

# Online Estimation and Inference for Robust Policy Evaluation in Reinforcement Learning

Weidong Liu<sup>\*</sup> Jiyuan Tu<sup>†</sup> Xi Chen<sup>‡</sup> Yichen Zhang<sup>§</sup>

## Abstract

Reinforcement learning has emerged as one of the prominent topics attracting attention in modern statistical learning, with policy evaluation being a key component. Unlike the traditional machine learning literature on this topic, our work emphasizes statistical inference for the model parameters and value functions of reinforcement learning algorithms. While most existing analyses assume random rewards to follow standard distributions, we embrace the concept of robust statistics in reinforcement learning by simultaneously addressing issues of outlier contamination and heavy-tailed rewards within a unified framework. In this paper, we develop a fully online robust policy evaluation procedure, and establish the Bahadur-type representation of our estimator. Furthermore, we develop an online procedure to efficiently conduct statistical inference based on the asymptotic distribution. This paper connects robust statistics and statistical inference in reinforcement learning, offering a more versatile and reliable approach to online policy evaluation. Finally, we validate the efficacy of our algorithm through numerical experiments conducted in simulations and real-world reinforcement learning experiments.

*Keywords:* Statistical inference; dependent samples; online policy evaluation; robust inference; Bahadur representation.

---

<sup>\*</sup>School of Mathematical Sciences and MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University

<sup>†</sup>School of Mathematics, Shanghai University of Finance and Economics

<sup>‡</sup>Stern School of Business, New York University.

<sup>§</sup>Daniels School of Business, Purdue University

# 1 Introduction

Reinforcement learning has offered immense success and remarkable breakthroughs in a variety of application domains, including autonomous driving, precision medicine, recommendation systems, and robotics (to name a few, e.g., [Murphy, 2003](#); [Kormushev et al., 2013](#); [Mnih et al., 2015](#); [Shi et al., 2018](#)). From recommendation systems to mobile health (mHealth) intervention, reinforcement learning can be used to adaptively make personalized recommendations and optimize intervention strategies learned from retrospective behavioral and physiology data. While the achievements of reinforcement learning algorithms in applications are undisputed, the reproducibility of its results and reliability is still in many ways nascent.

Those recommendation and health applications enjoy great flexibility and affordability due to the development of reinforcement algorithms, despite calling for critical needs for a reliable and trustworthy uncertainty quantification for such implementation. The reliability of such implementations sometimes plays a life-threatening role in emerging applications. For example, in autonomous driving, it is critical to avoid deadly explorations based upon some uncertainty measures in the trial-and-error learning procedure. This substance also extends to other applications including precision medicine and autonomous robotics. From the statistical perspective, it is important to quantify the uncertainty of a point estimate with complementary hypothesis testing to reveal or justify the reliability of the learning procedure.

Policy evaluation plays a cornerstone role in typical reinforcement learning (RL) algorithms such as Temporal Difference (TD) learning. As one of the most commonly adopted algorithms for policy evaluation in RL, TD learning provides an estimator of the value function iteratively with regard to a given policy based on samples from a Markov chain. In large-scale RL tasks where the state space is infinitely expansive, a typical procedure to provide a scalable yet efficient estimation of the value function is via linear function approximation. This procedure can be formulated in a linear stochastic approximation problem ([Sutton, 1988](#); [Tsitsiklis and Van Roy, 1997](#); [Sutton et al., 2009](#); [Ramprasad et al., 2023](#)), which is designed to sequentially solve a deterministic equation  $A\theta = b$  by a matrix-vector pair of a sequence of unbiased random observations of  $(A_t, b_t)$  governed by an ergodic Markov chain.

The earliest and most prototypical stochastic approximation algorithm is the Robbins-Monro algorithm introduced by [Robbins and Monro \(1951\)](#) for solving a root-finding problem, where the function is represented as an expected value, e.g.,  $\mathbb{E}[f(\theta)] = 0$ . The algorithm has generated profound interest in the field of stochastic optimization and machine learning to minimize a loss function using random samples. When referring to an optimization problem, its first-order condition can be represented as  $\mathbb{E}[f(\theta)] = 0$ , and the corresponding Robbins-Monro algorithm is often referred to as first-order methods, or more widely known as stochastic gradient descent (SGD) in machine learning literature. It is well established in the literature that its averaged version ([Ruppert, 1988](#); [Polyak and Juditsky, 1992](#)), as an online first-order method, achieves optimal statistical efficiency when estimating the model parameters in statistical models, which apparently kills the interest in developing second-order methods that use additional information to help the convergence. That being said, it is often observed in practice that first-order algorithms entail significant accuracy loss on non-asymptotic convergence as well as severe instability in the choice of hyperparameters, specifically, the stepsizes (a.k.a. learning rate). In addition, the stepsize tuning further complicates the quantification of uncertainty associated with the algorithm output. Despite the known drawbacks above, first-order stochastic methods are historically favored in machine learning tasks due to their computational efficiency while they primarily focus on estimation. On the other hand, when the emphasis of the task lies on statistical inference, certain computation of the second-order information is generally inevitable during the inferential procedure, which shakes the supremacy of first-order methods over second-order algorithms.

In light of that, we propose a second-order online algorithm which utilizes second-order information to perform the policy evaluation sequentially. Meanwhile, our algorithm can be used for conducting statistical inference in an online fashion, allowing for the characterization of uncertainty in the estimation of the value function. Such a procedure generates no extra per-unit computation and storage cost beyond  $O(d^2)$ , which is at least the same as typical first-order stochastic methods featuring statistical inference (see, e.g., [Chen et al., 2020](#) for SGD and [Ramprasad et al., 2023](#) for TD). More importantly, we show theoretically that the proposed algorithm converges faster in terms of the remainder term compared with first-order stochastic approximation methods, and revealed significant discrepancies in numerical experiments. In addition, the proposed algorithm is

free from tuning stepsizes, which has been well-established as a substantial criticism of first-order algorithms.

Another challenge to the reliability of reinforcement learning algorithms lies in the modeling assumptions. Most algorithms in RL have been in the *optimism in the face of uncertainty* paradigm where such procedures are vulnerable to manipulation (see some earlier exploration in e.g., [Everitt et al., 2017](#); [Wang et al., 2020](#)). In practice, it is often unrealistic to believe that rewards on the entire trajectory follow exactly the same underlying model. Indeed, non-standard behavior of the rewards happens from time to time in practice. The model-misspecification and presence of outliers are indeed very common in an RL environment, especially that with a large time horizon  $T$ . It is of substantial interest to design a *robust* policy evaluation procedure. In pursuit of this, our proposed algorithm uses a smoothed Huber loss to replace the least-squares loss function used in classical TD learning, which is tailored to handle both outliers and heavy-tailed rewards. To model outlier observations of rewards in reinforcement learning, we bring the static  $\alpha$ -contamination model ([Huber, 1992](#)) to an online setting with dependent samples. In a static offline robust estimation problem, one aims to learn the distribution of interest, where a sample of size  $n$  is drawn i.i.d. from a mixture distribution  $(1 - \alpha_n)P + \alpha_n Q$ , and  $Q$  denotes an arbitrary outlier distribution. We adopt robust estimation in an online environment, where the observations are no longer independent, and the occurrence time of outliers is unknown. In contrast to the offline setting, future observations cannot be used in earlier periods in an online setting. Therefore, in the earlier periods, there is very limited information to help determine whether an observation is an outlier. In addition to this discrepancy between the online decision process and offline estimation, we further allow the outlier reward models to be potentially different for different time  $t$  instead of being from a fixed distribution, and such rewards may be arbitrary and even adversarially adaptive to historical information. In addition to the outlier model, our model also incorporates rewards with heavy-tailed distributions. This substantially relaxes the boundedness condition on the reward functions that is common in policy evaluation literature.

We summarize the challenges and contributions of this paper in the following facets.

- We propose a *fully online* method that leverages dependent samples to simultaneously per-

form policy evaluation and conduct statistical inference on the model parameters and value functions. Furthermore, we build a Bahadur-type representation of the proposed estimator, which includes the main term corresponding to the asymptotic normal distribution and a higher-order remainder term. Moreover, it shows that our algorithm matches the offline oracle and converges strictly faster to the asymptotic distribution than that of a prototypical first-order stochastic method such as TD learning.

- Compared to existing reinforcement learning literature, our proposed algorithm features an online generalization of the  $\alpha_n$ -contamination model where the rewards contain outliers or arbitrary corruptions. Our proposed algorithm is robust to adversarial corruptions which can be adaptive to the trajectory, as well as heavy-tailed distribution of the rewards. Due to the existence of outliers, we use a smooth Huber loss where the thresholding parameter is carefully specified to change over time to accommodate the online streaming data. From a theoretical standpoint, a robust policy evaluation procedure forces the update step from  $\theta_t$  to  $\theta_{t+1}$  to be a non-linear function of  $\theta_t$ , which brings in additional technical challenges compared to the analysis of classical TD learning algorithms (see e.g., [Ramprasad et al., 2023](#)) based on linear stochastic approximation.
- Our proposed algorithm is based on a dedicated averaged version of the second-order method, where in each iteration a surrogate Hessian is obtained and used in the update step. This second-order information enables the proposed algorithm to be free from stepsize tuning while still ensuring efficient implementation. Furthermore, our proposed algorithm stands out distinctly from conventional first-order stochastic approximation approaches which fall short of attaining the optimal offline remainder rate. On the other hand, while deterministic second-order methods do excel in offline scenarios, they lack the online adaptability crucial for real-time applications.

## 1.1 Related Works

Conducting statistical inference for model parameters in stochastic approximation has attracted great interest in the past decade, with a building foundation of the asymptotic distribution of the

averaged version of stochastic approximation first established in [Ruppert \(1988\)](#); [Polyak and Juditsky \(1992\)](#). The established asymptotic distribution has been brought to conduct online inference. For example, [Fang et al. \(2018\)](#) presented a perturbation-based resampling procedure. [Chen et al. \(2020\)](#) proposed two online procedures to estimate the asymptotic covariance matrix to conduct inference. [Chen et al. \(2021c\)](#) studied inference for zeroth-order stochastic approximation. [Shi et al. \(2021a\)](#) developed online inference procedures for high-dimensional problems. Other than those focused on the inference procedures, [Shao and Zhang \(2022\)](#) established the rate of convergence in the remainder term, and utilize it to establish Berry-Esseen bounds of the averaged SGD estimator. [Givchi and Palhang \(2015\)](#); [Mou et al. \(2021\)](#) analyzed averaged SGD with Markovian data. Second-order stochastic algorithms were analyzed in [Ruppert \(1985\)](#); [Schraudolph et al. \(2007\)](#); [Byrd et al. \(2016\)](#) and applied to TD learning in [Givchi and Palhang \(2015\)](#).

Under the online decision-making settings including bandit algorithms and reinforcement learning, a few existing works focused on statistical inference of the model parameters or uncertainty quantification of value functions. [Deshpande et al. \(2018\)](#); [Hadad et al. \(2021\)](#); [Zhang et al. \(2021, 2022\)](#) studied statistical inference with adaptively sampled data. [Zhan et al. \(2021\)](#); [Chen et al. \(2021a,b\)](#); [Dimakopoulou et al. \(2021\)](#); [Chen et al. \(2022\)](#); [Han et al. \(2022\)](#) proposed online inference procedures for bandit algorithms particularly. [Shen et al. \(2024\)](#) studied optimal policy evaluation and constructed valid confidence intervals for the value of the optimal policy. For reinforcement learning algorithms, [Thomas et al. \(2015\)](#) proposed high-confidence off-policy evaluation based on Bernstein inequality. [Hanna et al. \(2017\)](#) presented two bootstrap methods to compute confidence bounds for off-policy value estimates. [Dai et al. \(2020\)](#); [Feng et al. \(2021\)](#) construct confidence intervals for value functions based on optimization formulations, and [Jiang and Huang \(2020\)](#) derived a minimax value interval, both with i.i.d. sample. [Shi et al. \(2021b\)](#) proposed inference procedures for Q functions in RL via sieve approximations. [Hao et al. \(2021\)](#) studied multiplier bootstrap algorithms to offer uncertainty quantification for exploration in fitted Q-evaluation. [Syrkkanis and Zhan \(2023\)](#) studied a re-weighted Z-estimator on episodic RL data and conducted inference on the structural parameter.

The most relevant literature to ours is [Ramprasad et al. \(2023\)](#), who studied a bootstrap online statistical inference procedure under Markov noise using a quadratic SGD and demonstrated its

application in the classical TD (and Gradient TD) algorithms in RL. Our proposed procedure and analysis differ in at least two aspects. First, our proposed estimator is a Newton-type second-order approach that enjoys a faster convergence and optimality in the remainder rate. In addition, we show both analytically and numerically that the computation cost of our procedure is typically lower than [Ramprasad et al. \(2023\)](#) for an inference task. Second, our proposed algorithm is a robust alternative to TD algorithms, featuring a non-quadratic loss function to handle the potential outliers and heavy-tailed rewards. There exists limited RL literature on either outliers or heavy-tailed rewards. Recent works [Li and Sun \(2023\)](#); [Zhu et al. \(2024\)](#) studied online linear stochastic bandits and offline reinforcement learning in the presence of heavy-tailed rewards. However, their studies do not apply to online statistical inference.

## 1.2 Paper Organization and Notations

The remainder of this paper is organized as follows. In Sections 2 and 3, we present and discuss our proposed algorithm for robust policy evaluation in reinforcement learning. Theoretical results on convergence rates, asymptotic normality, and the Bahadur representation are presented in Section 3. In Section 4, we develop an estimator for the long-run covariance matrix to construct confidence intervals in an online fashion and provide its theoretical guarantee. Simulation experiments are provided in Section 5 to demonstrate the effectiveness of our method. Concluding remarks are given in Section 6. All proofs are deferred to the supplementary material.

For every vector  $\mathbf{v} = (v_1, \dots, v_d)^\top$ , denote  $|\mathbf{v}|_2 = \sqrt{\sum_{l=1}^d v_l^2}$ ,  $|\mathbf{v}|_1 = \sum_{l=1}^d |v_l|$ , and  $|\mathbf{v}|_\infty = \max_{1 \leq l \leq d} |v_l|$ . For simplicity, we denote  $\mathbb{S}^{d-1}$  and  $\mathbb{B}^d$  as the unit sphere and unit ball in  $\mathbb{R}^d$  centered at  $\mathbf{0}$ . Moreover, we use  $\text{supp}(\mathbf{v}) = \{1 \leq l \leq d \mid v_l \neq 0\}$  as the support of the vector  $\mathbf{v}$ . For every matrix  $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2}$ , define  $\|\mathbf{A}\| = \sup_{|\mathbf{v}|_2=1} |\mathbf{A}\mathbf{v}|_2$  as the matrix operator norms,  $\Lambda_{\max}(\mathbf{A})$  and  $\Lambda_{\min}(\mathbf{A})$  as the largest and smallest singular values of  $\mathbf{A}$  respectively. The symbols  $\lfloor x \rfloor$  ( $\lceil x \rceil$ ) denote the greatest integer (the smallest integer) not larger than (not less than)  $x$ . We denote  $(x)_+ = \max(0, x)$ . For two sequences  $a_n, b_n$ , we say  $a_n \asymp b_n$  when  $a_n = O(b_n)$  and  $b_n = O(a_n)$  hold at the same time. We say  $a_n \approx b_n$  if  $\lim_{n \rightarrow \infty} a_n/b_n = 1$ . For a sequence of random variables  $\{X_n\}_{n=1}^\infty$ , we denote  $X_n = O_{\mathbb{P}}(a_n)$  if there holds  $\lim_{C \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{P}(|X_n| > Ca_n) = 0$ , and

denote  $X_n = o_{\mathbb{P}}(a_n)$  if there holds  $\lim_{C \rightarrow 0} \limsup_{n \rightarrow \infty} \mathbb{P}(|X_n| > Ca_n) = 0$ . Lastly, the generic constants are assumed to be independent of  $n$  and  $d$ .

## 2 Online Robust Policy Evaluation in Reinforcement Learning

We first review the Least-squares temporal difference methods in RL. Consider a 4-tuple  $(\mathcal{S}, \mathcal{U}, \mathcal{P}, \mathcal{R})$ . Here  $\mathcal{S} = \{1, 2, \dots, N\}$  is the global finite state space,  $\mathcal{U}$  is the set of control (action),  $\mathcal{P}$  is the transition kernel, and  $\mathcal{R}$  is the reward function. One of the core steps in RL is to estimate the cumulative reward  $J^*$  (which is also called the state value function) for a given policy:

$$J^*(s) = \mathbb{E} \left[ \sum_{k=0}^{\infty} \gamma^k \mathcal{R}(s_k) \mid s_0 = s \right],$$

where  $\gamma \in [0, 1)$  is a given discount factor and  $s \in \mathcal{S}$  is any state. Here  $\{s_k\}$  denote the environment states which are usually modeled by a Markov chain. In real-world RL applications, the state space is often very large such that one cannot directly compute value estimates for every state in the state space. A common approach in the modern RL is to approximate the value function  $J^*(\cdot)$ , i.e., let

$$\tilde{J}(s, \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \boldsymbol{\phi}(s) = \sum_{l=1}^d \theta_l \phi_l(s),$$

be a linear approximation of  $J^*(\cdot)$ , where  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)^\top \in \mathbb{R}^d$  contains the model parameters, and  $\boldsymbol{\phi}(s) = (\phi_1(s), \dots, \phi_d(s))^\top \in \mathbb{R}^d$  is a set of feature vectors that corresponds to the state  $s \in \mathcal{S}$ . Here we write  $\boldsymbol{\phi}_l = (\phi_l(1), \dots, \phi_l(N))^\top$ ,  $1 \leq l \leq d$  are linearly independent vectors in  $\mathbb{R}^N$  and  $d \ll N$ . That is, we use a low dimensional linear approximation (with the basis vectors  $\{\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_d\}$ ) for  $J^*(\cdot)$ . Let the matrix  $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_d)$  and then  $\tilde{J} = \boldsymbol{\Phi} \boldsymbol{\theta}$ .

The state value function  $J^*$  satisfies the Bellman equation

$$J^*(s) = \mathcal{R}(s) + \gamma \mathbb{E} \left[ \sum_{k=1}^{\infty} \gamma^{k-1} \mathcal{R}(s_k) \mid s_0 = s \right] = \mathcal{R}(s) + \gamma \sum_{s'=1}^N p_{ss'} J^*(s'), \quad (1)$$

where  $s'$  is the next state transferred from  $s$ . Let  $\mathcal{P} = (p_{ss'})_{N \times N}$  be the probability transition matrix and define the Bellman operator  $\mathcal{T}$  by  $\mathcal{T}(Q) = \mathcal{R} + \gamma \mathcal{P}Q$ . The state function  $J^*$  is the



unique fixed point of the Bellman operator  $\mathcal{T}$ . When  $J^*$  is replaced by  $\tilde{J} = \Phi\theta$ , the Bellman equation may not hold.

The classical temporal difference attempts to find  $\theta$  such that  $\tilde{J}(s, \theta)$  well approximates the value function  $J^*(s)$ . Particularly, the TD algorithm updates

$$\theta_{t+1} = \theta_t - \eta_t \phi(s_t)[(\phi^\top(s_t) - \gamma \phi^\top(s_{t+1}))\theta_t - \mathcal{R}(s_t)], \quad t \geq 0, \quad (2)$$

where  $\eta_t$  is the step size which often requires careful tuning. We next illustrate the key point that (2) leads to a good estimator such that  $\tilde{J}(s, \theta_t)$  is close to  $J^*(s)$ . It can be shown that  $\theta_t$  converges to an unknown population parameter  $\theta^*$  that minimizes the expected squared difference of the Bellman equation,

$$\theta^* = \underset{u \in \mathbb{R}^d}{\operatorname{argmin}} \mathbb{E} |\phi^\top(s)u - (\mathcal{R}(s) + \gamma \phi^\top(s')\theta^*)|^2; \quad (3)$$

It is easy to see that such  $\theta^*$  exists and satisfies the following first-order condition,

$$\mathbb{E} \phi(s)[(\phi^\top(s) - \gamma \phi^\top(s'))\theta^* - \mathcal{R}(s)] = 0. \quad (4)$$

Under the condition that  $\mathbf{H} := \mathbb{E} \phi(s)(\phi^\top(s) - \gamma \phi^\top(s'))$  is positive definite in the sense that  $x^\top \mathbf{H} x > 0$  for all  $x \in \mathbb{R}^d$ ; see [Tsitsiklis and Van Roy \(1997\)](#), the solution of (4) writes

$$\theta^* = (\mathbb{E} \phi(s)(\phi^\top(s) - \gamma \phi^\top(s')))^{-1} \mathbb{E} \phi(s) \mathcal{R}(s).$$

The estimation equation (2) is a first-order stochastic algorithm that converges to the stochastic root-finding problem (4) using a sequence of observations  $\{s_t, r_t\}_{t \geq 1}$ . In this paper, we refer to it as the Least-squares temporal difference estimator. See [Kolter and Ng \(2009\)](#) for details on the properties of the Least-squares TD estimator.

Least-squares-based methods are oftentimes criticized due to their sensitivity to outliers in data. When there may exist outliers in some observations of the reward  $\mathcal{R}(s)$ , it is a natural call for interest to design a robust estimator of  $\theta^*$ . Following the widely-known classical literature ([Huber, 1964](#); [Charbonnier et al., 1994, 1997](#); [Hastie et al., 2009](#)), we replace the square loss  $|\cdot|^2$

by a smoothed (Pseudo) Huber loss  $f_\tau(x) = \tau^2(\sqrt{1 + (x/\tau)^2} - 1)$ , parametrized by a thresholding parameter  $\tau$ . We define a similar fixed point equation with the smoothed Huber loss by

$$\boldsymbol{\theta}_\tau^* = \underset{\mathbf{u} \in \mathbb{R}^d}{\operatorname{argmin}} \mathbb{E} f_\tau(\boldsymbol{\phi}^\top(s) \mathbf{u} - (\mathcal{R}(s) + \gamma \boldsymbol{\phi}^\top(s') \boldsymbol{\theta}_\tau^*)). \quad (5)$$

In this section, when we motivate the algorithm, we assume that the fixed point  $\boldsymbol{\theta}_\tau^*$  exists<sup>1</sup>. As the thresholding parameter  $\tau$  goes to infinity, the objective equation in (5) is close to the least-squares loss in (3), and  $\boldsymbol{\theta}_\tau^*$  should be close to  $\boldsymbol{\theta}^*$ . When  $\tau$  tends to 0, the problem (5) becomes  $\boldsymbol{\theta}_0^* = \underset{\mathbf{u} \in \mathbb{R}^d}{\operatorname{argmin}} \mathbb{E} |\boldsymbol{\phi}^\top(s) \mathbf{u} - (\mathcal{R}(s) + \gamma \boldsymbol{\phi}^\top(s') \boldsymbol{\theta}_0^*)|$ , with a nonsmooth least absolute deviation (LAD) loss, which is out of the scope of this paper. In this paper, we carefully specify  $\tau$  to balance the statistical efficiency and the effect of potential outliers in an online fashion (see Theorem 1).

By the first-order condition of (5), we obtain a similar estimation function as (4),

$$\mathbb{E} [\boldsymbol{\phi}(s) g_\tau((\boldsymbol{\phi}^\top(s) - \gamma \boldsymbol{\phi}^\top(s')) \boldsymbol{\theta}_\tau^* - \mathcal{R}(s))] = 0, \quad (6)$$

where the function  $g_\tau(x) = f'_\tau(x)$  is the differential of the smoothed Huber loss  $f_\tau(x)$ . Instead of using the first-order iteration with the estimation equation (6) as in TD (2), we propose a Newton-type iterative estimator, which avoids the tuning of the learning rate. Newton-type estimators are often referred to as second-order methods when discussed in convex optimization. Nonetheless, the equation (5) is just for illustration purposes, which cannot be directly optimized since  $\boldsymbol{\theta}_\tau^*$  appears in the objective function for minimization. On the contrary, our proposed method is a Newton-type method for solving the root-finding problem (6) using a sequence of the observations.

In this paper, we model two types of noise in observed rewards. The first is the classical Huber contamination model (Huber, 1992, 2004), where an  $\alpha_n$ -fraction of rewards comes from arbitrary distributions. The second is the heavy-tailed model, where the reward function may admit heavy-tailed distributions. In the following section, we show that our proposed estimator is robust to the aforementioned two noises.

---

<sup>1</sup>Here  $\boldsymbol{\theta}_\tau^*$  is used for the motivation only. Our algorithm and theory do not depend on the existence of  $\boldsymbol{\theta}_\tau^*$ .

### 3 Online Newton-type Method for Parameter Estimation

In the following, we introduce an online Newton-type method for estimating the parameter  $\boldsymbol{\theta}^*$  in the presence of outliers and heavy-tailed noise in the reward function  $\mathcal{R}(s)$ . For ease of presentation, we denote the observations by  $\mathbf{X}_i = \boldsymbol{\phi}(s_i)$ ,  $\mathbf{Z}_i = \boldsymbol{\phi}(s_i) - \gamma\boldsymbol{\phi}(s_{i+1})$ , and  $b_i = \mathcal{R}(s_i)$ , where  $s_i$  is the state at time  $i$ . Here  $(\mathbf{X}_i, \mathbf{Z}_i, b_i) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$  for each observation in the sample. Our objective is to propose an online estimator by the estimation function (6), which can be rewritten as

$$\mathbb{E}[\mathbf{X}g_\tau(\mathbf{Z}^\top \boldsymbol{\theta}_\tau^* - b)] = 0, \quad (7)$$

based on a sequence of dependent observations  $\{(\mathbf{X}_i, \mathbf{Z}_i, b_i)\}_{i \geq 1}$ .

At iteration  $n + 1$ , our proposed Newton-type estimator updates  $\hat{\boldsymbol{\theta}}_{n+1}$  by

$$\hat{\boldsymbol{\theta}}_{n+1} = \frac{1}{n+1} \sum_{i=0}^n \hat{\boldsymbol{\theta}}_i - \widehat{\mathbf{H}}_{n+1}^{-1} \frac{1}{n+1} \sum_{i=0}^n \mathbf{X}_{i+1} g_{\tau_{i+1}}(\mathbf{Z}_{i+1}^\top \hat{\boldsymbol{\theta}}_i - b_{i+1}). \quad (8)$$

Here  $\widehat{\mathbf{H}}_{n+1}$  is an empirical information matrix of the estimation equation (7), as

$$\widehat{\mathbf{H}}_{n+1} = \frac{1}{n+1} \sum_{i=0}^n \mathbf{X}_{i+1} \mathbf{Z}_{i+1}^\top g'_{\tau_{i+1}}(\mathbf{Z}_{i+1}^\top \hat{\boldsymbol{\theta}}_i - b_{i+1}). \quad (9)$$

where  $\tau_i$  is the thresholding parameter in the Huber loss. We let  $\tau_n$  tend to infinity to eliminate the bias generated by the smoothed Huber loss.

It is noteworthy to mention that the matrix  $\widehat{\mathbf{H}}_{n+1}$  is not the Hessian matrix of the objective function on the right-hand side of (5). As discussed in the previous section, (5) cannot be directly optimized, nor do they lead to M-estimation problems. Indeed, our proposed update (8) is a Newton-type method to find the root of (7) using a sequence of observations  $\{(\mathbf{X}_i, \mathbf{Z}_i, b_i)\}_{i \geq 1}$ . In the following section, we will show the use of the matrix  $\widehat{\mathbf{H}}_{n+1}$  helps for desirable convergence properties.

It should be noted that (8) can be implemented efficiently in a fully-online manner, i.e., without storage of the trajectory of historical information. Specifically, we write (8) as  $\hat{\boldsymbol{\theta}}_{n+1} = \bar{\boldsymbol{\theta}}_{n+1} - \widehat{\mathbf{H}}_{n+1}^{-1} \mathbf{G}_{n+1}$ , with each of the item on the right-hand side being a running average. It is easy to see

---

**Algorithm 1** Robust Online Policy Evaluation (ROPE).

---

**Input:** Online streaming data  $\{(\mathbf{X}_i, \mathbf{Z}_i, b_i) \mid i \geq 1\}$

- 1: Compute  $\widehat{\mathbf{H}}_{n_0}^{-1}$  and  $\mathbf{G}_{n_0}$  according to (12) respectively.
- 2: **for**  $n = n_0 + 1, n_0 + 2, \dots$  **do**
- 3:   Specify the thresholding parameter  $\tau_n$ .
- 4:   Compute  $\bar{\boldsymbol{\theta}}_n = (n-1)\bar{\boldsymbol{\theta}}_{n-1}/n + \widehat{\boldsymbol{\theta}}_{n-1}/n$ , and  $(\widehat{\mathbf{H}}_n^{-1}, \mathbf{G}_n)$  by (10) and (11).
- 5:   Update the parameter by

$$\widehat{\boldsymbol{\theta}}_n = \bar{\boldsymbol{\theta}}_n - \widehat{\mathbf{H}}_n^{-1} \mathbf{G}_n.$$

6: **end for**

---

that the averaged estimator  $\bar{\boldsymbol{\theta}}_{n+1} = \frac{1}{n+1}(n\bar{\boldsymbol{\theta}}_n + \widehat{\boldsymbol{\theta}}_n)$  and the vector

$$\begin{aligned} \mathbf{G}_{n+1} &= \frac{1}{n+1} \sum_{i=0}^n \mathbf{X}_{i+1} g_{\tau_{i+1}}(\mathbf{Z}_{i+1}^\top \widehat{\boldsymbol{\theta}}_i - b_{i+1}) \\ &= \frac{n}{n+1} \mathbf{G}_n + \frac{1}{n+1} \mathbf{X}_{n+1} g_{\tau_{n+1}}(\mathbf{Z}_{n+1}^\top \widehat{\boldsymbol{\theta}}_n - b_{n+1}) \end{aligned} \quad (10)$$

can both be updated online. In addition, the inverse  $\widehat{\mathbf{H}}_{n+1}^{-1}$  can be directly and efficiently computed online by the inverse recursion formulation. By the Sherman-Morrison formula, we have

$$\begin{aligned} \widehat{\mathbf{H}}_{n+1}^{-1} &= \frac{n+1}{n} \widehat{\mathbf{H}}_n^{-1} - \frac{n+1}{n^2} \widehat{\mathbf{H}}_n^{-1} \mathbf{X}_{n+1} \\ &\quad \times \left[ \frac{1}{n} \mathbf{Z}_{n+1}^\top \widehat{\mathbf{H}}_n^{-1} \mathbf{X}_{n+1} + \{g'_{\tau_{n+1}}(\mathbf{Z}_{n+1}^\top \widehat{\boldsymbol{\theta}}_n - b_{n+1})\}^{-1} \right]^{-1} \mathbf{Z}_{n+1}^\top \widehat{\mathbf{H}}_n^{-1}. \end{aligned} \quad (11)$$

Here we note that both terms  $\frac{1}{n} \mathbf{Z}_{n+1}^\top \widehat{\mathbf{H}}_n^{-1} \mathbf{X}_{n+1}$  and  $\{g'_{\tau_{n+1}}(\mathbf{Z}_{n+1}^\top \widehat{\boldsymbol{\theta}}_n - b_{n+1})\}^{-1}$  inside the brackets on the right-hand side of (11) are scalars in  $\mathbb{R}^1$ , not matrices.

The complete algorithm is presented in Algorithm 1. We refer to it as the Robust Online Policy Evaluation (ROPE). Compared with existing stochastic approximation algorithms for TD learning (Durmus et al., 2021; Mou et al., 2021; Ramprasad et al., 2023), our ROPE algorithm does not need to tune the step size.

Since performing iterations of  $\widehat{\mathbf{H}}_{n+1}^{-1}$  in (11) requires an initial invertible  $\widehat{\mathbf{H}}_{n_0}$ , we shall compute

from the first  $n_0$  samples that

$$\widehat{\mathbf{H}}_{n_0} = \frac{1}{n_0} \sum_{i=1}^{n_0} \mathbf{X}_i \mathbf{Z}_i^\top g'_{\tau_0}(\mathbf{Z}_i^\top \widehat{\boldsymbol{\theta}}_0 - b_i), \quad \mathbf{G}_{n_0} = \frac{1}{n_0} \sum_{i=1}^{n_0} \mathbf{X}_i g_{\tau_0}(\mathbf{Z}_i^\top \widehat{\boldsymbol{\theta}}_0 - b_i), \quad (12)$$

which serves as the initial quantities of (10)–(11) in order to compute the forthcoming iterations. Here  $\widehat{\boldsymbol{\theta}}_0$  is a given initial parameter, and  $\tau_0$  is a specified initial threshold level.

### 3.1 Convergence Rate of ROPE

Before illustrating how to conduct statistical inference on the parameter  $\boldsymbol{\theta}^*$ , we first provide theoretical results for our proposed ROPE method.

To characterize the weak dependence of the sequence of environment states, we use the concept of  $\phi$ -mixing. More precisely, we assume  $\{s_i, i \geq 1\}$  is a sequence of  $\phi$ -mixing dependent variables, which satisfy

$$|\mathbb{P}(B|A) - \mathbb{P}(B)| \leq \phi(k) \quad (13)$$

for all  $A \in \mathcal{F}_1^n, B \in \mathcal{F}_{n+k}^\infty$  and all  $n, k \geq 1$ , where  $\mathcal{F}_a^b = \sigma(s_i, a \leq i \leq b)$ . The  $\phi$ -mixing dependence covers the irreducible and aperiodic Markov chain, which is typically used to model the states sampling in reinforcement learning (RL) (Rosenblatt, 1956; Rio, 2017, and others). The mixing rate  $\phi(k)$  identifies the strength of dependence of the sequence. In this paper, we assume the following condition on the mixing rate.

**Condition (C1).** *The sequence  $\{(\mathbf{X}_i, \mathbf{Z}_i)\}$  is a stationary  $\phi$ -mixing sequence, where the mixing rate satisfies  $\phi(k) = O(\rho^k)$  for some  $0 < \rho < 1$ .*

Condition (C1) holds when  $\{s_i\}$  is an irreducible and aperiodic Markov chain with finite state space (Tsitsiklis and Van Roy, 1997). While the main techniques of our analysis could be extended to accommodate infinite and even continuous state spaces, we restrict our study to finite state spaces, since our proposed robust estimator targets  $\boldsymbol{\theta}^*$ , the fixed point of the projected Bellman equation. The deviation of this fixed point from that of the original Bellman equation, known as the approximation error, is well understood in the literature (Tsitsiklis and Van Roy, 1997; Bhandari et al., 2018) for finite state spaces in TD learning. In contrast, the approximation error under

continuous state spaces introduce additional challenges. Since our primary focus is inference for online policy evaluation under heavy-tailed rewards and outliers, we concentrate on establishing the statistical properties of the proposed estimator in the finite-state setting.

In addition, we also assume the following boundedness condition on  $\mathbf{X} = \phi(s)$  that is commonly required by the RL literature (Sutton and Barto, 2018; Ramprasad et al., 2023).

**Condition (C2).** *There exists a constant  $C_X > 0$  such that  $\max\{|\phi(s)|_2 \mid s \in \mathcal{S}\} \leq C_X$ .*

Condition (C2) imposes a uniform bound on the feature space that is independent of  $d$ , which ensures that our theoretical bounds scale appropriately with the feature dimension. Next, we assume the matrix  $\mathbf{H} = \mathbb{E}[\mathbf{X}\mathbf{Z}^\top]$  has a bounded condition number.

**Condition (C3).** *There exists a constant  $c > 0$  such that  $c \leq \Lambda_{\min}(\mathbf{H}) \leq \Lambda_{\max}(\mathbf{H}) \leq c^{-1}$ , where  $\Lambda_{\max}$  and  $\Lambda_{\min}$  denote the largest and the smallest singular values, respectively.*

When  $\{s_i\}$  is an irreducible and aperiodic Markov chain, Tsitsiklis and Van Roy (1997) proved that  $\mathbf{H}$  is positive definite, i.e.,  $x^\top \mathbf{H} x > 0$  for all  $x \neq 0$ . This indicates that  $\mathbf{H}$  is nonsingular and hence (C3) holds. Finally, we provide conditions on the temporal difference error  $\mathbf{Z}^\top \boldsymbol{\theta}^* - b$ .

**Condition (C4).** *The temporal-difference error  $\mathbf{Z}^\top \boldsymbol{\theta}^* - b$  comes from the distribution  $(1 - \alpha_n)P + \alpha_n Q$ , where  $\alpha_n \in [0, 1)$ . The distribution  $Q$  is an arbitrary distribution, and  $P$  satisfies that  $\mathbb{E}_P[|\mathbf{Z}^\top \boldsymbol{\theta}^* - b|^{1+\delta} \mid \mathbf{X}, \mathbf{Z}] \leq C_b$  holds uniformly for some constants  $\delta > 0$  and  $C_b > 0$ .*

We assume that the temporal-difference error comes from the Huber contamination model. The outlier distribution  $Q$  can be arbitrary and the true distribution  $P$  has  $(1 + \delta)$ -th order of moment ( $\delta > 0$ ), which does not necessarily indicate a variance exists when  $\delta \in (0, 1)$ . Therefore, our assumption largely extends the boundedness condition of the reward function  $\mathcal{R}(s)$  in RL literature (Sutton and Barto, 2018; Ramprasad et al., 2023). In Condition (C4),  $\alpha_n$  controls the ratio of outliers among  $n$  samples. Particularly,  $\alpha_n = m_n/n$  means that there are  $m_n$  outliers among  $n$  samples  $\{(\mathbf{X}_i, \mathbf{Z}_i, b_i), 1 \leq i \leq n\}$ . Given the above conditions, by selecting the thresholding parameter  $\tau_i$  as defined in (8), we can obtain the convergence rate for all the intermediate iterates  $\hat{\boldsymbol{\theta}}_i$ .

**Theorem 1.** Suppose that (C1) to (C4) hold and the thresholding parameter  $\tau_i = C_\tau \max(1, i^{\beta_1}/(\log i)^{\beta_2})$  (where  $\beta_1 \in [0, 1)$ ,  $\beta_2 \geq 0$ , and  $C_\tau > 0$ ). Assume  $n_0$  is sufficiently large and the initial value  $|\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*|_2 \leq c_0$  for some  $c_0 < 1$ . Then for every  $\nu > 0$ , there exist constants  $C, c > 0$  which are independent of the dimension  $d$  such that

$$\mathbb{P}\left(\cap_{i=n_0}^n \{|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}^*|_2 \geq C\sqrt{d}e_i\}\right) \geq 1 - cn_0^{-\nu},$$

where

$$e_n = \alpha_n \tau_n + \tau_n^{-\min(\delta, 2)} + \sqrt{\frac{\tau_n^{(1-\delta)+} \log n}{n}} + \frac{\tau_n \log^2 n}{n} + \frac{1}{\sqrt{d}}(c_0)^{2^{n-n_0}}. \quad (14)$$

Here  $\delta$  is defined in the moment condition in (C4).

The error bounds  $e_n$  in (14) contains five terms. In the sequel, we will discuss each of these terms individually. The first term  $\alpha_n \tau_n$  comes from the impact of outliers among the samples. Here the contamination ratio  $\alpha_n = m_n/n$  should tend to zero; otherwise, it is impossible to obtain a consistent estimator for  $\boldsymbol{\theta}^*$ . The second term  $\tau_n^{-\min(\delta, 2)}$  is due to the bias from smoothed Huber's loss for mean estimation. The third and the fourth term in (14) is the classical statistical rate due to the variance and boundedness incurred by thresholding, respectively. These two terms commonly appear in the Bernstein-type inequalities (see, e.g. [Bennett, 1962](#)). As we can see from the third term, the convergence rate has a phase transition between the regimes of finite variance  $\delta > 1$  and infinite variance  $0 < \delta \leq 1$  (see also [Corollary 3](#) below). This phenomenon has also been observed in different estimation problems with Huber loss; see [Sun et al. \(2020\)](#) and [Fan et al. \(2022\)](#). In the presence of a higher moment condition, specifically  $\delta \geq 5$ , we can eliminate the fourth term by applying a more delicate analysis (See [Proposition 5](#) below and [Proposition 8](#) in the supplementary material for more details). The last term represents how quickly the proposed iterative algorithm converges. Given an initial value  $\hat{\boldsymbol{\theta}}_0$  with an error  $c_0 < 1$ , the proposed second-order method enjoys a local quadratic convergence, i.e., the error evolves from  $c_0$  to  $(c_0)^{2^{n-n_0}}$  after  $n - n_0$  iterations. This term decreases super-exponentially fast, a characteristic convergence rate often encountered in the realm of second-order methods (see, e.g., [Nesterov, 2003](#)). Specifically, when the sample size, which is equivalent to the number of iterations, is reasonably large, the last term  $(c_0)^{2^{n-n_0}} = O(1/\sqrt{n})$  is

dominated by the statistical error in the other terms of (14). The assumption on the initial error  $|\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*|_2 \leq c_0$  for some  $c_0 < 1$  is mild. A valid initial value  $\hat{\boldsymbol{\theta}}_0$  can be obtained by the solution to the following estimation equation  $\sum_{i=0}^{n_0} \mathbf{X}_i g_\tau(\mathbf{Z}_i^\top \boldsymbol{\theta} - b_i) = 0$  using a subsample with fixed size  $n_0$  before running Algorithm 1 with a specified constant threshold  $\tau$ . The above equation can be solved efficiently by classical first-order root-finding algorithms.

To highlight the relationship between the convergence rate and the outlier rewards, we present the following corollary which gives a clear statement on how the rate depends on the number of outliers  $m_n$ .

**Corollary 2.** *Suppose the conditions in Theorem 1 hold with  $\delta \geq 2$ . Let the thresholding parameter  $\tau_i = C_\tau i^\beta$  with some  $\beta, C_\tau > 0$  and  $n_0$  tend to infinity. We have*

$$|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*|_2 = O_{\mathbb{P}} \left( \sqrt{d} \left( \sqrt{\frac{\log n}{n}} + \frac{\log^2 n}{n^{1-\beta}} + \frac{m_n}{n^{1-\beta}} + \frac{1}{n^{2\beta}} \right) \right),$$

where  $m_n$  is the number of outliers among the samples of size  $n$ .

The corollary shows that, when there are  $m_n = o(n^{1-\beta})$  outliers, our ROPE can still estimate  $\boldsymbol{\theta}^*$  consistently. Note that  $\beta$  in  $\tau_i$  is the exponent specified by the practitioner. A smaller  $\beta$  allows more outliers  $m_n$  among the samples. Furthermore, when there are  $m_n = o(n^{\frac{1}{2}-\beta})$  outliers and  $\frac{1}{4} \leq \beta \leq \frac{1}{2}$ , the ROPE estimator achieves the optimal rate  $|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*|_2 = O_{\mathbb{P}} \left( \sqrt{\frac{d \log n}{n}} \right)$ , up to a logarithm term.

The next corollary indicates the impact of the tail of rewards  $\mathcal{R}(s)$  on the convergence rate. For a clear presentation, we discuss the impact under the case without outliers, i.e.,  $\alpha_n = 0$ .

**Corollary 3.** *Suppose the conditions in Theorem 1 hold with the contamination rate  $\alpha_n = 0$  and let  $n_0$  tend to infinity.*

- When  $\delta \in (0, 1]$ , we specify  $\tau_i = C_\tau (i / \log i)^{1/(1+\delta)}$ . Then

$$|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*|_2 = O_{\mathbb{P}} \left( \sqrt{d} \left( \frac{\log n}{n} \right)^{\frac{\delta}{1+\delta}} \right).$$



- When  $\delta > 1$ , we specify  $\tau_i = C_\tau(i/\log i)^{1/2}$ . Then

$$|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*|_2 = O_{\mathbb{P}} \left( \sqrt{d} \left( \frac{\log n}{n} \right)^{1/2} \right).$$

Corollary 3 illustrates a sharp phase transition between the regimes of infinite variance  $\delta \in (0, 1]$  and finite variance  $\delta > 1$ , and such transition is smooth and optimal. When  $\delta \in (0, 1]$ , the optimal choice of  $\tau_i$  depends on  $\delta$ . In practice, when  $\delta$  is unknown, a classical inference method on the tail behavior, such as the Hill estimator (Hill, 1975), can be employed to estimate  $\delta$ . This estimation can be conducted during an initial pilot stage of data collection using a constant thresholding parameter  $\tau_0$ .

The rate of convergence established in Corollary 3 matches the offline oracle with independent samples in the estimation of linear regression model (Sun et al., 2020), ignoring the logarithm term. A similar corollary can be established for this phase transition of the convergence rate when the contamination rate  $\alpha_n > 0$ .

### 3.2 Asymptotic Normality and Bahadur Representation

In the following theorem, we give the Bahadur representation for the proposed estimator  $\hat{\boldsymbol{\theta}}_n$ . The Bahadur representation provides a more refined rate for the estimation error. Moreover, the asymptotic normality can be established by applying the central limit theorem to the main term in the representation.

**Theorem 4.** *Suppose the conditions in Theorem 1 hold, and (C4) holds with  $\delta \geq 5$ . Let  $n_0$  tend to infinity. We have for any nonzero  $\mathbf{v} \in \mathbb{R}^d$  with  $|\mathbf{v}|_2 \leq 1$ ,*

$$\begin{aligned} \mathbf{v}^\top (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) &= \mathbf{v}^\top \mathbf{H}^{-1} \frac{1}{n} \sum_{i \notin \mathcal{Q}_n} \mathbf{X}_i (\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i) \\ &\quad + O_{\mathbb{P}} \left( \sqrt{d} \tau_n \alpha_n + d e_{n-1}^2 \log^2 n + \sqrt{d} \tau_n^{-2} + \frac{\sqrt{d}}{(n \tau_n^2)^{2/5}} \right). \end{aligned} \tag{15}$$

Here  $e_n$  is defined in (14), and  $\mathcal{Q}_n$  denotes the index set of the outliers. Moreover, if the contami-

nation rate  $\alpha_n = o(1/(\sqrt{n}\tau_n))$  and  $n^{1/4} = o(\tau_n)$ , then

$$\frac{\sqrt{n}}{\sigma_{\mathbf{v}}} \mathbf{v}^\top (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}(0, 1), \text{ where } \sigma_{\mathbf{v}}^2 = \mathbf{v}^\top \mathbf{H}^{-1} \boldsymbol{\Sigma} (\mathbf{H}^\top)^{-1} \mathbf{v}, \quad (16)$$

as  $n \rightarrow \infty$ . Here

$$\boldsymbol{\Sigma} = \sum_{k=-\infty}^{\infty} \mathbb{E}[\mathbf{X}_0 \mathbf{X}_k^\top (\mathbf{Z}_0^\top \boldsymbol{\theta}^* - b_0)(\mathbf{Z}_k^\top \boldsymbol{\theta}^* - b_k)]. \quad (17)$$

Theorem 4 provides a Bahadur-type representation for the proposed estimator  $\hat{\boldsymbol{\theta}}_n$ , decomposing it into a leading term representing an asymptotic linear expansion in (15) and a higher-order remainder term. This decomposition is instrumental in establishing asymptotic distribution in (16) and thereby enabling the construction of valid confidence intervals in an online setting, as demonstrated in Section 4. Furthermore, in our framework, it facilitates the comparison of second-order convergence rates of the proposed method with those of first-order methods, as we will discuss in Remark 1 below, and may serve as a foundation for Berry–Esseen-type bounds on the convergence rate of the distribution, which we leave as a promising direction for further investigation. Additionally, this representation explicitly characterizes the orders of both the leading and remainder terms in relation to the model’s dimensionality. While the Bahadur representation was originally introduced in the context of quantile regression, we adopt its use following the recent literature on Huber-type loss problems (Sun et al., 2020), motivated by its role in robustness in reinforcement learning.

To achieve asymptotic normality, additional conditions  $\alpha_n = o(1/(\sqrt{n}\tau_n))$  and  $n^{1/4} = o(\tau_n)$  are required on the specification of the thresholding parameters  $\tau_n$ . These conditions easily hold when the practitioner specifies  $\tau_n = C_\tau n^\beta$  with  $\beta > 1/4$  and the number of outliers satisfies  $m_n = o(n^{1/2-\beta})$ . We further note that to establish the Bahadur representation in the above theorem, we require the sixth moment to exist (i.e.,  $\delta = 5$ ). This condition may be weakened and we leave this theoretical question to future investigation. It is worth noting that even for the case where there is no contamination, i.e.,  $\alpha_n = 0$ , our result is still new. To our knowledge, there is no literature that establishes the Bahadur representation for the TD method. To highlight the rate of convergence concisely in the remainder term, we provide the following corollary under  $\alpha_n = 0$ .

**Proposition 5.** *Suppose the conditions of Theorem 4 hold with  $\alpha_n = 0$  and  $\tau_i = C_\tau i^\beta$  for  $\beta \geq 3/4$  and some  $C_\tau > 0$ , we have for any nonzero  $\mathbf{v} \in \mathbb{R}^d$  with  $|\mathbf{v}|_2 \leq 1$ ,*

$$\mathbf{v}^\top (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) = \mathbf{v}^\top \mathbf{H}^{-1} \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i (\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i) + O_{\mathbb{P}}\left(\frac{d \log n}{n}\right).$$

In this proposition, we specify  $\beta \geq 3/4$  to ensure that the remainder term possesses an order of  $O_{\mathbb{P}}(d \log n/n)$ . Notably, this result is not a direct corollary of Theorems 1 and 4, since the rate of  $e_{n-1}$  in (14) of Theorem 1 is not sharp enough to guarantee the term  $d e_{n-1}^2 \log^2 n$  in (15) converges as fast as  $O(d \log n/n)$  in Proposition 5. To reach this fast rate of the remainder, we establish an improved rate of Theorem 1 under the stronger assumptions, achieved by eliminating the fourth term of (14). The detailed formal result of the improved rate of  $e_n$  is relegated to Proposition 8 in the supplementary material.

**Remark 1** (Acceleration of Convergence). *We discuss the order of the remainder term in the Bahadur representation and further demonstrate that our proposed method converges strictly faster to the asymptotic distribution than that of a prototypical first-order stochastic method such as TD learning. Existing analysis on the TD learning (Ramprasad et al., 2023) does not provide a Bahadur representation. Therefore, we compare it with its i.i.d. version instead. For estimators obtained with i.i.d. sample using SGD, the proof of Polyak and Juditsky (1992) indicates that the remainder term of SGD with learning  $cn^{-\eta}$ ,  $0 < \eta \leq 1$  is at the order of  $O_{\mathbb{P}}(1/n^{1-\eta/2} + 1/n^\eta)$ , which is upper bounded by  $O_{\mathbb{P}}(n^{-\frac{2}{3}})$ , when the dimension  $d$  is fixed. Our proposed method achieves a strictly faster rate of remainder,  $O(n^{-1} \log n)$ .*

## 4 Estimation of Long-Run Covariance Matrix and Online Statistical Inference

In the above section, we provide an online Newton-type algorithm to estimate the parameter  $\boldsymbol{\theta}^*$ . As we can see in Theorem 4, the proposed estimator has an asymptotic variance with a sandwich structure  $\mathbf{H}^{-1} \boldsymbol{\Sigma} (\mathbf{H}^\top)^{-1}$ . To conduct statistical inference simultaneously with ROPE method we proposed in Section 3, we propose an online plug-in estimator for this sandwich structure. Note

that the online estimator  $\widehat{\mathbf{H}}_n^{-1}$  of  $\mathbf{H}^{-1}$  is proposed and utilized in (11) of the estimation algorithm. It remains to construct an online estimator for  $\mathbf{\Sigma}$  in (17).

Developing an online estimator of  $\mathbf{\Sigma}$  with dependent samples is intricate compared to the case of independent ones. As the samples are dependent, the covariance matrix  $\mathbf{\Sigma}$  is an infinite sum of the series of time-lag covariance matrices. To estimate the above long-run covariance matrix  $\mathbf{\Sigma}$ , we first rewrite the definition of  $\mathbf{\Sigma}$  in (17) into the following form

$$\begin{aligned}\mathbf{\Sigma} = & \mathbb{E}[\mathbf{X}_0 \mathbf{X}_0^\top (\mathbf{Z}_0^\top \boldsymbol{\theta}^* - b_0)^2] + \sum_{k=-\infty}^{-1} \mathbb{E}[\mathbf{X}_0 \mathbf{X}_k^\top (\mathbf{Z}_0^\top \boldsymbol{\theta}^* - b_0)(\mathbf{Z}_k^\top \boldsymbol{\theta}^* - b_k)] \\ & + \sum_{k=1}^{\infty} \mathbb{E}[\mathbf{X}_0 \mathbf{X}_k^\top (\mathbf{Z}_0^\top \boldsymbol{\theta}^* - b_0)(\mathbf{Z}_k^\top \boldsymbol{\theta}^* - b_k)].\end{aligned}$$

Next, we replace each term  $\mathbb{E}[\mathbf{X}_0 \mathbf{X}_k^\top (\mathbf{Z}_0^\top \boldsymbol{\theta}^* - b_0)(\mathbf{Z}_k^\top \boldsymbol{\theta}^* - b_k)]$  in (17) with its empirical counterpart using the  $n$  samples. Meanwhile, to handle outliers, we replace  $(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i)$  by  $g_{\tau_i}(\mathbf{Z}_i^\top \widehat{\boldsymbol{\theta}}_{i-1} - b_i)$ . In addition, the infinite sum in (17) is not feasible for direct estimation. Nonetheless, Condition (C1) on the mixing rate of  $\{(\mathbf{X}_t, \mathbf{Z}_t)\}$  allows us to approximate it by effectively estimating the first  $\lceil \lambda \log n \rceil$  terms only, where  $\lambda$  is a pre-specified constant. In summary, we can construct the following estimator for the covariance matrix  $\mathbf{\Sigma}$ ,

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_n = & \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^\top (g_{\tau_i}(\mathbf{Z}_i^\top \widehat{\boldsymbol{\theta}}_{i-1} - b_i))^2 \\ & + \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^{\lceil \lambda \log i \rceil \wedge (i-1)} \mathbf{X}_i \mathbf{X}_{i-k}^\top g_{\tau_i}(\mathbf{Z}_i^\top \widehat{\boldsymbol{\theta}}_{i-1} - b_i) g_{\tau_{i-k}}(\mathbf{Z}_{i-k}^\top \widehat{\boldsymbol{\theta}}_{i-k-1} - b_{i-k}) \\ & + \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^{\lceil \lambda \log i \rceil \wedge (i-1)} \mathbf{X}_{i-k} \mathbf{X}_i^\top g_{\tau_{i-k}}(\mathbf{Z}_{i-k}^\top \widehat{\boldsymbol{\theta}}_{i-k-1} - b_{i-k}) g_{\tau_i}(\mathbf{Z}_i^\top \widehat{\boldsymbol{\theta}}_{i-1} - b_i).\end{aligned}\tag{18}$$

To perform a fully-online update for the estimator  $\widehat{\mathbf{\Sigma}}_n$ , we define

$$\mathbf{S}_j = \sum_{i=1}^j \mathbf{X}_i^\top g_{\tau_i}(\mathbf{Z}_i^\top \widehat{\boldsymbol{\theta}}_{i-1} - b_i), \quad j \geq 1,$$

and the above equation (18) can be rewritten as,

$$\begin{aligned}\widehat{\Sigma}_n = & \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^\top g_{\tau_i}^2(\mathbf{Z}_i^\top \widehat{\boldsymbol{\theta}}_{i-1} - b_i) + \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i g_{\tau_i}(\mathbf{Z}_i^\top \widehat{\boldsymbol{\theta}}_{i-1} - b_i)(\mathbf{S}_{i-1} - \mathbf{S}_{i-\lceil \lambda \log i \rceil \wedge (i-1)}) \\ & + \frac{1}{n} \sum_{i=1}^n (\mathbf{S}_{i-1} - \mathbf{S}_{i-\lceil \lambda \log i \rceil \wedge (i-1)})^\top \mathbf{X}_i^\top g_{\tau_i}(\mathbf{Z}_i^\top \widehat{\boldsymbol{\theta}}_{i-1} - b_i).\end{aligned}$$

In practice, at the  $n$ -th step, we keep only the terms  $\{\mathbf{S}_n, \dots, \mathbf{S}_{n-\lceil \lambda \log n \rceil \wedge (n-1)}\}$  in memory, and the covariance matrix can be updated by

$$\begin{aligned}\widehat{\Sigma}_n = & \frac{1}{n} \left[ \mathbf{X}_n \mathbf{X}_n^\top g_{\tau_n}^2(\mathbf{Z}_n^\top \widehat{\boldsymbol{\theta}}_{n-1} - b_n) + \mathbf{X}_n g_{\tau_n}(\mathbf{Z}_n^\top \widehat{\boldsymbol{\theta}}_{n-1} - b_n)(\mathbf{S}_{n-1} - \mathbf{S}_{n-\lceil \lambda \log n \rceil \wedge (n-1)}) \right. \\ & \left. + (\mathbf{S}_{n-1} - \mathbf{S}_{n-\lceil \lambda \log n \rceil \wedge (n-1)})^\top \mathbf{X}_n^\top g_{\tau_n}(\mathbf{Z}_n^\top \widehat{\boldsymbol{\theta}}_{n-1} - b_n) \right] + \frac{n-1}{n} \widehat{\Sigma}_{n-1}.\end{aligned}\quad (19)$$

The proposed estimator  $\widehat{\Sigma}_n$  complements the estimator  $\widehat{\mathbf{H}}_n^{-1}$  in (11) to construct an estimator of the sandwich form  $\mathbf{H}^{-1} \boldsymbol{\Sigma} (\mathbf{H}^\top)^{-1}$  that appears in the asymptotic distribution in Theorem 4.

Specifically, we can construct the confidence interval for the parameter  $\mathbf{v}^\top \boldsymbol{\theta}^*$  using the asymptotic distribution in (16) in the following way. For any unit vector  $\mathbf{v}$ , a confidence interval with nominal level  $(1 - \xi)$  is

$$\left[ \mathbf{v}^\top \widehat{\boldsymbol{\theta}}_n - q_{1-\xi/2} \widehat{\sigma}_{\mathbf{v}}, \mathbf{v}^\top \widehat{\boldsymbol{\theta}}_n + q_{1-\xi/2} \widehat{\sigma}_{\mathbf{v}} \right], \quad (20)$$

where  $\widehat{\sigma}_{\mathbf{v}}^2 = \mathbf{v}^\top \widehat{\mathbf{H}}_n^{-1} \widehat{\Sigma}_n (\widehat{\mathbf{H}}_n^\top)^{-1} \mathbf{v}$  and  $q_{1-\xi/2}$  denotes the  $(1 - \xi/2)$ -th quantile of a standard normal distribution.

**Remark 2.** We provide some insights into the computational complexity of our inference procedure. Both estimators  $\widehat{\mathbf{H}}_n^{-1}$  and  $\widehat{\Sigma}_n$  can be computed in an  $O(d^2)$  per-iteration computation cost, as discussed in (11). In addition, the computation  $\widehat{\sigma}_{\mathbf{v}}^2 = \mathbf{v}^\top \widehat{\mathbf{H}}_n^{-1} \widehat{\Sigma}_n (\widehat{\mathbf{H}}_n^\top)^{-1} \mathbf{v}$  requires three vector-matrix multiplication with the same computation complexity  $O(d^2)$ . On the contrary, the online bootstrap method proposed in Ramprasad et al. (2023) has a per-iteration complexity of at least  $O(Bd)$ , where the resampling size  $B$  in the bootstrap is usually much larger than  $d$  in practice.

The construction in (20) serves also for the uncertainty quantification of the value function  $\phi(s)^\top \boldsymbol{\theta}^*$  if one specifies  $\mathbf{v} = \phi(s)$ . Furthermore, the constructed confidence interval can be utilized

to implement early stopping, reducing costs associated with data collection. One potential approach is to adopt the early stopping rule outlined in [Xia et al. \(2023\)](#), which is based on predefined tolerance probability, stopping checkpoints and target accuracy level. This allows the robust policy evaluation algorithm to terminate efficiently once the desired precision is achieved, optimizing resource usage.

In the following theorem, we provide the theoretical guarantee of the confidence interval construction by showing that our online estimator of the covariance matrix,  $\widehat{\Sigma}_n$ , is consistent.

**Theorem 6.** *Suppose the conditions in Theorem 4 hold. The covariance estimator  $\widehat{\Sigma}_n$  defined in (19) satisfies*

$$\|\widehat{\Sigma}_n - \Sigma\| = O_{\mathbb{P}} \left( \sqrt{d}\tau_n^2\alpha_n + \sqrt{d}\tau_n^{-1} + \tau_n \sqrt{\frac{d \log n}{n}} + \frac{d\tau_n^2 \log^2 n}{n} \right).$$

Theorem 6 provides an upper bound on the estimation error  $\widehat{\Sigma}_n - \Sigma$ . To achieve the consistency on  $\widehat{\Sigma}_n$ , we specify the thresholding parameter  $\tau_n = C_{\tau}n^{\beta}$  for  $1/4 < \beta < 1/2$ , such that

$$\|\widehat{\Sigma}_n - \Sigma\| = O_{\mathbb{P}} \left( \sqrt{d}\tau_n^2\alpha_n + \frac{\sqrt{d \log n}}{n^{1/2-\beta}} \right).$$

In this case, as long as the fraction of outliers  $\alpha_n$  satisfies  $\alpha_n = o(1/(n^{2\beta}\sqrt{d}))$ , we obtain a consistent estimator of the matrix  $\Sigma$ . In other words, our proposed robust algorithm allows  $o(n^{1-2\beta})$  outliers, ignoring the dimension. We summarize the result in the following corollary.

**Corollary 7.** *Suppose the conditions of Theorem 4 hold, and the fraction of outliers satisfies  $\alpha_n = o(1/(n^{2\beta}\sqrt{d}))$ , where we specify  $\tau_i = C_{\tau}i^{\beta}$  and  $1/4 < \beta < 1/2$ . Then given a vector  $\mathbf{v} \in \mathbb{R}^d$  and a pre-specified confidence level  $1 - \xi$ , we have*

$$\lim_{n \rightarrow \infty} \mathbb{P} \left( \mathbf{v}^{\top} \boldsymbol{\theta}^* \in \left[ \mathbf{v}^{\top} \widehat{\boldsymbol{\theta}}_n - q_{1-\xi/2} \widehat{\sigma}_{\mathbf{v}}, \mathbf{v}^{\top} \widehat{\boldsymbol{\theta}}_n + q_{1-\xi/2} \widehat{\sigma}_{\mathbf{v}} \right] \right) = 1 - \xi,$$

where  $\widehat{\sigma}_{\mathbf{v}}^2 = \mathbf{v}^{\top} \widehat{\mathbf{H}}_n^{-1} \widehat{\Sigma}_n (\widehat{\mathbf{H}}_n^{\top})^{-1} \mathbf{v}$  and  $q_{1-\xi/2}$  denotes the  $(1 - \xi/2)$ -th quantile of a standard normal distribution.

## 5 Numerical Experiments

In this section, we assess the performance of our ROPE algorithm in numerical experiments. We construct all confidence intervals with a nominal coverage of 95%, and compare our method with the online bootstrap method with the linear stochastic approximation proposed in [Ramprasad et al. \(2023\)](#), which we refer to as LSA.

### 5.1 Parameter Inference for Infinite-Horizon MDP

In the first experiment, we focus on an infinite-horizon Markov Decision Process (MDP) setting. Specifically, we create an environment with a state space of size 50 and an action space of size 5. The dimension of the features is fixed at 10. The transition probability kernel of the MDP, the state features are randomly generated from  $\mathcal{N}(\mathbf{0}_p, \mathbb{I}_p)$ , and the policy under evaluation is generated by picking an action uniformly from the action space for each state. By employing the Bellman equation (1), we can compute the expected rewards at each state under the policy. To introduce variability, we add noise to the expected rewards, drawn from different distributions such as the standard normal distribution and  $t$  distribution with a degree of freedom of 2.25. The main objective of this experiment is parameter inference. Particularly, we construct a confidence interval for the first coordinate of the true parameter  $\theta^*$  to study the effect of the first feature on the value function.

We set the thresholding level as  $\tau_i = C(i/(\log i)^2)^\beta$  where  $\beta$  is a positive constant. We begin with an investigation of the influence of the parameters  $C$  and  $\beta$  on the coverage probability and width of the confidence interval in our ROPE algorithm. Figure 1 displays the results, revealing that the performance is relatively robust to variations in  $C$  and  $\beta$ , irrespective of the type of noise applied. Notably, although our theoretical guarantees are based on the noises with a 6-th order moment (See Theorem 4), the experiment indicates that our method exhibits a wider range of applicability.

In subsequent experiments, we specify  $C = 0.5$  and  $\beta = 1/3$  for the ROPE algorithm and compare it with the LSA method. Alongside the coverage probability and width of the confidence interval, we also assess the average running time for both methods. The findings are depicted in Figure 2. It is evident that, under a light-tailed noise that admits normal distribution, the

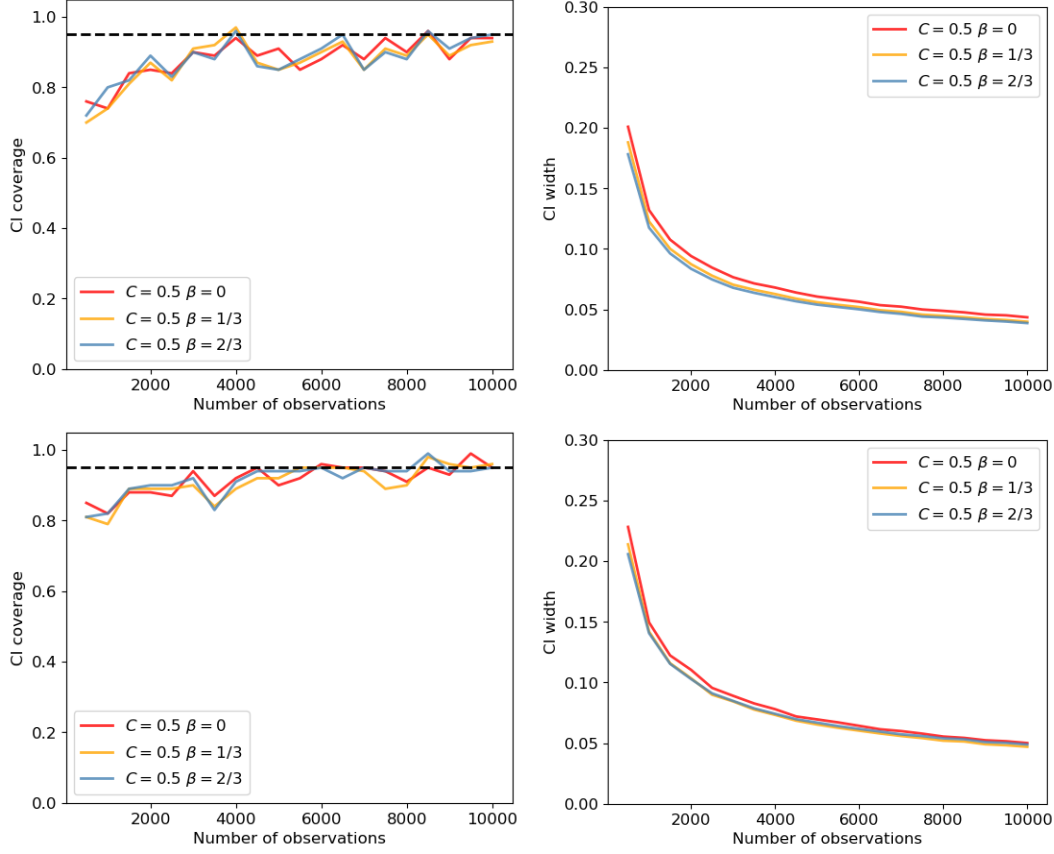


Figure 1: Coverage probability (the first column) and the width of confidence interval (the second column) of ROPE of various  $C$  and  $\beta$ . We specify the noise distribution as standard normal (the first row) and Student's  $t_{2.25}$  (the second row).

two methods yield comparable results, with the ROPE method consistently outperforming LSA in terms of the confidence interval width. When the noise exhibits heavy-tailed characteristics, the confidence interval width of LSA is notably larger than that of ROPE. Additionally, the runtime of ROPE is much shorter than that of LSA.

In the LSA method, the step size is determined by the expression  $\alpha i^{-\eta}$ , which relies on two positive hyperparameters  $\alpha$  and  $\eta$ . In this experiment, we investigate the sensitivity of LSA method on these parameters. For comparison, we also present the result of our ROPE algorithm, which does not require any hyperparameter tuning. In this particular experiment, we specify the noise distribution to be standard normal, and directly apply the online Newton step on the square loss (as opposed to the pseudo-Huber loss). Notably, Figure 3 illustrates that the LSA method is



sensitive to the parameter  $\alpha$ . Specifically, when  $\alpha$  takes on larger values, the LSA method generates significantly wider confidence intervals than that of LSA with well-tuned hyperparameters, which is undesirable in practical applications. Meanwhile, our ROPE method always generates confidence intervals with a valid width and comparable coverage rate.

## 5.2 Value Inference for FrozenLake RL Environment

In this section, we consider the FrozenLake environment provided by OpenAI gym, which involves a character navigating an  $8 \times 8$  grid world. The character’s objective is to start from the first tile of the grid and reach the end tile within each episode. If the character successfully reaches the target tile, a reward of 1 is obtained; otherwise, the reward is 0. In our setup, we generate the state features uniformly from  $[0, 1]^p$ , with a dimensionality of  $p = 4$ . The policy under evaluation is pre-trained using  $Q$ -learning, and the true parameter can be explicitly computed using the transition probability matrix. Under a contaminated reward model, we introduce random perturbations by replacing the true reward with a value uniformly sampled from the range  $[0, 100]$  with probability  $\alpha \in \{0, n^{-1}, 0.05n^{-1/2}\}$ . Our goal is to infer the initial state value.

Following the approach in Section 5.1, we proceed to examine the sensitivity of ROPE to the thresholding parameters  $C$  and  $\beta$ . We present the results in Figure 4, and observe that the performance of ROPE is influenced by both the contamination rate  $\alpha_n$  and the chosen thresholding parameter. Specifically, when  $\alpha_n$  is small, a larger thresholding parameter tends to yield better outcomes. However, as  $\alpha_n$  increases, the use of a large thresholding parameter can negatively impact performance.

Subsequently, we set the values  $C = 0.5, \beta = 1/3$  when  $\alpha_n = 0$  or  $n^{-1}$ , and  $C = 0.1, \beta = 1/3$  when  $\alpha_n = 0.05n^{-1/2}$  for ROPE algorithm, and compare its performance with that of the LSA method. Figure 5 displays the results, which clearly demonstrates that our ROPE method consistently outperforms LSA in terms of coverage rates, CI widths, and running time. The advantage of ROPE becomes more pronounced as the contamination rate  $\alpha_n$  increases.

In the next experiment, we investigate the impact of the initialization value  $n_0$  on the performance of the ROPE algorithm. We consider the corruption rate  $0.05n^{-3/4}$ , while varying the

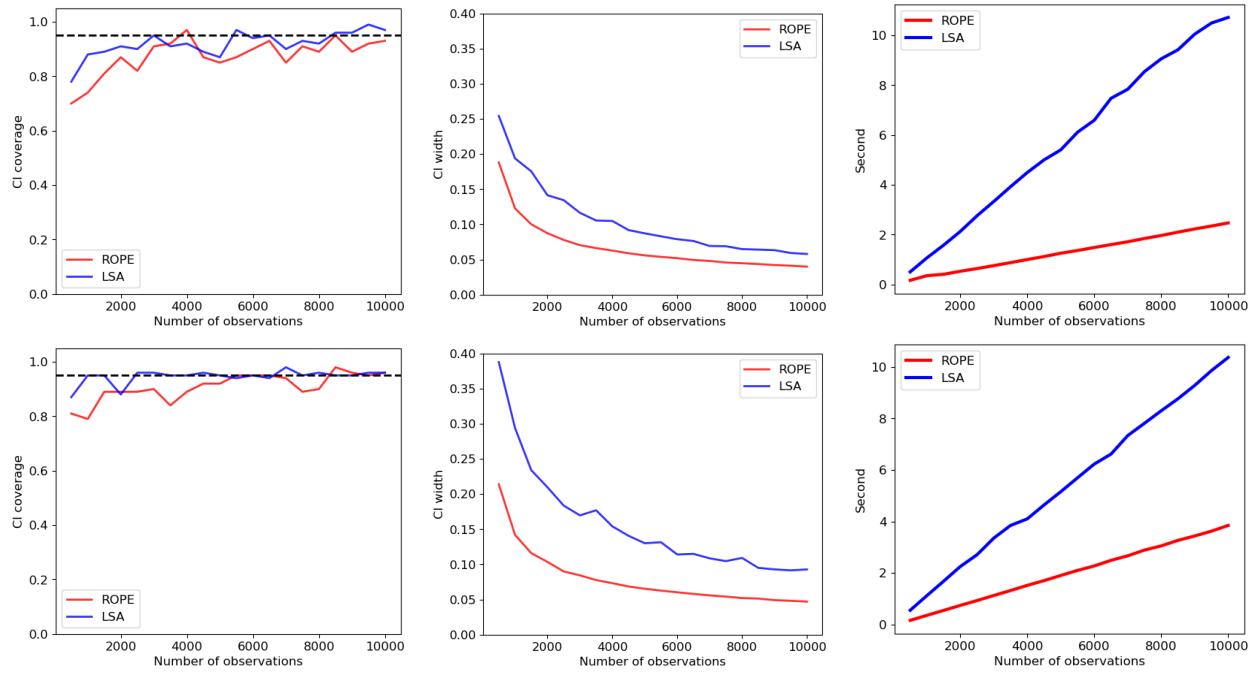


Figure 2: Coverage probability (the first column), the width of confidence interval (the second column), and computing time (the third column) of ROPE and LSA. We specify the noise distribution as standard normal (the first row) and Student's  $t_{2.25}$  (the second row).

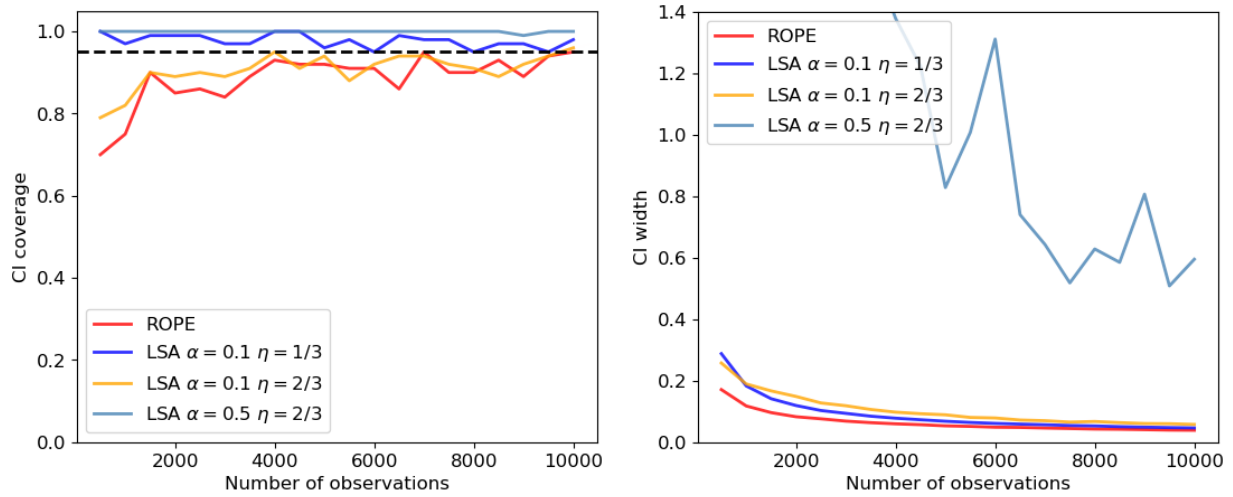


Figure 3: Coverage probability (left), and the width of confidence interval (right) of ROPE and LSA of various  $\alpha$  and  $\eta$ . We specify the noise distribution as standard normal.

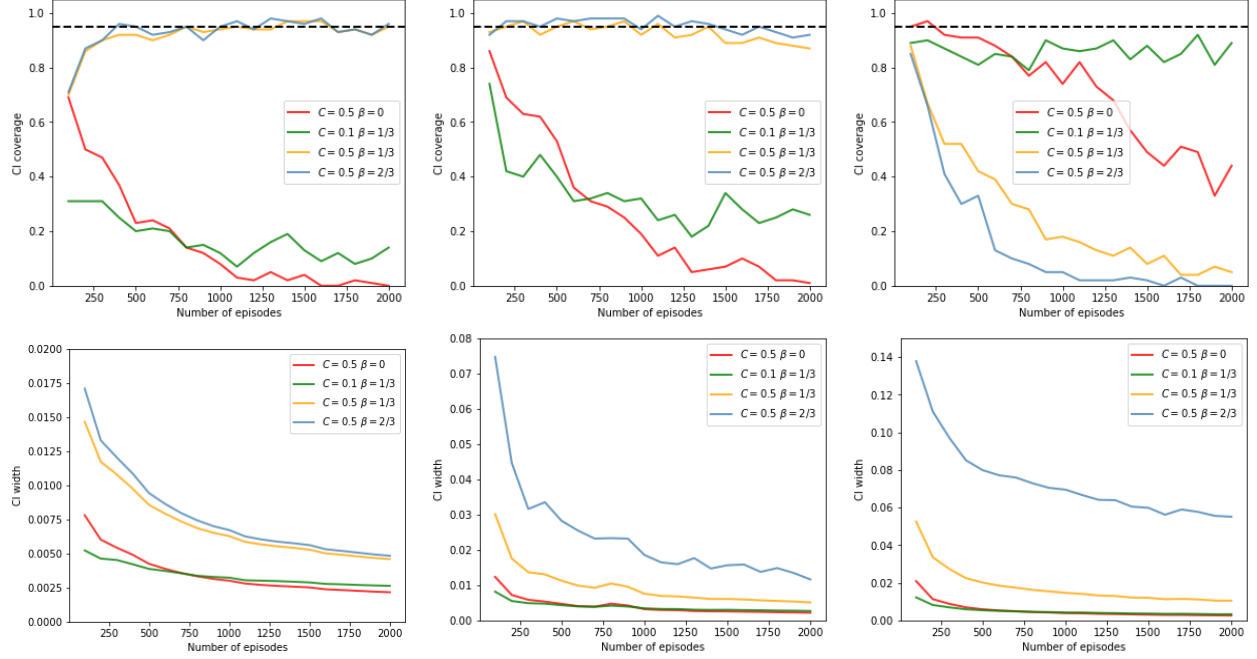


Figure 4: Coverage probability (the first row) and the width of confidence interval (the second row) of ROPE of various  $C$  and  $\beta$ . We set the contamination rate to be 0 (the first column),  $n^{-1}$  (the second column), and  $0.05n^{-1/2}$  (the third column), respectively.

number of episodes in the initialization by  $\{5, 10, 20\}$ . Each episode contains an average of 38 observations, leading to corresponding values of  $n_0$  approximately  $\{190, 380, 760\}$ . The parameters are set to  $C = 0.5$ ,  $\beta = 1/3$ . The results in Figure 6 indicate that the value of  $n_0$  has a minimal effect on  $\ell_2$ -errors and coverage rates.

### 5.3 Online Policy Evaluation on MIMIC-III Dataset

In this section, we demonstrate the effectiveness of the proposed ROPE algorithm by applying it to the Medical Information Mart for Intensive Care version III (MIMIC-III) database (Johnson et al., 2016). MIMIC-III is a freely accessible database containing de-identified critical care data collected from 2001 to 2012 across six ICUs at a Boston teaching hospital. Following Komorowski et al. (2018), we select the data pertaining to adult sepsis patients and extract a set of 43 features to characterize each patient, including demographics, Elixhauser premorbid status, vital signs, and laboratory values. We then apply feature clustering, resulting in 752 distinct states. Each state is assigned a feature vector, computed as the mean of the features within that state. We follow

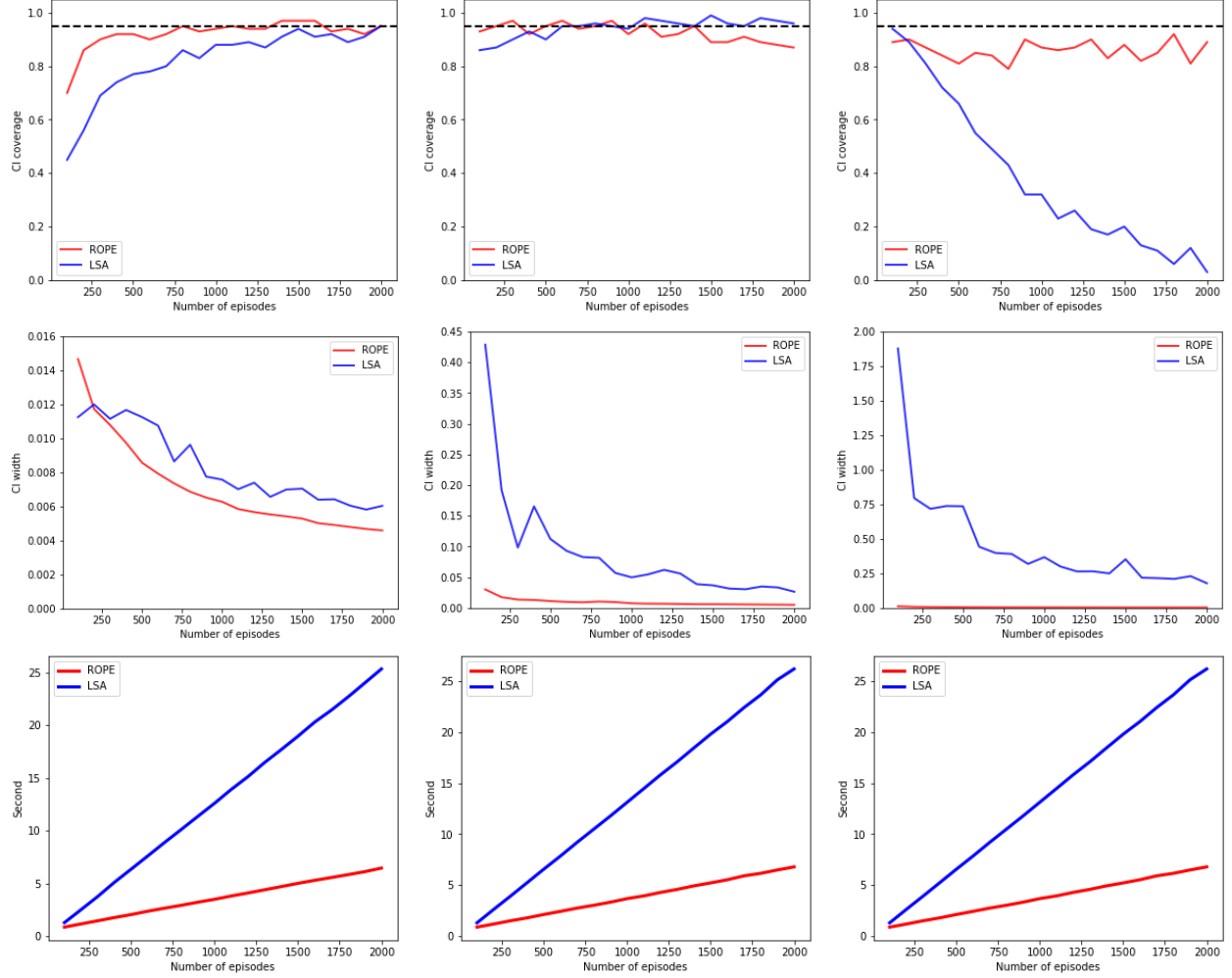


Figure 5: Coverage probability (the first row), the width of confidence interval (the second row), and computing time (the third row) of ROPE and LSA. We set the contamination rate to be 0 (the first column),  $n^{-1}$  (the second column), and  $0.05n^{-1/2}$  (the third column), respectively.

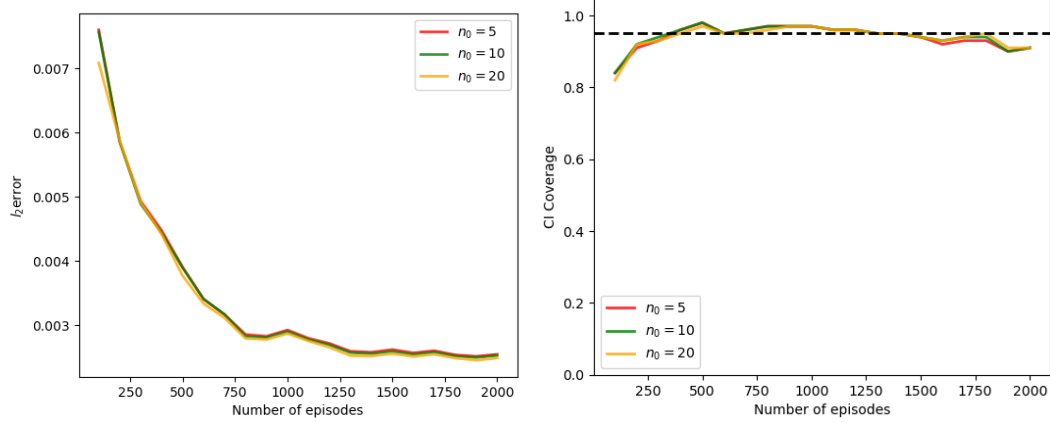


Figure 6: The  $\ell_2$  error (left) and coverage probability (right) of ROPE with different initialization using 5, 10, 20 episodes. We set the contamination rate to be  $0.05n^{-3/4}$ .

the approach of Komorowski et al. (2018) and Prasad et al. (2017) to assign rewards to each state based on patient health metrics and mortality outcomes. The reward at the end of each episode reflects hospital mortality, while the rewards assigned to intermediate observations incorporate physiological parameters such as heart rate, respiratory rate, and arterial pH (Prasad et al., 2017). The terminal reward has a substantially larger magnitude than the intermediate rewards, naturally inducing a heavy-tailed reward distribution. This design is common in healthcare applications, as it emphasizes critical but infrequent outcomes of great clinical importance. Our objective is to perform *on-policy* evaluation and construct confidence intervals for the value function corresponding to the initial state.

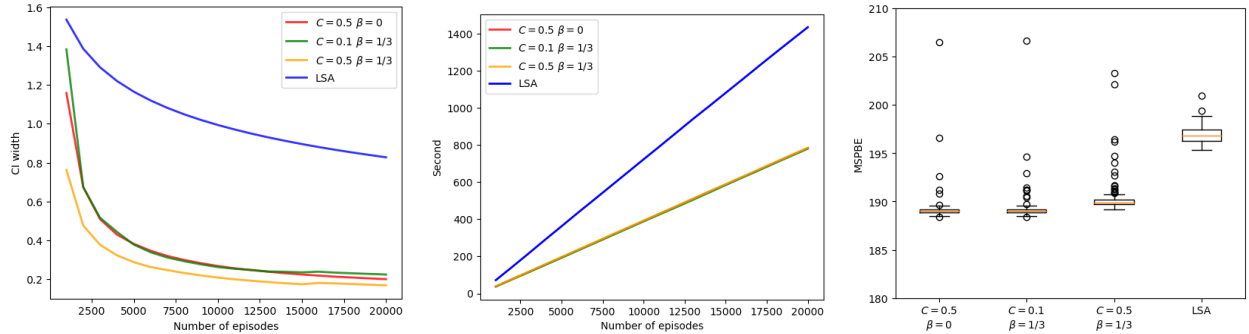


Figure 7: The width of confidence interval (the first column), computational time (the second column), and MSPBE (the third column) of ROPE and LSA for the MIMIC-III dataset.

The final processed data contains 20943 episodes, comprising a total 278598 observations. We

compare the LSA method with our ROPE method using various  $(C, \beta)$  pairs:  $(0.5, 0)$ ,  $(0.1, 1/3)$ , and  $(0.5, 1/3)$ , where ROPE is initialized with 100 episodes. The discount factor is set to  $\gamma = 0.99$ , and the experiment is repeated 100 times. Since the true value is unknown in real-data analysis, we report only the evolution of confidence interval width and computational time as the number of episodes increases. Additionally, we present a box plot of the Mean-Squared Projected Bellman Error (MSPBE) in the final episode. The results, shown in Figure 7, demonstrates that our proposed ROPE method outperforms LSA across all metrics. In particular, we observe that the unrobustified LSA method yields considerably wider confidence intervals than our ROPE method, and the dispersion of MSPBE values is also larger for LSA. These findings underscore the necessity of robust method in the presence of heavy-tailed noise.

## 6 Concluding Remarks

This paper introduces a robust online Newton-type algorithm designed for policy evaluation in reinforcement learning under the influence of heavy-tailed noise and outliers. We demonstrate the estimation consistency as well as the asymptotic normality of our estimator. We further establish the Bahadur representation and show that it converges strictly faster than that of a prototypical first-order method such as TD. To further conduct statistical inference, we propose an online approach for constructing confidence intervals. Our experimental results highlight the efficiency and robustness of our approach when compared to the existing online method.

While our current work focuses on TD learning for on-policy evaluation, our robust framework can be extended to Gradient Temporal-Difference (GTD) algorithms for off-policy evaluation. There is also potential to apply our approach to multi-step TD methods and  $TD(\lambda)$ , though this may introduce challenges due to increased dependencies. Additionally, we identify an extension to fitted-Q evaluation as a promising direction for future research. Beyond policy evaluation, extending our methodology to the reinforcement learning algorithms with policy optimization (e.g., deep Q-learning) presents significant potential but also introduces challenges, which we leave as open questions for future investigation.

## References

- Bennett, G. (1962), “Probability inequalities for the sum of independent random variables,” *Journal of the American Statistical Association*, 57, 33–45.
- Berkes, I. and Philipp, W. (1979), “Approximation theorems for independent and weakly dependent random vectors,” *The Annals of Probability*, 7, 29–54.
- Bhandari, J., Russo, D., and Singal, R. (2018), “A Finite Time Analysis of Temporal Difference Learning With Linear Function Approximation,” in *Conference On Learning Theory*.
- Billingsley, P. (1968), *Convergence of Probability Measures*, Wiley Series in Probability and Mathematical Statistics, Wiley.
- Byrd, R. H., Hansen, S. L., Nocedal, J., and Singer, Y. (2016), “A stochastic quasi-Newton method for large-scale optimization,” *SIAM Journal on Optimization*, 26, 1008–1031.
- Charbonnier, P., Blanc-Feraud, L., Aubert, G., and Barlaud, M. (1994), “Two deterministic half-quadratic regularization algorithms for computed imaging,” in *Proceedings of 1st International Conference on Image Processing*.
- (1997), “Deterministic edge-preserving regularization in computed imaging,” *IEEE Transactions on Image Processing*, 6, 298–311.
- Chen, H., Lu, W., and Song, R. (2021a), “Statistical inference for online decision making: In a contextual bandit setting,” *Journal of the American Statistical Association*, 116, 240–255.
- (2021b), “Statistical inference for online decision making via stochastic gradient descent,” *Journal of the American Statistical Association*, 116, 708–719.
- Chen, X., Lai, Z., Li, H., and Zhang, Y. (2021c), “Online Statistical Inference for Stochastic Optimization via Kiefer-Wolfowitz Methods,” *arXiv preprint arXiv:2102.03389*.
- (2022), “Online statistical inference for contextual bandits via stochastic gradient descent,” *arXiv preprint arXiv:2212.14883*.

- Chen, X., Lee, J. D., Tong, X. T., and Zhang, Y. (2020), “Statistical inference for model parameters in stochastic gradient descent,” *The Annals of Statistics*, 48, 251–273.
- Dai, B., Nachum, O., Chow, Y., Li, L., Szepesvári, C., and Schuurmans, D. (2020), “Coincidence: Off-policy confidence interval estimation,” *Advances in Neural Information Processing Systems*.
- Deshpande, Y., Mackey, L., Syrgkanis, V., and Taddy, M. (2018), “Accurate inference for adaptive linear models,” in *International Conference on Machine Learning*, PMLR, pp. 1194–1203.
- Dimakopoulou, M., Ren, Z., and Zhou, Z. (2021), “Online multi-armed bandits with adaptive inference,” *Advances in Neural Information Processing Systems*, 34, 1939–1951.
- Doukhan, P., Massart, P., and Rio, E. (1994), “The functional central limit theorem for strongly mixing processes,” in *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, vol. 30, pp. 63–82.
- Durmus, A., Moulines, E., Naumov, A., Samsonov, S., and Wai, H.-T. (2021), “On the stability of random matrix product with Markovian noise: Application to linear stochastic approximation and TD learning,” in *Proceedings of Thirty Fourth Conference on Learning Theory*.
- Everitt, T., Krakovna, V., Orseau, L., and Legg, S. (2017), “Reinforcement learning with a corrupted reward channel,” in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*.
- Fan, J., Guo, Y., and Jiang, B. (2022), “Adaptive Huber regression on Markov-dependent data,” *Stochastic Processes and their Applications*, 150, 802–818.
- Fang, Y., Xu, J., and Yang, L. (2018), “Online bootstrap confidence intervals for the stochastic gradient descent estimator,” *Journal of Machine Learning Research*, 19, 1–21.
- Feng, Y., Tang, Z., Zhang, N., and Liu, Q. (2021), “Non-asymptotic confidence intervals of off-policy evaluation: Primal and dual bounds,” in *International Conference on Learning Representations*.
- Freedman, D. A. (1975), “On tail probabilities for martingales,” *The Annals of Probability*, 3, 100–118.



- Givchi, A. and Palhang, M. (2015), “Quasi newton temporal difference learning,” in *Proceedings of the Sixth Asian Conference on Machine Learning*.
- Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S., and Athey, S. (2021), “Confidence intervals for policy evaluation in adaptive experiments,” *Proceedings of the National Academy of Sciences*, 118, e2014602118.
- Han, Q., Sun, W. W., and Zhang, Y. (2022), “Online statistical inference for matrix contextual bandit,” *arXiv preprint arXiv:2212.11385*.
- Hanna, J., Stone, P., and Niekum, S. (2017), “Bootstrapping with models: Confidence intervals for off-policy evaluation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Hao, B., Ji, X., Duan, Y., Lu, H., Szepesvari, C., and Wang, M. (2021), “Bootstrapping fitted q-evaluation for off-policy inference,” in *International Conference on Machine Learning*.
- Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, vol. 2, Springer.
- Hill, B. M. (1975), “A simple general approach to inference about the tail of a distribution,” *Ann. Statist.*, 3, 1163–1174.
- Huber, P. J. (1964), “Robust estimation of a location parameter,” *The Annals of Mathematical Statistics*, 35, 73–101.
- (1992), “Robust estimation of a location parameter,” in *Breakthroughs in Statistics*, Springer.
- (2004), *Robust Statistics*, vol. 523, John Wiley & Sons.
- Jiang, N. and Huang, J. (2020), “Minimax value interval for off-policy evaluation and policy optimization,” *Advances in Neural Information Processing Systems*.
- Johnson, A. E. W., Pollard, T. J., Shen, L., et al. (2016), “MIMIC-III, a freely accessible critical care database,” *Scientific Data*, 3, 160035.

- Kolter, J. Z. and Ng, A. Y. (2009), “Regularization and feature selection in least-squares temporal difference learning,” in *Proceedings of the 26th Annual International Conference on Machine Learning*.
- Komorowski, M., Celi, L. A., Badawi, O., et al. (2018), “The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care,” *Nature Medicine*, 24, 1716–1720.
- Kormushev, P., Calinon, S., and Caldwell, D. G. (2013), “Reinforcement learning in robotics: Applications and real-world challenges,” *Robotics*, 2, 122–148.
- Li, X. and Sun, Q. (2023), “Variance-aware robust reinforcement learning with linear function approximation under heavy-tailed rewards,” *arXiv e-prints*, arXiv:2303.05606.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015), “Human-level control through deep reinforcement learning,” *Nature*, 518, 529–533.
- Mou, W., Pananjady, A., Wainwright, M. J., and Bartlett, P. L. (2021), “Optimal and instance-dependent guarantees for Markovian linear stochastic approximation,” *arXiv e-prints*, arXiv:2112.12770.
- Murphy, S. A. (2003), “Optimal dynamic treatment regimes,” *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 65, 331–355.
- Nesterov, Y. (2003), *Introductory Lectures on Convex Optimization: A Basic Course*, vol. 87, Springer Science & Business Media.
- Polyak, B. T. and Juditsky, A. B. (1992), “Acceleration of stochastic approximation by averaging,” *SIAM Journal on Control and Optimization*, 30, 838–855.
- Prasad, N., Cheng, L.-F., Chivers, C., Draugelis, M., and Engelhardt, B. E. (2017), “A Reinforcement Learning Approach to Weaning of Mechanical Ventilation in Intensive Care Units,” in *33rd Conference on Uncertainty in Artificial Intelligence*.

- Ramprasad, P., Li, Y., Yang, Z., Wang, Z., Sun, W. W., and Cheng, G. (2023), “Online bootstrap inference for policy evaluation in reinforcement learning,” *Journal of the American Statistical Association*, 118, 2901–2914.
- Rio, E. (2017), *Asymptotic Theory of Weakly Dependent Random Processes*, vol. 80, Springer.
- Robbins, H. and Monro, S. (1951), “A stochastic approximation method,” *The Annals of Mathematical Statistics*, 22, 400–407.
- Rosenblatt, M. (1956), “A central limit theorem and a strong mixing condition,” *Proceedings of the National Academy of Sciences*, 42, 43–47.
- Ruppert, D. (1985), “A Newton-Raphson Version of the Multivariate Robbins-Monro Procedure,” *The Annals of Statistics*, 13, 236–245.
- (1988), “Efficient estimations from a slowly convergent Robbins-Monro process,” Tech. rep., Cornell University Operations Research and Industrial Engineering.
- Schraudolph, N. N., Yu, J., and Günter, S. (2007), “A stochastic quasi-Newton method for online convex optimization,” in *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*.
- Shao, Q.-M. and Zhang, Z.-S. (2022), “Berry–Esseen bounds for multivariate nonlinear statistics with applications to M-estimators and stochastic gradient descent algorithms,” *Bernoulli*, 28, 1548–1576.
- Shen, Y., Cai, H., and Song, R. (2024), “Doubly robust interval estimation for optimal policy evaluation in online learning,” *Journal of the American Statistical Association*, 1–20.
- Shi, C., Fan, A., Song, R., and Lu, W. (2018), “High-dimensional A-learning for optimal dynamic treatment regimes,” *The Annals of Statistics*, 46, 925–957.
- Shi, C., Song, R., Lu, W., and Li, R. (2021a), “Statistical inference for high-dimensional models via recursive online-score estimation,” *Journal of the American Statistical Association*, 116, 1307–1318.

- Shi, C., Zhang, S., Lu, W., and Song, R. (2021b), “Statistical inference of the value function for reinforcement learning in infinite-horizon settings,” *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 84, 765–793.
- Sun, Q., Zhou, W.-X., and Fan, J. (2020), “Adaptive Huber Regression,” *Journal of the American Statistical Association*, 115, 254–265.
- Sutton, R. (1988), “Learning to predict by the method of temporal differences,” *Machine Learning*, 3, 9–44.
- Sutton, R. and Barto, A. (2018), *Reinforcement Learning, second edition: An Introduction*, Adaptive Computation and Machine Learning series, MIT Press.
- Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, C., and Wiewiora, E. (2009), “Fast gradient-descent methods for temporal-difference learning with linear function approximation,” in *Proceedings of the 26th Annual International Conference on Machine Learning*.
- Syrsganis, V. and Zhan, R. (2023), “Post-Episodic Reinforcement Learning Inference,” *arXiv e-prints*, arXiv:2302.08854.
- Thomas, P., Theodorou, G., and Ghavamzadeh, M. (2015), “High confidence policy improvement,” in *International Conference on Machine Learning*, PMLR.
- Tsitsiklis, J. and Van Roy, B. (1997), “An analysis of temporal-difference learning with function approximation,” *IEEE Transactions on Automatic Control*, 42, 674–690.
- Vershynin, R. (2010), “Introduction to the non-asymptotic analysis of random matrices,” *arXiv preprint arXiv:1011.3027*.
- Wang, J., Liu, Y., and Li, B. (2020), “Reinforcement learning with perturbed rewards,” in *Proceedings of the AAAI conference on Artificial Intelligence*.
- Xia, E., Khamaru, K., Wainwright, M. J., and Jordan, M. I. (2023), “Instance-Dependent Confidence and Early Stopping for Reinforcement Learning,” *Journal of Machine Learning Research*, 24, 1–43.

- Zhan, R., Hadad, V., Hirshberg, D. A., and Athey, S. (2021), “Off-policy evaluation via adaptive weighting with data from contextual bandits,” in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2125–2135.
- Zhang, K., Janson, L., and Murphy, S. (2021), “Statistical inference with M-estimators on adaptively collected data,” *Advances in Neural Information Processing Systems*, 34, 7460–7471.
- (2022), “Statistical inference after adaptive sampling in non-Markovian environments,” *arXiv preprint arXiv:2202.07098*.
- Zhu, J., Wan, R., Qi, Z., Luo, S., and Shi, C. (2024), “Robust offline reinforcement learning with heavy-tailed rewards,” in *International Conference on Artificial Intelligence and Statistics*.

## 7 Appendix

### 7.1 Experiment of the Effect of Thresholding parameters

In this section, we extend the analysis from the infinite-horizon Markov Decision Process setting discussed in Section 5.1 by examining the effect of thresholding parameter  $\tau$  on our ROPE method. As stated in Theorem 1, the thresholding parameter is defined as  $\tau = C_\tau \max(1, i^{\beta_1}/(\log i)^{\beta_2})$ , and is determined by three factors  $C_\tau, \beta_1$  and  $\beta_2$ . In our experiments, we investigate how variations in each of these factors affect performance. We adopt the ideal parameter triple  $(C_\tau, \beta_1, \beta_2) = (0.5, 1/3, 2/3)$  and add only  $t_{2.25}$  noise to the reward. In each experiment, one parameter is varied while keep the other two remain fixed, and we report the corresponding coverage probability and confidence interval width. The results are presented in Figure 8 below.

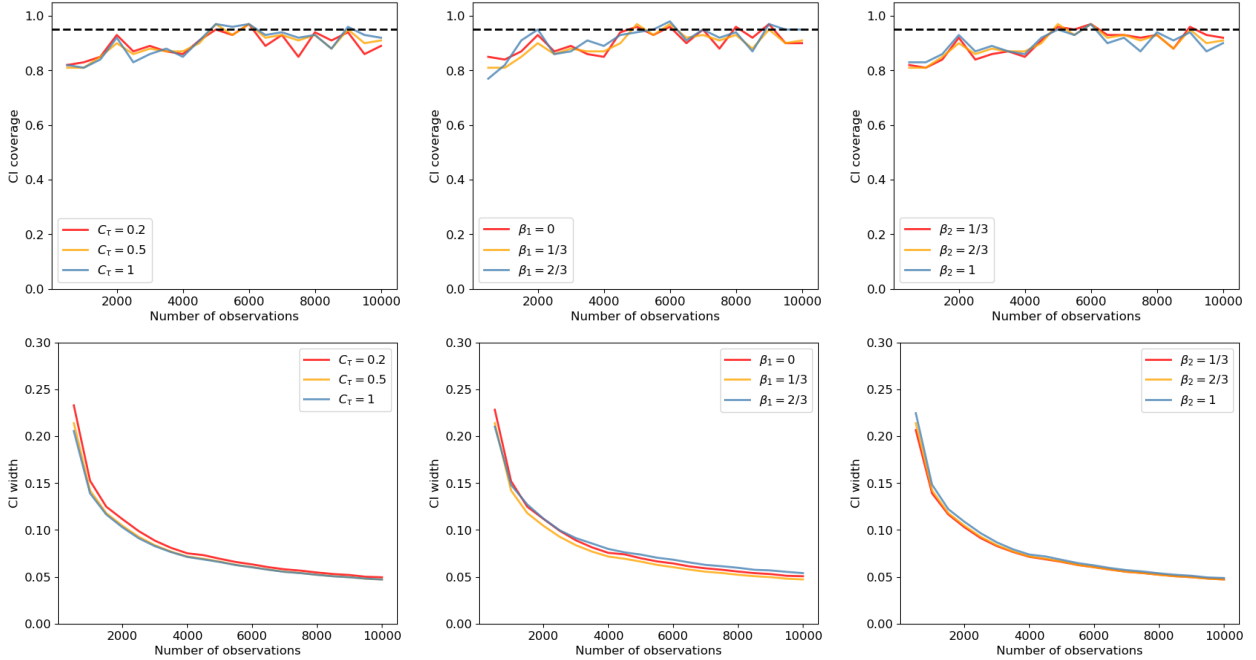


Figure 8: Coverage probability (the first row) and the width of confidence interval (the second row) ROPE under different thresholding parameters. We vary  $C_\tau$  (the first column),  $\beta_1$  (the second column) and  $\beta_2$  (the third column) while other factors are held constant.

From our experimental results, we observe that the factors  $C_\tau, \beta_1$  and  $\beta_2$  have a relatively limited impact on the performance under  $t_{2.25}$  noise. In particular, as illustrated in the first the third columns of Figure 8, a smaller thresholding parameter leads to a narrower confidence interval

width when  $\beta_1$  is held constant. Moreover, the results in the second column indicate that the confidence interval width does not vary monotonically with respect to  $\beta_1$ , with  $\beta_1 = 1/3$  emerging as the optimal choice.

## 7.2 Technical Lemmas

**Lemma 1.** *Let  $Y_1, \dots, Y_n$  be  $\phi$ -mixing sequence with  $\phi(k) = O(\rho^k)$ . Assume that  $|Y_i| \leq M$  and  $\mathbb{E}Y_i = 0$ . Then for any fixed  $\nu > 0$  and  $1 \leq k \leq n$ , we have some constant  $C$  such that*

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^k Y_i}{k}\right| \geq C\sqrt{\frac{\log n}{k}} + C\frac{\log^2 n}{k}\right) = O(n^{-\nu}).$$

*Proof.* We divide the  $k$ -tuple  $(1, \dots, k)$  into  $m_k$  different subsets  $H_1, \dots, H_{m_k}$ , where  $m_k = \lceil k/(\lambda \log n) \rceil$ . Here  $|H_i| = \lceil \lambda \log n \rceil$  for  $1 \leq i \leq m_k - 1$ , and  $|H_{m_k}| \leq \lceil \lambda \log n \rceil$ .  $\lambda$  is a sufficiently large constant which will be specified later. Then we have  $m_k \approx k/(\lambda \log n)$ . Without loss of generality, we assume that  $m_k$  is an even integer.

Let  $\xi_l = \frac{1}{\lceil \lambda \log n \rceil} \sum_{i \in H_{2l-1}} Y_i$ , then we know  $|\xi_l| \leq M$  and  $\mathbb{E}[\xi_l] = 0$  holds. For all  $B_1 \in \sigma(\xi_1, \dots, \xi_l)$  and  $B_2 \in \sigma(\xi_{l+1})$ , there holds  $|\mathbb{P}(B_2|B_1) - \mathbb{P}(B_2)| \leq \phi(\lceil \lambda \log n \rceil)$  for all  $l$ . By Theorem 2 of [Berkes and Philipp \(1979\)](#), there exists a sequence of independent variables  $\eta_l$ ,  $l \geq 1$  with  $\eta_l$  having the same distribution as  $\xi_l$ , and

$$\mathbb{P}\left(|\xi_l - \eta_l| \geq 6\phi(\lceil \lambda \log n \rceil)\right) \leq 6\phi(\lceil \lambda \log n \rceil).$$

For any  $\nu > 0$ , we can take  $\lambda \geq (\nu + 1)/|\log \rho|$  so that

$$\begin{aligned} & \mathbb{P}\left(\left|\frac{2}{m_k} \sum_{l=1}^{m_k/2} (\xi_l - \eta_l)\right| \geq C_1 n^{-\nu}\right) \\ & \leq \mathbb{P}\left(\left|\frac{2}{m_k} \sum_{l=1}^{m_k/2} (\xi_l - \eta_l)\right| \geq 6\phi(\lceil \lambda \log n \rceil)\right) \leq 3m_k \phi(\lceil \lambda \log n \rceil) \leq C_1 n^{-\nu}, \end{aligned} \tag{21}$$

where  $C_1 > 0$  is some constant. Next, we bound the variance of  $\xi_l$ . From equation (20.23) of

Billingsley (1968) we know that for arbitrary  $i, j$ , there is

$$|\mathbb{E}[Y_i Y_j] - \mathbb{E}[Y_i] \mathbb{E}[Y_j]| \leq 2\sqrt{\phi(|i-j|)} \sqrt{\mathbb{E}[Y_i^2] \mathbb{E}[Y_j^2]} \leq 2\rho^{|i-j|/2} M^2.$$

Then we compute that

$$\begin{aligned} \text{Var}(\eta_l) &= \text{Var}(\xi_l) = \frac{1}{\lceil \lambda \log n \rceil^2} \left[ \sum_{i,j \in H_{2l-1}} \{\mathbb{E}(Y_i Y_j) - \mathbb{E}(Y_i) \mathbb{E}(Y_j)\} \right] \\ &\leq \frac{2}{\lceil \lambda \log n \rceil^2} \sum_{i,j=1}^{\lceil \lambda \log n \rceil} \rho^{|i-j|/2} M^2 \\ &\leq \frac{2\lceil \lambda \log n \rceil}{\lceil \lambda \log n \rceil^2} \left( \frac{2}{1-\sqrt{\rho}} - 1 \right) M^2 \leq \frac{4}{(1-\sqrt{\rho})\lceil \lambda \log n \rceil} M^2. \end{aligned} \quad (22)$$

Notice that  $\eta_l$ 's are all uniformly bounded by  $M$ , we can apply Bernstein's inequality (Bennett, 1962) and yield

$$\begin{aligned} \mathbb{P}\left(\left|\frac{2}{m_k} \sum_{l=1}^{m_k/2} \eta_l\right| \geq t\right) &\leq 2 \exp\left(-\frac{m_k t^2/4}{\text{Var}(\eta_l)/2 + Mt/3}\right) \\ &\leq \exp\left(-\frac{m_k \lceil \lambda \log n \rceil t^2/4}{2M^2/(1-\sqrt{\rho}) + \lceil \lambda \log n \rceil Mt/3}\right). \end{aligned}$$

We take  $t = C(\sqrt{\log n/k} + \log^2 n/k)$  for some  $C$  large enough (note that  $k \approx m_k \lceil \lambda \log n \rceil$ ). Then we have that

$$\mathbb{P}\left(\left|\frac{2}{m_k} \sum_{l=1}^{m_k/2} \eta_l\right| \geq t\right) = O(n^{-\nu}). \quad (23)$$

Combining (21) and (23) we have that

$$\mathbb{P}\left(\left|\frac{2}{m_k} \sum_{l=1}^{m_k/2} \xi_l\right| \geq C\sqrt{\frac{\log n}{k}} + C\frac{\log^2 n}{k}\right) = O(n^{-\nu}). \quad (24)$$

Next we denote  $\tilde{\xi} = \frac{1}{\lceil \lambda \log n \rceil} \sum_{i \in H_{2l}} Y_i$ . We can similarly show that

$$\mathbb{P}\left(\left|\frac{2}{m_k} \sum_{l=1}^{m_k/2} \tilde{\xi}_l\right| \geq C\sqrt{\frac{\log n}{k}} + C\frac{\log^2 n}{k}\right) = O(n^{-\nu}). \quad (25)$$



Combining (24) and (25) we prove the desired result.  $\square$

**Lemma 2.** Let  $\mathbf{Y}_1, \dots, \mathbf{Y}_n$  be random vectors in  $\mathbb{R}^p$ . Let  $\mathcal{G}_i = \sigma(\mathbf{Y}_1, \dots, \mathbf{Y}_i)$ . Assume that

$$\begin{aligned} \mathbb{E}[\mathbf{Y}_i | \mathcal{G}_{i-1}] &= \mathbf{0}, \quad \max_{1 \leq i \leq n} \sup_{\mathbf{v} \in \mathbb{S}^p} |\mathbf{v}^\top \mathbf{Y}_i| \leq 1, \\ \sup_{\mathbf{v} \in \mathbb{S}^p} \sum_{i=1}^n \text{Var}[\mathbf{v}^\top \mathbf{Y}_i | \mathcal{G}_{i-1}] &\leq b_n. \end{aligned}$$

Then for every  $\nu > 0$ , there is

$$\mathbb{P}\left(\left|\sum_{i=1}^n \mathbf{Y}_i\right|_2 \geq C(d \log n + \sqrt{db_n \log n})\right) = O(n^{-\nu d}),$$

for some  $C$  sufficiently large.

*Proof.* Let  $\mathfrak{N}$  be the  $1/2$ -net of the unit ball  $\mathbb{S}^{d-1}$ . By Lemma 5.2 of [Vershynin \(2010\)](#) we know that  $|\mathfrak{N}| \leq 5^d$ . Then we have that

$$\begin{aligned} \left|\sum_{i=1}^n \mathbf{Y}_i\right|_2 &= \sup_{\mathbf{v} \in \mathbb{S}^{d-1}} \left|\sum_{i=1}^n \mathbf{v}^\top \mathbf{Y}_i\right| \\ &\leq \sup_{\tilde{\mathbf{v}} \in \mathfrak{N}} \left|\sum_{i=1}^n \tilde{\mathbf{v}}^\top \mathbf{Y}_i\right| + \sup_{|\mathbf{v} - \tilde{\mathbf{v}}|_2 \leq 1/2} \left|\sum_{i=1}^n (\mathbf{v} - \tilde{\mathbf{v}})^\top \mathbf{Y}_i\right|, \\ \Rightarrow \left|\sum_{i=1}^n \mathbf{Y}_i\right|_2 &\leq 2 \sup_{\tilde{\mathbf{v}} \in \mathfrak{N}} \left|\sum_{i=1}^n \tilde{\mathbf{v}}^\top \mathbf{Y}_i\right|. \end{aligned}$$

By Theorem 1.6 of [Freedman \(1975\)](#) we know that

$$\sup_{\tilde{\mathbf{v}} \in \mathbb{S}^{d-1}} \mathbb{P}\left(\left|\sum_{i=1}^n \tilde{\mathbf{v}}^\top \mathbf{Y}_i\right| \geq C_1(d \log n + \sqrt{db_n \log n})\right) = O(n^{-(\nu + \log 5)d})$$

for some large constant  $C' > 0$ . Therefore

$$\begin{aligned}
& \mathbb{P}\left(\left|\sum_{i=1}^n \mathbf{Y}_i\right|_2 \geq 2C_1(d \log n + \sqrt{db_n \log n})\right) \\
& \leq \mathbb{P}\left(\sup_{\tilde{\mathbf{v}} \in \mathfrak{N}} \left|\sum_{i=1}^n \tilde{\mathbf{v}}^\top \mathbf{Y}_i\right| \geq C_1(d \log n + \sqrt{db_n \log n})\right) \\
& \leq 5^d \sup_{\tilde{\mathbf{v}} \in \mathfrak{N}} \mathbb{P}\left(\left|\sum_{i=1}^n \tilde{\mathbf{v}}^\top \mathbf{Y}_i\right| \geq C_1(d \log n + \sqrt{db_n \log n})\right) = O(n^{-\nu d}),
\end{aligned}$$

which proves the lemma.  $\square$

**Lemma 3.** *Let  $\mathbf{Y}$  be a  $d \times d$  random matrix with  $\mathbb{E}[\mathbf{Y}] = \mathbf{0}$ ,  $\|\mathbf{Y}\| \leq 1$ ,  $\mathcal{F}$  be the  $\sigma$ -field generated by  $\mathbf{Y}$ . Then for any  $\sigma$ -field  $\mathcal{G}$ , there holds*

$$\mathbb{E}\left[\|\mathbb{E}[\mathbf{Y}|\mathcal{G}]\|\right] \leq d\sqrt{2\pi\phi},$$

where

$$\phi = \sup_{A \in \mathcal{F}, B \in \mathcal{G}} |\mathbb{P}(AB) - \mathbb{P}(A)\mathbb{P}(B)|.$$

*Proof.* By elementary inequality of matrix, we know that

$$|\mathbf{Y}|_\infty \leq \|\mathbf{Y}\| \leq 1, \quad \text{and} \quad \|\mathbf{Y}\| \leq \|\mathbf{Y}\|_F.$$

For every element  $Y_{i,j}$  of  $\mathbf{Y}$ , by Lemma 4.4.1 of [Berkes and Philipp \(1979\)](#) we have that

$$\mathbb{E}\left[|\mathbb{E}[Y_{ij}|\mathcal{G}]|^2\right] \leq \mathbb{E}\left[|\mathbb{E}[Y_{ij}|\mathcal{G}]|\right] \leq 2\pi\phi.$$

Therefore, we have that

$$\begin{aligned}
& \mathbb{E}\left[\|\mathbb{E}[\mathbf{Y}|\mathcal{G}]\|\right] \leq \mathbb{E}\left[\|\mathbb{E}[\mathbf{Y}|\mathcal{G}]\|_F\right], \\
& \leq \left\{\mathbb{E}\left[\|\mathbb{E}[\mathbf{Y}|\mathcal{G}]\|_F^2\right]\right\}^{1/2} \\
& \leq \left\{\mathbb{E}\left[\sum_{i,j=1}^d |\mathbb{E}[Y_{ij}|\mathcal{G}]|^2\right]\right\}^{1/2} \leq \sqrt{d^2 2\pi\phi} = d\sqrt{2\pi\phi}.
\end{aligned}$$

The proof is complete.  $\square$

**Lemma 4.** (*Approximation of pseudo-Huber gradient*) The gradient of pseudo-Huber loss  $g_\tau(x)$  satisfies

$$\begin{aligned} |g_\tau(x) - x| &\leq \frac{1}{2}\tau^{-\delta}|x|^{1+\delta}, \quad \text{for } \delta \in (0, 2] \\ |g'_\tau(x) - 1| &\leq \frac{5}{2}\tau^{-1-\delta}|x|^{1+\delta} \quad \text{for } \delta \in (0, 1] \end{aligned} \quad (26)$$

uniformly for  $|x| \leq \tau$ . And

$$|g_\tau(x) - x| \leq |x|, \quad |g'_\tau(x) - 1| \leq 1, \quad (27)$$

holds for all  $x$ .

*Proof.* It is easy to compute that

$$g_\tau(x) = \frac{x}{\sqrt{1 + x^2/\tau^2}}, \quad g'_\tau(x) = \frac{1}{(1 + x^2/\tau^2)^{3/2}}.$$

It is not hard to see that

$$|g_\tau(x) - x| \leq |x|, \quad |g'_\tau(x) - 1| \leq 1.$$

This proves (27). Furthermore, when  $|x| \leq \tau$ , we have

$$\begin{aligned} |g_\tau(x) - x| &= \frac{x^2/\tau^2}{(1 + \sqrt{1 + x^2/\tau^2})\sqrt{1 + x^2/\tau^2}} \cdot |x| \\ &\leq \frac{1}{2}\tau^{-\delta}|x|^{1+\delta} \end{aligned}$$

for all  $0 \leq \delta \leq 2$ . Similarly,

$$\begin{aligned} |g'_\tau(x) - 1| &= \frac{x^2/\tau^2(2 + \sqrt{1 + x^2/\tau^2} + x^2/\tau^2)}{(1 + \sqrt{1 + x^2/\tau^2})(1 + x^2/\tau^2)^{3/2}} \\ &\leq \frac{5}{2}\tau^{-1-\delta}|x|^{1+\delta}, \end{aligned}$$

for  $|x| \leq \tau$ . The proof is complete.  $\square$

**Lemma 5.** For any sequence  $\{a_k\}$  where  $a_k = (\log k)^{\beta_1}/k^{\beta_2}$  and  $\beta_1 \geq 0$ , there hold

$$n_0 a_{n_0} + \sum_{k=n_0+1}^n a_k \leq \begin{cases} C n a_n \log n & \text{if } \beta_2 \leq 1; \\ C (\log n)^{\beta_1} / n_0^{\beta_2-1} & \text{if } \beta_2 > 1. \end{cases}$$

Here the constant  $C$  only depends on the parameter  $\beta_2$ .

*Proof.* Directly compute that

$$n_0 a_{n_0} + \sum_{k=n_0+1}^n a_k \leq (\log n)^{\beta_1} \left( \frac{n_0}{n_0^{\beta_2}} + \sum_{k=n_0+1}^n \frac{1}{k^{\beta_2}} \right). \quad (28)$$

We know that

$$\sum_{k=n_0+1}^n \frac{1}{k^{\beta_2}} \leq \int_{n_0}^n \frac{1}{x^{\beta_2}} dx = \frac{1}{1-\beta_2} (n^{1-\beta_2} - n_0^{1-\beta_2}).$$

When  $\beta_2 < 1$ , (28) can be bounded by

$$\begin{aligned} n_0 a_{n_0} + \sum_{k=n_0+1}^n a_k &\leq (\log n)^{\beta_1} \left( n_0^{1-\beta_2} + \frac{1}{1-\beta_2} (n^{1-\beta_2} - n_0^{1-\beta_2}) \right) \\ &\leq (\log n)^{\beta_1} n^{1-\beta_2} \left( -\frac{\beta_2}{1-\beta_2} \left( \frac{n_0}{n} \right)^{1-\beta_2} + \frac{1}{1-\beta_2} \right) \\ &\leq \frac{1}{1-\beta_2} n a_n. \end{aligned}$$

When  $\beta_2 > 1$ , (28) can be bounded by

$$\begin{aligned} n_0 a_{n_0} + \sum_{k=n_0+1}^n a_k &\leq (\log n)^{\beta_1} \left( \frac{\beta_2}{\beta_2-1} \times \frac{1}{n_0^{\beta_2-1}} - \frac{1}{\beta_2-1} \times \frac{1}{n^{\beta_2-1}} \right) \\ &\leq \frac{\beta_2}{\beta_2-1} (\log n)^{\beta_1} \frac{1}{n_0^{\beta_2-1}}. \end{aligned}$$

When  $\beta_2 = 1$ , (28) can be bounded by

$$\begin{aligned} n_0 a_{n_0} + \sum_{k=n_0+1}^n a_k &\leq (\log n)^{\beta_1} (1 + \log n) \\ &\leq 2 n a_n \log n. \end{aligned}$$

□

### 7.3 Proof of Results in Section 2

*Proof of Theorem 1.* For each  $k \geq n_0$ , when  $\delta > 0$ , we denote

$$e_k = \alpha_k \tau_k + \tau_k^{-\min(\delta, 2)} + \sqrt{\frac{\tau_k^{(1-\delta)+} \log k}{k}} + \frac{\log^2 k \tau_k}{k} + \frac{1}{\sqrt{d}} (c_0)^{2^{k-n_0}}, \quad (29)$$

where  $\delta' = \min\{2, \delta\}$ . Given a sufficiently large constant  $\Psi > 0$ , we define the event

$$\begin{aligned} E_k = & \left\{ |\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*|_2 \leq \Psi \sqrt{d} e_k \right\} \cap \left\{ \|\widehat{\mathbf{H}}_k - \mathbf{H}\| \leq C_0 \Psi \sqrt{d} e_k \log k \right\} \\ & \cap \left\{ \left| \frac{1}{k} \sum_{i=1}^k \mathbf{X}_i g_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i) \right|_2 \leq C_0 \sqrt{d} e_k \right\}, \end{aligned} \quad (30)$$

where the  $C_0$  is a constant that does not depend on  $\Psi$ . We will show that  $\mathbb{P}(E_k^c, \cap_{i=n_0}^{k-1} E_i) \leq C_0 k^{-\nu}$ , where  $\nu > 1$  is some constant and  $C_0 > 0$  is some constant only depends on  $\nu$ . Let us prove it by induction on  $k$ . By the choice of initial parameter  $\widehat{\boldsymbol{\theta}}_0$ , we know (30) holds for  $k = n_0$ . For  $n = k+1$ , from (8) we have that

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{k+1} - \boldsymbol{\theta}^* = & \frac{1}{k+1} \sum_{i=1}^{k+1} (\widehat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}^*) - \widehat{\mathbf{H}}_{k+1}^{-1} \frac{1}{k+1} \sum_{i=1}^{k+1} \{ \mathbf{X}_i g_{\tau_i}(\mathbf{Z}_i^\top \widehat{\boldsymbol{\theta}}_{i-1} - b_i) - \mathbf{X}_i g_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i) \} \\ & - \widehat{\mathbf{H}}_{k+1}^{-1} \frac{1}{k+1} \sum_{i=1}^{k+1} \{ \mathbf{X}_i g_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i) \}. \end{aligned} \quad (31)$$

For the second term on the right-hand side of (31), we have that

$$\begin{aligned} & \frac{1}{k+1} \sum_{i=1}^{k+1} \{ \mathbf{X}_i g_{\tau_i}(\mathbf{Z}_i^\top \widehat{\boldsymbol{\theta}}_{i-1} - b_i) - \mathbf{X}_i g_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i) \} \\ = & \frac{1}{k+1} \sum_{i=1}^{k+1} \mathbf{X}_i \mathbf{Z}_i^\top g'_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i) (\widehat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}^*) \\ & + \frac{1}{k+1} \sum_{i=1}^{k+1} \int_0^1 (1-t) \mathbf{X}_i g''_{\tau_i} \{ \mathbf{Z}_i^\top (\boldsymbol{\theta}^* + t(\widehat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}^*)) - b_i \} (\mathbf{Z}_i^\top (\widehat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}^*))^2 dt \\ = & \frac{1}{k+1} \sum_{i=1}^{k+1} \mathbf{X}_i \mathbf{Z}_i^\top g'_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i) (\widehat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}^*) + \frac{1}{k+1} \sum_{i=1}^{k+1} O(1) |\widehat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}^*|_2^2 \end{aligned}$$

$$= \frac{1}{k+1} \sum_{i=1}^{k+1} \mathbf{X}_i \mathbf{Z}_i^\top g'_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i) (\hat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}^*) + O(1) \frac{1}{k+1} \sum_{i=1}^{k+1} |\hat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}^*|_2^2,$$

where the second equality holds because of the fact that  $|g''_{\tau_i}(x)| \leq 3/\tau_i = O(1)$ , the last equality uses the inductive hypothesis. Substitute it into (31), we have

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{k+1} - \boldsymbol{\theta}^* &= (\mathbf{I} - \widehat{\mathbf{H}}_{k+1}^{-1} \mathbf{H}) \frac{1}{k+1} \sum_{i=1}^{k+1} (\hat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}^*) - \widehat{\mathbf{H}}_{k+1}^{-1} \frac{1}{k+1} \sum_{i=1}^{k+1} (\mathbf{A}_i - \mathbf{H}) (\hat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}^*) \\ &\quad - \widehat{\mathbf{H}}_{k+1}^{-1} \frac{1}{k+1} \sum_{i=1}^{k+1} \{ \mathbf{X}_i g_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i) \} + O(1) \frac{1}{k+1} \sum_{i=1}^{k+1} |\hat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}^*|_2^2, \end{aligned} \quad (32)$$

where  $\mathbf{A}_i = \mathbf{X}_i \mathbf{Z}_i^\top g'_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i)$  and  $\mathbf{H} = \mathbb{E}[\mathbf{X} \mathbf{Z}^\top]$ .

Under event  $\cap_{i=n_0}^k E_i$ , by Lemma 5 there exists a constant  $C_1 > 0$  such that

$$\begin{aligned} &\mathbb{P}\left(\frac{1}{k+1} \sum_{i=1}^{k+1} |\hat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}^*|_2^2 \geq C_1 \Psi^2 d e_k^2 \log k, \cap_{i=n_0}^k E_i\right) \\ &\geq \mathbb{P}\left(\frac{1}{k+1} \sum_{i=1}^{k+1} |\hat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}^*|_2^2 \geq \frac{1}{k+1} \sum_{i=0}^k \Psi^2 d e_i^2, \cap_{i=n_0}^k E_i\right) = 0. \end{aligned} \quad (33)$$

Similarly, by Lemma 5 there exists a constant  $C_2 > 0$  such that

$$\begin{aligned} &\mathbb{P}\left(\frac{1}{k+1} \sum_{i=1}^{k+1} |\hat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}^*|_2 \geq C_2 \Psi \sqrt{d} e_k \log k, \cap_{i=n_0}^k E_i\right) \\ &\geq \mathbb{P}\left(\frac{1}{k+1} \sum_{i=1}^{k+1} |\hat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}^*|_2 \geq \frac{1}{k+1} \sum_{i=0}^k \Psi \sqrt{d} e_i, \cap_{i=n_0}^k E_i\right) = 0. \end{aligned} \quad (34)$$

By Lemma 6, (34) and (42) we know that under event  $\cap_{i=n_0}^k E_i$ , for every  $\nu > 0$  there exist constants  $C_1, C_3 > 0$  (which only depend on  $\nu$ ), such that

$$\begin{aligned} &\mathbb{P}\left(\left\|(\mathbf{I} - \widehat{\mathbf{H}}_{k+1}^{-1} \mathbf{H}) \frac{1}{k+1} \sum_{i=0}^k (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}^*)\right\|_2 \geq C_1 \Psi^2 d e_k^2 \log^2 k, \cap_{i=n_0}^k E_i\right) \\ &\leq \mathbb{P}\left(\|\widehat{\mathbf{H}}_{k+1}^{-1}\| \cdot \|\widehat{\mathbf{H}}_{k+1} - \mathbf{H}\| \cdot \frac{1}{k+1} \sum_{i=0}^k |\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}^*|_2 \geq C_1 \Psi^2 d e_k^2 \log^2 k, \cap_{i=n_0}^k E_i\right) \leq C_3 (k+1)^{-\nu}. \end{aligned}$$

By Lemma 8 we know that

$$\mathbb{P}\left(\left|\frac{1}{k+1}\sum_{i=1}^{k+1}(\mathbf{A}_i - \mathbf{H})(\widehat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}^*)\right|_2 \geq C_1(\alpha_k + \Psi^2 de_k^2 \log k), \cap_{i=n_0}^k E_i\right) \leq C_3(k+1)^{-\nu}. \quad (35)$$

In this part, we mainly consider the case where  $\delta > 0$  and  $\tau_k$  can be arbitrary. A special case where  $\delta > 4$  and  $\sqrt{i/\log^3 i} = O(\tau_i)$  will be presented in Proposition 8 below. By Lemma 7 we have that

$$\mathbb{P}\left(\left|\widehat{\mathbf{H}}_{k+1}^{-1} \frac{1}{k+1} \sum_{i=1}^{k+1} \{\mathbf{X}_i g(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i)\}\right|_2 \geq C_1 \sqrt{de_{k+1}}, \cap_{i=n_0}^k E_i\right) \leq C_3(k+1)^{-\nu};$$

By substituting the above inequalities into (32), we have that

$$\begin{aligned} & \mathbb{P}\left(|\widehat{\boldsymbol{\theta}}_{k+1} - \boldsymbol{\theta}^*|_2 \geq \Psi \sqrt{de_{k+1}}, \cap_{i=n_0}^k E_i\right) \\ & \leq \mathbb{P}\left(|\widehat{\boldsymbol{\theta}}_{k+1} - \boldsymbol{\theta}^*|_2 \geq C_1 \sqrt{de_{k+1}} + C_1 \alpha_k + 3C_1 \Psi^2 de_k^2 \log^2 k, \cap_{i=n_0}^k E_i\right) \\ & \leq 3C_3(k+1)^{-\nu}, \end{aligned} \quad (36)$$

where  $\Psi$  can be taken to be sufficiently large, given  $C_1$  fixed. Here the second inequality holds because  $\sqrt{de_k}$  can be sufficiently small for  $k \geq n_0$  by taking  $n_0$  sufficiently large. Meanwhile, as  $\alpha_k \leq (k+1)\alpha_{k+1}/k \leq (n_0+1)\alpha_{k+1}/n_0 = o(\tau_{k+1}\alpha_{k+1}) = o(e_{k+1})$ , the last two terms both have order of  $o(e_{k+1})$ . Obviously, the rest two events of (30) are contained in (36). Therefore, the inductive hypothesis that  $\mathbb{P}(E_{k+1}^c, \cap_{j=n_0}^k E_j) \leq 3C_3(k+1)^{-\nu}$ , is proved.

For  $k \geq n_0$ , there holds

$$\begin{aligned} & \mathbb{P}\left(\cap_{i=n_0}^k E_i\right) \\ & \geq \mathbb{P}\left(E_{n_0}\right) - \sum_{i=n_0+1}^k \mathbb{P}\left(E_i^c, \cap_{j=n_0}^{i-1} E_j\right) \\ & \geq 1 - \sum_{i=n_0+1}^k 3C_3 i^{-\nu} \geq 1 - \frac{3C_3}{\nu-1} \frac{1}{n_0^{\nu-1}}, \end{aligned}$$

which proves the theorem.  $\square$

**Lemma 6.** (*Bound of the pseudo-Huber Hessian matrix*) Under the same assumptions as in The-

orem 1, there exists uniform constants  $C$  and  $c$ , such that the Hessian matrix  $\widehat{\mathbf{H}}_{k+1}$  satisfies

$$\mathbb{P}\left(\|\widehat{\mathbf{H}}_{k+1} - \mathbf{H}\| \geq C\Psi\sqrt{d}e_k \log k, \cap_{i=n_0}^k E_i\right) \leq c(k+1)^{-\nu d}.$$

*Proof.* We note that

$$\|\widehat{\mathbf{H}}_{k+1} - \mathbf{H}\| \leq \frac{1}{k+1} \left\| \sum_{i \in \mathcal{Q}_k} \mathbf{X}_i \mathbf{Z}_i g'_{\tau_i}(\mathbf{Z}_i^\top \widehat{\boldsymbol{\theta}}_{i-1} - b_i) - \mathbf{H} \right\| + \frac{1}{k+1} \left\| \sum_{i \notin \mathcal{Q}_k} \mathbf{X}_i \mathbf{Z}_i g'_{\tau_i}(\mathbf{Z}_i^\top \widehat{\boldsymbol{\theta}}_{i-1} - b_i) - \mathbf{H} \right\|.$$

For the first term, there is

$$\begin{aligned} & \left\| \mathbf{X} \mathbf{Z}^\top g'_\tau(\mathbf{Z}^\top \boldsymbol{\theta} - b) \right\| \\ &= \sup_{\mathbf{u}, \mathbf{v} \in \mathbb{S}^{d-1}} \mathbf{u}^\top \mathbf{X} \mathbf{Z}^\top \mathbf{v} |g'_\tau(\mathbf{Z}^\top \boldsymbol{\theta} - b)| \leq M^2, \end{aligned}$$

holds for all  $\boldsymbol{\theta}$  and  $\tau \geq 1$ . Therefore we have that

$$\begin{aligned} & \frac{1}{k+1} \left\| \sum_{i \in \mathcal{Q}_k} \mathbf{X}_i \mathbf{Z}_i g'_{\tau_i}(\mathbf{Z}_i^\top \widehat{\boldsymbol{\theta}}_{i-1} - b_i) - \mathbf{H} \right\| \\ & \leq \frac{1}{k+1} \sum_{i \in \mathcal{Q}_k} \|\mathbf{X}_i \mathbf{Z}_i g'_{\tau_i}(\mathbf{Z}_i^\top \widehat{\boldsymbol{\theta}}_{i-1} - b_i)\| + \frac{1}{k+1} \sum_{i \in \mathcal{Q}_i} \|\mathbf{H}\| \leq 2M^2 \alpha_k. \end{aligned} \tag{37}$$

For the second term, we first prove that

$$\mathbb{P}\left(\left\| \frac{1}{k+1} \sum_{i \notin \mathcal{Q}_k} (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]) \right\| \geq C_1 \sqrt{\frac{d \log k}{k+1}}\right) = O((k+1)^{-\nu d}). \tag{38}$$

Let  $\mathfrak{N}$  be the  $1/4$ -net of the unit ball  $\mathbb{S}^{d-1}$ . By Lemma 5.2 of Vershynin (2010) we know that  $|\mathfrak{N}| \leq 9^d$ . Then we have that

$$\begin{aligned} \left\| \frac{1}{k+1} \sum_{i \notin \mathcal{Q}_k} (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]) \right\| &= \sup_{\mathbf{u}, \mathbf{v} \in \mathbb{S}^{d-1}} \left| \frac{1}{k+1} \sum_{i \notin \mathcal{Q}_k} \mathbf{u}^\top (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]) \mathbf{v} \right| \\ &\leq \sup_{\tilde{\mathbf{u}}, \tilde{\mathbf{v}} \in \mathfrak{N}} \left| \frac{1}{k+1} \sum_{i \notin \mathcal{Q}_k} \tilde{\mathbf{u}}^\top (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]) \tilde{\mathbf{v}} \right| \\ &\quad + \sup_{\tilde{\mathbf{u}} \in \mathfrak{N}} \sup_{|\mathbf{v} - \tilde{\mathbf{v}}|_2 \leq 1/4} \left| \frac{1}{k+1} \sum_{i \notin \mathcal{Q}_k} \tilde{\mathbf{u}}^\top (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]) (\mathbf{v} - \tilde{\mathbf{v}}) \right|_2 \end{aligned}$$



$$\begin{aligned}
& + \sup_{\mathbf{v} \in \mathbb{S}^{d-1}} \sup_{|\mathbf{u} - \tilde{\mathbf{u}}|_2 \leq 1/4} \left| \frac{1}{k+1} \sum_{i \notin \mathcal{Q}_k} (\mathbf{u} - \tilde{\mathbf{u}})^\top (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]) \mathbf{v} \right|_2 \\
\Rightarrow \left\| \frac{1}{k+1} \sum_{i \notin \mathcal{Q}_k} (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]) \right\| & \leq 2 \sup_{\tilde{\mathbf{u}}, \tilde{\mathbf{v}} \in \mathfrak{N}} \left| \frac{1}{k+1} \sum_{i \notin \mathcal{Q}_k} \tilde{\mathbf{u}}^\top (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]) \tilde{\mathbf{v}} \right|.
\end{aligned}$$

Applying Lemma 1 to every  $\frac{1}{k+1} \sum_{i \notin \mathcal{Q}_k} \tilde{\mathbf{u}}^\top (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]) \tilde{\mathbf{v}}$  we can obtain that

$$\sup_{\tilde{\mathbf{u}}, \tilde{\mathbf{v}} \in \mathfrak{N}} \mathbb{P} \left( \left| \frac{1}{k+1} \sum_{i \notin \mathcal{Q}_k} \tilde{\mathbf{u}}^\top (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]) \tilde{\mathbf{v}} \right|_2 \geq C_1 \sqrt{\frac{d \log k}{k+1}} \right) = O((k+1)^{-(\nu + \log 9)d})$$

for some constant  $C_1 > 0$ . Therefore

$$\begin{aligned}
& \mathbb{P} \left( \left\| \frac{1}{k+1} \sum_{i \notin \mathcal{Q}_k} (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]) \right\| \geq 2C_1 \sqrt{\frac{d \log k}{k+1}} \right) \\
& \leq \mathbb{P} \left( \sup_{\tilde{\mathbf{u}}, \tilde{\mathbf{v}} \in \mathfrak{N}} \left| \frac{1}{k+1} \sum_{i \notin \mathcal{Q}_k} \tilde{\mathbf{u}}^\top (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]) \tilde{\mathbf{v}} \right|_2 \geq C_1 \sqrt{\frac{d \log k}{k+1}} \right) \\
& \leq 9^{2d} \sup_{\tilde{\mathbf{u}}, \tilde{\mathbf{v}} \in \mathfrak{N}} \mathbb{P} \left( \left| \frac{1}{k+1} \sum_{i \notin \mathcal{Q}_k} \tilde{\mathbf{u}}^\top (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]) \tilde{\mathbf{v}} \right|_2 \geq C_1 \sqrt{\frac{d \log k}{k+1}} \right) = O((k+1)^{-\nu d}),
\end{aligned}$$

which proves (38). Next we prove that when  $(\mathbf{X}, \mathbf{Z}, b)$  are normal data, for every  $\tau > 0$ ,

$$\left\| \mathbb{E}[\mathbf{X} \mathbf{Z}^\top g'_\tau(\mathbf{Z}^\top \boldsymbol{\theta}^* - b)] - \mathbb{E}[\mathbf{X} \mathbf{Z}^\top] \right\| \leq C_2 \tau^{-\delta'}.$$

Indeed, denote  $\epsilon = \mathbf{Z}^\top \boldsymbol{\theta}^* - b$ , we have that

$$\begin{aligned}
& \left\| \mathbb{E}[\mathbf{X} \mathbf{Z}^\top g'_\tau(\mathbf{Z}^\top \boldsymbol{\theta}^* - b)] - \mathbb{E}[\mathbf{X} \mathbf{Z}^\top] \right\| \\
& \leq \sup_{\mathbf{u}, \mathbf{v}} \mathbb{E}[\mathbf{u}^\top \mathbf{X} \mathbf{Z}^\top \mathbf{v} |g'_\tau(\epsilon) - 1|] \\
& \leq \sup_{\mathbf{u}, \mathbf{v}} \mathbb{E}[\mathbf{u}^\top \mathbf{X} \mathbf{Z}^\top \mathbf{v} \mathbb{I}(|\epsilon| > \tau)] + \frac{5}{2} \mathbf{u}^\top \mathbf{X} \mathbf{Z}^\top \mathbf{v} \tau^{-1-\delta''} |\epsilon|^{1+\delta''} \mathbb{I}(|\epsilon| \leq \tau) \\
& \leq \sup_{\mathbf{u}, \mathbf{v}} \mathbb{E}[\mathbf{u}^\top \mathbf{X} \mathbf{Z}^\top \mathbf{v} \mathbb{P}(|\epsilon| > \tau | \mathbf{X}, \mathbf{Z})] + \frac{5}{2} \tau^{-1-\delta''} \sup_{\mathbf{u}, \mathbf{v}} \mathbb{E}[\mathbf{u}^\top \mathbf{X} \mathbf{Z}^\top \mathbf{v} \mathbb{E}[|\epsilon|^{1+\delta''} | \mathbf{X}, \mathbf{Z}]] \leq C_2 \tau^{-1-\delta''} \leq C_2 \tau^{-\delta'},
\end{aligned}$$

where the third line uses Lemma 4, and  $\delta'' = \min(1, \delta)$ . Then clearly we have that  $1 + \delta'' \geq \delta'$ .

Therefore, by the choice of  $\tau_i$  and Lemma 5, we have

$$\left\| \frac{1}{k+1} \sum_{i \notin \mathcal{Q}_k} \mathbb{E}[\mathbf{A}_i] - \mathbf{H} \right\| \leq \frac{1}{k+1} \sum_{i \notin \mathcal{Q}_k} \left\| \mathbb{E}[\mathbf{A}_i] - \mathbf{H} \right\| \leq C_2 \tau_{k+1}^{-\delta'}. \quad (39)$$

Under event  $\cap_{i=n_0}^k E_i$ , from (34) we know there is

$$\begin{aligned} \left\| \frac{1}{k+1} \sum_{i \notin \mathcal{Q}_k} \mathbf{A}_i - \widehat{\mathbf{H}}_k \right\| &= \left\| \frac{1}{k+1} \sum_{i \notin \mathcal{Q}_k} \mathbf{X}_i \mathbf{Z}_i^\top \{g'_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i) - g'_{\tau_i}(\mathbf{Z}_i^\top \widehat{\boldsymbol{\theta}}_{i-1} - b_i)\} \right\| \\ &\leq M^2 \frac{1}{k+1} \sum_{i \notin \mathcal{Q}_k} |\widehat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}^*|_2 \leq C_3 \Psi \sqrt{de_k \log k}, \end{aligned} \quad (40)$$

Therefore combining (38), (39) and (40) we have that

$$\begin{aligned} &\mathbb{P} \left( \frac{1}{k+1} \left\| \sum_{i \notin \mathcal{Q}_k} \mathbf{X}_i \mathbf{Z}_i g'_{\tau_i}(\mathbf{Z}_i^\top \widehat{\boldsymbol{\theta}}_{i-1} - b_i) - \mathbf{H} \right\| \geq C_4 \left( \sqrt{\frac{d \log k}{k+1}} + \tau_{k+1}^{-\delta'} + \Psi \sqrt{de_k \log k} \right) \right) \\ &= O((k+1)^{-\nu d}). \end{aligned} \quad (41)$$

Combining (37) and (41), we have that

$$\mathbb{P} \left( \left\| \widehat{\mathbf{H}}_{k+1} - \mathbf{H} \right\| \geq C_5 \left( \alpha_k + \sqrt{\frac{d \log k}{k+1}} + \tau_{k+1}^{-\delta'} + \sqrt{de_k \log k} \right) \right) = O((k+1)^{-\nu d}),$$

which proves the lemma.

As a corollary, we can show that under event  $\cap_{i=n_0}^k E_i$ ,

$$\left\| \widehat{\mathbf{H}}_{k+1}^{-1} \right\| = (\Lambda_{\min}(\widehat{\mathbf{H}}_{k+1}))^{-1} \leq (\Lambda_{\min}(\mathbf{H}) - \left\| \widehat{\mathbf{H}}_{k+1} - \mathbf{H} \right\|)^{-1} \leq 2/\Lambda_{\min}(\mathbf{H}). \quad (42)$$

□

**Lemma 7.** *Under the same assumptions as in Theorem 1, for every  $\nu > 0$ , there exist constants  $C$  and  $c$  such that:*

i) If  $\delta > 0$  and  $\tau_k$  can be arbitrary, then

$$\mathbb{P}\left(\left|\frac{1}{k+1}\sum_{i=1}^{k+1}\{\mathbf{X}_i g_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i)\}\right|_2 \geq C\sqrt{de_{k+1}}\right) \leq c(k+1)^{-\nu}.$$

Here  $e_{k+1}$  is defined in (29).

ii) If  $\delta > 1$  and  $\sqrt{k/\log^3 k} = O(\tau_k)$ , then

$$\mathbb{P}\left(\left|\frac{1}{k+1}\sum_{i=1}^{k+1}\{\mathbf{X}_i g_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i)\}\right|_2 \geq C\sqrt{de_{k+1}}\right) \leq c(k+1)^{-\min\{\nu, (\delta-1)/3\}}.$$

Here  $e_{k+1}$  is defined in (65).

*Proof. Proof of i).* We prove the bound coordinate-wisely. When  $(\mathbf{X}, \mathbf{Z}, b)$  has no outlier, for the  $l$ -th coordinate, we first prove that for every  $\tau > 0$ , there is

$$|\mathbb{E}[X_l g_\tau(\mathbf{Z}^\top \boldsymbol{\theta}^* - b)]| \leq C_1 \tau^{-\delta'}.$$

Indeed,  $\epsilon = \mathbf{Z}^\top \boldsymbol{\theta}^* - b$ , we have that

$$\begin{aligned} |\mathbb{E}[X_l g_\tau(\epsilon)]| &\leq |\mathbb{E}[X_l \epsilon]| + |\mathbb{E}[X_l |\epsilon - g_\tau(\epsilon)| \mathbb{I}(|\epsilon| \leq \tau)]| + |\mathbb{E}[X_l |g_\tau(\epsilon) - \epsilon| \mathbb{I}(|\epsilon| > \tau)]| \\ &\leq \frac{1}{2} \tau^{-\delta'} |\mathbb{E}[X_l |\epsilon|^{1+\delta'} \mathbb{I}(|\epsilon| \leq \tau)]| + |\mathbb{E}[X_l |\epsilon| \mathbb{I}(|\epsilon| > \tau)]| \\ &\leq \frac{1}{2} \tau^{-\delta'} |\mathbb{E}[X_l \mathbb{E}[|\epsilon|^{1+\delta'} \mathbb{I}(|\epsilon| \leq \tau) | \mathbf{X}, \mathbf{Z}]]| + |\mathbb{E}[X_l \mathbb{E}[|\epsilon| \mathbb{I}(|\epsilon| > \tau) | \mathbf{X}, \mathbf{Z}]]| \\ &\leq C_1 \tau^{-\delta'}. \end{aligned}$$

Therefore, by the choice of  $\tau_i$  and Lemma 5, we have

$$\left|\frac{1}{k+1}\sum_{i=1}^{k+1}\mathbb{E}[X_{i,l} g_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i)]\right| \leq \frac{1}{k+1}\sum_{i=1}^{k+1}|\mathbb{E}[X_{i,l} g_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i)]| \leq C_2 \tau_{k+1}^{-\delta'}. \quad (43)$$

Next, when all data are normal, we prove the rate of

$$\left| \frac{1}{k+1} \sum_{i=1}^{k+1} \{X_{i,l}g_{\tau_i}(\epsilon_i) - \mathbb{E}[X_{i,l}g_{\tau_i}(\epsilon_i)]\} \right|.$$

We basically rehash the proof in Lemma 1.

**Case 1:**  $\delta \in (0, 1]$ . We divide the  $k$ -tuple  $(1, \dots, k)$  into  $m_k$  different subsets  $H_1, \dots, H_{m_k}$ , where  $m_k = \lceil k/(\lambda \log k) \rceil$ . Here  $|H_i| = \lceil \lambda \log k \rceil$  for  $1 \leq i \leq m_k - 1$ , and  $|H_{m_k}| \leq \lceil \lambda \log k \rceil$ . Then we have  $m_k \approx k/(\lambda \log k)$ . Without loss of generality, we assume that  $m_k$  is an even integer.

Let  $Y_i = X_{i,l}g_{\tau_i}(\epsilon_i) - \mathbb{E}[X_{i,l}g_{\tau_i}(\epsilon_i)]$  and  $\xi_q = \frac{1}{\lceil \lambda \log k \rceil} \sum_{i \in H_{2q-1}} Y_i$ , then we know  $|\xi_q| \leq 2M\tau_{k+1}$  (since  $X_{i,l}g_{\tau_i}(\epsilon) \leq |\mathbf{X}|_{2\tau_i} \leq M\tau_{k+1}$ ) and  $\mathbb{E}[\xi_i] = 0$  holds. For all  $B_1 \in \sigma(\xi_1, \dots, \xi_q)$  and  $B_2 \in \sigma(\xi_{q+1})$ , there holds  $|\mathbb{P}(B_2|B_1) - \mathbb{P}(B_2)| \leq \phi(\lceil \lambda \log k \rceil)$  for all  $q$ . By Theorem 2 of [Berkes and Philipp \(1979\)](#), there exists a sequence of independent variables  $\eta_l$ ,  $l \geq 1$  with  $\eta_l$  having the same distribution as  $\xi_l$ , and

$$\mathbb{P}(|\xi_q - \eta_q| \geq 6\phi(\lceil \lambda \log k \rceil)) \leq 6\phi(\lceil \lambda \log k \rceil).$$

For any  $\nu > 0$ , we can take  $\lambda \geq (\nu + 1)/|\log \rho|$  so that

$$\begin{aligned} & \mathbb{P}\left(\left|\frac{2}{m_k} \sum_{q=1}^{m_k/2} (\xi_q - \eta_q)\right| \geq C_3 k^{-\nu}\right) \\ & \leq \mathbb{P}\left(\left|\frac{2}{m_k} \sum_{q=1}^{m_k/2} (\xi_q - \eta_q)\right| \geq 6\phi(\lceil \lambda \log k \rceil)\right) \leq 3m_k \phi(\lceil \lambda \log k \rceil) \leq C_3 k^{-\nu}, \end{aligned} \tag{44}$$

where  $C_3 > 0$  is some constant. Next, we bound the variance of  $\xi_q$ . From equation (20.23) of [Billingsley \(1968\)](#) we know that for arbitrary  $i, j$ , there is

$$|\mathbb{E}[Y_i Y_j]| \leq 2\sqrt{\phi(|i-j|)} \sqrt{\mathbb{E}[Y_i^2] \mathbb{E}[Y_j^2]} \leq 2C_4 \rho^{|i-j|/2} M^2 \tau_{k+1}^{1-\delta}.$$

Then we compute that

$$\begin{aligned}
\text{Var}(\eta_q) &= \text{Var}(\xi_q) = \frac{1}{\lceil \lambda \log k \rceil^2} \left[ \sum_{i,j \in H_{2q-1}} \{\mathbb{E}(Y_i Y_j)\} \right] \\
&\leq \frac{2C_4}{\lceil \lambda \log k \rceil^2} \sum_{i,j=1}^{\lceil \lambda \log k \rceil} \rho^{|i-j|/2} M^2 \tau_{k+1}^{1-\delta} \\
&\leq \frac{2\lceil \lambda \log k \rceil C_4}{\lceil \lambda \log k \rceil^2} \left( \frac{2}{1-\sqrt{\rho}} - 1 \right) M^2 \tau_{k+1}^{1-\delta} \leq \frac{4C_4}{(1-\sqrt{\rho})\lceil \lambda \log k \rceil} M^2 \tau_{k+1}^{1-\delta}.
\end{aligned}$$

Notice that  $\eta_q$ 's are all uniformly bounded by  $2M\tau_{k+1}$ , we can apply Bernstein's inequality ([Bennett, 1962](#)) and yield

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{2}{m_k} \sum_{q=1}^{m_k/2} \eta_q\right| \geq t\right) &\leq 2 \exp\left(-\frac{m_k t^2/4}{\text{Var}(\eta_q)/2 + M\tau_k t/3}\right) \\
&\leq \exp\left(-\frac{m_k \lceil \lambda \log k \rceil t^2/4}{2M^2 C_\epsilon \tau_{k+1}^{1-\delta}/(1-\sqrt{\rho}) + 2\lceil \lambda \log k \rceil M\tau_{k+1} t/3}\right).
\end{aligned}$$

We take  $t = C(\sqrt{\tau_{k+1}^{1-\delta} \log k/(k+1)} + \tau_{k+1} \log^2 k/(k+1))$  for some  $C$  large enough (note that  $k \approx m_k \lceil \lambda \log k \rceil$ ). Then we have that

$$\mathbb{P}\left(\left|\frac{2}{m_k} \sum_{q=1}^{m_k/2} \eta_q\right| \geq t\right) = O(k^{-\nu}). \tag{45}$$

Combining (44) and (45) we have that

$$\mathbb{P}\left(\left|\frac{2}{m_k} \sum_{q=1}^{m_k/2} \xi_q\right| \geq C\sqrt{\frac{\tau_{k+1}^{1-\delta} \log k}{k+1}} + C\frac{\log^2 k \tau_{k+1}}{k+1}\right) = O(k^{-\nu}). \tag{46}$$

Next we denote  $\tilde{\xi} = \frac{1}{\lceil \lambda \log k \rceil} \sum_{i \in H_{2q}} Y_i$ , then we can similarly show that

$$\mathbb{P}\left(\left|\frac{2}{m_k} \sum_{q=1}^{m_k/2} \tilde{\xi}_q\right| \geq C\sqrt{\frac{\tau_{k+1}^{1-\delta} \log k}{k+1}} + C\frac{\log^2 k \tau_{k+1}}{k+1}\right) = O(k^{-\nu}), \tag{47}$$

Combining (46) and (47) we can prove that

$$\mathbb{P}\left(\left|\frac{1}{k+1}\sum_{i=1}^{k+1}\{X_{i,l}g_{\tau_i}(\epsilon_i) - \mathbb{E}[X_{i,l}g_{\tau_i}(\epsilon_i)]\}\right| \geq C\left(\sqrt{\frac{\tau_{k+1}^{1-\delta}\log k}{k+1}} + \frac{\log^2 k \tau_{k+1}}{k+1}\right)\right) = O(k^{-\nu}). \quad (48)$$

By combining it with (43) we have that

$$\mathbb{P}\left(\left|\frac{1}{k+1}\sum_{i=1}^{k+1}\{\mathbf{X}_{i}g_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i)\}\right|_2 \geq C\left(\sqrt{d}\tau_{k+1}^{-\delta'} + \sqrt{\frac{d\tau_{k+1}^{1-\delta}\log k}{k+1}} + \frac{\sqrt{d}\log^2 k \tau_{k+1}}{k+1}\right)\right) = O(k^{-\nu}).$$

**Case 2:**  $\delta > 1$ . Similarly, we can obtain the rate

$$\mathbb{P}\left(\left|\frac{1}{k+1}\sum_{i=1}^{k+1}\{X_{i,l}g_{\tau_i}(\epsilon_i) - \mathbb{E}[X_{i,l}g_{\tau_i}(\epsilon_i)]\}\right| \geq C\left(\sqrt{\frac{\log k}{k+1}} + \frac{\log^2 k \tau_{k+1}}{k+1}\right)\right) = O(k^{-\nu}),$$

by directly plugging  $\tau_k$  into (48). Then we have that

$$\mathbb{P}\left(\left|\frac{1}{k+1}\sum_{i=1}^{k+1}\{\mathbf{X}_{i}g_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i)\}\right|_2 \geq C\left(\tau_{k+1}^{-\delta'} + \sqrt{\frac{\log k}{k+1}} + \frac{\log^2 k \tau_{k+1}}{k+1}\right)\right) = O(k^{-\nu}).$$

**Proof of ii).** When  $\delta > 1$  and  $\sqrt{k/\log^3 k} = O(\tau_k)$ , we define the thresholding level  $\tilde{\tau} = C_\tau \sqrt{k+1}/\log^{3/2} k$ , and consider the truncated random variables  $\mathbf{X}_{i}g_{\tau_i}(\epsilon_i)\mathbb{I}(|\epsilon_i| \leq \tilde{\tau})$ . Here  $C_\tau$  is a sufficiently large constant. Then we have that

$$\begin{aligned} & \mathbb{P}\left(\sum_{i=1}^{k+1}\mathbf{X}_{i}g_{\tau_i}(\epsilon_i) \neq \sum_{i=1}^{k+1}\mathbf{X}_{i}g_{\tau_i}(\epsilon_i)\mathbb{I}(|\epsilon_i| \leq \tilde{\tau})\right) \\ & \leq \mathbb{P}\left(\bigcup_{i=1}^{k+1}\{\mathbf{X}_{i}g_{\tau_i}(\epsilon_i) \neq \mathbf{X}_{i}g_{\tau_i}(\epsilon_i)\mathbb{I}(|\epsilon_i| \leq \tilde{\tau})\}\right) \\ & \leq (k+1) \max_{1 \leq i \leq k+1} \mathbb{P}\left(|\epsilon_i| > \tilde{\tau}\right) = O((k+1)^{-(\delta-1)/3}). \end{aligned} \quad (49)$$

Similarly as in (43), we have that

$$\left|\frac{1}{k+1}\sum_{i=1}^{k+1}\mathbb{E}[X_{i,l}g_{\tau_i}(\epsilon_i)\mathbb{I}(|\epsilon_i| \leq \tilde{\tau})]\right| \leq \frac{1}{k+1}\sum_{i=1}^{k+1}\left|\mathbb{E}[X_{i,l}g_{\tau_i}(\epsilon_i)\mathbb{I}(|\epsilon_i| \leq \tilde{\tau})]\right| \leq C(\tau_{k+1}^{-\delta'} + \tilde{\tau}^{-\delta}). \quad (50)$$

Similarly as in the proof of (48), for every  $\nu > 0$ , there exists  $C > 0$  such that

$$\mathbb{P}\left(\left|\frac{1}{k+1}\sum_{i=1}^{k+1}\{X_{i,l}g_{\tau_i}(\epsilon_i)\mathbb{I}(|\epsilon_i| \leq \tilde{\tau}) - \mathbb{E}[X_{i,l}g_{\tau_i}(\epsilon_i)\mathbb{I}(|\epsilon_i| \leq \tilde{\tau})]\}\right| \geq C\left(\sqrt{\frac{\log k}{k+1}} + \frac{\log^2 k \tilde{\tau}}{k+1}\right)\right) = O(k^{-\nu}) \quad (51)$$

By the choice of  $\tilde{\tau}$ , we know that

$$\sqrt{\frac{\log k}{k+1}} + \frac{\log^2 k \tilde{\tau}}{k+1} = O\left(\sqrt{\frac{\log k}{k+1}}\right).$$

Combining (49), (50) and (51) we have that

$$\mathbb{P}\left(\left|\frac{1}{k+1}\sum_{i=1}^{k+1}\{\mathbf{X}_{i}g_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i)\}\right|_2 \geq C\sqrt{d}\left(\tau_{k+1}^{-\delta'} + \tilde{\tau}^{-\delta} + \sqrt{\frac{\log k}{k+1}}\right)\right) = O((k+1)^{-\min\{\nu, (\delta-1)/3\}}).$$

When there are  $\alpha_k$  fraction of outliers, denote  $\mathcal{Q}_k$  as the index set, we have

$$\begin{aligned} & \left|\frac{1}{k+1}\sum_{i=1}^{k+1}\{\mathbf{X}_{i}g_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i)\}\right|_2 \\ & \leq \frac{1}{k+1}\left|\sum_{i \in \mathcal{Q}_k}\{\mathbf{X}_{i}g_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i)\}\right|_2 + \frac{1}{k+1}\left|\sum_{i \notin \mathcal{Q}_k}\{\mathbf{X}_{i}g_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i)\}\right|_2 \\ & = O(\sqrt{d}\alpha_k\tau_{k+1}) + \frac{1}{k+1}\left|\sum_{i \notin \mathcal{Q}_k}\{\mathbf{X}_{i}g_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i)\}\right|_2, \end{aligned}$$

which proves the lemma.  $\square$

**Lemma 8.** (Bound of the mixed term) Under the same assumptions as in Theorem 1, for every  $\nu > 0$ , there exist constants  $C$  and  $c$ , such that

$$\mathbb{P}\left(\left|\frac{1}{k+1}\sum_{i=1}^{k+1}(\mathbf{A}_i - \mathbf{H})(\hat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}^*)\right|_2 \geq C(\alpha_k + \Psi^2 d e_k^2 \log k), \cap_{i=n_0}^k E_i\right) \leq c(k+1)^{-\nu}.$$

*Proof.* Firstly, from (39) and Lemma 5, under the event  $\cap_{i=n_0}^k E_i$ , we can bound the term

$$\begin{aligned}
& \left| \frac{1}{k+1} \sum_{i=1}^{k+1} (\mathbb{E}[\mathbf{A}_i] - \mathbf{H})(\widehat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}^*) \right|_2 \\
& \leq \frac{1}{k+1} \sum_{i=1}^{k+1} \|\mathbb{E}[\mathbf{A}_i] - \mathbf{H}\| \cdot |\widehat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}^*|_2 \\
& \leq \frac{1}{k+1} \sum_{i=1}^{k+1} C\Psi\tau_i^{-\delta'} e_{i-1} = O(\tau_{k+1}^{-\delta'} \sqrt{de_k} \log k) = O(\sqrt{de_k^2} \log k).
\end{aligned} \tag{52}$$

For ease of notation, we first consider the case when there is no outlier. Similar as in the proof of Lemma 1, for each  $k$ , we evenly divide the tuple  $\{n_0, \dots, k\}$  into  $m_k$  subsets  $H_1, \dots, H_{m_k}$ , where  $m_k = \lceil (k - n_0) / \lceil \lambda \log k \rceil \rceil$ . Here  $|H_q| = \lceil \lambda \log k \rceil$  for  $1 \leq q \leq m_k - 1$ , and  $|H_{m_k}| \leq \lceil \lambda \log k \rceil$ .  $\lambda$  is a sufficiently large constant which will be specified later. Then we have  $m_k \approx (k - n_0) / (\lambda \log k)$ . We further denote  $H_0 = \{1, \dots, n_0\}$ . Without loss of generality, we assume that  $m_k$  is an even integer. For each  $i \in \{n_0, \dots, k\}$ , suppose  $i \in H_{l_i}$ , we construct the following random variable

$$\tilde{\boldsymbol{\theta}}_i = \frac{1}{i} \sum_{q=0}^{l_i-2} \sum_{j \in H_q} (\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}^*) + \mathbf{H}^{-1} \frac{1}{i} \sum_{q=0}^{l_i-2} \sum_{j \in H_q} \mathbf{X}_{j+1} g_{\tau_{j+1}} (\mathbf{Z}_{j+1}^\top \widehat{\boldsymbol{\theta}}_j - b_{j+1}). \tag{53}$$

When  $l_j = 1$ , we take the sum for the terms in  $H_0$ . For  $i \in H_0$ , we take  $\tilde{\boldsymbol{\theta}}_i = \widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}^* = \widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*$ .

Then by Lemma 9 below we have that,

$$\begin{aligned}
& \mathbb{P} \left( \left| \frac{1}{k+1} \sum_{i=1}^{k+1} (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i])(\widehat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}^*) - \frac{1}{k+1} \sum_{i=1}^{k+1} (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]) \tilde{\boldsymbol{\theta}}_{i-1} \right|_2 \geq C\Psi^2 de_k^2, \cap_{i=n_0}^k E_i \right) \\
& = O((k+1)^{-\nu}).
\end{aligned} \tag{54}$$

Moreover, from the proof of Lemma 9, we can obtain that under  $\cap_{i=n_0}^k E_i$ , there is

$$|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}^* - \tilde{\boldsymbol{\theta}}_i|_2 \leq \Psi \sqrt{de_i}. \tag{55}$$

It left to bound the term  $\frac{1}{k+1} \sum_{j=0}^k (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]) \tilde{\boldsymbol{\theta}}_i$ . To be more precise, we define the  $\sigma$ -field



$\mathcal{G}_l = \sigma((\mathbf{X}_i, \mathbf{Z}_i, b_i) : i \in \cup_{j=1}^{2l+1} H_j)$ , and construct the random variable

$$\boldsymbol{\xi}_l = \sum_{i \in H_{2l+1}} (\mathbf{A}_{i+1} - \mathbb{E}[\mathbf{A}_{i+1}]) \tilde{\boldsymbol{\theta}}_i$$

Then by (55) we know that

$$\mathbb{P}\left(\cup_i \left\{|\tilde{\boldsymbol{\theta}}_i|_2 \geq 2\Psi\sqrt{de_i}\right\}, \cap_{j=n_0}^k E_j\right) = O((k+1)^{-\nu}).$$

We further set

$$\tilde{\boldsymbol{\xi}}_l = \sum_{i \in H_{2l+1}} (\mathbf{A}_{i+1} - \mathbb{E}[\mathbf{A}_{i+1}]) \tilde{\boldsymbol{\theta}}_i \mathbb{I}\left\{|\tilde{\boldsymbol{\theta}}_i|_2 \leq 2\Psi\sqrt{de_i}\right\},$$

then there is

$$\mathbb{P}\left(\sum_{l=1}^{m_k/2} \boldsymbol{\xi}_l \neq \sum_{l=1}^{m_k/2} \tilde{\boldsymbol{\xi}}_l, \cap_{i=n_0}^k E_i\right) = O(k^{-\nu}). \quad (56)$$

Notice that  $\{\tilde{\boldsymbol{\xi}}_l - \mathbb{E}[\tilde{\boldsymbol{\xi}}_l | \mathcal{G}_{l-1}], l \geq 1\}$  are martingale differences, and there is  $\mathbb{E}[\tilde{\boldsymbol{\theta}}_i | \mathcal{G}_{l-1}] = \tilde{\boldsymbol{\theta}}_i$  for  $i \in H_{2l+1}$ . Therefore we have

$$\mathbb{E}[\tilde{\boldsymbol{\xi}}_l | \mathcal{G}_{l-1}] = \sum_{i \in H_{2l+1}} \mathbb{E}[(\mathbf{A}_{i+1} - \mathbb{E}[\mathbf{A}_{i+1}]) | \mathcal{G}_{l-1}] \tilde{\boldsymbol{\theta}}_i \mathbb{I}\left\{|\tilde{\boldsymbol{\theta}}_i|_2 \leq 2\Psi\sqrt{de_i}\right\}.$$

By Lemma 3 there is

$$\mathbb{E}\left[\left\|\mathbb{E}[\mathbf{A}_{i+1} - \mathbb{E}[\mathbf{A}_{i+1}] | \mathcal{G}_{l-1}]\right\|\right] \leq Cd\sqrt{\phi(\lceil \lambda \log k \rceil)} = O(k^{-\nu-2}),$$

for some  $\lambda$  large enough. Then Markov's inequality yields

$$\mathbb{P}\left(\left\|\frac{1}{k} \sum_{l=1}^{m_k/2} \mathbb{E}[\tilde{\boldsymbol{\xi}}_l | \mathcal{G}_{l-1}]\right\|_2 \geq C \frac{\sqrt{de_k}}{k}, \cap_{i=n_0}^k E_i\right) = O(k^{-\nu}). \quad (57)$$

It is direct to verify that

$$\begin{aligned} \sup_{\mathbf{v} \in \mathbb{S}^{d-1}} |\mathbf{v}^\top (\tilde{\boldsymbol{\xi}}_l - \mathbb{E}[\tilde{\boldsymbol{\xi}}_l | \mathcal{G}_{l-1}])| &\leq C\Psi \lambda \log k, \\ \sup_{\mathbf{v} \in \mathbb{S}^{d-1}} \sum_{l=1}^{m_k/2} \text{Var}[\mathbf{v}^\top \tilde{\boldsymbol{\xi}}_l | \mathcal{G}_{l-1}] &\leq \sup_{\mathbf{v} \in \mathbb{S}^{d-1}} \sum_{l=1}^{m_k/2} \mathbb{E}[\mathbf{v}^\top \tilde{\boldsymbol{\xi}}_l^2 | \mathcal{G}_{l-1}] \leq C\Psi^2 k de_k^2 \log k. \end{aligned}$$

Then we can apply Lemma 2 and yield

$$\mathbb{P}\left(\left|\sum_{l=1}^{m_k/2} (\tilde{\boldsymbol{\xi}}_l - \mathbb{E}[\tilde{\boldsymbol{\xi}}_l | \mathcal{G}_{l-1}])\right|_2 \geq C\Psi(d \log^2 k + de_k \sqrt{k \log^2 k})\right) = O(k^{-\nu d}).$$

Combining it with (57) and (56), we have that

$$\mathbb{P}\left(\left|\frac{1}{k+1} \sum_{l=1}^{m_k/2} \boldsymbol{\xi}_l\right|_2 \geq C\Psi\left(\frac{d \log^2 k}{k} + de_k \sqrt{\frac{\log^2 k}{k}}\right), \cap_{i=n_0}^k E_i\right) = O((k+1)^{-\nu})$$

A similar result holds for the even term. Note that  $\frac{d \log^2 k}{k} + de_k \sqrt{\frac{\log^2 k}{k}} = O(de_k^2)$ , and together with (54) we have

$$\mathbb{P}\left(\left|\frac{1}{k+1} \sum_{j=0}^k (\mathbf{A}_{i+1} - \mathbb{E}[\mathbf{A}_{i+1}]) (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}^*)\right|_2 \geq C\Psi^2 de_k^2 \log k, \cap_{i=n_0}^k E_i\right) = O((k+1)^{-\nu}).$$

When taking the outliers into consideration, we have that

$$\begin{aligned} &\left|\frac{1}{k+1} \sum_{j=0}^k (\mathbf{A}_{i+1} - \mathbb{E}[\mathbf{A}_{i+1}]) (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}^*)\right|_2 \\ &= \left|\frac{1}{k+1} \sum_{j \in \mathcal{Q}_k} (\mathbf{A}_{i+1} - \mathbb{E}[\mathbf{A}_{i+1}]) (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}^*)\right|_2 + \left|\frac{1}{k+1} \sum_{j \notin \mathcal{Q}_k} (\mathbf{A}_{i+1} - \mathbb{E}[\mathbf{A}_{i+1}]) (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}^*)\right|_2 \\ &= \left|\frac{1}{k+1} \sum_{j \notin \mathcal{Q}_k} (\mathbf{A}_{i+1} - \mathbb{E}[\mathbf{A}_{i+1}]) (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}^*)\right|_2 + O(\alpha_k), \end{aligned}$$

under the event  $\cap_{i=n_0}^k E_i$ , which proves the lemma by combining it with (52).  $\square$

**Lemma 9.** Let  $\tilde{\boldsymbol{\theta}}_i$  be defined in (53), for every  $\nu > 0$ , there exist constants  $C$  and  $c$ , such that

$$\begin{aligned} & \mathbb{P}\left(\left|\frac{1}{k+1} \sum_{i=1}^{k+1} (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i])(\hat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}^*) - \frac{1}{k+1} \sum_{i=1}^{k+1} (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i])\tilde{\boldsymbol{\theta}}_{i-1}\right|_2 \geq C\Psi^2 de_k^2, \cap_{i=n_0}^k E_i\right) \\ & \leq c(k+1)^{-\nu}. \end{aligned}$$

*Proof.* Under event  $\cap_{i=n_0}^k E_i$ , using Lemma 5, we have

$$\begin{aligned} & |\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}^* - \tilde{\boldsymbol{\theta}}_i|_2 \\ &= \left| \frac{1}{i} \sum_{j=(l_i-2)\lceil \lambda \log k \rceil + n_0 + 1}^{i-1} (\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}^*) \right|_2 + \left| \mathbf{H}^{-1} \frac{1}{i} \sum_{j=(l_i-2)\lceil \lambda \log k \rceil + n_0 + 1}^{i-1} \mathbf{X}_{j+1} g_{\tau_{j+1}}(\mathbf{Z}_{j+1}^\top \hat{\boldsymbol{\theta}}_j - b_{j+1}) \right|_2 \\ & \quad + \left| (\widehat{\mathbf{H}}_{i-1}^{-1} - \mathbf{H}^{-1}) \frac{1}{i} \sum_{j=0}^{i-1} \mathbf{X}_{j+1} g_{\tau_{j+1}}(\mathbf{Z}_{j+1}^\top \hat{\boldsymbol{\theta}}_j - b_{j+1}) \right|_2 \\ &= \left| (\widehat{\mathbf{H}}_{i-1}^{-1} - \mathbf{H}^{-1}) \frac{1}{i} \sum_{j=0}^{i-1} \mathbf{X}_{j+1} g_{\tau_{j+1}}(\mathbf{Z}_{j+1}^\top \hat{\boldsymbol{\theta}}_j - b_{j+1}) \right|_2 \\ & \quad + \left| \mathbf{H}^{-1} \frac{1}{i} \sum_{j=(l_i-2)\lceil \lambda \log k \rceil + n_0 + 1}^{i-1} \mathbf{X}_{j+1} g_{\tau_{j+1}}(\mathbf{Z}_{j+1}^\top \boldsymbol{\theta}^* - b_{j+1}) \right|_2 + O\left(\frac{e_i \log k}{i}\right). \end{aligned} \tag{58}$$

For the first term, by Lemma 6 and (42) we have that under event  $\cap_{i=n_0}^k E_i$ ,

$$\|\widehat{\mathbf{H}}_{i-1}^{-1} - \mathbf{H}^{-1}\| \leq \|\widehat{\mathbf{H}}_{i-1}^{-1}\| \cdot \|\widehat{\mathbf{H}}_{i-1} - \mathbf{H}\| \cdot \|\mathbf{H}^{-1}\| \leq C\Psi\sqrt{de_i}.$$

On the other hand, there is

$$\begin{aligned} & \left| \frac{1}{i} \sum_{j=0}^{i-1} \mathbf{X}_{j+1} g_{\tau_{j+1}}(\mathbf{Z}_{j+1}^\top \hat{\boldsymbol{\theta}}_j - b_{j+1}) \right|_2 \\ & \leq \left| \frac{1}{i} \sum_{j=0}^{i-1} \mathbf{X}_{j+1} g_{\tau_{j+1}}(\mathbf{Z}_{j+1}^\top \boldsymbol{\theta}^* - b_{j+1}) \right|_2 + M^2 \frac{1}{i} \sum_{j=0}^{i-1} |\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}^*|_2 \leq C\Psi\sqrt{de_i} \log i. \end{aligned}$$

Therefore, we have that

$$\begin{aligned}
& \frac{1}{k+1} \sum_{i=1}^{k+1} (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]) (\hat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}^*) - \frac{1}{k+1} \sum_{i=1}^{k+1} (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]) \tilde{\boldsymbol{\theta}}_{i-1} \\
&= \frac{1}{k+1} \sum_{i=1}^{k+1} (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]) \mathbf{H}^{-1} \frac{1}{i} \sum_{j=(l_i-2)\lceil \lambda \log k \rceil + n_0 + 1}^{i-1} \mathbf{X}_{j+1} g_{\tau_{j+1}}(\mathbf{Z}_{j+1}^\top \boldsymbol{\theta}^* - b_{j+1}) + O(\Psi^2 d e_i^2) \\
&= \frac{1}{k+1} \sum_{i=1}^{k+1} \left\{ \sum_{j=(l_i-2)\lceil \lambda \log k \rceil + n_0 + 1}^i \frac{1}{j} (\mathbf{A}_j - \mathbb{E}[\mathbf{A}_j]) \right\} \mathbf{H}^{-1} \mathbf{X}_i g_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i) + O(\Psi^2 d e_i^2). \tag{59}
\end{aligned}$$

Denote

$$\mathbf{Y}_i = \left\{ \sum_{j=(l_i-2)\lceil \lambda \log k \rceil + n_0 + 1}^i \frac{1}{j} (\mathbf{A}_j - \mathbb{E}[\mathbf{A}_j]) \right\} \mathbf{H}^{-1} \mathbf{X}_i g_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i).$$

It is direct to verify that

$$\begin{aligned}
|\mathbb{E}[\mathbf{Y}_i]|_2 &\leq \sqrt{\mathbb{E} \left\| \sum_{j=(l_i-2)\lceil \lambda \log k \rceil + n_0 + 1}^i \frac{1}{j} (\mathbf{A}_j - \mathbb{E}[\mathbf{A}_j]) \right\|^2 \mathbb{E} \left| \mathbf{H}^{-1} \mathbf{X}_i g_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i) \right|_2^2} \\
&= O\left( \frac{\sqrt{d} \log k}{i} \tau_i^{(1-\delta)+/2} \right) = O(\sqrt{d} e_i^2). \tag{60}
\end{aligned}$$

Therefore, it left bound

$$\left| \frac{1}{k+1} \sum_{i=1}^{k+1} (\mathbf{Y}_i - \mathbb{E}[\mathbf{Y}_i]) \right|_2.$$

To this end, we further denote

$$\boldsymbol{\xi}_l = \sum_{i \in H_{4l+1}} (\mathbf{Y}_i - \mathbb{E}[\mathbf{Y}_i]),$$

We define the  $\sigma$ -field  $\mathcal{G}_l = \sigma((\mathbf{X}_i, \mathbf{Z}_i, b_i) : i \in \cup_{j=1}^{4l+1} H_j)$ , by Lemma 3 we have that

$$\mathbb{E} \left[ \left| \mathbb{E}[\boldsymbol{\xi}_l | \mathcal{G}_{l-1}] \right|_2 \right] = O\left( \frac{\sqrt{d} \tau_l \log k}{l k^{\nu+2}} \right).$$

Therefore, by Markov's inequality

$$\mathbb{P}\left(\left|\frac{1}{k} \sum_{l=1}^{m_k/4} \mathbb{E}[\boldsymbol{\xi}_l | \mathcal{G}_{l-1}]\right|_2 \geq C \frac{\sqrt{d} \tau_k \log k}{k^2}, \cap_{i=n_0}^k E_i\right) = O(k^{-\nu}). \quad (61)$$

It is direct to verify that

$$\begin{aligned} \sup_{\mathbf{v} \in \mathbb{S}^{d-1}} |\mathbf{v}^\top (\boldsymbol{\xi}_l - \mathbb{E}[\boldsymbol{\xi}_l | \mathcal{G}_{l-1}])| &\leq C \sqrt{d} \log k, \\ \sup_{\mathbf{v} \in \mathbb{S}^{d-1}} \sum_{l=1}^{m_k/2} \text{Var}[\mathbf{v}^\top \boldsymbol{\xi}_l | \mathcal{G}_{l-1}] &\leq \sup_{\mathbf{v} \in \mathbb{S}^{d-1}} \sum_{l=1}^{m_k/2} \mathbb{E}[|\mathbf{v}^\top \boldsymbol{\xi}_l|^2 | \mathcal{G}_{l-1}] \\ &\leq C_1 \sum_{l=1}^{m_k/4} d \tau_l^{(1-\delta)+} \log^2 k / l^2 \leq C d \log^3 k. \end{aligned}$$

Then we can apply Lemma 2 and yield

$$\mathbb{P}\left(\left|\sum_{l=1}^{m_k/4} (\boldsymbol{\xi}_l - \mathbb{E}[\boldsymbol{\xi}_l | \mathcal{G}_{l-1}])\right|_2 \geq C(d^{3/2} \log^2 k + d \log^2 k)\right) = O(k^{-\nu d}).$$

Combining it with (60) and (61), we have that

$$\mathbb{P}\left(\left|\frac{1}{k+1} \sum_{l=1}^{m_k/2} \boldsymbol{\xi}_l\right|_2 \geq C \frac{d^{3/2} \log^2 k}{k}, \cap_{i=n_0}^k E_i\right) = O((k+1)^{-\nu})$$

A similar result holds for the average of  $\boldsymbol{\xi}_l = \sum_{i \in H_{4l+q}} (\mathbf{Y}_i - \mathbb{E}[\mathbf{Y}_i])$ , where  $q = 0, 2, 3$ . Substitute it into (59), we have that

$$\begin{aligned} &\mathbb{P}\left(\left|\frac{1}{k+1} \sum_{i=1}^{k+1} (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]) (\hat{\boldsymbol{\theta}}_{i-1} - \boldsymbol{\theta}^*) - \frac{1}{k+1} \sum_{i=1}^{k+1} (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]) \tilde{\boldsymbol{\theta}}_{i-1}\right|_2 \geq C \Psi^2 d e_k^2, \cap_{i=n_0}^k E_i\right) \\ &= O((k+1)^{-\nu}), \end{aligned}$$

which proves the lemma.  $\square$

## 7.4 Proof of Results in Section 3.2

*Proof of Theorem 4.* From the proof of Theorem 1, as  $n_0$  tends to infinity, we can obtain that

$$\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* = \mathbf{H}^{-1} \frac{1}{n} \sum_{i \notin \mathcal{Q}_n} \mathbf{X}_i g_{\tau_i}(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i) + O_{\mathbb{P}}\left(\sqrt{d} \tau_n \alpha_n + d e_{n-1}^2 \log^2 n\right), \quad (62)$$

where  $\mathbf{H} = \mathbb{E}[\mathbf{X} \mathbf{X}^\top]$ . Denote  $\epsilon_i = \mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i$ , we first bound the term

$$\begin{aligned} & \left| \frac{1}{n} \sum_{i \notin \mathcal{Q}_n} \mathbf{X}_i (g_{\tau_i}(\epsilon_i) - \epsilon_i) \right|_2 \\ & \leq \left| \frac{1}{n} \sum_{i \notin \mathcal{Q}_n} \mathbf{X}_i (g_{\tau_i}(\epsilon_i) - \epsilon_i) - \frac{1}{n} \sum_{i \notin \mathcal{Q}_n} \mathbb{E}[\mathbf{X}_i (g_{\tau_i}(\epsilon_i) - \epsilon_i)] \right|_2 + \frac{1}{n} \sum_{i \notin \mathcal{Q}_n} \left| \mathbb{E}[\mathbf{X}_i (g_{\tau_i}(\epsilon_i) - \epsilon_i)] \right|_2 \\ & = \left| \frac{1}{n} \sum_{i \notin \mathcal{Q}_n} \mathbf{X}_i (g_{\tau_i}(\epsilon_i) - \epsilon_i) - \frac{1}{n} \sum_{i \notin \mathcal{Q}_n} \mathbb{E}[\mathbf{X}_i (g_{\tau_i}(\epsilon_i) - \epsilon_i)] \right|_2 + O(\sqrt{d} \tau_n^{-2}), \end{aligned} \quad (63)$$

where the last inequality uses the moment assumption on  $\epsilon$  and similar argument as in (43). To bound the first term, we directly compute its variance and use Chebyshev's inequality. More precisely, by equation (20.23) of Billingsley (1968), for arbitrary  $i, j$  and each coordinate  $l$ , there is

$$\begin{aligned} & \left| \mathbb{E} \left[ \left\{ X_{i,l}(g_{\tau_i}(\epsilon_i) - \epsilon_i) - \mathbb{E}[X_{i,l}(g_{\tau_i}(\epsilon_i) - \epsilon_i)] \right\} \left\{ X_{j,l}(g_{\tau_j}(\epsilon_j) - \epsilon_j) - \mathbb{E}[X_{j,l}(g_{\tau_j}(\epsilon_j) - \epsilon_j)] \right\} \right] \right| \\ & \leq 2\rho^{|i-j|/2} \sqrt{\mathbb{E}[X_{i,l}^2(g_{\tau_i}(\epsilon_i) - \epsilon_i)^2] \mathbb{E}[X_{j,l}^2(g_{\tau_j}(\epsilon_j) - \epsilon_j)^2]} \\ & \leq 2C_0 \rho^{|i-j|/2} \frac{1}{\tau_i^2 \tau_j^2} \sqrt{\mathbb{E}[|\epsilon_i|^6] \mathbb{E}[|\epsilon_j|^6]} \leq C_1 \frac{\rho^{|i-j|/2}}{\tau_i^2 \tau_j^2}. \end{aligned}$$

Therefore we have that for every unit vector  $\mathbf{v} \in \mathbb{S}^{d-1}$ , there is

$$\begin{aligned} \text{Var} \left[ \frac{1}{n} \sum_{i \notin \mathcal{Q}_n} \mathbf{X}_i (g_{\tau_i}(\epsilon_i) - \epsilon_i) \right] & \leq \frac{C_1}{n^2} \sum_{i,j=1}^n \frac{\rho^{|i-j|/2}}{\tau_i^2 \tau_j^2} \\ & \leq \frac{C_1}{n^2} \sum_{i=1}^n \frac{2}{\tau_i^2} \left( \sum_{j=0}^n \rho^{j/2} \right) \leq C_2 \frac{1}{n \tau_n^2}, \end{aligned}$$

for some constant  $C_2 > 0$ . Therefore, by Chebyshev's inequality, we have that

$$\left| \frac{1}{n} \sum_{i \notin \mathcal{Q}_n} \mathbf{X}_i(g_{\tau_i}(\epsilon_i) - \epsilon_i) - \frac{1}{n} \sum_{i \notin \mathcal{Q}_n} \mathbb{E}[\mathbf{X}_i(g_{\tau_i}(\epsilon_i) - \epsilon_i)] \right|_2 = O_{\mathbb{P}}\left(\frac{\sqrt{d}}{(n\tau_n^2)^{2/5}}\right). \quad (64)$$

Combining (62), (63) and (64), given a unit vector  $\mathbf{v} \in \mathbb{S}^{d-1}$ , we have

$$\begin{aligned} \mathbf{v}^\top (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) &= \mathbf{v}^\top \mathbf{H}^{-1} \frac{1}{n} \sum_{i \notin \mathcal{Q}_n} \mathbf{X}_i \epsilon_i + O_{\mathbb{P}}\left(\sqrt{d}\tau_n \alpha_n + de_{n-1}^2 \log^2 n + \sqrt{d}\tau_n^{-2} + \frac{\sqrt{d}}{(n\tau_n^2)^{2/5}}\right), \\ &= \frac{1}{n} \sum_{i \notin \mathcal{Q}_n} \mathbf{v}^\top \mathbf{H}^{-1} \mathbf{X}_i \epsilon_i + o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right), \end{aligned}$$

where the remainder becomes  $o_{\mathbb{P}}(1/\sqrt{n})$  under some rate constraints. Here the main term is the average of strict stationary and strong mixing sequence  $\mathbf{v}^\top \mathbf{H}^{-1} \mathbf{X}_i \epsilon_i$ . By Lemma 10, the conditions in Theorem 1 of Doukhan et al. (1994) is fulfilled. Then we can apply Theorem 1 of Doukhan et al. (1994) and yield

$$\frac{\sqrt{n}}{\sigma_{\mathbf{v}}} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \rightarrow \mathcal{N}(0, 1),$$

where

$$\begin{aligned} \sigma_{\mathbf{v}}^2 &= \sum_{i=-\infty}^{\infty} \text{Cov}(\mathbf{v}^\top \mathbf{H}^{-1} \mathbf{X}_0 \epsilon_0, \mathbf{v}^\top \mathbf{H}^{-1} \mathbf{X}_i^\top \epsilon_i) \\ &= \mathbf{v}^\top \mathbf{H}^{-1} \boldsymbol{\Sigma} (\mathbf{H}^\top)^{-1} \mathbf{v}, \end{aligned}$$

and  $\boldsymbol{\Sigma}$  is defined in (17). Therefore the theorem is proved.  $\square$

**Lemma 10.** *Let the random variable  $X$  satisfies  $\mathbb{E}[|X|^{1+\delta}]$  for some  $\delta > 1$ , and  $\phi(k) = O(\rho^k)$  (where  $\rho < 1$ ). Then we have that*

$$\int_0^1 \phi^{-1}(u) Q_X^2(u) du < \infty,$$

where  $\phi^{-1}(u)$  is the inverse function of  $\phi(k)$ , i.e.,  $\phi^{-1}(u) = \log u / \log \rho$ , and  $Q_X(u) = \inf\{t : \mathbb{P}(|X| \geq t) \leq u\}$ .

*Proof.* By Markov's inequality, we have that

$$\begin{aligned}\mathbb{P}(|X| \geq Q_X(u)) &\leq u \leq \frac{\mathbb{E}[|X|^{1+\delta}]}{(Q_X(u))^{1+\delta}}, \\ \Rightarrow Q_X(u) &\leq \left( \frac{\mathbb{E}[|X|^{1+\delta}]}{u} \right)^{1/(1+\delta)}.\end{aligned}$$

Therefore, we have that

$$\int_0^1 \phi^{-1}(u) Q_X^2(u) du \leq \int_0^1 \frac{\log u}{\log \rho} \left( \frac{\mathbb{E}[|X|^{1+\delta}]}{u} \right)^{2/(1+\delta)} du < \infty,$$

as long as  $\delta > 1$ , which proves the lemma.  $\square$

**Proposition 8.** *Suppose the conditions in Theorem 1 hold. When  $\delta > 4$  and  $\sqrt{i/\log^3 i} = O(\tau_i)$ , for every  $\nu > 0$ , there exists constants  $C, c > 0$  such that*

$$\mathbb{P}\left(\cap_{i=n_0}^n \{|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}^*|_2 \geq C\sqrt{de_i}\}\right) \geq 1 - cn_0^{-\min\{\nu, (\delta-4)/3\}},$$

where

$$e_n = \alpha_n \tau_n + \sqrt{\frac{\log n}{n}} + \frac{1}{\sqrt{d}} (c_0)^{2^{n-n_0}}. \quad (65)$$

Here  $\delta$  is defined in the moment condition in (C4).

*Proof.* When  $\delta > 4$  and  $\sqrt{k} = O(\tau_k)$ , we denote

$$e_k = \alpha_k \tau_k + \sqrt{\frac{\log k}{k}} + \frac{1}{\sqrt{d}} (c_0)^{2^{k-n_0}},$$

for  $k \geq n_0$ . We continue from equation (33) in the proof of Theorem 1.

By ii) of Lemma 7, there is

$$\mathbb{P}\left(\left|\widehat{\mathbf{H}}_{k+1}^{-1} \frac{1}{k+1} \sum_{i=1}^{k+1} \{\mathbf{X}_i g(\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i)\}\right|_2 \geq C_1 \sqrt{de_{k+1}}, \cap_{i=n_0}^k E_i\right) \leq C_3 (k+1)^{-(\delta-1)/3}.$$



Then we have that

$$\mathbb{P}\left(|\widehat{\boldsymbol{\theta}}_{k+1} - \boldsymbol{\theta}^*|_2 \geq \Psi \sqrt{de_{k+1}}, \cap_{i=n_0}^k E_i\right) \leq 3C_3(k+1)^{-\min\{\nu, (\delta-1)/3\}},$$

which yields

$$\begin{aligned} \mathbb{P}\left(\cap_{i=n_0}^k E_i\right) &\geq 1 - \sum_{i=n_0+1}^k 3C_3 i^{-\min\{\nu, (\delta-1)/3\}} \\ &\geq 1 - 3C_3 \max\left\{\frac{1}{\nu-1}, \frac{2}{\delta-4}\right\} n_0^{-\min\{\nu-1, (\delta-4)/3\}}. \end{aligned}$$

Therefore, the theorem is proved.  $\square$

*Proof of Proposition 5.* By Proposition 8 and Theorem 4, we know that  $\widehat{\boldsymbol{\theta}}_n$  admits the Bahadur representation

$$\mathbf{v}^\top (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) = \mathbf{v}^\top \mathbf{H}^{-1} \frac{1}{n} \sum_{i \notin \mathcal{Q}_n} \mathbf{X}_i (\mathbf{Z}_i^\top \boldsymbol{\theta}^* - b_i) + O_{\mathbb{P}}\left(de_{n-1}^2 \log^2 n + \sqrt{d} \tau_n^{-2} + \frac{\sqrt{d}}{(n\tau_n^2)^{2/5}}\right),$$

where  $\alpha_n = 0$  and  $e_{n-1}$  is given in (65). When  $\tau_i = Ci^\beta$  for  $\beta \geq 3/4$  we have that the remainder term has an order  $O_{\mathbb{P}}(d \log n/n)$  when  $n_0 \rightarrow \infty$ , which proves the proposition.  $\square$

## 7.5 Proof of Results in Section 4

*Proof of Theorem 6.* For simplicity we denote  $\boldsymbol{\Gamma}_k = \text{Cov}(\mathbf{X}_0 \epsilon_0, \mathbf{X}_{-k} \epsilon_{-k})$ ,  $\mathbf{Y}_{i,k} = \mathbf{X}_i g_{\tau_i}(\epsilon_i) \mathbf{X}_{i-k}^\top g_{\tau_{i-k}}(\epsilon_{i-k})$ .

Then by (17) we know

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}_0 + \sum_{k=1}^{\infty} (\boldsymbol{\Gamma}_k + \boldsymbol{\Gamma}_k^\top).$$

Under event (30), for  $k = 0, \dots, \lceil \lambda \log n \rceil$ , we have that

$$\begin{aligned} &\left\| \frac{1}{n} \sum_{i=\lceil e^{k/\lambda} \rceil}^n \mathbf{Y}_{i,k} - \frac{1}{n} \sum_{i=\lceil e^{k/\lambda} \rceil}^n \mathbf{X}_{i-k} g_{\tau_{i-k}}(\mathbf{Z}_{i-k}^\top \widehat{\boldsymbol{\theta}}_{i-k-1} - b_{i-k}) \mathbf{X}_i^\top g_{\tau_i}(\mathbf{Z}_i^\top \widehat{\boldsymbol{\theta}}_{i-1} - b_i) \right\| \\ &= O\left(\frac{1}{n} \sum_{i=\lceil e^{k/\lambda} \rceil}^n \tau_i |\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}^*|_2\right) = O_{\mathbb{P}}\left(\sqrt{d} \tau_n e_n\right). \end{aligned}$$

Therefore we have that

$$\left\| \widehat{\Sigma}_n - \left( \frac{1}{n} \sum_{i=1}^n \mathbf{Y}_{i,0} + \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^{\lceil \lambda \log i \rceil} (\mathbf{Y}_{i,k} + \mathbf{Y}_{i,k}^\top) \right) \right\| = O_{\mathbb{P}}(\sqrt{d} \tau_n e_n). \quad (66)$$

We first prove

$$\left\| \Sigma - \left( \Gamma_0 + \sum_{k=1}^{\lceil \lambda \log n/2 \rceil} (\Gamma_k + \Gamma_k^\top) \right) \right\| = O(n^{-\nu}), \quad (67)$$

for some  $\nu > 0$ . By equation (20.23) of [Billingsley \(1968\)](#), for each  $k$ , there is

$$\begin{aligned} \|\Gamma_k\| &= \sup_{\mathbf{u}, \mathbf{v} \in \mathbb{S}^{d-1}} \mathbb{E} \left[ |\mathbf{u}^\top \mathbf{X}_0 \epsilon_0 \mathbf{v}^\top \mathbf{X}_{-k} \epsilon_{-k}| \right] \\ &\leq 2 \sqrt{\phi(|k|)} \sup_{\mathbf{u}, \mathbf{v} \in \mathbb{S}^{d-1}} \sqrt{\mathbb{E}[|\mathbf{u}^\top \mathbf{X}_0 \epsilon_0|^2] \mathbb{E}[|\mathbf{v}^\top \mathbf{X}_{-k} \epsilon_{-k}|^2]} = O(\rho^{|k|/2}). \end{aligned}$$

Therefore we can obtain that

$$\begin{aligned} &\left\| \Sigma - \left( \Gamma_0 + \sum_{k=1}^{\lceil \lambda \log n/2 \rceil} (\Gamma_k + \Gamma_k^\top) \right) \right\| \\ &\leq 2 \sum_{k=\lceil \lambda \log n/2 \rceil + 1}^{\infty} \|\Gamma_k\| = O\left( \sum_{k=\lceil \lambda \log n/2 \rceil + 1}^{\infty} \rho^{|k|/2} \right) = O(\rho^{\lceil \lambda \log n \rceil / 4}) = O(n^{-\nu}), \end{aligned}$$

for  $\lambda \geq 4\nu/|\log \rho|$ , which proves (67). Next we prove that for  $k = 0, \dots, \lceil \lambda \log n \rceil$ , there holds

$$\begin{aligned} &\left\| \frac{1}{n} \sum_{i=\lceil e^{k/\lambda} \rceil}^n \mathbf{Y}_{i,k} - \frac{1}{n} \sum_{i, i-k \notin \mathcal{Q}_n} \mathbb{E}[\mathbf{Y}_{i,k}] \right\| \\ &= \left\| \frac{1}{n} \sum_{i, i-k \notin \mathcal{Q}_n} \mathbf{Y}_{i,k} - \frac{1}{n} \sum_{i, i-k \notin \mathcal{Q}_n} \mathbb{E}[\mathbf{Y}_{i,k}] \right\| + O_{\mathbb{P}}(\tau_n^2 \alpha_n) \\ &= O_{\mathbb{P}}\left( \tau_n^2 \alpha_n + \sqrt{\frac{d \log^2 n}{n}} + \frac{d \tau_n^2 \log^2 n}{n} \right). \end{aligned} \quad (68)$$

By the proof of Lemma 6, we know that

$$\left\| \frac{1}{n} \sum_{i, i-k \notin \mathcal{Q}_n} \mathbf{Y}_{i,k} - \frac{1}{n} \sum_{i, i-k \notin \mathcal{Q}_n} \mathbb{E}[\mathbf{Y}_{i,k}] \right\| \leq 2 \sup_{\mathbf{u}, \mathbf{v} \in \mathfrak{N}} \left| \frac{1}{n} \sum_{i, i-k \notin \mathcal{Q}_n} \mathbf{u}^\top (\mathbf{Y}_{i,k} - \mathbb{E}[\mathbf{Y}_{i,k}]) \mathbf{v} \right|,$$

where  $\mathfrak{N}$  is a  $1/4$ -net of  $\mathbb{S}^{d-1}$ . For  $k = 0, \dots, \lceil \lambda \log n \rceil - 1$ , Denote  $\tilde{\mathcal{F}}_{k,a}^b = \sigma(\mathbf{Y}_{i,k}, a \leq i \leq b)$ , then by (13) we know that

$$|\mathbb{P}(B|A) - \mathbb{P}(B)| \leq \phi((j-k)_+),$$

for all  $A \in \tilde{\mathcal{F}}_{k,1}^n, B \in \tilde{\mathcal{F}}_{k,n+j}^\infty$  for all  $n, j \geq 0$ . We basically rehash the proof in Lemma 1 and obtain that

$$\sup_{\mathbf{u}, \mathbf{v} \in \mathfrak{N}} \mathbb{P} \left( \left| \frac{1}{n} \sum_{i, i-k \notin \mathcal{Q}_n} \mathbf{u}^\top (\mathbf{Y}_{i,k} - \mathbb{E}[\mathbf{Y}_{i,k}]) \mathbf{v} \right| \geq C \left( \sqrt{\frac{d \log^2 n}{n}} + \frac{d \tau_n^2 \log^2 n}{n} \right) \right) = O(n^{-(\gamma+2 \log 9)d}).$$

Here the only difference is that  $\text{Var}(\eta_l) = O(1)$  in (22), and  $\mathbf{u}^\top \mathbf{Y}_{i,k} \mathbf{v}$  is bounded by  $\tau_n^2$ . Therefore (68) can be proved.

Last, we prove that

$$\left\| \frac{1}{n} \sum_{i, i-k \notin \mathcal{Q}_n} \mathbb{E}[\mathbf{Y}_{i,k}] - \mathbf{\Gamma}_k \right\| = O(\alpha_n + \tau_n^{-2}) \quad (69)$$

To see this, we compute that

$$\begin{aligned} & \left\| \mathbb{E}[\mathbf{Y}_{i,k}] - \mathbf{\Gamma}_k \right\| \\ & \leq \left\| \mathbb{E}[\mathbf{X}_i g_{\tau_i}(\epsilon_i) \mathbf{X}_{i-k}^\top g_{\tau_{i-k}}(\epsilon_{i-k})] - \mathbb{E}[\mathbf{X}_0 \epsilon_0 \mathbf{X}_{-k}^\top g_{\tau_{i-k}}(\epsilon_{-k})] \right\| \\ & \quad + \left\| \mathbb{E}[\mathbf{X}_i \epsilon_i \mathbf{X}_{i-k}^\top g_{\tau_{i-k}}(\epsilon_{i-k})] - \mathbb{E}[\mathbf{X}_0 \epsilon_0 \mathbf{X}_{-k}^\top \epsilon_{-k}] \right\| \end{aligned}$$

For the second term,

$$\begin{aligned} & \left\| \mathbb{E}[\mathbf{X}_i \epsilon_i \mathbf{X}_{i-k}^\top g_{\tau_{i-k}}(\epsilon_{i-k})] - \mathbb{E}[\mathbf{X}_0 \epsilon_0 \mathbf{X}_{-k}^\top \epsilon_{-k}] \right\| \\ & = \sup_{\mathbf{u}, \mathbf{v} \in \mathbb{S}^{d-1}} \left| \mathbb{E}[\mathbf{u}^\top \mathbf{X}_i \epsilon_i \mathbf{X}_{i-k}^\top \mathbf{v} \{g_{\tau_{i-k}}(\epsilon_{i-k}) - \epsilon_{i-k}\}] \right| \\ & \leq \sup_{\mathbf{u}, \mathbf{v} \in \mathbb{S}^{d-1}} \sqrt{\mathbb{E}[|\mathbf{u}^\top \mathbf{X}_i \epsilon_i|^2] \mathbb{E}[|\mathbf{X}_{i-k}^\top \mathbf{v} \{g_{\tau_{i-k}}(\epsilon_{i-k}) - \epsilon_{i-k}\}|^2]} \leq C_1 \tau_{i-k}^{-2}, \end{aligned}$$

for some constants  $C_1 > 0$ . Therefore (69) is proved.

Combining (66), (67), (68) and (69) we have that

$$\begin{aligned}
& \|\widehat{\Sigma}_n - \Sigma\| \\
\leq & \left\| \widehat{\Sigma}_n - \left( \frac{1}{n} \sum_{i=1}^n \mathbf{Y}_{i,0} + \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^{\lceil \lambda \log i \rceil} (\mathbf{Y}_{i,k} + \mathbf{Y}_{i,k}^\top) \right) \right\| + \left\| \Sigma - \left( \mathbf{\Gamma}_0 + \sum_{k=1}^{\lceil \lambda \log n/2 \rceil} (\mathbf{\Gamma}_k + \mathbf{\Gamma}_k^\top) \right) \right\| \\
& + 2 \sum_{k=0}^{\lceil \lambda \log n/2 \rceil} \left\| \frac{1}{n} \sum_{i=\lceil e^{k/\lambda} \rceil}^n \mathbf{Y}_{i,k} - \frac{1}{n} \sum_{i, i-k \notin Q_n} \mathbb{E}[\mathbf{Y}_{i,k}] \right\| + 2 \sum_{k=0}^{\lceil \lambda \log n/2 \rceil} \left\| \frac{1}{n} \sum_{i, i-k \notin Q_n} \mathbb{E}[\mathbf{Y}_{i,k}] - \mathbf{\Gamma}_k \right\| \\
& + 2 \sum_{k=\lceil \lambda \log n/2 \rceil+1}^{\lceil \lambda \log n \rceil} \left\| \frac{1}{n} \sum_{i=\lceil e^{k/\lambda} \rceil}^n \mathbf{Y}_{i,k} - \frac{1}{n} \sum_{i, i-k \notin Q_n} \mathbb{E}[\mathbf{Y}_{i,k}] \right\| + 2 \sum_{k=\lceil \lambda \log n/2 \rceil+1}^{\lceil \lambda \log n \rceil} \left\| \frac{1}{n} \sum_{i, i-k \notin Q_n} \mathbb{E}[\mathbf{Y}_{i,k}] - \mathbf{\Gamma}_k \right\| \\
& + 2 \sum_{k=\lceil \lambda \log n/2 \rceil+1}^{\lceil \lambda \log n \rceil} \left\| \mathbf{\Gamma}_k \right\| \\
= & O_{\mathbb{P}} \left( \sqrt{d} \tau_n e_n + n^{-\nu} + \tau_n^2 \alpha_n + \sqrt{\frac{d \log^2 n}{n}} + \frac{d \tau_n^2 \log^2 n}{n} + \alpha_n + \tau_n^{-2} \right) \\
= & O_{\mathbb{P}} \left( \sqrt{d} \tau_n^2 \alpha_n + \sqrt{d} \tau_n^{-1} + \tau_n \sqrt{\frac{d \log n}{n}} + \frac{d \tau_n^2 \log^2 n}{n} \right),
\end{aligned}$$

which proves the theorem.  $\square$