PRIVACY-AWARE DATA INTEGRATION FOR ENHANCED QUANTILE INFERENCE UNDER HETEROGENEITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantile estimation and inference play essential roles in diverse scientific and industrial applications, and their accuracy can often be enhanced by integrating auxiliary data from multiple sites. However, developing efficient aggregation methods for quantile inference under potential privacy constraints, particularly with heterogeneous datasets, remains challenging. To address these issues, we propose a systematic framework for quantile estimation and inference under potential local differential privacy (LDP). The key idea is to construct weighted estimators by adaptively aggregating quantile estimates from target and source sites. The adaptive weights are determined by minimizing the asymptotic variance, incorporating an additional ℓ_2 penalty to account for parameter shift. A parallel stochastic gradient descent algorithm under LDP constraints is developed for weight estimation and valid inference. Additionally, we introduce a conservative weighted estimator to ensure robust inference across diverse heterogeneous scenarios. Rigorous theoretical analysis establishes the consistency, normality, and effectiveness of the proposed methods. Extensive numerical studies and real data application corroborate our theoretical findings.

1 INTRODUCTION

Motivation. Quantile estimation and inference are critically important in many scientific and industrial applications (Chernozhukov & Fernández-Val, 2011; Huang et al., 2017; Kallus et al., 2024; Deuber et al., 2024; Yadlowsky et al., 2025). They offer robust summaries of data distributions, particularly for heavy-tailed or extreme outcomes. For instance, financial institutions often rely on quantile-based measures like value-at-risk to evaluate investment risks (Chen, 2008; Barbaglia et al., 2023). Given their substantial impact on risk management and decision-making, enhancing the accuracy and efficiency of quantile inference has drawn considerable attention. One promising direction for improvement is to leverage information through data integration from auxiliary datasets collected by multiple organizations (or sites) (Wang et al., 2019; Cai et al., 2024a; Han et al., 2025). However, integrating data from different sites may encounter privacy concerns, as many datasets contain sensitive personal or proprietary information protected by ethical standards and legal regulations (Dwork et al., 2006; Cai et al., 2024c). In practice, privacy requirements vary across sites: hospital consortia, financial networks, and federated platforms (e.g., smartphones or autonomous vehicles) often operate under different jurisdictional and organizational rules. Consequently, some sites may release data under record-level local differential privacy, whereas others may share unperturbed summaries or aggregates (Konečnỳ et al., 2016; Hard et al., 2018; Li et al., 2020; Nguyen et al., 2022). Consequently, how to effectively integrate auxiliary data sources to enhance the quantile estimation and inference, while satisfying potential privacy constraints through privacy-preserving techniques such as differential privacy, has become an important research problem.

Challenge. Various data integration methods have been proposed in recent literature; see Section 2 for a detailed discussion. However, these approaches still have several limitations when applied to privacy-aware quantile inference. First, existing integration methods typically construct weighted estimators by combining estimators from the target and additional datasets. They determine weights by minimizing criteria related to asymptotic variance (Li et al., 2022a; Cai et al., 2024a). Nevertheless, commonly used variance estimation techniques, such as the classical sample variance or plug-in methods (Zhu et al., 2021; Li et al., 2023; Huang et al., 2022; Gu & Chen, 2023; Han et al., 2025;

Guo et al., 2025), are no longer feasible under local differential privacy constraints. In addition, even without considering privacy constraints, most existing methods mainly focus on estimating mean parameters or optimizing smooth loss functions (Li et al., 2013; Lee et al., 2017; Chen & Xie, 2014; Li et al., 2022a; 2023). These methods rely on smoothness assumptions that are typically violated in quantile problems. Second, current data integration methods generally impose restrictive assumptions on auxiliary data sources. They often require that the source and target parameters are either identical (Lee et al., 2017; Duan et al., 2020; Zhu et al., 2021; Wang & Shen, 2024), or their differences are distinctly separated by a margin bounded away from zero (Huang et al., 2022; Li et al., 2022a; Cai et al., 2024b;c). Such assumptions cannot guarantee valid inference across the diverse heterogeneous scenarios encountered in practice.

Contributions. To address the above challenges, we propose a systematic framework to enhance quantile estimation and inference when both target and auxiliary source datasets may require LDP. The key idea is to construct weighted estimators by adaptively combining quantile estimates derived from the target and source sites. We determine the adaptive weights by minimizing the asymptotic variance of the weighted estimator, incorporating an additional ℓ_2 penalty to regularize the parameter shift. To implement this approach, we develop a parallel stochastic gradient descent (PSGD) algorithm under LDP constraints to estimate these weights and facilitate valid statistical inference. In addition, we propose a conservative weighted estimator to ensure robust inference across a wide range of potential heterogeneous scenarios. **Methodologically,** we develop a general and systematic framework for privacy-aware quantile estimation and inference via data integration. Our framework introduces multiple weighted estimators that can effectively improve estimation accuracy and inference reliability for the target quantile parameter under appropriate conditions. Moreover, the framework is broadly applicable across diverse heterogeneous scenarios. **Theoretically,** we provide rigorous guarantees for the proposed methods. Specifically, we: (i) establish consistency of the variance estimator obtained from the proposed PSGD algorithm, providing solid theoretical support for the weighted estimators and subsequent inference; (ii) establish consistency and asymptotic normality of the resulting weighted quantile estimators under diverse heterogeneous scenarios; and (iii) demonstrate that our approach consistently improves estimation and inference compared to using the target site alone under mild conditions, as long as the source sites contain useful information.

2 Related Work

Data integration. Recently, statistical data integration methods have attracted growing interest. Under the assumption of parameter homogeneity, existing studies primarily develop aggregation strategies that minimize appropriately defined asymptotic variance criteria of parameter estimators (Li et al., 2013; Chen & Xie, 2014; Wang et al., 2019; Zhu et al., 2021; Gu & Chen, 2023). When potential parameter shift exists, aggregation strategies in the existing literature typically also consider biases between parameters from auxiliary sources and the target parameter to mitigate adverse effects (Li et al., 2022a; 2023; Cai et al., 2024a;c;b; Han et al., 2025). To mitigate privacy concerns, classical methods aggregate summarized statistics (e.g., parameter estimates) rather than raw data (Chen et al., 2006; Lee et al., 2017; Duan et al., 2020; Guo et al., 2025; Bai et al., 2024), while recent approaches incorporate differential privacy constraints to achieve stronger privacy guarantees (Cai et al., 2024a;c;b). In particular, statistical data integration methods have also been extensively developed for quantile problems. The existing literature mainly focuses on estimation (Hu et al., 2021; Jiang & Yu, 2021; Tan et al., 2022; Pillutla et al., 2024; Wang & Shen, 2024; Shi et al., 2025). Recently, a few studies have also addressed inference problems (Huang et al., 2022; Bai et al., 2024).

Local differential privacy (LDP). Differential privacy (DP) bounds how much a statistic can change if one record is modified, formalizing "plausible deniability" Dwork et al. (2006). Variants such as Rényi DP, zCDP, and concentrated DP sharpen composition and enabled releases like the 2020 U.S. Census. DP's Achilles' heel is its reliance on a trusted curator; breaches, subpoenas, or misconfigurations can expose raw data Narayanan & Shmatikov (2008). Local DP (LDP) removes that trust by randomizing data at the source, generalizing randomized response Kasiviswanathan et al. (2011); Duchi et al. (2013). Pan-DP further shows only locally perturbed data withstand repeated intrusions, aligning pan-DP with LDP Amin et al. (2020). LDP is now deployed in Chrome telemetry, Safari domain statistics, and Windows Defender reporting. These applications demonstrate that curator-free privacy can coexist with high-utility analytics, spurring research on utility-optimal protocols, adaptive privacy budgeting, and federated inference.

3 METHODOLOGY

3.1 PROBLEM DESCRIPTION

We begin by introducing the model setup and notation. Due to page limitations, a complete list of notation is provided in Appendix B. Consider a total of N observations stored across a fixed set of $K+1$ sites, indexed by $\{0, 1, \ldots, K\}$. Denote the sample size at the k-th site by n_k, with $\sum_{k=0}^{K} n_k = N$, and assume that $n_k \asymp N/K$. At each site k, observations $\{X_{k,t}\}_{t=1}^{n_k} \subseteq \mathbb{R}$ are independently generated from an unknown distribution $\mathcal{P}_k$. The parameter of interest at each site is the quantile at a specified level $\tau \in (0,1)$. Specifically, define the check loss function as: $\ell(x,\theta) = (x-\theta)(\tau - \mathbf{1}(x \le \theta))$. Then, the quantile parameter at the k-th site is expressed as:

$$\theta_k = \arg\min_{\theta} \mathbb{E}_{x\sim\mathcal{P}_k}\{\ell(x,\theta)\}. \tag{3.1}$$

To estimate the parameter θ_k in practice, one typically minimizes the empirical counterpart of the objective function, which is given by $\widehat{\theta}_k = \arg\min_\theta \sum_{t=1}^{n_k} \ell(X_{k,t},\theta)$. Under regular conditions, it is assumed that $\widehat{\theta}_k$ admits the following asymptotic rule:

$$\sqrt{n_k}\left(\widehat{\theta}_k - \theta_k\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\tau(1-\tau)}{f_k^2(\theta_k)}\right), \tag{3.2}$$

where f_k denotes the probability density function associated with the distribution $\mathcal{P}_k$. Various algorithms are available for solving this empirical optimization problem, facilitating both estimation and statistical inference for θ_k. The classical and simplest method is based on order statistics (Van der Vaart, 2000). Specifically, the quantile parameter θ_k at site k can be directly estimated by taking the corresponding empirical quantile. Inference is typically conducted by plugging in a density estimator, such as a kernel density estimator, for the unknown probability density function evaluated at τ. While this method is straightforward and efficient, it directly utilizes raw data, limiting its applicability in sensitive scenarios. An alternative classical approach is Averaged Stochastic Gradient Descent (ASGD) (Polyak & Juditsky, 1992; Chen et al., 2023). Starting from an initial estimator $\widehat{\theta}_{k,0}$, ASGD iteratively updates the estimator at each site k as follows:

$$\widehat{\theta}_{k,t+1} = \widehat{\theta}_{k,t} - \eta_{k,t}\left\{\tau - \mathbf{1}\left(X_{k,t+1} \le \widehat{\theta}_{k,t}\right)\right\}, \tag{3.3}$$

where $0 \le \eta_{k,t} \le 1$ denotes the learning rate. The final estimator is computed as the average of all iterates as $\widehat{\theta}_k = n_k^{-1}\sum_{t=1}^{n_k}\widehat{\theta}_{k,t}$. Statistical inference can be conveniently implemented using self-normalized methods (Li et al., 2022b; Lee et al., 2022). Compared to the order-statistics-based method, the iterative nature of ASGD makes it naturally amenable to incorporating privacy-preserving mechanisms. Specifically, we consider here the ASGD algorithm under LDP constraints (Liu et al., 2023). Before introducing the detailed algorithm, we first provide formal definitions for DP and LDP.

Definition 1 (DP, see (Dwork et al., 2006)). *A randomized algorithm $\mathcal{A}$, taking a dataset consisting of individuals as its input, is (ϵ,δ)-differentially private if, for any pair of datasets S and S' that differ in the record of a single individual and any event E, satisfies $\mathbb{P}[\mathcal{A}(S) \in E] \le e^{\epsilon}\mathbb{P}[\mathcal{A}(S') \in E] + \delta$. When $\delta = 0$, $\mathcal{A}$ is called ϵ-differentially private (ϵ-DP).*

Definition 2 (LDP, see Joseph et al. (2019)). *An (ϵ,δ)-randomizer $R : X \to Y$ satisfies (ϵ,δ)-LDP if, for any event E and any input data point $X \ne X'$, $\mathbb{P}[R(X) \in E] \le e^{\epsilon}\mathbb{P}[R(X') \in E] + \delta$.*

Next, we modify the classical ASGD procedure in (3.3) by incorporating a local randomization step into the binary indicator function $\mathbf{1}(X_{k,t+1} \le \widehat{\theta}_{k,t})$. To be more precise, at the t-th iteration, we issue a query to the private data point $X_{k,t+1}$. In response, with probability r_k, we receive the true binary indicator, and with probability $1-r_k$, we receive an random variable $v \sim \text{Bernoulli}(0.5)$; see detailed Algorithm A.1 in Appendix. Here, the response rate r_k controls the level of privacy protection, with smaller values corresponding to stronger privacy guarantees. When $r_k = 1$, the method reduces to the standard non-private case. Since the observed binary variable is now a randomized version of the original indicator, it is necessary to execute bias-correction to ensure an unbiased gradient estimate. Let $\widehat{\zeta}_{k,t}$ denote the perturbed binary variable observed at iteration t and site k. Under this LDP mechanism, the iterative updating formula in (3.3) becomes:

$$\widehat{\theta}_{k,t+1} = \widehat{\theta}_{k,t} - \eta_{k,t}\left\{\frac{1+r_k-2r_k\tau}{2}\widehat{\zeta}_{k,t} - \frac{1-r_k+2r_k\tau}{2}\left(1-\widehat{\zeta}_{k,t}\right)\right\}. \tag{3.4}$$

Similarly as in the classical ASGD method, the final estimator is obtained by averaging the iterates over n_k steps. Statistical inference can then be performed using self-normalization techniques adapted to this LDP setting, see (Liu et al., 2023).

The methods described above provide feasible algorithms for solving the optimization problem (3.1) at each site k, with possible LDP constraints. Specifically, the classical order-statistics approach can be applied directly if privacy protection is unnecessary, whereas the ASGD-based method should be employed in an LDP setting. However, the parameter estimation and inference at each site can potentially be improved further by appropriately aggregating data across multiple sites, especially if other sites contain useful and relevant information, such as sharing the same underlying quantile parameter. Without loss of generality, we treat site 0 as the target site and the remaining K sites as source sites. An important and natural question thus arises.

> Under possible LDP constraints, how can information from these K source sites be efficiently leveraged to enhance estimation and inference of the target's quantile parameter θ_0?

Figure 1 illustrates the challenges of data integration under LDP through a simple example.

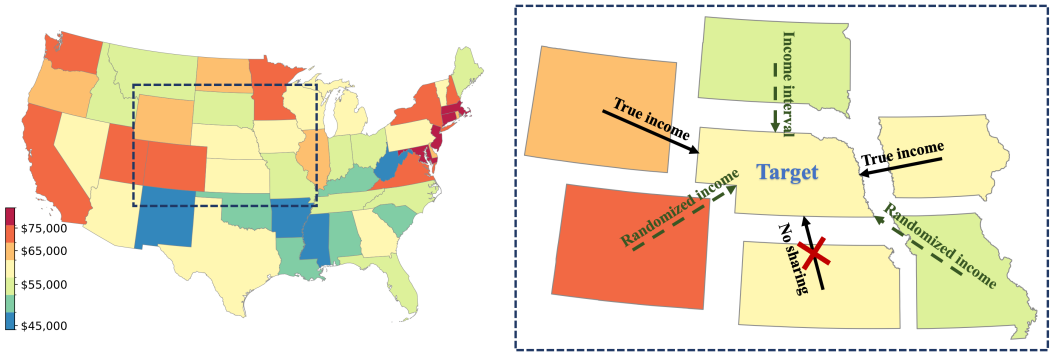


Figure 1: Consider improving the estimation of household median annual income at a target state in the United States by leveraging information from surrounding states. Two main challenges arise: (1) median incomes at different source states may differ from that of the target state; and (2) source states may face privacy-preserving requirements, which can vary across states.

Remark 1. *Beyond the proposed LDP mechanism in* (3.4)*, one can also achieve LDP by directly adding noise to the local stochastic gradients (Song et al., 2013). We compare this variant (refer to DP-SGD) with our proposed method in Appendix A.1. The results show that DP-SGD yields larger mean squared errors and wider confidence intervals than our approach.*

3.2 Adaptive Weighted Estimator

A natural way to utilize information from multiple sites is to combine the estimators from the target and source sites into a weighted estimator. Specifically, for each site k, let $\widehat{\theta}_k$ denote the estimator of θ_k derived by one of the previously discussed methods, and let w_k be the corresponding weight satisfying $w_k \geq 0$ for all $0 \leq k \leq K$ and $\sum_{k=0}^{K} w_k = 1$. We then define the weighted estimator $\widehat{\theta}(\mathbf{w}) = \sum_{k=0}^{K} w_k \widehat{\theta}_k$, where $\mathbf{w} = \{w_k\}_{k=0}^{K}$. Our goal is to determine weights $\{w_k\}$ that maximize the efficiency of the weighted estimator while controlling for the negative impact arising from heterogeneity in data distributions and parameters between the target and source sites. To this end, we introduce the following loss function with respect to the weights $\mathbf{w}$:

$$\mathcal{L}(\mathbf{w}) = \sum_{k=0}^{K} w_k^2 \sigma_k^2 / n_k + \lambda \sum_{k=0}^{K} w_k^2 b_k^2,$$

where σ_k^2/n_k is the asymptotic variance of the estimator $\widehat{\theta}_k$. We will rigorously prove that $\sigma_k^2 = \{4r_k^2 f_k^2(\theta_k)\}^{-1}\{1 - r_k^2(2\tau - 1)^2\}$ in subsequent theoretical analysis. The bias term $b_k = \theta_k - \theta_0$

represents the parameter shift of the k-th source site relative to the target site 0. The tuning parameter $\lambda \geq 0$ controls the trade-off between variance and bias. In particular, setting $\lambda = 0$ yields classical inverse-variance weighting (Zhu et al., 2021; Shi et al., 2023), while setting $\lambda = 1$ approximately corresponds to minimizing the mean squared error of the estimator (Li et al., 2023). Minimizing $\mathcal{L}(\mathbf{w})$ with respect to $\mathbf{w}$ yields the oracle weights $\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w})$, which has the following explicit closed-form solution:

$$w_k^* = \left\{ \sum_{j=0}^{K} \left(\frac{\sigma_j^2}{n_j} + \lambda b_j^2 \right)^{-1} \right\}^{-1} \left(\frac{\sigma_k^2}{n_k} + \lambda b_k^2 \right)^{-1}, \quad 0 \leq k \leq K. \tag{3.5}$$

Here oracle weights refer to the ideal weights that theoretically minimize the asymptotic variance of the estimator and assume the true site-specific asymptotic variances are known. The weights defined in (3.5) are adaptive in the sense that they automatically adjust according to the characteristics of each source site. Specifically, sites with higher noise levels (e.g., lower response rates r_k) or larger parameter shifts (i.e., larger biases b_k) are assigned smaller weights. Conversely, sites with lower noise levels and smaller parameter shifts receive relatively larger weights. This enables the weighted estimator to efficiently prioritize sources that are more informative and relevant to the target site.

In practice, the oracle weights defined in (3.5) involve unknown parameters and thus must be estimated from data. A natural estimator for the bias term is given by: $\widehat{b}_k = \widehat{\theta}_k - \widehat{\theta}_0$. For estimating the variance term σ_k^2, the classical plug-in method is commonly employed in the literature when an explicit form of σ_k^2 is available (Han et al., 2025). However, in the scenario considered here, some sites may require LDP protection. In such cases, as indicated by the asymptotic result (3.2), the plug-in approach is infeasible because the raw data necessary to estimate f_k is unavailable. To address this challenge, we propose a PSGD under LDP constraints to automatically estimate the variance σ_k^2 when raw data are not directly accessible. The key idea of PSGD is to partition the local data at each site into multiple subsets and then run independent SGD procedures, referred to as chains, in parallel on these subsets, enabling automatic estimation of the variance term.

Specifically, at each local site k, the original data $\{X_{k,t}\}_{t=1}^{n_k}$ is randomly partitioned into M_k subsets. Each subset corresponds to an i.i.d. SGD chain. We denote the data within the m-th chain by $\{X_{k,t}^{(m)}\}_{t=1}^{\lfloor n_k/M_k \rfloor}$ for $1 \leq m \leq M_k$. For each chain m, the PSGD algorithm initializes an estimator $\widehat{\theta}_{k,0}^{(m)} = \widehat{\theta}_{k,0}$ for $0 \leq k \leq K$ and updates it iteratively as follows:

$$\widehat{\theta}_{k,t+1}^{(m)} = \widehat{\theta}_{k,t}^{(m)} - \eta_{k,t} \left\{ \frac{1 + r_k - 2r_k\tau}{2} \widehat{\zeta}_{k,t}^{(m)} - \frac{1 - r_k + 2r_k\tau}{2} \left(1 - \widehat{\zeta}_{k,t}^{(m)} \right) \right\}, \tag{3.6}$$

where $\widehat{\zeta}_{k,t}^{(m)}$ denotes the locally randomized version of the indicator $\mathbf{1}(X_{k,t+1}^{(m)} \leq \widehat{\theta}_{k,t}^{(m)})$. After completing the iterations within each chain, we compute the chain-specific estimator by $\widehat{\theta}_k^{(m)} = (\lfloor n_k/M_k \rfloor)^{-1} \sum_{t=1}^{\lfloor n_k/M_k \rfloor} \widehat{\theta}_{k,t}^{(m)}$. The final estimator at site k is then obtained by averaging across the M_k chain-specific estimators: $\widehat{\theta}_k = M_k^{-1} \sum_{m=1}^{M_k} \widehat{\theta}_k^{(m)}$. Note that these chain-specific estimators $\{\widehat{\theta}_k^{(m)}\}$ are independent of each other. This inspires the following variance estimator:

$$\widehat{\sigma}_k^2 = (M_k - 1)^{-1} \sum_{m=1}^{M_k} \lfloor n_k/M_k \rfloor \left(\widehat{\theta}_k^{(m)} - \widehat{\theta}_k \right)^2.$$

Next, we estimate the oracle weights defined in equation (3.5) by replacing the unknown parameters σ_k^2 and b_k with their estimators $\widehat{\sigma}_k^2$ and $\widehat{b}_k$, respectively. We denote these estimated weights as $\{\widehat{w}_k\}$. Subsequently, the resulting weighted estimator for the target parameter θ_0 and corresponding variance estimator are obtained as

$$\widehat{\theta}_{\text{est}} = \sum_{k=0}^{K} \widehat{w}_k \widehat{\theta}_k, \quad \widehat{\sigma}_{\text{est}}^2 = \sum_{k=0}^{K} \frac{N}{n_k} \widehat{w}_k \widehat{\sigma}_k^2.$$

We will theoretically show that $\widehat{\sigma}_{\text{est}}^2/N$ is a consistent estimator of $\text{Var}(\widehat{\theta}_{\text{est}})$. Thus, statistical inference for $\widehat{\theta}_{\text{est}}$ can be readily conducted by constructing a $(1-\alpha)$-confidence interval:

$[\widehat{\theta}_{\text{est}} - \mathcal{Z}_{\alpha/2}\widehat{\sigma}_{\text{est}}/\sqrt{N}, \widehat{\theta}_{\text{est}} + \mathcal{Z}_{\alpha/2}\widehat{\sigma}_{\text{est}}/\sqrt{N}]$, where $\mathcal{Z}_{\alpha/2}$ is the upper $(\alpha/2)$-quantile of the standard normal distribution. The complete pseudo code for the data integration algorithm is described below in Algorithm A.2.

Remark 2. *It should be noted that inference based on PSGD has recently been studied in the literature (Zhu et al., 2024), but existing work primarily focuses on smooth loss functions and does not provide consistent variance estimation. Extending these results to non-smooth quantile loss with consistent variance estimation is itself an important and challenging problem.*

3.3 Conservative Weighted Estimator

Algorithm A.2 provides an adaptive weighted method capable of automatic inference under possible LDP constraints. However, as demonstrated in the subsequent theoretical analysis, the validity of this inference depends on certain conditions regarding the parameter shift between the target and source sites. Specifically, we require the bias b_k to be either vanishing or clearly distinguishable (i.e., significantly smaller or larger than $N^{-1/2}$). Although this requirement is weaker than most existing assumptions in the literature (Shi et al., 2023; Gu & Chen, 2023; Han et al., 2025), it neglects an important intermediate scenario in which the bias b_k is exactly of order $N^{-1/2}$. Unfortunately, even the oracle weights fail to yield valid statistical inference in such intermediate cases. The intuition is as follows. Consider the simple scenario where there is only one source site $(K = 1)$. If b_1 is significantly smaller than $N^{-1/2}$, the bias of the weighted estimator can be safely ignored relative to its variance. If b_1 is significantly larger than $N^{-1/2}$, the oracle weights naturally assign minimal weight to site 1, thereby reducing its negative impact. However, if b_1 is exactly of order $N^{-1/2}$, it becomes comparable to the standard errors of $\widehat{\theta}_0$ and $\widehat{\theta}_1$. In this situation, the oracle weights fail to sufficiently down-weight site 1. Consequently, the final weighted estimator retains a bias of the same order as its variance, invalidating statistical inference. Our proposed weighted estimator inevitably faces the same issue. Figure 2 illustrates the inference performance when the estimator's bias is either negligible or comparable to its variance.

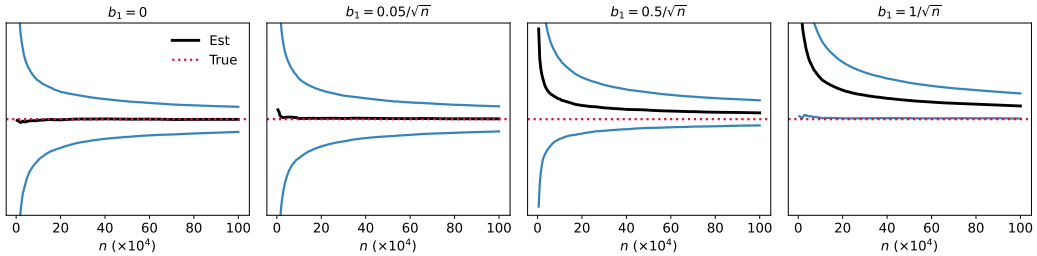


Figure 2: Illustration of confidence intervals under varying levels of bias. Each panel shows an estimator (black line) with bias b_1 and known variance $1/n$. The corresponding confidence interval with 95% significance level is represented by the blue lines. The red dashed line represents the true parameter value. The bias increases gradually from left to right. In the rightmost panel, the confidence interval no longer covers the true parameter, indicating that inference becomes invalid.

To address this issue, we propose a conservative approach based on the adaptive weighting method in Section 3.2. The key idea is to construct a conservative estimate of b_k, which tends to further down-weight biased source sites. Specifically, we replace the direct bias estimates with conservative upper bounds as

$$\widetilde{b}_k = \left|\widehat{b}_k\right| + \mathcal{C}\sqrt{\widehat{\sigma}_k^2/n_k + \widehat{\sigma}_0^2/n_0}, \;\; \mathcal{C} > 0,$$

where $\widehat{\sigma}_k^2/n_k + \widehat{\sigma}_0^2/n_0$ is the asymptotic variance of $\widehat{b}_k$, and $\mathcal{C}$ controls the conservativeness level. We then derive the conservative weights $\widetilde{w}_k$ for $0 \le k \le K$ by replacing σ_k^2 and b_k in equation (3.5) with $\widehat{\sigma}_k^2$ and $\widetilde{b}_k$, respectively, and setting $\widetilde{b}_0 = 0$. Accordingly, the conservative weighted estimator for the target parameter θ_0 is defined as: $\widehat{\theta}_{\text{cons}} = \sum_{k=0}^{K} \widetilde{w}_k \widehat{\theta}_k$. The corresponding variance estimator for $\widehat{\theta}_{\text{cons}}$ is given by $\widehat{\sigma}_{\text{cons}}^2 = \sum_{k=0}^{K} N\widetilde{w}_k^2 \widehat{\sigma}_k^2/n_k$.

4 THEORETICAL PROPERTIES

The objective of this section is two-fold. First, we establish theoretical guarantees for the proposed PSGD algorithm. Second, we investigate the statistical properties of the proposed weighted estimators. To this end, the following standard technical conditions are needed.

Assumption 1 (Property of f_k)**.** *For every $0 \leq k \leq K$, $f_k(\cdot)$ is continuous and $f_k(\theta_k) > 0$. In addition, $|f_k'(\cdot)|$ is uniformly bounded by C for some constant $C > 0$.*

Assumption 2 (Decaying learning rate)**.** *The learning rate $\eta_{k,t}$ in equation* (3.6) *satisfies $\eta_{k,t} \asymp t^{-\beta}$ for some constant $\beta \in (1/2, 1)$.*

Assumption 3 (Number of chains)**.** *Assume that $M_k^{(\beta+1/2)\wedge(2-\beta)} \lesssim N^{(\beta-1/2)\wedge(1-\beta)}$.*

Remark 3. *Assumption 1 imposes standard regularity conditions on the probability density function. Assumption 2 requires decay learning rate, commonly assumed in the SGD literature (Polyak & Juditsky, 1992; Lee et al., 2022; Li et al., 2022b). Assumption 3 restricts the growth rate of the number of PSGD chains M_k relative to the sample size, ensuring accuracy of the final averaged estimator.*

Let $\{\widehat{\theta}_k\}$ and $\{\widehat{\sigma}_k^2\}$ be the estimators produced by the PSGD in Algorithm A.2. It is worth noting that the entire algorithm to solve for $\widehat{\theta}_k, \widehat{\sigma}_k^2$ relies on iterations involving $\widehat{\theta}_{k,t}^{(m)}$, which is a linear function of the perturbed gradient. Consequently, following the arguments presented in (Liu et al., 2023), PSGD Algorithm satisfies the definition of LDP, as summarized in Proposition 4.1. Subsequently, the asymptotic properties of these estimators are established in Theorem 4.1.

Proposition 4.1 (Differential privacy)**.** *The PSGD algorithm for the k-th site satisfies $(\epsilon_k, 0)$-LDP with $\epsilon_k = \log\{(1+r_k)/(1-r_k)\}$. Therefore, the Algorithm A.2 integrating $K+1$ sites LDP data is $(\max_{0\leq k\leq K} \epsilon_k, 0)$- LDP.*

Theorem 4.1 (Asymptotic normality)**.** *Under Assumptions 1 – 3, for each $k = 0, \ldots, K$, we have $\mathbb{E}|\widehat{b}_k^2 - b_k^2| \lesssim 1/N + |b_k|/\sqrt{N}$, $\widehat{\sigma}_k^2 - \sigma_k^2 = \mathcal{O}_p(1)$,and $\widehat{\theta}_k$ satisfies the following asymptotic normality: $\sqrt{N}\left(\widehat{\theta}_k - \theta_k\right) \xrightarrow{d} \mathcal{N}\left(0, N\sigma_k^2/n_k\right)$.*

Theorem 4.1 establishes the consistency and asymptotic normality of the estimator at the k-th site. We find that larger values of r_k lead to smaller asymptotic variance. In particular, when $r_k = 1$, the asymptotic variance coincides with the classical non-private quantile estimation result (3.2). Therefore, we generally employ the PSGD estimator across the K source sites.

Next, we provide theoretical guarantees for the weighted estimators. It is worth noting that our theory is not restricted to estimators obtained from the PSGD algorithm. In fact, it applies broadly as long as consistent variance estimators are available. To accommodate this general setting, we introduce the following assumptions.

Assumption 4 (Regularity of estimators)**.** *For each $0 \leq k \leq K$, the estimators $\widehat{\sigma}_k^2$ and $\widehat{\theta}_k$ satisfies that $\widehat{\sigma}_k^2 - \sigma_k^2 = \mathcal{O}_p(1)$, and $\sqrt{N}\big(\widehat{\theta}_k - \theta_k\big) \xrightarrow{d} \mathcal{N}\big(0, N\sigma_k^2/n_k\big)$.*

Assumption 5 (Bias scale)**.** *Assume the bias b_k for each $1 \leq k \leq K$ satisfies at least one of the following conditions: (1) Vanishing bias: $b_k \ll N^{-1/2}$ or (2) Distinguishable bias: $b_k \gg N^{-1/2}$.*

Assumption 6 (The choice of λ)**.** *Assume that $\lambda = \mathcal{O}(1)$ and $\lambda b_k \gg N^{-1/2}$ for every $b_k \gg N^{-1/2}$.*

Remark 4. *Assumption 4 is a general regularity condition that accommodates various estimators, including for example, the proposed PSGD method and the classical order statistics-based estimator in the non-private case (Van der Vaart, 2000). Assumption 5 allows the parameter shift to either vanish faster than $N^{-1/2}$ or be significantly larger than $N^{-1/2}$, covering a wide range of heterogeneous scenarios. This condition relaxes existing assumptions in the literature, which typically require either zero parameter shift (Zhu et al., 2021; Gu & Chen, 2023) or a parameter shift bounded away from zero (Li et al., 2022a; Han et al., 2025). Assumption 6 specifies conditions on the tuning parameter λ.*

In our theoretical analysis, we investigate the properties of the proposed estimators under various choices of λ. Specifically, we establish the following theorem to facilitate valid statistical inference.

Theorem 4.2 (Consistency and asymptotic normality)**.** *Under Assumption 4, consider the following three scenarios: (a) $\lambda = 0$, the bias scale satisfies Assumption 5 (1); (b) λ is bounded away from 0, the bias scale satisfies Assumption 5 (2); (c) λ satisfies Assumption 6, the bias scale satisfies Assumption 5. Then we have (1) $\widehat{w}_k - w_k^* = o_p(1)$ for $1 \le k \le K$ and (2) $\sqrt{N/\widehat{\sigma}^2_{\text{est}}}(\widehat{\theta}_{\text{est}} - \theta_0) \xrightarrow{d} \mathcal{N}(0,1)$.*

It should be noted that our inference is constructed using normal quantiles, which yields narrower confidence intervals compared to self-normalization methods (Liu et al., 2023) at the same confidence level, even without data integration. We next present the following theorem to demonstrate the efficiency of data integration:

Theorem 4.3 (Improved efficiency)**.** *Under Assumptions 4 – 6, further assume there exists $b_k \ll N^{-1/2}$ for some $1 \le k \le K$, the asymptotic variance of $\widehat{\theta}_{\text{est}}$ is strictly smaller than that of $\widehat{\theta}_0$.*

Theorem 4.3 shows that as long as at least one auxiliary source site has a bias significantly smaller than $N^{-1/2}$, the weighted estimator achieves a strictly smaller asymptotic variance compared to $\widehat{\theta}_0$. Consequently, our method enhances estimation efficiency and yields narrower confidence intervals at the same confidence level compared to relying solely on the target site's data. Finally, we summarize the asymptotic behaviour of the proposed conservative estimator as follows.

Theorem 4.4 (Asymptotic normality of $\widehat{\theta}_{\text{cons}}$)**.** *Under Assumption 4, further assume that λ is bounded away from 0, $\mathcal{C} \to \infty$ and $\mathcal{C}\sqrt{\widehat{\sigma}_k^2 - \sigma_k^2} = \mathcal{O}_p(1)$ for $0 \le k \le K$, then we have $\sqrt{N/\widehat{\sigma}^2_{\text{cons}}}(\widehat{\theta}_{\text{cons}} - \theta_0) \xrightarrow{d} \mathcal{N}(0,1)$.*

Unlike Theorem 4.2, Theorem 4.4 adds no extra bias-scale constraints, so the conservative method remains robust across more heterogeneous settings.

5 Experiments

In this section, we examine the finite-sample performance of the proposed data integration method on both synthetic and real data. In synthetic data, we fix quantile levels at $\tau = 0.25, 0.5, 0.75$. Data at each site are generated from either the Normal distribution $\mathcal{N}(\mu_k, 1)$ or the Cauchy distribution $\mathcal{C}(\mu_k, 1)$ with the target site fixed at $\mu_0 = 0$. We set the number of sites as $K = 3$ and the response rate to $r_k = 0.5$ for $0 \le k \le K$. The target sample size n_0 ranges from $20,000$ to $200,000$, and the source sample sizes are three times larger. The number of local chains M_k varies between 8 and 20. The learning rate for chain m at site k in the t-th iteration is set as $\eta_{k,t_m} = 1/t_m^{0.6}$. Each experiment is replicated $1,000$ times. We consider the following estimators for comparison:

- **ADP(0):** The proposed adaptive weighted estimator with $\lambda = 0$, which ignores parameter shift.
- **ADP(1):** The proposed adaptive weighted estimator with $\lambda = 1$.
- **ADP(cv):** The proposed adaptive weighted estimator with λ selected via cross-validation.
- **ADP(cons):** The proposed conservative weighted estimator with $\lambda = 1$ and $\mathcal{C} = 1.96$.
- **Target (Liu et al., 2023):** An ASGD-based estimator using only the target site's data under potential LDP constraints. Inference is conducted via self-normalization.

We evaluate these estimators using three metrics: mean squared error on the log scale (log MSE), empirical coverage probability (ECP), and average confidence interval (CI) length. Complete implementation details and definitions of these metrics can be found in Appendix A.2. Due to space constraints, we present selected results in the main text and defer additional results to Appendix A.2.

First, we evaluate finite-sample performance under scenarios of either vanishing or distinguishable bias, in order to verify Theorems 4.2 and 4.3. Specifically, for each $1 \le k \le 3$, we set μ_k to be either 0, $0.1/\sqrt{n_0}$, or $100/\sqrt{n_0}$. This setup creates 10 distinct bias levels ranging from complete homogeneity (level 1) to strong heterogeneity (level 10); see Table A.1 in Appendix A.2 for detailed descriptions. Results are presented in Figure 3. We find that (1) ADP(0) achieves the lowest log MSE and shortest confidence interval length at relatively small bias levels. ADP(cv) outperforms ADP(1) at smaller bias levels and performs comparably at larger bias levels. (2) At relatively large bias levels, ADP(0) becomes severely biased, causing its ECP to fall significantly below 95%. In

contrast, ADP(cv) and ADP(1) consistently outperform Target except under extreme heterogeneity (level 10). These results validate our theoretical results very well. Next, we examine finite-sample performance under more general bias scenarios by varying μ_k from $\exp(-5)$ to 1 for $1 \leq k \leq 3$, presented on a logarithmic scale for clearer illustration. The results are shown in Figure 3 and the results highlight the robustness of the conservative estimator across diverse heterogeneity scenarios.

In the experiments above, we fixed the response rate at 0.5 for all sites to clearly evaluate the impact of bias and validate our theoretical results. We also conducted experiments with varying response rates to further enrich our analysis. To further strengthen our simulation study, we conducted additional experiments, including (i) sensitivity analysis of learning rates and (ii) evaluations of finite-sample performance under smaller chains M_0 and varying numbers of sites K. Detailed results are provided in Appendices A.2. Finally, we also evaluate our method on a real-world dataset: the Government Salary Dataset (Plečko et al., 2024) in Appendix A.3.

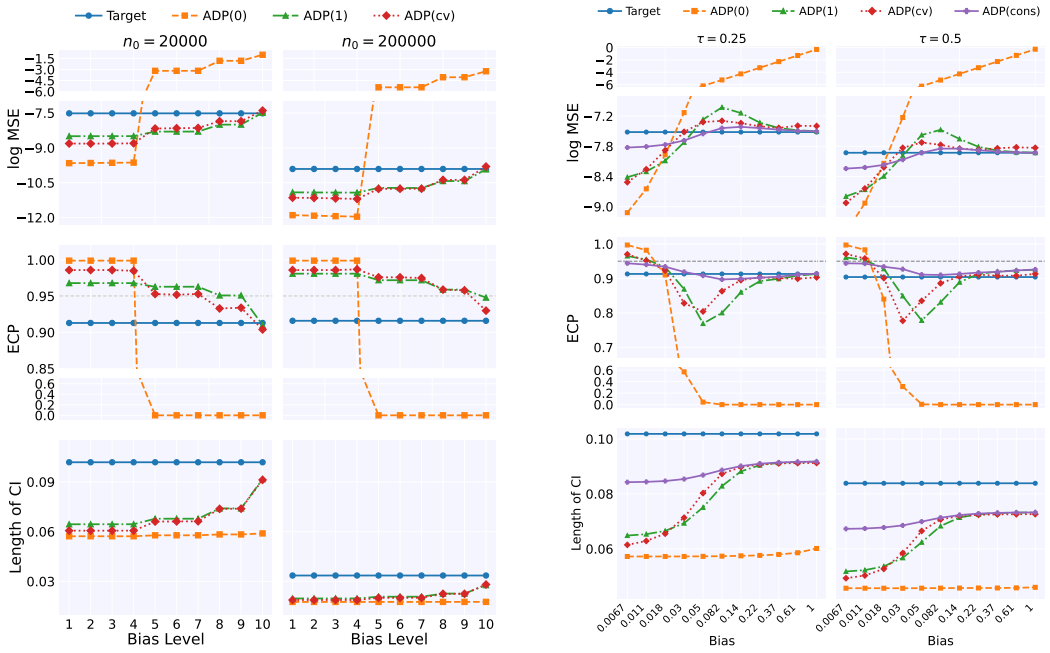


Figure 3: Simulation results under different heterogeneity scenarios. The left panel considers scenarios of either vanishing or distinguishable bias. Bias levels range from complete homogeneity to strong heterogeneity. The quantile level is fixed at $\tau = 0.25$. The right panel considers a broader range of bias values from $\exp(-5)$ to 1. The target sample size is fixed at $n_0 = 20,000$. Results are shown for various quantile levels. Data for both panels are generated from normal distributions, with the response rate fixed at $r_k = 0.5$ for all sites.

6 Concluding Remark

In summary, we propose a unified, privacy-aware framework leveraging auxiliary data to enhance quantile estimation and inference under local differential privacy. By optimally weighting estimators via a PSGD algorithm and penalizing parameter shift, our approach systematically reduces variance and ensures robustness through a conservative alternative. Theoretical results establish consistency and asymptotic normality across diverse heterogeneity settings, demonstrating improved efficiency and reliability over target-only methods, thus offering a principled and practical solution for privacy-preserving quantile inference. However, there also exist some limitations in our framework. First, our asymptotic theory requires the number of chains M_k to diverge; thus, additional investigation into fixed-chain scenarios or non-asymptotic results is necessary. Second, the current framework assumes a fixed number of sites (K); extending the methodology to accommodate diverging K remains open.

REPRODUCIBILITY STATEMENT

All numerical experiments and real-data analyses are fully reproducible via the code included in the submitted anonymized supplementary materials.

REFERENCES

Kareem Amin, Matthew Joseph, and Jieming Mao. Pan-private uniformity testing. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 183–218. PMLR, 09–12 Jul 2020.

Ruiqi Bai, Yijiao Zhang, Hanbo Yang, and Zhongyi Zhu. Transfer learning for high-dimensional quantile regression with distribution shift. *arXiv preprint arXiv:2411.19933*, 2024.

Luca Barbaglia, Sergio Consoli, and Sebastiano Manzan. Forecasting with economic news. *Journal of Business & Economic Statistics*, 41(3):708–719, 2023.

T Tony Cai, Abhinav Chakraborty, and Lasse Vuursteen. Federated nonparametric hypothesis testing with differential privacy constraints: Optimal rates and adaptive tests. *arXiv preprint arXiv:2406.06749*, 2024a.

T Tony Cai, Abhinav Chakraborty, and Lasse Vuursteen. Optimal federated learning for nonparametric regression with heterogeneous distributed differential privacy constraints. *arXiv preprint arXiv:2406.06755*, 2024b.

Tony Cai, Abhinav Chakraborty, and Lasse Vuursteen. Optimal federated learning for functional mean estimation under heterogeneous privacy constraints. *arXiv preprint arXiv:2412.18992*, 2024c.

Likai Chen, Georg Keilbar, and Wei Biao Wu. Recursive quantile estimation: Non-asymptotic confidence bounds. *Journal of Machine Learning Research*, 24(91):1–25, 2023.

Song Xi Chen. Nonparametric estimation of expected shortfall. *Journal of financial econometrics*, 6(1):87–107, 2008.

Xueying Chen and Min-ge Xie. A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, pp. 1655–1684, 2014.

Yixin Chen, Guozhu Dong, Jiawei Han, Jian Pei, Benjamin W Wah, and Jianyong Wang. Regression cubes with lossless compression and aggregation. *IEEE Transactions on Knowledge and Data Engineering*, 18(12):1585–1599, 2006.

Victor Chernozhukov and Iván Fernández-Val. Inference for extremal conditional quantile models, with an application to market and birthweight risks. *The Review of Economic Studies*, 78(2): 559–589, 2011.

Miklos Csörgo and Pál Révész. *Strong approximations in probability and statistics*. Academic press, 1981.

David Deuber, Jinzhou Li, Sebastian Engelke, and Marloes H Maathuis. Estimation and inference of extremal quantile treatment effects for heavy-tailed distributions. *Journal of the American Statistical Association*, 119(547):2206–2216, 2024.

Rui Duan, Mary Regina Boland, Zixuan Liu, Yue Liu, Howard H Chang, Hua Xu, Haitao Chu, Christopher H Schmid, Christopher B Forrest, John H Holmes, et al. Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. *Journal of the American Medical Informatics Association*, 27(3):376–385, 2020.

John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 429–438. IEEE, 2013.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503. Springer, 2006.

Sébastien Gadat and Fabien Panloup. Optimal non-asymptotic analysis of the ruppert–polyak averaging stochastic algorithm. *Stochastic Processes and their Applications*, 156:312–348, 2023.

Jennifer Gillenwater, Matthew Joseph, and Alex Kulesza. Differentially private quantiles. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3713–3722. PMLR, 18–24 Jul 2021.

Jia Gu and Song Xi Chen. Distributed statistical inference under heterogeneity. *Journal of Machine Learning Research*, 24(387):1–57, 2023.

Zijian Guo, Xiudi Li, Larry Han, and Tianxi Cai. Robust inference for federated meta-learning. *Journal of the American Statistical Association*, pp. 1–16, 2025.

Larry Han, Jue Hou, Kelly Cho, Rui Duan, and Tianxi Cai. Federated adaptive causal estimation (face) of target treatment effects. *Journal of the American Statistical Association*, (just-accepted): 1–25, 2025.

Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Franccoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

Aijun Hu, Yuling Jiao, Yanyan Liu, Yueyong Shi, and Yuanshan Wu. Distributed quantile regression for massive heterogeneous data. *Neurocomputing*, 448:249–262, 2021.

Jiayu Huang, Mingqiu Wang, and Yuanshan Wu. Estimation and inference for transfer learning with high-dimensional quantile regression. *arXiv preprint arXiv:2211.14578*, 2022.

Qi Huang, Hanze Zhang, Jiaqing Chen, and MJJBB He. Quantile regression models and their applications: A review. *Journal of Biometrics & Biostatistics*, 8(3):1–6, 2017.

Rong Jiang and Keming Yu. Smoothing quantile regression for a distributed system. *Neurocomputing*, 466:311–326, 2021.

Matthew Joseph, Jieming Mao, Seth Neel, and Aaron Roth. The role of interactivity in local differential privacy. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 94–105, 2019. doi: 10.1109/FOCS.2019.00015.

Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. Localized debiased machine learning: Efficient inference on quantile treatment effects and beyond. *Journal of Machine Learning Research*, 25 (16):1–59, 2024.

Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

Jason D Lee, Qiang Liu, Yuekai Sun, and Jonathan E Taylor. Communication-efficient sparse regression. *Journal of Machine Learning Research*, 18(5):1–30, 2017.

Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin. Fast and robust online inference with stochastic gradient descent via random scaling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7381–7389, 2022.

Runze Li, Dennis KJ Lin, and Bing Li. Statistical inference in massive data sets. *Applied Stochastic Models in Business and Industry*, 29(5):399–409, 2013.

Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):149–173, 2022a.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020.

Ting Li, Chengchun Shi, Qianglin Wen, Yang Sui, Yongli Qin, Chunbo Lai, and Hongtu Zhu. Combining experimental and historical data for policy evaluation. In *Forty-first International Conference on Machine Learning*, 2023.

Xiang Li, Jiadong Liang, Xiangyu Chang, and Zhihua Zhang. Statistical estimation and online inference via local sgd. In *Conference on Learning Theory*, pp. 1613–1661. PMLR, 2022b.

Yi Liu, Qirui Hu, Lei Ding, and Linglong Kong. Online local differential private quantile inference via self-normalization. In *International Conference on Machine Learning*, pp. 21698–21714. PMLR, 2023.

Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125. IEEE, 2008.

Anh Nguyen, Tuong Do, Minh Tran, Binh X Nguyen, Chien Duong, Tu Phan, Erman Tjiputra, and Quang D Tran. Deep federated learning for autonomous driving. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1824–1830. IEEE, 2022.

Kriti Pillutla, Yannis Laguel, Jerome Malick, and Zaid Harchaoui. Federated learning with superquantile aggregation for heterogeneous data. *Machine Learning*, 2024. doi: 10.1007/s10994-023-06332-x.

Drago Plečko, Nicolas Bennett, and Nicolai Meinshausen. fairadapt: Causal reasoning for fair data preprocessing. *Journal of Statistical Software*, 110:1–35, 2024.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

Robert J Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, New Jersey, 2009.

Jianwei Shi, Yue Wang, Zhongyi Zhu, and Heng Lian. Decentralized learning of quantile regression: A smoothing approach. *Journal of Computational and Graphical Statistics*, pp. 1–11, 2025.

Xu Shi, Ziyang Pan, and Wang Miao. Data integration in causal inference. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(1):e1581, 2023.

Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE global conference on signal and information processing*, pp. 245–248. IEEE, 2013.

Keng Meng Tan, Heather Battey, and Wen-Xin Zhou. Communication-constrained distributed quantile regression with optimal statistical guarantees. *Journal of Machine Learning Research*, 23(99):1–53, 2022. URL `https://www.jmlr.org/papers/volume23/21-1223/21-1223.pdf`.

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

Caixing Wang and Ziliang Shen. Distributed high-dimensional quantile regression: estimation efficiency and support recovery. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 51415–51441, 2024.

Xiaozhou Wang, Zhuoyi Yang, Xi Chen, and Weidong Liu. Distributed inference for linear support vector machine. *Journal of machine learning research*, 20(113):1–41, 2019.

Chuhan Xie, Kaicheng Jin, Jiadong Liang, and Zhihua Zhang. Asymptotic time-uniform inference for parameters in averaged stochastic approximation. *arXiv preprint arXiv:2410.15057*, 2024.

Steve Yadlowsky, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager. Evaluating treatment prioritization rules via rank-weighted average treatment effects. *Journal of the American Statistical Association*, 120(549):38–51, 2025.

Wanrong Zhu, Zhipeng Lou, Ziyang Wei, and Wei Biao Wu. High confidence level inference is almost free using parallel stochastic optimization. *arXiv preprint arXiv:2401.09346*, 2024.

Xuening Zhu, Feng Li, and Hansheng Wang. Least-square approximation for a distributed system. *Journal of Computational and Graphical Statistics*, 30(4):1004–1018, 2021.

A Additional Discussion and Results

Algorithm A.1 Locally Randomized Compare (Liu et al., 2023)

Require: Inquiry θ, response rate r, private data x
Ensure: A randomized binary response
1: Sample $u \sim \text{Bernoulli}(r)$
2: Sample $v \sim \text{Bernoulli}(0.5)$
3: **if** $u = 1$ **then**
4: **return** $\mathbf{1}_{\theta > x}$
5: **else**
6: **return** v
7: **end if**

Algorithm A.2 Privacy-Aware Quantile Inference via Data Integration

Input: Learning rates $\{\eta_{k,t}\}$, sample sizes $\{n_k\}$, number of chains $\{M_k\}$, target quantile τ, truthful response rates $\{r_k\}$, tuning parameter λ, and significance level α.
Output: $\widehat{\theta}_{\text{est}}$ and $[\widehat{\theta}_{\text{est}} - \mathcal{Z}_{\alpha/2}\widehat{\sigma}_{\text{est}}/\sqrt{N}, \widehat{\theta}_{\text{est}} + \mathcal{Z}_{\alpha/2}\widehat{\sigma}_{\text{est}}/\sqrt{N}]$.
Initialization: set $\widehat{\theta}_{k,0}^{(m)} \leftarrow 0$ for all k, m.
for $0 \le k \le K$ **do**
for $1 \le m \le M_k$ **do**
for $1 \le t \le \lfloor n_k/M_k \rfloor$ **do**
Obtain the locally randomizer $\widehat{\zeta}_{k,t}^{(m)} = \text{LRC}\left(\widehat{\theta}_{k,t}^{(m)}, r_k, X_{k,t+1}^{(m)}\right)$ using Algorithm A.1.
Compute $\widehat{\theta}_{k,t}^{(m)}$ according to equation (3.6).
end for
Compute the chain-specific estimator by $\widehat{\theta}_k^{(m)} = (\lfloor n_k/M_k \rfloor)^{-1} \sum_{t=1}^{\lfloor n_k/M_k \rfloor} \widehat{\theta}_{k,t}^{(m)}$.
end for
Compute the final estimator and corresponding variance estimator at site k by
$\widehat{\theta}_k = M_k^{-1} \sum_{m=1}^{M_k} \widehat{\theta}_k^{(m)}, \quad \widehat{\sigma}_k^2 = \frac{1}{M_k - 1} \sum_{m=1}^{M_k} \lfloor n_k/M_k \rfloor (\widehat{\theta}_k^{(m)} - \widehat{\theta}_k)^2$.
end for
Compute the weighted estimator and its variance estimator by $\widehat{\theta}_{\text{est}} = \sum_{k=0}^{K} \widehat{w}_k \widehat{\theta}_k$ and
$\widehat{\sigma}_{\text{est}}^2 = \sum_{k=0}^{K} N \widehat{w}_k \widehat{\sigma}_k^2 / n_k$, where $\widehat{w}_k = \left\{ \sum_{j=0}^{K} \left(\frac{\widehat{\sigma}_j^2}{n_j} + \lambda \widehat{b}_j^2 \right)^{-1} \right\}^{-1} \left(\frac{\widehat{\sigma}_k^2}{n_k} + \lambda \widehat{b}_k^2 \right)^{-1}$ and
$N = \sum_k n_k$.

A.1 Additional Discussion

Lasso penalty. In this paper, we utlize an ℓ_2 penalty to regularize parameter shifts between the target and auxiliary sites, resulting in an algorithm that is easy to implement and has well-established theoretical properties. However, alternative penalties, such as the ℓ_1 penalty, can also be employed within our framework. Specifically, under an ℓ_1 penalty, the corresponding loss function with respect to the weights $\mathbf{w}$ can be expressed as:

$$\mathcal{L}_{\text{lasso}}(\mathbf{w}) = \sum_{k=0}^{K} w_k^2 \frac{\widehat{\sigma}_k^2}{n_k} + \lambda \sum_{k=0}^{K} |w_k| \widehat{b}_k^2.$$

This optimization problem can be efficiently solved using standard algorithms, such as the alternating direction method of multipliers (ADMM). Let $\widehat{w}_k^{\text{lasso}}$ denote the resulting adaptive weights. The

final Lasso-weighted estimator is then given by: $\widehat{\theta}_{\text{lasso}} = \sum_{k=0}^{K} \widehat{w}_k^{\text{lasso}} \widehat{\theta}_k$. The variance estimator is given by $\widehat{\sigma}^2_{\text{lasso}} = \sum_{k=0}^{K} N \widehat{w}_k^{\text{lasso}} \widehat{\sigma}^2_k / n_k$. Under certain regularity conditions, we expect the Lasso-based estimator to exhibit analogous theoretical properties. However, similar to the adaptive weights derived from the ℓ_2 penalty, the Lasso method also cannot handle scenarios where the bias b_k is exactly of order $N^{-1/2}$. Our numerical experiments indicate that the performance of the Lasso method is similar to that of the non-conservative adaptive approach. We investigate its finite-sample performance under general bias scenarios and varying response rates with λ selected via cross-validation. the corresponding results are illustrated in Figures A.5 – A.10.

DP-SGD. As noted in Remark 1, beyond the proposed LDP mechanism in (3.4), one can also achieve LDP by directly adding noise to the local stochastic gradients (Song et al., 2013). Specifically, modify the update in (3.3) to

$$\widehat{\theta}_{k,t+1} = \widehat{\theta}_{k,t} - \eta_{k,t} \left\{ \tau - \mathbf{1}\left(X_{k,t+1} \leq \widehat{\theta}_{k,t} \right) + Z_{k,t+1} \right\},$$

where $Z_t^k \sim \text{Laplace}(0, b)$ with scale $b = 1/\log\{(1+r_k)/(1-r_k)\}$. We compare our method with DP-SGD. We evaluate DP-SGD under the same settings as in the left panel of Figure 2 in Section 5. The results in Figure A.1 show that DP-SGD yields larger log MSE and wider confidence intervals than our proposed methods.

Conservative variance estimator. Note that $\sigma_k^2 = 4r_k^2 f_k^2(\theta_k)^{-1}\{1 - r_k^2(2\tau-1)^2\}$ can be large when r_k is relatively small. Thus, the variance estimation might become unstable when the number of chains is limited. This inspires us to adopt a conservative variance estimator as well. Specifically, a conservative upper bound for σ_k^2 at significance level α is given by: $\widetilde{\sigma}_k^2 = (M_k - 1)\widehat{\sigma}_k^2 / \chi^2_{\alpha, M_k - 1}$, where χ^2_{α, M_k-1} denotes the α-quantile of the chi-squared distribution with $M_k - 1$ degrees of freedom. We study its performance under varying response rates, the results are summarized in Figures Results are summarized in Figures A.7 -– A.10.

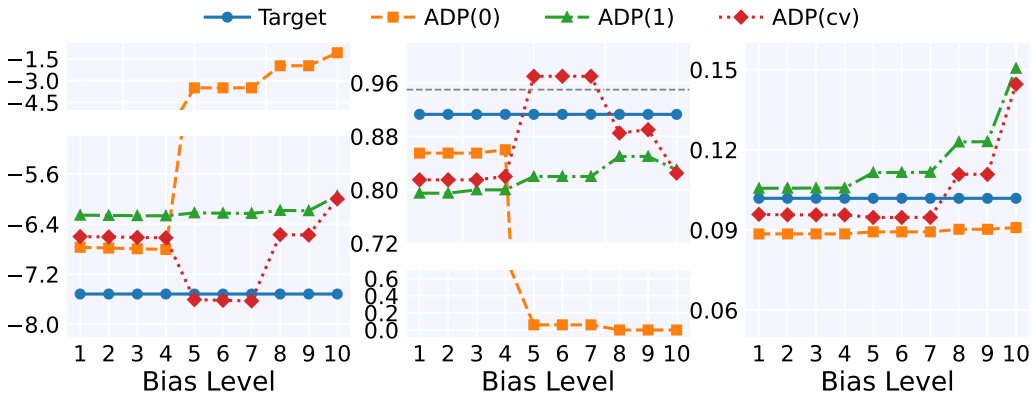


Figure A.1: Simulation results for DP-SGD method under a broader range of bias values, from $\exp(-5)$ to 1. The left, middle, and right panels present log MSE, ECP, and CI length, respectively. The target sample size is fixed at $n_0 = 20{,}000$, with the target quantile set to $\tau = 0.25$. Data are generated from Normal distributions with a fixed response rate of $r_k = 0.5$ across all sites.

A.2 Complete Experimental Details and Results

We provide here additional experimental details not described in the main text. For all experiments, we use 36 Intel(R) Xeon(R) Gold 6271 CPUs, equipped with a total of 128GB of RAM and 500GB of storage. The experiments are implemented using Python 3.12, and the computational time required to generate each figure is approximately 3 to 8 hours. The details of the LDP-based algorithm are provided in Algorithm A.1. The complete Algorithm of the proposed method is given in Algorithm A.2.

The three performance measures considered in our numerical studies are defined explicitly as follows. Let $\theta^{(r)}$ denote the estimated quantile obtained in the r-th replication, and let $\text{CI}^{(r)}$ represent

the corresponding 95% confidence interval (CI). We evaluate the estimators using the following metrics:

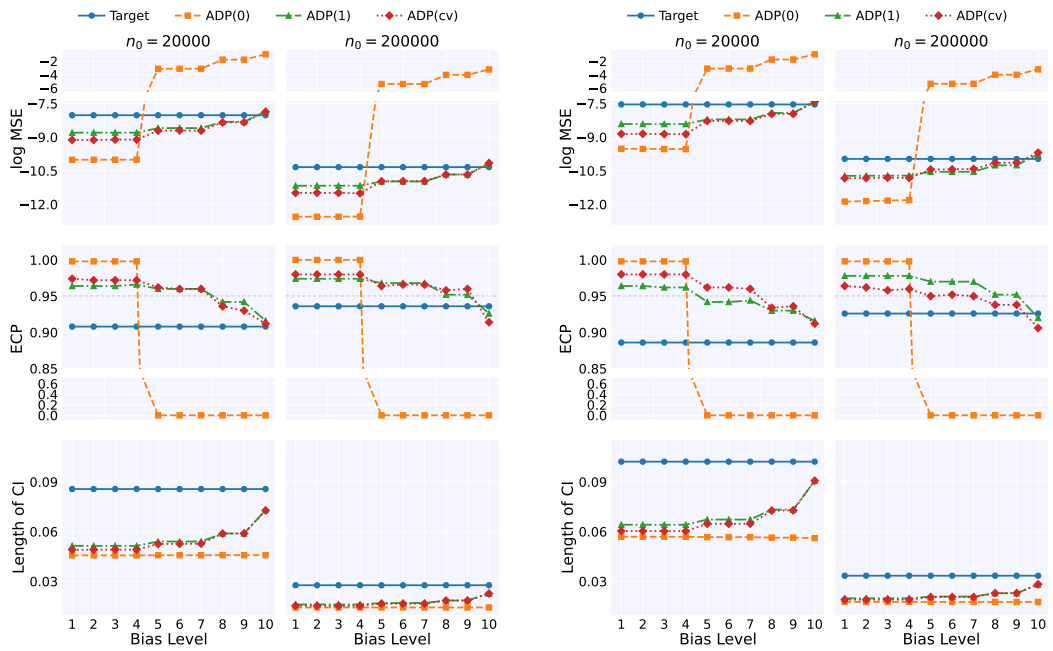
- **Mean Squared Error on the Log Scale (log MSE)**: $\log \text{MSE} = \log\left(R^{-1}\sum_{r=1}^{R}|\theta^{(r)}-\theta_0|^2\right)$,
- **Empirical Coverage Probability (ECP)**: $\text{ECP} = R^{-1}\sum_{r=1}^{R}\mathbf{1}\{\theta_0 \in \text{CI}^{(r)}\}$.
- **Average Confidence Interval Length (CI length)**: $\text{CI length} = R^{-1}\sum_{r=1}^{R}|\,\text{CI}^{(r)}\,|$.

Level	Description of source sites	Biases $b_k,\ k=1,\ldots,3$
1	All source sites unbiased	$b_1=b_2=b_3=0$
2	One weakly biased source	$b_1=b_2=0,\ b_3=0.1/\sqrt{n_0}$
3	Two weakly biased sources	$b_1=0,\ b_2=b_3=0.1/\sqrt{n_0}$
4	Three weakly biased sources	$b_1=b_2=b_3=0.1/\sqrt{n_0}$
5	One strongly biased source	$b_1=b_2=0,\ b_3=100/\sqrt{n_0}$
6	One weak + one strong	$b_1=0,\ b_2=0.1/\sqrt{n_0},\ b_3=100/\sqrt{n_0}$
7	Two weak + one strong	$b_1=b_2=0.1/\sqrt{n_0},\ b_3=100/\sqrt{n_0}$
8	Two strongly biased sources	$b_1=0,\ b_2=b_3=100/\sqrt{n_0}$
9	One weak + two strong	$b_1=0.1/\sqrt{n_0},\ b_2=b_3=100/\sqrt{n_0}$
10	All source sites strongly biased	$b_1=b_2=b_3=100/\sqrt{n_0}$

Table A.1: Bias levels for source sites. Target site has $b_0=0$.

Next, we present additional experimental results not shown in the main text. First, for scenarios where the bias is either vanishing or distinguishable, further results are shown in Figures A.2–A.4. Second, for the general bias scenario, additional results are in Figures A.4 – A.6. We observe that these additional results are qualitatively consistent with those presented in the main text. Furthermore, we investigate the effect of different response rates on estimation performance. For this analysis, we set $\mu_k=0$ for $0\le k\le 3$. For the target data, we consider both a scenario with no privacy protection ($r_0=1$) and scenarios with strong privacy protection ($r_0=0.25$ for normal data and $r_0=0.4$ for Cauchy data). Here, we also evaluate the performance of the conservative variance estimation (ADP(consvar)) introduced in Appendix A.1.The corresponding weights $\widehat{w}_k^{\text{cvar}}$ are computed by replacing σ_k^2 and b_k in (3.5) with $\widetilde{\sigma}_k^2$ and $\widehat{b}_k$, respectively. The resulting point estimator and variance estimator are then given by: $\widehat{\theta}_{\text{cvar}}=\sum_{k=0}^{K}\widehat{w}_k^{\text{cvar}}\widehat{\theta}_k$, and $\widehat{\sigma}^2_{\text{cvar}}=\sum_{k=0}^{K}N\widehat{w}_k^{\text{cvar}}\widetilde{\sigma}_k^2/n_k$. In our experiments, we set $\alpha=10^{-4}$. For the source sites, we vary the response rates (r_k) from r_0 to 0.9, while all other settings remain identical to previous experiments. Results are summarized in Figures A.7 -– A.10. For the scenario without privacy protection for the target data (Figures A.7 and A.8), we find that all estimators enhance estimation and inference performance compared to the Target when the source sites have relatively high response rates. However, these estimators demonstrate limited improvement over Target when r_k is relatively low, primarily because the limited number of chains results in imprecise variance estimation under inherently large variance conditions. In contrast, the conservative variance estimator consistently outperforms Target across all considered response rates, highlighting its robustness, especially at lower response rates. In addition, for scenarios where the target data has strong privacy protection (Figures A.9 and A.10), almost all estimators improve performance over Target when $r_k > r_0$.

Finally, we report (i) sensitivity analysis of the learning rate parameter β, and (ii) evaluations of finite-sample performance under smaller number of chains ($M_0=6$) and varying numbers of sites K. Detailed results are presented in Figure A.11 and Table A.2. In these simulations, the target sample size is fixed at $n_0=20{,}000$, and the data are generated from Normal distributions with a fixed response rate of $r_k=0.5$ for all sites. The results demonstrate the robustness of our proposed method.

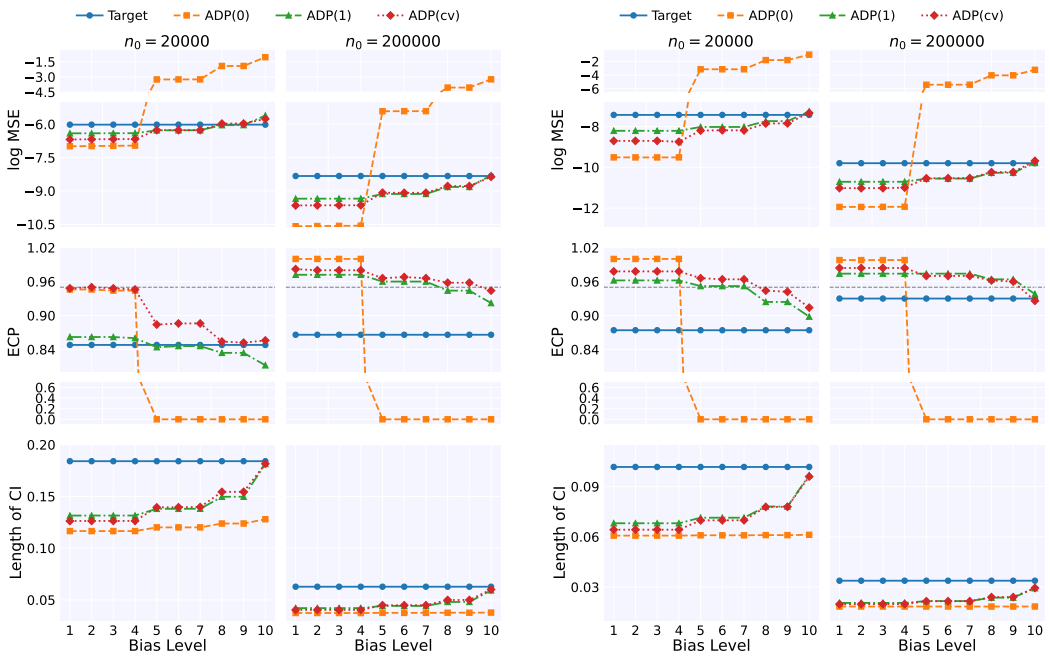


((i)) $\tau = 0.5$ ((ii)) $\tau = 0.75$

Figure A.2: Simulation results under scenarios of either vanishing or distinguishable bias. Bias levels range from complete homogeneity to strong heterogeneity. Results are reported for various target sample sizes (n_0). Data are generated from Normal distributions, with the response rate fixed at $r_k = 0.5$ for all sites.

β	Method	0.007	0.011	0.018	0.030	0.135	0.223	0.368	0.607	1.000
0.65	ADP(0)	99.5	98.7	87.3	39.0	0.0	0.0	0.0	0.0	0.0
	ADP(1)	96.9	96.2	93.6	85.8	88.2	91.0	92.0	92.1	92.3
	ADP(cons)	94.0	94.2	93.8	93.2	91.2	91.7	92.0	92.2	92.4
	ADP(cv)	97.1	95.8	90.1	77.7	90.6	91.1	90.8	90.8	91.3
0.70	ADP(0)	99.6	98.8	90.9	51.9	0.0	0.0	0.0	0.0	0.0
	ADP(1)	96.1	95.6	94.0	87.5	87.4	90.6	91.8	91.8	91.7
	ADP(cons)	94.0	93.9	93.6	92.9	90.9	91.8	91.7	91.7	91.7
	ADP(cv)	97.1	95.8	90.1	77.7	90.6	91.1	90.8	90.8	91.3
0.75	ADP(0)	99.6	99.1	95.7	71.5	0.0	0.0	0.0	0.0	0.0
	ADP(1)	95.6	95.6	94.5	89.9	85.3	89.1	90.5	91.3	91.5
	ADP(cons)	93.8	93.6	93.6	93.0	90.0	90.2	90.6	91.3	91.4
	ADP(cv)	97.1	95.8	90.1	77.7	90.6	91.1	90.8	90.8	91.3

Table A.2: Empirical coverage probability across a broader range of bias values, from $\exp(-5)$ to 1, under varying learning-rate parameters β. The target sample size is fixed at $n_0 = 20{,}000$, with the target quantile set to $\tau = 0.5$. Data are generated from Normal distributions with a fixed response rate of $r_k = 0.5$ across all sites.



((i)) $\tau = 0.25$ ((ii)) $\tau = 0.5$

Figure A.3: Simulation results under scenarios of either vanishing or distinguishable bias. Bias levels range from complete homogeneity to strong heterogeneity. Results are reported for various target sample sizes (n_0). Data are generated from Cauchy distributions, with the response rate fixed at $r_k = 0.5$ for all sites.

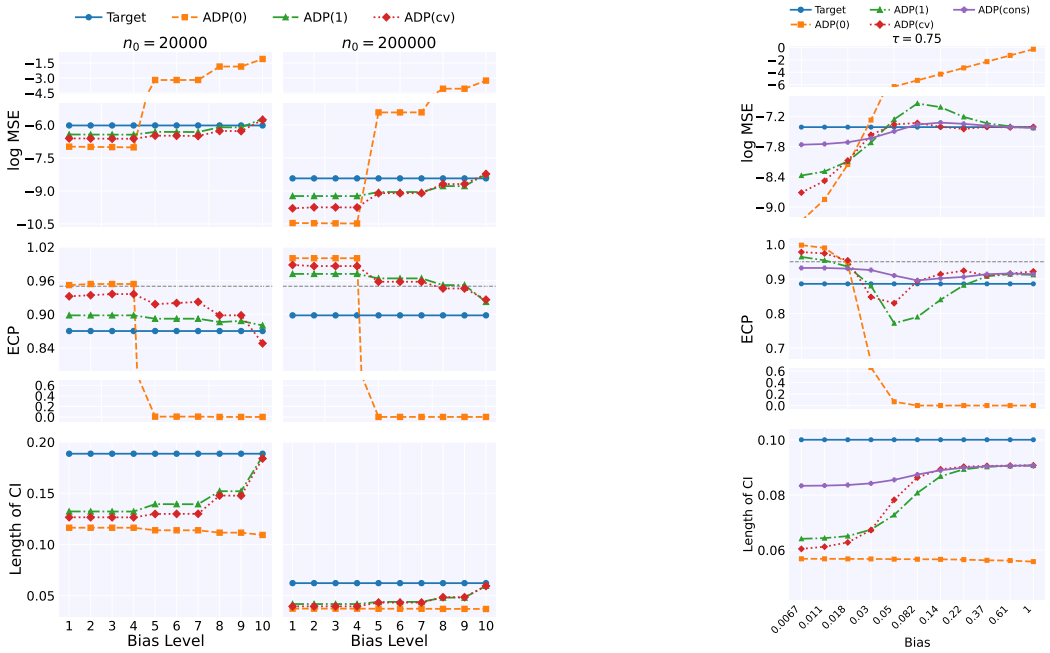


Figure A.4: Simulation results under different heterogeneity scenarios are presented. The left panel considers scenarios of either vanishing or distinguishable bias. Bias levels range from complete homogeneity to strong heterogeneity. Results are reported for various target sample sizes (n_0). Data are generated from Cauchy distributions. The quantile level is fixed at $\tau = 0.75$. The right panel considers a broader range of bias values from $\exp(-5)$ to 1. The target sample size is fixed at $n_0 = 20,000$. Results are shown for various quantile levels. Data are generated from Normal distributions. Both panels use a fixed response rate of $r_k = 0.5$ for all sites.

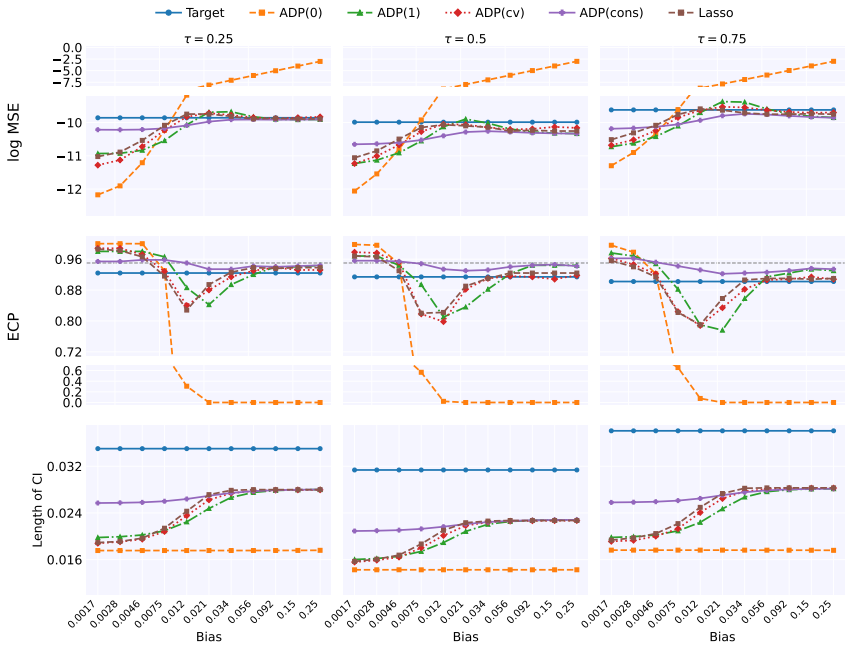


Figure A.5: Simulation results under a broader range of bias values from $\exp(-5)/4$ to 0.25. The target sample size is fixed at $n_0 = 200,000$. Results are shown for various quantile levels. Data are generated from Normal distributions with a fixed response rate of $r_k = 0.5$ for all sites.

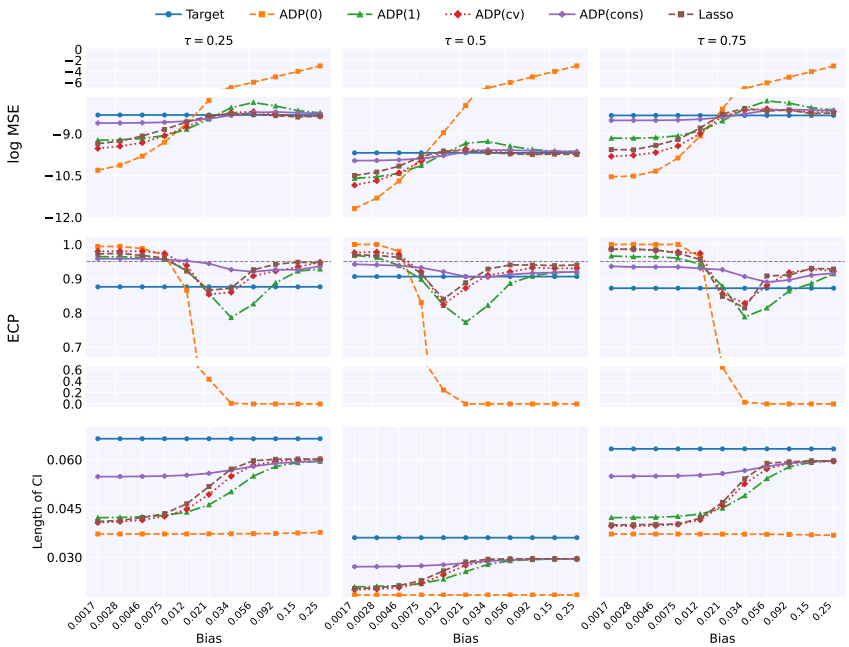


Figure A.6: Simulation results under a broader range of bias values from $\exp(-5)/4$ to 0.25. The target sample size is fixed at $n_0 = 200,000$. Results are shown for various quantile levels. Data are generated from Cauchy distributions with a fixed response rate of $r_k = 0.5$ for all sites.

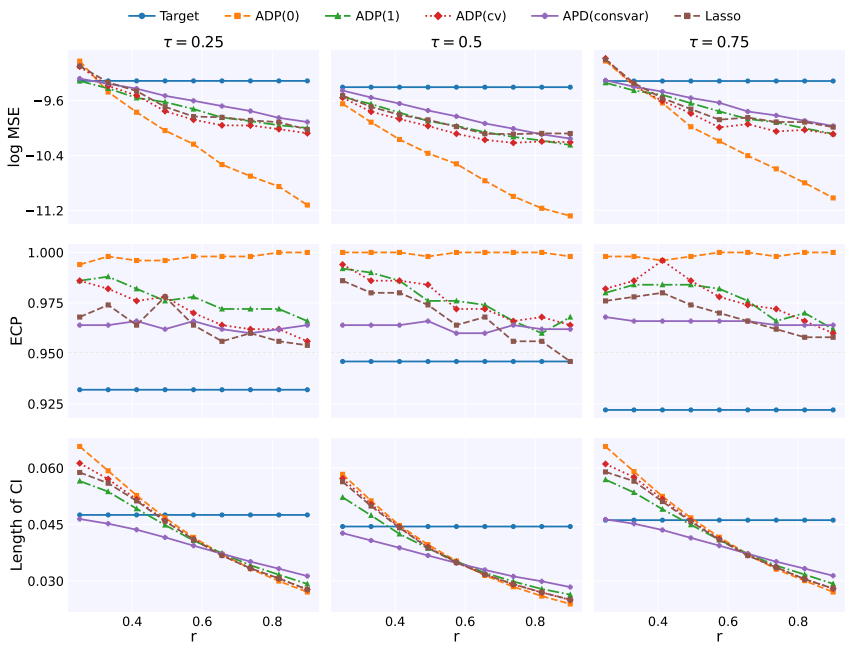


Figure A.7: Simulation results are presented across varying response rates and different quantile levels. Results are reported under the Normal distribution with $r_0 = 1$, $n_0 = 20,000$, and fixed $b_k = 0$ for all sites.

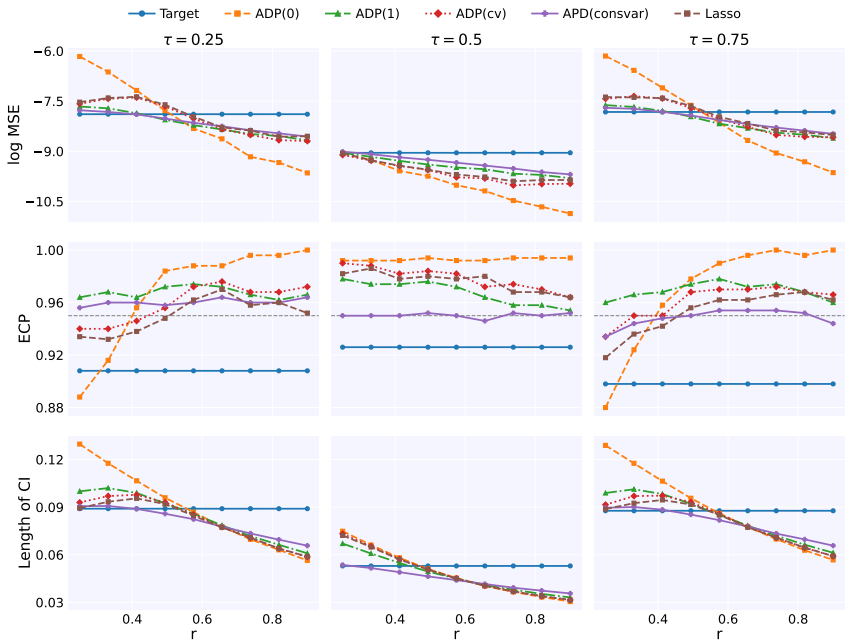


Figure A.8: Simulation results are presented across varying response rates and different quantile levels. Results are reported under the Cauchy distribution with $r_0 = 1$, $n_0 = 20,000$, and fixed $b_k = 0$ for all sites.

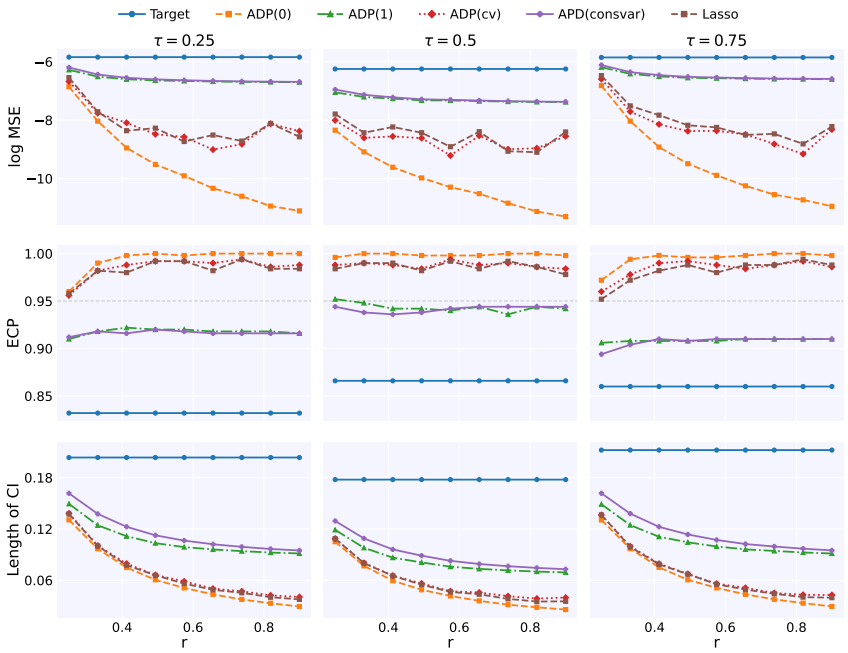


Figure A.9: Simulation results are presented across varying response rates and different quantile levels. Results are reported under the normal distribution with $r_0 = 0.25$, $n_0 = 20,000$, and fixed $b_k = 0$ for all sites.

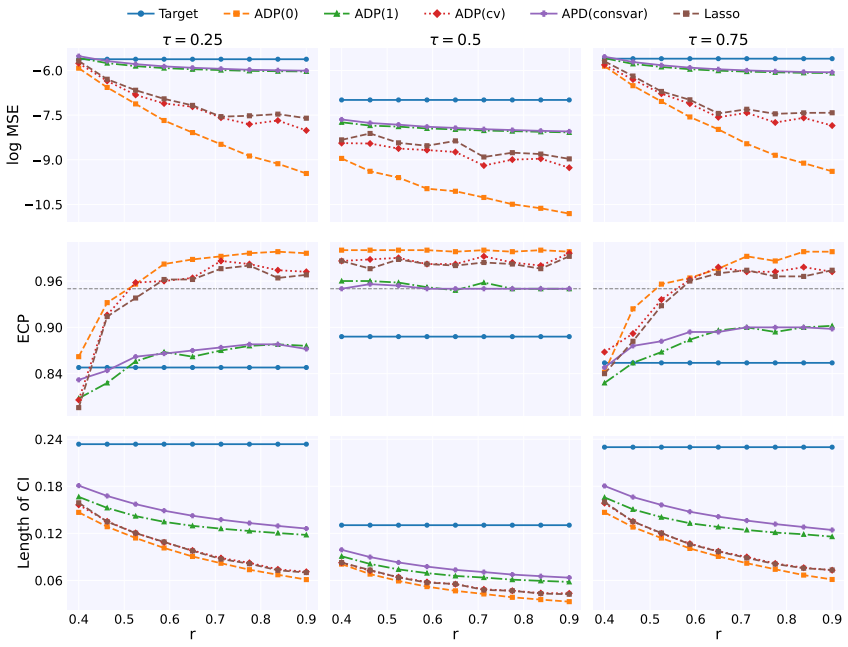


Figure A.10: Simulation results are presented across varying response rates and different quantile levels. Results are reported under the Cauchy distribution with $r_0 = 0.4$, $n_0 = 20,000$, and fixed $b_k = 0$ for all sites.

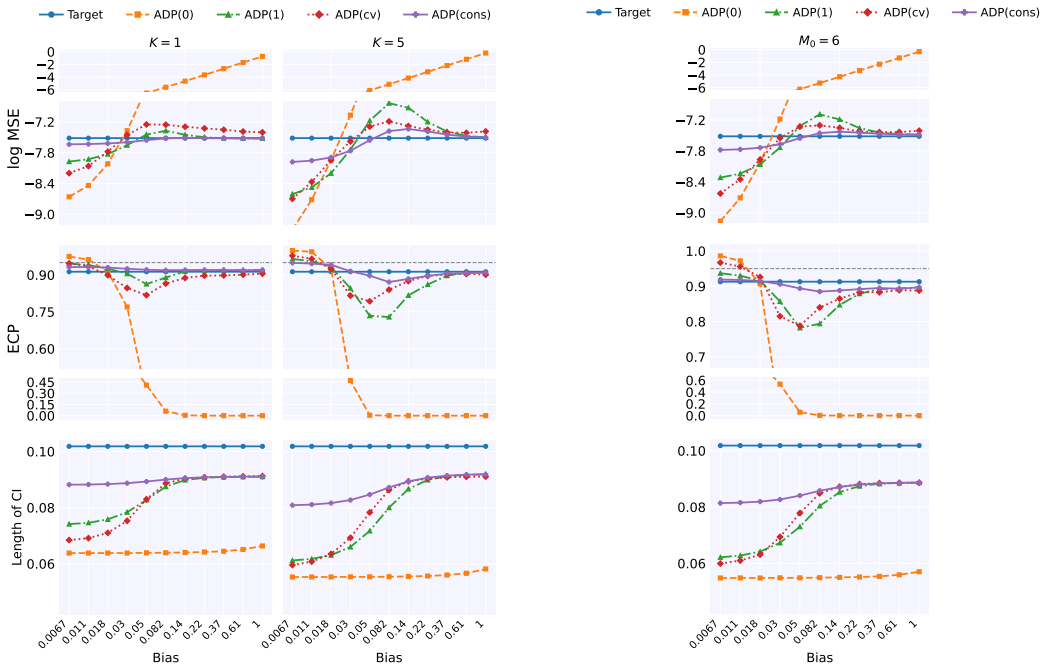


Figure A.11: Simulation results across a broader range of bias values from $\exp(-5)$ to 1, under varying K (left panel) and smaller M_0 (right panel). The target sample size is fixed at $n_0 = 20{,}000$, with the target quantile set to $\tau = 0.25$. Data are generated from Normal distributions with a fixed response rate of $r_k = 0.5$ across all sites.

A.3 REAL DATA

We evaluate our method on a real-world dataset widely employed in privacy research: the Government Salary Dataset (Plečko et al., 2024). This dataset is derived from the 2018 American Community Survey conducted by the U.S. Census Bureau and contains over 200,000 records with annual income (USD) as the response. Because income is sensitive personal financial information (Gillenwater et al., 2021), we treat it as privacy-protected data. To reflect the dataset's geographic structure, we partitioned the data by "economic region", treating each region as a site. The original data include nine regions: we select one region (sample size 27,387) as the target site and use three larger regions (Southeast, Far West, and Mideast) as source sites. For all sites, we fix the number of chains at $M_k = 10$, all other hyperparameters follow Section 5. We log-transform the response for estimation and back-transform all reported quantities.

We target quantile levels $\tau \in \{0.25, 0.5\}$ and consider two privacy settings: (1) a homogeneous truthful response rate r for all sites, and (2) heterogeneous rates uniformly distributed on $[0.7, 1.0]$. We report point estimates (Est) and confidence-interval lengths (CI Len) for our two main approaches, ADP(cv) and ADP(cons), and include the target-only estimator for reference. The results are summarized in Table A.3. As expected, the target-only intervals are the longest, while ADP(cv) yields the shortest. When privacy constraints are relaxed, intervals for both ADP(cv) and ADP(cons) become shorter, consistent with our simulations. In most cases, the target quantiles fall within our proposed intervals, indicating strong practical performance on real data.

	$\tau = 0.25$			$\tau = 0.5$		
Metric	ADP(cv)	ADP(cons)	Target	ADP(cv)	ADP(cons)	Target
			$r_k = 0.7$			
Est	26815	23721	27442	46358	43403	45374
CI Len	1815	1910	3014	1435	1459	1950
			hetero r_k			
Est	28495	23724	27442	48089	44620	45374
CI Len	1302	1896	3014	1140	1410	1950

Table A.3: Real-data analysis under different target quantile and truthful response rates. The number of chains per site is fixed at $M_k = 10$. $r_k = 0.7$ denotes a common truthful-response rate across sites, while "hetero r_k" denotes site-specific rates range from 0.7 to 1.0.

B TECHNICAL LEMMAS

We first introduce some notation used throughout the paper. For two positive sequences a_n and b_n, write $a_n \ll b_n$ or $a_n = \mathrm{o}(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$. Similarly, write $a_n \lesssim b_n$ or $a_n = O(b_n)$ if there exists a constant $C < \infty$ such that $a_n/b_n \le C$ for all sufficiently large n. Moreover, write $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold simultaneously. For two random sequences $\{X_n\}$ and $\{Y_n\}$, write $X_n = \mathrm{o}_p(Y_n)$ if $P(|X_n/Y_n| > \varepsilon) \to 0$ for any $\varepsilon > 0$. Write $X_n = O_p(Y_n)$ if for every $\varepsilon > 0$, there exists a constant $M > 0$ such that $\limsup_{n\to\infty} P\big(|X_n|/|Y_n| > M\big) < \varepsilon$. Write $X_n \xrightarrow{d} X$ if the random sequence X_n converges in distribution to a random variable X. Finally, $\lfloor x \rfloor$ denotes the largest integer less than or equal to x.

Lemma B.1. *Denote $Z_k^{(m)} = \sqrt{(n_k/M_k)}(\widehat{\theta}_k^{(m)} - \theta_k)$. We have*

$$\mathbb{E}(Z_k^{(m)})^2 - \sigma_k^2 = \mathcal{O}\Big(1/(n_k/M_k)^{(\beta-1/2)\wedge(1-\beta)}\Big).$$

$$\mathbb{E}(Z_k^{(m)}) = \mathcal{O}\Big(1/(n_k/M_k)^{(\beta/2-1/4)\wedge(1/2-\beta/2)}\Big).$$

Proof. The proof is shown in Corollary 6 in Gadat & Panloup (2023). □

Lemma B.2. *For each $0 \le k \le K$, define the weighted estimator as $\widehat{\theta}(\mathbf{w}) = \sum_{k=0}^{K} w_k \widehat{\theta}_k$, let $\theta^*(\mathbf{w}) = \sum_{k=0}^{K} w_k \theta_k$ and $\sigma^2(\mathbf{w}) = \sum_{k=0}^{K} \frac{N}{n_k} w_k^2 \sigma_k^2$ represent its corresponding true parameter*

and variance, respectively. Then under assumption 4, we have:

$$\sqrt{N}\left(\widehat{\theta}(\mathbf{w})-\theta^*(\mathbf{w})\right)\xrightarrow{d}\mathcal{N}\left(0,\sigma^2(\mathbf{w})\right).$$

Proof. Lemma B.2 follows naturally as a corollary of Theorem 4.1. □

Lemma B.3. *For every* $0\le k\le K$, *we have*

$$\mathbb{E}|\widehat{b}_k-b_k| \lesssim \frac{1}{\sqrt{N}} \tag{B.1}$$

$$\mathbb{E}|\widehat{b}_k^2-b_k^2| \lesssim \frac{1}{N}+|b_k|\frac{1}{\sqrt{N}} \tag{B.2}$$

Proof. Denote $Z_k^{(m)}=\sqrt{(n_k/M_k)}(\widehat{\theta}_k^{(m)}-\theta_k)$, $\bar{Z}_k=M_k^{-1}\sum_{m=1}^{M_k}Z_k^{(m)}$ for every $0\le k\le K$.

(i): Proof of equation (B.1)**.** Recall that $b_k=\theta_k-\theta_0$ and $\widehat{b}_k=\widehat{\theta}_k-\widehat{\theta}_0$, with $\widehat{\theta}_k=\frac{1}{M_K}\sum_{m=1}^{M_K}\widehat{\theta}_k^{(m)}$. First, note that

$$\mathbb{E}|\widehat{b}_k-b_k|\lesssim\sqrt{\mathbb{E}|\widehat{b}_k-b_k|^2}.$$

It then suffices to analysize $\mathbb{E}|\widehat{b}_k-b_k|^2$. It could be proved that under Assumption 3

$$\begin{aligned}\mathbb{E}|\widehat{b}_k-b_k|^2 &\lesssim \mathbb{E}\big(\widehat{\theta}_0-\theta_0\big)^2+\mathbb{E}\big(\widehat{\theta}_k-\theta_k\big)^2=\frac{M_0}{n_0}\big\{\mathbb{E}^2(\bar{Z}_0)+\mathrm{var}(\bar{Z}_0)\big\}\\ &\quad+\frac{M_k}{n_k}\big\{\mathbb{E}^2(\bar{Z}_k)+\mathrm{var}(\bar{Z}_k)\big\}\lesssim\frac{M_0}{n_0}\Big(1/(n_0/M_0)^{(\beta-1/2)\wedge(1-\beta)}\Big)\\ &\quad+\frac{M_k}{n_k}\Big(1/(n_k/M_k)^{(\beta-1/2)\wedge(1-\beta)}\Big)+\frac{1}{N}\lesssim\frac{1}{N}.\end{aligned}$$

This yields $\mathbb{E}|\widehat{b}_k-b_k|\lesssim 1/\sqrt{N}$.

(ii): Proof of equation (B.2)**.** It could be verified that

$$\begin{aligned}\mathbb{E}|\widehat{b}_k^2-b_k^2| &\le \mathbb{E}|\widehat{b}_k-b_k|^2+2b_k\mathbb{E}|\widehat{b}_k-b_k|\lesssim\mathbb{E}|\widehat{b}_k-b_k|^2+|b_k|\sqrt{\mathbb{E}|\widehat{b}_k-b_k|^2}\\ &\lesssim \frac{1}{N}+\frac{|b_k|}{\sqrt{N}}.\end{aligned}$$

This finishes the whole Lemma proof. □

Lemma B.4 (Consistency)**.** *Under Assumption 4, consider the following three scenarios: (a)* $\lambda=0$*, and the bias scale satisfies Assumption 5 (1); (b)* λ *is bounded away from* 0*, and the bias scale satisfies Assumption 5 (2); (c)* λ *satisfies Assumption 6, and the bias scale satisfies Assumption 5. Then, for each* $0\le k\le K$*, with probability* $1-o(1)$*, we have*

$$|\widehat{w}_k-w_k^*|\lesssim\left\{\sum_{k=0}^{K}\prod_{j\neq k}\Big(\frac{\sigma_j^2}{N}+\lambda b_j^2\Big)\right\}^{-1}\left\{\sum_{k=0}^{K}\sum_{i\neq k}\Big(\frac{\lambda|b_i|}{\sqrt{N}}+\frac{o(1)}{N}\Big)\prod_{j\neq i}^{j\neq k}\Big(\frac{\sigma_j^2}{N}+\lambda b_j^2\Big)\right\}=o(1).$$

C Proof of the Main Theoretical Results

Without loss of generality, we simply assume that $\lfloor n_k/M_k\rfloor=n_k/M_k=:T_k$.

C.1 Proof of Theorem 4.1

First, we have $\mathbb{E}|\widehat{b}_k^2-b_k^2|\lesssim 1/N+|b_k|/\sqrt{N}$ in Lemma B.3. Next, for any fixed $k=0,\dots,K$, one rewrites

$$\widehat{\theta}_k-\theta_k=\frac{1}{M_k}\sum_{m=1}^{M_k}\frac{1}{T_k}\sum_{t=1}^{T_k}(\widehat{\theta}_{k,t}^{(m)}-\theta_k)=:\frac{1}{M_k}\sum_{m=1}^{M_k}\mathcal{T}_k^{(m)},$$

where

$$\mathcal{T}_k^{(m)} = \widehat{\theta}_k^{(m)} - \theta_k = \frac{1}{T_k}\sum_{t=1}^{T_k}(\widehat{\theta}_{k,t}^{(m)} - \theta_k).$$

Let $U_{k,t}^{(m)}$ and $V_{k,t}^{(m)}$ be i.i.d. Bernoulli variables, mutually independent and also independent of $X_{k,t}^{(m)}$, with

$$\mathbb{P}(U_{k,t}^{(m)} = 1) = r_k, \quad \mathbb{P}(U_{k,t}^{(m)} = 0) = 1 - r_k, \qquad \mathbb{P}(V_{k,t}^{(m)} = 1) = \mathbb{P}(V_{k,t}^{(m)} = 0) = \tfrac{1}{2}.$$

Then it could be found that equation (3.6) could be rewritten as

$$\widehat{\theta}_{k,t+1}^{(m)} = \widehat{\theta}_{k,t}^{(m)} - \eta_{k,t} G_k(\widehat{\theta}_{k,t}^{(m)}, \zeta_{k,t+1}^{(m)}),$$

where $\zeta_{k,t}^{(m)} = \left(X_{k,t}^{(m)}, U_{k,t}^{(m)}, V_{k,t}^{(m)}\right)^\top$, and

$$\begin{aligned} G_k(\theta, \zeta_{k,t}^{(m)}) = &\frac{1 + r_k - 2r_k\tau}{2}\left\{\mathbf{1}(X_{k,t}^{(m)} \le \theta)U_{k,t}^{(m)} + (1 - U_{k,t}^{(m)})(1 - V_{k,t}^{(m)})\right\} \\ &- \frac{1 - r_k + 2r_k\tau}{2}\left\{\mathbf{1}(X_{k,t}^{(m)} > \theta)U_{k,t}^{(m)} + (1 - U_{k,t}^{(m)})V_{k,t}^{(m)}\right\}. \end{aligned}$$

Define

$$\varepsilon_{k,t}^{(m)} = g_k(\widehat{\theta}_{k,t-1}^{(m)}) - G_k(\widehat{\theta}_{k,t-1}^{(m)}, \zeta_{k,t}^{(m)}), \;\; \widetilde{\varepsilon}_{k,t}^{(m)} = g_k(\theta_k) - G_k(\theta_k, \zeta_{k,t}^{(m)}).$$

Elementary calculation shows that

$$\begin{aligned} \mathbb{E}(\varepsilon_{k,t}^{(m)2} | \mathcal{F}_{k,t-1}^{(m)}) &= \frac{1 + r_k - 2r_k F_k(\widehat{\theta}_{k,t-1}^{(m)})}{2} - \left(\frac{1 + r_k - 2r_k F_k(\widehat{\theta}_{k,t-1}^{(m)})}{2}\right)^2 \\ &= \frac{1 - r_k^2\left\{2F_k(\widehat{\theta}_{k,t-1}^{(m)}) - 1\right\}^2}{4} \\ &\xrightarrow{\mathbb{P}} \frac{1 - r_k^2(2\tau - 1)^2}{4}, \end{aligned}$$

where the convergence in probability holds by the consistency of the quantile estimation and the continuous mapping theorem. Denote $\Delta_{k,t}^{(m)} = \widehat{\theta}_{k,t}^{(m)} - \theta_k$, $H_k = r_k f_k(\theta_k)$, $B_{k,t} = 1 - \eta_{k,t}H_k$, $A_{k,j}^t = \sum_{s=j}^t \left(\prod_{i=j+1}^s B_{k,i}\right)\eta_{k,i}$ for any $j \le t$. We decompose that

$$\begin{aligned} \mathcal{T}_k^{(m)} &= \frac{1}{T_k}\sum_{j=1}^{T_k}(\widehat{\theta}_{k,j}^{(m)} - \theta_k) = \frac{1}{T_k}\sum_{j=1}^{T_k}\Delta_{k,j}^{(m)} \\ &= \frac{1}{T_k}A_{k,0}^{T_k-1}B_{k,0}\Delta_{k,0} + \frac{1}{T_k}\sum_{j=0}^{T_k-1}A_{k,j}^{T_k-1}r_{k,j}^{(m)} + \frac{1}{T_k}\sum_{j=0}^{T_k-1}\left(A_{k,j}^{T_k-1} - H_k^{-1}\right)\varepsilon_{k,j+1}^{(m)} \\ &+ \frac{1}{T_k}\sum_{j=0}^{T_k-1}H_k^{-1}\left(\varepsilon_{k,j+1}^{(m)} - \widetilde{\varepsilon}_{k,j+1}^{(m)}\right) + \frac{1}{T_k}\sum_{j=0}^{T_k-1}H_k^{-1}\widetilde{\varepsilon}_{k,j+1}^{(m)} \\ &=: \mathcal{T}_{k,1}^{(m)} + \mathcal{T}_{k,2}^{(m)} + \mathcal{T}_{k,3}^{(m)} + \mathcal{T}_{k,4}^{(m)} + \mathcal{T}_{k,5}^{(m)}, \end{aligned}$$

in which

$$r_{k,j}^{(m)} = H_k(\widehat{\theta}_{k,j}^{(m)} - \theta_k) - g_k(\widehat{\theta}_{k,j}^{(m)}).$$

For $\mathcal{T}_{k,1}^{(m)}$: According to Lemma C.4 of Xie et al. (2024), one has $\left|A_{k,0}^{t-1}\right| \le C_0$ uniformly for all $t \ge 1$. Further observe that $\Delta_{k,0}^{(m)} \equiv \widehat{\theta}_{k,0} - \theta_k$ for all $1 \le m \le M_k$, thus one obtains

$$\left|\frac{1}{M_k}\sum_{m=1}^{M_k}\mathcal{T}_{k,1}^{(m)}\right| = \mathcal{O}\left(M_k/n_k\right).$$

For $\mathcal{T}_{k,2}^{(m)}$: Theorem 5 of Gadat & Panloup (2023) shows that

$$\max_{1\le m\le M_k}\mathbb{E}\left|\widehat{\theta}_{k,t}^{(m)}-\theta_k\right|^2\lesssim\eta_{k,t}.$$

According to the uniform boundedness of $\left|A_{k,j}^{t-1}\right|$ and the fact that $|r_{k,t}^{(m)}|\lesssim|\widehat{\theta}_{k,t}^{(m)}-\theta_k|^2$,

$$\begin{aligned}\mathbb{E}\left|\frac{1}{M_k}\sum_{m=1}^{M_k}\mathcal{T}_{k,2}^{(m)}\right|&\lesssim\mathbb{E}\frac{1}{M_k}\sum_{m=1}^{M_k}\frac{1}{T_k}\sum_{t=0}^{T_k-1}\left|r_{k,t}^{(m)}\right|\\&\le\frac{1}{M_k}\sum_{m=1}^{M_k}\frac{1}{T_k}\sum_{t=0}^{T_k-1}\mathbb{E}|\widehat{\theta}_{k,t}^{(m)}-\theta_k|^2\\&=\mathcal{O}\left((M_k/n_k)^a\right).\end{aligned}$$

Hence,

$$\left|\frac{1}{M_k}\sum_{m=1}^{M_k}\mathcal{T}_{k,2}^{(m)}\right|=\mathcal{O}_p\left((M_k/n_k)^a\right).$$

For $\mathcal{T}_{k,3}^{(m)}$: For any fixed $p>0$, note that $\max_{1\le m\le M_k}\mathbb{E}|\varepsilon_{k,j}^{(m)}|^{2p}=\mathbb{E}|\varepsilon_{1,j}^{(m)}|^{2p}$ is bounded. Following the arguments in Xie et al. (2024), one has $\left\|\mathcal{T}_{3,k}^{(m)}\right\|_{2p}=\mathcal{O}\left((n_k/M_k)^{-1+a/2}\right)$. Then, using the Lemma A in Chapter 9.2.6 of Serfling (2009) and the independence over m, one has that

$$\left\|\frac{1}{M_k}\sum_{m=1}^{M_k}\mathcal{T}_{k,3}^{(m)}\right\|_{2p}=\mathcal{O}\left(M_k^{-1/2}(n_k/M_k)^{-1+a/2}\right),$$

which implies that

$$\left|\frac{1}{M_k}\sum_{m=1}^{M_k}\mathcal{T}_{k,3}^{(m)}\right|=\mathcal{O}_p\left(M_k^{-1/2}(n_k/M_k)^{-1+a/2}\right).$$

For $\mathcal{T}_{k,4}^{(m)}$: Observe that $\varepsilon_{k,j}^{(m)}-\widetilde{\varepsilon}_{k,j}^{(m)}=g_k(\widehat{\theta}_{k,j-1}^{(m)})-G_k(\widehat{\theta}_{k,j-1}^{(m)},\zeta_{k,j}^{(m)})+G_k(\theta_k,\zeta_{k,j}^{(m)})$, and

$$\begin{aligned}\mathbb{E}\left(\varepsilon_{k,j}^{(m)}-\widetilde{\varepsilon}_{k,j}^{(m)}\right)^2&\lesssim\mathbb{E}g_k^2(\widehat{\theta}_{k,j-1}^{(m)})+\mathbb{E}\left\{G_k(\widehat{\theta}_{k,j-1}^{(m)},\zeta_{k,j})-G_k(\theta_k,\zeta_{k,j})\right\}^2\\&\lesssim\mathbb{E}\left|\widehat{\theta}_{k,j-1}^{(m)}-\theta_k\right|^2+\mathbb{E}\left|\widehat{\theta}_{k,j-1}^{(m)}-\theta_k\right|\\&\lesssim\mathbb{E}\left|\widehat{\theta}_{k,j-1}^{(m)}-\theta_k\right|^2+\left\{\mathbb{E}\left|\widehat{\theta}_{k,j-1}^{(m)}-\theta_k\right|^2\right\}^{1/2}\lesssim\eta_{k,t}^{1/2},\end{aligned}$$

where the last inequality holds by Theorem 5 of Gadat & Panloup (2023), and the constant does not depend on m.

Observe that $\sum_{t=1}^{T_k}H_k^{-1}\left(\varepsilon_{k,t}^{(m)}-\widetilde{\varepsilon}_{k,t}^{(m)}\right)$ is a martingale for each m (independent over $1\le m\le M_k$), Burkholder's inequality entails that

$$\begin{aligned}\left\|\sum_{t=1}^{T_k}H_k^{-1}\left(\varepsilon_{k,t}^{(m)}-\widetilde{\varepsilon}_{k,t}^{(m)}\right)\right\|_2&\lesssim\left\{\sum_{t=1}^{T_k}\left\|H_k^{-1}\left(\varepsilon_{k,t}^{(m)}-\widetilde{\varepsilon}_{k,t}^{(m)}\right)\right\|_2^2\right\}^{1/2}\\&\lesssim\left(\sum_{t=1}^{T_k}\eta_{k,t}^{1/2}\right)^{1/2}\lesssim(n_k/M_k)^{\{1-a/2\}/2}.\end{aligned}$$

Hence,

$$\left\|\mathcal{T}_{k,4}^{(m)}\right\|_2=\left\|\frac{1}{T_k}\sum_{j=1}^{T_k}H^{-1}\left(\varepsilon_{k,j}^{(m)}-\widetilde{\varepsilon}_{k,j}^{(m)}\right)\right\|_2=\mathcal{O}\left((n_k/M_k)^{-1/2-a/4}\right),$$

which further implies that

$$\frac{1}{M_k}\sum_{m=1}^{M_k}\mathcal{T}_{k,4}^{(m)} = \mathcal{O}_p\left(M_k^{-1/2}(n_k/M_k)^{-1/2-a/4}\right).$$

For $\mathcal{T}_{k,5}^{(m)}$: Elementary calculation shows that

$$\mathbb{E}\widetilde{\varepsilon}_{k,j}^{(m)2} = \frac{1-r_k^2(2\tau-1)^2}{4} =: S_k.$$

Applying Theorem 2.6.7 of Csörgo & Révész (1981) with $H(x)=x^{2p}$ and $x_n = n_k^{\beta_0}$, there exist i.i.d. standard normal $\widetilde{Z}_{k,i}$'s and some $a_k, C_k > 0$ (depending on the distribution of $H_k^{-1}\widetilde{\varepsilon}_{k,j}^{(m)}$) such that

$$\mathbb{P}\left(\left|\sum_{m=1}^{M_k}\sum_{t=1}^{T_k}\frac{H_k^{-1}\widetilde{\varepsilon}_{k,t}^{(m)}}{\sqrt{H_k^{-1}S_kH_k^{-1}}} - \sum_{i=1}^{n_k}\widetilde{Z}_{k,i}\right| > n_k^{\beta_0}\right) \le C_k a_k^{-2p} n_k^{1-2p\beta_0}.$$

Thus,

$$\mathbb{P}\left(\left|\frac{1}{n_k}\sum_{m=1}^{M_k}\sum_{t=1}^{T_k}\frac{H_k^{-1}\widetilde{\varepsilon}_{k,t}^{(m)}}{\sqrt{H_k^{-1}S_kH_k^{-1}}} - \frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{Z}_{k,i}\right| > n_k^{-1+\beta_0}\right) \lesssim n_k^{1-2p\beta_0}.$$

For $p>2$, one selects $\beta_0 \in (1/p, 1/2)$, the Borel-Cantelli lemma leads to

$$\left|\frac{1}{M_k}\sum_{m=1}^{M_k}\mathcal{T}_{k,5}^{(m)} - \frac{1}{n_k}\sum_{i=1}^{n_k}Z_{k,i}\right| = \mathcal{O}_{a.s.}\left(n_k^{-1+\beta_0}\right),$$

where $Z_{k,i}$'s are i.i.d. normal r.v.'s with mean zero and covariance $H_k^{-1}S_kH_k^{-1}$.

Therefore, we obtain that

$$\left|\widehat{\theta}_k - \theta_k - \frac{1}{n_k}\sum_{i=1}^{n_k}Z_{k,i}\right| = \mathcal{o}_p\left(n_k^{-1/2}\right),$$

which completes the proof of the weak convergence result.

As for the consistency of $\widehat{\sigma}_k^2$, we rewrite

$$\begin{aligned}\widehat{\sigma}_k^2 &= \frac{n_k}{M_k-1}\frac{1}{M_k}\sum_{m=1}^{M_k}\left(\widehat{\theta}_k^{(m)}-\theta_k\right)^2 - \frac{n_k}{M_k-1}\left\{\frac{1}{M_k}\sum_{m=1}^{M_k}\left(\widehat{\theta}_k^{(m)}-\theta_k\right)\right\}^2 \\ &= \frac{n_k}{M_k-1}\frac{1}{M_k}\sum_{m=1}^{M_k}\mathcal{T}_k^{(m)2} - \frac{n_k}{M_k-1}\left\{\frac{1}{M_k}\sum_{m=1}^{M_k}\mathcal{T}_k^{(m)}\right\}^2.\end{aligned}$$

Recall the definitions of $\mathcal{T}_{k,j}^{(m)}$ for $1\le j\le 5$.

For $\mathcal{T}_{k,1}^{(m)}$: According to Lemma C.4 of Xie et al. (2024), one has $\left|A_{k,0}^{t-1}\right| \le C_0$ uniformly for all $t\ge 1$. Further observe that $\Delta_{k,0}^{(m)} \equiv \widehat{\theta}_{k,0} - \theta_k$ for all $1\le m\le M_k$, thus one obtains $\left\|\mathcal{T}_{k,1}^{(m)}\right\|_2 = \mathcal{O}\left(T_k^{-1}\right)$, where the constant does not depend on m.

For $\mathcal{T}_{k,2}^{(m)}$: Consider that

$$\left|\frac{1}{T_k}\sum_{j=0}^{T_k-1}A_{k,j}^{T_k-1}r_{k,j}^{(m)}\right| \lesssim \frac{1}{T_k}\sum_{j=0}^{T_k-1}\left|r_{k,j}^{(m)}\right| \lesssim \frac{1}{T_k}\sum_{j=0}^{T_k-1}\left|\widehat{\theta}_{k,j}^{(m)}-\theta_k\right|^2.$$

Then,

$$\left\| \frac{1}{T_k} \sum_{j=0}^{T_k-1} A_{k,j}^{T_k-1} r_{k,j}^{(m)} \right\|_2 \lesssim \frac{1}{T_k} \sum_{j=0}^{T_k-1} \left\| \widehat{\theta}_{k,j}^{(m)} - \theta_k \right\|_4^2.$$

Applying Theorem 5 of Gadat & Panloup (2023), we have

$$\mathbb{E} \left| \widehat{\theta}_{k,j}^{(m)} - \theta_k \right|^4 \lesssim \eta_{k,j}^2.$$

Since $\eta_{k,j} \asymp j^{-a}$ with $a > 1/2$, it follows that $\left\| \mathcal{T}_{k,2}^{(m)} \right\|_2 = o\left(T_k^{-1/2}\right)$.

For $\mathcal{T}_{k,3}$: As shown in the previous arguments, one has $\left\| \mathcal{T}_{3,k}^{(m)} \right\|_{2p} = \mathcal{O}\left(T_k^{-1+a/2}\right) = o\left(T_k^{-1/2}\right)$, since $a < 1$.

For $\mathcal{T}_{k,4}$: Observe that $\sum_{j=1}^{T_k} H_k^{-1} \left(\varepsilon_{k,j}^{(m)} - \widetilde{\varepsilon}_{k,j}^{(m)} \right)$ is a martingale for each m (independent over $1 \le m \le M_k$), Burkholder's inequality entails that

$$\begin{aligned} \left\| \sum_{j=1}^{T_k} H_k^{-1} \left(\varepsilon_{k,j}^{(m)} - \widetilde{\varepsilon}_{k,j}^{(m)} \right) \right\|_2 &\lesssim \left\{ \sum_{j=1}^{T_k} \left\| H_k^{-1} \left(\varepsilon_{k,j}^{(m)} - \widetilde{\varepsilon}_{k,j}^{(m)} \right) \right\|_2^2 \right\}^{1/2} \\ &\lesssim \left(\sum_{j=1}^{T_k} \eta_{k,j}^{1/2} \right)^{1/2} \lesssim T_k^{\{1-a/2\}/2}. \end{aligned}$$

Hence, for any $1 \le m \le M_k$,

$$\left\| \mathcal{T}_{k,4}^{(m)} \right\|_2 = \left\| \frac{1}{T_k} \sum_{j=1}^{T_k} H_k^{-1} \left(\varepsilon_{k,j}^{(m)} - \widetilde{\varepsilon}_{k,j}^{(m)} \right) \right\|_2 = \mathcal{O}\left(T_k^{-1/2-a/4} \right).$$

For $\mathcal{T}_{k,5}$: Applying Theorem 2.6.7 of Csörgo & Révész (1981) with $H(x) = x^{2p}$ and $x_n = v T_k^{\beta_0}$, there exist i.i.d. standard normal $\widetilde{Z}_{k,j}^{(m)}$'s and some $a_k, C_k > 0$ (depending on the distribution of $H_k^{-1} \widetilde{\varepsilon}_{k,j}^{(m)}$) such that

$$\mathbb{P}\left(\left| \sum_{j=1}^{T_k} \frac{H_k^{-1} \widetilde{\varepsilon}_{k,j}^{(m)}}{\sqrt{H_k^{-1} S_k H_k^{-1}}} - \sum_{j=1}^{T_k} \widetilde{Z}_{k,j}^{(m)} \right| > v T_k^{\beta_0} \right) \le C_k a_k^{-2p} v^{-2p} T_k^{1-2p\beta_0}.$$

Thus,

$$\mathbb{P}\left(\left| \frac{1}{T_k} \sum_{j=1}^{T_k} \frac{H_k^{-1} \widetilde{\varepsilon}_{k,j}^{(m)}}{\sqrt{H_k^{-1} S_k H_k^{-1}}} - \frac{1}{T_k} \sum_{j=1}^{T_k} \widetilde{Z}_{k,j}^{(m)} \right| > v T_k^{-1+\beta_0} \right) \lesssim v^{-2p} T_k^{1-2p\beta_0}.$$

Since $\mathbb{E}X^p = \int_0^\infty p v^{p-1} \mathbb{P}(|X| > v) dv$, we also have

$$\left\| \mathcal{T}_{k,5}^{(m)} - \frac{1}{T_k} \sum_{j=1}^{T_k} Z_{k,j}^{(m)} \right\|_2 = \mathcal{O}\left(T_k^{-1+\beta_0} \right).$$

Acoording to the above results, we show that

$$\left\| \mathcal{T}_k^{(m)} - \frac{1}{T_k} \sum_{j=1}^{T_k} Z_{k,j} \right\|_2 = o\left(T_k^{-1/2} \right),$$

which implies

$$\left\|\mathcal{T}_k^{(m)}\right\|_2 = \left\|\frac{1}{\sqrt{T_k}}\sum_{j=1}^{T_k} Z_{k,j}\right\|_2 + o(1) < \infty.$$

The SLLN (i.i.d.) and the continuous mapping theorem further yields that

$$\frac{1}{M_k}\sum_{m=1}^{M_k}\left(\sqrt{T_k}\mathcal{T}_k^{(m)}\right)^2 \xrightarrow{a.s.} \sigma_k^2, \quad \left(\frac{1}{M_k}\sum_{m=1}^{M_k}\sqrt{T_k}\mathcal{T}_k^{(m)}\right)^2 \xrightarrow{a.s.} 0,$$

which completes the proof of consistency of $\widehat{\sigma}_k^2$.

C.2 Proof of Lemma B.4

Under Assumption 4, it is easy to obtain that for every $0 \le k \le K$,

$$|\widehat{b}_k - b_k| = \mathcal{O}_p\Big(\frac{1}{\sqrt{N}}\Big); \quad |\widehat{b}_k^2 - b_k^2| = \mathcal{O}_p\Big(\frac{1}{N}\Big) + |b_k|\mathcal{O}_p\Big(\frac{1}{\sqrt{N}}\Big). \tag{C.1}$$

Next, define the notation $\lesssim_p$ as follows: $a \lesssim_p b$ indicates that $a \lesssim b$ holds with probability $1 - o(1)$. Recall the definitions of w_k^* and $\widehat{w}_k$. Define $\Delta_k = \frac{\sigma_k^2}{n_k} + \lambda b_k^2$, and $\widehat{\Delta}_k = \frac{\widehat{\sigma}_k^2}{n_k} + \lambda \widehat{b}_k^2$. Then, we have

$$w_k^* = \frac{\Delta_k^{-1}}{\sum_{j=0}^K \Delta_j^{-1}}, \quad \widehat{w}_k = \frac{\widehat{\Delta}_k^{-1}}{\sum_{j=0}^K \widehat{\Delta}_j^{-1}}.$$

We first consider the case $K = 2$ and scenario (c). i.e., λ satisfies Assumption 6, and the bias scale satisfies Assumption 5. In this case, w_k^* and $\widehat{w}_k$ can be rewritten as

$$w_k^* = \frac{\Delta_{k_1}\Delta_{k_2}}{\Delta_0\Delta_1 + \Delta_0\Delta_2 + \Delta_1\Delta_2} = \frac{\Delta_{k_1}\Delta_{k_2}}{\sum_{k=0}^2\prod_{j\neq k}\Delta_j}, \widehat{w}_k = \frac{\widehat{\Delta}_{k_1}\widehat{\Delta}_{k_2}}{\sum_{k=0}^2\prod_{j\neq k}\widehat{\Delta}_j} \text{ for } k_1 \neq k_2, k_1 \neq k, k_2 \neq k.$$

It can then be verified that

$$\begin{aligned}
|\widehat{w}_k - w_k^*| &= \Big|\frac{\Delta_{k_1}\Delta_{k_2}}{\sum_{k=0}^2\prod_{j\neq k}\Delta_j} - \frac{\widehat{\Delta}_{k_1}\widehat{\Delta}_{k_2}}{\sum_{k=0}^2\prod_{j\neq k}\widehat{\Delta}_j}\Big| \\
&= \frac{|(\widehat{\Delta}_{k_1}\widehat{\Delta}_{k_2} - \Delta_{k_1}\Delta_{k_2})(\sum_{k=0}^2\prod_{j\neq k}\Delta_j) + \Delta_{k_1}\Delta_{k_2}\{\sum_{k=0}^2(\prod_{j\neq k}\Delta_j - \prod_{j\neq k}\widehat{\Delta}_j)\}|}{(\sum_{k=0}^2\prod_{j\neq k}\widehat{\Delta}_j)(\sum_{k=0}^2\prod_{j\neq k}\widehat{\Delta}_j)} \\
&\le \frac{|\widehat{\Delta}_{k_1}\widehat{\Delta}_{k_2} - \Delta_{k_1}\Delta_{k_2}| + |\sum_{k=0}^2(\prod_{j\neq k}\Delta_j - \prod_{j\neq k}\widehat{\Delta}_j)|}{\sum_{k=0}^2\prod_{j\neq k}\Delta_j}\frac{\sum_{k=0}^2\prod_{j\neq k}\Delta_j}{\sum_{k=0}^2\prod_{j\neq k}\widehat{\Delta}_j}.
\end{aligned}$$

The inequality holds because $\Delta_{k_1}\Delta_{k_2} \le \sum_{k=0}^2\prod_{j\neq k}\Delta_j$. Under assumption 4, we know that $\widehat{\Delta}_j$ is a consistent estimator of Δ_j. Therefore, we obtain

$$|\widehat{w}_k - w_k^*| \lesssim_p \frac{|\widehat{\Delta}_{k_1}\widehat{\Delta}_{k_2} - \Delta_{k_1}\Delta_{k_2}| + |\sum_{k=0}^2(\prod_{j\neq k}\Delta_j - \prod_{j\neq k}\widehat{\Delta}_j)|}{\sum_{k=0}^2\prod_{j\neq k}\Delta_j} \lesssim \frac{\sum_{k=0}^2|\prod_{j\neq k}\widehat{\Delta}_j - \prod_{j\neq k}\Delta_j|}{\sum_{k=0}^2\prod_{j\neq k}\Delta_j}. \tag{C.2}$$

We now analyze the numerator in equation (C.2). Under assumption 4 and equation (C.1), we have

$$|\widehat{\Delta}_k - \Delta_k| \lesssim \frac{|\widehat{\sigma}_k^2 - \sigma_k^2|}{N} + \lambda|\widehat{b}_k^2 - b_k^2| \lesssim_p \frac{o(1)}{N} + \frac{\lambda|b_k|}{\sqrt{N}}. \tag{C.3}$$

It can thus be shown that

$$\begin{aligned}
&|\widehat{\Delta}_{k_1}\widehat{\Delta}_{k_2} - \Delta_{k_1}\Delta_{k_2}| \le |\Delta_{k_1}(\Delta_{k_2} - \widehat{\Delta}_{k_2})| + |\Delta_{k_2}(\Delta_{k_1} - \widehat{\Delta}_{k_1})| + |(\widehat{\Delta}_{k_2} - \Delta_{k_2})(\Delta_{k_1} - \widehat{\Delta}_{k_1})| \\
&\lesssim_p \Big(\frac{\sigma_{k_1}^2}{N} + \lambda b_{k_1}^2\Big)\Big\{\frac{\lambda|b_{k_2}|}{\sqrt{N}} + \frac{o(1)}{N}\Big\} + \Big(\frac{\sigma_{k_2}^2}{N} + \lambda b_{k_2}^2\Big)\Big\{\frac{\lambda|b_{k_1}|}{\sqrt{N}} + \frac{o(1)}{N}\Big\} + \Big\{\frac{\lambda|b_{k_1}|}{\sqrt{N}} + \frac{o(1)}{N}\Big\}\times
\end{aligned}$$

$$\Big\{\frac{\lambda|b_{k_2}|}{\sqrt{N}}+\frac{\mathcal{O}(1)}{N}\Big\}=\mathcal{E}_1+\mathcal{E}_2+\mathcal{E}_3.$$

If $|b_{k_1}|\ll N^{-1/2}$, we have

$$\frac{\lambda|b_{k_1}|}{\sqrt{N}}+\frac{\mathcal{O}(1)}{N}\ll\frac{\sigma_{k_1}^2}{N}+\lambda b_{k_1}^2,$$

which implies $\mathcal{E}_3\ll\mathcal{E}_1$. Similarly, if $|b_{k_2}|\ll N^{-1/2}$, we have $\mathcal{E}_3\ll\mathcal{E}_2$. If both $\lambda|b_{k_1}|\gg N^{-1/2}$ and $\lambda|b_{k_2}|\gg N^{-1/2}$, then

$$\frac{\mathcal{E}_3}{\mathcal{E}_1}=\mathcal{O}\Big(\frac{\lambda^2|b_{k_1}||b_{k_2}|}{N}\times\frac{\sqrt{N}}{\lambda b_{k_1}^2\lambda|b_{k_2}|}\Big)=\mathcal{O}\Big(\frac{1}{\sqrt{N}|b_{k_1}|}\Big)=o(1).$$

Thus, we simplify to obtain

$$\begin{aligned}\big|\widehat{\Delta}_{k_1}\widehat{\Delta}_{k_2}-\Delta_{k_1}\Delta_{k_2}\big| &\lesssim_p \big|\Delta_{k_1}(\widehat{\Delta}_{k_2}-\Delta_{k_2})\big|+\big|\Delta_{k_2}(\widehat{\Delta}_{k_1}-\Delta_{k_1})\big|\\ &\lesssim_p \Big(\frac{\sigma_{k_1}^2}{N}+\lambda b_{k_1}^2\Big)\Big\{\frac{\lambda|b_{k_2}|}{\sqrt{N}}+\frac{\mathcal{O}(1)}{N}\Big\}+\Big(\frac{\sigma_{k_2}^2}{N}+\lambda b_{k_2}^2\Big)\Big\{\frac{\lambda|b_{k_1}|}{\sqrt{N}}+\frac{\mathcal{O}(1)}{N}\Big\}.\end{aligned}$$

Next, we analyze the denominator in equation (C.2). By definition, it could be shown that

$$\sum_{k=0}^{2}\prod_{j\neq k}\Delta_j\asymp\sum_{k=0}^{2}\prod_{j\neq k}\Big(\frac{\sigma_j^2}{N}+\lambda b_j^2\Big).$$

Substituting this result back into equation (C.2), we obtain

$$|\widehat{w}_k-w_k^*|\lesssim_p\Big\{\sum_{k=0}^{2}\prod_{j\neq k}\Big(\frac{\sigma_j^2}{N}+\lambda b_j^2\Big)\Big\}^{-1}\Big\{\sum_{k=0}^{2}\sum_{i\neq k}\Big(\frac{\lambda|b_i|}{\sqrt{N}}+\frac{o(1)}{N}\Big)\prod_{j\neq i}^{j\neq k}\Big(\frac{\sigma_j^2}{N}+\lambda b_j^2\Big)\Big\}.$$

We now consider different cases. First, note that we have $b_0=0$. For $k=1,2$, we discuss the following scenarios: (1) When $b_k\ll N^{-1/2}$ for $k=1,2$, the denominator is of order $1/N^2$, while the numerator is of order $o(1)/N^2$. (2) When $\lambda b_k\gg N^{-1/2}$ for one particular k and $b_j\ll N^{-1/2}$ for $j\neq k$, the denominator has order $\frac{\lambda b_k^2}{N}$, while the numerator has order $(\lambda b_k^2)o(1)/N+(\lambda|b_k|)/(N^{3/2})$. (3) When $\lambda b_k\gg N^{-1/2}$ for both $k=1,2$, the denominator has order $(\lambda b_1^2)(\lambda b_2^2)$, while the numerator is of order $(\lambda b_1^2)(\lambda|b_2|/\sqrt{N})+(\lambda b_2^2)(\lambda|b_1|/\sqrt{N})$. In each scenario above, we can verify that $|\widehat{w}_k-w_k^*|=o_p(1)$.

For the general case with arbitrary K, the proof follows analogously. The detailed argument is omitted here, but it similarly leads to the result:

$$|\widehat{w}_k-w_k^*|\lesssim_p\Big\{\sum_{k=0}^{K}\prod_{j\neq k}\Big(\frac{\sigma_j^2}{N}+\lambda b_j^2\Big)\Big\}^{-1}\Big\{\sum_{k=0}^{K}\sum_{i\neq k}\Big(\frac{\lambda|b_i|}{\sqrt{N}}+\frac{o(1)}{N}\Big)\prod_{j\neq i}^{j\neq k}\Big(\frac{\sigma_j^2}{N}+\lambda b_j^2\Big)\Big\}.$$

Under Assumption 5, we therefore conclude that $|\widehat{w}_k-w_k^*|=o_p(1)$. Finally, noting that scenarios (a) and (b) are in fact special cases of scenario (c), we complete the proof of the theorem.

C.3 Proof of Theorem 4.2

The proof of the consistency of $\widehat{w}_k$ is provided in Appendix C.2. Next, we establish the asymptotic properties of the proposed estimator $\widehat{\theta}_{\text{est}}$. The proof is divided into three parts. In the first part, we prove that:

$$\sqrt{N}(\widehat{\theta}(\mathbf{w}^*)-\theta_0)\to_d\mathcal{N}\big(0,\sigma^2(\mathbf{w}^*)\big).$$

In the second part, we establish that:

$$\sqrt{N}(\widehat{\theta}(\mathbf{w}^*)-\widehat{\theta}_{\text{est}})=o_p(1).$$

Combining the results from the first two parts immediately yields:

$$\sqrt{N}(\widehat{\theta}_{\text{est}}-\theta_0)\to_d\mathcal{N}\big(0,\sigma^2(\mathbf{w}^*)\big).$$

In the third part, we prove:

$$\widehat{\sigma}^2_{\text{est}} \to_p \sigma^2(\mathbf{w}^*).$$

By combining these three parts, we directly obtain Theorem 4.2.

PART 1. Recall that the estimator $\widehat{\theta}(\mathbf{w})$ can be written as:

$$\widehat{\theta}(\mathbf{w}) = \sum_{k=0}^{K} w_k \widehat{\theta}_k = \widehat{\theta}_0 + \sum_{k=1}^{K} w_k \widehat{b}_k,$$

where $\widehat{b}_k = \widehat{\theta}_k - \widehat{\theta}_0$. Define the true weighted parameter $\theta^*(\mathbf{w}) = \sum_{k=0}^{K} w_k \theta_k = \theta_0 + \sum_{k=1}^{K} w_k b_k$. We first show that $\sqrt{N}(\theta^*(\mathbf{w}^*) - \theta_0) = o(1)$. To see this, consider two cases based on Assumption 5: Frist, for $b_k \ll N^{-1/2}$, it follows immediately that: $\sqrt{N} b_k = o(1)$. In addition, for $b_k \gg N^{-1/2}$, recall the oracle weight definition:

$$w_k^* = \frac{(\frac{\sigma_k^2}{n_k} + \lambda b_k^2)^{-1}}{(\frac{\sigma_0^2}{n_0})^{-1} + \sum_{j=1}^{K} (\frac{\sigma_j^2}{n_j} + \lambda b_j^2)^{-1}}.$$

Under Assumption 6, we obtain: $w_k^* = O(1/(\lambda b_k^2 N)) = o(1)$, which implies:

$$\sqrt{N} w_k^* b_k = O(\sqrt{N} \frac{1}{\lambda b_k^2 N} b_k) = O(\frac{1}{\lambda |b_k| \sqrt{N}}) = o(1). \tag{C.4}$$

Therefore, under Assumption 5, we have:

$$\sqrt{N}(\theta^*(\mathbf{w}) - \theta_0) = o(1).$$

Combining this with Lemma B.2, we conclude that:

$$\sqrt{N}(\widehat{\theta}(\mathbf{w}^*) - \theta_0) \to_d \mathcal{N}(0, \sigma^2(\mathbf{w}^*)).$$

This finishes PART 1.

PART 2. Next, we analyze: $\sqrt{N}\{\widehat{\theta}(\mathbf{w}^*) - \widehat{\theta}_{\text{est}}\}$. By definition, we have

$$\widehat{\theta}(\mathbf{w}^*) - \widehat{\theta}_{\text{est}} = \widehat{\theta}_0 + \sum_{k=1}^{K} w_k^* (\widehat{\theta}_k - \widehat{\theta}_0) - \Big\{ \widehat{\theta}_0 + \sum_{k=1}^{K} \widehat{w}_k (\widehat{\theta}_k - \widehat{\theta}_0) \Big\} = \sum_{k=1}^{K} (w_k^* - \widehat{w}_k)(\widehat{\theta}_k - \widehat{\theta}_0).$$

Again, we consider two scenarios: When $b_k \ll N^{-1/2}$, it can be verified from equation (B.1) that:

$$\sqrt{N}(\widehat{\theta}_k - \widehat{\theta}_0) = \sqrt{N}(\widehat{b}_k - b_k) + \sqrt{N} b_k = O_p(1).$$

Using Lemma B.4 and Assumption 3, we thus have:

$$\sqrt{N}(w_k^* - \widehat{w}_k)(\widehat{\theta}_k - \widehat{\theta}_0) = o_p(1).$$

For the case $b_k \gg N^{-1/2}$, we similarly decompose:

$$\sqrt{N}(w_k^* - \widehat{w}_k)(\widehat{\theta}_k - \widehat{\theta}_0) = \sqrt{N}(w_k^* - \widehat{w}_k)(\widehat{\theta}_k - \theta_k) - \sqrt{N}(w_k^* - \widehat{w}_k)(\widehat{\theta}_0 - \theta_0) + \sqrt{N}(w_k^* - \widehat{w}_k) b_k.$$

By Lemmas B.2 and Lemma B.4, the first two terms are $o_p(1)$. Thus, it suffices to study $\sqrt{N}(w_k^* - \widehat{w}_k) b_k$. Note that

$$\sqrt{N}(w_k^* - \widehat{w}_k) b_k = \sqrt{N} w_k^* b_k - \sqrt{N} \widehat{w}_k (b_k - \widehat{b}_k) - \sqrt{N} \widehat{w}_k \widehat{b}_k.$$

First, from the analysis in PART 1, we already have $\sqrt{N} w_k^* b_k = o_p(1)$ by equation (C.4). Next, recall:

$$\widehat{w}_k = \frac{(\frac{\widehat{\sigma}_k^2}{n_k} + \lambda \widehat{b}_k^2)^{-1}}{\frac{\widehat{\sigma}_0^2}{n_0} + \sum_{j=1}^{K} (\frac{\widehat{\sigma}_j^2}{n_j} + \lambda \widehat{b}_j^2)^{-1}}.$$

Using equation (B.1) again, it can be shown that when $b_k \gg N^{-1/2}$, $1/(\widehat{b}_k \sqrt{N}) = o_p(1)$ when $b_k \gg N^{-1/2}$. As a result, it could be shown that

$$\widehat{w}_k = O_p\Big(\frac{1}{\lambda \widehat{b}_k^2 N}\Big)$$

since $\widehat{\sigma}_k^2 - \sigma_k^2 = o_p(1)$. This yields

$$\sqrt{N}\widehat{w}_k\widehat{b}_k = O_p\Big(\frac{1}{\lambda\widehat{b}_k\sqrt{N}}\Big) = o_p(1).$$

Hence, under Assumption 5, we obtain:

$$\sqrt{N}\Big(\widehat{\theta}(\mathbf{w}^*) - \widehat{\theta}_{\text{est}}\Big) = o_p(1),$$

which completes PART 2.

PART 3. Finally, from Assumption 4, we have: $\widehat{\sigma}_k^2 - \sigma_k^2 = o_p(1)$. Combining this with Lemma B.4, we have: $\widehat{\sigma}_{\text{est}}^2 \to_p \sigma^2(\mathbf{w}^*)$, which completes PART 3. This finished the whole theorem proof.

C.4 PROOF OF THEOREM 4.3

Recall that the loss function is defined as: $\mathcal{L}(\mathbf{w}) = \sum_{k=0}^K w_k^2 \frac{\sigma_k^2}{n_k} + \lambda \sum_{k=0}^K w_k^2 b_k^2$. The corresponding optimization problem is given by:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}), \quad \text{subject to} \quad w_k \geq 0 \text{ for all } k, \quad \sum_{k=0}^K w_k = 1. \tag{C.5}$$

Clearly, $(1, 0, \ldots, 0)$ is a feasible solution to problem (C.5), with the corresponding loss equal to σ_0^2/n_0. As a consequence, to prove Theorem 4.3, it suffices to show that there exists a feasible solution $\mathbf{w}'$ satisfying $\mathcal{L}(\mathbf{w}') < \sigma_0^2/n_0$.

Suppose that the bias scale of the kth site satisfies $b_k \ll N^{-1/2}$. We define $\mathbf{w}' = (w_j')$ as follows:

$$w_0' = \frac{\sigma_k^2/n_k}{\sigma_0^2/n_0 + \sigma_k^2/n_k}, \quad w_k' = \frac{\sigma_0^2/n_0}{\sigma_0^2/n_0 + \sigma_k^2/n_k}, \quad w_j' = 0 \text{ for all } j \neq 0, k.$$

Clearly, $\mathbf{w}'$ is feasible, and it could be computed that:

$$\mathcal{L}(\mathbf{w}') = \frac{(\sigma_0^2/n_0)(\sigma_k^2/n_k)}{\sigma_0^2/n_0 + \sigma_k^2/n_k} + \lambda w_k'^2 b_k^2 = \frac{1}{\frac{1}{\sigma_0^2/n_0} + \frac{1}{\sigma_k^2/n_k}}(1 + o(1)).$$

This yields:

$$\frac{\mathcal{L}(\mathbf{w}')}{\sigma_0^2/n_0} = \frac{\sigma_k^2/n_k}{\sigma_0^2/n_0 + \sigma_k^2/n_k}(1 + o(1)) < 1.$$

This finishes the whole theorem proof.

C.5 PROOF OF THEOREM 4.4.

Note that the conservative weight $\widetilde{w}_k$ does not necessarily converge to the oracle weight w_k^*. Denote $U_0 = 0$, $U_k = \mathcal{C}\sqrt{\sigma_k^2/n_k + \sigma_0^2/n_0}$ for $1 \leq k \leq K$, we introduce the auxiliary weights

$$\widetilde{w}_k^* = \frac{\Big\{\frac{\sigma_k^2}{n_k} + \lambda\big(|b_k| + U_k\big)^2\Big\}^{-1}}{\Big(\frac{\sigma_0^2}{n_0}\Big)^{-1} + \sum_j \Big\{\frac{\sigma_j^2}{n_j} + \lambda\big(|b_j| + U_j\big)^2\Big\}^{-1}} \text{ for } 0 \leq k \leq K.$$

Using these weights, we define the auxiliary estimator $\widehat{\theta}\big(\widetilde{\mathbf{w}}\big) = \sum_{k=0}^K \widetilde{w}_k \widehat{\theta}_k$. The proof has two steps. in the first step, We show that $\big|\widetilde{w}_k - \widetilde{w}_k^*\big| = o_p(1)$. In the second step, we establish Theorem 4.4 with the help of $\widetilde{w}_k^*$ and $\widehat{\theta}\big(\widetilde{\mathbf{w}}^*\big)$.

STEP 1. Recall the definition of $\widetilde{w}_k$, denote $\widehat{U}_0 = 0$, $\widehat{U}_k = \mathcal{C}\sqrt{\widehat{\sigma}_k^2/n_k + \widehat{\sigma}_0^2/n_0}$ for $1 \leq k \leq K$, we have

$$\widetilde{w}_k = \frac{\Big\{\frac{\widehat{\sigma}_k^2}{n_k} + \lambda\big(|\widehat{b}_k| + \widehat{U}_k\big)^2\Big\}^{-1}}{\Big(\frac{\widehat{\sigma}_0^2}{n_0}\Big)^{-1} + \sum_j \Big\{\frac{\widehat{\sigma}_j^2}{n_j} + \lambda\big(|\widehat{b}_j| + \widehat{U}_j\big)^2\Big\}^{-1}} \text{ for } 0 \leq k \leq K.$$

Then we have

$$\begin{aligned}\big|\widetilde{b}_k^2-(|b_k|+U_k)^2\big| &= \big|(|\widehat{b}_k|+\widehat{U}_k)^2-(|b_k|+U_k)^2\big| \\ &\le \big\{(|\widehat{b}_k|+\widehat{U}_k)-(|b_k|+U_k)\big\}^2+(|b_k|+U_k)\big\{(|\widehat{b}_k|+\widehat{U}_k)-(|b_k|+U_k)\big\}.\end{aligned} \tag{C.6}$$

It could be proved that

$$|\widehat{U}_k-U_k|^2\le\mathcal{C}^2\Big|\Big(\frac{\widehat{\sigma}_k^2}{n_k}+\frac{\widehat{\sigma}_0^2}{n_0}\Big)-\Big(\frac{\sigma_k^2}{n_k}+\frac{\sigma_0^2}{n_0}\Big)\Big|\le\mathcal{C}^2\Big(\Big|\frac{\widehat{\sigma}_k^2}{n_k}-\frac{\sigma_k^2}{n_k}\Big|+\Big|\frac{\widehat{\sigma}_0^2}{n_0}-\frac{\sigma_0^2}{n_0}\Big|\Big).$$

This yields

$$|\widehat{U}_k-U_k|\le\mathcal{C}\Big(\Big|\frac{\widehat{\sigma}_k^2}{n_k}-\frac{\sigma_k^2}{n_k}\Big|+\Big|\frac{\widehat{\sigma}_0^2}{n_0}-\frac{\sigma_0^2}{n_0}\Big|\Big)^{1/2}.$$

Because we assume $\mathcal{C}\to\infty$ and $\mathcal{C}\sqrt{\widehat{\sigma}_k^2-\sigma_k^2}=\mathcal{O}_p(1)$ for every $k=0,\ldots,K$, then we have

$$\Big|(|\widehat{b}_k|+\widehat{U}_k)-(|b_k|+U_k)\Big|\le\Big||\widehat{b}_k|-|b_k|\Big|+|\widehat{U}_k-U_k|\le|\widehat{b}_k-b_k|+|\widehat{U}_k-U_k|=\mathcal{O}_p\big(\frac{1}{\sqrt{N}}\big).$$

Substituting the bound above into equation (C.6) yields

$$\Big|(|\widehat{b}_k|+\widehat{U}_k)^2-(|b_k|+U_k)^2\Big|=\mathcal{O}_p\big(\frac{1}{N}\big)+\Big\{|b_k|+\mathcal{C}\Omega_p\big(\frac{1}{\sqrt{N}}\big)\Big\}\mathcal{O}_p\big(\frac{1}{\sqrt{N}}\big).$$

Next, by repeating the key steps in the proof of Lemma B.4, we can readily show that $\widetilde{w}_k-\widetilde{w}_k^*=o_p(1)$. The detailed proof is therefore omitted.

STEP 2. Note that $|b_k|+U_k\gg N^{-1/2}$, then it could be verified that

$$\widetilde{w}_k^*\lesssim\frac{\lambda^{-1}\Big\{(|b_k|+U_k)^2\Big\}^{-1}}{N+\sum_j\Big\{\lambda\big(|b_j|+U_j\big)^2\Big\}^{-1}}=\mathcal{O}\Big(\frac{1}{\lambda(|b_k|+U_k)^2N}\Big)=o(1),$$

and

$$\sqrt{N}\widetilde{w}_k^*b_k=\mathcal{O}\Big(\frac{b_k}{\lambda(|b_k|+U_k)^2\sqrt{N}}\Big)=o(1).$$

Similarly, we can prove that

$$\widetilde{w}_k=O_p\Big(\frac{1}{\lambda(|\widehat{b}_k|+\widehat{U}_k)^2N}\Big)$$

and

$$\sqrt{N}\widetilde{w}_k\widehat{b}_k=O_p\Big(\frac{\widehat{b}_k}{\lambda(|\widehat{b}_k|+\widehat{U}_k)^2\sqrt{N}}\Big)=o_p(1).$$

Next, by repating the key steps in the proof of Theorem 4.2, it could be verifeid that

$$\sqrt{N}(\widehat{\theta}(\widetilde{\mathbf{w}}^*)-\theta_0)\to_d\mathcal{N}\big(0,\sigma^2(\widetilde{\mathbf{w}}^*)\big),\quad\sqrt{N}(\widehat{\theta}(\widetilde{\mathbf{w}}^*)-\widehat{\theta}_{\text{cons}})=o_p(1).$$

Finally, note that

$$\widehat{\sigma}_{\text{cons}}^2\to_p\sigma^2(\widetilde{\mathbf{w}}^*)$$

by Theorem 4.1 and using the fact that $\widetilde{w}_k-\widetilde{w}_k^*=o_p(1)$. By combining these three results, we finished the whole proof.