp \left(\frac{NT}{\log(NT)}\right)^{d/(2p+d)}$, which satisfies the constraint, we have

$$\|\widehat{Q}^\pi - Q^\pi\|_\infty = O_p\left(\left(\frac{\log(NT)}{NT}\right)^{p/(2p+d)}\right).$$

Next, we present the proof related to the derivative case. Note that by the previous result, we have

$$\|Q^\pi - \widetilde{Q}^\pi\|_\infty = O_p(J^{-p/d}).$$

In addition, by Bernstein inequalities in approximation theory, we have

$$\|\partial^\alpha Q\|_\infty = O(J^{\|\alpha\|_{\ell_1}/d})\|Q\|_\infty,$$

for all $Q \in \Psi_J$. Hence we can show that by Lemma 3.1 and 5.3 Result (2),

$$\begin{aligned}
\|\partial^\alpha \widetilde{Q}^\pi - \partial^\alpha Q^\pi\|_\infty &\le \|\partial^\alpha \widetilde{Q}^\pi - \partial^\alpha(\overline{\Pi}_J Q^\pi)\|_\infty + \|\partial^\alpha Q^\pi - \partial^\alpha(\overline{\Pi}_J Q^\pi)\|_\infty \\
&\le O(J^{\|\alpha\|_{\ell_1}/d})\|\widetilde{Q}^\pi - \overline{\Pi}_J Q^\pi\|_\infty + \|\partial^\alpha Q^\pi - \partial^\alpha(\overline{\Pi}_J Q^\pi)\|_\infty \\
&\le O(J^{\|\alpha\|_{\ell_1}/d})\|\widetilde{h}^\pi - \Pi_J h_0^\pi\|_\infty + \|\partial^\alpha Q^\pi - \partial^\alpha(\overline{\Pi}_J Q^\pi)\|_\infty \\
&\le O_p(J^{-(p-\|\alpha\|_{\ell_1})/d}) + \|\partial^\alpha Q^\pi - \partial^\alpha Q_J^\pi\|_\infty + \|\partial^\alpha Q_J^\pi - \partial^\alpha(\overline{\Pi}_J Q^\pi)\|_\infty \\
&\le O_p(J^{-(p-\|\alpha\|_{\ell_1})/d}) + \|\partial^\alpha Q^\pi - \partial^\alpha Q_J^\pi\|_\infty + O(J^{\|\alpha\|_{\ell_1}/d})\|Q_J^\pi - Q^\pi\|_\infty.
\end{aligned}$$

By choosing Q_J^π such that $\|Q_J^\pi - Q^\pi\|_\infty = O(J^{-p/d})$ and $\|\partial^\alpha Q_J^\pi - \partial^\alpha(\overline{\Pi}_J Q^\pi)\|_\infty = O(J^{-(p-\|\alpha\|_{\ell_1})/d})$, we have

$$\|\partial^\alpha \widetilde{Q}^\pi - \partial^\alpha Q^\pi\|_\infty = O_p(J^{-(p-\|\alpha\|_{\ell_1})/d}).$$

Finally, we can derive that

$$\begin{aligned}
\|\partial^\alpha \widehat{Q}^\pi - \partial^\alpha Q^\pi\|_\infty &\le \|\partial^\alpha \widehat{Q}^\pi - \partial^\alpha \widetilde{Q}^\pi\|_\infty + \|\partial^\alpha \widetilde{Q}^\pi - \partial^\alpha Q^\pi\|_\infty \\
&\le O(J^{\|\alpha\|_{\ell_1}/d})\|\widehat{Q}^\pi - \widetilde{Q}^\pi\|_\infty + \|\partial^\alpha \widetilde{Q}^\pi - \partial^\alpha Q^\pi\|_\infty \\
&\le O(J^{\|\alpha\|_{\ell_1}/d})O_p\Big(\xi_J\sqrt{(\log J)/(NT)}\Big) + O_p(J^{-(p-\|\alpha\|_{\ell_1})/d}),
\end{aligned}$$

where the last inequality is given by Lemma 5.3 (1). This concludes our proof.

E. Proof of Lemma 5.3 Result (1)

The proof consists of three steps.

Step 1: Decompose the difference between $\widehat{c}$ and $\widetilde{c}$.

$$\begin{aligned}
\widehat{c} - \widetilde{c} &= [\Gamma_\pi^\top B(B^\top B)^- B^\top \Gamma_\pi]^- \Gamma_\pi^\top B(B^\top B)^- B^\top(\mathbf{R} - H_0) \\
&= [{\Sigma^\pi}^\top G_b^{-1}\Sigma^\pi]^{-1}{\Sigma^\pi}^\top G_b^{-1} B^\top\Big(\frac{\mathbf{R} - H_0}{NT}\Big) \\
&+ \Big(-[{\Sigma^\pi}^\top G_b^{-1}\Sigma^\pi]^{-1}{\Sigma^\pi}^\top G_b^{-1} + [{\widehat{\Sigma}^\pi}^\top \widehat{G}_b^- \widehat{\Sigma}^\pi]^- {\widehat{\Sigma}^\pi}^\top \widehat{G}_b^-\Big) B^\top\Big(\frac{\mathbf{R} - H_0}{NT}\Big) \\
&= (I) + (II),
\end{aligned}$$

where

$$\widehat{\Sigma}^{\pi} = \frac{B^{\top}\Gamma_{\pi}}{NT} \quad \text{and} \quad \widehat{G}_b = \frac{B^{\top}B}{NT}.$$

Step 2: Bound the first term (I). Define an event

$$\mathcal{E}_{NT} = \left\{ \left\| \frac{[G_b]^{-1/2} B^{\top} B [G_b]^{-1/2}}{NT} - I_K \right\| \leq \frac{1}{2} \right\},$$

where I_K is an identity matrix with size K. By Lemma H.2 (b), we have

$$\left\| \frac{[G_b]^{-1/2} B^{\top} B [G_b]^{-1/2}}{NT} - I_K \right\| = O_p(\zeta_b \sqrt{\log(NT)\log(K)/(NT)}),$$

as long as $\zeta\sqrt{\log(NT)\log(K)/(NT)} = o(1)$. Hence we obtain that $\Pr(\mathcal{E}_{NT}^c) = o(1)$ by the assumption in Lemma 5.3 that $\zeta^2\sqrt{\log(NT)\log(K)/(NT)} = O(1)$ and $\zeta \geq \sqrt{J}$.

Now for any $x > 0$, we can show that

$$P(\|(I)\|_{\ell_\infty} > x) \tag{30}$$

$$\leq \sum_{j=1}^{J} P\left(\left| \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=0}^{T-1} q_j^{\pi}(S_{i,t}, A_{i,t}) \left(R_{i,t} - h_0^{\pi}(S_{i,t}, A_{i,t}, S_{i,t+1}) \right) \right| > x, \mathcal{E}_{NT} \right) + \Pr(\mathcal{E}_{NT}^c), \tag{31}$$

where $q_j^{\pi}(S_{i,t}, A_{i,t}) = \left\{ [\Sigma^{\pi\,\top} G_b^{-1} \Sigma^{\pi}]^{-1} \Sigma^{\pi\,\top} G_b^{-1} b^K(S_{i,t}, A_{i,t}) \right\}_j$ (j-th element of a vector). Note that

$$\mathbb{E}\left[R_{i,t} - h_0^{\pi}(S_{i,t}, A_{i,t}, S_{i,t+1}) \mid S_{i,t}, A_{i,t} \right] = 0,$$

by the Bellman equation (1). Therefore the sequence

$$\{ q_j^{\pi}(S_{i,t}, A_{i,t}) \left(R_{i,t} - h_0^{\pi}(S_{i,t}, A_{i,t}, S_{i,t+1}) \right) \}_{0 \leq t \leq (T-1), 1 \leq i \leq N}$$

forms a mean 0 martingale. We aim to apply Freedman's inequality. Firstly, by Assumption 2.2 on the reward, we have

$$|R_{i,t} - h_0^{\pi}(S_{i,t}, A_{i,t}, S_{i,t+1})| \leq \frac{2R_{\max}}{1-\gamma}.$$

In addition, we can show that

$$\begin{aligned}
&|q_j^{\pi}(S_{i,t}, A_{i,t})| \\
\leq &\| [\Sigma^{\pi\,\top} G_b^{-1} \Sigma^{\pi}]^{-1} \Sigma^{\pi\,\top} G_b^{-1} b^K(S_{i,t}, A_{i,t}) \|_{\ell_2} \\
\leq &\| [G_\kappa^{\pi}]^{-1/2} \|_{\ell_2} \| [[G_\kappa^{\pi}]^{-1/2} \Sigma^{\pi\,\top} G_b^{-1} \Sigma^{\pi} [G_\kappa^{\pi}]^{-1/2}]^{-1} [G_\kappa^{\pi}]^{-1/2} \Sigma^{\pi\,\top} G_b^{-1/2} \|_{\ell_2} \| G_b^{-1/2} b^K(S_{i,t}, A_{i,t}) \|_{\ell_2} \\
\leq &\frac{\zeta}{s_{JK}\sqrt{e_J}},
\end{aligned}$$

where

$$s_{JK}^{-1} = \sup_{h \in \Theta_J^{\pi}} \frac{\|h\|_{L^2(S,A,S')}}{\|\Pi_K \mathcal{T} h\|_{L^2(S,A)}} = s_{\min}(G_b^{-\frac{1}{2}} \Sigma_{\pi} [G_\kappa^{\pi}]^{-1/2}]),$$

and $s_{\min}$ refers to the minimum singular value. One can show that $s_{JK}^{-1} \leq \tau_J \lesssim 1$ by Lemma A.1 of (Chen & Christensen, 2018) using Assumption 5.2 (a)

Secondly, we can show that conditioning on $\mathcal{E}_{NT}$,

$$\sum_{i=1}^{N} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\{ q_j^{\pi}(S_{i,t}, A_{i,t}) \left(R_{i,t} - h_0^{\pi}(S_{i,t}, A_{i,t}, S_{i,t+1}) \right) \right\}^2 \mid S_{i,t}, A_{i,t} \right] \leq \frac{6NTR_{\max}^2}{(1-\gamma)^2 s_{JK}^2 e_J}.$$

This relies on the following argument. Conditioning on $\mathcal{E}_{NT}$, for every j,

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{t=0}^{T-1}\mathbb{E}\left[(q_j^\pi(S_{i,t},A_{i,t}))^2 \mid S_{i,t},A_{i,t}\right]\\
=&\sum_{i=1}^{N}\sum_{t=0}^{T-1}\|\{[\Sigma^{\pi\top}G_b^{-1}\Sigma^\pi]^{-1}\Sigma^{\pi\top}G_b^{-1/2}\}_{j\bullet}G_b^{-1/2}b^K(S_{i,t},A_{i,t})\|_{\ell_2}^2\\
\leq&\frac{3NT}{2}\|\{[\Sigma^{\pi\top}G_b^{-1}\Sigma^\pi]^{-1}\Sigma^{\pi\top}G_b^{-1/2}\}_{j\bullet}\|_{\ell_2}^2\\
\leq&\frac{3NT}{2}\|[\Sigma^{\pi\top}G_b^{-1}\Sigma^\pi]^{-1}\|_{\ell_2}^2\\
\leq&\frac{3NT}{2}\|[G_\kappa^\pi]^{-1/2}[[G_\kappa^\pi]^{-1/2}\Sigma^{\pi\top}G_b^{-1}\Sigma^\pi[G_\kappa^\pi]^{-1/2}]^{-1}[G_\kappa^\pi]^{-1/2}\|_{\ell_2}^2\\
\leq&\frac{3NT}{2s_{JK}^2e_J},
\end{aligned}$$

where the first inequality is given by the event $\mathcal{E}_{NT}$. Then by Freedman's inequality (e.g., Theorem 1.1 of (Tropp, 2011)), we can show,

$$\|(I)\|_{\ell_\infty}=O_p\left(\sqrt{\frac{\log J}{NTe_J}}\right), \tag{32}$$

as long as $\zeta\sqrt{\log(J)/(NT)}=o(1)$.

Define

$$(G_b^{-1/2}\Sigma^\pi)_l^- = \left[[\Sigma^\pi]^\top(G_b)^{-1}\Sigma^\pi\right]^{-1}[\Sigma^\pi]^\top(G_b)^{-1/2},$$

and similarly for $(\widehat{G}_b^{-1/2}\widehat{\Sigma}^\pi)_l^-$.

Step 3: We bound the second term (II). Relying on Lemmas H.5 (a) and H.3, we have

$$\begin{aligned}
&\|(\mathrm{II})\|_{\ell_\infty} &(33)\\
\leq&\|(G_b^{-1/2}\widehat{\Sigma}^\pi)_l^-\widehat{G}_b^{-1/2}G_b^{1/2}-(G_b^{-1/2}\Sigma^\pi)_l^-\|_{\ell^2}\|G_b^{-1/2}B^\top(R-H_0)/(NT)\|_{\ell^2} &(34)\\
=&O_p\Big(s_{JK}^{-2}\zeta\sqrt{(\log(NT)\log J)/(NTe_J)}\Big)O_p(\frac{R_{\max}}{1-\gamma}\sqrt{\frac{K}{NT}}) &(35)\\
=&O_p\Big(\sqrt{\log(J)/(NTe_J)}\Big), &(36)
\end{aligned}$$

by the assumption in Lemma 5.3 (1) that $\zeta^2\sqrt{\log(NT)}/\sqrt{NT}=O(1)$ and the fact that $\zeta\geq\sqrt{K}$ and $s_{JK}^{-1}\leq\tau_J\lesssim 1$. This completes the proof of Lemma 5.3(1) by noting that $\sup_{s\in\mathcal{S},a\in\mathcal{A}}\|\psi^J(s,a)\|_{\ell_1}=\xi_J$ by definition.

F. Proof of Lemma 5.3 Result (2)

We first prove the following Lemma.

Lemma F.1. *Suppose that $\zeta^2\sqrt{\log(J)\log(NT)}/\sqrt{NT}=O(1)$ and let Assumptions 5.1-5.2 hold. Then $\|\widetilde{h}^\pi-\Pi_Jh_0^\pi\|_\infty\leq O_p(1)\times\|h_0^\pi-\Pi_Jh_0^\pi\|_\infty$.*

The proof follows similarly as Lemma A.3 of (Chen & Christensen, 2018). Note that the difference between $\widetilde{h}^\pi$ and $\Pi_Jh_0^\pi$ can be decomposed as

$$\begin{aligned}
&\widetilde{h}^\pi(s,a,s')-\Pi_Jh_0^\pi(s,a,s')\\
=&\widetilde{\Pi}(h_0^\pi-\Pi_Jh_0^\pi)(s,a,s')\\
+&(\kappa_\pi^J(s,a,s'))^\top(G_b^{-1/2}\Sigma^\pi)_l^-\{G_b^{-1/2}(B^\top(H_0-\Gamma_\pi c_J)/(NT)-E[b^K(S,A)(h_0^\pi(S,A,S')-h_J^\pi(S,A,S'))])\}\\
+&(\kappa_\pi^J(s,a,s'))^\top\{(\widehat{G}_b^{-1/2}\widehat{\Sigma}^\pi)_l^-\widehat{G}_b^{-1/2}G_b^{-1/2}-(G_b^{-1/2}\Sigma^\pi)_l^-\}G_b^{-1/2}B^\top(H_0-\Gamma_\pi c_J)/(NT)\\
=&(I)+(II)+(III).
\end{aligned}$$

For (I), by Assumption 5.2 (b), we can show that $\|(I)\|_\infty \lesssim \|h_0^\pi - \Pi_J h_0^\pi\|_\infty$. For (II), by Lemma H.4, we can show

$$\|(II)\|_\infty \le \zeta_{\kappa,J}^\pi s_{JK}^{-1} O_p\big(\zeta_{b,K}\sqrt{\frac{\log(NT)\log(K)}{NT}}\big)\|h_0^\pi - \Pi_J h_0^\pi\|_\infty \tag{37}$$

$$= O_p\big(\zeta^2\sqrt{\frac{\log(NT)\log(K)}{NT}}\big)\|h_0^\pi - \Pi_J h_0^\pi\|_\infty = O(1)\|h_0^\pi - \Pi_J h_0^\pi\|_\infty, \tag{38}$$

where we use the fact that $\zeta \ge \max\{\zeta_{b,K}, \zeta_{\kappa,J}\}$, $s_{JK}^{-1} \le \tau_J \lesssim 1$ and that $\zeta^2\sqrt{\log(J)\log(NT)}/\sqrt{NT} = O(1)$. For (III) term, by Lemma H.5(b), we can show that

$$\begin{aligned}
&\|(III)\|_\infty \\
\le&\zeta_{\kappa,J}\|(\widehat{G}_b^{-1/2}\widehat{\Sigma}^\pi)_l^- \widehat{G}_b^{-1/2} G_b^{-1/2} - (G_b^{-1/2}\Sigma^\pi)\|_{\ell_2}\|G_b^{-1/2}B^\top(H_0 - \Gamma_\pi c_J)/(NT)\|_{\ell_2} \\
\le&\zeta_{\kappa,J} O_p\big(\zeta\sqrt{\frac{\log J\log(NT)}{NT}}\big)\{O_p\big(\zeta_{b,K}\sqrt{\frac{\log(NT)\log K}{NT}}\big)\|h_0^\pi - \Pi_J h_0^\pi\|_\infty + \|\Pi_K \mathcal{T}(h_0^\pi - \Pi_J h_0^\pi)\|_{L^2(S,A)}\} \\
\le&\zeta_{\kappa,J} O_p\big(\zeta\sqrt{\frac{\log J\log(NT)}{NT}}\big)\{O_p\big(\zeta_{b,K}\sqrt{\frac{\log(NT)\log K}{NT}}\big)\|h_0^\pi - \Pi_J h_0^\pi\|_\infty + \|(h_0^\pi - \Pi_J h_0^\pi)\|_{L^2(S,A)}\} \\
=&O_p\big(\zeta^2\sqrt{\frac{\log(J)\log(NT)}{NT}}\big)\|(h_0^\pi - \Pi_J h_0^\pi)\|_{L^2(S,A)} \\
=&O_p(1)\|h_0^\pi - \Pi_J h_0^\pi\|_{L^2(S,A)},
\end{aligned}$$

by the condition that $\zeta^2\sqrt{\log(J)\log(NT)}/\sqrt{NT} = O(1)$.

Now, we return to Result (2) of Lemma 5.3. By Lemma 3.1, we can see that

$$\begin{aligned}
&\|\widetilde{Q}^\pi - Q^\pi\|_\infty \lesssim \|\widetilde{h}^\pi - h_0^\pi\|_\infty \\
\le&\|\widetilde{h}^\pi - \Pi_J h_0^\pi\|_\infty + \|h_0^\pi - \Pi_J h_0^\pi\|_\infty \\
=&O_p(1)\|h_0^\pi - \Pi_J h_0^\pi\|_\infty, \\
\le&O_p(1)\|Q^\pi - \overline{\Pi}_J Q^\pi\|_\infty,
\end{aligned}$$

which concludes our proof.

G. Proof of Theorem 5.5

The idea of proof is similar to that in Lemma 5.3 (1) and Theorem 5.4. By triangle inequality, we have $\|\widehat{Q}^\pi - Q^\pi\|_2 \le \|\widehat{Q}^\pi - \widetilde{Q}^\pi\|_2 + \|\widetilde{Q}^\pi - \overline{\Pi}_J Q^\pi\|_2 + \|Q^\pi - \overline{\Pi}_J Q^\pi\|_2$. In the following **Step 1-3**, we first bound $\|\widehat{h}^\pi - \widetilde{h}^\pi\|_2$ since $\|\widehat{Q}^\pi - \widetilde{Q}^\pi\|_2 \lesssim \|\widehat{h}^\pi - \widetilde{h}^\pi\|_2$. The last step is to bound $\|\widetilde{Q}^\pi - \overline{\Pi}_J Q^\pi\|_2$.

Step 1: Decompose the difference between $\widehat{h}^\pi(s,a,s')$ and $\widetilde{h}^\pi(s,a,s')$ as follows.

$$\begin{aligned}
&(\kappa_\pi^J(s,a,s'))^\top\widehat{c} - (\psi^J(s,a))^\top\widetilde{c} = (\psi^J(s,a))^\top[\Gamma_\pi^\top B(B^\top B)^- B^\top\Gamma_\pi]^-\Gamma_\pi^\top B(B^\top B)^- B^\top(\mathbf{R} - H_0) \\
=&(\kappa_\pi^J(s,a,s'))^\top[\Sigma^{\pi\top} G_b^{-1}\Sigma^\pi]^{-1}\Sigma^{\pi\top} G_b^{-1} B^\top\big(\frac{\mathbf{R} - H_0}{NT}\big) \\
+&(\kappa_\pi^J(s,a,s'))^\top\left(-[\Sigma^{\pi\top} G_b^{-1}\Sigma^\pi]^{-1}\Sigma^{\pi\top} G_b^{-1} + [\widehat{\Sigma}^{\pi\top}\widehat{G}_b^-\widehat{\Sigma}^\pi]^-\widehat{\Sigma}^{\pi\top}\widehat{G}_b^-\right) B^\top\big(\frac{\mathbf{R} - H_0}{NT}\big) \\
=&(I) + (II),
\end{aligned}$$

where

$$\widehat{\Sigma}^\pi = \frac{B^\top\Gamma_\pi}{NT} \quad \text{and} \quad \widehat{G}_b = \frac{B^\top B}{NT}.$$

Step 2: Bound the first term (I). Note that

$$\begin{aligned}\|(I)\|_2 &= \|(\kappa_\pi^J(\bullet,\bullet,\bullet))^\top[\Sigma^{\pi\top}G_b^{-1}\Sigma^\pi]^{-1}\Sigma^{\pi\top}G_b^{-1}B^\top(\frac{\mathbf{R}-H_0}{NT})\|_2\\ &= \|[G_\kappa^\pi]^{1/2}[\Sigma^{\pi\top}G_b^{-1}\Sigma^\pi]^{-1}\Sigma^{\pi\top}G_b^{-1}B^\top(\frac{\mathbf{R}-H_0}{NT})\|_{\ell_2}\\ &\le s_{KJ}^{-1}\|G_b^{-1/2}B^\top(\frac{\mathbf{R}-H_0}{NT})\|_{\ell_2} = O_p(\sqrt{\frac{K}{NT}}),\end{aligned}$$

where the last inequality is given by Lemma H.3 and $s_{JK}^{-1}\lesssim 1$.

Step 3: we bound the second term (II). Relying on Lemmas H.5 (b) and H.3, we have

$$\|(\mathrm{II})\|_2 \tag{39}$$
$$\le\|[G_\kappa^\pi]^{1/2}\{(G_b^{-1/2}\widehat{\Sigma}^\pi)_l^-\widehat{G}_b^{-1/2}G_b^{1/2}-(G_b^{-1/2}\Sigma^\pi)_l^-\}\|_{\ell^2}\|G_b^{-1/2}B^\top(\mathbf{R}-H_0)/(NT)\|_{\ell^2} \tag{40}$$
$$=O_p\Big(s_{JK}^{-2}\zeta\sqrt{(\log(NT)\log J)/(NT)}\Big)O_p(\frac{R_{\max}}{1-\gamma}\sqrt{\frac{K}{NT}}) \tag{41}$$
$$=O_p\Big(\sqrt{K/(NT)}\Big), \tag{42}$$

by the assumption in Theorem 5.5 that $\zeta\sqrt{\log(NT)\log(J)}/\sqrt{NT}) = o(1)$ and $s_{JK}^{-1}\lesssim 1$.

Step 4: In the remaining proof, we show the bound for $\|\widetilde{Q}^\pi-\overline{\Pi}_J Q^\pi\|_2$. By Theorem 3.3, we can show that,

$$(1+\gamma)\|\widetilde{Q}^\pi-\overline{\Pi}_J Q^\pi\|_2 \lesssim \|\widetilde{h}^\pi-\Pi_J h^\pi\|_2.$$

Then by a similar proof in Appendix F, we can show that as long as $\zeta\sqrt{\log(NT)\log(J)}/\sqrt{NT}) = O(1)$,

$$\|\widetilde{Q}^\pi-\overline{\Pi}_J Q^\pi\|_2 \le O_p(1)\times\|Q^\pi-\overline{\Pi}_J Q^\pi\|_2 = O_p(1)\times J^{-p/d},$$

where we use the existing result on the approximation error of the linear sieve in the last equation. Summarizing **Step 1-4** together, we obtain the statements in Theorem 5.5. Finally, we conclude our proof by the similar argument in the proof of Theorem 5.4 for the derivatives case.

H. Technical Lemmas

Lemma H.1. *For any policy π, under Assumptions 2.1, 2.3, 3.2 and 5.1, we have*

$$e_J \gtrsim \frac{p_{\min}^2}{p_{\max}}(1-\gamma)^2\omega_J$$

for every $J\ge 1$, where $\omega_J = \lambda_{\min}(\mathbb{E}\left[\psi^J(S,A)(\psi^J(S,A))^\top\right])$

Proof. By definition,

$$e_J = \lambda_{\min}\left\{\mathbb{E}\left[\left(\psi^J(S,A)-\gamma\psi_\pi^J(S')\right)\left(\psi^J(S,A)-\gamma\psi_\pi^J(S')\right)^\top\right]\right\}.$$

Applying Theorem 3.3 with $Q_1(s,a) = (\psi^J(S,A))^\top x$ and $Q_2(s,a)=0$ for every $s\in\mathcal{S}$ and $a\in\mathcal{A}$ (recall that the sieve space is a subset of $L^2(S,A)$), we have

$$\begin{aligned}&x^\top\mathbb{E}\left[\left(\psi^J(S,A)-\gamma\psi_\pi^J(S')\right)\left(\psi^J(S,A)-\gamma\psi_\pi^J(S')\right)^\top\right]x\\ \ge& x^\top\mathbb{E}\left[\left(\psi^J(S,A)-\gamma\mathbb{E}\left[\psi_\pi^J(S')\mid S,A\right]\right)\left(\psi^J(S,A)-\gamma\mathbb{E}\left[\psi_\pi^J(S')\mid S,A\right]\right)^\top\right]x\\ =&\|(\psi^J(S,A)-\gamma\mathbb{E}\left[\psi_\pi^J(S')\mid S,A\right])^\top x\|_2^2\\ \ge&\frac{p_{\min}}{p_{\max}}(1-\gamma)^2\|(\psi^J(S,A))^\top x\|_2^2 \ge \frac{p_{\min}}{p_{\max}}(1-\gamma)^2\omega_J\|x\|_{\ell_2}^2,\end{aligned}$$

where the first inequality is given by Jensen's inequality and the last inequality is by the definition of ω_J. □

By examining the proof, we can see that the above lemma also holds for $\bar{d}_T^{\pi^b}$ without Assumption 5.1.

Next we present several technical lemmas adapted from (Chen & Christensen, 2018). Define the *orthonormalized* matrix estimators

$$\begin{aligned}\widehat{G}_b^o &= G_b^{-1/2}\widehat{G}_b G_b^{-1/2}\\ \widehat{G}_\kappa^{\pi,\,o} &= [G_\kappa^\pi]^{-1/2}\widehat{G}_\kappa^\pi [G_\kappa^\pi]^{-1/2}\\ \widehat{\Sigma}^{\pi,\,o} &= G_b^{-1/2}\widehat{\Sigma}^\pi [G_\kappa^\pi]^{-1/2},\end{aligned}$$

where $\widehat{G}_\kappa^\pi = \frac{\Gamma_\pi^\top \Gamma_\pi}{NT}$. Let $G_b^o = I_K$, $G_\kappa^{\pi\,o} = I_J$ and $\Sigma^{\pi\,o}$ denote their corresponding expected values.

Lemma H.2. *Under Assumption 5.1, the following three bounds hold.*

$$\begin{aligned}\|\widehat{G}_\kappa^{\pi,\,o} - G_\kappa^{\pi,\,o}\|_{\ell^2} &= O_p(\zeta_{\kappa,J}^\pi \sqrt{(\log(NT)\log(J))/(NT)})\\ \|\widehat{G}_b^o - G_b^o\|_{\ell^2} &= O_p(\zeta_{b,K}\sqrt{(\log(NT)\log K)/(NT)})\\ \|\widehat{\Sigma}^{\pi,\,o} - \Sigma^{\pi,\,o}\|_{\ell^2} &= O_p(\max(\zeta_{b,K},\zeta_{\kappa,J}^\pi)\sqrt{(\log(NT)\log K)/(NT)})\,.\end{aligned}$$

as $N,T,J,K\to\infty$ as long as $\zeta\sqrt{(\log(NT)\log(J)/NT} = o(1)$.

Proof. The proof follows similarly from Lemma 2.2 of (Chen & Christensen, 2015). The basic idea is to use Berbee's coupling lemma (e.g., Theorem 4.2 of (Chen & Christensen, 2015)) and matrix Bernstein's inequality (e.g., (Tropp, 2015)). For brevity, we only show the proof of the second statement in Lemma H.2, while others are similar.

Let $X_{i,t} = G_b^{-1/2} b^K(S_{i,t},A_{i,t})[b^K(S_{i,t},A_{i,t})]^\top G_b^{-1/2}/(NT) - I_K/NT$ and $\mathbb{E}[X_{i,t}] = 0_{K\times K}$. Denote the upper bound of mixing coefficient as $\beta(w) = \beta_0\exp(-\beta_1 w)$, where β_0 and β_1 are given in Assumption 5.1. By Berbee's lemma, for a fixed i with $1\le i\le N$ and some integer w, the stochastic process $\{X_{i,t}\}_{t\ge 0}$ can be coupled by a process $Y_{i,t}^*$ such that $Y_{i,k} = \{X_{i,(k-1)w+j}\}_{0\le j<w}$ and $Y_{i,k}^* = \{X_{i,(k-1)w+j}^*\}_{0\le j<w}$ are identically distributed for each $k\ge 1$ and $P(Y_{i,k}\neq Y_{i,k}^*)\le\beta(w)$. In addition, the sequence $\{\{Y_{i,k}^*\}\mid k = 2z, z\ge 1\}$ are independent and so are the sequence $\{\{Y_{i,k}^*\}\mid k = 2z+1, z\ge 0\}$. Denote I_e as the indices of the corresponding even number block and I_o as indices of the corresponding odd number blocks in $\{0,\cdots,T-1\}$. Let I_r be the indices in the remainders, i.e., $I_r = \{\lfloor T/w\rfloor w,\cdots,T-1\}$ and thus $\mathrm{Card}(I_r) < w$. We construct a coupled stochastic process for every $1\le i\le N$ trajectory. Now by triangle inequality, we can show that for $x>0$

$$\begin{aligned}&\Pr(\|\sum_{i=1}^N\sum_{t=0}^{T-1} X_{i,t}\|_{\ell_2}\ge 4x)\\ \le&\Pr(\|\sum_{i=1}^N\sum_{t=0}^{\lfloor T/w\rfloor w-1} X_{i,t}^*\|_{\ell_2}\ge 2x) + \Pr(\|\sum_{i=1}^N\sum_{t\in I_r} X_{i,t}\|_{\ell_2}\ge x)) + \Pr(\|\sum_{i=1}^N\sum_{t=0}^{\lfloor T/w\rfloor w-1}(X_{i,t}^* - X_{i,t})\|_{\ell_2}\ge x)\\ \le&\Pr(\|\sum_{i=1}^N\sum_{t\in I_o} X_{i,t}^*\|_{\ell_2}\ge x) + \Pr(\|\sum_{i=1}^N\sum_{t\in I_e} X_{i,t}^*\|_{\ell_2}\ge x) + \Pr(\|\sum_{i=1}^N\sum_{t\in I_r} X_{i,t}\|_{\ell_2}\ge x) + \frac{NT\beta(w)}{w}.\end{aligned}$$

By choosing $w = c\log(NT)$ for sufficiently large c, we can show that

$$\frac{NT\beta(w)}{w}\lesssim\frac{1}{NT}.$$

For the term $\Pr(\|\sum_{i=1}^N\sum_{t\in I_o} X_{i,t}^*\|_{\ell_2}\ge x)$, notice that $\sum_{i=1}^N\sum_{t\in I_o} X_{i,t}^*$ has been decomposed into the sum of fewer than $N\times\lfloor T\rfloor/w$ independent matrices, i.e., $Z_{i,k}^* = \sum_{t=(k-1)w}^{kw-1} X_{i,t}^*, k\ge 1$. One can show that $\|Z_{i,k}^*\|_{\ell_2}\le\frac{w(\zeta^2+1)}{NT} = w\bar{R}$ and $\max(\|Z_{i,k}^*[Z_{i,k}^*]^\top\|_2, \|[Z_{i,k}^*]^\top Z_{i,k}^*\|_2)\le\frac{w^2(\zeta^2+1)}{(NT)^2} = w^2\sigma^2$. Then by matrix Berstein's inequality, we have

$$\Pr(\|\sum_{i=1}^N\sum_{t\in I_o} X_{i,t}^*\|_{\ell_2}\ge x)\le 2K\exp\left(\frac{-x^2/2}{(NT)w\sigma^2 + w\bar{R}x/3}\right).$$

Then we can bound this probability towards 0 as $K \to \infty$ by choosing $x = C\sigma\sqrt{wNT\log(K)}$ for sufficiently large C with the condition given in the statement that $\bar{R}\sqrt{w\log(K)} = o(\sigma\sqrt{NT})$, i.e., $\zeta\sqrt{\log(NT)\log(K)}/\sqrt{NT} = o(1)$. Similar argument can be applied to $\Pr(\|\sum_{i=1}^N \sum_{t\in I_e} X^*_{i,t}\|_{\ell_2} \geq x)$.

Next, we derive an upper bound for $\Pr(\|\sum_{i=1}^N \sum_{t\in I_r} X_{i,t}\|_{\ell_2} \geq \bar{x})$ for some $\bar{x} > 0$. By Bernstein's inequality and $\sum_{t\in I_r} X_{i,t}$ are independent for $1 \leq i \leq N$, we have

$$\Pr(\|\sum_{i=1}^N \sum_{t\in I_r} X_{i,t}\|_{\ell_2} \geq \bar{x}) \leq 2K \exp\left(\frac{-\bar{x}^2/2}{Nw^2\sigma^2 + w\bar{R}\bar{x}/3}\right).$$

By choosing $\bar{x} = C_1\sigma\sqrt{NTw\log(K)}$ for sufficiently large C_1, we can show that

$$\Pr(\|\sum_{i=1}^N \sum_{t\in I_r} X_{i,t}\|_{\ell_2} \geq \bar{x}) \lesssim K^{-C_1(T/w)+1},$$

as long as $\zeta\sqrt{\log(NT)\log(K)}/\sqrt{NT} = o(1)$. Without loss of generality, we can assume $w \leq T$, which completes the proof in the second statement of Lemma H.2 as the probability converges to 0 as long as $K \to \infty$. Otherwise the result in the statement can be obtained directly by using the Bernstein's inequality in the i.i.d setting without using Berbee's lemma. Other statements follow similarly. □

Lemma H.3. *Under Assumptions 2.1 and 2.2,* $\|G_b^{-1/2}B^\top(\mathbf{R} - H_0)/(NT)\|_{\ell^2} = O_p(\frac{R_{\max}}{1-\gamma}\sqrt{\frac{K}{NT}})$.

Proof. We apply the Markov inequality. Note that

$$\begin{aligned}&\|G_b^{-1/2}B^\top(\mathbf{R} - H_0)/(NT)\|_2^2\\ \leq &\frac{4R_{\max}^2}{(1-\gamma)^2}K/NT,\end{aligned}$$

because all the terms in $G_b^{-1/2}B^\top(\mathbf{R} - H_0)/(NT)$ are uncorrelated by the Bellman equation (1). Hence the proof completes. □

Lemma H.4. *Let* $h_J^\pi(s,a,s') = \kappa_\pi^J(s,a,s')^\top c_J$ *for any deterministic* $c_J \in \mathbb{R}^J$ *and* $H_J = (h_J^\pi(S_{1,0},A_{1,0},S_{1,1}), h_J^\pi(S_{1,1},A_{1,1},S_{1,2}),\ldots,h_J^\pi(S_{N,T-1},A_{N,T-1},S_{N,T}))^\top = \Gamma_\pi c_J$. *Under Assumptions 5.1,*

$$\begin{aligned}&\|G_b^{-1/2}(B^\top(H_0 - H_J)/(NT) - E[b^K(S,A)(h_0^\pi(S,A,S') - h_J^\pi(S,A,S'))])\|_{\ell^2}\\ =\;& O_p\left(\sqrt{\frac{\zeta_{b,K}\log(NT)\log(K)}{NT}} \times \|h_0^\pi - h_J\|_\infty\right).\end{aligned}$$

provided $\sqrt{\frac{\log(NT)\log(K)}{NT}} = o(1)$.

Proof. We again use Berbee's coupling lemma and matrix Bernstein's inequality (e.g., (Dedecker & Louhichi, 2002; Chen & Christensen, 2015)) and get the result. The argument is similar to that in the proof of Lemma H.2. In particular, let

$$Z_{i,t} = G_b^{-1/2}b^K(S_{i,t},A_{i,t})(h_0^\pi(S_{i,t},A_{i,t},S_{i,t+1}) - h_J^\pi(S_{i,t},A_{i,t},S_{i,t+1})).$$

It can be seen that $\|Z_{i,t}\|_{\ell_2} \leq \zeta_{b,K}\|h_0^\pi - h_J\|_\infty$ and

$$\max\{\mathbb{E}[Z_{i,t}^\top Z_{i,t}], \mathbb{E}[Z_{i,t}Z_{i,t}^\top]\} \leq \zeta_{b,K}^2\|h_0^\pi - h_J\|_\infty^2,$$

which gives the result. □

Lemma H.5. *Let $s_{JK}^{-1}\zeta\sqrt{(\log(NT)\log J)/(NT)} = o(1)$, and Assumption 5.1 is satisfied. Then:*

$$(a) \qquad \|(\widehat{G}_b^{-1/2}\widehat{\Sigma}^\pi)_l^- \widehat{G}_b^{-1/2} G_b^{1/2} - (G_b^{-1/2}\Sigma^\pi)_l^- \|_{\ell^2} = O_p\Big(s_{JK}^{-2}\zeta\sqrt{(\log(NT)\log J)/(NTe_J)}\Big)$$

$$(b) \quad \|[G_\kappa^\pi]^{1/2}\{(\widehat{G}_b^{-1/2}\widehat{\Sigma}^\pi)_l^- \widehat{G}_b^{-1/2} G_b^{1/2} - (G_b^{-1/2}\Sigma^\pi)_l^-\}\|_{\ell^2} = O_p\Big(s_{JK}^{-2}\zeta\sqrt{(\log(NT)\log J)/(NT))}\Big)$$

where

$$(G_b^{-1/2}\Sigma^\pi)_l^- = \left[[\Sigma^\pi]^\top (G_b)^{-1}\Sigma^\pi\right]^{-1}\Sigma^\pi]^\top (G_b)^{-1/2},$$

and similarly for $(\widehat{G}_b^{-1/2}\widehat{\Sigma}^\pi)_l^-$.

Proof. We use the similar proof as Lemma F.10 of (Chen & Christensen, 2018) with Berbee's coupling lemma again. The argument is also similar to that in the proof of Lemma H.2. We omit here for brevity. □