Supplement to "Distributed Mean Dimension Reduction through Semi-parametric Approaches"

Zhengtian Zhu, Wangli Xu and Liping Zhu

Renmin University of China

Supplementary Material

The online Supplementary Material contains descriptions of the pooled algorithms, additional simulations, and technical proofs of all theorems.

S1. The pooled algorithms

S1.1 The first pooled algorithm with dense solutions

For now we assume all observations $\{(\mathbf{x}_i, Y_i), i = 1, \ldots, N\}$ are scattered at a single machine. To implement (2.2), we must replace all unknowns with their sample counterparts. Towards this goal, Ma and Zhu (2014) suggested estimating $m(\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha})$, its first derivative $\mathbf{m}_1(\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha})$, $E\{w(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}\}$, and $E\{\mathbf{x}w(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}\}$ via nonparametric treatment, where $\boldsymbol{\alpha}$ is an intermediate estimate. In particular, $\widehat{m}(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\alpha})$ and $\widehat{\mathbf{m}}_1(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\alpha})$ can be simultaneously

obtained through a local linear approximation procedure,

$$(\widehat{b}_k, \widehat{\mathbf{b}}_k) \stackrel{\text{def}}{=} \arg\min_{b_k, \mathbf{b}_k} \sum_{i=1, i\neq k}^{N} \{Y_i - b_k - (\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\alpha} - \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\alpha})\mathbf{b}_k\}^2 K_{h_{1,*}}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\alpha} - \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\alpha}),$$

where $K_{h_{1,*}}(\cdot) = K(\cdot/h_{1,*})/h_{1,*}^d$, K is the multiplication of d univariate kernel functions, $h_{1,*}$ is a bandwidth. Let $\widehat{m}(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\alpha}) = \widehat{b}_k$ and $\widehat{\mathbf{m}}_1(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\alpha}) = \widehat{\mathbf{b}}_k$. Ma and Zhu (2014) suggested estimating $E\{w(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}\}$ and $E\{\mathbf{x}w(\mathbf{x}) \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}\}$ through usual kernel smoothers. For now we assume $w(\mathbf{x})$ is known. We can simply specify $w(\mathbf{x})$ as $w^*(\mathbf{x})$, or assume it has a parametric form $w(\mathbf{x}, \theta)$. We can also estimate $w(\mathbf{x})$ directly with kernel smoothers. As long as $w(\mathbf{x})$ is correctly specified (Ma and Zhu, 2014) or consistently estimated (Luo and Cai, 2016) , the resulting estimate of the pooled algorithm described below is semiparametrically efficient, despite the fact that the convergence rate of nonparametric estimate of $w(\mathbf{x})$ is pretty slow in high dimensions. Even if it is misspecified or estimated inconsistently, the resulting solution remains to be consistent. To be specific, we let $h_{2,*}$ and $h_{3,*}$ be two bandwidths. Define

$$\widehat{E}\{w(\mathbf{x}_k) \mid \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\alpha}\} \stackrel{\text{def}}{=} \frac{\sum_{i=1, i\neq k}^{N} K_{h_{2,*}}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\alpha} - \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\alpha}) w(\mathbf{x}_i)}{\sum_{i=1, i\neq k}^{N} K_{h_{2,*}}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\alpha} - \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\alpha})}, \text{ and,}$$

$$\widehat{E}\{\mathbf{x}_k w(\mathbf{x}_k) \mid \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\alpha}\} \stackrel{\text{def}}{=} \frac{\sum_{i=1, i\neq k}^{N} K_{h_{3,*}}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\alpha} - \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\alpha}) \{\mathbf{x}_i w(\mathbf{x}_i)\}}{\sum_{i=1, i\neq k}^{N} K_{h_{3,*}}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\alpha} - \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\alpha})}.$$

We further define

$$\widehat{\mathbf{x}}_k(\boldsymbol{\alpha}) \stackrel{\text{def}}{=} \text{vecl}\left\{\left[\mathbf{x}_k - \frac{\widehat{E}\{\mathbf{x}_k w(\mathbf{x}_k) \mid \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\alpha}\}}{\widehat{E}\{w(\mathbf{x}_k) \mid \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\alpha}\}}\right]\widehat{\mathbf{m}}_1^{\mathrm{T}}(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\alpha})\right\}.$$

Let $\widehat{\mathbf{S}}\{\mathbf{x}_k, Y_k, \boldsymbol{\alpha}, w(\mathbf{x}_k)\} \stackrel{\text{def}}{=} \{Y_k - \widehat{m}(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\alpha})\}w(\mathbf{x}_k)\widehat{\mathbf{x}}_k(\boldsymbol{\alpha})$. To implement (2.2) we simply replace $E\Big[\mathbf{S}\{\mathbf{x}, Y, \boldsymbol{\alpha}, w(\mathbf{x})\}\Big]$ and $\mathbf{H}(\boldsymbol{\alpha})$ with their respective sample averages,

$$\begin{aligned}\widehat{E}_{\text{pool},1}\Big[\mathbf{S}\{\mathbf{x}, Y, \boldsymbol{\alpha}, w(\mathbf{x})\}\Big] &\stackrel{\text{def}}{=} N^{-1}\sum_{k=1}^{N}\widehat{\mathbf{S}}\{\mathbf{x}_k, Y_k, \boldsymbol{\alpha}, w(\mathbf{x}_k)\}, \text{ and}\\ \widehat{\mathbf{H}}(\boldsymbol{\alpha}) &\stackrel{\text{def}}{=} N^{-1}\sum_{k=1}^{N}\Big[w(\mathbf{x}_k)\{\widehat{\mathbf{x}}_k(\boldsymbol{\alpha})\}\{\widehat{\mathbf{x}}_k(\boldsymbol{\alpha})\}^{\mathrm{T}}\Big].\end{aligned}$$

Starting from $\boldsymbol{\beta}^{(0)}$, we iterate the Newton–Raphson algorithm as follows,

$$\begin{aligned}&\text{vecl}(\boldsymbol{\beta}_{\text{pool},1}^{(t+1)}) \stackrel{\text{def}}{=} \qquad\qquad (S1.1)\\ &\text{vecl}(\boldsymbol{\beta}_{\text{pool},1}^{(t)}) + \Big\{\widehat{\mathbf{H}}(\boldsymbol{\beta}_{\text{pool},1}^{(t)})\Big\}^{-1}\widehat{E}_{\text{pool},1}\Big[\mathbf{S}\{\mathbf{x}, Y, \boldsymbol{\beta}_{\text{pool},1}^{(t)}, w(\mathbf{x})\}\Big].\end{aligned}$$

We denote the final solution by $\widehat{\boldsymbol{\beta}}_{\text{pool},1}$.

S1.2 The second pooled algorithm with sparse solutions

Under the least squares framework (2.4), we can incorporate penalties into the loss functions to produce sparse solutions. In particular, we define

$$\widehat{Y}(\boldsymbol{\alpha}) \stackrel{\text{def}}{=} \{\widehat{\mathbf{x}}(\boldsymbol{\alpha})\}^{\mathrm{T}}\,\text{vecl}(\boldsymbol{\alpha}) + \{Y - \widehat{m}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha})\}. \qquad (S1.2)$$

We start from an initial value $\boldsymbol{\beta}^{(0)}$, which may be sparse or not. We incorporate the least absolute shrinkage and selection operator (Tibshirani,

1996) into the least squares framework. Suppose, after t iterations, we have $\boldsymbol{\beta}_{\text{pool},2}^{(t)}$. We proceed to update it with $\boldsymbol{\beta}_{\text{pool},2}^{(t+1)}$, the minimizer of the following penalized least squares,

$$(2N)^{-1}\sum_{k=1}^{N}\{\widehat{Y}_k(\boldsymbol{\beta}_{\text{pool},2}^{(t)}) - \widehat{\mathbf{x}}_k(\boldsymbol{\beta}_{\text{pool},2}^{(t)})^{\mathrm{T}}\text{vecl}(\boldsymbol{\alpha})\}^2 w(\mathbf{x}_k) + \lambda_N\|\boldsymbol{\alpha}\|_1. \quad \text{(S1.3)}$$

We iterate (S1.3) until convergence. The final solution is denoted by $\widehat{\boldsymbol{\beta}}_{\text{pool},2}$.

S1.3 The third pooled algorithm under orthogonality constraints

We define the pooled algorithm under orthogonality constraint as follows. Suppose $\boldsymbol{\beta}_{\text{pool},3}^{(t)}$ is an orthonormal matrix, which is the pooled estimate at the t-th iteration. We use the definitions of $\widehat{\mathbf{x}}_k\big(\boldsymbol{\beta}_{\text{pool},3}^{(t)}\big)$ and $\widehat{Y}_k\big(\boldsymbol{\beta}_{\text{pool},3}^{(t)}\big)$ in Section 5, and define

$$\boldsymbol{\beta}_{\text{pool},3}^{(t+1)} \overset{\text{def}}{=} \arg\min_{\boldsymbol{\alpha}}\left[\sum_{k=1}^{N}\{\widehat{Y}_k(\boldsymbol{\beta}_{\text{pool},3}^{(t)}) - \widehat{\mathbf{x}}_k(\boldsymbol{\beta}_{\text{pool},3}^{(t)})^{\mathrm{T}}\text{vec}(\boldsymbol{\alpha})\}^2 w(\mathbf{x}_k)\right],$$

for $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha} = \mathbf{I}_{d\times d}$. We iterate the above minimization process until convergence. The final solution is denoted by $\widehat{\boldsymbol{\beta}}_{\text{pool},3}$. In this pooled algorithm, we also use non-monotone line search of Barzilai and Borwein (1988) to choose the step size $\tau^{(t)}$.

S2. Additional simulation studies

In this subsection, we provide additional simulation studies for three distributed algorithms and other competitors.

S2.1 Simulation study for algorithm 1

For algorithm 1, we generate data from the following two examples.

Example 1. We generate $\mathbf{x}$ from a multivariate normal distribution with mean zero and covariance matrix $\mathbf{\Sigma} = \left(0.5^{|i-j|}\right)_{p\times p}$, where $p = 6$. We generate Y from a normal distribution with mean $m(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = (\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}+1)$ and variance $\sigma^2(\mathbf{x}) = \{(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})^2 + 1\}/2$, where $\boldsymbol{\beta} = (1, 0.5, 1, 1.5, 2, -2)^{\mathrm{T}}$, $d = 1$.

Example 2. We generate $\mathbf{x}$ independently from a uniform distribution defined on $[-2, 2]$. We generate Y from a normal distribution with mean $m(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = \exp(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_1)+(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_2)^2$ and variance $\sigma^2(\mathbf{x}) = \log\{(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_1)^2+(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_2)^2+2\}$, where $\boldsymbol{\beta}_1 = (1, 0, 0.5, 1, 1.5, 2)^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = (0, 1, -0.5, 1, -1.5, 2)^{\mathrm{T}}$, and $d = 2$.

We run 500 replicates to compare the performance of the following estimates:

1. $\widehat{\boldsymbol{\beta}}_{\mathrm{pool},1}(w)$: The pooled estimate that pools all observations together and uses the true weight $w(\mathbf{x}) = \{\sigma^2(\mathbf{x})\}^{-1}$. This serves as a benchmark.

2. $\widehat{\boldsymbol{\beta}}_{\text{dist},1}(w)$: The distributed estimate that uses $w(\mathbf{x}) = \{\sigma^2(\mathbf{x})\}^{-1}$.

3. $\widehat{\boldsymbol{\beta}}_{\text{dist},1}(w^*)$: The distributed estimate that uses $w^*(\mathbf{x}) = 1$.

4. $\widehat{\boldsymbol{\beta}}_{\text{dist},1}(\widehat{w})$: The distributed estimate that estimates $w(\mathbf{x})$ with kernel smoother $\widehat{w}_j(\mathbf{x})$ at the jth machine, for $j = 1, \ldots m$.

5. OSIR1: The online sliced inverse regression via perturbation method (Cai et al., 2020).

6. OSIR2: The online sliced inverse regression via gradient descent optimization (Cai et al., 2020).

7. DKPCA: The distributed kernel principal component analysis (Balcan et al., 2016).

Let $\boldsymbol{\beta}$ be a basis matrix of the central mean subspace, and $\widehat{\boldsymbol{\beta}}$ be its estimate. To assess the estimation accuracy of $\widehat{\boldsymbol{\beta}}$, we use the Euclidean distance between $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$, defined as the Frobenius norm of the matrix $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})^{-1}\widehat{\boldsymbol{\beta}}^{\mathrm{T}} - \boldsymbol{\beta}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathrm{T}}$. A smaller distance indicates a better estimate.

Throughout, we fix the total sample size $N = 2500$. We consider three combinations, $(n, m) = (500, 5)$, $(250, 10)$, and $(100, 25)$, where m is the number of machines. We choose the initial value $\boldsymbol{\beta}^{(0)}$ by using a minimum average variance estimation (Xia et al., 2002). We choose the bandwidths

using a "rule-of-thumb" approach because the semiparametric estimating equations approach is not sensitive to the bandwidth selections (Ma and Zhu, 2014). In particular, we set $h_1 = h_2 = h_3 = cn^{-1/(4+d)}$. To estimate $w(\mathbf{x})$ with kernel smoother $\widehat{w}_j(\mathbf{x})$, we follow Luo and Cai (2016) and set $h_4 = cn^{-1/(2p)}$, where c is the average of marginal standard deviations of the p covariates.

Tables 1 summarizes the averages and the standard deviations (in the parentheses) of the distances. Not surprisingly, $\widehat{\boldsymbol{\beta}}_{\text{pool},1}(w)$ performs the best among the distributed estimates from algorithm 1. It has the smallest biases and standard deviations across all scenarios. It can be clearly seen that, the biases and standard deviations of three distributed estimates, $\widehat{\boldsymbol{\beta}}_{\text{dist},1}(w)$, $\widehat{\boldsymbol{\beta}}_{\text{dist},1}(w^*)$ and $\widehat{\boldsymbol{\beta}}_{\text{dist},1}(\widehat{w})$, increase with the number of machines. In addition, $\widehat{\boldsymbol{\beta}}_{\text{dist},1}(w)$ seems the most efficient, followed by $\widehat{\boldsymbol{\beta}}_{\text{dist},1}(\widehat{w})$. It is not surprising that $\widehat{\boldsymbol{\beta}}_{\text{dist},1}(w^*)$ is the least efficient among the distributed estimates from algorithm 1 because the weight function is misspecified. The online sliced inverse regressions, OSIR1 and OSIR2, have comparable performance as the distributed estimates in Example 1 where the linearity condition is satisfied. However, these two online estimates are much worse in Example 2 where the linearity condition is violated. The distributed kernel principal component analysis is not surprisingly the worst because it is an

Table 1: The averages (aver) and the standard deviations (in the parentheses, std) of the distance of various estimates.

(n, m)	(500,5)		(250, 10)		(100, 25)	
	aver	std	aver	std	aver	std
	Example 1					
$\widehat{\boldsymbol{\beta}}_{\text{pool},1}(w)$	0.092 (0.033)					
$\widehat{\boldsymbol{\beta}}_{\text{dist},1}(w)$	0.094	(0.037)	0.099	(0.041)	0.103	(0.045)
$\widehat{\boldsymbol{\beta}}_{\text{dist},1}(\widehat{w})$	0.105	(0.042)	0.111	(0.048)	0.118	(0.052)
$\widehat{\boldsymbol{\beta}}_{\text{dist},1}(w^*)$	0.109	(0.048)	0.122	(0.066)	0.143	(0.075)
OSIR1	0.099 (0.037)					
OSIR2	0.106 (0.041)					
DKPCA	0.198	(0.072)	0.236	(0.085)	0.257	(0.091)
	Example 2					
$\widehat{\boldsymbol{\beta}}_{\text{pool},1}(w)$	0.109 (0.038)					
$\widehat{\boldsymbol{\beta}}_{\text{dist},1}(w)$	0.115	(0.044)	0.122	(0.048)	0.139	(0.065)
$\widehat{\boldsymbol{\beta}}_{\text{dist},1}(\widehat{w})$	0.119	(0.049)	0.125	(0.052)	0.146	(0.069)
$\widehat{\boldsymbol{\beta}}_{\text{dist},1}(w^*)$	0.138	(0.063)	0.147	(0.071)	0.175	(0.077)
OSIR1	0.241 (0.091)					
OSIR2	0.167 (0.062)					
DKPCA	0.321	(0.095)	0.364	(0.113)	0.478	(0.134)

unsupervised method that completely ignores the response observations.

S2.2 Simulation study for algorithm 2

For algorithm 2, we generate data from the following two examples. In both examples, $p = 500$.

Example 3. In this example, $d = 1$, $\boldsymbol{\beta}$ is a p-vector with its first five components being $(1, 1, -1, 1, 1)^{\mathrm{T}}$ and all other entries being identically zero. We generate $\mathbf{x}$ independently from a uniform distribution defined on $[-3^{1/2}, 3^{1/2}]$. We generate Y from a normal distribution with mean $m(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = (\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$, and variance function $\sigma^2(\mathbf{x}) = \exp(X_1)$, where X_1 is the first coordinate of $\mathbf{x}$.

Example 4. In this example, $d = 2$, both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are p-vectors with their first six components being $(1, 0, 1, 1, 1, 1)^{\mathrm{T}}$ and $(0, 1, 1, -1, 1, -1)^{\mathrm{T}}$, respectively, and all other entries being identically zero. We generate $\mathbf{x}$ from a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma} = \left(0.5^{|i-j|}\right)_{p\times p}$. We generate Y from a normal distribution with mean function $m(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = (\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_1)/\{0.5 + (1.5 + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_2)^2\}$, and variance function $\sigma^2(\mathbf{x}) = \{0.1 + m^2(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})/5\}$.

We implement the following estimates and run 500 replicates.

1. $\widehat{\boldsymbol{\beta}}_{\mathrm{pool},2}(w)$: The regularized pooled estimate that aggregates all obser-

vations together and uses the true weight $w(\mathbf{x}) = \{\sigma^2(\mathbf{x})\}^{-1}$. This serves as a benchmark.

2. $\widehat{\boldsymbol{\beta}}_{\text{dist},2}(w)$: The regularized distributed estimate that uses $w(\mathbf{x}) = \{\sigma^2(\mathbf{x})\}^{-1}$.

3. $\widehat{\boldsymbol{\beta}}_{\text{dist},2}(w^*)$: The regularized distributed estimate that misspecifies $w(\mathbf{x})$ as $w^*(\mathbf{x}) = 1$.

4. DMDR: Distributed mean dimension reduction by Zhu and Zhu (2022).

5. DKPCA: Communication efficient distributed kernel principal component analysis by Balcan et al. (2016).

We evaluate the support recovery performance through the F_1-score. It is defined as

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}},$$

which ranges from 0 to 1. A larger F_1-score implies better support recovery.

Throughout we fix the total sample size $N = 2500$. We consider three combinations, $(n, m) = (500, 5)$, $(250, 10)$, and $(100, 25)$, where m is the number of machines. We choose the initial value $\boldsymbol{\beta}^{(0)}$ by sparse sliced inverse regression (Lin et al., 2019). We set the bandwidths in the same way as in Section 3.3.

Table 2: The average (aver) and the standard deviations (in the parentheses, std) of the distances and F_1-scores. The data is generated from example 3 and 4.

(n, m)	(500,5)		(250, 10)		(100, 25)	
	distance	score	distance	score	distance	score
	Example 3					
$\widehat{\boldsymbol{\beta}}_{\text{pool},2}(w)$	0.19 (0.02)	1.00 (0.00)				
$\widehat{\boldsymbol{\beta}}_{\text{dist},2}(w)$	0.22 (0.03)	1.00 (0.00)	0.27 (0.06)	0.97 (0.08)	0.32 (0.10)	0.95 (0.11)
$\widehat{\boldsymbol{\beta}}_{\text{dist},2}(w*)$	0.26 (0.05)	0.96 (0.09)	0.31 (0.09)	0.92 (0.12)	0.39 (0.14)	0.88 (0.18)
DMDR	0.24 (0.04)	1.00 (0.00)	0.29 (0.06)	0.97 (0.11)	0.37 (0.09)	0.91 (0.15)
DKPCA	0.43 (0.08)	0.49 (0.02)	0.57 (0.12)	0.46 (0.03)	0.66 (0.17)	0.44 (0.03)
	Example 4					
$\widehat{\boldsymbol{\beta}}_{\text{pool},2}(w)$	0.25 (0.02)	1.00 (0.00)				
$\widehat{\boldsymbol{\beta}}_{\text{dist},2}(w)$	0.31 (0.04)	0.99 (0.01)	0.37 (0.08)	0.94 (0.02)	0.39 (0.14)	0.91 (0.04)
$\widehat{\boldsymbol{\beta}}_{\text{dist},2}(w*)$	0.35 (0.07)	0.95 (0.10)	0.43 (0.09)	0.90 (0.14)	0.52 (0.13)	0.82 (0.21)
DKPCA	0.46 (0.10)	0.47 (0.03)	0.59 (0.14)	0.44 (0.04)	0.70 (0.15)	0.41 (0.06)

Table 2 summarizes the averages and the standard deviations of the distances and F_1-scores. Not surprisingly, $\widehat{\boldsymbol{\beta}}_{\text{pool},2}(w)$ performs the best across all scenarios among proposed estimates, followed by $\widehat{\boldsymbol{\beta}}_{\text{dist},2}(w)$. Though $\widehat{\boldsymbol{\beta}}_{\text{dist},2}(w^*)$ performs the worst, it is apparently consistent. The distributed dimension reduction method proposed by Zhu and Zhu (2022) performs worse than our approach. The distributed kernel principal component analysis algorithm has larger distances and fails to recover the support.

S2.3 Simulation study for algorithm 3

We use example 1 and 2 to illustrate the performance of algorithm 3. We run 500 replicates and the consider the following proposed estimates.

1. $\widehat{\boldsymbol{\beta}}_{\text{pool},3}(w)$: The pooled estimate that aggregates all observations together and uses the true weight $w(\mathbf{x}) = \{\sigma^2(\mathbf{x})\}^{-1}$. This serves as a benchmark for algorithm 3.

2. $\widehat{\boldsymbol{\beta}}_{\text{dist},3}(w)$: The distributed estimate that uses $w(\mathbf{x}) = \{\sigma^2(\mathbf{x})\}^{-1}$.

3. $\widehat{\boldsymbol{\beta}}_{\text{dist},3}(w^*)$: The distributed estimate that misspecifies $w(\mathbf{x})$ as $w^*(\mathbf{x}) = 1$.

4. $\widehat{\boldsymbol{\beta}}_{\text{dist},3}(\widehat{w})$: The distributed estimate that estimates $w(\mathbf{x})$ with kernel smoother.

Table 3: The average (aver) and the standard deviations (in the parentheses, std) of the distance of various distributed estimates. The data is generated from example 1 and 2.

(n, m)	(500,5)		(250, 10)		(100, 25)	
	aver	std	aver	std	aver	std
	Example 1					
$\widehat{\boldsymbol{\beta}}_{\text{pool},3}(w)$	0.085 (0.028)					
$\widehat{\boldsymbol{\beta}}_{\text{dist},3}(w)$	0.090	(0.034)	0.092	(0.039)	0.096	(0.041)
$\widehat{\boldsymbol{\beta}}_{\text{dist},3}(\widehat{w})$	0.101	(0.039)	0.105	(0.046)	0.115	(0.054)
$\widehat{\boldsymbol{\beta}}_{\text{dist},3}(w^*)$	0.106	(0.044)	0.117	(0.057)	0.134	(0.068)
	Example 2					
$\widehat{\boldsymbol{\beta}}_{\text{pool},3}(w)$	0.101 (0.031)					
$\widehat{\boldsymbol{\beta}}_{\text{dist},3}(w)$	0.109	(0.033)	0.117	(0.045)	0.134	(0.062)
$\widehat{\boldsymbol{\beta}}_{\text{dist},3}(\widehat{w})$	0.113	(0.041)	0.121	(0.051)	0.141	(0.059)
$\widehat{\boldsymbol{\beta}}_{\text{dist},3}(w^*)$	0.126	(0.055)	0134	(0.062)	0.168	(0.071)

Tables 3 summarizes the averages and the standard deviations (in the parentheses) of the distances. It can be clearly seen that, among all distributed estimates from algorithm 3, $\widehat{\boldsymbol{\beta}}_{\text{dist},3}(w)$ performs the best across all scenarios, followed by $\widehat{\boldsymbol{\beta}}_{\text{dist},3}(\widehat{w})$. Not surprisingly, $\widehat{\boldsymbol{\beta}}_{\text{dist},3}(w^*)$ is the least efficient.

S3. Technical Proofs

S3.1 Proof of Theorem 1

We first provide some auxiliary lemmas that will be used in the proof of Theorem 1.

Lemma 1. *(Mack and Silverman, 1982) Let* $\{(\mathbf{v}_i, Y_i), i = 1, \ldots, n\}$ *be independent and identically distributed random observations and let* $\mathbf{v} \in \mathbb{R}^d$*. Assume that there exist* $r > 1$ *such that* $E(|Y|^r) < \infty$ *and* $\sup_{\mathbf{v}} |Y|^s f(\mathbf{v}, Y)\mathrm{d}Y < \infty$ *where* f *denotes the joint density of* $(\mathbf{v}, Y)$*. The kernel function* K *satisfies the Condition (C1). Then*

$$\sup_{\mathbf{v}} \left| n^{-1} \sum_{i=1}^{n} [K_h(\mathbf{v}_i - \mathbf{v})Y_i - E\{K_h(\mathbf{v}_i - \mathbf{v})Y_i\}] \right| = O_p\left[\left\{ \frac{\log(1/h)}{nh^d} \right\} \right] \quad \text{(S3.1)}$$

provided that $n^{2\delta-1}h \to \infty$ *for some* $\delta < 1 - r^{-1}$*.*

Lemma 2. *(Zhu and Fang, 1996) Let* $\{(\mathbf{v}_i, Y_i), i = 1, \ldots, n\}$ *be independent and identically distributed random observations and let* $\mathbf{v} \in \mathbb{R}^d$*. Assume that the* $(q-1)$*-th derivative of mean function* $m(\mathbf{v}) = E(Y \mid \mathbf{v})$ *is locally Lipschitz continuous. Then*

$$\sup_{\mathbf{v}} |E\{K_h(\mathbf{v}_i - \mathbf{v})Y_i\} - m(\mathbf{v})f(\mathbf{v})| = O(h^q). \quad \text{(S3.2)}$$

Lemma 3. *(Ma and Zhu, 2014) Let* $\{(\mathbf{v}_i, Y_i), i = 1, \ldots, n\}$ *be a random sample. Define* $m(\mathbf{v}) \stackrel{def}{=} E(Y \mid \mathbf{v})$ *and* $\mathbf{m}_1(\mathbf{v}) \stackrel{def}{=} d\{m(\mathbf{v})\}/d\mathbf{v}$ *for* $\mathbf{v} \in \mathbb{R}^d$*.*

The local linear estimate for $m(\mathbf{v})$ *and* $\mathbf{m}_1(\mathbf{v})$ *is defined as*

$$\{\widehat{m}(\mathbf{v}_i), \widehat{\mathbf{m}}_1(\mathbf{v}_i)\} \stackrel{def}{=} \arg\min_{a_i, \mathbf{b}_i} \left[\sum_{k=1}^{n} \{Y_k - a_i - \mathbf{b}_i^T(\mathbf{v}_k - \mathbf{v}_i)\}^2 K_h(\mathbf{v}_k - \mathbf{v}_i) \right].$$

Define $\varepsilon \stackrel{def}{=} Y - m(\mathbf{v})$. *Assume the kernel function* K *satisfies the Condition (C1). The density function of* $\mathbf{v}$, *denoted by* $f(\mathbf{v})$ *is bounded away from zero and infinity, and the* $(q-1)$*-th derivative of* $m(\mathbf{v})$ *is locally Lipschitz continuous. Then*

$$\begin{aligned} & n^{-1} \sum_{i=1}^{n} \varepsilon_i \begin{pmatrix} \widehat{m}(\mathbf{v}_i) - m(\mathbf{v}_i) \\ \widehat{\mathbf{m}}_1(\mathbf{v}_i) - \mathbf{m}_1(\mathbf{v}_i) \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & h\mathbf{I}_{d\times d} \end{pmatrix} \\ & = O_p\left\{ h^q/n^{1/2} + h^{2q} + \log^2 n/(nh^d) \right\}. \end{aligned} \tag{S3.3}$$

Before analyzing the approximate Newton distributed estimate, we establish some auxiliary results. We begin with defining three "good" events. Recall that the initial value condition (C8) and the moment condition (C5) guarantee the existence of a ball $U_\rho \stackrel{\text{def}}{=} \{\boldsymbol{\beta}_0 \in \mathbb{R}^{p\times d} : \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}\| < \rho\}$ such that $\|\widehat{\mathbf{H}}(\boldsymbol{\alpha}_1; \mathbf{x}) - \widehat{\mathbf{H}}(\boldsymbol{\alpha}_2; \mathbf{x})\|_2 \le L(\mathbf{x}, Y)\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|_2$ for all $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in U_\rho$ and any $(\mathbf{x}, Y)$, where $E\{L(\mathbf{x}, Y)\} \le L^4$. In addition, the covariate condition (C4) guarantees that $\mathbf{H}(\boldsymbol{\beta}) \ge \lambda \mathbf{I}_{(p-d)d\times(p-d)d}$. Now, choosing the potentially

smaller radius $\delta_\rho \stackrel{\text{def}}{=} \min\{\rho, \rho\lambda/4L\}$, we can define following events

$$\begin{aligned}
\mathcal{E}_1 &\stackrel{\text{def}}{=} \left\{ n^{-1} \sum_{i=1}^{n} L_j(\mathbf{x}_i, Y_i) \leq 2L,\ j = 1, \ldots, m \right\}, \qquad \text{(S3.4)} \\
\mathcal{E}_2 &\stackrel{\text{def}}{=} \left\{ \left\| \widehat{\mathbf{H}}_j(\boldsymbol{\beta}) - \mathbf{H}_j(\boldsymbol{\beta}) \right\|_2 \leq \rho\lambda/4,\ j = 1, \ldots, m \right\} \\
\mathcal{E}_3 &\stackrel{\text{def}}{=} \left\{ \left| \widehat{E}_j \left[\mathbf{S}\{\mathbf{x}, Y, \boldsymbol{\beta}, w(\mathbf{x})\} \right] \right|_2 \leq (1-\rho)\lambda\delta_\rho/4,\ j = 1, \ldots, m \right\}
\end{aligned}$$

We first show these "good" events hold with high probability.

Lemma 4. *Under Conditions (C1)-(C8), there exist constants C such that*

$$\begin{aligned}
E\left\{ \left| \widehat{E}_j \left[\mathbf{S}\{\mathbf{x}, Y, \boldsymbol{\beta}, w(\mathbf{x})\} \right] \right|_2^4 \right\} &\leq CG^4/n^2, \\
E\left\{ \left\| \widehat{\mathbf{H}}_j(\boldsymbol{\beta}) - \mathbf{H}_j(\boldsymbol{\beta}) \right\|_2^4 \right\} &\leq C(\log 2p)^4 H^4/n^2,
\end{aligned}$$

for $j = 1, \ldots, m$.

The proof of Lemma 4 follows the similar arguments of Lemma 7 in Zhang et al. (2013), thus we omit it here.

Proof. As an immediate consequence of Lemma 4, we see that the events $\mathcal{E}_2$ and $\mathcal{E}_3$ occur with high probability. We further define $\mathcal{E} \stackrel{\text{def}}{=} \mathcal{E}_1 \bigcap \mathcal{E}_2 \bigcap \mathcal{E}_3$, by the Boole's law and the union bound we have $\text{pr}(\mathcal{E}^c) = \text{pr}(\mathcal{E}_1^c \bigcup \mathcal{E}_2^c \bigcup \mathcal{E}_3^c) \leq \text{pr}(\mathcal{E}_1^c) + \text{pr}(\mathcal{E}_2^c) + \text{pr}(\mathcal{E}_3^c)$. For the "bad" event $\mathcal{E}_1^c$, we know that $\text{pr}(\mathcal{E}_1^c) \leq 2^4 E[|n^{-1}\sum_{i=1}^n L_1(\mathbf{x}_i) - E\{L(\mathbf{x})\}|^4/L^4] \leq C_1/n^2$. For $\mathcal{E}_2^c$, it is direct to show $\text{pr}(\mathcal{E}_2^c) \leq 2^4 E\left\{ \left\| \widehat{\mathbf{H}}_j(\boldsymbol{\beta}) - \mathbf{H}_j(\boldsymbol{\beta}) \right\|_2^4 \right\} / \rho^4\lambda^4 \leq C_2(\log 2p)^4 H^4/n^2$. For $\mathcal{E}_3^c$, we

have $\mathrm{pr}(\mathcal{E}_3^c) \leq 2^4 E\left\{\left|\widehat{E}_1\left[\mathbf{S}\{\mathbf{x}, Y, \boldsymbol{\beta}, w(\mathbf{x})\}\right]\right|_2^4 / \{(1-\rho)^4\lambda^4\delta_\rho^4\}\right\} \leq C_3 G^4/n^2$. Here C_1, C_2, C_3 are some universal constants. Consequently, we find that $\mathrm{pr}(\mathcal{E}^c) = O(n^{-2})$. □

Recall the definition $\mathcal{E} = \mathcal{E}_1 \bigcap \mathcal{E}_2 \bigcap \mathcal{E}_3$. We then give a deterministic result under the "good" event $\mathcal{E}$.

Lemma 5. *Under Conditions (C1)-(C8) and event $\mathcal{E}$, we have*

$$\lambda_{\min}\left\{\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}}_{pool,1})\right\} \geq (1-\rho)\lambda/2,\ \left\|\widehat{\mathbf{H}}(\boldsymbol{\beta}_{dist,1}^{(0)}) - \widehat{\mathbf{H}}_1(\boldsymbol{\beta}_{dist,1}^{(0)})\right\|_2 \tag{S3.5}$$
$$\leq \left\{2L\rho/\lambda^2 + (\rho+4)/4\lambda\right\}\left\{\left\|\widehat{\mathbf{H}}(\boldsymbol{\beta}) - \widehat{\mathbf{H}}_1(\boldsymbol{\beta})\right\|_2 + 4L\left|\boldsymbol{\beta}_{dist,1}^{(0)} - \boldsymbol{\beta}\right|_2\right\}.$$

Proof. We can bound $\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}}_{\mathrm{pool},1})$ as $\lambda_{\min}\left\{\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}}_{\mathrm{pool},1})\right\} \geq \lambda_{\min}\{\mathbf{H}(\boldsymbol{\beta})\} - \|\widehat{\mathbf{H}}(\boldsymbol{\beta}) - \mathbf{H}(\boldsymbol{\beta})\|_2 - \|\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}}_{\mathrm{pool},1}) - \widehat{\mathbf{H}}(\boldsymbol{\beta})\|_2 \geq \lambda - \rho\lambda/2 - 2L|\widehat{\boldsymbol{\beta}}_{\mathrm{pool},1} - \boldsymbol{\beta}|_2 \geq (1-\rho)\lambda/2$.

To bound the term $\|\widehat{\mathbf{H}}(\boldsymbol{\beta}_{\mathrm{dist},1}^{(0)}) - \widehat{\mathbf{H}}_1(\boldsymbol{\beta}_{\mathrm{dist},1}^{(0)})\|_2$, we make use of the following inequality, for any matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$,

$$\left\|(\mathbf{A}+\Delta\mathbf{A})^{-1} - \mathbf{A}^{-1}\right\|_2 \leq \left\|\mathbf{A}^{-1}\right\|_2^2 \|\Delta\mathbf{A}\|_2 \tag{S3.6}$$

We choose $\mathbf{A} = \mathbf{H}(\boldsymbol{\beta})$ and $\Delta\mathbf{A} = \widehat{\mathbf{H}}(\boldsymbol{\beta}_{\mathrm{dist},1}^{(0)}) - \mathbf{H}(\boldsymbol{\beta})$, then $\|\Delta\mathbf{A}\|_2 \leq \|\widehat{\mathbf{H}}(\boldsymbol{\beta}_{\mathrm{dist},1}^{(0)}) - \widehat{\mathbf{H}}(\boldsymbol{\beta})\|_2 + \|\widehat{\mathbf{H}}(\boldsymbol{\beta}) - \mathbf{H}(\boldsymbol{\beta})\|_2$. Therefore, $\|\widehat{\mathbf{H}}^{-1}(\boldsymbol{\beta}_{\mathrm{dist},1}^{(0)})\|_2 \leq \|\widehat{\mathbf{H}}^{-1}(\boldsymbol{\beta}_{\mathrm{dist},1}^{(0)}) - \mathbf{H}^{-1}(\boldsymbol{\beta})\|_2 + \lambda^{-1} \leq 2L\lambda^{-2}|\boldsymbol{\beta}_{\mathrm{dist},1}^{(0)} - \boldsymbol{\beta}|_2 + \lambda^{-2}\|\widehat{\mathbf{H}}(\boldsymbol{\beta}) - \mathbf{H}(\boldsymbol{\beta})\|_2 + \lambda^{-1} \leq 2L\lambda^{-2}\rho + \lambda^{-1} + \lambda^{-1}\rho/4$. We then choose $\mathbf{A} = \widehat{\mathbf{H}}(\boldsymbol{\beta}_{\mathrm{dist},1}^{(0)})$ and

$\Delta\mathbf{A} = \widehat{\mathbf{H}}_1(\boldsymbol{\beta}^{(0)}_{\text{dist},1}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}^{(0)}_{\text{dist},1})$. Using inequality (S3.6) again, we obtain $\|\widehat{\mathbf{H}}(\boldsymbol{\beta}^{(0)}_{\text{dist},1}) - \widehat{\mathbf{H}}_1(\boldsymbol{\beta}^{(0)}_{\text{dist},1})\|_2 \leq \|\widehat{\mathbf{H}}(\boldsymbol{\beta}^{(0)}_{\text{dist},1})\|_2^2 \|\widehat{\mathbf{H}}_1(\boldsymbol{\beta}^{(0)}_{\text{dist},1}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}^{(0)}_{\text{dist},1})\|_2$. By the triangle inequality, we know that $\|\widehat{\mathbf{H}}_1(\boldsymbol{\beta}^{(0)}_{\text{dist},1}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}^{(0)}_{\text{dist},1})\|_2 \leq \|\widehat{\mathbf{H}}_1(\boldsymbol{\beta}) - \widehat{\mathbf{H}}(\boldsymbol{\beta})\|_2 + \|\widehat{\mathbf{H}}(\boldsymbol{\beta}^{(0)}_{\text{dist},1}) - \widehat{\mathbf{H}}(\boldsymbol{\beta})\|_2 + \|\widehat{\mathbf{H}}_1(\boldsymbol{\beta}^{(0)}_{\text{dist},1}) - \widehat{\mathbf{H}}_1(\boldsymbol{\beta})\|_2$. As a consequence, $\|\widehat{\mathbf{H}}(\boldsymbol{\beta}^{(0)}_{\text{dist},1}) - \widehat{\mathbf{H}}_1(\boldsymbol{\beta}^{(0)}_{\text{dist},1})\|_2 \leq \|\widehat{\mathbf{H}}(\boldsymbol{\beta}^{(0)}_{\text{dist},1})\|_2^2 \{\|\widehat{\mathbf{H}}_1(\boldsymbol{\beta}) - \widehat{\mathbf{H}}(\boldsymbol{\beta})\|_2 + 4L|\boldsymbol{\beta}^{(0)}_{\text{dist},1} - \boldsymbol{\beta}|_2\}$. With the result of $\|\widehat{\mathbf{H}}^{-1}(\boldsymbol{\beta}^{(0)}_{\text{dist},1})\|_2$, we complete the proof. □

With Lemma 4 and 5, we proceed to prove Theorem 1.

Proof. We begin with defining the global one-step Newton-Raphson estimate,

$$\text{vecl}(\boldsymbol{\beta}^{(1)}_{\text{pool},1}) = \text{vecl}(\boldsymbol{\beta}^{(0)}) + \left\{\widehat{\mathbf{H}}(\boldsymbol{\beta}^{(0)})\right\}^{-1} \widehat{E}_{\text{pool},1}\Big[\mathbf{S}\{\mathbf{x}, Y, \boldsymbol{\beta}^{(0)}.w(\mathbf{x})\}\Big].$$

Here we assume that the global and distributed estimate share the same initial value $\boldsymbol{\beta}^{(0)}$. Recall the update scheme (3.2), we know that

$$\text{vecl}(\boldsymbol{\beta}^{(1)}_{\text{dist},1}) = \text{vecl}(\boldsymbol{\beta}^{(0)}) + \left\{\widehat{\mathbf{H}}_1(\boldsymbol{\beta}^{(0)})\right\}^{-1} \widehat{E}_{\text{dist},1}\Big[\mathbf{S}\{\mathbf{x}, Y, \boldsymbol{\beta}^{(0)}, w(\mathbf{x})\}\Big].$$

The error can be decomposed as $\text{vecl}(\boldsymbol{\beta}^{(1)}_{\text{dist},1} - \widehat{\boldsymbol{\beta}}_{\text{pool},1}) = \text{vecl}(\boldsymbol{\beta}^{(1)}_{\text{dist},1} - \boldsymbol{\beta}^{(1)}_{\text{pool},1}) + \text{vecl}(\boldsymbol{\beta}^{(1)}_{\text{pool},1} - \widehat{\boldsymbol{\beta}}_{\text{pool},1})$.

We analyze the two terms separately. The first term $\text{vecl}(\boldsymbol{\beta}^{(1)}_{\text{dist},1} - \boldsymbol{\beta}^{(1)}_{\text{pool},1})$ can be expressed as $\left\{\widehat{\mathbf{H}}_1(\boldsymbol{\beta}^{(0)})\right\}^{-1} \widehat{E}_{\text{dist},1}\Big[\mathbf{S}\{\mathbf{x}, Y, \boldsymbol{\beta}^{(0)}, w(\mathbf{x})\}\Big] - \left\{\widehat{\mathbf{H}}(\boldsymbol{\beta}^{(0)})\right\}^{-1}$

$\widehat{E}_{\text{pool},1}\Big[\mathbf{S}\{\mathbf{x}, Y, \boldsymbol{\beta}^{(0)}, w(\mathbf{x})\}\Big] = T_1 + T_2$, where $T_1 \overset{\text{def}}{=} \Big\{\widehat{\mathbf{H}}_1(\boldsymbol{\beta}^{(0)})\Big\}^{-1}$ $\widehat{E}_{\text{dist},1}\Big[\mathbf{S}\{\mathbf{x}, Y, \boldsymbol{\beta}^{(0)}, w(\mathbf{x})\}\Big] - \Big\{\widehat{\mathbf{H}}_1(\boldsymbol{\beta}^{(0)})\Big\}^{-1} \widehat{E}_{\text{pool},1}\Big[\mathbf{S}\{\mathbf{x}, Y, \boldsymbol{\beta}^{(0)}, w(\mathbf{x})\}\Big]$ and $T_2 \overset{\text{def}}{=} \Big\{\widehat{\mathbf{H}}_1(\boldsymbol{\beta}^{(0)})\Big\}^{-1} \widehat{E}_{\text{pool},1}\Big[\mathbf{S}\{\mathbf{x}, Y, \boldsymbol{\beta}^{(0)}, w(\mathbf{x})\}\Big] - \Big\{\widehat{\mathbf{H}}(\boldsymbol{\beta}^{(0)})\Big\}^{-1} \widehat{E}_{\text{pool},1}\Big[\mathbf{S}\{\mathbf{x}, Y, \boldsymbol{\beta}^{(0)}, w(\mathbf{x})\}\Big]$. For term T_1, we obtain $\widehat{E}_j\Big[\mathbf{S}\{\mathbf{x}, Y, \boldsymbol{\beta}^{(0)}, w(\mathbf{x})\}\Big] - E\Big[\mathbf{S}\{\mathbf{x}, Y, \boldsymbol{\beta}^{(0)}, w(\mathbf{x})\}\Big] = O_p\{h_1^q(h_1^q + h_2^q + h_3^q + h_4^q) + h_1^q/n^{1/2} + \log n/\{n(h_1^q h_2^q)^{1/2}\} + \log n/\{n(h_1^q h_3^q)^{1/2}\} + \log n/\{n(h_1^q h_4^q)^{1/2}\} + \log n/(nh_1^q)\}$ by the proof in Ma and Zhu (2014). With the simple averaging aggregation procedure $\widehat{E}_{\text{dist},1}\Big[\mathbf{S}\{\mathbf{x}, Y, \boldsymbol{\beta}^{(0)}, w(\mathbf{x})\}\Big] = m^{-1}\sum_{j=1}^{m} \widehat{E}_j\Big[\mathbf{S}\{\mathbf{x}, Y, \boldsymbol{\beta}^{(0)}, w(\mathbf{x})\}\Big]$, the bias term remains the same while the variance term becomes m times smaller. Specifically, we have $\widehat{E}_{\text{dist},1}\Big[\mathbf{S}\{\mathbf{x}, Y, \boldsymbol{\beta}^{(0)}, w(\mathbf{x})\}\Big] - E\Big[\mathbf{S}\{\mathbf{x}, Y, \boldsymbol{\beta}^{(0)}, w(\mathbf{x})\}\Big] = O_p\{h_1^q(h_1^q + h_2^q + h_3^q + h_4^q) + h_1^q/n^{1/2} + \log n/\{N(h_1^q h_2^q)^{1/2}\} + \log n/\{N(h_1^q h_3^q)^{1/2}\} + \log n/\{N(h_1^q h_4^q)^{1/2}\} + \log n/(Nh_1^q)\}$. With the bandwidths condition, we conclude that $T_1 = O_p(N^{-1/2})$. For term T_2, recall that $\widehat{E}_{\text{pool},1}\Big[\mathbf{S}\{\mathbf{x}, Y, \widehat{\boldsymbol{\beta}}_{\text{pool},1}, w(\mathbf{x})\}\Big] = \mathbf{0}$. Then $T_2 = \left[\Big\{\widehat{\mathbf{H}}_1(\boldsymbol{\beta}^{(0)})\Big\}^{-1} - \Big\{\widehat{\mathbf{H}}(\boldsymbol{\beta}^{(0)})\Big\}^{-1}\right]\Big\{\widehat{E}_{\text{pool},1}\Big[\mathbf{S}\{\mathbf{x}, Y, \boldsymbol{\beta}^{(0)}, w(\mathbf{x})\}\Big] - \widehat{E}_{\text{pool},1}\Big[\mathbf{S}\{\mathbf{x}, Y, \widehat{\boldsymbol{\beta}}_{\text{pool},1}, w(\mathbf{x})\}\Big]\Big\}$. By the proof of Ma and Zhu (2014), we achieve $\widehat{E}_{\text{pool},1}\Big[\mathbf{S}\{\mathbf{x}, Y, \boldsymbol{\beta}^{(0)}, w(\mathbf{x})\}\Big] - \widehat{E}_{\text{pool},1}\Big[\mathbf{S}\{\mathbf{x}, Y, \widehat{\boldsymbol{\beta}}_{\text{pool},1}, w(\mathbf{x})\}\Big] = \text{vecl}(\boldsymbol{\beta}^{(0)} - \widehat{\boldsymbol{\beta}}_{\text{pool},1}) + o_p(N^{-1/2})$. With these arguments, we show that $\Big|\text{vecl}(\boldsymbol{\beta}^{(1)}_{\text{dist},1} - \boldsymbol{\beta}^{(1)}_{\text{pool},1})\Big|_2 \leq C\left\|\Big\{\widehat{\mathbf{H}}_1(\boldsymbol{\beta}^{(0)})\Big\}^{-1} - \Big\{\widehat{\mathbf{H}}(\boldsymbol{\beta}^{(0)})\Big\}^{-1}\right\|_2 \Big|\text{vecl}(\boldsymbol{\beta}^{(0)} - \widehat{\boldsymbol{\beta}}_{\text{pool},1})\Big|_2$. The second term $\text{vecl}(\boldsymbol{\beta}^{(1)}_{\text{pool},1} - \widehat{\boldsymbol{\beta}}_{\text{pool},1})$ can be analyzed by using Theorem 5.3 in Bubeck et al. (2015), which yields $\Big|\text{vecl}(\boldsymbol{\beta}^{(1)}_{\text{pool},1} - \widehat{\boldsymbol{\beta}}_{\text{pool},1})\Big|_2 \leq L/\lambda_{\min}\Big\{\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}}_{\text{pool},1})\Big\}\Big|\text{vecl}(\boldsymbol{\beta}^{(0)} - \widehat{\boldsymbol{\beta}}_{\text{pool},1})\Big|_2^2$.

With Lemma 4 and 5, we conclude that $\left|\text{vecl}(\boldsymbol{\beta}^{(1)}_{\text{dist},1} - \widehat{\boldsymbol{\beta}}_{\text{pool},1})\right|_2 \leq C' \left\{ \left|\text{vecl}(\boldsymbol{\beta}^{(0)} - \widehat{\boldsymbol{\beta}}_{\text{pool},1})\right|_2 + \left|\text{vecl}(\widehat{\boldsymbol{\beta}}_{\text{pool},1} - \boldsymbol{\beta})\right|_2 + \left\|\widehat{\mathbf{H}}(\boldsymbol{\beta}) - \widehat{\mathbf{H}}_1(\boldsymbol{\beta})\right\|_2 \right\} \left|\text{vecl}(\boldsymbol{\beta}^{(0)} - \widehat{\boldsymbol{\beta}}_{\text{pool},1})\right|_2$. Then we turn to control $\left\|\widehat{\mathbf{H}}(\boldsymbol{\beta}) - \widehat{\mathbf{H}}_1(\boldsymbol{\beta})\right\|_2$. Recall that

$$\widehat{\mathbf{H}}_1(\boldsymbol{\beta}) \;=\; n^{-1} \sum_{k=1}^{n} \left[w(\mathbf{x}_{k,1}) \left\{ \widehat{\mathbf{x}}_{k,1}(\boldsymbol{\beta}) \right\} \left\{ \widehat{\mathbf{x}}_{k,1}(\boldsymbol{\beta}) \right\}^{\mathrm{T}} \right],$$

where

$$\widehat{\mathbf{x}}_{k,1}(\boldsymbol{\beta}) \;=\; \text{vecl} \left\{ \left[\mathbf{x}_{k,1} - \frac{\widehat{E}\{\mathbf{x}_{k,1} w(\mathbf{x}_{k,1}) \mid \mathbf{x}_{k,1}^{\mathrm{T}} \boldsymbol{\beta}\}}{\widehat{E}\{w(\mathbf{x}_{k,1}) \mid \mathbf{x}_{k,1}^{\mathrm{T}} \boldsymbol{\beta}\}} \right] \widehat{\mathbf{m}}_1^{\mathrm{T}}(\mathbf{x}_{k,1}^{\mathrm{T}} \boldsymbol{\beta}) \right\}.$$

We further define

$$\widetilde{\mathbf{H}}_1(\boldsymbol{\beta}) \;\stackrel{\text{def}}{=}\; n^{-1} \sum_{k=1}^{n} \left[w(\mathbf{x}_{k,1}) \left\{ \widetilde{\mathbf{x}}_{k,1}(\boldsymbol{\beta}) \right\} \left\{ \widetilde{\mathbf{x}}_{k,1}(\boldsymbol{\beta}) \right\}^{\mathrm{T}} \right],$$

where

$$\widetilde{\mathbf{x}}_{k,1}(\boldsymbol{\beta}) \;\stackrel{\text{def}}{=}\; \text{vecl} \left\{ \left[\mathbf{x}_{k,1} - \frac{E\{\mathbf{x}_{k,1} w(\mathbf{x}_{k,1}) \mid \mathbf{x}_{k,1}^{\mathrm{T}} \boldsymbol{\beta}\}}{E\{w(\mathbf{x}_{k,1}) \mid \mathbf{x}_{k,1}^{\mathrm{T}} \boldsymbol{\beta}\}} \right] \mathbf{m}_1^{\mathrm{T}}(\mathbf{x}_{k,1}^{\mathrm{T}} \boldsymbol{\beta}) \right\}.$$

Invoking the triangle inequality, we get $\left\|\widehat{\mathbf{H}}_1(\boldsymbol{\beta}) - \mathbf{H}(\boldsymbol{\beta})\right\|_2 \leq \left\|\widehat{\mathbf{H}}_1(\boldsymbol{\beta}) - \widetilde{\mathbf{H}}_1(\boldsymbol{\beta})\right\|_2 + \left\|\widetilde{\mathbf{H}}_1(\boldsymbol{\beta}) - \mathbf{H}(\boldsymbol{\beta})\right\|_2$. We will show $\left\|\widehat{\mathbf{H}}_1(\boldsymbol{\beta}) - \widetilde{\mathbf{H}}_1(\boldsymbol{\beta})\right\|_2 = o_p(n^{-1/2})$ and $\left\|\widetilde{\mathbf{H}}_1(\boldsymbol{\beta}) - \mathbf{H}(\boldsymbol{\beta})\right\|_2 = O_p(n^{-1/2})$.

To control $\left\|\widehat{\mathbf{H}}_1(\boldsymbol{\beta}) - \widetilde{\mathbf{H}}_1(\boldsymbol{\beta})\right\|_2$, we introduce an intermediate variable $\overline{\mathbf{x}}_{k,1}(\boldsymbol{\beta})$,

$$\overline{\mathbf{x}}_{k,1}(\boldsymbol{\beta}) \;\stackrel{\text{def}}{=}\; \text{vecl} \left\{ \left[\mathbf{x}_{k,1} - \frac{E\{\mathbf{x}_{k,1} w(\mathbf{x}_{k,1}) \mid \mathbf{x}_{k,1}^{\mathrm{T}} \boldsymbol{\beta}\}}{E\{w(\mathbf{x}_{k,1}) \mid \mathbf{x}_{k,1}^{\mathrm{T}} \boldsymbol{\beta}\}} \right] \widehat{\mathbf{m}}_1^{\mathrm{T}}(\mathbf{x}_{k,1}^{\mathrm{T}} \boldsymbol{\beta}) \right\}.$$

We then have $\left\|\widehat{\mathbf{H}}_1(\boldsymbol{\beta}) - \widetilde{\mathbf{H}}_1(\boldsymbol{\beta})\right\|_2 \leq 2M_1 + M_2 + 2M_3 + M_4$, where $M_1 \stackrel{\text{def}}{=} \left\|n^{-1}\sum_{k=1}^{n}\left[w(\mathbf{x}_{k,1})\left\{\widehat{\mathbf{x}}_{k,1}(\boldsymbol{\beta}) - \overline{\mathbf{x}}_{k,1}(\boldsymbol{\beta})\right\}\left\{\overline{\mathbf{x}}_{k,1}(\boldsymbol{\beta})\right\}^{\mathrm{T}}\right]\right\|_2$, $M_2 \stackrel{\text{def}}{=} \left\|n^{-1}\sum_{k=1}^{n}\left[w(\mathbf{x}_{k,1})\left\{\widehat{\mathbf{x}}_{k,1}(\boldsymbol{\beta}) - \overline{\mathbf{x}}_{k,1}(\boldsymbol{\beta})\right\}\left\{\widehat{\mathbf{x}}_{k,1}(\boldsymbol{\beta}) - \overline{\mathbf{x}}_{k,1}(\boldsymbol{\beta})\right\}^{\mathrm{T}}\right]\right\|_2$, $M_3 \stackrel{\text{def}}{=} \left\|n^{-1}\sum_{k=1}^{n}\left[w(\mathbf{x}_{k,1})\left\{\overline{\mathbf{x}}_{k,1}(\boldsymbol{\beta}) - \widetilde{\mathbf{x}}_{k,1}(\boldsymbol{\beta})\right\}\left\{\widetilde{\mathbf{x}}_{k,1}(\boldsymbol{\beta})\right\}^{\mathrm{T}}\right]\right\|_2$, $M_4 \stackrel{\text{def}}{=} \left\|n^{-1}\sum_{k=1}^{n}\left[w(\mathbf{x}_{k,1})\left\{\overline{\mathbf{x}}_{k,1}(\boldsymbol{\beta}) - \widetilde{\mathbf{x}}_{k,1}(\boldsymbol{\beta})\right\}\left\{\overline{\mathbf{x}}_{k,1}(\boldsymbol{\beta}) - \widetilde{\mathbf{x}}_{k,1}(\boldsymbol{\beta})\right\}^{\mathrm{T}}\right]\right\|_2$. By Lemma 3 and condition (C5), we know that M_1, M_2, M_3 and M_4 are of order $o_p(n^{-1/2})$. Thus $\left\|\widehat{\mathbf{H}}_1(\boldsymbol{\beta}) - \widetilde{\mathbf{H}}_1(\boldsymbol{\beta})\right\|_2 = o_p(n^{-1/2})$. Next we show that $\left\|\widetilde{\mathbf{H}}_1(\boldsymbol{\beta}) - \mathbf{H}(\boldsymbol{\beta})\right\|_2 = O_p(n^{-1/2})$.

Notice that $\left\|\widetilde{\mathbf{H}}_1(\boldsymbol{\beta}) - \mathbf{H}(\boldsymbol{\beta})\right\|_2 = \sup_{\mathbf{u}\in\mathcal{S}^{(p-d)d-1}} g(\mathbf{u})$, where $g(\mathbf{u}) = \mathbf{u}^{\mathrm{T}}\Big(n^{-1}\sum_{k=1}^{n}\Big[w(\mathbf{x}_{k,1})\{\widetilde{\mathbf{x}}_{k,1}(\boldsymbol{\beta})\}\{\widetilde{\mathbf{x}}_{k,1}(\boldsymbol{\beta})\}^{\mathrm{T}}\Big] - E\Big[w(\mathbf{x})\{\widetilde{\mathbf{x}}(\boldsymbol{\beta})\}\{\widetilde{\mathbf{x}}(\boldsymbol{\beta})\}^{\mathrm{T}}\Big]\Big)\mathbf{u}$. Let $\mathcal{N}$ be a 1/4-covering of cone $\mathcal{S}^{(p-d)d-1}$ with cardinality $|\mathcal{N}| \leq 9^{(p-d)d}$. We further denote $\widehat{\mathbf{u}} = \arg\max_{\mathbf{u}} g(\mathbf{u})$. If we can find such $\widetilde{\mathbf{u}} \in \mathcal{N}$ that $\|\widehat{\mathbf{u}} - \widetilde{\mathbf{u}}\| \leq 1/4$, then we have $|g(\widetilde{\mathbf{u}}) - g(\widehat{\mathbf{u}})| \leq g(\widehat{\mathbf{u}})/2$. Thus $\sup_{\mathbf{u}\in\mathcal{S}^{p-1}} g(\mathbf{u}) \leq 2\sup_{\mathbf{u}\in\mathcal{N}^{p-1}} g(\mathbf{u})$. For any $\mathbf{u} \in \mathcal{N}$ and $\varepsilon \geq 0$, we further have $\Pr(g(\mathbf{u}) \geq \varepsilon/2) \leq 2\exp\{-c_2\min(\varepsilon, \varepsilon^2)n\}$ by Bernstein's inequality. Therefore $\Pr(\left\|\widetilde{\mathbf{H}}_1(\boldsymbol{\beta}) - \mathbf{H}(\boldsymbol{\beta})\right\|_2 \geq \varepsilon) \leq 2\exp\{c_1(p-d)d - c_2\min(\varepsilon, \varepsilon^2)n\}$. By choosing $\varepsilon = \{(p-d)d/n\}^{1/2}$, we have $\left\|\widetilde{\mathbf{H}}_1(\boldsymbol{\beta}) - \mathbf{H}(\boldsymbol{\beta})\right\|_2 = O_p(n^{-1/2})$. Aggregating the above arguments, we get $\left\|\widehat{\mathbf{H}}_1(\boldsymbol{\beta}) - \mathbf{H}(\boldsymbol{\beta})\right\|_2 = O_p(n^{-1/2})$. By a similar procedure, we know that $\left\|\widehat{\mathbf{H}}(\boldsymbol{\beta}) - \mathbf{H}(\boldsymbol{\beta})\right\|_2 = O_p(N^{-1/2})$.

Putting the pieces together, we obtain $\left|\mathrm{vecl}(\boldsymbol{\beta}^{(1)}_{\mathrm{dist},1} - \widehat{\boldsymbol{\beta}}_{\mathrm{pool},1})\right|_2 \leq C/n^{1/2}$

$\left|\text{vecl}(\boldsymbol{\beta}^{(0)}_{\text{dist},1} - \widehat{\boldsymbol{\beta}}_{\text{pool},1})\right|_2$. Applying the triangle inequality, we know that $\left|\text{vecl}(\boldsymbol{\beta}^{(1)}_{\text{dist},1} - \boldsymbol{\beta})\right|_2 \leq C(n^{-1} + N^{-1/2})$ with high probability. By replacing the initial value, we can apply the one-round result recursively and achieve $\left|\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},1} - \boldsymbol{\beta})\right|_2 \leq C(n^{-(t+1)/2} + N^{-1/2})$ with high probability. □

S3.2 Proof of Theorem 2

We first provide some technical lemmas that will be used in the proof of Theorem 2.

Lemma 6. *We define*

$$\widetilde{\mathbf{H}}_1(\boldsymbol{\beta}) \overset{def}{=} n^{-1}\sum_{k=1}^{n}\left[w(\mathbf{x}_{k,1})\left\{\widetilde{\mathbf{x}}_{k,1}(\boldsymbol{\beta})\right\}\left\{\widetilde{\mathbf{x}}_{k,1}(\boldsymbol{\beta})\right\}^T\right], \tag{S3.7}$$

and

$$\widetilde{\mathbf{H}}(\boldsymbol{\beta}) \overset{def}{=} N^{-1}\sum_{k=1}^{N}\left[w(\mathbf{x}_{k})\left\{\widetilde{\mathbf{x}}_{k}(\boldsymbol{\beta})\right\}\left\{\widetilde{\mathbf{x}}_{k}(\boldsymbol{\beta})\right\}^T\right], \tag{S3.8}$$

where

$$\widetilde{\mathbf{x}}_k(\boldsymbol{\beta}) \overset{def}{=} vecl\left\{\left[\mathbf{x}_k - \frac{E\{\mathbf{x}_k w(\mathbf{x}_k) \mid \mathbf{x}_k^T\boldsymbol{\beta}\}}{E\{w(\mathbf{x}_k) \mid \mathbf{x}_k^T\boldsymbol{\beta}\}}\right]\mathbf{m}_1^T(\mathbf{x}_k^T\boldsymbol{\beta})\right\}.$$

Under Conditions (C1)-(C4) and (C5')-(C8'), we have

$$\begin{aligned}\left\|\widetilde{\mathbf{H}}_1(\boldsymbol{\beta}) - \mathbf{H}(\boldsymbol{\beta})\right\|_\infty &\leq 2C_1^{-1/2}(\log p/n)^{1/2},\\ \left\|\widetilde{\mathbf{H}}(\boldsymbol{\beta}) - \mathbf{H}(\boldsymbol{\beta})\right\|_\infty &\leq 2C_1^{-1/2}(\log p/N)^{1/2},\end{aligned} \tag{S3.9}$$

with probability at least $1 - 2/p^2$.

Proof. We let $G_{i,j}^k \stackrel{\text{def}}{=} \{\widetilde{\mathbf{x}}_{k,1}(\boldsymbol{\beta})\}_i\{\widetilde{\mathbf{x}}_{k,1}(\boldsymbol{\beta})\}_j - \mathbf{H}_{i,j}(\boldsymbol{\beta}), k = 1,\ldots,n, i = 1,\ldots,(p-d)d, j = 1,\ldots,(p-d)d$, where $\mathbf{H}_{i,j}(\boldsymbol{\beta})$ is the element in i-th row and j-th column of $\mathbf{H}(\boldsymbol{\beta})$. We can verify that $\widetilde{\mathbf{x}}_k(\boldsymbol{\beta})$ is a sub-gaussian random variable. By Remark 5.18 of Vershynin (2018), we know that both $\{\widetilde{\mathbf{x}}_{k,1}(\boldsymbol{\beta})\}_i\{\widetilde{\mathbf{x}}_{k,1}(\boldsymbol{\beta})\}_j$ and $G_{i,j}^k$ are sub-exponential random variables. According to Corollary 5.17 of Vershynin (2018), we have $\text{pr}(|n^{-1}\sum_{k=1}^n G_{i,j}^k| \geq t) \leq 2\exp\{-\min(C_1t^2, C_2t)n\}$, where C_1 and C_2 are generic constants independent of pair (i,j). By a union bound over all (i,j) pairs,

$$\text{pr}\left(\left\|\widetilde{\mathbf{H}}_1(\boldsymbol{\beta}) - \mathbf{H}(\boldsymbol{\beta})\right\|_\infty \geq t\right) \leq 2p^2\exp\{-\min(C_1t^2, C_2t)n\}.$$

Through assuming $n \geq 4C_1C_2^{-1}\log p$, we have $\left\|\widetilde{\mathbf{H}}_1(\boldsymbol{\beta}) - \mathbf{H}(\boldsymbol{\beta})\right\|_\infty$ $\leq 2C_1^{-1/2}(\log p/n)^{1/2}$ with high probability by setting $t = 2C_1^{-1/2}(\log p/n)^{1/2}$. Following similar arguments, we can establish (S3.9). □

Lemma 7. *We define*

$$\widetilde{Y}_k(\boldsymbol{\beta}) \stackrel{\text{def}}{=} \{\widetilde{\mathbf{x}}_k(\boldsymbol{\beta})\}vecl(\boldsymbol{\beta}) + \{Y_k - m(\mathbf{x}_k^T\boldsymbol{\beta})\}. \tag{S3.10}$$

Under Conditions (C1)-(C4) and (C5')-(C8'), we have

$$\left|N^{-1}\sum_{j=1}^m\sum_{i=1}^n\left[\widetilde{\mathbf{x}}_{i,j}(\boldsymbol{\beta})\widetilde{Y}_{i,j}(\boldsymbol{\beta}) - E\left\{\widetilde{\mathbf{x}}_{i,j}(\boldsymbol{\beta})\widetilde{Y}_{i,j}(\boldsymbol{\beta})\right\}\right]\right|_\infty$$
$$\leq 2C_1^{-1/2}(\log p/N)^{1/2}, \tag{S3.11}$$

with probability at least $1 - 2/p^3$.

Proof. We let $G_k^{i,j} \stackrel{\text{def}}{=} \{\widetilde{\mathbf{x}}_{i,j}(\boldsymbol{\beta})\}_k \widetilde{Y}_{i,j}(\boldsymbol{\beta}) - E\left[\{\widetilde{\mathbf{x}}_{i,j}(\boldsymbol{\beta})\}_k \widetilde{Y}_{i,j}(\boldsymbol{\beta})\right], i = 1, \ldots, n, j = 1, \ldots, m, k = 1, \ldots, (p-d)d$. By Remark 5.18 of Vershynin (2018), $G_k^{i,j}$ is a sub-exponential variable. Through applying Corollary 5.17 of Vershynin (2018), we know that $\text{pr}(|N^{-1}\sum_{j=1}^{m}\sum_{i=1}^{n} G_k^{i,j}| \geq t) \leq 2\exp\{-\min(C_1 t^2, C_2 t)N\}$, where C_1 and C_2 are generic constants independent of k. By a union bound over k, we have

$$\begin{aligned}
&\text{pr}\left\{\left|N^{-1}\sum_{j=1}^{m}\sum_{i=1}^{n}\left[\widetilde{\mathbf{x}}_{i,j}(\boldsymbol{\beta})\widetilde{Y}_{i,j}(\boldsymbol{\beta}) - E\left\{\widetilde{\mathbf{x}}_{i,j}(\boldsymbol{\beta})\widetilde{Y}_{i,j}(\boldsymbol{\beta})\right\}\right]\right|_{\infty} \geq t\right\} \\
&\leq 2p\exp\{-\min(C_1 t^2, C_2 t)N\}.
\end{aligned}$$

We set $t = 2C_1^{-1/2}(\log p/N)^{1/2}$ under the assumption $N \geq 4C_1 C_2^{-1}\log p$ and complete our proof. □

Lemma 8. *Under Conditions (C1)-(C4) and (C5')-(C8'), if* $n \geq 2C_2^2(s + \log p)$*, we have*

$$\begin{aligned}
\left\|\widetilde{\mathbf{H}}_{1,\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta}) - \mathbf{H}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})\right\|_{\text{op}} &\leq 2^{1/2}C_2 c_0\{(s + \log p/n)\}^{1/2}, \\
\left\|\widetilde{\mathbf{H}}^{-1}_{1,\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta}) - \mathbf{H}^{-1}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})\right\|_{\text{op}} &\leq 2^{1/2}C_2 c_0^3\{(s + \log p/n)\}^{1/2}, \quad \text{(S3.12)}
\end{aligned}$$

with probability at least $1 - 2p^{-C_1 C_2^2}$*, where* C_1 *and* C_2 *are generic positive constants depending on the sub-gaussian norm of* $\mathbf{x}_{i,1}$*.*

Proof. By Remark 5.40 in Vershynin (2018), for $t \geq 0$, with probability at least $1 - 2p^{-C_1 C_2^2}$, we know that $\left\|\widetilde{\mathbf{H}}_{1,\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta}) - \mathbf{H}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})\right\|_{\text{op}} \leq \max(\delta, \delta^2)$,

where $\delta = C_2(s/n)^{1/2} + t/n^{1/2}$, and C_2 depends on the maximum of $\mathbf{x}_{i,1}\mathbf{x}_{i,1}^{\mathrm{T}}$ in sub-exponential norm. Set $t = C_2(\log p)^{1/2}$. If $n \geq 2C_2^2(s + \log p)$, we establish $\left\|\widetilde{\mathbf{H}}_{1,\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta}) - \mathbf{H}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})\right\|_{\mathrm{op}} \leq 2^{1/2}C_2c_0\{(s+\log p/n)\}^{1/2}$ with high probability.

For any $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{s\times s}$, $\|(\mathbf{A}+\mathbf{B})^{-1} - \mathbf{A}^{-1}\|_{\mathrm{op}} \leq \|\mathbf{A}^{-1}\|_{\mathrm{op}}^2 \|\mathbf{B}\|_{\mathrm{op}}$. By setting $\mathbf{A} = \mathbf{H}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})$ and $\mathbf{B} = \widetilde{\mathbf{H}}_{1,\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta}) - \mathbf{H}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})$, we obtain that $\left\|\widetilde{\mathbf{H}}_{1,\mathcal{S}\times\mathcal{S}}^{-1}(\boldsymbol{\beta}) - \mathbf{H}_{\mathcal{S}\times\mathcal{S}}^{-1}(\boldsymbol{\beta})\right\|_{\mathrm{op}} \leq \left\|\mathbf{H}_{\mathcal{S}\times\mathcal{S}}^{-1}(\boldsymbol{\beta})\right\|_{\mathrm{op}}^2 \left\|\widetilde{\mathbf{H}}_{1,\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta}) - \mathbf{H}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})\right\|_{\mathrm{op}}$. Together with the result of $\left\|\widetilde{\mathbf{H}}_{1,\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta}) - \mathbf{H}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})\right\|_{\mathrm{op}}$, we then complete the proof. □

We then provide a deterministic result, which shows that the estimation error upper bound of $\boldsymbol{\beta}_{\mathrm{dist},2}^{(t)}$ is actually proportional to the regularization parameter $\lambda_N^{(t)}$.

Lemma 9. *Let* $\mathcal{C} \stackrel{def}{=} \{\boldsymbol{\delta} \in \mathbb{R}^p : |\boldsymbol{\delta}|_1 \leq 4s^{1/2}|\boldsymbol{\delta}|_2\}$. *In addition to Conditions (C1)-(C4) and (C5')-(C8'), we further assume*

$$\left|\widehat{\mathbf{H}}_1(\boldsymbol{\beta}_{dist,2}^{(t-1)})vecl(\boldsymbol{\beta}) - \mathbf{z}_N(\boldsymbol{\beta}_{dist,2}^{(t-1)}) - \{\widehat{\mathbf{H}}_1(\boldsymbol{\beta}_{dist,2}^{(t-1)}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}_{dist,2}^{(t-1)})\}vecl(\boldsymbol{\beta}_{dist,2}^{(t-1)})\right|_\infty \leq \lambda_N^{(t)}/2, \tag{S3.13}$$

and

$$\boldsymbol{\delta}^{T}\widehat{\mathbf{H}}_1(\boldsymbol{\beta}_{dist,2}^{(t-1)})\boldsymbol{\delta} \geq \gamma|\boldsymbol{\delta}|_2^2, \tag{S3.14}$$

hold for any $\boldsymbol{\delta} \in \mathcal{C}$ *and some constant* $\gamma > 0$. *Then we have*

$$\left| vecl(\boldsymbol{\beta}^{(t)}_{dist,2} - \boldsymbol{\beta}) \right|_2 \leq 6s^{1/2}\gamma^{-1}\lambda^{(t)}_N, \tag{S3.15}$$

and

$$\left| vecl(\boldsymbol{\beta}^{(t)}_{dist,2} - \boldsymbol{\beta}) \right|_1 \leq 24s^{1/2}\gamma^{-1}\lambda^{(t)}_N, \tag{S3.16}$$

where $t \geq 1$.

Proof. For simplicity, we define $\mathbf{A} \overset{\text{def}}{=} \widehat{\mathbf{H}}_1(\boldsymbol{\beta}^{(t-1)}_{\text{dist},2})$ and $\mathbf{b} \overset{\text{def}}{=} \mathbf{z}_N(\boldsymbol{\beta}^{(t-1)}_{\text{dist},2}) + (\widehat{\mathbf{H}}_1(\boldsymbol{\beta}^{(t-1)}_{\text{dist},2}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}^{(t-1)}_{\text{dist},2}))\text{vecl}(\boldsymbol{\beta}^{(t-1)}_{\text{dist},2})$. By the definition of $\boldsymbol{\beta}$, we have $|\boldsymbol{\beta}|_1 = |\boldsymbol{\beta}_{\mathcal{S}}|_1$. We define

$$\begin{aligned}\mathcal{Q} \overset{\text{def}}{=}& \ \text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2})^{\text{T}}\mathbf{A}\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2})/2 - \text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2})^{\text{T}}\mathbf{b} \\ &-\{\text{vecl}(\boldsymbol{\beta})^{\text{T}}\mathbf{A}\text{vecl}(\boldsymbol{\beta})/2 - \text{vecl}(\boldsymbol{\beta})\mathbf{b}\}.\end{aligned}$$

On one hand, by definition of $\boldsymbol{\beta}^{(t)}_{\text{dist},2}$, we have $\mathcal{Q} \leq \lambda^{(t)}_N\{|\text{vecl}(\boldsymbol{\beta})|_1 - |\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2})|_1\}$. Since $|\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2})|_1 = |\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2,\mathcal{S}})|_1 + |\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2,\mathcal{S}^c})|_1$, we have $|\text{vecl}(\boldsymbol{\beta})|_1 - |\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2})|_1 \leq |\text{vecl}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}_{\text{dist},2})_{\mathcal{S}}|_1 - |\text{vecl}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}_{\text{dist},2})_{\mathcal{S}^c}|_1$. By these arguments, we achieve the upper bound of $\mathcal{Q}$,

$$\mathcal{Q} \leq \lambda^{(t)}_N|\text{vecl}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}_{\text{dist},2})_{\mathcal{S}}|_1 - \lambda_N|\text{vecl}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}_{\text{dist},2})_{\mathcal{S}^c}|_1. \tag{S3.17}$$

On the other hand, since $\mathbf{A}$ is non-negative definite, $\text{vecl}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}_{\text{dist},2})^{\text{T}}\mathbf{A}\text{vecl}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}_{\text{dist},2}) \geq 0$. This implies $\mathcal{Q} \geq (\mathbf{A}\text{vecl}(\boldsymbol{\beta}) - \mathbf{b})^{\text{T}}\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2} - \boldsymbol{\beta})$. By Hölder

inequality, $(\mathbf{A}\text{vecl}(\boldsymbol{\beta})-\mathbf{b})^{\mathrm{T}}\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2}-\boldsymbol{\beta}) \geq -|\mathbf{A}\text{vecl}(\boldsymbol{\beta})-\mathbf{b}|_\infty|\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2}-\boldsymbol{\beta})|_1$. With assumption (S3.13), we achieve the lower bound of $\mathcal{Q}$,

$$\mathcal{Q} \geq -\lambda_N^{(t)}|\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2}-\boldsymbol{\beta})|_1/2. \tag{S3.18}$$

Combine (S3.17) and (S3.18) to obtain $|\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2}-\boldsymbol{\beta})_{\mathcal{S}^c}|_1 \leq 3|\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2}-\boldsymbol{\beta})_{\mathcal{S}}|_1$. This implies $|\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2}-\boldsymbol{\beta})|_1 \leq 4|\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2}-\boldsymbol{\beta})_{\mathcal{S}}|_1 \leq 4s^{1/2}|\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2}-\boldsymbol{\beta})_{\mathcal{S}}|_2 \leq 4s^{1/2}|\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2}-\boldsymbol{\beta})|_2$. Consequently, $\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2}-\boldsymbol{\beta}) \in \mathcal{C}$.

By the first order condition of the surrogate loss function (4.3), we have $\mathbf{A}\text{vecl}(\boldsymbol{\beta})^{(t)}_{\text{dist},2} - \mathbf{b} \in \lambda_N\partial|\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2})|_1$. This implies $|\mathbf{A}\text{vecl}(\boldsymbol{\beta})^{(t)}_{\text{dist},2} - \mathbf{b}|_\infty \leq \lambda_N^{(t)}$. Combine (S3.13) to obtain $|\mathbf{A}\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2}-\boldsymbol{\beta})|_\infty \leq 3/2\lambda_N$. This, together with (S3.14) entails that $\gamma|\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2}-\boldsymbol{\beta})|_2^2 \leq \text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2}-\boldsymbol{\beta})^{\mathrm{T}}\mathbf{A}\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2}-\boldsymbol{\beta})$. By Hölder inequality, $\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2}-\boldsymbol{\beta})^{\mathrm{T}}\mathbf{A}\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2}-\boldsymbol{\beta}) \leq |\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2}-\boldsymbol{\beta})|_1|\mathbf{A}\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2}-\boldsymbol{\beta})|_\infty$. Consequently, $|\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2}-\boldsymbol{\beta})|_2^2 \leq 6s^{1/2}\gamma^{-1}\lambda_N^{(t)}|\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2}-\boldsymbol{\beta})|_2$. Accordingly, we have $|\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2}-\boldsymbol{\beta})|_2 \leq 6s^{1/2}\gamma^{-1}\lambda_N^{(t)}$ and $|\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2}-\boldsymbol{\beta})|_1 \leq 24s^{1/2}\gamma^{-1}\lambda_N^{(t)}$. This completes the proof. □

The following lemmas show that the conditions (S3.13) and (S3.14) in Lemma 9 hold with high probability.

Lemma 10. *Under Conditions (C1)-(C4) and (C5')-(C8'), we have for*

$t \geq 1$

$$\left|\mathbf{z}_N(\boldsymbol{\beta}_{dist,2}^{(t-1)}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}_{dist,2}^{(t-1)})vecl(\boldsymbol{\beta})\right|_\infty \leq 4C_1^{-1/2}(\log p/N)^{1/2}, \qquad \text{(S3.19)}$$

holds with probability at least $1 - 2/p^3 - 2/p^2$.

Proof. By the triangular inequality, $\left|\mathbf{z}_N(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)})\text{vecl}(\boldsymbol{\beta})\right|_\infty \leq Q_1 + Q_2$, where $Q_1 \stackrel{\text{def}}{=} \left|\widehat{\mathbf{H}}(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)})\text{vecl}(\boldsymbol{\beta}) - \mathbf{H}(\boldsymbol{\beta})\text{vecl}(\boldsymbol{\beta})\right|_\infty$ and $Q_2 \stackrel{\text{def}}{=} \left|\mathbf{H}(\boldsymbol{\beta})\text{vecl}(\boldsymbol{\beta}) - \mathbf{z}_N(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)})\right|_\infty$.

We first control term Q_1. It is direct to show

$$\begin{aligned} Q_1 \leq & \left|\widehat{\mathbf{H}}(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)})\text{vecl}(\boldsymbol{\beta}) - \widehat{\mathbf{H}}(\boldsymbol{\beta})\text{vecl}(\boldsymbol{\beta})\right|_\infty \\ & + \left|\widehat{\mathbf{H}}(\boldsymbol{\beta})\text{vecl}(\boldsymbol{\beta}) - \widetilde{\mathbf{H}}(\boldsymbol{\beta})\text{vecl}(\boldsymbol{\beta})\right|_\infty \\ & + \left|\widetilde{\mathbf{H}}(\boldsymbol{\beta})\text{vecl}(\boldsymbol{\beta}) - \mathbf{H}(\boldsymbol{\beta})\text{vecl}(\boldsymbol{\beta})\right|_\infty . \end{aligned}$$

By the similar arguments in Lemma S.2 of Ma et al. (2019), we know that $\left|\widehat{\mathbf{H}}(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)})\text{vecl}(\boldsymbol{\beta}) - \widehat{\mathbf{H}}(\boldsymbol{\beta})\text{vecl}(\boldsymbol{\beta})\right|_\infty = o_p\{(\log p/N)^{1/2}\}$. Applying Lemma S.1 of Ma et al. (2019), we have $\left|\widehat{\mathbf{H}}(\boldsymbol{\beta})\text{vecl}(\boldsymbol{\beta}) - \widetilde{\mathbf{H}}(\boldsymbol{\beta})\text{vecl}(\boldsymbol{\beta})\right|_\infty = o_p\{(\log p/N)^{1/2}\}$. According to Lemma 6, we establish $\left|\widetilde{\mathbf{H}}(\boldsymbol{\beta})\text{vecl}(\boldsymbol{\beta}) - \mathbf{H}(\boldsymbol{\beta})\text{vecl}(\boldsymbol{\beta})\right|_\infty \leq 2C_1^{-1/2}(\log p/N)^{1/2}$ with probability at least $1 - 2/p^2$.

By the definition of Q_2, we have

$$Q_2 = \left|N^{-1}\sum_{j=1}^{m}\sum_{i=1}^{n}\left[\widehat{\mathbf{x}}_{i,j}(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)})\widehat{Y}_{i,j}(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)}) - E\left\{\widetilde{\mathbf{x}}_{i,j}(\boldsymbol{\beta})\widetilde{Y}_{i,j}(\boldsymbol{\beta})\right\}\right]\right|_\infty .$$

Applying triangular inequality repeatedly, we conclude $Q_2 \leq T_1 + T_2 + T_3 + T_4 + T_5$, where

$$T_1 \stackrel{\text{def}}{=} \left| N^{-1} \sum_{j=1}^{m} \sum_{i=1}^{n} \left[\widehat{\mathbf{x}}_{i,j}(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)}) \widehat{Y}_{i,j}(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)}) - \widehat{\mathbf{x}}_{i,j}(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)}) \widehat{Y}_{i,j}(\boldsymbol{\beta}) \right] \right|_{\infty},$$

$$T_2 \stackrel{\text{def}}{=} \left| N^{-1} \sum_{j=1}^{m} \sum_{i=1}^{n} \left[\widehat{\mathbf{x}}_{i,j}(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)}) \widehat{Y}_{i,j}(\boldsymbol{\beta}) - \widehat{\mathbf{x}}_{i,j}(\boldsymbol{\beta}) \widehat{Y}_{i,j}(\boldsymbol{\beta}) \right] \right|_{\infty},$$

$$T_3 \stackrel{\text{def}}{=} \left| N^{-1} \sum_{j=1}^{m} \sum_{i=1}^{n} \left[\widehat{\mathbf{x}}_{i,j}(\boldsymbol{\beta}) \widehat{Y}_{i,j}(\boldsymbol{\beta}) - \widetilde{\mathbf{x}}_{i,j}(\boldsymbol{\beta}) \widehat{Y}_{i,j}(\boldsymbol{\beta}) \right] \right|_{\infty},$$

$$T_4 \stackrel{\text{def}}{=} \left| N^{-1} \sum_{j=1}^{m} \sum_{i=1}^{n} \left[\widetilde{\mathbf{x}}_{i,j}(\boldsymbol{\beta}) \widehat{Y}_{i,j}(\boldsymbol{\beta}) - \widetilde{\mathbf{x}}_{i,j}(\boldsymbol{\beta}) \widetilde{Y}_{i,j}(\boldsymbol{\beta}) \right] \right|_{\infty},$$

$$T_5 \stackrel{\text{def}}{=} \left| N^{-1} \sum_{j=1}^{m} \sum_{i=1}^{n} \left[\widetilde{\mathbf{x}}_{i,j}(\boldsymbol{\beta}) \widetilde{Y}_{i,j}(\boldsymbol{\beta}) - E\left\{ \widetilde{\mathbf{x}}_{i,j}(\boldsymbol{\beta}) \widetilde{Y}_{i,j}(\boldsymbol{\beta}) \right\} \right] \right|_{\infty}.$$

With Lemma S.2 of Ma et al. (2019), we have term T_1 and T_2 are of order $o_p\{(\log p/N)^{1/2}\}$. By Lemma S.1 of Ma et al. (2019), we know that term T_3 and T_4 are of order $o_p\{(\log p/N)^{1/2}\}$. Applying Lemma 7, we establish $T_5 \leq 2C_1^{-1/2}(\log p/N)^{1/2}$ with probability at least $1 - 2/p^3$. Combing the results of Q_1 and Q_2, we complete our proof. □

Lemma 11. *Under Conditions (C1)-(C4) and (C5')-(C8'), we set*

$$\lambda_N^{(t)} = 8C_0C_1^{-1/2}\left\{(\log p/N)^{1/2} + (\log p/n)^{1/2}|vecl(\boldsymbol{\beta}_{dist,2}^{(t-1)} - \boldsymbol{\beta})|_1 + |vecl(\boldsymbol{\beta}_{dist,2}^{(t-1)} - \boldsymbol{\beta})|_2^2\right\},$$

define the event $\mathcal{E}^{(t)}$

$$\left|\widehat{\mathbf{H}}_1(\boldsymbol{\beta}_{dist,2}^{(t-1)})vecl(\boldsymbol{\beta}) - \mathbf{z}_N(\boldsymbol{\beta}_{dist,2}^{(t-1)}) - \{\widehat{\mathbf{H}}_1(\boldsymbol{\beta}_{dist,2}^{(t-1)}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}_{dist,2}^{(t-1)})\}vecl(\boldsymbol{\beta}_{dist,2}^{(t-1)})\right|_{\infty}$$
$$\leq \lambda_N^{(t)}/2.$$

Then $\bigcap_{t\geq 1}\mathcal{E}^{(t)}$ *holds with probability at least* $1-2/p^3-2/p^2$.

Proof. We define the event

$$\mathcal{E}_0 \overset{\text{def}}{=} \left\{\left|\mathbf{z}_N(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)})\text{vecl}(\boldsymbol{\beta})\right|_\infty \leq 4C_1^{-1/2}(\log p/N)^{1/2}\right\}$$
$$\bigcap\left\{\left|\widehat{\mathbf{H}}_1(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)})\right|_\infty \leq 4C_1^{-1/2}(\log p/N)^{1/2}\right\}.$$

For the bound $\left|\widehat{\mathbf{H}}_1(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)})\right|_\infty$, we know that

$$\begin{aligned}
&\left|\widehat{\mathbf{H}}_1(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)})\right|_\infty \\
&\leq \left|\widehat{\mathbf{H}}_1(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)}) - \widehat{\mathbf{H}}_1(\boldsymbol{\beta})\right|_\infty + \left|\widehat{\mathbf{H}}_1(\boldsymbol{\beta}) - \widetilde{\mathbf{H}}_1(\boldsymbol{\beta})\right|_\infty + \left|\widetilde{\mathbf{H}}_1(\boldsymbol{\beta}) - \mathbf{H}(\boldsymbol{\beta})\right|_\infty \\
&+ \left|\widehat{\mathbf{H}}(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)}) - \widehat{\mathbf{H}}(\boldsymbol{\beta})\right|_\infty + \left|\widehat{\mathbf{H}}(\boldsymbol{\beta}) - \widetilde{\mathbf{H}}(\boldsymbol{\beta})\right|_\infty + \left|\widetilde{\mathbf{H}}(\boldsymbol{\beta}) - \mathbf{H}(\boldsymbol{\beta})\right|_\infty.
\end{aligned}$$

Then with Lemma 6, 10, Lemma S.1 and S.2 of Ma et al. (2019), we can establish $\text{pr}(\mathcal{E}_0) \geq 1-2/p^3-2/p^2$.

We then show $\mathcal{E}_0 \subset \bigcap_{t\geq 1}\mathcal{E}^{(t)}$. Actually, by the triangular inequality, we get $\left|\widehat{\mathbf{H}}_1(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)})\text{vecl}(\boldsymbol{\beta}) - \mathbf{z}_N(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)}) - \{\widehat{\mathbf{H}}_1(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)})\}\text{vecl}(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)})\right|_\infty \leq$ $\left|\mathbf{z}_N(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)})\text{vecl}(\boldsymbol{\beta})\right|_\infty + \left|\{\widehat{\mathbf{H}}_1(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)})\}\{\text{vecl}(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)} - \boldsymbol{\beta})\}\right|_\infty$. By Hölder inequality, we further have $\left|\{\widehat{\mathbf{H}}_1(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)})\}\{\text{vecl}(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)} - \boldsymbol{\beta})\}\right|_\infty \leq$ $\left|\{\widehat{\mathbf{H}}_1(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)})\}\right|_\infty \left|\{\text{vecl}(\boldsymbol{\beta}_{\text{dist},2}^{(t-1)} - \boldsymbol{\beta})\}\right|_1$. With the choice of the regularization parameter $\lambda_N^{(t)}$, we achieve $\mathcal{E}_0 \subset \mathcal{E}^{(t)}$ for each $t \geq 1$. Finally, we conclude $\mathcal{E}_0 \subset \bigcap_{t\geq 1}\mathcal{E}^{(t)}$ and $\mathcal{E}_0$ holds with high probability, which completes the proof. □

Lemma 12. *Under Conditions (C1)-(C4) and (C5')-(C8'),*

$$\boldsymbol{\delta}^{T}\widehat{\mathbf{H}}_1(\boldsymbol{\beta}_{dist,2}^{(t-1)})\boldsymbol{\delta} \geq \gamma|\boldsymbol{\delta}|_2^2,$$

holds with probability at least $1 - 4/p^2$.

Proof. We define events

$$\mathcal{E}_1^{(t)} \stackrel{\text{def}}{=} \boldsymbol{\delta}^{\mathrm{T}}\widehat{\mathbf{H}}_1(\boldsymbol{\beta}_{\mathrm{dist},2}^{(t-1)})\boldsymbol{\delta} \geq \gamma|\boldsymbol{\delta}|_2^2,$$

and

$$\mathcal{E}_2 \stackrel{\text{def}}{=} \boldsymbol{\delta}^{\mathrm{T}}\widetilde{\mathbf{H}}_1(\boldsymbol{\beta})\boldsymbol{\delta} \geq \gamma|\boldsymbol{\delta}|_2^2.$$

We know that $\mathcal{E}_2 \subset \mathcal{E}_1^{(t)}$ for each $t \geq 1$. By the similar arguments in Lemma D.3 of Huang et al. (2018), we achieve

$$\frac{\lambda_{\min}\{\mathbf{H}(\boldsymbol{\beta})\}}{68[4\kappa\lambda_{\min}\{\mathbf{H}(\boldsymbol{\beta})\}+1]^2}\boldsymbol{\delta}^{\mathrm{T}}\widetilde{\mathbf{H}}_1(\boldsymbol{\beta})\boldsymbol{\delta} \geq |\boldsymbol{\delta}|_2^2$$

hold with probability at least $1 - 4/p^2$. We conclude $\mathcal{E}_2 \subset \bigcap_{t\geq 1}\mathcal{E}_1^{(t)}$ and $\mathcal{E}_2$ holds with high probability, which completes the proof. □

At last, we present the proof of Theorem 2.

Proof. We first give the error bound of the distributed estimate after one round of iteration. With Lemma 9, we conclude that the error bound is proportional to the regularization parameter $\lambda_N^{(1)}$. Furthermore, according

to Lemma 10, 11 and 12, the conditions in Lemma 9 are satisfied with high probability. It is direct to show that

$$\left|\text{vecl}(\boldsymbol{\beta}^{(1)}_{\text{dist},2} - \boldsymbol{\beta})\right|_2 = O_p\left\{(s\log p/N)^{1/2} + s^{3/2}\log p/n\right\}.$$

We then turn to the results after multi rounds of iteration. The proof proceeds by recursively applying the error bound result with one round iteration. Let $a \stackrel{\text{def}}{=} 4C_a s(\log p/N)^{1/2}$ and $b \stackrel{\text{def}}{=} 4C_b s(\log p/n)^{1/2}$. If we treat $\text{vecl}(\boldsymbol{\beta}^{(t-1)}_{\text{dist},2}$ as the initial value, we will achieve

$$\begin{aligned}
&\left|\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2} - \boldsymbol{\beta})\right|_1 \le a + b\left|\text{vecl}(\boldsymbol{\beta}^{(t-1)}_{\text{dist},2} - \boldsymbol{\beta})\right|_1, \\
&\left|\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2} - \boldsymbol{\beta})\right|_2 \le 1/4 s^{-1/2}\left\{a + b\left|\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2} - \boldsymbol{\beta})\right|_1\right\}.
\end{aligned}$$

Through applying the inequality recursively and sum a geometric sequence, we get

$$\begin{aligned}
&\left|\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2} - \boldsymbol{\beta})\right|_1 \le (1-b^t)(1-b)^{-1}a + b^t\left|\text{vecl}(\boldsymbol{\beta}^{(0)}_{\text{dist},2} - \boldsymbol{\beta})\right|_1, \\
&\left|\text{vecl}(\boldsymbol{\beta}^{(t)}_{\text{dist},2} - \boldsymbol{\beta})\right|_2 \le 1/4 s^{-1/2}\left\{(1-b^{t+1})(1-b)^{-1}a + b^{t+1}\left|\text{vecl}(\boldsymbol{\beta}^{(0)}_{\text{dist},2} - \boldsymbol{\beta})\right|_1\right\}.
\end{aligned}$$

By the definition of a and b, we complete the proof. □

S3.3 Proof of Theorem 3

We use the primal-dual witness construction (Wainwright, 2009, PDW for short) to prove Theorem 3. We first present the result after one round iteration, which mainly includes the following three steps:

1. Let $\widetilde{\boldsymbol{\beta}}$ be the solution of the following semi–definite programming (SDP):

$$\widetilde{\boldsymbol{\beta}} \stackrel{\text{def}}{=} \arg\min_{\boldsymbol{\alpha}_{\mathcal{S}^c}=0} \Big[\text{vecl}(\boldsymbol{\alpha})^{\text{T}}\widehat{\mathbf{H}}_1(\boldsymbol{\beta}^{(0)}_{\text{dist},2})\text{vecl}(\boldsymbol{\alpha})/2 + \lambda_N^{(1)}|\boldsymbol{\alpha}|_1 \quad \text{(S3.20)}$$
$$-\text{vecl}(\boldsymbol{\alpha})^{\text{T}}\{\mathbf{z}_N(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) + (\widehat{\mathbf{H}}_1(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2}))\text{vecl}(\boldsymbol{\beta}^{(0)}_{\text{dist},2})\}\Big].$$

The solution $\widetilde{\boldsymbol{\beta}}$ is unique as long as $\widehat{\mathbf{H}}_{1,\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2})$ is invertible almost surely.

2. Let $\widetilde{\mathbf{Z}}$ be the subgradient, which satisfies the zero-mean condition:

$$\widehat{\mathbf{H}}_1(\boldsymbol{\beta}^{(0)}_{\text{dist},2})\text{vecl}(\widetilde{\boldsymbol{\beta}}) + \lambda_N^{(1)}\widetilde{\mathbf{Z}} \quad \text{(S3.21)}$$
$$-\{\mathbf{z}_N(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) + (\widehat{\mathbf{H}}_1(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2}))\text{vecl}(\boldsymbol{\beta}^{(0)}_{\text{dist},2})\} = 0.$$

In particular, $\widetilde{\mathbf{Z}}_{\mathcal{S}} \in |\widetilde{\boldsymbol{\beta}}_{\mathcal{S}}|_1$ and satisfies

$$\widehat{\mathbf{H}}_{1,\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2})\text{vecl}(\widetilde{\boldsymbol{\beta}})_{\mathcal{S}} + \lambda_N^{(1)}\widetilde{\mathbf{Z}}_{\mathcal{S}} \quad \text{(S3.22)}$$
$$-\{\mathbf{z}_N(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) + (\widehat{\mathbf{H}}_1(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2}))\text{vecl}(\boldsymbol{\beta}^{(0)}_{\text{dist},2})\}_{\mathcal{S}} = 0.$$

3. Construct $\widetilde{\mathbf{Z}}_{\mathcal{S}^c}$ as

$$\widetilde{\mathbf{Z}}_{\mathcal{S}^c} \stackrel{\text{def}}{=} -\lambda_N^{(1)}\Big[\{\widehat{\mathbf{H}}_1(\boldsymbol{\beta}^{(0)}_{\text{dist},2})\text{vecl}(\widetilde{\boldsymbol{\beta}})\}_{\mathcal{S}^c} \quad \text{(S3.23)}$$
$$-\big\{\mathbf{z}_N(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) + (\widehat{\mathbf{H}}_1(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2}))\text{vecl}(\boldsymbol{\beta}^{(0)}_{\text{dist},2})\big\}_{\mathcal{S}^c}\Big].$$

Define $Z_j \stackrel{\text{def}}{=} (\widetilde{\mathbf{Z}}_{\mathcal{S}^c})_j$ for $j \in \mathcal{S}^c$. We check if $|Z_j| < 1$ uniformly for all $j \in \mathcal{S}^c$.

The first and second steps are obvious. We require Lemma 13 to prove the third step.

Lemma 13. *Assume the conditions in Theorem 3. With probability approaching one, we have* $|Z_j| \leq v$ *uniformly for* $j \in \mathcal{S}^c$*, for some* $0 < v < 1$.

Proof. Define

$$\mathcal{D}_1 \stackrel{\text{def}}{=} \widehat{\mathbf{H}}_{1,\mathcal{S}^c\times\mathcal{S}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2})\ \widehat{\mathbf{H}}^{-1}_{1,\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2})\ \mathcal{I},$$

$$\mathcal{D}_2 \stackrel{\text{def}}{=} \widehat{\mathbf{H}}_{1,\mathcal{S}^c\times\mathcal{S}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2})\ \widehat{\mathbf{H}}^{-1}_{1,\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2})\ \widetilde{\mathbf{Z}}_{\mathcal{S}},$$

$$\mathcal{D}_3 \stackrel{\text{def}}{=} \lambda_N^{-1}\{\mathbf{z}_N(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2})\text{vecl}(\boldsymbol{\beta})\}_{\mathcal{S}^c},$$

$$\mathcal{D}_4 \stackrel{\text{def}}{=} \lambda_N^{-1}(\widehat{\mathbf{H}}_{\mathcal{S}^c\times\{1,\ldots,(p-d)d\}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) - \widehat{\mathbf{H}}_{1,\mathcal{S}^c\times\{1,\ldots,(p-d)d\}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2}))\text{vecl}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}_{\text{dist},2}),$$

where $\mathcal{I} \stackrel{\text{def}}{=} -\lambda_N^{-1}\big\{(\mathbf{z}_N(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2})\text{vecl}(\boldsymbol{\beta}))_{\mathcal{S}} + (\widehat{\mathbf{H}}_{\mathcal{S}\times\{1,\ldots,(p-d)d\}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) - \widehat{\mathbf{H}}_{1,\mathcal{S}\times\{1,\ldots,(p-d)d\}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2}))\text{vecl}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}_{\text{dist},2})\big\}$. By definition in (S3.23), we have $\widetilde{\mathbf{Z}}_{\mathcal{S}^c} = \mathcal{D}_1 + \mathcal{D}_2 + \mathcal{D}_3 + \mathcal{D}_4$.

We study $|\mathcal{D}_1|_\infty$, $|\mathcal{D}_2|_\infty$, $|\mathcal{D}_3|_\infty$ and $|\mathcal{D}_4|_\infty$, respectively. We first deal with $\mathcal{D}_1$. By triangle inequality, $|\mathcal{D}_1|_\infty \leq \|\widehat{\mathbf{H}}_{1,\mathcal{S}^c\times\mathcal{S}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2})\widehat{\mathbf{H}}^{-1}_{1,\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2})\|_\infty\, |\mathcal{I}|_\infty$. Let $\mathcal{E}_1 \stackrel{\text{def}}{=} \{\widehat{\mathbf{H}}_{1,\mathcal{S}^c\times\mathcal{S}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) - \mathbf{H}_{\mathcal{S}^c\times\mathcal{S}}(\boldsymbol{\beta})\}\{\widehat{\mathbf{H}}^{-1}_{1,\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) - \mathbf{H}^{-1}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})\}$, $\mathcal{E}_2 \stackrel{\text{def}}{=} \mathbf{H}_{\mathcal{S}^c\times\mathcal{S}}(\boldsymbol{\beta})\{\widehat{\mathbf{H}}^{-1}_{1,\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) - \mathbf{H}^{-1}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})\}$, $\mathcal{E}_3 \stackrel{\text{def}}{=} \{\widehat{\mathbf{H}}_{1,\mathcal{S}^c\times\mathcal{S}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) - \mathbf{H}_{\mathcal{S}^c\times\mathcal{S}}(\boldsymbol{\beta})\}\mathbf{H}^{-1}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})$, and $\mathcal{E}_4 \stackrel{\text{def}}{=} \mathbf{H}_{\mathcal{S}^c\times\mathcal{S}}(\boldsymbol{\beta})\mathbf{H}^{-1}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})$. It follows immediately that $\widehat{\mathbf{H}}_{1,\mathcal{S}^c\times\mathcal{S}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2})\widehat{\mathbf{H}}^{-1}_{1,\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) = \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3 + \mathcal{E}_4$. By the definition of the infinity norm and the operator norm, $\|\widehat{\mathbf{H}}_{1,\mathcal{S}^c\times\mathcal{S}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) - \mathbf{H}_{\mathcal{S}^c\times\mathcal{S}}(\boldsymbol{\beta})\|_\infty \leq s|\widehat{\mathbf{H}}_{1,\mathcal{S}^c\times\mathcal{S}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) - \mathbf{H}_{\mathcal{S}^c\times\mathcal{S}}(\boldsymbol{\beta})|_\infty$

and $\|\widehat{\mathbf{H}}^{-1}_{1,\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) - \mathbf{H}^{-1}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})\|_\infty \leq s^{1/2}\|\widehat{\mathbf{H}}^{-1}_{1,\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) - \mathbf{H}^{-1}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})\|_{\text{op}}$. It follows that $\|\mathcal{E}_1\|_\infty \leq s^{3/2}\|\widehat{\mathbf{H}}_{1,\mathcal{S}^c\times\mathcal{S}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) - \mathbf{H}_{\mathcal{S}^c\times\mathcal{S}}(\boldsymbol{\beta})\|_\infty \|\widehat{\mathbf{H}}^{-1}_{1,\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) - \mathbf{H}^{-1}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})\|_{\text{op}}$. Lemma 8 indicates that $\|\mathcal{E}_1\|_\infty = O_p[s^{3/2}\{(s+\log N)/n\}^{1/2}\{(\log N/n)^{1/2}\}]$. Therefore, $\|\mathcal{E}_1\|_\infty = O_p\{s^2(\log N)/n\}$. Similarly, we can show that $\|\mathcal{E}_2\|_\infty = O_p[s\{(s+\log N)/n\}^{1/2}]$ and $\|\mathcal{E}_3\|_\infty = O_p\{s^{3/2}(\log N/n)^{1/2}\}$. We have, $|\mathcal{D}_1|_\infty \leq (o_p(1) + \|\mathcal{E}_4\|_\infty)\,|\mathcal{I}|_\infty$. By the assumption of Theorem 3, we know that $\|\mathcal{E}_4\|_\infty \leq 1-\alpha$. The proof of Theorem 2 indicates that $\lambda_N|\mathbf{z}_N(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2})\text{vecl}(\boldsymbol{\beta})|_\infty$ is small enough through setting C_0 in λ_N sufficiently large. Assumption on the initial value ensures that $\text{pr}(\text{supp}(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) \subseteq \mathcal{S}) \to 1$. Consequently, $\lambda_N^{-1}|\{\widehat{\mathbf{H}}_{1,\mathcal{S}\times\{1,\ldots,(p-d)d\}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2}) - \widehat{\mathbf{H}}_{\mathcal{S}\times\{1,\ldots,(p-d)d\}}(\boldsymbol{\beta}^{(0)}_{\text{dist},2})\}\text{vecl}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}_{\text{dist},2})|_\infty = O_p\{\lambda_N^{-1}(s\log N/n)^{1/2}|\text{vecl}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}_{\text{dist},2})|_2\} \leq \alpha/12$. In other words, $|\mathcal{D}_1|_\infty \leq \alpha/12$ with probability approaching one. Following similar arguments, we have $|\mathcal{D}_3|_\infty \leq \alpha/12$ and $|\mathcal{D}_4|_\infty \leq \alpha/12$ with probability approaching one. Based on the result $\|\mathcal{E}_4\|_\infty \leq 1-\alpha$ and assumption $|\widetilde{\mathbf{Z}}_{\mathcal{S}}|_\infty \leq 1$, we have $|\mathcal{D}_2|_\infty \leq 1 - \alpha/2$ with probability approaching one. These completes the proof of Lemma 13. □

With Lemma 13, we give the proof of Theorem 3.

Proof. On one hand, we have $\boldsymbol{\beta}^{(1)}_{\text{dist},2} = \widetilde{\boldsymbol{\beta}}$ with probability approaching one. Moreover, by the PDW construction, we have, $\text{pr}\big\{\mathcal{S}(\boldsymbol{\beta}^{(1)}_{\text{dist},2}) \subseteq \mathcal{S}\big\} \to 1$.

On the other hand, $\mathrm{vecl}(\widetilde{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathcal{S}}) = \mathbf{H}^{-1}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})(\mathcal{M}_1 + \mathcal{M}_2 + \mathcal{M}_3 + \mathcal{M}_4)$, where $\mathcal{M}_1 = -\lambda^{(1)}_N \widetilde{\mathbf{Z}}_{\mathcal{S}}$, $\mathcal{M}_2 = -\{\widehat{\mathbf{H}}_{1,\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta}^{(0)}_{\mathrm{dist},2}) - \mathbf{H}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})\}\{\mathrm{vecl}(\widetilde{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathcal{S}})\}$, $\mathcal{M}_3 = -\{\widehat{\mathbf{H}}_{1,\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta}^{(0)}_{\mathrm{dist},2}) - \widehat{\mathbf{H}}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta}^{(0)}_{\mathrm{dist},2})\}\{\mathrm{vecl}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}_{\mathrm{dist},2})\}_{\mathcal{S}}$, $\mathcal{M}_4 = \{\mathbf{z}_N(\boldsymbol{\beta}^{(0)}_{\mathrm{dist},2}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}^{(0)}_{\mathrm{dist},2})\mathrm{vecl}(\boldsymbol{\beta})\}_{\mathcal{S}}$. By the definition of the infinity norm of a matrix, we have $|\mathrm{vecl}(\widetilde{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathcal{S}})|_\infty \le \|\mathbf{H}^{-1}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})\|_\infty(|\mathcal{M}_1|_\infty + |\mathcal{M}_2|_\infty + |\mathcal{M}_3|_\infty + |\mathcal{M}_4|_\infty)$.

Next we study the infinity norms of $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$, respectively. By the definition of $\widetilde{\mathbf{Z}}_{\mathcal{S}}$, we have $|\mathcal{M}_1|_\infty \le \lambda^{(1)}_N$. By the definition of the ℓ_2–norm, we have $|\mathcal{M}_2|_\infty \le |\mathcal{M}_2|_2$. By the definition of the operator norm, we have $|\mathcal{M}_2|_2 \le \|\widehat{\mathbf{H}}_{1,\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta}^{(0)}_{\mathrm{dist},2}) - \mathbf{H}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})\|_{\mathrm{op}}|\mathrm{vecl}(\widetilde{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathcal{S}})|_2$. By Lemma 8, we have $\|\mathbf{H}^{-1}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})\|_\infty |\mathcal{M}_2|_\infty \le C\|\mathbf{H}^{-1}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})\|_\infty\{s(s + \log N)/n\}^{1/2}|\mathrm{vecl}(\widetilde{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathcal{S}})|_\infty$. Following similar arguments, we have $\|\mathbf{H}^{-1}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})\|_\infty|\mathcal{M}_3|_\infty \le C\|\mathbf{H}^{-1}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})\|_\infty|(\mathrm{vecl}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}_{\mathrm{dist},2})_{\mathcal{S}}|_1(\log p/n)^{1/2}$ and $\|\mathbf{H}^{-1}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})\|_\infty|\mathcal{M}_4|_\infty \le C\|\mathbf{H}^{-1}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})\|_\infty(\log p/N)^{1/2}$. Choosing $\lambda^{(1)}_N$ properly, we have $|\mathrm{vecl}(\widetilde{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathcal{S}})|_\infty \le C\|\mathbf{H}^{-1}_{\mathcal{S}\times\mathcal{S}}(\boldsymbol{\beta})\|_\infty \Big\{(\log p/N)^{1/2} + |(\mathrm{vecl}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}_{\mathrm{dist},2})_{\mathcal{S}}|_1(\log p/n)^{1/2}\Big\}$. This, together with the lower bound condition of $\boldsymbol{\beta}_{\mathcal{S}}$, entails that $\mathrm{pr}(\mathcal{S}(\boldsymbol{\beta}^{(1)}_{\mathrm{dist},2}) \supseteq \mathcal{S}) \to 1$, and accordingly, $\mathrm{pr}(\mathcal{S}(\boldsymbol{\beta}^{(1)}_{\mathrm{dist},2}) = \mathcal{S}) \to 1$.

For $t \ge 2$, we replace the initial estimate with $\boldsymbol{\beta}^{(t-1)}_{\mathrm{dist},2}$ and update the lower bound condition on $\boldsymbol{\beta}_{\mathcal{S}}$. By similar arguments, the proof is completed.

□

References

Balcan, M. F., Y. Liang, L. Song, D. Woodruff, and B. Xie (2016). Communication efficient distributed kernel principal component analysis. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 725–734.

Barzilai, J. and J. M. Borwein (1988). Two-point step size gradient methods. *Journal of Numerical Analysis 8*(1), 141–148.

Bubeck, S. et al. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning 8*(3-4), 231–357.

Cai, Z., R. Li, and L. Zhu (2020). Online sufficient dimension reduction through sliced inverse regression. *Journal of Machine Learning Research 21*(10), 1–25.

Huang, J., Y. Jiao, X. Lu, and L. Zhu (2018). Robust decoding from 1-bit compressive sampling with ordinary and regularized least squares. *SIAM Journal on Scientific Computing 40*(4), A2062–A2086.

Lin, Q., Z. Zhao, and J. S. Liu (2019). Sparse sliced inverse regression via lasso. *Journal of the American Statistical Association 114*(528), 1726–1739.

Luo, W. and X. Cai (2016). A new estimator for efficient dimension reduction in regression. *Journal of Multivariate Analysis 145*, 236–249.

Ma, S., L. Zhu, Z. Zhang, C.-L. Tsai, and R. J. Carroll (2019). A robust and efficient approach to causal inference based on sparse sufficient dimension reduction. *The Annals of Statistics 47*(3), 1505 – 1535.

Ma, Y. and L. Zhu (2014). On estimation efficiency of the central mean subspace. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(5), 885–901.

Mack, Y.-p. and B. W. Silverman (1982). Weak and strong uniform consistency of kernel regression estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 61*(3), 405–415.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 58*(1), 267–288.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science / Roman Vershynin*, Volume 47 of *Cambridge series in statistical and probabilistic mathematics*. Cambridge: Cambridge University Press.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory 55*(5), 2183–2202.

Xia, Y., H. Tong, W. K. Li, and L.-X. Zhu (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(3), 363–410.

Zhang, Y., J. C. Duchi, and M. J. Wainwright (2013). Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research 14*(1), 3321–3363.

Zhu, L. and K. Fang (1996). Asymptotics for kernel estimate of sliced inverse regression. *The Annals of Statistics 24*(3), 1053–1068.

Zhu, Z. and L. Zhu (2022). Distributed dimension reduction with nearly oracle rate. *Statistical Analysis and Data Mining 15*(6), 692–706.