SUPPLEMENT TO "DISTRIBUTED ESTIMATION AND INFERENCE FOR SEMI-PARAMETRIC BINARY RESPONSE MODELS"

The supplementary material is organized as follows. In Appendix A, we provide the detailed theoretical results for the high-dimensional `(mSMSE)`, as well as data-dependent choices of the unknown parameters. In Appendix B, we give the technical proofs for all the theoretical results. In Appendix C, we discuss the asymptotic performance of our estimator over a class of models. Additional discussions and results in the simulations are given in Appendix D.

APPENDIX A: THEORETICAL RESULTS OF THE HIGH-DIMENSIONAL MULTI-ROUND SMSE

In this section, we give the complete theoretical analysis of the high-dimensional multi-round SMSE in Algorithm 3. First, we restate the conditions as the following three assumptions.

ASSUMPTION 11. *Assume that the initial value $\widehat{\boldsymbol{\beta}}^{(0)}$ satisfies*

$$\left\|\widehat{\boldsymbol{\beta}}^{(0)}-\boldsymbol{\beta}^*\right\|_2=O_{\mathbb{P}}\left(\delta_{m,0}\right),\quad \left\|\widehat{\boldsymbol{\beta}}^{(0)}-\boldsymbol{\beta}^*\right\|_1=O_{\mathbb{P}}\left(\sqrt{s}\delta_{m,0}\right), \tag{30}$$

for some $\delta_{m,0}=o(1)$.

ASSUMPTION 12. *Assume that the dimension $p=O\left(n^{\nu}\right)$ for some $\nu>0$, the local sample size $m=O(n^c)$ for some $0<c<1$, and the sparsity $s=O\left(m^r\right)$ for some $0<r<1/4$.*

ASSUMPTION 13. *Assume that the covariates are uniformly bounded, i.e., there exists $\overline{B}$ such that $\|\boldsymbol{Z}\|_{\infty}\leq\overline{B}$. Further, assume that the covariates have finite second moment, i.e., $\sup_{\|\boldsymbol{v}\|_2=1}\mathbb{E}\left[(\boldsymbol{v}^{\top}\boldsymbol{Z})^2\right]<+\infty$.*

Assumption 11 requires that the error of the initial value can be upper bounded in both ℓ_1 and ℓ_2 norm. Moreover, the bound of the ℓ_1 error is of the same order as $\sqrt{s}$ times the bound of the ℓ_2 error. This can be achieved by the path-following method proposed by Feng et al. (2022), with $\delta_{m,0}=(s\log p/m)^{\alpha/(2\alpha+1)}$. Assumption 12 restricts the dimension p, the local size m, and the sparsity s. The first two requirements on p and m imply that $\log p=O\left(\log n\right)$ and $\log m=O(\log n)$, which is necessary for ensuring that the algorithm converges in finite iterations. The third assumption $S=o(m^r)$ for $0<r<1/4$ is also necessary to ensure the consistency of our estimator, which is further discussed in Remark 7 below. Assumption 13 assumes the uniform boundedness of the predictors, which can be achieved by scaling each element of $\boldsymbol{Z}$ into $[-1,1]$, without impacting $\boldsymbol{\beta}^*$ and Y.

REMARK 7. The assumption $S=o(m^r)$ for $0<r<1/4$ arises from the requirement that $s^{3/2}\delta_{m,0}=o(1)$, which ensures that the estimator can be iteratively refined in the algorithm. As shown in Equation (27) in the main text, the convergence rate of $\widehat{\boldsymbol{\beta}}^{(1)}$ contains a bias term $s^{3/2}\delta_{m,0}^2$, which is an improvement from the initial error $\delta_{m,0}$ if and only if $s^{3/2}\delta_{m,0}=o(1)$. Plugging in the initial error $\delta_{m,0}=(s\log p/m)^{\alpha/(2\alpha+1)}$ yields the requirement $s=o(m^{1/4})$.

This requirement can be relaxed to $s=o(m^{1/2})$ if we assume the sub-Gaussianity of the covariate $\boldsymbol{Z}$ in the high-dimensional settings, since the bias term is reduced to $s^{1/2}\delta_{m,0}^2$

under this assumption. The condition $s = o(m^{1/2})$ is also necessary for ensuring the restricted eigenvalue condition for the local Hessian $V_{m,1}$.

Furthermore, we note that this assumption $s = o(m^{1/4})$ is essentially $s = o\big(m_1^{1/4}\big)$, where m_1 is the local sample size on the first machine, since we compute the initial estimator $\widehat{\boldsymbol{\beta}}^{(0)}$ and the Hessian matrix $V_{m_1,1}$ on the first machine. For the ease of presentation, we assume the sample size m is identical on each local machine, i.e., $m_1 = m = n/L$. If there is a machine where its local sample size is larger, we can choose this machine as the first machine. Further, if a certain amount of the entire data is allowed to be pooled to one machine, we can use that pooled size as the m_1.

Under these assumptions, we formally restate Theorem 5.2:

THEOREM A.1. *For $t = 1, 2, \ldots, T$, define $h^* := \big(\frac{\log p}{n}\big)^{\frac{1}{2\alpha+1}}$ and $r_m := \sqrt{\frac{s^2 \log p}{m(h^*)^3}} + s^{3/2}\delta_{m,0}$. Assume Assumptions 1–4 and 11–13 hold. Then there exists a constant α_0 such that, by choosing a kernel $H'(\cdot)$ with order $\alpha > \alpha_0$, bandwidth $h_t \equiv h^*$, and parameters*

$$\lambda_n^{(t)} = C_\lambda \left[\left(\frac{\log p}{n} \right)^{\frac{\alpha}{2\alpha+1}} + \frac{1}{\sqrt{s}} (r_m)^t \delta_{m,0} \right],$$

with a sufficiently large constant C_λ, we have that $r_m = o(1)$,

$$\left\| \widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^* \right\|_2 = O_{\mathbb{P}}\left(\sqrt{s} \lambda_n^{(t)} \right) \quad \text{and} \quad \left\| \widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^* \right\|_1 \leq 2\sqrt{s} \left\| \widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^* \right\|_2, \tag{31}$$

with probability tending to one.

The proof of Theorem A.1 is given in Section B.5. The upper bound $\sqrt{s}\lambda_n^{(t)}$ in Theorem A.1 denotes the order of the ℓ_2-error of $\widehat{\boldsymbol{\beta}}^{(t)}$, which contains two components:

$$\sqrt{s}\lambda_n^{(t)} = \sqrt{s}\,(\log p/n)^{\alpha/(2\alpha+1)} + (r_m)^t \delta_{m,0}.$$

The first term $\sqrt{s}\,(\log p/n)^{\alpha/(2\alpha+1)}$ is the best rate our method can achieve, and the second term comes from the error of the initial estimator. As long as $r_m = o(1)$, the second term decreases exponentially as t increases, which implies that $\widehat{\boldsymbol{\beta}}^{(t)}$ converges after at most $O(\log n)$ iterations. Concretely, the number of required iterations is

$$\frac{\frac{\alpha}{2\alpha+1}\,(\log n - \log\log p) - \frac{1}{2}\log s + \log \delta_{m,0}}{[-\log(r_m)]}, \tag{32}$$

which is larger than that in the low-dimensional case. Under the Assumption 12, it's easy to see that (32) can be upper bounded by a finite number.

The condition $r_m = o(1)$ can be ensured by choosing a kernel function with order higher than a certain constant α_0. See Remark 8 for explanation.

REMARK 8. To make sure that $r_m = \frac{s^2 \log p}{m(h^*)^3} + s^{3/2}\delta_{m,0} = o(1)$ and $\sqrt{s}\delta_{m,0} = O\big((h^*)^{3/2}\big)$, we need to choose a kernel $H'(\cdot)$ with order α such that $\alpha > \alpha_0 := \max\big\{\frac{3r}{2(1-4r)}, \frac{3}{2c(1-2r)} + \frac{r}{2(1-2r)}\big\}$, where $(\log m)/(\log n) < c < 1$ and $(\log s)/(\log m) < r < 1/4$ are supposed in Assumption 12. Here we plug in the rate $\delta_{m,0} = (s \log p/m)^{\alpha/(2\alpha+1)}$ obtained by the path-following algorithm in Feng et al. (2022). If the order of the kernel is not sufficiently high (i.e., less than α_0), Algorithm 3 still works by choosing $h_m = m^{\frac{r-2\alpha(1-2r)}{3(2\alpha+1)}+\varepsilon}$ for some small constant $\varepsilon > 0$, and the corresponding convergence rate will be $\sqrt{s}h_m^\alpha$.

In Remark 9, we explain the dependency of the parameter $\lambda_n^{(t)}$ on s and provide solutions in cases where s is unknown in practice.

REMARK 9. In Theorem A.1, the tuning parameter $\lambda_n^{(t)}$ for the Dantzig Selector depends on s. This dependence arises from the requirement that for any iteration t, the parameter $\lambda_n^{(t)}$ has to satisfy $\left\|V_{m,1}^{(t)}\boldsymbol{\beta}^* - \left(V_{m,1}^{(t)}\widehat{\boldsymbol{\beta}}^{(t-1)} - U_n^{(t)}\right)\right\|_\infty \le \lambda_n^{(t)}$, i.e., the true parameter $\boldsymbol{\beta}^*$ must lie in the feasible set of the Dantzig Selector (26). This is necessary for ensuring the consistency of the Dantzig Selector. When $t=1$, as shown in the proof of Theorem 5.1 in Section B.5, the order of $\left\|V_{m,1}^{(1)}\boldsymbol{\beta}^* - \left(V_{m,1}^{(1)}\widehat{\boldsymbol{\beta}}^{(0)} - U_n^{(1)}\right)\right\|_\infty$ is $O_{\mathbb{P}}\Big(s\delta_{m,0}^2 + h^\alpha + \sqrt{\frac{\log p}{nh}} + \sqrt{\frac{s\log p}{mh^3}}\delta_{m,0}\Big)$, which depends on s and leads to the choice of $\lambda_n^{(1)}$ in Theorem 5.1. The dependence of $\lambda_n^{(t)}$ for $t>1$ on s follows from a similar reason.

We also note that, if the sparsity s is unknown in practice, one can apply the Lepski's method provided in Section A.1 to estimate s and obtain the same convergence rate as in Theorem A.1. Furthermore, if an upper bound for s is known in practice, say $s = O(m^r)$ for some constant $0 < r < 1/4$, then it is possible to replace s in the above equation with the upper bound, which leads to a slower algorithmic convergence (i.e., requires more rounds to converge) but does not affect the final rate $\sqrt{s}\left(\frac{\log p}{n}\right)^{\frac{\alpha}{2\alpha+1}}$ that our algorithm can achieve.

A.1. Data Dependent Method for Unknown Parameters. In Algorithm 3, the choice of regularization parameters $\{\lambda_n^{(t)}\}$ and the bandwidth parameters $\{h_t\}$ depends on the sparsity s and the smoothness α, which might be unknown in practice. In this section, we provide data-adaptive estimation methods to deal with either unknown s or α using the Lepski's approach (Lepskii, 1991; Feng et al., 2022). We show that the estimators obtained by the data-adaptive method achieve the same convergence rate as that in Algorithm 3.

Unknown s**.** We first consider the case where s is unknown and α is known. Let $\mathcal{S} := \{2^q : q = 0, 1, \ldots, \lfloor \log_2(p) \rfloor\}$, $\delta_{m,0,s'} := \left(\frac{s'\log p}{m}\right)^{\frac{\alpha}{2\alpha+1}}$, and $h^* = \left(\frac{\log p}{n}\right)^{\frac{1}{2\alpha+1}}$. For $s' \in \mathcal{S}$ and $t = 1, 2, \ldots, T$, define $r_{m,s'} := \sqrt{\frac{(s')^2\log p}{m(h^*)^3}} + (s')^{3/2}\delta_{m,0,s'}$ and $\lambda_{n,s'}^{(t)} = C_\lambda\Big[\left(\frac{\log p}{n}\right)^{\frac{\alpha}{2\alpha+1}} + \frac{1}{\sqrt{s'}}(r_{m.s'})^t\delta_{m,0}\Big]$, with a sufficiently large constant C_λ. In the tth iteration, let $\widehat{\boldsymbol{\beta}}_{s'}^{(t)}$ denote the estimator obtained by solving (26) with parameter $\lambda_{n,s'}^{(t)}$ for any $s' \in \mathcal{S}$. Then, with a sufficiently large constant $\overline{C}$, we estimate the sparsity as

(33)
$$\widehat{s}^{(t)} := \min\left\{\widetilde{s} \in \mathcal{S} : \left\|\widehat{\boldsymbol{\beta}}_{\widetilde{s}}^{(t)} - \widehat{\boldsymbol{\beta}}_{s'}^{(t)}\right\|_2 \le \overline{C}\sqrt{s'}\lambda_{n,s'}^{(t)},\ \left\|\widehat{\boldsymbol{\beta}}_{\widetilde{s}}^{(t)} - \widehat{\boldsymbol{\beta}}_{s'}^{(t)}\right\|_1 \le \overline{C}s'\lambda_{n,s'}^{(t)},\ \forall s' > \widetilde{s}\right\}.$$

The convergence rate of $\widehat{\boldsymbol{\beta}}_{\widehat{s}^{(t)}}^{(t)}$ is given by the following theorem.

THEOREM A.2. *Assume the conditions in Theorem 5.2 hold, and then for $t = 1, 2, \ldots, T$, we have*

$$\left\|\widehat{\boldsymbol{\beta}}_{\widehat{s}^{(t)}}^{(t)} - \boldsymbol{\beta}^*\right\|_2 = O_{\mathbb{P}}(\delta_{m,t}), \quad \left\|\widehat{\boldsymbol{\beta}}_{\widehat{s}^{(t)}}^{(t)} - \boldsymbol{\beta}^*\right\|_1 = O_{\mathbb{P}}(\sqrt{s}\delta_{m,t}) \tag{34}$$

where $\delta_{m,t} := \sqrt{s}\left(\frac{\log p}{n}\right)^{\frac{\alpha}{2\alpha+1}} + (r_m)^t\delta_{m,0}$ is the same as the error rate in Theorem 5.2.

Theorem A.2 shows that the data-adaptive estimator $\widehat{\boldsymbol{\beta}}_{\widehat{s}^{(t)}}^{(t)}$ obtained by the Lepski's method achieves the same convergence rate as $\widehat{\boldsymbol{\beta}}^{(t)}$ in Algorithm 3. The proof of Theorem A.2 is provided in Section B.5.

Unknown α. Now we present a data adaptive method for tuning the bandwidth h when the smoothing parameter α is unknown and the sparsity s is known. In this part, we assume that the initial error bound $\delta_{m,0}$ does not depend on α, for example, $\left(\frac{s\log p}{m}\right)^{\frac{1}{3}}$. We then define the bandwidth set as $\mathcal{D}=\left\{2^{-q}, q=0,1,\ldots,\min\left\{\lfloor-\frac{2}{3}\log_2(\sqrt{s}\delta_{m,0})\rfloor, \lfloor-\frac{1}{3}\log_2(s^2\log p/m)\rfloor\right\}\right\}$, which ensures that for all $h'\in\mathcal{D}$, the assumptions $\sqrt{s}\delta_{m,0}=O\big((h')^{\frac{3}{2}}\big)$ and $\frac{s^2\log p}{m(h')^3}=o(1)$ are satisfied. For $h'\in\mathcal{D}$ and $t=1,2,\ldots,T$, define $r_{m,h'}:=\sqrt{\frac{s^2\log p}{m(h')^3}}+s^{3/2}\delta_{m,0}$ and $\lambda_{n,h'}^{(t)}=C_\lambda\left[\sqrt{\frac{\log p}{nh'}}+(h')^\alpha+\frac{1}{\sqrt{s}}(r_{m,h'})^t\delta_{m,0}\right]$, with a sufficiently large constant C_λ. In the tth iteration, let $\widehat{\boldsymbol{\beta}}_{h'}^{(t)}$ denote the estimator obtained by solving (26) with bandwidth h' and parameter $\lambda_{n,h'}^{(t)}$ for any $h'\in\mathcal{D}$. Then, with a sufficiently large constant $\overline{C}$, we estimate the optimal bandwidth as

(35)
$$\widehat{h}^{(t)}:=\max\left\{\widetilde{h}\in\mathcal{D}:\left\|\widehat{\boldsymbol{\beta}}_{\widetilde{h}}^{(t)}-\widehat{\boldsymbol{\beta}}_{h'}^{(t)}\right\|_2\le\overline{C}\sqrt{s}\lambda_{n,h'}^{(t)},\ \left\|\widehat{\boldsymbol{\beta}}_{\widetilde{h}}^{(t)}-\widehat{\boldsymbol{\beta}}_{h'}^{(t)}\right\|_1\le\overline{C}s\lambda_{n,h'}^{(t)},\ \forall h'<\widetilde{h}\right\}.$$

Similar to Theorem A.2, the convergence rate of $\widehat{\boldsymbol{\beta}}_{\widehat{h}^{(t)}}^{(t)}$ is given by the following theorem.

THEOREM A.3. *Assume the conditions in Theorem 5.2 hold, and further assume that $s^{3/2}\delta_{m,0}=o(1)$. Then for $t=1,2,\ldots,T$, we have*

$$\left\|\widehat{\boldsymbol{\beta}}_{\widehat{h}^{(t)}}^{(t)}-\boldsymbol{\beta}^*\right\|_2=O_{\mathbb{P}}\left(\delta_{m,t}\right)\text{ and }\left\|\widehat{\boldsymbol{\beta}}_{\widehat{h}^{(t)}}^{(t)}-\boldsymbol{\beta}^*\right\|_1=O_{\mathbb{P}}\left(\sqrt{s}\delta_{m,t}\right),\tag{36}$$

where $\delta_{m,t}=\sqrt{s}\left(\frac{\log p}{n}\right)^{\frac{\alpha}{2\alpha+1}}+(r_m)^t\delta_{m,0}$ is the same as the error rate in Theorem 5.2.

Theorem A.3 shows that the estimator $\widehat{\boldsymbol{\beta}}_{\widehat{h}^{(t)}}^{(t)}$ obtained by the Lepski's method achieves the same convergence rate as $\widehat{\boldsymbol{\beta}}^{(t)}$ in Algorithm 3. The proof of Theorem A.3 is provided in Section B.5.

APPENDIX B: TECHNICAL PROOF OF THE THEORETICAL RESULTS

B.1. Proof of the Results for the ℓ_2 Error Bound of `(mSMSE)`.

Proof of Proposition 3.2

We first restate Proposition 3.2 in a detailed version. The first step of `(mSMSE)` can be written as

$$\widehat{\boldsymbol{\beta}}^{(1)}-\boldsymbol{\beta}^*=\left(V_{n,h_1}\big(\widehat{\boldsymbol{\beta}}^{(0)}\big)\right)^{-1}U_{n,h_1}\big(\widehat{\boldsymbol{\beta}}^{(0)}\big),\tag{37}$$

where

(38)
$$V_{n,h}(\boldsymbol{\beta})=\nabla^2F_h(\boldsymbol{\beta})=\frac{1}{nh^2}\sum_{i=1}^n(-y_i)H''\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}}{h}\right)\boldsymbol{z}_i\boldsymbol{z}_i^\top,$$

(39)
$$U_{n,h}(\boldsymbol{\beta})=\frac{1}{nh^2}\sum_{i=1}^n(-y_i)H''\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}}{h}\right)\boldsymbol{z}_i\boldsymbol{z}_i^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^*)-\frac{1}{nh}\sum_{i=1}^n(-y_i)H'\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}}{h}\right)\boldsymbol{z}_i.$$

PROPOSITION 3.2. Assume Assumptions 1–5 hold. Further assume that $\left\|\widehat{\boldsymbol{\beta}}^{(0)}-\boldsymbol{\beta}^*\right\|_2 = O_{\mathbb{P}}\left(\delta_{m,0}\right)$, $h_1 = o(1)$ and $\frac{p\log n}{nh_1^3} = o(1)$. We then have

$$\left\|U_{n,h_1}\left(\widehat{\boldsymbol{\beta}}^{(0)}\right)\right\|_2 = O_{\mathbb{P}}\left(\delta_{m,0}^2 + h_1^\alpha + \sqrt{\frac{p}{nh_1}} + \delta_{m,0}\sqrt{\frac{p\log n}{nh_1^3}}\right), \tag{40}$$

$$\left\|V_{n,h_1}\left(\widehat{\boldsymbol{\beta}}^{(0)}\right) - V\right\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{p\log n}{nh_1^3}} + \delta_{m,0} + h_1^\alpha\right), \tag{41}$$

and therefore

$$\left\|\widehat{\boldsymbol{\beta}}^{(1)}-\boldsymbol{\beta}^*\right\|_2 = O_{\mathbb{P}}\left(\delta_{m,0}^2 + h_1^\alpha + \sqrt{\frac{p}{nh_1}} + \delta_{m,0}\sqrt{\frac{p\log n}{nh_1^3}}\right). \tag{42}$$

PROOF OF PROPOSITION 3.2. Throughout the whole proof, without loss of generality, we assume that $\left\|\widehat{\boldsymbol{\beta}}^{(0)}-\boldsymbol{\beta}^*\right\|_2 \le \delta_{m,0}$ with probability approaching one, i.e., we assume the constant in $O_{\mathbb{P}}(\delta_{m,0})$ to be 1. For simplicity, we replace the notation h_1 with h.

Proof of (40)

We first prove (40). It suffices to show that

$$\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\le\delta_{m,0}} \|U_{n,h}(\boldsymbol{\beta})\|_2 = O_{\mathbb{P}}\left(\delta_{m,0}^2 + h^\alpha + \sqrt{\frac{p}{nh}} + \delta_{m,0}\sqrt{\frac{p\log n}{nh^3}}\right), \tag{43}$$

which implies (40) since $\left\|\widehat{\boldsymbol{\beta}}^{(0)}-\boldsymbol{\beta}^*\right\|_2 \le \delta_{m,0}$ with probability approaching one.

By the proof of Lemma 3 in Cai et al. (2010), there exists $\boldsymbol{v}_1,...,\boldsymbol{v}_{5^p} \in \mathbb{R}^p$, s.t. for any $\boldsymbol{v}$ in the unit sphere $\mathbb{S}^{p-1} = \{\boldsymbol{v}\in\mathbb{R}^p : \|\boldsymbol{v}\|_2 = 1\}$, there exists $j_v \in [5^p]$ satisfying $\|\boldsymbol{v}-\boldsymbol{v}_{j_v}\|_2 \le 1/2$. Then we have

$$\|U_{n,h}(\boldsymbol{\beta})\|_2 = \sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \left|\boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta})\right| \le \sup_{j_v\in[5^p]} \left|\boldsymbol{v}_{j_v}^\top U_{n,h}(\boldsymbol{\beta})\right| + \frac{1}{2}\|U_{n,h}(\boldsymbol{\beta})\|_2,$$

and thus

$$\|U_{n,h}(\boldsymbol{\beta})\|_2 \le \sup_{j_v\in[5^p]} 2\left|\boldsymbol{v}_{j_v}^\top U_{n,h}(\boldsymbol{\beta})\right|.$$

Therefore, to show (43), it suffices to show that

$$\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\le\delta_{m,0}} \sup_{j_v\in[5^p]} \left|\boldsymbol{v}_{j_v}^\top U_{n,h}(\boldsymbol{\beta})\right| = O_{\mathbb{P}}\left(\sqrt{\frac{p}{nh}} + \delta_{m,0}\sqrt{\frac{p\log n}{nh^3}} + \delta_{m,0}^2 + h^\alpha\right). \tag{44}$$

Recall the definition

$$\begin{aligned}
&U_{n,h}(\boldsymbol{\beta}) \\
&= \frac{1}{nh^2}\sum_{i=1}^n (-y_i) H''\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}}{h}\right)\boldsymbol{z}_i\boldsymbol{z}_i^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^*) - \frac{1}{nh}\sum_{i=1}^n (-y_i) H'\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}}{h}\right)\boldsymbol{z}_i \\
&=: \frac{1}{n}\sum_{i=1}^n U_{h,i}(\boldsymbol{\beta}),
\end{aligned}$$

where

(45)
$$U_{h,i}(\boldsymbol{\beta}) := \frac{1}{h^2}(-y_i) H''\left(\frac{x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}}{h}\right) \boldsymbol{z}_i \boldsymbol{z}_i^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^*) - \frac{1}{h}(-y_i) H'\left(\frac{x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}}{h}\right) \boldsymbol{z}_i.$$

For any $\boldsymbol{v}$ in $\mathbb{S}^{p-1}$, we have the following decomposition:

(46)
$$\begin{aligned}
&\boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta}) \\
&= (1-\mathbb{E})\left[\boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta}) - \boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta}^*)\right] + (1-\mathbb{E})\left[\boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta}^*)\right] + \mathbb{E}\left[\boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta})\right] \\
&= \frac{1}{n}\sum_{i=1}^n \phi^U_{i,\boldsymbol{v}}(\boldsymbol{\beta}) + (1-\mathbb{E})\left[\boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta}^*)\right] + \mathbb{E}\left[\boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta})\right]
\end{aligned}$$

where $\phi^U_{i,\boldsymbol{v}}(\boldsymbol{\beta}) := (1-\mathbb{E})\left[\boldsymbol{v}^\top U_{h,i}(\boldsymbol{\beta}) - \boldsymbol{v}^\top U_{h,i}(\boldsymbol{\beta}^*)\right]$. We will separately bound the three terms in (46). through the following three steps.

Step 1

We will show that, for some sufficiently large constant $\gamma > 0$, there exists $C_\phi > 0$ such that

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \le \delta_{m,0}} \sup_{j_v \in [5^p]} \left|\frac{1}{n}\sum_{i=1}^n \phi^U_{i,\boldsymbol{v}_{j_v}}(\boldsymbol{\beta})\right| \le C_\phi \delta_{m,0}\sqrt{\frac{p\log n}{nh^3}}, \tag{47}$$

with probability at least $1 - n^{-\gamma/2} - 2(5n^{-\gamma})^p$.

For any positive γ and each $j \in \{1,\dots,p\}$, divide the interval $[\beta_j^* - \delta_{m,0}, \beta_j^* + \delta_{m,0}]$ into n^γ small intervals, each with length $\frac{2\delta_{m,0}}{n^\gamma}$. This division creates n^γ intervals on each dimension, and the direct product of those intervals divides the hypercube $\{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_\infty \le \delta_{m,0}\}$ into $n^{\gamma p}$ small hypercubes. By arbitrarily picking a point on each small hypercube, we can find $\{\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_{n^{\gamma p}}\} \subset \mathbb{R}^p$, such that for all $\boldsymbol{\beta}$ in the ball $\{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \le \delta_{m,0}\}$ (which is a subset of $\{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_\infty \le \delta_{m,0}\}$), there exists $j_\beta \in [n^{\gamma p}]$ such that $\left\|\boldsymbol{\beta} - \boldsymbol{\beta}_{j_\beta}\right\|_\infty \le \frac{2\delta_{m,0}}{n^\gamma}$.

Assumption 1 ensures that $H''(x)$, $H'(x)$ are both bounded and Lipschitz continuous, and thus we have, for any $\boldsymbol{v}$ such that $\|\boldsymbol{v}\|_2 = 1$,

$$\begin{aligned}
&\left|\boldsymbol{v}^\top U_{h,i}(\boldsymbol{\beta}) - \boldsymbol{v}^\top U_{h,i}(\boldsymbol{\beta}_{j_\beta})\right| \\
&\le \frac{\left|\boldsymbol{v}^\top \boldsymbol{z}_i\right|}{h^2}\left|\boldsymbol{z}_i^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^*)\left[H''\left(\frac{x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}}{h}\right) - H''\left(\frac{x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}_{j_\beta}}{h}\right)\right] + \boldsymbol{z}_i^\top (\boldsymbol{\beta} - \boldsymbol{\beta}_{j_\beta}) H''\left(\frac{x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}_{j_\beta}}{h}\right)\right| \\
&\quad + \frac{\left|\boldsymbol{v}^\top \boldsymbol{z}_i\right|}{h}\left|H'\left(\frac{x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}_{j_\beta}}{h}\right) - H'\left(\frac{x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}}{h}\right)\right| \\
&\le \frac{2C_H p^{1/2}\delta_{m,0}^2 \|\boldsymbol{z}_i\|_2^3}{n^\gamma h^3} + \frac{4C_H p^{1/2}\delta_{m,0}\|\boldsymbol{z}_i\|_2^2}{n^\gamma h^2},
\end{aligned}$$

where C_H is a constant that is larger than the upper bounds and Lipschitz constants of $H''(x)$ and $H'(x)$. Therefore,

$$\begin{aligned}
&\sup_{j_v \in [5^p]} \sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \le \delta_{m,0}} \inf_{j_\beta \in [n^{\gamma p}]} \left|\frac{1}{n}\sum_{i=1}^n \phi^U_{i,\boldsymbol{v}_{j_v}}(\boldsymbol{\beta}) - \frac{1}{n}\sum_{i=1}^n \phi^U_{i,\boldsymbol{v}_{j_v}}(\boldsymbol{\beta}_{j_\beta})\right| \\
&\le \left|(1-\mathbb{E})\frac{1}{n}\sum_{i=1}^n \left[\boldsymbol{v}_{j_v}^\top U_{h,i}(\boldsymbol{\beta}) - \boldsymbol{v}_{j_v}^\top U_{h,i}(\boldsymbol{\beta}_{j_\beta})\right]\right|
\end{aligned}$$

$$\leq 8C_H \left(\frac{p^{1/2}\delta_{m,0}^2 \sum_{i=1}^n \|\boldsymbol{z}_i\|_2^3}{n^{\gamma+1}h^3} + \frac{p^{1/2}\delta_{m,0} \sum_{i=1}^n \|\boldsymbol{z}_i\|_2^2}{n^{\gamma+1}h^2} \right).$$

The assumption $\sup_{\|\boldsymbol{v}\|_2\leq 1} \mathbb{E}\exp(\eta(\boldsymbol{v}^\top \boldsymbol{z}_i)^2) < \infty$ implies that $\sup_{i,j} \mathbb{E}|z_{i,j}|^3 < \infty$. By Markov's inequality, it holds that $\frac{1}{n}\sum_{i=1}^n \|\boldsymbol{z}_i\|_3^3 = \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^p |z_{i,j}|^3 \leq 2n^{\gamma/2} p \sup_{i,j} \mathbb{E}|z_{i,j}|^3$, with probability at least $1-\frac{1}{2}n^{-\gamma/2}$. Since $\|\boldsymbol{z}_i\|_2^3 \leq n^{1/2}\|\boldsymbol{z}_i\|_3^3$, we obtain that $\frac{p^{1/2}\delta_{m,0}^2 \sum_{i=1}^n \|\boldsymbol{z}_i\|_2^3}{n^{\gamma+1}h^3} \leq 2\sup_{i,j}\mathbb{E}|z_{i,j}|^3 \frac{p^{3/2}\delta_{m,0}^2}{n^{(\gamma-1)/2}h^3}$. Similarly, $\frac{p^{1/2}\delta_{m,0} \sum_{i=1}^n \|\boldsymbol{z}_i\|_2^2}{n^{\gamma+1}h^2} \leq 2\sup_{i,j}\mathbb{E}|z_{i,j}|^2 \frac{p^{3/2}\delta_{m,0}}{n^{\gamma/2}h^2}$, with probability at least $1-\frac{1}{2}n^{-\gamma/2}$. Using the assumption $p\log n/(nh^3) = o(1)$, we can choose γ to be sufficiently large such that

$$\frac{p^{3/2}\delta_{m,0}^2}{n^{(\gamma-1)/2}h^3} + \frac{p^{3/2}\delta_{m,0}}{n^{\gamma/2}h^2} = o\left(\delta_{m,0}\sqrt{\frac{p\log n}{nh^3}}\right),$$

and then we have, with probability at least $1-n^{-\gamma/2}$,

$$\begin{aligned} &\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\leq\delta_{m,0}} \sup_{j_v\in[5^p]} \left| \frac{1}{n}\sum_{i=1}^n \phi_{i,\boldsymbol{v}_{j_v}}^U(\boldsymbol{\beta}) \right| - \sup_{j_\beta\in[n^{\gamma p}]} \sup_{j_v\in[5^p]} \left| \frac{1}{n}\sum_{i=1}^n \phi_{i,\boldsymbol{v}_{j_v}}^U(\boldsymbol{\beta}_{j_\beta}) \right| \\ &= o\left(\delta_{m,0}\sqrt{\frac{p\log n}{nh^3}}\right). \end{aligned} \tag{48}$$

Now let $\boldsymbol{v}$ be a fixed vector in $\mathbb{S}^{p-1}$ and $\boldsymbol{\beta}$ be a fixed vector in $\{\boldsymbol{\beta} : \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \leq \delta_{m,0}\}$. Recall that $\zeta = X + \boldsymbol{Z}^\top\boldsymbol{\beta}^*$ and $\rho(\cdot\,|\,\boldsymbol{Z})$ denotes the density of ζ given $\boldsymbol{Z}$. Let $\mathbb{E}_{\cdot|\boldsymbol{Z}}$ denote the expectation conditional on $\boldsymbol{Z}$. By Assumption 2, the density function $\rho(\cdot\,|\,\boldsymbol{Z})$ is bounded, and hence,

$$\begin{aligned} &\mathbb{E}_{\cdot|\boldsymbol{Z}}\left[H''\left(\frac{X+\boldsymbol{Z}^\top\boldsymbol{\beta}}{h}\right)\right]^2 = \mathbb{E}_{\cdot|\boldsymbol{Z}}\left[H''\left(\frac{\boldsymbol{Z}^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^*)+\zeta}{h}\right)\right]^2 \\ &= h\int_{-1}^1 H''(\xi)^2 \rho\left(\xi h - \boldsymbol{Z}^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\,|\,\boldsymbol{Z}\right)\mathrm{d}\xi = O(h), \end{aligned} \tag{49}$$

where we also use the facts that $h = o(1)$ and $H''(x)$ is bounded. Similarly, for some $\breve{\boldsymbol{\beta}}$ between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$,

$$\begin{aligned} \mathbb{E}_{\cdot|\boldsymbol{Z}}\left[H'\left(\frac{X+\boldsymbol{Z}^\top\boldsymbol{\beta}}{h}\right) - H'\left(\frac{X+\boldsymbol{Z}^\top\boldsymbol{\beta}^*}{h}\right)\right]^2 &= \mathbb{E}_{\cdot|\boldsymbol{Z}}\left[\frac{\boldsymbol{Z}^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^*)}{h}H''\left(\frac{X+\boldsymbol{Z}^\top\breve{\boldsymbol{\beta}}}{h}\right)\right]^2 \\ &= O\left(\frac{\left|\boldsymbol{Z}^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\right|^2}{h}\right). \end{aligned} \tag{50}$$

Let

$$\begin{aligned} &\widetilde{\phi}_{i,\boldsymbol{v}}^U(\boldsymbol{\beta}) := \frac{h^2}{\delta_{m,0}}\phi_{i,\boldsymbol{v}}^U(\boldsymbol{\beta}) \\ &= (1-\mathbb{E})(-y_i)(\boldsymbol{v}^\top\boldsymbol{z}_i)\left\{\frac{\boldsymbol{z}_i^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^*)}{\delta_{m,0}}H''\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}}{h}\right) - \frac{h}{\delta_{m,0}}\left[H'\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}}{h}\right) - H'\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}^*}{h}\right)\right]\right\}, \end{aligned}$$

Then, using the inequalities $e^x \le 1 + x + x^2 e^{\max(x,0)}$, $1 + x \le e^x$, and the fact that $\mathbb{E}\left[\widetilde{\phi}^U_{i,\boldsymbol{v}}(\boldsymbol{\beta})\right] = 0$, for any $b, t > 0$, we have

$$
\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^n \widetilde{\phi}^U_{i,\boldsymbol{v}}(\boldsymbol{\beta}) > b\right) \\
&= \mathbb{P}\left[\exp\left(t\sum_{i=1}^n \widetilde{\phi}^U_{i,\boldsymbol{v}}(\boldsymbol{\beta})\right) > e^{tb}\right] \\
&\le e^{-tb}\prod_{i=1}^n \mathbb{E}\left[\exp\left(t\widetilde{\phi}^U_{i,\boldsymbol{v}}(\boldsymbol{\beta})\right)\right] \qquad (51)\\
&\le e^{-tb}\prod_{i=1}^n \mathbb{E}\left[1 + \left(t\widetilde{\phi}^U_{i,\boldsymbol{v}}(\boldsymbol{\beta})\right)^2 \exp\left(t\left|\widetilde{\phi}^U_{i,\boldsymbol{v}}(\boldsymbol{\beta})\right|\right)\right] \\
&\le \exp\left(-tb + t^2\sum_{i=1}^n \mathbb{E}\left[\left(\widetilde{\phi}^U_{i,\boldsymbol{v}}(\boldsymbol{\beta})\right)^2 \exp\left(t\left|\widetilde{\phi}^U_{i,\boldsymbol{v}}(\boldsymbol{\beta})\right|\right)\right]\right).
\end{aligned}
$$

We note that the technique used in (51) is similar to Lemma 1 in Cai and Liu (2011). By (49) and (50),

$$
\begin{aligned}
&\mathbb{E}\left[\left(\widetilde{\phi}^U_{i,\boldsymbol{v}}(\boldsymbol{\beta})\right)^2 \exp\left(t\left|\widetilde{\phi}^U_{i,\boldsymbol{v}}(\boldsymbol{\beta})\right|\right)\right] \\
&\lesssim h\mathbb{E}\left[(\boldsymbol{v}^\top \boldsymbol{z}_i)^2(\boldsymbol{u}^\top \boldsymbol{z}_i)^2 \exp\left(C'_H t\left|\boldsymbol{v}^\top \boldsymbol{z}_i\right|\left|\boldsymbol{u}^\top \boldsymbol{z}_i\right|\right)\right],
\end{aligned}
$$

where $\boldsymbol{u} := (\boldsymbol{\beta} - \boldsymbol{\beta}^*)/\left\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\right\|_2$ is a unit vector, and C'_H is a constant depending on $H(\cdot)$. By the assumption that $\sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E}\left[\exp(\eta(\boldsymbol{v}^\top \boldsymbol{z}_i)^2)\right] < +\infty$ and Cauchy-Schwartz inequality, we obtain that

$$
\mathbb{E}\left[\left(\widetilde{\phi}^U_{i,\boldsymbol{v}}(\boldsymbol{\beta})\right)^2 \exp\left(t\left|\widetilde{\phi}^U_{i,\boldsymbol{v}}(\boldsymbol{\beta})\right|\right)\right] \le C_1^2 h,
$$

for some absolute constant C_1 and sufficiently small t. In particular, let $t = \sqrt{\frac{\gamma_1 p \log n}{4C_1^2 nh}}$ and $b = C_1\sqrt{nh\gamma_1 p \log n}$, where γ_1 is an arbitrary positive constant. By (51),

$$
\mathbb{P}\left(\sum_{i=1}^n \widetilde{\phi}^U_{i,\boldsymbol{v}}(\boldsymbol{\beta}) > C_1\sqrt{nh\gamma_1 p \log n}\right) \le \exp\left[-\frac{1}{4}\gamma_1 p \log n\right] = n^{-\gamma_1 p/4},
$$

which implies that

$$
\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \phi^U_{i,\boldsymbol{v}}(\boldsymbol{\beta})\right| > C_1\sqrt{\frac{\delta_{m,0}^2 \gamma_1 p \log n}{nh^3}}\right) \le 2n^{-\gamma_1 p/4}.
$$

Let $\gamma_1 = 8\gamma$ and $C_\phi = C_1\sqrt{\gamma_1}$. The above inequality is true for any $\boldsymbol{\beta} \in \mathbb{R}^p$ that satisfies $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \le \delta_{m,0}$ and any $\boldsymbol{v} \in \mathbb{S}^{p-1}$. In particular, for any $j_\beta \in [n^{\gamma p}]$ and $j_v \in [5^p]$, with probability at least $1 - 2n^{-2\gamma p}$,

$$
\left|\frac{1}{n}\sum_{i=1}^n \phi^U_{i,\boldsymbol{v}_{j_v}}\left(\boldsymbol{\beta}_{j_\beta}\right)\right| \le C_\phi \delta_{m,0}\sqrt{\frac{p\log n}{nh^3}},
$$

which implies that with probability at least $1-2(5n^{-\gamma})^p$,

$$\sup_{j_\beta\in[n^{\gamma p}]}\sup_{j_v\in[5^p]}\left|\frac{1}{n}\sum_{i=1}^n\phi^U_{i,\boldsymbol{v}_{j_v}}(\boldsymbol{\beta}_{j_\beta})\right|\le C_\phi\delta_{m,0}\sqrt{\frac{p\log n}{nh^3}}.$$

Combining with (48), we obtain (47).

Step 2 We will show that, for any constant $\gamma_2>0$, there exists a constant $C^*>0$, such that

$$\sup_{j_v\in[5^p]}\left|(1-\mathbb{E})\boldsymbol{v}_{j_v}^\top U_{n,h}(\boldsymbol{\beta}^*)\right|\le C^*\sqrt{\frac{p}{nh}}, \tag{52}$$

with probability at least $1-2(5e^{-\gamma_2/4})^p$.

Let $\boldsymbol{v}$ be any fixed vector in $\mathbb{S}^{p-1}$. Note that

$$(1-\mathbb{E})\boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta}^*)=\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\left[\boldsymbol{v}^\top U_{h,i}(\boldsymbol{\beta}^*)\right]=\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\frac{(\boldsymbol{v}^\top\boldsymbol{z}_i)y_i}{h}H'\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}^*}{h}\right),$$

and $(1-\mathbb{E})\left[\boldsymbol{v}^\top U_{h,i}(\boldsymbol{\beta}^*)\right]$ are i.i.d. among different i. By Assumption 2, the density function of $\zeta=X+\boldsymbol{Z}^\top\boldsymbol{\beta}^*$ satisfies that $\rho^{(1)}(\cdot\,|\,\boldsymbol{Z})$ is bounded uniformly for all $\boldsymbol{Z}$, which implies that $\rho(t\,|\,\boldsymbol{Z})=\rho(0\,|\,\boldsymbol{Z})+O(t)$. The constant in $O(t)$ is the same for all t. Therefore,

$$\begin{aligned}
&\mathbb{E}_{\cdot\,|\,\boldsymbol{Z}}\left[\boldsymbol{v}^\top U_{h,i}(\boldsymbol{\beta}^*)\right]^2\\
=&\frac{(\boldsymbol{v}^\top\boldsymbol{Z})^2}{h^2}\mathbb{E}_{\cdot\,|\,\boldsymbol{Z}}\left[H'\left(\frac{X+\boldsymbol{Z}^\top\boldsymbol{\beta}^*}{h}\right)\right]^2\\
=&\int_{-1}^1\frac{(\boldsymbol{v}^\top\boldsymbol{Z})^2}{h}\left[H'(\xi)\right]^2\rho(\xi h\,|\,\boldsymbol{Z})\,\mathrm{d}\xi\\
=&\frac{(\boldsymbol{v}^\top\boldsymbol{Z})^2}{h}\int_{-1}^1\left[H'(\xi)\right]^2\rho(0\,|\,\boldsymbol{Z})\,\mathrm{d}\xi+O\left((\boldsymbol{v}^\top\boldsymbol{Z})^2\right)\\
=&O\left(\frac{(\boldsymbol{v}^\top\boldsymbol{Z})^2}{h}\right).
\end{aligned} \tag{53}$$

Let $\widetilde{\phi}^{U,*}_{i,\boldsymbol{v}}=h(1-\mathbb{E})\left[\boldsymbol{v}^\top U_{h,i}(\boldsymbol{\beta}^*)\right]$. Analogous to the proof in **Step 1** for the bound of $\widetilde{\phi}^U_{i,\boldsymbol{v}}(\boldsymbol{\beta})$, we have

$$\mathbb{P}\left(\sum_{i=1}^n\widetilde{\phi}^{U,*}_{i,\boldsymbol{v}}>b\right)\le\exp\left(-tb+nt^2C_2^2h\right),$$

for some absolute constant C_2 and small enough t. Letting $b=C_2\sqrt{\gamma_2nph}$ and $t=\sqrt{\frac{\gamma_2p}{4C_2^2nh}}$ leads to

$$\mathbb{P}\left(\sum_{i=1}^n\widetilde{\phi}^{U,*}_{i,\boldsymbol{v}}>C_2\sqrt{\gamma_2nhp}\right)\le\exp\left[-\frac{1}{4}\gamma_2p\right],$$

which leads to

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^n(1-\mathbb{E})\left[\boldsymbol{v}^\top U_{h,i}(\boldsymbol{\beta}^*)\right]\right|>C_2\sqrt{\frac{\gamma_2p}{nh}}\right)\le 2e^{-\gamma_2p/4}.$$

The above inequality is true for all $\boldsymbol{v}$ in $\mathbb{S}^{p-1}$. Therefore, with probability at least $1-2(5e^{-\gamma_2/4})^p$,

$$\sup_{j_v\in[5^p]}\left|(1-\mathbb{E})\boldsymbol{v}_{j_v}^\top U_{n,h}\left(\boldsymbol{\beta}^*\right)\right|\leq C^*\sqrt{\frac{p}{nh}},$$

where $C^*=C_2\sqrt{\gamma_2}$.

Step 3

We will show that,

$$\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\leq\delta_{m,0}}\mathbb{E}\left[\boldsymbol{v}^\top U_{n,h}\left(\boldsymbol{\beta}\right)\right]\leq C_E(\delta_{m,0}^2+h^\alpha), \tag{54}$$

where C_E is an absolute constant.

By Assumptions 2 and 3, for almost every $\boldsymbol{Z}$,

$$\rho\left(t\,|\,\boldsymbol{Z}\right)=\sum_{k=0}^{\alpha-1}\frac{1}{k!}\rho^{(k)}\left(0\,|\,\boldsymbol{Z}\right)t^k+\frac{1}{\alpha!}\rho^{(\alpha)}\left(t'\,|\,\boldsymbol{Z}\right)t^\alpha,$$

$$F\left(-t\,|\,\boldsymbol{Z}\right)=\frac{1}{2}+\sum_{k=1}^{\alpha}\frac{1}{k!}F^{(k)}\left(0\,|\,\boldsymbol{Z}\right)(-t)^k+\frac{1}{(\alpha+1)!}F^{(\alpha+1)}\left(t''\,|\,\boldsymbol{Z}\right)(-t)^{\alpha+1},$$

where t',t'' are between 0 and t. Therefore,

$$\left(2F\left(-t\,|\,\boldsymbol{Z}\right)-1\right)\rho\left(t\,|\,\boldsymbol{Z}\right)=\sum_{k=1}^{2\alpha+1}M_k\left(\boldsymbol{Z}\right)t^k, \tag{55}$$

where $M_k\left(\boldsymbol{Z}\right)$'s are constants depending on ρ, F, $\boldsymbol{Z}$, and t. Since $\rho^{(k)}\left(\cdot\,|\,\boldsymbol{Z}\right)$ and $F^{(k)}\left(\cdot\,|\,\boldsymbol{Z}\right)$ are bounded for all k and almost all $\boldsymbol{Z}$, we know there exists a constant M such that $\sup_{k,\boldsymbol{Z},t}\left|M_k\left(\boldsymbol{Z}\right)\right|\leq M$ for all t. (In particular, we can obtain that

$$M_1(\boldsymbol{Z})=2F^{(1)}(0\,|\,\boldsymbol{Z})\rho(0\,|\,\boldsymbol{Z}),\quad M_\alpha\left(\boldsymbol{Z}\right)=\sum_{k=1}^{\alpha}\frac{2(-1)^{-k}}{(\alpha-k)!k!}F^{(k)}\left(0\,|\,\boldsymbol{Z}\right)\rho^{(\alpha-k)}\left(0\,|\,\boldsymbol{Z}\right),$$

which will be used in the proof of the following theorems.)

By Assumption 1, when $x>1$ or $x<-1$, $H'\left(x\right)=H''\left(x\right)=0$. The kernel $H'\left(x\right)=\int_{-1}^{x}H''\left(t\right)\mathrm{d}t$ is bounded, satisfying $\int_{-1}^{1}H'\left(x\right)\mathrm{d}x=1$ and $\int_{-1}^{1}x^kH'\left(x\right)\mathrm{d}x=0$ for any $1\leq k\leq\alpha-1$. Using integration by parts, we have $\int_{-1}^{1}xH''\left(x\right)\mathrm{d}x=-1$ and $\int_{-1}^{1}x^kH''\left(x\right)\mathrm{d}x=0$ for $k=0$ and $2\leq k\leq\alpha$.

Now we are ready to compute the expectation of $\boldsymbol{v}^\top U_{n,h}\left(\boldsymbol{\beta}\right)$ for any $\boldsymbol{v}\in\mathbb{S}^{p-1}$ and $\boldsymbol{\beta}\in\{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\leq\delta_{m,0}\}$. Since $\mathbb{E}\left[\boldsymbol{v}^\top U_{n,h}\left(\boldsymbol{\beta}\right)\right]=\mathbb{E}\left[\boldsymbol{v}^\top U_{h,i}\left(\boldsymbol{\beta}\right)\right]$ for all i, we omit i in the following computation of expectation. Let $\mathbb{E}_{\cdot|\boldsymbol{Z}}$ denote the expectation conditional on $\boldsymbol{Z}$.

Define $\Delta(\boldsymbol{\beta}) := \boldsymbol{\beta} - \boldsymbol{\beta}^*$ and recall that $\zeta = X + \boldsymbol{Z}^\top \boldsymbol{\beta}^*$.

(56)

$$
\begin{aligned}
&\mathbb{E}_{\cdot|\boldsymbol{Z}}\left[\boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta})\right] \\
=&\boldsymbol{Z}^\top \boldsymbol{v} \cdot \mathbb{E}_{\cdot|\boldsymbol{Z}}\left[\frac{\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})}{h^2}\left[2\mathbb{I}\left(X + \boldsymbol{Z}^\top\boldsymbol{\beta}^* + \epsilon < 0\right) - 1\right] H''\left(\frac{X + \boldsymbol{Z}^\top\boldsymbol{\beta}}{h}\right)\right. \\
&\qquad \left. - \frac{1}{h}\left[2\mathbb{I}\left(X + \boldsymbol{Z}^\top\boldsymbol{\beta}^* + \epsilon < 0\right) - 1\right] H'\left(\frac{X + \boldsymbol{Z}^\top\boldsymbol{\beta}}{h}\right)\right] \\
=&\boldsymbol{Z}^\top \boldsymbol{v} \cdot \mathbb{E}_{\cdot|\boldsymbol{Z}}\left\{\left[2\mathbb{I}(\zeta + \epsilon < 0) - 1\right]\left[\frac{\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})}{h^2} H''\left(\frac{\zeta + \boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})}{h}\right) - \frac{1}{h} H'\left(\frac{\zeta + \boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})}{h}\right)\right]\right\} \\
=&\left(\boldsymbol{Z}^\top \boldsymbol{v}\right)\int_{-1}^{1}\left[2F\left(\boldsymbol{Z}^\top\Delta(\boldsymbol{\beta}) - \xi h \,|\, \boldsymbol{Z}\right) - 1\right]\rho\left(\xi h - \boldsymbol{Z}^\top\Delta(\boldsymbol{\beta}) \,|\, \boldsymbol{Z}\right) \\
&\quad \cdot\left(\frac{\boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})}{h} H''(\xi) - H'(\xi)\right) \mathrm{d}\xi \qquad \left(\text{by changing variable } \xi = \frac{\zeta + \boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})}{h}\right) \\
=&\left(\boldsymbol{Z}^\top \boldsymbol{v}\right)\sum_{k=1}^{2\alpha+1} M_k(\boldsymbol{Z})\int_{-1}^{1}\left(\xi h - \boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})\right)^k\left(\frac{\boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})}{h} H''(\xi) - H'(\xi)\right)\mathrm{d}\xi.
\end{aligned}
$$

When $1 \le k \le \alpha - 1$,

(57)

$$
\begin{aligned}
&\int_{-1}^{1}\left(\xi h - \boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})\right)^k\left(\frac{\boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})}{h} H''(\xi) - H'(\xi)\right)\mathrm{d}\xi \\
=&\sum_{k'=0}^{k}\binom{k}{k'} h^{k'}\left(-\boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})\right)^{k-k'}\left[\left(\boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})/h\right)\int_{-1}^{1}\xi^{k'} H''(\xi)\,\mathrm{d}\xi - \int_{-1}^{1}\xi^{k'} H'(\xi)\,\mathrm{d}\xi\right] \\
=&(k-1)\left(-\boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})\right)^k,
\end{aligned}
$$

When $k = \alpha$,

$$
\begin{aligned}
&\int_{-1}^{1}\left(\xi h - \boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})\right)^\alpha\left(\frac{\boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})}{h} H''(\xi) - H'(\xi)\right)\mathrm{d}\xi \\
=&\sum_{k'=0}^{\alpha}\binom{\alpha}{k'} h^{k'}\left(-\boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})\right)^{\alpha-k'}\left[\left(\boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})/h\right)\int_{-1}^{1}\xi^{k'} H''(\xi)\,\mathrm{d}\xi - \int_{-1}^{1}\xi^{k'} H'(\xi)\,\mathrm{d}\xi\right] \\
=&(\alpha-1)\left(-\boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})\right)^\alpha - \pi_U h^\alpha,
\end{aligned}
$$

where $\pi_U = \int_{-1}^{1}\xi^\alpha H'(\xi)\,\mathrm{d}\xi$ is defined in Assumption 1.

When $\alpha + 1 \le k \le 2\alpha + 1$, using the fact that H', H'' are both bounded and the inequality $a^{k'}b^{k-k'} \le (a+b)^k \le 2^{k-1}(a^k + b^k)$, we have

$$
\begin{aligned}
&\int_{-1}^{1}\left(\xi h - \boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})\right)^k\left(\frac{\boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})}{h} H''(\xi) - H'(\xi)\right)\mathrm{d}\xi \\
=&\sum_{k'=0}^{k}\binom{k}{k'} h^{k'}\left(-\boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})\right)^{k-k'}\left[\left(\boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})/h\right)\int_{-1}^{1}\xi^{k'} H''(\xi)\,\mathrm{d}\xi - \int_{-1}^{1}\xi^{k'} H'(\xi)\,\mathrm{d}\xi\right]
\end{aligned}
$$

$$
\begin{aligned}
=&(k-1)\left(-\boldsymbol{Z}^{\top}\Delta\left(\boldsymbol{\beta}\right)\right)^{k}-\sum_{k'=\alpha+1}^{k}\binom{k}{k'}h^{k'-1}\left(-\boldsymbol{Z}^{\top}\Delta\left(\boldsymbol{\beta}\right)\right)^{k-k'+1}\left[\int_{-1}^{1}\xi^{k'}H''\left(\xi\right)\mathrm{d}\xi\right]\\
&-\sum_{k'=\alpha}^{k}\binom{k}{k'}h^{k'}\left(-\boldsymbol{Z}^{\top}\Delta\left(\boldsymbol{\beta}\right)\right)^{k-k'}\left[\int_{-1}^{1}\xi^{k'}H'\left(\xi\right)\mathrm{d}\xi\right]\\
=&O\left[h^{k}+(\boldsymbol{Z}^{\top}\Delta(\boldsymbol{\beta}))^{k}\right].
\end{aligned}
$$

Note that $\left|\boldsymbol{Z}^{\top}\Delta(\boldsymbol{\beta})\right|\leq\delta_{m,0}\left|\boldsymbol{Z}^{\top}\boldsymbol{u}\right|$, where $\boldsymbol{u}=\Delta(\boldsymbol{\beta})/\left\|\Delta(\boldsymbol{\beta})\right\|_{2}$. By Assumption 5, it holds that $\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\mathbb{E}(\boldsymbol{Z}^{\top}\boldsymbol{v})^{2k}\leq\infty$ for all positive integer k. Adding that $M_{k}(\boldsymbol{Z})$'s are uniformly bounded , we finally obtain that

$$
\left|\mathbb{E}\left[\boldsymbol{v}^{\top}U_{n,h}\left(\boldsymbol{\beta}\right)\right]\right|\leq C_{E}\left(\delta_{m,0}^{2}+h^{\alpha}\right), \tag{58}
$$

where C_E is a constant not depending on $\boldsymbol{v}$ and $\boldsymbol{\beta}$. This leads to (54).

Combining the three steps with (44) completes the proof of (40).

Proof of (41)

The proof of (41) is similar to that of (40). We use the same $\{\boldsymbol{v}_{j_v}\}$ and $\{\boldsymbol{\beta}_{j_\beta}\}$ as in the proof of (40). By the proof of Lemma 3 in Cai et al. (2010),

$$
\left\|A\right\|_{2}\leq4\sup_{j_{v}\in[5^{p}]}\left|\boldsymbol{v}_{j_{v}}^{\top}A\boldsymbol{v}_{j_{v}}\right|,
$$

for any symmetric $A\in\mathbb{R}^{p\times p}$. Therefore, it suffices to bound $\sup_{j_v\in[5^p]}\left|\boldsymbol{v}_{j_v}^{\top}\left[V_{n,h}\left(\boldsymbol{\beta}^{(0)}\right)-V\right]\boldsymbol{v}_{j_v}\right|$. By the choice of $\left\{\boldsymbol{\beta}_{j_\beta}\right\}$, for all $\boldsymbol{\beta}$ in the ball $\{\boldsymbol{\beta}:\left\|\boldsymbol{\beta}-\boldsymbol{\beta}^{*}\right\|_{2}\leq\delta_{m,0}\}$, there exists $j_\beta\in[n^{\gamma p}]$ such that $\left\|\boldsymbol{\beta}-\boldsymbol{\beta}_{j_\beta}\right\|_{\infty}\leq\frac{2\delta_{m,0}}{n^{\gamma}}$. Recall

$$
V_{n,h}\left(\boldsymbol{\beta}\right)=\frac{1}{nh^{2}}\sum_{i=1}^{n}\left(-y_{i}\right)H''\left(\frac{x_{i}+\boldsymbol{z}_{i}^{\top}\boldsymbol{\beta}}{h}\right)\boldsymbol{z}_{i}\boldsymbol{z}_{i}^{\top}=:\frac{1}{n}\sum_{i=1}^{n}V_{h,i}\left(\boldsymbol{\beta}\right),
$$

where

$$
V_{h,i}\left(\boldsymbol{\beta}\right):=\frac{1}{h^{2}}\left(-y_{i}\right)H''\left(\frac{x_{i}+\boldsymbol{z}_{i}^{\top}\boldsymbol{\beta}}{h}\right)\boldsymbol{z}_{i}\boldsymbol{z}_{i}^{\top}. \tag{59}
$$

By the Lipschitz property of $H''(x)$, we have

$$
\begin{aligned}
&\sup_{j_{v}\in[5^{p}]}\sup_{\boldsymbol{\beta}:\left\|\boldsymbol{\beta}-\boldsymbol{\beta}^{*}\right\|_{2}\leq\delta_{m,0}}\inf_{j_{\beta}\in[n^{\gamma p}]}\left|\boldsymbol{v}_{j_{v}}^{\top}V_{n,h}\left(\boldsymbol{\beta}\right)\boldsymbol{v}_{j_{v}}-\boldsymbol{v}_{j_{v}}^{\top}V_{n,h}\left(\boldsymbol{\beta}_{j_{\beta}}\right)\boldsymbol{v}_{j_{v}}\right|\\
\leq&\sup_{j_{v}\in[5^{p}]}\sup_{\boldsymbol{\beta}:\left\|\boldsymbol{\beta}-\boldsymbol{\beta}^{*}\right\|_{2}\leq\delta_{m,0}}\inf_{j_{\beta}\in[n^{\gamma p}]}\frac{1}{n}\sum_{i=1}^{n}\frac{\left(\boldsymbol{z}_{i}^{\top}\boldsymbol{v}_{j_{v}}\right)^{2}}{h^{3}}\left|\boldsymbol{z}_{i}^{\top}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_{j_{\beta}}\right)\right|\\
=&O\left(\sum_{i=1}^{n}\frac{\delta_{m,0}\left\|\boldsymbol{z}_{i}\right\|_{2}^{3}}{n^{\gamma+1}h^{3}}\right).
\end{aligned}
$$

In the **Step 1** of the proof of (40), we show that $\frac{p^{1/2}\delta_{m,0}\sum_{i=1}^{n}\left\|\boldsymbol{z}_{i}\right\|_{2}^{3}}{n^{\gamma+1}h^{3}}\leq\sup_{i,j}\mathbb{E}|z_{i,j}|^{3}\frac{p^{3/2}\delta_{m,0}}{n^{(\gamma-1)/2}h^{3}}$, with probability and least $1-n^{-\gamma/2}$. By taking $\gamma>0$ large enough such that

$$
\frac{p^{3/2}\delta_{m,0}}{n^{(\gamma-1)/2}h^{3}}=o\left(\sqrt{\frac{p\log n}{nh^{3}}}\right),
$$

we obtain

(60)
$$\sup_{j_v\in[5^p]}\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\le\delta_{m,0}}\left|\boldsymbol{v}_{j_v}^\top\left[V_{n,h}\left(\boldsymbol{\beta}\right)-V\right]\boldsymbol{v}_{j_v}\right|-\sup_{j_v\in[5^p]}\sup_{j_\beta\in[n^{\gamma p}]}\left|\boldsymbol{v}_{j_v}^\top\left[V_{n,h}\left(\boldsymbol{\beta}_{j_\beta}\right)-V\right]\boldsymbol{v}_{j_v}\right|$$
$$=o\left(\sqrt{\frac{p\log n}{nh^3}}\right).$$

Note that $(1-\mathbb{E})\left(\boldsymbol{v}^\top\left[V_{n,h}\left(\boldsymbol{\beta}\right)-V\right]\boldsymbol{v}\right)=\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\boldsymbol{v}^\top V_{h,i}(\boldsymbol{\beta})\boldsymbol{v}$ and let

$$\phi_{i,\boldsymbol{v}}^V(\boldsymbol{\beta}):=h^2(1-\mathbb{E})\boldsymbol{v}^\top V_{h,i}(\boldsymbol{\beta})\boldsymbol{v}=(1-\mathbb{E})(\boldsymbol{v}^\top\boldsymbol{z}_i)^2\left(-y_i\right)H''\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}}{h}\right).$$

By repeating the procedure in **Step 1** of the proof for (40), we obtain

$$\mathbb{P}\left(\phi_{i,\boldsymbol{v}}^V(\boldsymbol{\beta})>b\right)\le\exp\left(-tb+t^2C_3^2nh\right),$$

for some absolute constant C_3 and small enough $t>0$. Let $b=C_3\sqrt{\gamma_3nhp\log n}$ and $t=b/(2C_3^2nh)$, where γ_3 is an arbitrary positive constant to be determined, it holds that

$$\mathbb{P}\left(\sum_{i=1}^n\phi_{i,\boldsymbol{v}}^V(\boldsymbol{\beta})>C_3\sqrt{\gamma_3nhp\log n}\right)\le n^{-\gamma_3p/4},$$

and hence

$$\mathbb{P}\left(\left|(1-\mathbb{E})\left(\boldsymbol{v}^\top\left[V_{n,h}\left(\boldsymbol{\beta}\right)-V\right]\boldsymbol{v}\right)\right|>C_3\sqrt{\frac{\gamma_3p\log n}{nh^3}}\right)\le2n^{-\gamma_3p/4}.$$

The above inequality is true for all $\boldsymbol{v}_{j_v}$ and $\boldsymbol{\beta}_{j_\beta}$, and thus

$$\mathbb{P}\left(\sup_{j_v\in[5^p]}\sup_{j_\beta\in[n^{\gamma p}]}\left|(1-\mathbb{E})\left(\boldsymbol{v}_{j_v}^\top\left[V_{n,h}\left(\boldsymbol{\beta}_{j_\beta}\right)-V\right]\boldsymbol{v}_{j_v}\right)\right|>C_3\sqrt{\frac{\gamma_3p\log n}{nh^3}}\right)\le2(5n^{\gamma-\gamma_3/4})^p.$$

Letting $\gamma_3=8\gamma$ and combining the above inequality with (60) lead to that, for any sufficiently large $\gamma>0$, there exists a constant C, such that

$$\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\le\delta_{m,0}}\left|(1-\mathbb{E})\left(\boldsymbol{v}^\top\left[V_{n,h}\left(\boldsymbol{\beta}\right)-V\right]\boldsymbol{v}\right)\right|\le C\sqrt{\frac{p\log n}{nh^3}},\tag{61}$$

with probability at least $1-n^{-\gamma/2}-2(5n^{-\gamma})^p$.

In the following proof, we first consider any fixed $\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{v}\in\mathbb{R}^p$ that satisfy $\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\le\delta_{m,0}$ and $\|\boldsymbol{v}\|_2=1$, and then apply the result to the specific $\boldsymbol{v}_{j_v}$ and $\boldsymbol{\beta}_{j_\beta}$. The computation of $\mathbb{E}\left[\boldsymbol{v}^\top\left[V_{n,h}\left(\boldsymbol{\beta}\right)-V\right]\boldsymbol{v}\right]$ is similar to **Step 3** in the proof of (40), where we obtain that

$$\left(2F\left(-t\,|\,\boldsymbol{Z}\right)-1\right)\rho\left(t\,|\,\boldsymbol{Z}\right)=\sum_{k=1}^{2\alpha+1}M_k\left(\boldsymbol{Z}\right)t^k,$$

with uniformly bounded $M_k(\boldsymbol{Z})$ and $M_1(\boldsymbol{Z}) = 2F'(0 \mid \boldsymbol{Z})\rho(0 \mid \boldsymbol{Z})$. Recall that $\int_{-1}^{1} xH''(x)\,\mathrm{d}x = -1$ and $\int_{-1}^{1} x^k H''(x)\,\mathrm{d}x = 0$ for $k = 0$ and $2 \le k \le \alpha$.

(62)

$$
\begin{aligned}
&\mathbb{E}_{\cdot|\boldsymbol{Z}}\left(\boldsymbol{v}^\top V_{n,h}(\boldsymbol{\beta})\,\boldsymbol{v}\right)\\
=&\frac{\left(\boldsymbol{v}^\top \boldsymbol{z}\right)^2}{h^2}\mathbb{E}_{\cdot|\boldsymbol{Z}}\left[2\mathbb{I}\left(\zeta+\epsilon<0\right)-1\right]H''\left(\frac{\boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})+\zeta}{h}\right)\\
=&\frac{\left(\boldsymbol{v}^\top \boldsymbol{Z}\right)^2}{h}\int_{-1}^{1}\left(2F\left(\boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})-\xi h\mid\boldsymbol{Z}\right)-1\right)\rho\left(\xi h-\boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})\mid\boldsymbol{Z}\right)H''(\xi)\,\mathrm{d}\xi\\
=&-2\left(\boldsymbol{v}^\top\boldsymbol{Z}\right)^2F'(0\mid\boldsymbol{Z})\,\rho(0\mid\boldsymbol{Z})\int_{-1}^{1}\xi H''(\xi)\,\mathrm{d}\xi+\frac{\left(\boldsymbol{v}^\top\boldsymbol{Z}\right)^2}{h}\int_{-1}^{1}\sum_{k=2}^{2\alpha+1}M_k(\boldsymbol{Z})\left(\xi h-\boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})\right)^k H''(\xi)\,\mathrm{d}\xi\\
=&2\left(\boldsymbol{v}^\top\boldsymbol{Z}\right)^2F'(0\mid\boldsymbol{Z})\,\rho(0\mid\boldsymbol{Z})+\left(\boldsymbol{v}^\top\boldsymbol{Z}\right)^2O\left(h^\alpha+(\boldsymbol{Z}^\top\boldsymbol{u})\delta_{m,0}\right),
\end{aligned}
$$

where $\boldsymbol{u} = \Delta(\boldsymbol{\beta})/\left\|\Delta(\boldsymbol{\beta})\right\|_2$, and the last inequality follows from that, for $k \ge 2$,

$$
\begin{aligned}
&\frac{\left(\boldsymbol{v}^\top\boldsymbol{Z}\right)^2}{h}\int_{-1}^{1}\left(\xi h-\boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})\right)^k H''(\xi)\,\mathrm{d}\xi\\
=&\left(\boldsymbol{v}^\top\boldsymbol{Z}\right)^2\sum_{k'=0}^{k}\binom{k}{k'}h^{k'-1}\left(-\boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})\right)^{k-k'}\int_{-1}^{1}\xi^{k'}H''(\xi)\,\mathrm{d}\xi\\
=&\left(\boldsymbol{v}^\top\boldsymbol{Z}\right)^2\left[(-k)\left(-\boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})\right)^{k-1}+\sum_{k'=\alpha+1}^{k}\binom{k}{k'}h^{k'-1}\left(-\boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})\right)^{k-k'}\int_{-1}^{1}\xi^{k'}H''(\xi)\,\mathrm{d}\xi\right]\\
\lesssim&\left(\boldsymbol{v}^\top\boldsymbol{Z}\right)^2\left[h^\alpha+\left|\boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})\right|^{k-1}\right].
\end{aligned}
$$

Note that the constant hidden in $O(\cdot)$ does not depend on $\boldsymbol{v}$ or $\boldsymbol{\beta}$. By Assumption 5 and $V = 2\mathbb{E}\left[\rho(0\mid\boldsymbol{Z})F'(0\mid\boldsymbol{Z})\boldsymbol{Z}\boldsymbol{Z}^\top\right]$, we obtain

$$
\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\le\delta_{m,0}}\left|\mathbb{E}\left(\boldsymbol{v}^\top\left[V_{n,h}(\boldsymbol{\beta})-V\right]\boldsymbol{v}\right)\right| = O\left(\delta_{m,0}+h^\alpha\right), \tag{63}
$$

which completes the proof for (41) together with (61).

Finally, (37), (40), and (41) directly yield (42) (i.e., (11)), given the assumption that $\Lambda_{\min}(V) > c_1^{-1}$ for some $c_1 > 0$.

□

Proof of Theorem 3.3

PROOF. Without loss of generality, we assume $\lambda_h = 1$. Then we have

$$
h_t = \max\left\{(p/n)^{1/(2\alpha+1)}, (p/m)^{2^t/(3\alpha)}\right\} \ge (p/n)^{1/(2\alpha+1)},
$$

which implies the assumption $\frac{p\log n}{nh_t^3} = o(1)$ holds for any t, since $\frac{p\log n}{nh_t^3} \le (p/n)^{\frac{2\alpha-2}{2\alpha+1}}\log n = O\left(n^{-(1-c_2)\frac{2\alpha-2}{2\alpha+1}}\log n\right) \to 0$ for $p = O(m^{c_2}) = O(n^{c_2})$ and $\alpha \ge 2$. Moreover, for any t, $h_t^\alpha = \max\left\{(p/m)^{2^t/3}, (p/n)^{\alpha/(2\alpha+1)}\right\}$ and $\sqrt{p/nh_t} \le (p/n)^{\alpha/(2\alpha+1)}$.

We first show that, for any t,

$$\left\|\widehat{\boldsymbol{\beta}}^{(t)}-\boldsymbol{\beta}^*\right\|_2=O_{\mathbb{P}}\left((p/n)^{\alpha/(2\alpha+1)}\sqrt{\log n}+(p/m)^{2^t/3}\right). \tag{64}$$

Recall that by (11), if $\left\|\widehat{\boldsymbol{\beta}}^{(t-1)}-\boldsymbol{\beta}^*\right\|_2=O_{\mathbb{P}}\left(\delta_{m,t-1}\right)$ and $h_t=o(1)$, we have that

$$\left\|\widehat{\boldsymbol{\beta}}^{(t)}-\boldsymbol{\beta}^*\right\|_2=O_{\mathbb{P}}\left(\delta_{m,t-1}^2+h_t^\alpha+\sqrt{\frac{p}{nh_t}}+\delta_{m,t-1}\sqrt{\frac{p\log n}{nh_t^3}}\right), \tag{65}$$

When $t=1$, $\delta_{m,0}=(p/m)^{1/3}\le h_1=(p/m)^{\frac{2}{3\alpha}}$, which implies $\delta_{m,0}\sqrt{\frac{p\log n}{nh_1^3}}=O\left(\sqrt{\frac{p\log n}{nh_1}}\right)$. Then (64) holds since

$$\left\|\widehat{\boldsymbol{\beta}}^{(1)}-\boldsymbol{\beta}^*\right\|_2=O_{\mathbb{P}}\left(\sqrt{\frac{p\log n}{nh_1}}+\left(\frac{p}{m}\right)^{2/3}+h_1^\alpha\right)=O_{\mathbb{P}}\left(\left(\frac{p}{n}\right)^{\frac{\alpha}{2\alpha+1}}\sqrt{\log n}+\left(\frac{p}{m}\right)^{2/3}\right).$$

Assume (64) holds for $t-1$, i.e.,

$$\left\|\widehat{\boldsymbol{\beta}}^{(t-1)}-\boldsymbol{\beta}^*\right\|_2=O_{\mathbb{P}}\left((p/n)^{\alpha/(2\alpha+1)}\sqrt{\log n}+(p/m)^{2^{t-1}/3}\right).$$

Then $\delta_{m,t-1}=(p/n)^{\alpha/(2\alpha+1)}\sqrt{\log n}+(p/m)^{2^{t-1}/3}$. Since $(p/n)^{\alpha/(2\alpha+1)}\sqrt{\log n}\ll(p/n)^{\frac{1}{2\alpha+1}}\le h_t$ and $(p/m)^{2^{t-1}/3}\le(p/m)^{\frac{2^t}{3\alpha}}\le h_t$, it holds that $\delta_{m,t-1}=O\left(h_t\right)$. Then

$$\left\|\widehat{\boldsymbol{\beta}}^{(t)}-\boldsymbol{\beta}^*\right\|_2=O_{\mathbb{P}}\left(\sqrt{\frac{p\log n}{nh_t}}+\delta_{m,t-1}^2+h_t^\alpha\right)=O_{\mathbb{P}}\left(\left(\frac{p}{n}\right)^{\frac{\alpha}{2\alpha+1}}\sqrt{\log n}+(p/m)^{2^t/3}\right),$$

since $h_t^\alpha=\max\left\{(p/m)^{2^t/3},(p/n)^{\alpha/(2\alpha+1)}\right\}$ and $\sqrt{p/nh_t}\le(p/n)^{\alpha/(2\alpha+1)}$. Therefore, we have proved that (64) holds for all t by induction.

To see (12), note that plugging $\delta_{m,t-1}=(p/n)^{\alpha/(2\alpha+1)}\sqrt{\log n}+(p/m)^{2^{t-1}/3}$ into (65) yields that

$$\left\|\widehat{\boldsymbol{\beta}}^{(t)}-\boldsymbol{\beta}^*\right\|_2=O_{\mathbb{P}}\left(\left(\frac{p}{n}\right)^{\frac{2\alpha}{2\alpha+1}}\log n+\left(\frac{p}{m}\right)^{\frac{2^t}{3}}+h_t^\alpha+\sqrt{\frac{p}{nh_t}}+\left(\left(\frac{p}{n}\right)^{\frac{\alpha}{2\alpha+1}}\sqrt{\log n}+\left(\frac{p}{m}\right)^{\frac{2^{t-1}}{3}}\right)\sqrt{\frac{p\log n}{nh_t^3}}\right). \tag{66}$$

Since $\alpha>1$,

$$\left(\frac{p}{n}\right)^{\frac{\alpha}{2\alpha+1}}\sqrt{\log n}\cdot\sqrt{\frac{p\log n}{nh_t^3}}=\left(\frac{p}{n}\right)^{\frac{\alpha}{2\alpha+1}}(\log n)\cdot\sqrt{\frac{p}{nh_t^3}}\le\left(\frac{p}{n}\right)^{\frac{\alpha}{2\alpha+1}}(\log n)\cdot(p/n)^{\frac{\alpha-1}{2\alpha+1}}\lesssim\left(\frac{p}{n}\right)^{\frac{\alpha}{2\alpha+1}}.$$

Therefore, the rate in (66) is upper bounded by

$$O\left(\left(\frac{p}{n}\right)^{\frac{\alpha}{2\alpha+1}}+\left(\frac{p}{m}\right)^{\frac{2^t}{3}}+\left(\frac{p}{m}\right)^{\frac{2^{t-1}}{3}}\left(\frac{p}{n}\right)^{\frac{\alpha-1}{2\alpha+1}}\sqrt{\log n}\right),$$

which completes the proof. □

B.2. Theoretical Results for the Inference of `(mSMSE)`**.** To show the asymptotic normality, we first prove an important Lemma about $U_{n,h}(\boldsymbol{\beta})$ defined by (39).

LEMMA B.1. *Under Assumptions 1–5, if* $h=o(1)$, $\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\le\delta$, $\delta=o\left(\min\{h^{\alpha/2},h/\sqrt{p\log n}\}\right)$ *then*

$$\mathbb{E}\left[\boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta})\right]=\boldsymbol{v}^\top Uh^\alpha+o\left(h^\alpha\right), \tag{67}$$

and

$$(1-\mathbb{E})\sqrt{nh}\left[\boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta})\right]/\sqrt{\boldsymbol{v}^\top V_s \boldsymbol{v}} \xrightarrow{d} \mathcal{N}(0,1), \tag{68}$$

for any $\boldsymbol{v}\in\mathbb{R}^p\setminus\{\mathbf{0}\}$.

PROOF OF LEMMA B.1. Recall that

$$M_\alpha(\boldsymbol{Z}) = \sum_{k=1}^{\alpha} \frac{2(-1)^k}{(\alpha-k)!k!} F^{(k)}(0\,|\,\boldsymbol{Z})\,\rho^{(\alpha-k)}(0\,|\,\boldsymbol{Z}),$$

and

$$U := \pi_U \mathbb{E}\left(\sum_{k=1}^{\alpha} \frac{2(-1)^{k+1}}{k!\,(\alpha-k)!} F^{(k)}(0\,|\,\boldsymbol{Z})\,\rho^{(\alpha-k)}(0\,|\,\boldsymbol{Z})\,\boldsymbol{Z}\right),$$

where π_U is defined in Assumption 1. For any $\boldsymbol{v}\in\mathbb{S}^{p-1}$, the computation in (56) yields that

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta})\right] &= -\mathbb{E}\left[\left(\boldsymbol{Z}^\top\boldsymbol{v}\right)\pi_U M_\alpha(\boldsymbol{Z})\,h^\alpha\right] + O\left(\delta^2+h^{\alpha+1}\right)\\ &= \boldsymbol{v}^\top U\cdot h^\alpha + o(h^\alpha),\end{aligned}$$

where $o(\cdot)$ hides a constant that does not depend on $\boldsymbol{\beta}$. This completes the proof of (67).

To show (68), further recall that in **Step 2** of the proof of (40), we show that

$$\mathbb{E}_{\cdot\,|\,\boldsymbol{Z}}\left[\boldsymbol{v}^\top U_{h,i}(\boldsymbol{\beta}^*)\right]^2 = \frac{\left(\boldsymbol{v}^\top\boldsymbol{Z}\right)^2}{h}\int_{-1}^{1}\left[H'(\xi)\right]^2\rho(0\,|\,\boldsymbol{Z})\,\mathrm{d}\xi + O\left((\boldsymbol{v}^\top\boldsymbol{Z})^2\right).$$

Since $V_s := \pi_V\mathbb{E}\rho(0\,|\,\boldsymbol{Z})\,\boldsymbol{Z}\boldsymbol{Z}^\top$, $\pi_V := \int_{-1}^{1}\left[H'(\xi)\right]^2\mathrm{d}\xi$, and $\mathbb{E}(\boldsymbol{v}^\top\boldsymbol{Z})^2<\infty$, it holds that

$$\mathbb{E}\left[\boldsymbol{v}^\top U_{h,i}(\boldsymbol{\beta}^*)\right]^2 = \frac{1}{h}\boldsymbol{v}^\top V_s\boldsymbol{v} + O(1).$$

Therefore,

$$\begin{aligned}&\mathrm{var}\left[\sqrt{h}\boldsymbol{v}^\top U_{h,i}(\boldsymbol{\beta}^*)\right]\\ &= h\left\{\mathbb{E}\left[\boldsymbol{v}^\top U_{h,i}(\boldsymbol{\beta}^*)\right]^2 - \left(\mathbb{E}\left[\boldsymbol{v}^\top U_{h,i}(\boldsymbol{\beta}^*)\right]\right)^2\right\}\\ &= \boldsymbol{v}^\top V_s\boldsymbol{v} + o(1),\end{aligned} \tag{69}$$

By CLT and Slusky's Theorem,

$$\sqrt{nh}(1-\mathbb{E})\left[\boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta}^*)\right]/\sqrt{\boldsymbol{v}^\top V_s\boldsymbol{v}} \xrightarrow{d} \mathcal{N}(0,1).$$

Furthermore, in **Step 1**, we show that

$$\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\le\delta}\left|(1-\mathbb{E})\left[\boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta}) - \boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta}^*)\right]\right| = O_{\mathbb{P}}\left(\delta\sqrt{\frac{p\log n}{nh^3}}\right),$$

which yields (68) if $\delta = o\left(h/\sqrt{p\log n}\right)$.

□

Proof of Theorem 3.4

We now give the proof for Theorem 3.4 using Lemma B.1.

PROOF. By Theorem 3.3, when $T \geq \log_2\left(\frac{3\alpha}{2\alpha+1}\cdot\frac{\log(n/p)}{\log(m/p)}\right)$, it holds that $h_T = (p/n)^{\frac{1}{2\alpha+1}}$ and $(p/m)^{2^T/3} \lesssim (p/n)^{\alpha/(2\alpha+1)}$. By taking $h_{T+1} = (\lambda_h/n)^{1/(2\alpha+1)}$, we have

$$\left\|\widehat{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\right\|_2 = O_{\mathbb{P}}\left((p/n)^{\frac{\alpha}{2\alpha+1}}\right) = o_{\mathbb{P}}\left(\max\{h_{T+1}^{\alpha/2}, h_{T+1}/\sqrt{p\log n}\}\right),$$

where we use the assumption $p = o\left(n^{\frac{2(\alpha-1)}{4\alpha+1}}(\log n)^{-\frac{2\alpha+1}{4\alpha+1}}\right)$. Hence, the assumptions of Lemma B.1 hold for $\widehat{\boldsymbol{\beta}}^{(T)}$ and h_{T+1}. Since $\sqrt{nh_{T+1}} = \sqrt{\lambda_h}h_{T+1}^{-\alpha}$, we obtain by Lemma B.1 that

$$\frac{\sqrt{nh_{T+1}}\cdot\boldsymbol{\vartheta}^\top V^{-1}U_{n,h_{T+1}}\left(\widehat{\boldsymbol{\beta}}^{(T)}\right) - \sqrt{\lambda_h}\boldsymbol{\vartheta}^\top V^{-1}U}{\sqrt{\boldsymbol{\vartheta}^\top V^{-1}V_s V\boldsymbol{\vartheta}}} \xrightarrow{d} \mathcal{N}(0,1).$$

It is easy to verify that $p\log n/(nh_{T+1}^3) = o(1)$ and $h_{T+1} = O(\delta_{m,T})$. By (41) and Assumption 4, $\left\|V_{n,h_{T+1}}^{-1}\left(\widehat{\boldsymbol{\beta}}^{(T)}\right) - V^{-1}\right\|_2 = o_{\mathbb{P}}(1)$, which yields

$$\frac{\sqrt{nh_{T+1}}\cdot\boldsymbol{\vartheta}^\top V_{n,h_{T+1}}^{-1}\left(\widehat{\boldsymbol{\beta}}^{(T)}\right)U_{n,h_{T+1}}\left(\widehat{\boldsymbol{\beta}}^{(T)}\right) - \sqrt{\lambda_h}\boldsymbol{\vartheta}^\top V^{-1}U}{\sqrt{\boldsymbol{\vartheta}^\top V^{-1}V_s V^{-1}\boldsymbol{\vartheta}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Plugging in $h_{T+1} = \left(\frac{\lambda_h}{n}\right)^{\frac{1}{2\alpha+1}}$ and $\widehat{\boldsymbol{\beta}}^{(T+1)} - \boldsymbol{\beta}^* = V_{n,h_{T+1}}^{-1}\left(\widehat{\boldsymbol{\beta}}^{(T)}\right)U_{n,h_{T+1}}\left(\widehat{\boldsymbol{\beta}}^{(T)}\right)$ leads to (14). Using the asymptotic bias and variance, we have

$$\begin{aligned}&\mathbb{E}\left[\left(\boldsymbol{\vartheta}^\top\widehat{\boldsymbol{\beta}}^{(T+1)} - \boldsymbol{\vartheta}^\top\boldsymbol{\beta}^*\right)^2\right]\\ &\asymp n^{-\frac{2\alpha}{2\alpha+1}}\left[\lambda_h^{-\frac{1}{2\alpha+1}}\boldsymbol{\vartheta}^\top V^{-1}V_s V^{-1}\boldsymbol{\vartheta} + \lambda_h^{\frac{2\alpha}{2\alpha+1}}U^\top V^{-1}\boldsymbol{\vartheta}\boldsymbol{\vartheta}^\top V^{-1}U\right],\end{aligned}$$

by minimizing which it is straightforward to obtain the optimal λ_h^* given in (15). □

Estimators for V, U and V_s

Now we formally define the estimators for V, U and V_s. Proposition B.1 in Section B.1 has already implies that, when $T \geq \log_2\left(\frac{3\alpha}{2\alpha+1}\cdot\frac{\log(n/p)}{\log(m/p)}\right)$, it holds that $V_{n,h_T}\left(\widehat{\boldsymbol{\beta}}^{(T)}\right) \xrightarrow{p} V$, where $V_{n,h}(\boldsymbol{\beta})$ is defined in (38), so we can use $\widehat{V} := V_{n,h_T}\left(\widehat{\boldsymbol{\beta}}^{(T)}\right)$ to estimate V. It remains to provide estimators for U and V_s.

THEOREM B.2. *Assume assumptions in Theorem 3.3 hold. Let $h_\kappa = (p/n)^{\frac{\kappa}{2\alpha+1}}$ for some $0 < \kappa < 1$. Define*

$$\widehat{U} := \frac{1}{nh_\kappa^{\alpha+1}}\sum_{i=1}^n y_i H'\left(\frac{x_i + \boldsymbol{z}_i^\top\widehat{\boldsymbol{\beta}}^{(T)}}{h_\kappa}\right)\boldsymbol{z}_i, \tag{70}$$

$$\widehat{V}_s := \frac{1}{nh_T}\sum_{i=1}^n\left[H'\left(\frac{x_i + \boldsymbol{z}_i^\top\widehat{\boldsymbol{\beta}}^{(T)}}{h_T}\right)\right]^2\boldsymbol{z}_i\boldsymbol{z}_i^\top. \tag{71}$$

When $T \geq \log_2\left(\frac{3\alpha}{2\alpha+1}\cdot\frac{\log(n/p)}{\log(m/p)}\right)$, we have

$$\left\|\widehat{U} - U\right\|_2 = o_{\mathbb{P}}(1), \quad \left\|\widehat{V}_s - V_s\right\|_2 = o_{\mathbb{P}}(1). \tag{72}$$

PROOF. When $T \geq \log_2\left(\frac{3\alpha}{2\alpha+1} \cdot \frac{\log(n/p)}{\log(m/p)}\right)$, we have

$$(p/n)^{\frac{\kappa}{2\alpha+1}} = h_\kappa \gg h_T = (p/n)^{\frac{1}{2\alpha+1}},$$

and

$$\left\|\widehat{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\right\|_2 = O_{\mathbb{P}}(\delta_{m,T}), \quad \delta_{m,T} = (p/n)^{\frac{\alpha}{2\alpha+1}} = o\left(h_\kappa^\alpha\right).$$

It is easy to verify that $p\log n/(nh_\kappa^3) = o(1)$. By the proof of Equation (40), we have that

$$\left\|U_{n,h_\kappa}\left(\widehat{\boldsymbol{\beta}}^{(T)}\right) - Uh_\kappa^\alpha\right\|_2 = O_{\mathbb{P}}\left((p/n)^{\alpha/(2\alpha+1)}\right).$$

Also, it holds that $\left\|V_{n,h_\kappa}\left(\widehat{\boldsymbol{\beta}}^{(T)}\right) - V\right\|_2 = o_{\mathbb{P}}(1)$ by (41). Note that

$$\widehat{U} = (h_\kappa)^{-\alpha}\left[U_{n,h_\kappa}\left(\widehat{\boldsymbol{\beta}}^{(T)}\right) - V_{n,h_\kappa}\left(\widehat{\boldsymbol{\beta}}^{(T)}\right)\left(\widehat{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\right)\right],$$

and $(h_\kappa)^{-\alpha}(p/n)^{\alpha/(2\alpha+1)} = o(1)$, which yields $\left\|\widehat{U} - U\right\|_2 = o_{\mathbb{P}}(1)$.

To prove $\left\|\widehat{V}_s - V_s\right\|_2 = o_{\mathbb{P}}(1)$, recall that

$$\widehat{V}_s := \frac{1}{nh_T}\sum_{i=1}^n \left[H'\left(\frac{x_i + \boldsymbol{z}_i^\top \widehat{\boldsymbol{\beta}}^{(T)}}{h_T}\right)\right]^2 \boldsymbol{z}_i\boldsymbol{z}_i^\top.$$

For any $\boldsymbol{v} \in \mathbb{S}^p$,

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{v}^\top \widehat{V}_s \boldsymbol{v}\right] &= \mathbb{E}\left(\mathbb{E}_{\cdot\,|\,\boldsymbol{Z}} \boldsymbol{v}^\top \widehat{V}_s \boldsymbol{v}\right) \\
&= \mathbb{E}\left[\left(\boldsymbol{Z}^\top \boldsymbol{v}\right)^2 \int_{-1}^1 \left[H'(\xi)\right]^2 \rho\left(\xi h_T - \boldsymbol{Z}^\top\left(\widehat{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\right) \,|\, \boldsymbol{Z}\right) \mathrm{d}\xi\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{Z}^\top \boldsymbol{v}\right)^2 \pi_V \rho(0\,|\,\boldsymbol{Z}) + O\left(h_T + \delta_{m,T}(\boldsymbol{Z}^\top \boldsymbol{u})\right)\right] \\
&= \boldsymbol{v}^\top V_s \boldsymbol{v} + o(1),
\end{aligned}$$

where $\boldsymbol{u} = (\widehat{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*)/\left\|\widehat{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\right\|_2$ and hence $\mathbb{E}[(\boldsymbol{Z}^\top \boldsymbol{u})^2] < \infty$. Also, since $H'(x)$ is bounded, $\operatorname{var}\left(\boldsymbol{v}^\top \widehat{V}_s \boldsymbol{v}\right) = O\left(\frac{1}{nh_T^2}\right) = o(1)$. Then $\left\|\widehat{V}_s - V_s\right\|_2 = o_{\mathbb{P}}(1)$ is proved by Chebyshev's inequality.

□

B.3. Proof of the Results for `(Avg-SMSE)`.

Proof of Theorem 3.1

PROOF. Since $\widehat{\boldsymbol{\beta}}_{\mathrm{SMSE},\ell}$ is the minimizer of $F_{h,\ell}(\boldsymbol{\beta}) = \sum_{i\in\mathcal{H}_\ell}(-y_i) H\left(\frac{x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}}{h}\right)$, we have $\nabla_{\boldsymbol{\beta}} F_{h,\ell}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{SMSE},\ell}\right) = 0$. By Taylor's expansion of $\nabla_{\boldsymbol{\beta}} F_{h,\ell}$ at $\boldsymbol{\beta}^*$, we have

$$0 = \nabla_{\boldsymbol{\beta}} F_{h,\ell}(\boldsymbol{\beta}^*) + \nabla^2_{\boldsymbol{\beta}} F_{h,\ell}\left(\breve{\boldsymbol{\beta}}_\ell\right)\left(\widehat{\boldsymbol{\beta}}_{\mathrm{SMSE},\ell} - \boldsymbol{\beta}^*\right),$$

where $\breve{\boldsymbol{\beta}}_\ell$ is between $\widehat{\boldsymbol{\beta}}_{\mathrm{SMSE},\ell}$ and $\boldsymbol{\beta}^*$.

Define

$$U_{m,\ell,h}(\boldsymbol{\beta}) := \frac{1}{m}\sum_{i\in\mathcal{H}_\ell} U_{h,i}(\boldsymbol{\beta}), \quad V_{m,\ell,h}(\boldsymbol{\beta}) := \frac{1}{m}\sum_{i\in\mathcal{H}_\ell} V_{h,i}(\boldsymbol{\beta}),$$

where $U_{h,i}(\boldsymbol{\beta})$ and $V_{h,i}(\boldsymbol{\beta})$ are defined in (45) and (59). By definition, $\nabla^2_{\boldsymbol{\beta}} F_{h,\ell}\left(\breve{\boldsymbol{\beta}}_\ell\right) = V_{m,\ell,h}\left(\breve{\boldsymbol{\beta}}_\ell\right)$ and $\nabla_{\boldsymbol{\beta}} F_{h,\ell}(\boldsymbol{\beta}^*) = -U_{m,\ell,h}(\boldsymbol{\beta}^*)$, which yields

$$\widehat{\boldsymbol{\beta}}_{\mathrm{SMSE},\ell} - \boldsymbol{\beta}^* = V^{-1}_{m,\ell,h}\left(\breve{\boldsymbol{\beta}}_\ell\right) U_{m,\ell,h}(\boldsymbol{\beta}^*).$$

Following the proof of Proposition 3.2, if $p\log m/(mh^3) = o(1)$, we have that for any sufficiently large γ, there exists a constant C such that

$$\sup_\ell \sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|\le\delta} \|V - V_{m,\ell,h}(\boldsymbol{\beta})\|_2 \le C\left(\sqrt{\frac{p\log m}{mh^3}} + h^\alpha + \delta\right),$$

with probability at least $1 - Lm^{-\gamma/2} - 2L(5m^{-\gamma})^p$. The assumptions $L = o\left(m^{2(\alpha-1)/3}/(p\log m)^{(2\alpha+1)/3}\right)$ and $h = (\lambda_h/n)^{1/(2\alpha+1)}$ ensure that $p\log m/(mh^3) = o(1)$ and $L(m^{-\gamma/2} - 2(5m^{-\gamma})^p) = o(1)$ for sufficiently large γ. Moreover, Theorem 1 in Horowitz (1992) showed that $\left\|\widehat{\boldsymbol{\beta}}_{\mathrm{SMSE},\ell} - \boldsymbol{\beta}^*\right\|_2 = o(1)$ almost surely for all ℓ, and thus there exists a uniform high-probability bound for $V^{-1} - V^{-1}_{m,\ell,h}\left(\breve{\boldsymbol{\beta}}_\ell\right)$ over all machines:

$$\sup_\ell \left\|V^{-1} - V^{-1}_{m,\ell,h}\left(\breve{\boldsymbol{\beta}}_\ell\right)\right\|_2 = o_{\mathbb{P}}(1),$$

which implies that

$$\widehat{\boldsymbol{\beta}}_{(\texttt{Avg-SMSE})} - \boldsymbol{\beta}^* = \frac{1}{L}\sum_{\ell=1}^{L}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{SMSE},\ell} - \boldsymbol{\beta}^*\right) = V^{-1}U_{n,h}(\boldsymbol{\beta}^*) + U_{n,h}(\boldsymbol{\beta}^*)\, o_{\mathbb{P}}(1),$$

using the facts that $U_{n,h}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n U_{h,i}(\boldsymbol{\beta})$ and $n = mL$.

By Lemma B.1, for any $\boldsymbol{\vartheta} \in \mathbb{S}^{p-1}\setminus\{0\}$,

$$\mathbb{E}\left[\boldsymbol{\vartheta}^\top V^{-1}U_{n,h}(\boldsymbol{\beta}^*)\right] = \boldsymbol{\vartheta}^\top V^{-1}Uh^\alpha + o(h^\alpha),$$

and

$$\sqrt{mLh}\cdot(1-\mathbb{E})\left[\boldsymbol{\vartheta}^\top V^{-1}U_{n,h}(\boldsymbol{\beta}^*)\right]/\sqrt{\boldsymbol{\vartheta}^\top V^{-1}V_sV^{-1}\boldsymbol{\vartheta}} \xrightarrow{d} \mathcal{N}(0,1).$$

When $h = \left(\frac{\lambda_h}{n}\right)^{\frac{1}{2\alpha+1}}$, we have $\sqrt{\lambda_h}h^{-\alpha} = \sqrt{mLh}$, and thus

$$\begin{aligned}
&\frac{\sqrt{mLh}\boldsymbol{\vartheta}^\top\left(\widehat{\boldsymbol{\beta}}_{(\texttt{Avg-SMSE})} - \boldsymbol{\beta}^*\right) - \sqrt{\lambda_h}\boldsymbol{\vartheta}^\top V^{-1}U}{\sqrt{\boldsymbol{\vartheta}^\top V^{-1}V_sV^{-1}\boldsymbol{\vartheta}}} \\
&= \sqrt{mLh}(1-\mathbb{E})\left[\boldsymbol{\vartheta}^\top V^{-1}U_{n,h}(\boldsymbol{\beta}^*)\right]/\sqrt{\boldsymbol{\vartheta}^\top V^{-1}V_sV^{-1}\boldsymbol{\vartheta}} + o(1) \\
&\quad + o_{\mathbb{P}}(1)\sqrt{mLh}\left(\boldsymbol{\vartheta}^\top U_{n,h}(\boldsymbol{\beta}^*) - \mathbb{E}\boldsymbol{\vartheta}^\top U_{n,h}(\boldsymbol{\beta}^*) + \mathbb{E}\boldsymbol{\vartheta}^\top U_{n,h}(\boldsymbol{\beta}^*)\right)/\sqrt{\boldsymbol{\vartheta}^\top V^{-1}V_sV^{-1}\boldsymbol{\vartheta}} \\
&\xrightarrow{d} \mathcal{N}(0,1),
\end{aligned}$$

which proves (9). Furthermore, if $h \gtrsim \left(\frac{1}{mL}\right)^{\frac{1}{2\alpha+1}}$, we have $h^{-\alpha} \lesssim \sqrt{mLh}$, and thus

$$\boldsymbol{\vartheta}^\top\left(\widehat{\boldsymbol{\beta}}_{(\texttt{Avg-SMSE})} - \boldsymbol{\beta}^*\right) - \boldsymbol{\vartheta}^\top V^{-1}U = o_{\mathbb{P}}(h^\alpha).$$

If $h \lesssim \left(\frac{1}{mL}\right)^{\frac{1}{2\alpha+1}}$ but still satisfies $\frac{p\log m}{mh^3} = o(1)$, we have $h^{-\alpha} \gtrsim \sqrt{mLh}$, and thus

$$\sqrt{mLh}\mathbb{E}\left[\boldsymbol{\vartheta}^\top U_{n,h}(\boldsymbol{\beta}^*)\right] = o_{\mathbb{P}}(1),$$

which yields

$$\sqrt{mLh}\boldsymbol{\vartheta}^\top\left(\widehat{\boldsymbol{\beta}}_{(\texttt{Avg}-\texttt{SMSE})} - \boldsymbol{\beta}^*\right) / \sqrt{\boldsymbol{\vartheta}^\top V^{-1} V_s V^{-1}\boldsymbol{\vartheta}} \xrightarrow{d} \mathcal{N}(0,1).$$

□

B.4. Proof of the Results for Data Heterogeneity.

Proof of Theorem 4.1

PROOF. In the proof of Theorem 3.1, we show that $\widehat{\boldsymbol{\beta}}_{\mathrm{SMSE},\ell} - \boldsymbol{\beta}^* = V_{m_\ell,\ell,h}^{-1}(\breve{\boldsymbol{\beta}}_\ell) U_{m_\ell,\ell,h}(\boldsymbol{\beta}^*)$ for any ℓ, where $\breve{\boldsymbol{\beta}}_\ell$ is between $\boldsymbol{\beta}^*$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{SMSE},\ell}$. The proof of (41) shows that, if $p\log m_\ell/(m_\ell h^3) = o(1)$, then for any sufficiently large γ, there exists a constant C_ℓ such that

$$\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|\le\delta} \|V_\ell - V_{m,\ell,h}(\boldsymbol{\beta})\|_2 \le C_\ell\left(\sqrt{\frac{p\log m_\ell}{m_\ell h^3}} + h^\alpha + \delta\right),$$

with probability at least $1 - m_\ell^{-\gamma/2} - 2\left(5m_\ell^{-\gamma}\right)^p$. The dependence of C_ℓ on ℓ is due to the distributions ρ_ℓ and F_ℓ and their derivatives, which can be uniformly upper bounded over all ℓ by Assumptions 6 and 7. Moreover, the constant γ does not depend on ℓ. Therefore, if $h = n^{-1/(2\alpha+1)}$ and $\min_\ell m_\ell \gtrsim pn^{3/(2\alpha+1)}\log n$, which satisfies $p\log\min m_\ell/(\min m_\ell h^3) = o(1)$, then for any sufficiently large γ, there exists a constant C (hidden in $o(1)$) such that

$$\sup_\ell \left\|V_\ell^{-1} - V_{m_\ell,\ell,h}^{-1}(\breve{\boldsymbol{\beta}}_\ell)\right\|_2 = o(1),$$

with probability at least $1 - \sum_{\ell=1}^L m_\ell^{-\gamma/2} - 2\sum_{\ell=1}^L\left(5m_\ell^{-\gamma}\right)^p$. The condition $\min_\ell m_\ell \gtrsim pn^{3/(2\alpha+1)}\log n$ also ensures that $\sum_{\ell=1}^L m_\ell^{-\gamma/2} + 2\sum_{\ell=1}^L\left(5m_\ell^{-\gamma}\right)^p = o(1)$ for sufficiently large γ. Therefore, we obtain that

$$\widehat{\boldsymbol{\beta}}_{(\texttt{wAvg}-\texttt{SMSE})} - \boldsymbol{\beta}^* = \sum_{\ell=1}^L W_\ell V_\ell^{-1} U_{m_\ell,\ell,h}(\boldsymbol{\beta}^*) + o_{\mathbb{P}}(1)\sum_{\ell=1}^L W_\ell U_{m_\ell,\ell,h}(\boldsymbol{\beta}^*). \tag{73}$$

By Step 3 in the proof of (40) and the uniformness in Assumptions 6 and 7, for any $\boldsymbol{v}$ satisfying $\|\boldsymbol{v}\|_2 = 1$, it holds that

$$\mathbb{E}\left[\sum_{\ell=1}^L \boldsymbol{v}^\top W_\ell V_\ell^{-1} U_{m_\ell,\ell,h}(\boldsymbol{\beta}^*)\right] = \sum_{\ell=1}^L \boldsymbol{v}^\top W_\ell V_\ell^{-1} U_\ell h^\alpha + o(h^\alpha) = \boldsymbol{v}^\top B_{(\texttt{wAvg}-\texttt{SMSE})} + o(h^\alpha), \tag{74}$$

where $B_{(\texttt{wAvg}-\texttt{SMSE})} := \sum_{\ell=1}^L W_\ell V_\ell^{-1} U_\ell h^\alpha$.

Define $\Phi_i := (1-\mathbb{E})m_\ell^{-1}\boldsymbol{v}^\top W_\ell V_\ell^{-1} U_{h,i}(\boldsymbol{\beta}^*)$ for $i \in \mathcal{H}_\ell$. (Notice that the definitions of Φ_i may be different in the proof of different theorems.) Then

$$(1-\mathbb{E})\left[\sum_{\ell=1}^L \boldsymbol{v}^\top W_\ell V_\ell^{-1} U_{m_\ell,\ell,h}(\boldsymbol{\beta}^*)\right] = (1-\mathbb{E})\left[\sum_{\ell=1}^L\sum_{i\in\mathcal{H}_\ell} m_\ell^{-1}\boldsymbol{v}^\top W_\ell V_\ell^{-1} U_{h,i}(\boldsymbol{\beta}^*)\right] = \sum_{i=1}^n \Phi_i. \tag{75}$$

Recall that by (69),

$$s_n^2 := \sum_{i=1}^{n} \operatorname{var}\left[\Phi_i\right] = h^{-1} \sum_{\ell=1}^{L} m_\ell^{-1} \boldsymbol{v}^\top W_\ell V_\ell^{-1} \left[V_{s,\ell} + o(1)\right] V_\ell^{-1} W_\ell^\top \boldsymbol{v}.$$

Similar to the proof of (69), we have that

$$\mathbb{E}\left[\Phi_i^4\right] \lesssim \frac{\|W_\ell\|_2^4}{m_\ell^4 h^3}, \qquad \text{when } i \in \mathcal{H}_\ell.$$

By Assumptions 8 and 9, the Lindeberg's condition can be verified as follows:

$$\frac{1}{s_n^2} \sum_{i=1}^{n} \mathbb{E}\left[(\Phi_i)^2 \mathbb{I}\left(|\Phi_i| > \varepsilon s_n\right)\right] \le \sum_{i=1}^{n} \frac{\mathbb{E}\left[\Phi_i^4\right]}{\varepsilon^2 s_n^4} \lesssim \frac{\sum_{\ell=1}^{L} \left[\|W_\ell\|_2^4 / m_\ell^3\right]}{h \sum_{\ell=1}^{L} \left[\|W_\ell\|_2^2 / m_\ell\right]} \asymp \frac{1}{nh} \to 0.$$

Therefore, since $s_n^2 \to \boldsymbol{v}^\top \Sigma_{(\texttt{wAvg-SMSE})} \boldsymbol{v}$, we have

$$\left(\boldsymbol{v}^\top \Sigma_{(\texttt{wAvg-SMSE})} \boldsymbol{v}\right)^{-1/2} \sum_{i=1}^{n} \Phi_i \xrightarrow{d} \mathcal{N}(0,1),$$

where $\Sigma_{(\texttt{wAvg-SMSE})} := h^{-1} \sum_{\ell=1}^{L} m_\ell^{-1} W_\ell V_\ell^{-1} V_{s,\ell} V_\ell^{-1} W_\ell^\top$. Now we do the following decomposition.

$$\begin{aligned} &\left(\boldsymbol{v}^\top \Sigma_{(\texttt{wAvg-SMSE})} \boldsymbol{v}\right)^{-1/2} \boldsymbol{v}^\top \left(\widehat{\boldsymbol{\beta}}_{(\texttt{wAvg-SMSE})} - \boldsymbol{\beta}^* - B_{(\texttt{wAvg-SMSE})}\right) \\ &= \left(\boldsymbol{v}^\top \Sigma_{(\texttt{wAvg-SMSE})} \boldsymbol{v}\right)^{-1/2} \sum_{i=1}^{n} \Phi_i + \left(\boldsymbol{v}^\top \Sigma_{(\texttt{wAvg-SMSE})} \boldsymbol{v}\right)^{-1/2} o(h^\alpha) \\ &\quad + o_{\mathbb{P}}(1) \left(\boldsymbol{v}^\top \Sigma_{(\texttt{wAvg-SMSE})} \boldsymbol{v}\right)^{-1/2} \sum_{\ell=1}^{L} \boldsymbol{v}^\top W_\ell U_{m_\ell,\ell,h_\ell}(\boldsymbol{\beta}^*). \end{aligned} \tag{76}$$

Assumptions 8 and 9 ensures that

$$\left(\boldsymbol{v}^\top \Sigma_{(\texttt{wAvg-SMSE})} \boldsymbol{v}\right)^{1/2} \asymp 1/\sqrt{nh} \asymp h^\alpha,$$

with $h = n^{-1/(2\alpha+1)}$, which implies that the second term on the RHS of (76) is $o(1)$. Meanwhile, it is straightforward to show that $\left(\boldsymbol{v}^\top \Sigma_{(\texttt{wAvg-SMSE})} \boldsymbol{v}\right)^{-1/2} \sum_{\ell=1}^{L} W_\ell U_{m_\ell,\ell,h_\ell}(\boldsymbol{\beta}^*)$ is bounded in probability by letting $V_\ell = I_{p\times p}$ in the previous analysis. Hence,

$$\left(\boldsymbol{v}^\top \Sigma_{(\texttt{wAvg-SMSE})} \boldsymbol{v}\right)^{-1/2} \boldsymbol{v}^\top \left(\widehat{\boldsymbol{\beta}}_{(\texttt{wAvg-SMSE})} - \boldsymbol{\beta}^* - B_{(\texttt{wAvg-SMSE})}\right) = \left(\boldsymbol{v}^\top \Sigma_{(\texttt{wAvg-SMSE})} \boldsymbol{v}\right)^{-1/2} \sum_{i=1}^{n} \Phi_i + o_{\mathbb{P}}(1),$$

which converges to $\mathcal{N}(0,1)$ in distribution. Then rewrite

$$\Sigma_{(\texttt{wAvg-SMSE})} = n^{-2\alpha/(2\alpha+1)} \sum_{\ell=1}^{L} \frac{n}{m_\ell} W_\ell V_\ell^{-1} V_{s,\ell} V_\ell^{-1} W_\ell^\top, \quad B_{(\texttt{wAvg-SMSE})} = n^{-\alpha/(2\alpha+1)} \sum_{\ell=1}^{L} W_\ell V_\ell^{-1} U_\ell,$$

which completes the proof of Theorem 4.1.

Additionally, we show that, for any vector $\boldsymbol{v}$, $\boldsymbol{v}^\top \Sigma_{(\texttt{wAvg-SMSE})} \boldsymbol{v}$ is minimized at

$$W_\ell^{*,(\texttt{wAvg-SMSE})} = \left(\sum_{\ell=1}^{L} m_\ell V_\ell V_{s,\ell}^{-1} V_\ell\right)^{-1} m_\ell V_\ell V_{s,\ell}^{-1} V_\ell.$$

We are to solve the following optimization problem:

$$\min_{W_\ell} \sum_{\ell=1}^{L} \frac{n}{m_\ell} \boldsymbol{v}^\top W_\ell V_\ell^{-1} V_{s,\ell} V_\ell^{-1} W_\ell^\top \boldsymbol{v}, \text{ s.t.} \sum_{\ell=1}^{L} W_\ell = I_{p\times p}.$$

The Lagrangian is

$$\sum_{\ell=1}^{L} \frac{n}{m_\ell} \boldsymbol{v}^\top W_\ell V_\ell^{-1} V_{s,\ell} V_\ell^{-1} W_\ell^\top \boldsymbol{v} + \left\langle \Lambda, \sum_{\ell=1}^{L} W_\ell - I_{p\times p} \right\rangle.$$

By taking derivative w.r.t W_ℓ and letting the derivative be zero, we obtain $2\frac{n}{m_\ell} \boldsymbol{v}\boldsymbol{v}^\top W_\ell V_\ell^{-1} V_{s,\ell} V_\ell^{-1} + \Lambda = \mathbf{0}$ or $2n\boldsymbol{v}\boldsymbol{v}^\top W_\ell = -\Lambda m_\ell V_\ell V_{s,\ell}^{-1} V_\ell$. By the constraint that $\sum_{\ell=1}^{L} W_\ell = I_{p\times p}$, we obtain that a sufficient condition for $\{W_\ell^*\}$ to be a minimizer is

$$\boldsymbol{v}\boldsymbol{v}^\top W_\ell^* = \boldsymbol{v}\boldsymbol{v}^\top \big(\sum_{\ell=1}^{L} m_\ell V_\ell V_{s,\ell}^{-1} V_\ell \big)^{-1} m_\ell V_\ell V_{s,\ell}^{-1} V_\ell,$$

and it is clear that the weight $\left\{ W_\ell^{*,(\texttt{wAvg}-\texttt{SMSE})} \right\}$ defined above satisfies this condition for all $\boldsymbol{v}$.

□

Proof of Theorem 4.2

PROOF. For (wmSMSE), we have

$$\widehat{\boldsymbol{\beta}}_{(\texttt{wmSMSE})}^{(t)} - \boldsymbol{\beta}^* = \left(\sum_{\ell=1}^{L} \sum_{i\in\mathcal{H}_\ell} m_\ell^{-1} W_\ell V_{h_t,i} \left(\widehat{\boldsymbol{\beta}}^{(t-1)} \right) \right)^{-1} \left(\sum_{\ell=1}^{L} \sum_{i\in\mathcal{H}_\ell} m_\ell^{-1} W_\ell U_{h_t,i} \left(\widehat{\boldsymbol{\beta}}^{(t-1)} \right) \right).$$

By Assumption 9, $W_\ell/m_\ell \asymp 1/n$, and hence, analogous the proof of Proposition 3.2, we can show that, if $\left\| \widehat{\boldsymbol{\beta}}_{(\texttt{wmSMSE})}^{(t-1)} - \boldsymbol{\beta}^* \right\|_2 = O_{\mathbb{P}}(\delta_{t-1})$, then

$$\left\| \widehat{\boldsymbol{\beta}}_{(\texttt{wmSMSE})}^{(t)} - \boldsymbol{\beta}^* \right\|_2 = O_{\mathbb{P}} \left(\sqrt{\frac{p}{nh_t}} + \delta_{t-1} \sqrt{\frac{p\log n}{nh_t^3}} + \delta_{t-1}^2 + h_t^\alpha \right). \tag{77}$$

Using the same analysis as in the proof of Theorem 3.3, by taking $h_t = \max\left\{ (p/n)^{1/(2\alpha+1)}, \delta_0^{2^t/\alpha} \right\}$ for $t = 1, \ldots, T$ and $T > \log_2 \left(\frac{\alpha}{2\alpha+1} \frac{\log(n/p)}{\log(1/\delta_0)} \right)$, we have $h_T = (p/n)^{1/(2\alpha+1)}$ and $\delta_T = (p/n)^{\alpha/(2\alpha+1)}$, which satisfies $\delta_T = o\left(\max\left\{ h_{T+1}^{\alpha/2}, h_{T+1}/\sqrt{p\log n} \right\} \right)$ with $h_{T+1} = n^{-1/(2\alpha+1)}$ and $p = o\left(n^{\frac{2(\alpha-1)}{4\alpha+1}} (\log n)^{-\frac{2\alpha+1}{4\alpha+1}} \right)$. Therefore, similar to the proof of (67), for any $\boldsymbol{v} \in \mathbb{S}^{p-1}$, we have

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \le \delta_T} \mathbb{E} \left(\sum_{\ell=1}^{L} \sum_{i\in\mathcal{H}_\ell} m_\ell^{-1} \boldsymbol{v}^\top W_\ell U_{h_{T+1},i}(\boldsymbol{\beta}) \right) = \boldsymbol{v}^\top \overline{U}_W h_{T+1}^\alpha + o\left(h_{T+1}^\alpha \right),$$

where $\overline{U}_W := \sum_{\ell=1}^{L} W_\ell U_\ell$. By **Step 1** in the proof of (40) and Assumption 9, it holds that

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \le \delta_T} (1-\mathbb{E}) \left(\sum_{\ell=1}^{L} \sum_{i\in\mathcal{H}_\ell} m_\ell^{-1} \boldsymbol{v}^\top W_\ell \left[U_{h_{T+1},i}(\boldsymbol{\beta}) - U_{h_{T+1},i}(\boldsymbol{\beta}^*) \right] \right)$$

$$= O_{\mathbb{P}} \left(\delta_T \sqrt{\frac{p\log n}{nh_{T+1}^3}} \right) = o_{\mathbb{P}} \left(\frac{1}{\sqrt{nh_{T+1}}} \right).$$

Define $\Phi_i := (1-\mathbb{E})\left(m_\ell^{-1}\boldsymbol{v}^\top W_\ell U_{h,i}(\boldsymbol{\beta}^*)\right)$ for $i \in \mathcal{H}_\ell$. Then

$$s_n^2 := \sum_{i=1}^n \operatorname{var}\left[\Phi_i\right] = h^{-1}\sum_{\ell=1}^L m_\ell^{-1}\boldsymbol{v}^\top W_\ell\left[V_{s,\ell}+o(1)\right]W_\ell^\top\boldsymbol{v}.$$

Similar to the proof of Theorem 4.1, we have $\left(h^{-1}\sum_{\ell=1}^L m_\ell^{-1}\boldsymbol{v}^\top W_\ell V_{s,\ell}W_\ell^\top\boldsymbol{v}\right)^{-1/2}\sum_{i=1}^n\Phi_i \xrightarrow{d}$ $\mathcal{N}(0,1)$ and $h_{T+1}^{-1}\sum_{\ell=1}^L m_\ell^{-1}\boldsymbol{v}^\top W_\ell V_{s,\ell}W_\ell^\top\boldsymbol{v} \asymp 1/\sqrt{nh_{T+1}} = h_{T+1}^\alpha$, and thus

$$\begin{aligned}
&\left(h_{T+1}^{-1}\sum_{\ell=1}^L m_\ell^{-1}\boldsymbol{v}^\top W_\ell V_{s,\ell}W_\ell^\top\boldsymbol{v}\right)^{-1/2}\left(\sum_{\ell=1}^L\sum_{i\in\mathcal{H}_\ell} m_\ell^{-1}\boldsymbol{v}^\top W_\ell U_{h_{T+1},i}\left(\widehat{\boldsymbol{\beta}}^{(T)}\right)-\boldsymbol{v}^\top\overline{U}_W h_{T+1}^\alpha\right)\\
&=\left(h_{T+1}^{-1}\sum_{\ell=1}^L m_\ell^{-1}\boldsymbol{v}^\top W_\ell V_{s,\ell}W_\ell^\top\boldsymbol{v}\right)^{-1/2}\left(\sum_{\ell=1}^L\sum_{i\in\mathcal{H}_\ell}(1-\mathbb{E})m_\ell^{-1}\boldsymbol{v}^\top W_\ell U_{h_{T+1},i}\left(\boldsymbol{\beta}^*\right)\right)\\
&\quad+\left(h_{T+1}^{-1}\sum_{\ell=1}^L m_\ell^{-1}\boldsymbol{v}^\top W_\ell V_{s,\ell}W_\ell^\top\boldsymbol{v}\right)^{-1/2}\left(\sum_{\ell=1}^L\sum_{i\in\mathcal{H}_\ell}(1-\mathbb{E})m_\ell^{-1}\boldsymbol{v}^\top W_\ell\left[U_{h_{T+1},i}\left(\widehat{\boldsymbol{\beta}}^{(T)}\right)-U_{h_{T+1},i}\left(\boldsymbol{\beta}^*\right)\right]\right)\\
&\quad+\left(h_{T+1}^{-1}\sum_{\ell=1}^L m_\ell^{-1}\boldsymbol{v}^\top W_\ell V_{s,\ell}W_\ell^\top\boldsymbol{v}\right)^{-1/2}\mathbb{E}\left(\sum_{\ell=1}^L\sum_{i\in\mathcal{H}_\ell} m_\ell^{-1}\boldsymbol{v}^\top W_\ell U_{h_{T+1},i}\left(\widehat{\boldsymbol{\beta}}^{(T)}\right)\right)-\boldsymbol{v}^\top\overline{U}_W h_{T+1}^\alpha\\
&=\left(h_{T+1}^{-1}\sum_{\ell=1}^L m_\ell^{-1}\boldsymbol{v}^\top W_\ell V_{s,\ell}W_\ell^\top\boldsymbol{v}\right)^{-1/2}\left(\sum_{\ell=1}^L\sum_{i\in\mathcal{H}_\ell}(1-\mathbb{E})m_\ell^{-1}\boldsymbol{v}^\top W_\ell U_{h_{T+1},i}\left(\boldsymbol{\beta}^*\right)\right)+o_{\mathbb{P}}(1)\xrightarrow{d}\mathcal{N}(0,1).
\end{aligned}$$

Note that $h_{T+1}^{-1}\sum_{\ell=1}^L m_\ell^{-1}W_\ell V_{s,\ell}W_\ell^\top = n^{-\frac{2\alpha}{2\alpha+1}}\sum_{\ell=1}^L\frac{n}{m_\ell}W_\ell V_{s,\ell}W_\ell^\top$, $n^{\alpha/(2\alpha+1)} = h_{T+1}^\alpha$, and Assumption 9 ensures that $n\sum_{\ell=1}^L m_\ell^{-1}W_\ell V_{s,\ell}W_\ell^\top$ is a finite matrix. Then we rewrite the above result as

$$\frac{n^{\alpha/(2\alpha+1)}\sum_{\ell=1}^L\sum_{i\in\mathcal{H}_\ell} m_\ell^{-1}\boldsymbol{v}^\top W_\ell U_{h_{T+1},i}\left(\widehat{\boldsymbol{\beta}}^{(T)}\right)-\boldsymbol{v}^\top\overline{U}_W}{\sqrt{n\sum_{\ell=1}^L m_\ell^{-1}\boldsymbol{v}^\top W_\ell V_{s,\ell}W_\ell^\top\boldsymbol{v}}}\xrightarrow{d}\mathcal{N}(0,1).$$

Combining with

$$\widehat{\boldsymbol{\beta}}_{(\texttt{wmSMSE})}^{(T+1)}-\boldsymbol{\beta}^* = \left(\sum_{\ell=1}^L\sum_{i\in\mathcal{H}_\ell} m_\ell^{-1}W_\ell V_{h_{T+1},i}\left(\widehat{\boldsymbol{\beta}}^{(T)}\right)\right)^{-1}\left(\sum_{\ell=1}^L\sum_{i\in\mathcal{H}_\ell} m_\ell^{-1}W_\ell U_{h_{T+1},i}\left(\widehat{\boldsymbol{\beta}}^{(T)}\right)\right),$$

and

$$\left\|\sum_{\ell=1}^L\sum_{i\in\mathcal{H}_\ell} m_\ell^{-1}W_\ell V_{h_{T+1},i}\left(\widehat{\boldsymbol{\beta}}^{(T)}\right)-\overline{V}_W\right\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{p\log n}{nh_{T+1}^3}}+\delta_T+h_{T+1}^\alpha\right),$$

by Slutsky's Theorem, (20) is true.

□

B.4.1. *Proof for Theorem 4.3.* We first restate Theorem 4.3 to explicitly show the dependence of the convergence rate on the constant $\omega \in (0,1)$.

THEOREM B.3 (Restatement of Theorem 4.3). *Assume the assumptions in Theorem 3.3 hold. Further assume that $\varepsilon \le \varepsilon_0$ for some constant $\varepsilon_0 < 1$. By choosing $h_0 = \left(\frac{p \log L}{\omega m}\right)^{\frac{1}{2\alpha+1}}$ and $h_t = \max\left\{\delta_{m,0}^{2^t/\alpha}, \left(\frac{p}{(1-\omega)n}\right)^{\frac{1}{2\alpha+1}}\right\}$ for $t = 1, 2, \ldots, T$, we have that*

(78)
$$\left\|\widehat{\boldsymbol{\beta}}_1^{(t)} - \boldsymbol{\beta}_1^*\right\|_2 = O_{\mathbb{P}}\left(\left(\frac{p}{(1-\varepsilon)(1-\omega)n}\right)^{\frac{\alpha}{2\alpha+1}} + \delta_{m,0}^{2^t} + \delta_{m,0}^{2^t-1}\left(\frac{p}{(1-\varepsilon)(1-\omega)n}\right)^{\frac{\alpha-1}{2\alpha+1}}\sqrt{\log n} + \varepsilon\delta_{m,0}^2\right),$$

where $\delta_{m,0} = \left(\frac{p \log L}{\omega m}\right)^{\frac{\alpha}{2\alpha+1}}$.

PROOF. Throughout the whole proof, let $|\mathcal{S}|$ denote the cardinality for any set $\mathcal{S}$. For $\ell = 1, \ldots, L$, denote $\mathcal{H}_\ell^{(0)}$ to be an arbitrary subset of $\mathcal{H}_\ell$ with size $\left|\mathcal{H}_\ell^{(0)}\right| = \omega\left|\mathcal{H}_\ell\right| = \omega m$, and let $\widehat{\boldsymbol{\beta}}_{\ell,\mathrm{SMSE}}^{(0)} = \operatorname{argmin}_{\boldsymbol{\beta}} \frac{1}{\omega m}\sum_{\ell \in \mathcal{H}_\ell^{(0)}}(-y_i)H\left(\frac{x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}}{h_0}\right)$, i.e., the SMSE computed using data in $\mathcal{H}_\ell^{(0)}$ with bandwidth $h_0 = (p \log L/\omega m)^{1/(2\alpha+1)}$. For a constant C_0, define

$$\mathcal{A}_{C_0} := \left\{\ell : \left\|\widehat{\boldsymbol{\beta}}_{\ell,\mathrm{SMSE}}^{(0)} - \widehat{\boldsymbol{\beta}}_{1,\mathrm{SMSE}}^{(0)}\right\|_2 \le C_0\delta_{m,0}\right\}, \mathcal{A}^* := \left\{\ell : \boldsymbol{\beta}_\ell^* = \boldsymbol{\beta}_1^*\right\},$$

where $\delta_{m,0} = (p \log L/\omega m)^{\alpha/(2\alpha+1)}$. We first show the following lemma:

LEMMA B.4. *Under the assumptions in Theorem 4.3, there exist constants C and C_0 such that the following event holds with probability approaching one:*

$$E_0 := \left\{\sup_\ell \left\|\widehat{\boldsymbol{\beta}}_{\ell,\mathrm{SMSE}}^{(0)} - \boldsymbol{\beta}_\ell^*\right\|_2 \le C\delta_{m,0}, \quad \mathcal{A}^* \subset \mathcal{A}_{C_0}, \quad \sup_{\ell \in \mathcal{A}_{C_0}\setminus\mathcal{A}^*} \|\boldsymbol{\beta}_\ell^* - \boldsymbol{\beta}_1^*\|_2 \le C\delta_{m,0}\right\}.$$

PROOF FOR LEMMA B.4. We first modify the definitions in the proof of Theorem 3.1 and Proposition 3.2. For $i \in \mathcal{H}_\ell$, define

$$U_{h,i,\ell}(\boldsymbol{\beta}) := \frac{1}{h^2}(-y_i)H''\left(\frac{x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}}{h}\right)\boldsymbol{z}_i\boldsymbol{z}_i^\top(\boldsymbol{\beta} - \boldsymbol{\beta}_\ell^*) - \frac{1}{h}(-y_i)H'\left(\frac{x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}}{h}\right)\boldsymbol{z}_i,$$

and

$$V_{h,i}(\boldsymbol{\beta}) := \frac{1}{h^2}(-y_i)H''\left(\frac{x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}}{h}\right)\boldsymbol{z}_i\boldsymbol{z}_i^\top.$$

Define

$$U_{m,\ell,h}(\boldsymbol{\beta}) = \frac{1}{\omega m}\sum_{i \in \mathcal{H}_\ell^{(0)}} U_{h,i,\ell}(\boldsymbol{\beta}), \quad V_{m,\ell,h}(\boldsymbol{\beta}) := \frac{1}{\omega m}\sum_{i \in \mathcal{H}_\ell^{(0)}} V_{h,i}(\boldsymbol{\beta}).$$

We will show that, for any sufficiently large γ, there exists a constant C, such that

(79)
$$\sup_\ell \|(1 - \mathbb{E})U_{m,\ell,h_0}(\boldsymbol{\beta}_\ell^*)\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{p \log L}{\omega m h_0}}\right),$$

(80)
$$\sup_\ell \sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta} - \boldsymbol{\beta}_\ell^*\|_2 \le \delta} \|V_{m,\ell,h_0}(\boldsymbol{\beta}) - V\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{p \log m}{\omega m h_0^3}} + h_0^\alpha + \delta\right).$$

To show (79), using the same vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{5^p} \in \mathbb{R}^p$ as in the proof of Proposition 3.2, it suffices to show that, for any sufficiently large γ, there exists a constant C^*, such that

(81)
$$\sup_\ell \sup_{j_v \in [5^p]} \left|\boldsymbol{v}_{j_v}^\top (1 - \mathbb{E})U_{m,\ell,h_0}(\boldsymbol{\beta}_\ell^*)\right| \le C^*\sqrt{\frac{p \log L}{\omega m h_0}}.$$

with probability at least $1-2L(5L^{-\gamma})^p$.

Let $\boldsymbol{v}$ be any fixed vector in $\mathbb{S}^{p-1}$. Note that

$$(1-\mathbb{E})\boldsymbol{v}^\top U_{m,\ell,h_0}(\boldsymbol{\beta}_\ell^*) = \frac{1}{\omega m}\sum_{i\in\mathcal{H}_\ell^{(0)}}(1-\mathbb{E})\left[\boldsymbol{v}^\top U_{h_0,i,\ell}(\boldsymbol{\beta}_\ell^*)\right] = \frac{1}{\omega m}\sum_{i\in\mathcal{H}_\ell^{(0)}}(1-\mathbb{E})\frac{(\boldsymbol{v}^\top \boldsymbol{z}_i)y_i}{h}H'\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}_\ell^*}{h}\right),$$

and $(1-\mathbb{E})\left[\boldsymbol{v}^\top U_{h,i}(\boldsymbol{\beta}_\ell^*)\right]$ are i.i.d. among different $i\in\mathcal{H}_\ell^{(0)}$. By the assumptions, the density function of $\zeta_\ell = X+\boldsymbol{Z}^\top\boldsymbol{\beta}_\ell^*$, denoted by $\rho(\cdot\,|\,\boldsymbol{Z})$, is the same for all ℓ and is bounded uniformly for all $\boldsymbol{Z}$. Therefore,

$$\begin{aligned}
&\mathbb{E}_{\cdot\,|\,\boldsymbol{Z}}\left[\boldsymbol{v}^\top U_{h_0,i,\ell}(\boldsymbol{\beta}_\ell^*)\right]^2 \\
=&\frac{(\boldsymbol{v}^\top\boldsymbol{Z})^2}{h_0^2}\mathbb{E}_{\cdot\,|\,\boldsymbol{Z}}\left[H'\left(\frac{X+\boldsymbol{Z}^\top\boldsymbol{\beta}_\ell^*}{h_0}\right)\right]^2 \qquad (82)\\
=&\int_{-1}^{1}\frac{(\boldsymbol{v}^\top\boldsymbol{Z})^2}{h_0}\left[H'(\xi)\right]^2\rho(\xi h_0\,|\,\boldsymbol{Z})\,\mathrm{d}\xi \le C_{\rho,H}\left(\frac{(\boldsymbol{v}^\top\boldsymbol{Z})^2}{h_0}\right),
\end{aligned}$$

where $C_{\rho,H}$ is a constant not depending on ℓ. Let $\widetilde{\phi}_{i,\boldsymbol{v}}^{U,*} = h(1-\mathbb{E})\left[\boldsymbol{v}^\top U_{h_0,i,\ell}(\boldsymbol{\beta}_\ell^*)\right]$. By (51), it holds that

$$\mathbb{P}\left(\sum_{i\in\mathcal{H}_\ell^{(0)}}\widetilde{\phi}_{i,\boldsymbol{v}}^{U,*}>b\right)\le\exp\left(-tb+t^2\sum_{i\in\mathcal{H}_\ell^{(0)}}\mathbb{E}\left[\left(\widetilde{\phi}_{i,\boldsymbol{v}}^{U,*}\right)^2\exp\left(t\left|\widetilde{\phi}_{i,\boldsymbol{v}}^{U,*}\right|\right)\right]\right).$$

Combining with (82) and Assumption 5, there exists a constant $C_{\rho,H,\eta}$ that does not depend on ℓ, such that

$$\mathbb{P}\left(\sum_{i\in\mathcal{H}_\ell^{(0)}}\widetilde{\phi}_{i,\boldsymbol{v}}^{U,*}>b\right)\le\exp\left(-tb+\omega mt^2C_{\rho,H,\eta}^2h_0\right),$$

for sufficiently small t. Letting $b=C_{\rho,H,\eta}\sqrt{\gamma\omega mph_0\log L}$ and $t=\sqrt{\frac{\gamma p\log L}{4C_{\rho,H,\eta}^2\omega mh_0}}$ leads to

$$\mathbb{P}\left(\sum_{i\in\mathcal{H}_\ell^{(0)}}\widetilde{\phi}_{i,\boldsymbol{v}}^{U,*}>C_{\rho,H,\eta}\sqrt{\gamma mh_0p\log L}\right)\le\exp\left[-\frac{1}{4}\gamma p\log L\right],$$

which yields

$$\mathbb{P}\left(\frac{1}{\omega m}\left|\sum_{i\in\mathcal{H}_\ell^{(0)}}(1-\mathbb{E})\left[\boldsymbol{v}^\top U_{h_0,i,\ell}(\boldsymbol{\beta}_\ell^*)\right]\right|>C_{\rho,H,\eta}\sqrt{\frac{\gamma p\log L}{\omega mh_0}}\right)\le 2L^{-\gamma p/4}.$$

The above inequality is true for all $\boldsymbol{v}$ in $\mathbb{S}^{p-1}$. Therefore, with probability at least $1-2L(5L^{-\gamma/4})^p$,

$$\sup_\ell\sup_{j_v\in[5^p]}\left|(1-\mathbb{E})\boldsymbol{v}_{j_v}^\top U_{m,\ell,h_0}(\boldsymbol{\beta}^*)\right|\le C^*\sqrt{\frac{p\log L}{\omega mh_0}},$$

where $C^*=C_{\rho,H,\eta}\sqrt{\gamma}$. Since γ can be arbitrarily large, this completes the proof of (79).

To show (80), we use the same set of $\{\boldsymbol{v}_{j_v}\}_{j_v\in[5^p]}$ as above. By the proof of Lemma 3 in Cai et al. (2010), $\|A\|_2\le 4\sup_{j_v\in[5^p]}\left|\boldsymbol{v}_{j_v}^\top A\boldsymbol{v}_{j_v}\right|$, for any symmetric $A\in\mathbb{R}^{p\times p}$. Therefore,

it suffices to bound $\sup_{j_v\in[5^p]}\left|\boldsymbol{v}_{j_v}^\top\left[V_{m,\ell,h_0}\left(\boldsymbol{\beta}\right)-V\right]\boldsymbol{v}_{j_v}\right|$. Using the same strategy as in **Step 1** of the proof of (40), we can find a set $\{\boldsymbol{\beta}_{j_\beta}\}_{j_\beta\in[(\omega m)^{\gamma p}]}$ satisfying that, for all $\boldsymbol{\beta}$ in the ball $\{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\le\delta\}$, there exists $j_\beta\in[(\omega m)^{\gamma p}]$ such that $\left\|\boldsymbol{\beta}-\boldsymbol{\beta}_{j_\beta}\right\|_\infty\le\frac{2\delta}{(\omega m)^\gamma}$. By the Lipschitz property of $H''(x)$, we have

$$\begin{aligned}
&\sup_{j_v\in[5^p]}\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\le\delta}\inf_{j_\beta\in[(\omega m)^{\gamma p}]}\left|\boldsymbol{v}_{j_v}^\top V_{m,\ell,h_0}\left(\boldsymbol{\beta}\right)\boldsymbol{v}_{j_v}-\boldsymbol{v}_{j_v}^\top V_{m,\ell,h_0}\left(\boldsymbol{\beta}_{j_\beta}\right)\boldsymbol{v}_{j_v}\right|\\
\le&\sup_{j_v\in[5^p]}\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\le\delta}\inf_{j_\beta\in[(\omega m)^{\gamma p}]}\frac{1}{m}\sum_{i\in\mathcal{H}_\ell^{(0)}}\frac{\left(\boldsymbol{z}_i^\top\boldsymbol{v}_{j_v}\right)^2}{h_0^3}\left|\boldsymbol{z}_i^\top\left(\boldsymbol{\beta}-\boldsymbol{\beta}_{j_\beta}\right)\right|\\
=&O\left(\sum_{i\in\mathcal{H}_\ell^{(0)}}\frac{\delta\left\|\boldsymbol{z}_i\right\|_2^3}{(\omega m)^{\gamma+1}h_0^3}\right).
\end{aligned}$$

Analogous to the **Step 1** of the proof of (40), we can show that $\frac{1}{(\omega m)^{\gamma+1}}\sum_{i\in\mathcal{H}_\ell^{(0)}}\left\|\boldsymbol{z}_i\right\|_2^3\le\frac{p\sup_{i,j}\mathbb{E}|z_{i,j}|^3}{(\omega m)^{(\gamma-1)/2}}$, with probability and least $1-(\omega m)^{-\gamma/2}$. By taking $\gamma>0$ large enough such that

$$\frac{p^{3/2}\delta}{(\omega m)^{(\gamma-1)/2}h_0^3}=o\left(\sqrt{\frac{p\log m}{\omega m h_0^3}}\right),$$

we obtain

(83)

$$\begin{aligned}
&\sup_{j_v\in[5^p]}\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\le\delta}\left|\boldsymbol{v}_{j_v}^\top\left[V_{m,\ell,h_0}\left(\boldsymbol{\beta}\right)-V\right]\boldsymbol{v}_{j_v}\right|-\sup_{j_v\in[5^p]}\sup_{j_\beta\in[m^{\gamma p}]}\left|\boldsymbol{v}_{j_v}^\top\left[V_{m,\ell,h_0}\left(\boldsymbol{\beta}_{j_\beta}\right)-V\right]\boldsymbol{v}_{j_v}\right|\\
=&o\left(\sqrt{\frac{p\log m}{\omega m h_0^3}}\right).
\end{aligned}$$

For any fixed $\boldsymbol{\beta}$, let

$$\phi_{i,\boldsymbol{v}}^V(\boldsymbol{\beta}):=h_0^2(1-\mathbb{E})\boldsymbol{v}^\top V_{h_0,i}(\boldsymbol{\beta})\boldsymbol{v}=(1-\mathbb{E})(\boldsymbol{v}^\top\boldsymbol{z}_i)^2\left(-y_i\right)H''\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}}{h_0}\right).$$

By repeating the procedure in the proof for (79), we obtain

$$\mathbb{P}\left(\sum_{i\in\mathcal{H}_\ell^{(0)}}\phi_{i,\boldsymbol{v}}^V(\boldsymbol{\beta})>b\right)\le\exp\left(-tb+t^2C_{\rho,H,\eta}^2\omega m h_0\right),$$

for some absolute constant $C_{\rho,H,\eta}$ and sufficiently small $t>0$. Let $b=C_{\rho,H,\eta}\sqrt{8\gamma\omega m h_0 p\log m}$ and $t=b/(2C_{\rho,H,\eta}^2\omega m h_0)$, where γ is an arbitrary positive constant to be determined, it holds that

$$\mathbb{P}\left(\phi_{i,\boldsymbol{v}}^V(\boldsymbol{\beta})>C_{\rho,H,\eta}\sqrt{8\gamma\omega m h_0 p\log m}\right)\le m^{-2\gamma p},$$

and hence

$$\mathbb{P}\left(\left|(1-\mathbb{E})\left(\boldsymbol{v}^\top\left[V_{m,\ell,h_0}\left(\boldsymbol{\beta}\right)-V\right]\boldsymbol{v}\right)\right|>C_{\rho,H,\eta}\sqrt{\frac{8\gamma p\log m}{\omega m h_0^3}}\right)\le 2(\omega m)^{-2\gamma p}.$$

The above inequality is true for all $\boldsymbol{v}_{j_v}$ and $\boldsymbol{\beta}_{j_\beta}$, and thus

$$\mathbb{P}\left(\sup_{j_v\in[5^p]}\sup_{j_\beta\in[(\omega m)^{\gamma p}]}\left|(1-\mathbb{E})\left(\boldsymbol{v}_{j_v}^\top\left[V_{m,\ell,h_0}\left(\boldsymbol{\beta}_{j_\beta}\right)-V\right]\boldsymbol{v}_{j_v}\right)\right|>C_{\rho,H,\eta}\sqrt{\frac{8\gamma p\log m}{\omega m h_0^3}}\right)\leq 2(5(\omega m)^{-\gamma})^p.$$

Then we obtain, for any sufficiently large $\gamma>0$, there exists a constant C_V, such that

$$\sup_\ell\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\leq\delta}\left|(1-\mathbb{E})\left(\boldsymbol{v}^\top\left[V_{m,\ell,h_0}\left(\boldsymbol{\beta}\right)-V\right]\boldsymbol{v}\right)\right|\leq C_V\sqrt{\frac{p\log m}{\omega m h_0^3}},$$

with probability at least $1-L(\omega m)^{-\gamma/2}-2L(5(\omega m)^{-\gamma})^p$. By the assumption that $m>n^{c_3}$ for some $0<c_3<1$, there exists a constant γ such that $L(\omega m)^{-\gamma/2}+2L(5(\omega m)^{-\gamma})^p=o(1)$. Therefore, it holds that

$$\sup_\ell\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\leq\delta}\left|(1-\mathbb{E})\left(\boldsymbol{v}^\top\left[V_{m,\ell,h_0}\left(\boldsymbol{\beta}\right)-V\right]\boldsymbol{v}\right)\right|=O_{\mathbb{P}}\left(\sqrt{\frac{p\log m}{\omega m h_0^3}}\right).$$

By the proof of (41), we have

$$\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\leq\delta}\left|\mathbb{E}\left(\boldsymbol{v}^\top\left[V_{m,\ell,h}\left(\boldsymbol{\beta}\right)-V\right]\boldsymbol{v}\right)\right|=O\left(\delta+h^\alpha\right),\tag{84}$$

where the constant hidden in $O(\cdot)$ is the same for all ℓ. This completes the proof for (80).

By **Step 3** in the proof of (40), we have $\sup_\ell\|\mathbb{E}[U_{m,\ell,h_0}(\boldsymbol{\beta}_\ell^*)]\|_2=O(h_0^\alpha)$. Then the proof of Theorem 3.1 indicates that Equations (79) and (80) imply

$$\sup_\ell\left\|\widehat{\boldsymbol{\beta}}_{\ell,\mathrm{SMSE}}^{(0)}-\boldsymbol{\beta}_\ell^*\right\|_2\leq C'\left(\sqrt{\frac{p\log L}{\omega m h_0}}+h_0^\alpha\right)=2C'\left(\frac{p\log L}{\omega m}\right)^{\alpha/(2\alpha+1)}=2C'\delta_{m,0},\tag{85}$$

for some absolute constant C', with probability approaching one. Furthermore, the definition of $\mathcal{A}_{C_0}$ with $C_0>4C'$ ensures that $\mathcal{A}^*\subset\mathcal{A}$ if (85) holds. Moreover, for any $\ell\in\mathcal{A}\setminus\mathcal{A}^*$, we have that $\|\boldsymbol{\beta}_\ell^*-\boldsymbol{\beta}_1^*\|_2\leq 2\sup_\ell\left\|\widehat{\boldsymbol{\beta}}_{\ell,\mathrm{SMSE}}^{(0)}-\boldsymbol{\beta}_\ell^*\right\|_2+\sup_\ell\left\|\widehat{\boldsymbol{\beta}}_{\ell,\mathrm{SMSE}}^{(0)}-\widehat{\boldsymbol{\beta}}_{1,\mathrm{SMSE}}^{(0)}\right\|_2\leq(C_0+4C')\delta_{m,0}$ if (85) holds. Therefore, the event E_0 holds with probability approaching one. □

Let $\widehat{\boldsymbol{\beta}}_1^{(0)}=\widehat{\boldsymbol{\beta}}_{1,\mathrm{SMSE}}^{(0)}$. Without loss of generality, lat $C=C_0=1$ in E_0. For simplicity, denote $\mathcal{A}=\mathcal{A}_{C_0}$. Then under E_0, it holds that

$$\sup_{\ell\in\mathcal{A}\setminus\mathcal{A}^*}\left\|\widehat{\boldsymbol{\beta}}_1^{(0)}-\boldsymbol{\beta}_\ell^*\right\|_2\leq\sup_{\ell\in\mathcal{A}\setminus\mathcal{A}^*}\left\|\widehat{\boldsymbol{\beta}}_{1,\mathrm{SMSE}}^{(0)}-\widehat{\boldsymbol{\beta}}_{\ell,\mathrm{SMSE}}^{(0)}\right\|_2+\sup_\ell\left\|\widehat{\boldsymbol{\beta}}_{\ell,\mathrm{SMSE}}^{(0)}-\boldsymbol{\beta}_\ell^*\right\|_2\leq 2\delta_{m,0},\tag{86}$$

and

$$\sup_{\ell\in\mathcal{A}^*}\left\|\widehat{\boldsymbol{\beta}}_1^{(0)}-\boldsymbol{\beta}_\ell^*\right\|_2=\left\|\widehat{\boldsymbol{\beta}}_{1,\mathrm{SMSE}}^{(0)}-\boldsymbol{\beta}_1^*\right\|_2\leq\delta_{m,0}.\tag{87}$$

Define $\widetilde{\mathcal{H}}_\ell=\mathcal{H}_\ell\setminus\mathcal{H}_\ell^{(0)}$. Recall that the algorithm is

$$\widehat{\boldsymbol{\beta}}_1^{(t+1)}=\widehat{\boldsymbol{\beta}}_1^{(t)}-\left[\frac{1}{(1-\omega)m|\mathcal{A}|}\sum_{\ell\in\mathcal{A}}\sum_{i\in\widetilde{\mathcal{H}}_\ell}V_{h,i}\left(\widehat{\boldsymbol{\beta}}_1^{(t)}\right)\right]^{-1}\left[\frac{1}{|\mathcal{A}|(1-\omega)mh}\sum_{\ell\in\mathcal{A}}\sum_{i\in\widetilde{\mathcal{H}}_\ell}(-y_i)\,H'\left(\frac{x_i+\boldsymbol{z}_i^\top\widehat{\boldsymbol{\beta}}_1^{(t)}}{h}\right)\boldsymbol{z}_i\right],$$

which implies that

$$\widehat{\boldsymbol{\beta}}_1^{(t+1)}-\boldsymbol{\beta}_1^*=\left[\widetilde{V}_{\mathcal{A}}\left(\widehat{\boldsymbol{\beta}}_1^{(t)}\right)\right]^{-1}\left[\frac{1}{|\mathcal{A}|}\sum_{\ell\in\mathcal{A}}\widetilde{U}_{m,\ell,h}\left(\widehat{\boldsymbol{\beta}}_1^{(t)}\right)\right],$$

where $|\mathcal{A}|$ denote the cardinality of $\mathcal{A}$,

$$\widetilde{V}_{\mathcal{A}}(\boldsymbol{\beta}) := \frac{1}{(1-\omega)m|\mathcal{A}|}\sum_{\ell\in\mathcal{A}}\sum_{i\in\widetilde{\mathcal{H}}_\ell} V_{h,i}\left(\boldsymbol{\beta}\right),$$

and

$$\begin{aligned}\widetilde{U}_{m,\ell,h}(\boldsymbol{\beta}) &:= \frac{1}{(1-\omega)m}\sum_{i\in\widetilde{\mathcal{H}}_\ell} V_{h,i}\left(\boldsymbol{\beta}\right)\left(\boldsymbol{\beta}-\boldsymbol{\beta}_1^*\right) - \frac{1}{(1-\omega)mh}\sum_{i\in\widetilde{\mathcal{H}}_\ell}\left(-y_i\right)H'\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}}{h}\right)\boldsymbol{z}_i\\ &= \frac{1}{(1-\omega)mh^2}\sum_{i\in\widetilde{\mathcal{H}}_\ell}\left(-y_i\right)H''\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}}{h}\right)\boldsymbol{z}_i\boldsymbol{z}_i^\top\left(\boldsymbol{\beta}-\boldsymbol{\beta}_1^*\right)\\ &\quad - \frac{1}{(1-\omega)mh}\sum_{i\in\widetilde{\mathcal{H}}_\ell}\left(-y_i\right)H'\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}}{h}\right)\boldsymbol{z}_i.\end{aligned}$$

Let δ be a quantity that satisfies $\delta = o(\delta_{m,0})$. Under E_0, it holds that $\|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_\ell^*\|_2 \le \delta_{m,0}$ for all $\ell \in \mathcal{A}\setminus\mathcal{A}^*$. Also, the data points in $\{\widetilde{\mathcal{H}}_\ell\}_{\ell=1}^L$ are independent to E_1. Following the proof for (80), it is straightforward to show that, when $h = o(1)$ and $p\log n/(nh^3) = o(1)$, with probability tending to one,

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_1^*\|_2\le\delta}\left\|\widetilde{V}_{\mathcal{A}}(\boldsymbol{\beta}) - V\right\|_2 \lesssim \sqrt{\frac{p\log n}{(1-\varepsilon)(1-\omega)nh^3}} + h^\alpha + \delta + \varepsilon\delta_{m,0},$$

using the facts that $|\mathcal{A}| \ge (1-\varepsilon)L$, $|\mathcal{A}\setminus\mathcal{A}^*| \le \varepsilon L$.

For the numerator, we have

$$\begin{aligned}\frac{1}{|\mathcal{A}|}\sum_{\ell\in\mathcal{A}}\widetilde{U}_{m,\ell,h}\big(\widehat{\boldsymbol{\beta}}_1^{(t)}\big) &= \frac{1}{|\mathcal{A}|}\sum_{\ell\in\mathcal{A}}(1-\mathbb{E})\widetilde{U}_{m,\ell,h}(\boldsymbol{\beta}_1^*) + \frac{1}{|\mathcal{A}|}\sum_{\ell\in\mathcal{A}}(1-\mathbb{E})\left[\widetilde{U}_{m,\ell,h}\big(\widehat{\boldsymbol{\beta}}_1^{(t)}\big) - \widetilde{U}_{m,\ell,h}(\boldsymbol{\beta}_1^*)\right]\\ &\quad + \frac{1}{|\mathcal{A}|}\sum_{\ell\in\mathcal{A}}\mathbb{E}\left[\widetilde{U}_{m,\ell,h}\big(\widehat{\boldsymbol{\beta}}_1^{(t)}\big)\right].\end{aligned}$$

Note that by the uniformly boundedness of $\rho(\cdot\,|\,\boldsymbol{Z})$, on each machine ℓ, Equations (49), (50), and (53) hold for all $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_1^*$ such that $\|\boldsymbol{\beta}-\boldsymbol{\beta}_1^*\|_2 \le \delta$ and $\|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_\ell^*\|_2 \le \delta_{m,0}$. Moreover, the constants hidden in the big O notations are uniform for all ℓ, $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_1^*$. Therefore, following the **Step 1** and **Step 2** in the proof of Proposition 3.2, we have that

$$\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}_1^*\|_2\le\delta}\left\|\frac{1}{|\mathcal{A}|}\sum_{\ell\in\mathcal{A}}(1-\mathbb{E})\left[\widetilde{U}_{m,\ell,h}(\boldsymbol{\beta}) - \widetilde{U}_{m,\ell,h}(\boldsymbol{\beta}_1^*)\right]\right\|_2 = O_{\mathbb{P}}\left(\delta\sqrt{\frac{p\log n}{(1-\varepsilon)(1-\omega)nh^3}}\right),$$

and

$$\left\|\frac{1}{|\mathcal{A}|}\sum_{\ell\in\mathcal{A}}(1-\mathbb{E})\widetilde{U}_{m,\ell,h}(\boldsymbol{\beta}_1^*)\right\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{p}{(1-\varepsilon)(1-\omega)nh}}\right).$$

By **Step 3** in the proof of Proposition 3.2, we have

$$\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}_1^*\|_2\le\delta}\frac{1}{|\mathcal{A}|}\sum_{\ell\in\mathcal{A}}\left\|\mathbb{E}\left[\widetilde{U}_{m,\ell,h}\big(\boldsymbol{\beta}\big)\right]\right\|_2 = O\left(h^\alpha + \delta^2 + \varepsilon\delta_{m,0}^2\right),$$

where again we use the facts that $|\mathcal{A} \setminus \mathcal{A}^*| \le \varepsilon L$ and $\|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_\ell^*\|_2 \le \delta_{m,0}$ for all $\ell \in \mathcal{A} \setminus \mathcal{A}^*$ under E_0. The above bounds imply that

$$\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*\|_2\le\delta} \left\| \frac{1}{|\mathcal{A}|} \sum_{\ell\in\mathcal{A}} \widetilde{U}_{m,\ell,h}(\boldsymbol{\beta}) \right\|_2$$
$$= O_{\mathbb{P}}\left(\sqrt{\frac{p}{(1-\varepsilon)(1-\omega)nh}} + \delta\sqrt{\frac{p\log n}{(1-\varepsilon)(1-\omega)nh^3}} + \delta^2 + h^\alpha + \varepsilon\delta_{m,0}^2 \right),$$

where $\delta_{m,0} = \left(\frac{p\log L}{\omega m}\right)^{\alpha/(2\alpha+1)}$. Adding that $\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_1^*\|_2\le\delta} \left\|\widetilde{V}_{\mathcal{A}}(\boldsymbol{\beta}) - V\right\|_2 = o_{\mathbb{P}}(1)$, we obtain that

(88)
$$\left\|\widehat{\boldsymbol{\beta}}_1^{(t+1)} - \boldsymbol{\beta}_1^*\right\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{p}{(1-\varepsilon)(1-\omega)nh_{t+1}}} + \delta_{m,t}\sqrt{\frac{p\log n}{(1-\varepsilon)(1-\omega)nh_{t+1}^3}} + \delta_{m,t}^2 + h_{t+1}^\alpha + \varepsilon\delta_{m,0}^2 \right),$$

if $\left\|\widehat{\boldsymbol{\beta}}_1^{(t)} - \boldsymbol{\beta}_1\right\|_2 = O_{\mathbb{P}}(\delta_{m,t})$. Since ω is a constant, and $\varepsilon \le \varepsilon_0$ for some constant $\varepsilon_0 < 1$, we omit the factor $(1-\varepsilon)(1-\omega)$ on n in the sequel.

Now we choose

$$h_t = \max\left\{\delta_{m,0}^{2^t/\alpha}, (p/n)^{1/(2\alpha+1)}\right\} \ge (p/n)^{1/(2\alpha+1)},$$

which implies the $\frac{p\log n}{nh_t^3} = o(1)$ holds for any t, since

$$\frac{p\log n}{nh_t^3} \le (p/n)^{\frac{2\alpha-2}{2\alpha+1}}\log n = O\big(n^{-(1-c_2)\frac{2\alpha-2}{2\alpha+1}}\log n\big) \to 0,$$

for $p = O(m^{c_2}) = O(n^{c_2})$ and $\alpha \ge 2$. Moreover, for any t, $h_t^\alpha = \max\left\{\delta_{m,0}^{2^t}, (p/n)^{\alpha/(2\alpha+1)}\right\}$ and $\sqrt{p/nh_t} \le (p/n)^{\alpha/(2\alpha+1)}$.

We now show that, for any t,

(89)
$$\left\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\right\|_2 = O_{\mathbb{P}}\left((p/n)^{\alpha/(2\alpha+1)}\sqrt{\log n} + \delta_{m,0}^{2^t} + \varepsilon\delta_{m,0}^2\right).$$

When $t = 1$, $\delta_{m,0} \le h_1 = \delta_{m,0}^{\frac{2}{\alpha}}$. Then (89) holds since $\delta_{m,0}\sqrt{\frac{p\log n}{nh_1^3}} = O\left(\sqrt{\frac{p\log n}{nh_1}}\right)$ and

$$\left\|\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*\right\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{p\log n}{nh_1}} + \delta_{m,0}^2 + h_1^\alpha + \varepsilon\delta_{m,0}^2\right) = O_{\mathbb{P}}\left(\left(\frac{p}{n}\right)^{\frac{\alpha}{2\alpha+1}}\sqrt{\log n} + \delta_{m,0}^2 + \varepsilon\delta_{m,0}^2\right).$$

Assume (89) holds for $t-1$, i.e.,

$$\left\|\widehat{\boldsymbol{\beta}}^{(t-1)} - \boldsymbol{\beta}^*\right\|_2 = O_{\mathbb{P}}\left((p/n)^{\alpha/(2\alpha+1)}\sqrt{\log n} + \delta_{m,0}^{2^{t-1}} + \varepsilon\delta_{m,0}^2\right).$$

Then $\delta_{m,t-1} = (p/n)^{\alpha/(2\alpha+1)}\sqrt{\log n} + \delta_{m,0}^{2^{t-1}}$. Since $(p/n)^{\alpha/(2\alpha+1)}\sqrt{\log n} \ll (p/n)^{\frac{1}{2\alpha+1}} \le h_t$ and $\delta_{m,0}^{2^{t-1}} \le \delta_{m,0}^{2^t/\alpha} \le h_t$, by (88),

$$\left\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\right\|_2 = O_{\mathbb{P}}\left(\left(\frac{p}{n}\right)^{\frac{\alpha}{2\alpha+1}}\sqrt{\log n} + \delta_{m,0}^{2^t} + \varepsilon\delta_{m,0}^2\right),$$

where we use $h_t^\alpha = \max\left\{\delta_{m,0}^{2^t}, (p/n)^{\alpha/(2\alpha+1)}\right\}$ and $\sqrt{p/nh_t} \le (p/n)^{\alpha/(2\alpha+1)}$. Therefore, we prove that (89) holds for all t by induction.

Plugging $\delta_{m,t-1} = (p/n)^{\alpha/(2\alpha+1)}\sqrt{\log n} + \delta_{m,0}^{2^{t-1}} + \varepsilon\delta_{m,0}^2$ into (88) yields that

$$\begin{aligned}&\left\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\right\|_2 \\ &= O_{\mathbb{P}}\left(\left(\frac{p}{n}\right)^{\frac{2\alpha}{2\alpha+1}}\log n + \delta_{m,0}^{2^t} + h_t^\alpha + \sqrt{\frac{p}{nh_t}} + \left(\left(\frac{p}{n}\right)^{\frac{\alpha}{2\alpha+1}}\sqrt{\log n} + \delta_{m,0}^{2^{t-1}}\right)\sqrt{\frac{p\log n}{nh_t^3}} + \varepsilon\delta_{m,0}^2\right).\end{aligned}$$

Since $\alpha > 1$,

$$\left(\frac{p}{n}\right)^{\frac{\alpha}{2\alpha+1}}\sqrt{\log n}\cdot\sqrt{\frac{p\log n}{nh_t^3}} = \left(\frac{p}{n}\right)^{\frac{\alpha}{2\alpha+1}}(\log n)\cdot\sqrt{\frac{p}{nh_t^3}} \le \left(\frac{p}{n}\right)^{\frac{\alpha}{2\alpha+1}}(\log n)\cdot(p/n)^{\frac{\alpha-1}{2\alpha+1}} \lesssim \left(\frac{p}{n}\right)^{\frac{\alpha}{2\alpha+1}}.$$

Therefore, the rate in (66) is upper bounded by

$$O\left(\left(\frac{p}{n}\right)^{\frac{\alpha}{2\alpha+1}} + \delta_{m,0}^{2^t} + \delta_{m,0}^{2^{t-1}}\left(\frac{p}{n}\right)^{\frac{\alpha-1}{2\alpha+1}}\sqrt{\log n} + \varepsilon\delta_{m,0}^2\right),$$

which leads to that

$$\left\|\widehat{\boldsymbol{\beta}}_1^{(T)} - \boldsymbol{\beta}_1^*\right\|_2 = O_{\mathbb{P}}\left((p/n)^{\alpha/(2\alpha+1)} + \varepsilon\delta_{m,0}^2\right),$$

for sufficiently large T.

□

B.5. Proof of the Results for the High-dimensional `(mSMSE)`**.** Before starting the proof, we first formalize our notation. Define

$$V_n(\boldsymbol{\beta}) = \frac{1}{nh^2}\sum_{i=1}^n (-y_i)\, H''\left(\frac{x_i + \boldsymbol{z}_i^\top\boldsymbol{\beta}}{h}\right)\boldsymbol{z}_i\boldsymbol{z}_i^\top. \tag{90}$$

$$V_{m,\ell}(\boldsymbol{\beta}) = \frac{1}{mh^2}\sum_{i\in\mathcal{H}_\ell} (-y_i)\, H''\left(\frac{x_i + \boldsymbol{z}_i^\top\boldsymbol{\beta}}{h}\right)\boldsymbol{z}_i\boldsymbol{z}_i^\top. \tag{91}$$

$$U_n(\boldsymbol{\beta}) = \frac{1}{nh}\sum_{i=1}^n (-y_i)\, H'\left(\frac{x_i + \boldsymbol{z}_i^\top\boldsymbol{\beta}}{h}\right)\boldsymbol{z}_i. \tag{92}$$

As claimed before, we will not compute $V_n(\boldsymbol{\beta})$ in the algorithm, but it is an important intermediate quantity in the theoretical analysis. Without loss of generality, we only consider $V_{m,1}(\boldsymbol{\beta})$ in the sequel. We first give the convergence property of $V_n\left(\widehat{\boldsymbol{\beta}}^{(0)}\right)$, $V_{m,1}\left(\widehat{\boldsymbol{\beta}}^{(0)}\right)$ and $U_n\left(\widehat{\boldsymbol{\beta}}^{(0)}\right)$, which is crucial for deriving the convergence rate of $\widehat{\boldsymbol{\beta}}^{(1)}$. Note that we omit the dependence of these quantities on the bandwidth h in the notation.

LEMMA B.5. *Assume Assumptions 1–4, 11–13 hold. Further assume that* $\frac{\log m}{mh^3} = o(1)$, $\sqrt{s}\delta_{m,0} = O\left(h^{3/2}\right)$ *($\delta_{m,0}$ is defined in Assumption 11) and* $h = o(1)$*, we have the following results:*

$$\left\|(1-\mathbb{E})\left[V_n\left(\widehat{\boldsymbol{\beta}}^{(0)}\right)\right]\right\|_{\max} = O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{nh^3}}\right), \tag{93}$$

$$\sup_{\|\boldsymbol{v}_1\|_2=\|\boldsymbol{v}_2\|_2=1} \boldsymbol{v}_1^\top \left(\mathbb{E}\left[V_n\left(\widehat{\boldsymbol{\beta}}^{(0)}\right)\right] - V \right) \boldsymbol{v}_2 = O_{\mathbb{P}}\left(\sqrt{s}\delta_{m,0} + h^\alpha\right), \tag{94}$$

$$\left\| (1-\mathbb{E})\left[V_{m,1}\left(\widehat{\boldsymbol{\beta}}^{(0)}\right)\right] \right\|_{\max} = O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{mh^3}}\right), \tag{95}$$

$$\sup_{\|\boldsymbol{v}_1\|_2=\|\boldsymbol{v}_2\|_2=1} \boldsymbol{v}_1^\top \left(\mathbb{E}\left[V_{m,1}\left(\widehat{\boldsymbol{\beta}}^{(0)}\right)\right] - V \right) \boldsymbol{v}_2 = O_{\mathbb{P}}\left(\sqrt{s}\delta_{m,0} + h^\alpha\right). \tag{96}$$

Additionally, define

$$\Psi_n\left(\widehat{\boldsymbol{\beta}}^{(0)}\right) := U_n\left(\widehat{\boldsymbol{\beta}}^{(0)}\right) - V_n\left(\widehat{\boldsymbol{\beta}}^{(0)}\right)\left(\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\right),$$

and then we have

$$\left\| (1-\mathbb{E})\Psi_n\left(\widehat{\boldsymbol{\beta}}^{(0)}\right) \right\|_\infty = O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{nh}}\right), \tag{97}$$

and

$$\sup_{\|\boldsymbol{v}\|_2=1} \left| \boldsymbol{v}^\top \mathbb{E}\left[\Psi_n\left(\widehat{\boldsymbol{\beta}}^{(0)}\right)\right] \right| = O_{\mathbb{P}}\left(s\delta_{m,0}^2 + h^\alpha\right). \tag{98}$$

PROOF. **Proof of (93):**
Recall our definitions

$$V_n\left(\widehat{\boldsymbol{\beta}}^{(0)}\right) = \frac{1}{nh^2}\sum_{i=1}^n (-y_i) H''\left(\frac{x_i + \boldsymbol{z}_i^\top \widehat{\boldsymbol{\beta}}^{(0)}}{h}\right) \boldsymbol{z}_i \boldsymbol{z}_i^\top,$$

and

$$V = -2\mathbb{E}\left[\rho\left(0 \mid \boldsymbol{Z}\right) F'\left(0 \mid \boldsymbol{Z}\right) \boldsymbol{Z}\boldsymbol{Z}^\top\right].$$

By Assumption 11, there exists constant C_{ℓ_1}, C_{ℓ_2} such that $\mathbb{P}\left(\widehat{\boldsymbol{\beta}}^{(0)} \in \Theta\right) \to 1$, where

$$\Theta := \left\{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \le C_{\ell_1}\delta_{m,0}, \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \le C_{\ell_2}\sqrt{s}\delta_{m,0}\right\}.$$

Without loss of generality, we assume $C_{\ell_1} = C_{\ell_2} = 1$ and $\widehat{\boldsymbol{\beta}}^{(0)} \in \Theta$ in the following proof.
For each $(j_1, j_2) \in \{1, .., p\} \times \{1, .., p\}$, define

$$V_{n,j_1,j_2}\left(\boldsymbol{\beta}\right) := \frac{1}{nh^2}\sum_{i=1}^n (-y_i) H''\left(\frac{x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}}{h}\right) z_{i,j_1} z_{i,j_2},$$

$$V_{j_1,j_2} := -2\mathbb{E}\left[\rho\left(0 \mid \boldsymbol{Z}\right) F'\left(0 \mid \boldsymbol{Z}\right) Z_{j_1} Z_{j_2}\right],$$

$$\phi^V_{ij_1j_2}\left(\boldsymbol{\beta}\right) := \frac{-y_i}{h^2} H''\left(\frac{x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}}{h}\right) z_{i,j_1} z_{i,j_2} + \frac{y_i}{h^2} H''\left(\frac{x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}^*}{h}\right) z_{i,j_1} z_{i,j_2},$$

and

$$\Phi^V_{n,j_1,j_2} := \sup_{\boldsymbol{\beta}\in\Theta} \left|(1-\mathbb{E})\left[V_{n,j_1,j_2}\left(\boldsymbol{\beta}\right) - V_{n,j_1,j_2}\left(\boldsymbol{\beta}^*\right)\right]\right| = \sup_{\boldsymbol{\beta}\in\Theta} \left| (1-\mathbb{E})\frac{1}{n}\sum_{i=1}^n \phi^V_{ij_1j_2}\left(\boldsymbol{\beta}\right) \right|.$$

Since

$$
\begin{aligned}
&\left\|(1-\mathbb{E})\left[V_n\left(\widehat{\boldsymbol{\beta}}^{(0)}\right)\right]\right\|_{\max} \\
&=\sup_{j_1,j_2}\left|(1-\mathbb{E})\left[V_{n,j_1,j_2}\left(\widehat{\boldsymbol{\beta}}^{(0)}\right)\right]\right| \\
&\leq \sup_{j_1,j_2}\Phi^V_{n,j_1,j_2}+\sup_{j_1,j_2}\left|(1-\mathbb{E})\,V_{n,j_1,j_2}\left(\boldsymbol{\beta}^*\right)\right|,
\end{aligned} \tag{99}
$$

we break the proof of (93) into two steps, separately controlling the two terms in (99).

Step 1:

$$
\sup_{j_1,j_2}\Phi^V_{n,j_1,j_2}=O_{\mathbb{P}}\left(\delta_{m,0}\sqrt{\frac{s\log p}{nh^6}}\right)\overset{\sqrt{s}\delta_{m,0}=O\left(h^{3/2}\right)}{=\!=\!=\!=}O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{nh^3}}\right). \tag{100}
$$

The proof in this step is analogous to the proof of Lemma B.1 in Luo et al. (2022). Since $|z_{i,j}|$ is upper bounded by $\overline{B}$, for any i and $\boldsymbol{\beta}\in\Theta$, we have $\boldsymbol{z}_i^\top\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\leq\overline{B}\left\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\right\|_1\leq\overline{B}\sqrt{s}\delta_{m,0}$. Since $H''(x)$ is Lipschitz, we have

$$
\overline{\phi}:=\sup_{i,j_1,j_2}\sup_{\boldsymbol{\beta}\in\Theta}\left|\phi^V_{ij_1j_2}\left(\boldsymbol{\beta}\right)\right|=O\left(\frac{\sqrt{s}\delta_{m,0}}{h^3}\right). \tag{101}
$$

Since $\rho\left(\cdot\,|\,\boldsymbol{Z}\right)$ is bounded, we also have

$$
\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\Theta}\mathbb{E}_{\cdot|\boldsymbol{Z}}\left[H''\left(\frac{X+\boldsymbol{Z}^\top\boldsymbol{\beta}}{h}\right)-H''\left(\frac{X+\boldsymbol{Z}^\top\boldsymbol{\beta}^*}{h}\right)\right]^2 \\
&=\sup_{\boldsymbol{\beta}\in\Theta}h\int_{-1}^{1}\left[H''\left(\xi+\frac{\boldsymbol{Z}^\top\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)}{h}\right)-H''\left(\xi\right)\right]^2\rho\left(h\xi\,|\,\boldsymbol{Z}\right)\mathrm{d}\xi \\
&=O\left(s\delta^2_{m,0}/h\right),
\end{aligned} \tag{102}
$$

which implies that

$$
\sup_{i,j_1,j_2}\sup_{\boldsymbol{\beta}\in\Theta}\mathbb{E}\left[\left|\phi^V_{ij_1j_2}\left(\boldsymbol{\beta}\right)\right|^2\right]=O\left(s\delta^2_{m,0}/h^5\right). \tag{103}
$$

Define $\sigma_1,...,\sigma_n$ to be independent Rademacher variables, i.e., binary variables that are uniformly distributed on $\{-1,+1\}$. By Rademacher symmetrization,

$$
\mathbb{E}\Phi^V_{n,j_1,j_2}\leq 2\mathbb{E}\sup_{\boldsymbol{\beta}\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n\sigma_i\phi^V_{ij_1j_2}\left(\boldsymbol{\beta}\right)\right|.
$$

Further, as

$$
\phi^V_{ij_1j_2}\left(\boldsymbol{\beta}\right):=\frac{-y_i}{h^2}H''\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}^*+\boldsymbol{z}_i^\top\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)}{h}\right)z_{i,j_1}z_{i,j_2}+\frac{y_i}{h^2}H''\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}^*}{h}\right)z_{i,j_1}z_{i,j_2},
$$

we can be view $\phi^V_{ij_1j_2}\left(\boldsymbol{\beta}\right)$ as a function of $\boldsymbol{z}_i^\top\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)$ with Lipschitz constant $\asymp\overline{B}^2/h^3$. By Talagrand's Lemma,

$$
\mathbb{E}_\sigma\sup_{\boldsymbol{\beta}\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n\sigma_i\phi^V_{ij_1j_2}\left(\boldsymbol{\beta}\right)\right|\lesssim\frac{1}{h^3}\mathbb{E}_\sigma\sup_{\boldsymbol{\beta}\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n\sigma_i\boldsymbol{z}_i^\top\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\right|\lesssim\left(\frac{\sqrt{s}\delta_{m,0}}{nh^3}\right)\mathbb{E}_\sigma\left\|\sum_{i=1}^n\sigma_i\boldsymbol{z}_i\right\|_\infty.
$$

Since $\sigma_i z_{i,j}\in\left[-\overline{B},+\overline{B}\right]$ for all $j\in\{1,...,p\}$, by Hoeffding's inequality,

$$
\mathbb{P}\left(\left|\sum_{i=1}^n\sigma_i z_{i,j}\right|\geq\sqrt{4\overline{B}^2n\log\max\{n,p\}}\right)\leq 2\exp\left(-\frac{4n\overline{B}^2\log\max\{n,p\}}{2n\overline{B}^2}\right)=\frac{2}{\max\{n,p\}^2},
$$

which implies that with probability at least $1-\frac{2}{\max\{n,p\}}$,

$$\left\|\sum_{i=1}^{n}\sigma_i \boldsymbol{z}_i\right\|_{\infty} \leq 2\overline{B}\sqrt{n\log\max\{n,p\}}.$$

Assumption 12 supposes that $\log p = O(\log n)$, and thus we obtain

$$\mathbb{E}_{\sigma}\left\|\sum_{i=1}^{n}\sigma_i \boldsymbol{z}_i\right\|_{\infty} \leq 2\overline{B}\sqrt{n\log\max\{n,p\}}\left(1-\frac{2}{\max\{n,p\}}\right)+\frac{2\overline{B}}{\max\{n,p\}} = O\left(\sqrt{n\log p}\right),$$

and hence

$$\mathbb{E}\Phi^{V}_{n,j_1,j_2} = O\left(\delta_{m,0}\sqrt{\frac{s\log p}{nh^6}}\right). \tag{104}$$

To show (100), we use Theorem 7.3 in Bousquet (2003), which is restated as the following Lemma B.6.

LEMMA B.6 (Bousquet (2003)). *Assume $\{\boldsymbol{z}_i\}_{i=1}^n$ are identically distributed random variables. Let $\mathcal{F}$ be a set of countable real-value functions such that all functions $f\in\mathcal{F}$ are measurable, square-integrable and satisfy $\mathbb{E}f(\boldsymbol{z}_i)=0$. Assume $\sup_{f,\boldsymbol{z}} f(\boldsymbol{z})\leq 1$. Define*

$$\Upsilon := \sup_{f\in\mathcal{F}}\sum_{i=1}^{n} f(\boldsymbol{z}_i).$$

If $\sum_{i=1}^n \sup_{f\in\mathcal{F}}\mathbb{E}f^2(\boldsymbol{z}_i)\leq n\sigma^2$, then for all $x>0$, we have

$$\mathbb{P}\left(\Upsilon > \mathbb{E}\Upsilon + \sqrt{2x(n\sigma^2+2\mathbb{E}\Upsilon)}+\frac{x}{3}\right) < e^{-x}.$$

Note that Lemma B.6 requires $\mathcal{F}$ to be countable. We first apply Lemma B.6 to prove (100) on rational $\boldsymbol{\beta}$, i.e.,

$$\sup_{j_1,j_2}\sup_{\boldsymbol{\beta}\in\Theta\cap\mathbb{Q}^p}\left|(1-\mathbb{E})\frac{1}{n}\sum_{i=1}^{n}\phi^{V}_{ij_1j_2}(\boldsymbol{\beta})\right| = O_{\mathbb{P}}\left(\delta_{m,0}\sqrt{\frac{s\log n}{nh^6}}\right).$$

Fix j_1, j_2 and take

$$\mathcal{F} := \left\{f_{\boldsymbol{\beta}}(\boldsymbol{z}_i) = \frac{(1-\mathbb{E})\phi^{V}_{ij_1j_2}(\boldsymbol{\beta})}{2\overline{\phi}} : \boldsymbol{\beta}\in\Theta\cap\mathbb{Q}^p\right\}.$$

By (101), (103) and (104), we have $f(\boldsymbol{z}_i)\leq 1$,

$$\sum_{i=1}^{n}\sup_{f_{\boldsymbol{\beta}}\in\mathcal{F}}\mathbb{E}f^2_{\boldsymbol{\beta}}(\boldsymbol{z}_i) = O\left(\frac{ns\delta^2_{m,0}}{\overline{\phi}^2h^5}\right),$$

and

$$\mathbb{E}\Upsilon = O\left(\frac{\delta_{m,0}}{\overline{\phi}h^3}\sqrt{ns\log p}\right).$$

By Lemma B.6, for all $x>0$, with probability $1-e^{-x}$,

$$\frac{1}{n}\sup_{\boldsymbol{\beta}\in\Theta\cap\mathbb{Q}^p}(1-\mathbb{E})\sum_{i=1}^{n}\phi^{V}_{ij_1j_2}(\boldsymbol{\beta}) = O\left(\frac{\delta_{m,0}}{h^3}\sqrt{\frac{s\log p}{n}}+\sqrt{2x\frac{s\delta^2_{m,0}}{nh^5}+4x\frac{\overline{\phi}\delta_{m,0}\sqrt{s\log p}}{n^{3/2}h^3}}+\frac{\overline{\phi}x}{3n}\right).$$

By taking $x = 3\log\max\{n,p\}$, plugging (101) in and using that $\log p = O(\log n)$ again, the above bound can be written as

$$O\left[\delta_{m,0}\left(\sqrt{\frac{s\log p}{nh^6}}+\sqrt{\frac{s\log p}{nh^5}+\frac{s\log p}{n^{3/2}h^6}}+\frac{\sqrt{s}\log p}{nh^3}\right)\right] = O\left(\delta_{m,0}\sqrt{\frac{s\log p}{nh^6}}\right).$$

For the same reason, with probability $1-1/\max\{n,p\}^3$,

$$\frac{1}{n}\sup_{\boldsymbol{\beta}\in\Theta\cap\mathbb{Q}^p}(1-\mathbb{E})\sum_{i=1}^{n}\left[-\phi^V_{ij_1j_2}(\boldsymbol{\beta})\right] = O\left(\delta_{m,0}\sqrt{\frac{s\log p}{nh^6}}\right).$$

Therefore, with probability $1-2/\max\{n,p\}^3$,

$$\sup_{\boldsymbol{\beta}\in\Theta\cap\mathbb{Q}^p}\left|(1-\mathbb{E})\frac{1}{n}\sum_{i=1}^{n}\phi^V_{ij_1j_2}(\boldsymbol{\beta})\right| = O\left(\delta_{m,0}\sqrt{\frac{s\log p}{nh^6}}\right).$$

By the continuity of $\phi^V_{ij_1j_2}(\boldsymbol{\beta})$,

$$\sup_{\boldsymbol{\beta}\in\Theta}\left|(1-\mathbb{E})\frac{1}{n}\sum_{i=1}^{n}\phi^V_{ij_1j_2}(\boldsymbol{\beta})\right| = O\left(\delta_{m,0}\sqrt{\frac{s\log p}{nh^6}}\right),$$

with the same probability. This is true for any j_1, j_2, hence

$$\sup_{j_1,j_2}\Phi^V_{n,j_1,j_2} = O\left(\delta_{m,0}\sqrt{\frac{s\log p}{nh^6}}\right) \overset{\sqrt{s}\delta_{m,0}=O\left(h^{3/2}\right)}{=\!=\!=} O\left(\sqrt{\frac{\log p}{nh^3}}\right),$$

with probability at least $1-2/\max\{n,p\}$, which completes the proof of (100).

Step 2:

$$\sup_{j_1,j_2}\left|(1-\mathbb{E})V_{n,j_1,j_2}(\boldsymbol{\beta}^*)\right| = O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{nh^3}}\right). \tag{105}$$

Recall that

$$V_{n,j_1,j_2}(\boldsymbol{\beta}^*) := \frac{1}{nh^2}\sum_{i=1}^{n}(-y_i)H''\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}^*}{h}\right)z_{i,j_1}z_{i,j_2}.$$

We have

$$\sup_i\left|\frac{-y_i}{h^2}H''\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}^*}{h}\right)z_{i,j_1}z_{i,j_2}\right| = O\left(1/h^2\right),$$

and

$$\sup_i\mathbb{E}\left|\frac{-y_i}{h^2}H''\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}^*}{h}\right)z_{i,j_1}z_{i,j_2}\right|^2 = O\left(1/h^3\right),$$

as $H''(x)$, $|z_{i,j}|$ is bounded and

$$\mathbb{E}_{\cdot\,|\,\boldsymbol{Z}}\left[H''\left(\frac{X+\boldsymbol{Z}^\top\boldsymbol{\beta}^*}{h}\right)\right]^2 = h\int_{-1}^{1}\left[H''(\xi)\right]^2\rho(\xi h\,|\,\boldsymbol{Z})\,\mathrm{d}\xi = O(h). \tag{106}$$

By Bernstein's inequality, there exists a constant $C>0$,

$$\mathbb{P}\left(\left|(1-\mathbb{E})V_{n,j_1,j_2}(\boldsymbol{\beta}^*)\right| \geq \sqrt{\frac{C_2\log\max\{n,p\}}{nh^3}}\right) \leq 2\exp\left(\frac{-\frac{C}{2}\log\max\{n,p\}}{1+\frac{1}{3}\sqrt{C\log\max\{n,p\}/(nh)}}\right).$$

Our assumptions $\log m/mh^3 = o(1)$, $m > n^c$ and $p = O(n^\gamma)$ ensure that $\log \max\{n,p\}/(nh) = o(1)$, and hence we can take a sufficiently large C to make

$$2\exp\left(\frac{-\frac{C}{2}\log\max\{n,p\}}{1+\frac{1}{3}\sqrt{C\log\max\{n,p\}/(nh)}}\right) \le \frac{2}{\max\{n,p\}^3}.$$

This implies that

$$\mathbb{P}\left(\sup_{j_1,j_2}\left|\left(1-\mathbb{E}\right)V_{n,j_1,j_2}\left(\boldsymbol{\beta}^*\right)\right| \le \sqrt{\frac{C\log\max\{n,p\}}{nh^3}}\right) \ge 1-\frac{2}{\max\{n,p\}},$$

which proves (105). Together with (99) and (100), we conclude the proof of (93).

Proof of (94):

Recall that, by Equation (55), for almost every $\boldsymbol{Z}$,

$$\begin{aligned}&\left(2F\left(-t\,|\,\boldsymbol{Z}\right)-1\right)\rho\left(t\,|\,\boldsymbol{Z}\right)\\ &=2F^{(1)}\left(0\,|\,\boldsymbol{Z}\right)\rho\left(0\,|\,\boldsymbol{Z}\right)t+\sum_{k=2}^{2\alpha+1}M_k\left(\boldsymbol{Z}\right)t^k,\end{aligned}\tag{107}$$

where $M_k(\boldsymbol{Z})$ is a constant depending on $\boldsymbol{Z}$, t' and t''. Since $\rho^{(k)}(\cdot\,|\,\boldsymbol{Z})$ and $F^{(k)}(\cdot\,|\,\boldsymbol{Z})$ are bounded around 0 for all k, we know there exists a constant M such that $\sup_k |M_k(\boldsymbol{Z})| \le M$ for all $\boldsymbol{Z}, t', t''$. In the following computation, we let $t = \xi h - \boldsymbol{Z}^\top\Delta(\boldsymbol{\beta})$, where $\Delta(\boldsymbol{\beta}) := \boldsymbol{\beta}-\boldsymbol{\beta}^*$.

Recall that when $x>1$ or $x<-1$, $H'(x) = H''(x) = 0$. The kernel $H'(x)$ is bounded, $\int_{-1}^1 H'(x)\,\mathrm{d}x = 1$, and $\int_{-1}^1 x^k H'(x)\,\mathrm{d}x = 0$ for all $1\le k\le\alpha-1$. Moreover, $\int_{-1}^1 xH''(x)\,\mathrm{d}x = -1$ and $\int_{-1}^1 x^k H''(x)\,\mathrm{d}x = 0$ for $k=0$ and $2\le k\le\alpha$. Also, recall that $\zeta = X+\boldsymbol{Z}^\top\boldsymbol{\beta}^*$ and $-y = -\mathrm{sign}(y^*) = -\mathrm{sign}(\boldsymbol{Z}+\epsilon) = 2\mathbb{I}(\boldsymbol{Z}+\epsilon<0)-1$.

For all $\boldsymbol{v}_1, \boldsymbol{v}_2$ that satisfies $\|\boldsymbol{v}_1\|_2 = \|\boldsymbol{v}_2\|_2 = 1$ and $\boldsymbol{\beta}\in\Theta$,

$$\begin{aligned}&\mathbb{E}_{\cdot|\boldsymbol{Z}}\left[\frac{\left(\boldsymbol{v}_1^\top\boldsymbol{Z}\right)\left(\boldsymbol{v}_2^\top\boldsymbol{Z}\right)}{h^2}\left(-Y\right)H''\left(\frac{X+\boldsymbol{Z}^\top\boldsymbol{\beta}}{h}\right)\right]\\ =&\frac{\left(\boldsymbol{v}_1^\top\boldsymbol{Z}\right)\left(\boldsymbol{v}_2^\top\boldsymbol{Z}\right)}{h^2}\mathbb{E}_{\cdot\,|\,\boldsymbol{Z}}\left[2\mathbb{I}\left(\boldsymbol{Z}+\epsilon<0\right)-1\right]H''\left(\frac{\boldsymbol{Z}^\top\Delta\left(\boldsymbol{\beta}\right)+\zeta}{h}\right)\\ =&\frac{\left(\boldsymbol{v}_1^\top\boldsymbol{Z}\right)\left(\boldsymbol{v}_2^\top\boldsymbol{Z}\right)}{h}\int_{-1}^1\left[2F\left(\boldsymbol{Z}^\top\Delta\left(\boldsymbol{\beta}\right)-\xi h\,|\,\boldsymbol{Z}\right)-1\right]\rho\left(\xi h-\boldsymbol{Z}^\top\Delta\left(\boldsymbol{\beta}\right)\,|\,\boldsymbol{Z}\right)H''\left(\xi\right)\mathrm{d}\xi\\ =&\frac{\left(\boldsymbol{v}_1^\top\boldsymbol{Z}\right)\left(\boldsymbol{v}_2^\top\boldsymbol{Z}\right)}{h}2F^{(1)}\left(0\,|\,\boldsymbol{Z}\right)\rho\left(0\,|\,\boldsymbol{Z}\right)\int_{-1}^1\left(\xi h-\boldsymbol{Z}^\top\Delta\left(\boldsymbol{\beta}\right)\right)H''\left(\xi\right)\mathrm{d}\xi\\ &+\frac{\left(\boldsymbol{v}_1^\top\boldsymbol{Z}\right)\left(\boldsymbol{v}_2^\top\boldsymbol{Z}\right)}{h}\cdot\sum_{k=2}^{2\alpha+1}M_k\left(\boldsymbol{Z}\right)\int_{-1}^1\left(\xi h-\boldsymbol{Z}^\top\Delta\left(\boldsymbol{\beta}\right)\right)^k H''\left(\xi\right)\mathrm{d}\xi\end{aligned}\tag{108}$$

For $1\le k\le\alpha$,

$$\int_{-1}^1\left(\xi h-\boldsymbol{Z}^\top\Delta\left(\boldsymbol{\beta}\right)\right)^k H''\left(\xi\right)\mathrm{d}\xi = \sum_{k'=0}^{k}h^{k'}\left(\boldsymbol{Z}^\top\Delta\left(\boldsymbol{\beta}\right)\right)^{k-k'}\int_{-1}^1\xi^{k'}H''\left(\xi\right)\mathrm{d}\xi = h\left(\boldsymbol{Z}^\top\Delta\left(\boldsymbol{\beta}\right)\right)^{k-1}.$$

For $\alpha+1\le k\le 2\alpha+1$, since $H''(x)$ is bounded,

$$\left|\int_{-1}^1\left(\xi h-\boldsymbol{Z}^\top\Delta\left(\boldsymbol{\beta}\right)\right)^k H''\left(\xi\right)\mathrm{d}\xi\right| \le \int_{-1}^1 2^{k-1}\left(\left|\xi h\right|^k+\left|\boldsymbol{Z}^\top\Delta\left(\boldsymbol{\beta}\right)\right|^k\right)\left|H''\left(\xi\right)\right|\mathrm{d}\xi$$

$$\leq 2^{2\alpha+1} \sup_x \left|H''(x)\right| \left[h^{\alpha+1} + \left|\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})\right|^k\right]$$
$$\leq 2^{2\alpha+2} \left(1+\overline{B}^k\right) \sup_x \left|H''(x)\right| h^{\alpha+1}.$$

The last inequality holds because $|Z_j| \leq \overline{B}$, $\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta}) \leq \overline{B}\sqrt{s}\delta_{m,0}$, $\sqrt{s}\delta_{m,0} = O\left(h^{3/2}\right)$ and $h = o(1)$. Hence,

$$\mathbb{E}_{\cdot|\boldsymbol{Z}}\left[\frac{\left(\boldsymbol{v}_1^\top \boldsymbol{Z}\right)\left(\boldsymbol{v}_2^\top \boldsymbol{Z}\right)}{h}(-y) H''\left(\frac{X+\boldsymbol{Z}^\top\boldsymbol{\beta}}{h}\right)\right] - \frac{\left(\boldsymbol{v}_1^\top \boldsymbol{Z}\right)\left(\boldsymbol{v}_2^\top \boldsymbol{Z}\right)}{h} 2F^{(1)}(0\,|\,\boldsymbol{Z})\,\rho(0\,|\,\boldsymbol{Z})$$
$$\leq \left|\frac{\left(\boldsymbol{v}_1^\top \boldsymbol{Z}\right)\left(\boldsymbol{v}_2^\top \boldsymbol{Z}\right)}{h} \sum_{k=2}^{\alpha} M_k(\boldsymbol{Z})\, h \left(\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})\right)^{k-1}\right|$$
$$+ \left|\frac{\left(\boldsymbol{v}_1^\top \boldsymbol{Z}\right)\left(\boldsymbol{v}_2^\top \boldsymbol{Z}\right)}{h} 2^{2\alpha+2} \sup_x \left|H''(x)\right| \sum_{k=\alpha+1}^{2\alpha+1} \left(1+\overline{B}^k\right) M_k(\boldsymbol{Z})\, h^{\alpha+1}\right|$$
$$\leq C\left(\boldsymbol{v}_1^\top \boldsymbol{Z}\right)\left(\boldsymbol{v}_2^\top \boldsymbol{Z}\right)\left(\sqrt{s}\delta_{m,0} + h^\alpha\right),$$

where C is a constant not depending on $\boldsymbol{\beta}$ and $\boldsymbol{Z}$. Therefore, by the assumption that $\boldsymbol{Z}$ has finite second moment and Cauchy-Schwarz inequality, we obtain that

(109)

$$\sup_{\boldsymbol{\beta}\in\Theta} \sup_{\|\boldsymbol{v}_1\|_2 = \|\boldsymbol{v}_2\|_2 = 1} \boldsymbol{v}_1^\top \left(\mathbb{E}\left[V_n(\boldsymbol{\beta})\right] - V\right) \boldsymbol{v}_2$$
$$= \sup_{\boldsymbol{\beta}\in\Theta} \sup_{\|\boldsymbol{v}_1\|_2 = \|\boldsymbol{v}_2\|_2 = 1} \mathbb{E}\left[\frac{\left(\boldsymbol{v}_1^\top \boldsymbol{Z}\right)\left(\boldsymbol{v}_2^\top \boldsymbol{Z}\right)}{h}(-y) H''\left(\frac{\boldsymbol{Z}^\top\boldsymbol{\beta}}{h}\right) - \frac{\left(\boldsymbol{v}_1^\top \boldsymbol{Z}\right)\left(\boldsymbol{v}_2^\top \boldsymbol{Z}\right)}{h} 2F^{(1)}(0\,|\,\boldsymbol{Z})\,\rho(0\,|\,\boldsymbol{Z})\right]$$
$$\lesssim \sqrt{s}\delta_{m,0} + h^\alpha,$$

which completes the proof of (94).

Proof of (95) and (96):

Equation (95) and (96) can be shown in the same way as above by replacing all the n with m.

Proof of (97):

The proof of (97) is analogous to that of (93). We will omit some details since they are the same. For each $j \in \{1, ..., p\}$, define

$$U_{n,h,j}(\boldsymbol{\beta}) := \frac{1}{nh^2}\sum_{i=1}^n (-y_i) H''\left(\frac{x_i + \boldsymbol{z}_i^\top\boldsymbol{\beta}}{h}\right) \boldsymbol{z}_i^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^*)\, z_{i,j}$$
$$- \frac{1}{nh}\sum_{i=1}^n (-y_i) H'\left(\frac{x_i + \boldsymbol{z}_i^\top\boldsymbol{\beta}}{h}\right) z_{i,j}.$$

Then we have

$$\left\|(1-\mathbb{E})\Psi_n\left(\widehat{\boldsymbol{\beta}}^{(0)}\right)\right\|_\infty$$
$$= \left\|(1-\mathbb{E})\left[\frac{1}{nh^2}\sum_{i=1}^n (-y_i) H''\left(\frac{x_i + \boldsymbol{z}_i^\top\widehat{\boldsymbol{\beta}}^{(0)}}{h}\right)\boldsymbol{z}_i^\top\left(\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\right)\boldsymbol{z}_i - \frac{1}{nh}\sum_{i=1}^n (-y_i) H'\left(\frac{x_i + \boldsymbol{z}_i^\top\boldsymbol{\beta}^{(0)}}{h}\right)\boldsymbol{z}_i\right]\right\|_\infty$$
$$\leq \sup_j \sup_{\boldsymbol{\beta}\in\Theta} \left|(1-\mathbb{E})\left(U_{n,h,j}(\boldsymbol{\beta}) - U_{n,h,j}(\boldsymbol{\beta}^*)\right)\right| + \sup_j \left|(1-\mathbb{E})\, U_{n,h,j}(\boldsymbol{\beta}^*)\right|.$$

Define

$$\phi_{i,j}^{U}(\boldsymbol{\beta}) := \left| \frac{-y_i}{h^2} H''\left(\frac{x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}}{h}\right) \boldsymbol{z}_i^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^*) z_{i,j} - \frac{-y_i}{h}\left[H'\left(\frac{x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}}{h}\right) - H'\left(\frac{x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}^*}{h}\right)\right] z_{i,j} \right|.$$

$$\Phi_{n,h,j}^{U} := \sup_{\boldsymbol{\beta}\in\Theta} \left|(1-\mathbb{E})\left(U_{n,h,j}(\boldsymbol{\beta}) - U_{n,h,j}(\boldsymbol{\beta}^*)\right)\right| = \sup_{\boldsymbol{\beta}\in\Theta} \left|(1-\mathbb{E})\frac{1}{n}\sum_{i=1}^{n}\phi_{i,j}^{U}(\boldsymbol{\beta})\right|.$$

Similar to the analysis of $\phi_{i,j}^{V}$, we have

$$\sup_{i,j}\sup_{\boldsymbol{\beta}\in\Theta}\left|\phi_{i,j}^{U}(\boldsymbol{\beta})\right| = O\left(\frac{\sqrt{s}\delta_{m,0}}{h^2}\right),$$

and

$$\sup_{i,j}\sup_{\boldsymbol{\beta}\in\Theta}\mathbb{E}\left[\phi_{i,j}^{U}(\boldsymbol{\beta})\right]^2 = O\left(\frac{s\delta_{m,0}^2}{h^3}\right).$$

By Rademacher symmetrization, Talagrand's concentration principle and Hoeffding's inequality,

$$\mathbb{E}\Phi_{n,h,j}^{U} \le 2\mathbb{E}\sup_{\boldsymbol{\beta}\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\phi_{i,j}^{U}(\boldsymbol{\beta})\right| \lesssim \left(\frac{\sqrt{s}\delta_{m,0}}{nh^2}\right)\cdot\mathbb{E}\left(\mathbb{E}_\sigma\left\|\sum_{i=1}^{n}\sigma_i\boldsymbol{z}_i^\top\right\|_\infty\right) \lesssim \delta_{m,0}\sqrt{\frac{s\log p}{nh^4}}.$$

(Details are the same as the proof of (104), while the only difference is that $\phi_{ij}^{U}(\boldsymbol{\beta})$ is a Lipschitz function of $\boldsymbol{z}_i^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^*)$ with Lipschitz constant $\asymp 1/h^2$, instead of $1/h^3$ for $\phi_{ij}^{V}(\boldsymbol{\beta})$.)

Using Lemma B.6 again, we can show that

$$\sup_j \Phi_{n,h,j}^{U} = O_{\mathbb{P}}\left(\delta_{m,0}\sqrt{\frac{s\log p}{nh^4}}\right) \underbrace{=}_{\sqrt{s}\delta_{m,0}=O\left(h^{3/2}\right)} O\left(\sqrt{\frac{\log p}{nh}}\right).$$

Similar to the proof of (105), we have

$$\sup_{\boldsymbol{\beta}\in\Theta}\sup_{i,j}\left|\frac{-y_i}{h}H'\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}^*}{h}\right)x_{i,j}\right| = O(1/h),$$

and

$$\sup_{\boldsymbol{\beta}\in\Theta}\sup_{i,j}\mathbb{E}\left|\frac{-y_i}{h}H'\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}^*}{h}\right)x_{i,j}\right|^2 = O(1/h).$$

By Bernstein's inequality,

$$\sup_j\left|(1-\mathbb{E})U_{n,h,j}(\boldsymbol{\beta}^*)\right| = O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{nh}}\right).$$

Proof of (98):

Recall equation (55) and $\pi_U = \int_{-1}^{1} x^\alpha H'(x)\,\mathrm{d}x \neq 0$. For any $\boldsymbol{v} \in \mathbb{R}^p$ that satisfies $\|\boldsymbol{v}\|_2 = 1$ and $\boldsymbol{\beta} \in \Theta$, we have

(110)

$$
\begin{aligned}
&\mathbb{E}_{\cdot|\boldsymbol{Z}} \boldsymbol{v}^\top \Psi_n(\boldsymbol{\beta}) \\
=&(\boldsymbol{v}^\top \boldsymbol{Z}) \cdot \mathbb{E}_{\cdot|\boldsymbol{Z}} \left[\frac{\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})}{h^2} [2\mathbb{I}(\zeta + \epsilon < 0) - 1] H''\left(\frac{X + \boldsymbol{Z}^\top \boldsymbol{\beta}}{h} \right) - \frac{1}{h} [2\mathbb{I}(\zeta + \epsilon < 0) - 1] H'\left(\frac{X + \boldsymbol{Z}^\top \boldsymbol{\beta}}{h} \right) \right] \\
=&(\boldsymbol{v}^\top \boldsymbol{Z}) \int_{-1}^{1} \left[2F\left(\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta}) - \xi h \,|\, \boldsymbol{Z} \right) - 1 \right] \rho\left(\xi h - \boldsymbol{Z}^\top \Delta(\boldsymbol{\beta}) \,|\, \boldsymbol{Z} \right) \\
&\cdot \left(\frac{\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})}{h} H''(\xi) - H'(\xi) \right) \mathrm{d}\xi \\
=&(\boldsymbol{v}^\top \boldsymbol{Z}) \sum_{k=1}^{2\alpha+1} M_k(\boldsymbol{Z}) \int_{-1}^{1} \left(\xi h - \boldsymbol{Z}^\top \Delta(\boldsymbol{\beta}) \right)^k \left(\frac{\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})}{h} H''(\xi) - H'(\xi) \right) \mathrm{d}\xi
\end{aligned}
$$

For $1 \leq k \leq \alpha - 1$,

$$
\begin{aligned}
&\sup_{\boldsymbol{\beta} \in \Theta} \left| \int_{-1}^{1} \left(\xi h - \boldsymbol{Z}^\top \Delta(\boldsymbol{\beta}) \right)^k \left(\frac{\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})}{h} H''(\xi) - H'(\xi) \right) \mathrm{d}\xi \right| \\
=&\sup_{\boldsymbol{\beta} \in \Theta} \left| \sum_{k'=0}^{k} \binom{k}{k'} h^{k'} \left(-\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta}) \right)^{k-k'} \left[\left(\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta}) / h \right) \int_{-1}^{1} \xi^{k'} H''(\xi)\,\mathrm{d}\xi - \int_{-1}^{1} \xi^{k'} H'(\xi)\,\mathrm{d}\xi \right] \right| \\
=&\,(k-1) \left| -\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta}) \right|^k = O\left(s\delta_{m,0}^2 \right).
\end{aligned}
$$

For $k = \alpha$,

$$
\begin{aligned}
&\sup_{\boldsymbol{\beta} \in \Theta} \left| \int_{-1}^{1} \left(\xi h - \boldsymbol{Z}^\top \Delta(\boldsymbol{\beta}) \right)^\alpha \left(\frac{\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})}{h} H''(\xi) - H'(\xi) \right) \mathrm{d}\xi \right| \\
=&\sup_{\boldsymbol{\beta} \in \Theta} \left| \sum_{k=0}^{\alpha} \binom{\alpha}{k} h^{k} \left(-\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta}) \right)^{\alpha-k} \left[\left(\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta}) / h \right) \int_{-1}^{1} \xi^{k} H''(\xi)\,\mathrm{d}\xi - \int_{-1}^{1} \xi^{k} H'(\xi)\,\mathrm{d}\xi \right] \right| \\
\leq&\,(\alpha-1) \left| \boldsymbol{Z}^\top \Delta(\boldsymbol{\beta}) \right|^\alpha + |\pi_U h^\alpha| = O\left[h^\alpha + \left(\sqrt{s}\delta_{m,0} \right)^\alpha \right].
\end{aligned}
$$

For $\alpha + 1 \leq k \leq 2\alpha + 1$,

$$
\sup_{\boldsymbol{\beta} \in \Theta} \left| \int_{-1}^{1} \left(\xi h - \boldsymbol{Z}^\top \Delta(\boldsymbol{\beta}) \right)^k \left(\frac{\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})}{h} H''(\xi) - H'(\xi) \right) \mathrm{d}\xi \right| = O\left[h^{\alpha+1} + \left(\sqrt{s}\delta_{m,0} \right)^{\alpha+1} \right].
$$

Therefore, by the assumption that $\boldsymbol{Z}$ has finite second moment and Cauchy-Schwarz inequality, we obtain that

$$
\mathbb{E}\left[\boldsymbol{v}^\top \Psi_n(\boldsymbol{\beta}) \right] \lesssim \mathbb{E}\left[(\boldsymbol{v}^\top \boldsymbol{Z}) \left(s\delta_{m,0}^2 + h^\alpha \right) \right] \lesssim s\delta_{m,0}^2 + h^\alpha,
$$

which completes the proof of (98).

□

Proof of Theorem 5.1

Now we are ready to prove the 1-step error for $\widehat{\boldsymbol{\beta}}^{(1)}$.

PROOF. For simplicity, we replace $V_{m,1}\left(\widehat{\boldsymbol{\beta}}^{(0)}\right)$, $V_n\left(\widehat{\boldsymbol{\beta}}^{(0)}\right)$, $U_n\left(\widehat{\boldsymbol{\beta}}^{(0)}\right)$, and $\lambda_n^{(1)}$ by $V_{m,1}$, V_n, U_n, and λ_n, respectively. Then, by Algorithm 3,

$$\widehat{\boldsymbol{\beta}}^{(1)} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \|\boldsymbol{\beta}\|_1 : \left\| V_{m,1}\boldsymbol{\beta} - \left(V_{m,1}\widehat{\boldsymbol{\beta}}^{(0)} - U_n\right)\right\|_\infty \le \lambda_n \right\}.$$

Hence,

$$\left\| V_{m,1}\widehat{\boldsymbol{\beta}}^{(1)} - \left(V_{m,1}\widehat{\boldsymbol{\beta}}^{(0)} - U_n\right)\right\|_\infty \le \lambda_n. \tag{111}$$

Using Lemma B.5, with probability tending to one, we have

$$\begin{aligned}
&\left\|(1-\mathbb{E})\left[V_{m,1}\boldsymbol{\beta}^* - \left(V_{m,1}\widehat{\boldsymbol{\beta}}^{(0)} - U_n\right)\right]\right\|_\infty \\
&\le \left\|(1-\mathbb{E})\left[U_n - V_n\left(\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\right)\right]\right\|_\infty + \left\|(1-\mathbb{E})\left(V_{m,1}-V_n\right)\left(\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\right)\right\|_\infty \\
&\le \left\|(1-\mathbb{E})\left[U_n - V_n\left(\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\right)\right]\right\|_\infty + \|(1-\mathbb{E})(V_{m,1}-V_n)\|_{\max}\left\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\right\|_1 \\
&\le C_\lambda\left(\sqrt{\frac{\log p}{nh}} + \sqrt{\frac{s\log p}{mh^3}}\delta_{m,0}\right).
\end{aligned} \tag{112}$$

and

$$\begin{aligned}
&\left\|\mathbb{E}\left[V_{m,1}\boldsymbol{\beta}^* - \left(V_{m,1}\widehat{\boldsymbol{\beta}}^{(0)} - U_n\right)\right]\right\|_2 \\
&\le \left\|\mathbb{E}\left[U_n - V_n\left(\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\right)\right]\right\|_2 + \left\|\mathbb{E}\left(V_{m,1}-V_n\right)\left(\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\right)\right\|_2 \\
&\le C_\lambda\left(s\delta_{m,0}^2 + h^\alpha\right).
\end{aligned} \tag{113}$$

for some large enough constant C_λ. By letting $\lambda_n = C_\lambda\left[\sqrt{\frac{\log p}{nh}} + \sqrt{\frac{s\log p}{mh^3}}\delta_{m,0} + s\delta_{m,0}^2 + h^\alpha\right]$, Equations (112) and (113) implies that

$$\left\| V_{m,1}\boldsymbol{\beta}^* - \left(V_{m,1}\widehat{\boldsymbol{\beta}}^{(0)} - U_n\right)\right\|_\infty \le \lambda_n. \tag{114}$$

Combining it with (111), we obtain that, with probability tending to 1,

$$\left\| V_{m,1}\left(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^{(1)}\right)\right\|_\infty \le 2\lambda_n. \tag{115}$$

Moreover, by the optimality of $\widehat{\boldsymbol{\beta}}^{(1)}$, it holds that $\left\|\widehat{\boldsymbol{\beta}}^{(1)}\right\|_1 \le \|\boldsymbol{\beta}^*\|_1$, which implies

$$\left\|\widehat{\boldsymbol{\beta}}_S^{(1)}\right\|_1 + \left\|\widehat{\boldsymbol{\beta}}_{S^c}^{(1)}\right\|_1 = \left\|\widehat{\boldsymbol{\beta}}^{(1)}\right\|_1 \le \|\boldsymbol{\beta}^*\|_1 = \|\boldsymbol{\beta}_S^*\|_1.$$

Therefore,

$$\left\|\left(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^{(1)}\right)_{S^c}\right\|_1 = \left\|\widehat{\boldsymbol{\beta}}_{S^c}^{(1)}\right\|_1 \le \left\|\boldsymbol{\beta}_S^* - \widehat{\boldsymbol{\beta}}_S^{(1)}\right\|_1 = \left\|\left(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^{(1)}\right)_S\right\|_1,$$

and hence,

$$\left\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^{(1)}\right\|_1 \le 2\left\|\left(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^{(1)}\right)_S\right\|_1 \le 2\sqrt{s}\left\|\left(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^{(1)}\right)_S\right\|_2 \le 2\sqrt{s}\left\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^{(1)}\right\|_2.$$

Let $\boldsymbol{\delta} := \widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*$. So far we have shown that, with probability tending to one, $\|\boldsymbol{\delta}\|_1 \leq 2\sqrt{s}\,\|\boldsymbol{\delta}\|_2$ and $\|V_{m,1}\boldsymbol{\delta}\|_\infty \leq 2\lambda_{m,0}$. Therefore,

$$
\begin{aligned}
\boldsymbol{\delta}^\top V_{m,1}\boldsymbol{\delta} &= \boldsymbol{\delta}^\top V\boldsymbol{\delta} + \boldsymbol{\delta}^\top\left(V_{m,1} - \mathbb{E}\left[V_{m,1}\right]\right)\boldsymbol{\delta} + \boldsymbol{\delta}^\top\left(\mathbb{E}\left[V_{m,1}\right] - V\right)\boldsymbol{\delta} \\
&\geq \Lambda_{\min}\left(V\right)\|\boldsymbol{\delta}\|_2^2 - \|(1-\mathbb{E})V_{m,1}\|_{\max}\|\boldsymbol{\delta}\|_1^2 + \boldsymbol{\delta}^\top\left(\mathbb{E}\left[V_{m,1}\right] - V\right)\boldsymbol{\delta} \\
&\geq \Lambda_{\min}\left(V\right)\|\boldsymbol{\delta}\|_2^2 - s\,\|(1-\mathbb{E})V_{m,1}\|_{\max}\|\boldsymbol{\delta}\|_2^2 - \left|\boldsymbol{\delta}^\top\left(\mathbb{E}\left[V_{m,1}\right] - V\right)\boldsymbol{\delta}\right|.
\end{aligned} \tag{116}
$$

By (95),

$$
s\,\|(1-\mathbb{E})V_{m,1}\|_{\max} = O_{\mathbb{P}}\left(\sqrt{\frac{s^2\log p}{mh^3}}\right) = o_{\mathbb{P}}\left(1\right).
$$

By (96),

$$
\left|\boldsymbol{\delta}^\top\left(\mathbb{E}\left[V_{m,1}\right] - V\right)\boldsymbol{\delta}\right| \lesssim \left(\sqrt{s}\delta_{m,0} + h^\alpha\right)\|\boldsymbol{\delta}\|_2^2 = o(\|\boldsymbol{\delta}\|_2^2).
$$

Therefore, (116) leads to

$$
\boldsymbol{\delta}^\top V_{m,1}\boldsymbol{\delta} \geq \Lambda_{\min}\left(V\right)\|\boldsymbol{\delta}\|_2^2 - o_{\mathbb{P}}\left(\|\boldsymbol{\delta}\|_2^2\right) \geq \left(\Lambda_{\min}\left(V\right)/2\right)\|\boldsymbol{\delta}\|_2^2,
$$

with probability tending to one.

On the other hand, combining (111), (114) and (113) yields that

$$
\begin{aligned}
&\boldsymbol{\delta}^\top V_{m,1}\boldsymbol{\delta} \\
&= \boldsymbol{\delta}^\top\left\{V_{m,1}\widehat{\boldsymbol{\beta}}^{(1)} - \left(V_{m,1}\widehat{\boldsymbol{\beta}}^{(0)} - U_n\right) - (1-\mathbb{E})\left[V_{m,1}\boldsymbol{\beta}^* - \left(V_{m,1}\widehat{\boldsymbol{\beta}}^{(0)} - U_n\right)\right]\right\} \\
&\quad - \boldsymbol{\delta}^\top\mathbb{E}\left[V_{m,1}\boldsymbol{\beta}^* - \left(V_{m,1}\widehat{\boldsymbol{\beta}}^{(0)} - U_n\right)\right] \\
&\leq C_\lambda\left(\sqrt{\frac{\log p}{nh}} + \sqrt{\frac{s\log p}{mh^3}}\delta_{m,0}\right)\|\boldsymbol{\delta}\|_1 + C_\lambda(s\delta_{m,0}^2 + h^\alpha)\|\boldsymbol{\delta}\|_2 \\
&\leq C_\lambda\left(\sqrt{\frac{s\log p}{nh}} + \sqrt{\frac{s^2\log p}{mh^3}}\delta_{m,0} + s\delta_{m,0}^2 + h^\alpha\right)\|\boldsymbol{\delta}\|_2.
\end{aligned} \tag{117}
$$

Combining the two inequalities above completes the proof. □

Proof of Theorem A.1

PROOF. First note that $\sqrt{\frac{s\log p}{nh}} + \sqrt{s}h^\alpha \asymp \sqrt{s}\left(\frac{\log p}{n}\right)^{\frac{\alpha}{2\alpha+1}}$ when $h = h^* = \left(\frac{s\log p}{n}\right)^{\frac{1}{2\alpha+1}}$. Let $\alpha_0 = \max\left\{\frac{3}{2c(1-2r)} + \frac{r}{2(1-2r)}, \frac{3r}{2(1-4r)}\right\}$ and $\delta_{m,0} = (s\log p/m)^{\alpha/(2\alpha+1)}$. Since $s = O\left(m^r\right)$ and $n = O\left(m^{1/c}\right)$ for some $0 < c < 1$ and $0 < r < 1/4$, the condition $\alpha > \alpha_0$ guarantees $\frac{s^2\log p}{m(h^*)^3} = o\left(1\right)$, $s^{3/2}\delta_{m,0} = o(1)$, (which implies $r_m = o(1)$), $s\delta_{m,0}^2 = O((h^*)^3)$, and $s(h^*)^\alpha = o\left(1\right)$. Adding the requirement of $\delta_{m,0}$, the assumptions in Theorem 5.1 hold, which proves Theorem A.1 when $t = 1$.

Now we show Theorem A.1 by induction. Define $\delta_{m,t} = \sqrt{s}\left(\frac{\log p}{n}\right)^{\frac{\alpha}{2\alpha+1}} + (r_m)^t\delta_{m,0} = \sqrt{s}\lambda_n^{(t)}$. Assume that $\left\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\right\|_2 = O_{\mathbb{P}}(\delta_{m,t})$ and $\left\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\right\|_1 = O_{\mathbb{P}}(\sqrt{s}\delta_{m,t})$ is true for

t. Note that $\sqrt{s}\left(\frac{\log p}{n}\right)^{\frac{\alpha}{2\alpha+1}} < \delta_{m,0}$, which implies $\delta_{m,t} < \delta_{m,0}$, and thus $s^{3/2}\delta_{m,t} = o(1)$ and $\sqrt{s}\delta_{m,t} = O\left((h^*)^{3/2}\right)$. Then by Theorem 5.1, when

$$\lambda_n^{(t+1)} = C_\lambda\left(\sqrt{\frac{\log p}{nh^*}} + \delta_{m,t}\sqrt{\frac{s\log p}{m(h^*)^3}} + s\delta_{m,t}^2 + (h^*)^\alpha\right),$$

we have

$$\begin{aligned}\left\|\widehat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\right\|_2 &= O_{\mathbb{P}}\left[\sqrt{s}\left(\frac{\log p}{n}\right)^{\frac{\alpha}{2\alpha+1}} + \left(\sqrt{\frac{s^2\log p}{m(h^*)^3}} + s^{3/2}\delta_{m,t}\right)\delta_{m,t}\right] \\ &= O_{\mathbb{P}}\left[\sqrt{s}\left(\frac{\log p}{n}\right)^{\frac{\alpha}{2\alpha+1}} + \left(\sqrt{\frac{s^2\log p}{m(h^*)^3}} + s^{3/2}\delta_{m,0}\right)^{t+1}\delta_{m,0}\right],\end{aligned}$$

and

$$\left\|\widehat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\right\|_1 \le 2\sqrt{s}\left\|\widehat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^*\right\|_2,$$

with probability tending to 1. This completes the proof. □

Proof of Data Adaptive Methods for Unknown Parameters

PROOF OF THEOREM A.2. First we prove the theorem for $t = 1$. Define $s^* = 2^{\lfloor \log_2(s)\rfloor+1} \in \mathcal{S}$ and define a good event

$$\begin{aligned}E_{m,n} := &\left\{\left\|V_{m,1}\boldsymbol{\beta}^* - \left(V_{m,1}\widehat{\boldsymbol{\beta}}^{(0)} - U_n\right)\right\|_\infty \le \lambda_{n,s^*}^{(1)}\right\} \cap \\ &\left\{\boldsymbol{\delta}^\top V_{m,1}\boldsymbol{\delta} \ge \frac{\Lambda_{\min}(V)}{2}\|\boldsymbol{\delta}\|_2^2, \quad \forall\,\boldsymbol{\delta}\,\text{such that}\,\|\boldsymbol{\delta}\|_1 \le 2\sqrt{s}\,\|\boldsymbol{\delta}\|_2\right\}.\end{aligned}$$

Since $s < s^* \le 2s$, Lemma B.5 and Equations (112), (113), and (116) ensure that $\mathbb{P}(E_{m,n}) \to 1$. Note that $\lambda_{n,s'}^{(t)}$ increases in s', which implies that, under $E_{m,n}$,

$$\left\|V_{m,1}\boldsymbol{\beta}^* - \left(V_{m,1}\widehat{\boldsymbol{\beta}}^{(0)} - U_n\right)\right\|_\infty \le \lambda_{n,s'}^{(1)},$$

for any $\lambda_{n,s'}^{(1)}$ with $s' \ge s^*$. Therefore, following the proof of Theorem 5.1, we have that, under $E_{m,n}$,

$$\left\|\widehat{\boldsymbol{\beta}}_{s'}^{(1)} - \boldsymbol{\beta}^*\right\|_2 \lesssim \sqrt{s'}\lambda_{n,s'}^{(1)} \quad \text{and} \quad \left\|\widehat{\boldsymbol{\beta}}_{s'}^{(1)} - \boldsymbol{\beta}^*\right\|_1 \lesssim s'\lambda_{n,s'}^{(1)},$$

for all $s' \ge s^*$, which further implies

$$\left\|\widehat{\boldsymbol{\beta}}_{s^*}^{(1)} - \widehat{\boldsymbol{\beta}}_{s'}^{(1)}\right\|_2 \le \left\|\widehat{\boldsymbol{\beta}}_{s^*}^{(1)} - \boldsymbol{\beta}^*\right\|_2 + \left\|\widehat{\boldsymbol{\beta}}_{s'}^{(1)} - \boldsymbol{\beta}^*\right\|_2 \lesssim \sqrt{s'}\lambda_{n,s'}^{(1)},$$

and

$$\left\|\widehat{\boldsymbol{\beta}}_{s^*}^{(1)} - \widehat{\boldsymbol{\beta}}_{s'}^{(1)}\right\|_1 \le \left\|\widehat{\boldsymbol{\beta}}_{s^*}^{(1)} - \boldsymbol{\beta}^*\right\|_1 + \left\|\widehat{\boldsymbol{\beta}}_{s'}^{(1)} - \boldsymbol{\beta}^*\right\|_1 \lesssim s'\lambda_{n,s'}^{(1)}.$$

By the definition of $\widehat{s}^{(1)}$, we obtain that $\widehat{s}^{(1)} \le s^*$, and hence

$$\left\|\widehat{\boldsymbol{\beta}}_{\widehat{s}^{(1)}}^{(1)} - \widehat{\boldsymbol{\beta}}_{s^*}^{(1)}\right\|_2 \lesssim \sqrt{s^*}\lambda_{n,s^*}^{(1)} \asymp \delta_{m,1} \quad \text{and} \quad \left\|\widehat{\boldsymbol{\beta}}_{\widehat{s}^{(1)}}^{(1)} - \widehat{\boldsymbol{\beta}}_{s^*}^{(1)}\right\|_1 \lesssim s^*\lambda_{n,s^*}^{(1)} \asymp \sqrt{s}\delta_{m,1},$$

since $s^* \asymp s$. Adding that $\left\|\widehat{\boldsymbol{\beta}}_{s^*}^{(1)} - \boldsymbol{\beta}^*\right\|_2 \lesssim \sqrt{s^*}\lambda_{n,s^*}^{(1)}$ and $\left\|\widehat{\boldsymbol{\beta}}_{s^*}^{(1)} - \boldsymbol{\beta}^*\right\|_1 \lesssim s^*\lambda_{n,s^*}^{(1)}$, we complete the proof of (34) for $t = 1$. The proof for $t \ge 2$ is straightforward by combining the proof for Theorem A.1 and that for $t = 1$. □

PROOF OF THEOREM A.3. First we prove the theorem for $t = 1$. Recall that $h^* = (\log p/n)^{1/(2\alpha+1)}$ and define $h^{**} = 2^{\lfloor \log_2(h^*) \rfloor - 1} \in \mathcal{D}$. Then $h^*/2 < h^{**} \leq h^*$. For any $h' \in \mathcal{D}$, by the proof of Lemma B.5, Equations (93)–(98) hold for h' with probability at least $1 - 12/p$, which, by the proof of Theorem 5.1, further implies that

$$\left\| \widehat{\boldsymbol{\beta}}_{h'}^{(1)} - \boldsymbol{\beta}^* \right\|_2 \lesssim \sqrt{s} \lambda_{n,h'}^{(1)} \quad \text{and} \quad \left\| \widehat{\boldsymbol{\beta}}_{h'}^{(1)} - \boldsymbol{\beta}^* \right\|_1 \lesssim s \lambda_{n,h'}^{(1)}. \tag{118}$$

Therefore, since the cardinality of $\mathcal{D}$ is less than $\log_2(m)$, we have that, with probability at least $1 - 12\log_2(m)/p$, Equation (118) holds for all $h' \in \mathcal{D}$. In particular, when $h' \leq h^{**}$, we have that

$$\left\| \widehat{\boldsymbol{\beta}}_{h^{**}}^{(1)} - \widehat{\boldsymbol{\beta}}_{h'}^{(1)} \right\|_2 \leq \left\| \widehat{\boldsymbol{\beta}}_{h^{**}}^{(1)} - \boldsymbol{\beta}^* \right\|_2 + \left\| \widehat{\boldsymbol{\beta}}_{h'}^{(1)} - \boldsymbol{\beta}^* \right\|_2 \lesssim \sqrt{s} \lambda_{n,h'}^{(1)},$$

and

$$\left\| \widehat{\boldsymbol{\beta}}_{h^{**}}^{(1)} - \widehat{\boldsymbol{\beta}}_{h'}^{(1)} \right\|_1 \leq \left\| \widehat{\boldsymbol{\beta}}_{h^{**}}^{(1)} - \boldsymbol{\beta}^* \right\|_1 + \left\| \widehat{\boldsymbol{\beta}}_{h'}^{(1)} - \boldsymbol{\beta}^* \right\|_1 \lesssim s \lambda_{n,h'}^{(1)},$$

where we use the fact that $\sqrt{\frac{\log p}{nh'}} \geq (h')^\alpha$ for all $h' \leq h^{**} \leq h^*$, and thus

$$\begin{aligned} \lambda_{n,h'}^{(t)} &\asymp \left(\sqrt{\frac{\log p}{nh'}} + (h')^\alpha + \frac{1}{\sqrt{s}} (r_{m,h'})^t \delta_{m,0} \right) \\ &\asymp \left(\sqrt{\frac{\log p}{nh'}} + \frac{1}{\sqrt{s}} (r_{m,h'})^t \delta_{m,0} \right), \end{aligned}$$

which decreases in h'. By the definition of $\widehat{h}^{(1)}$, we obtain that $\widehat{h}^{(1)} \geq h^{**}$ and

$$\left\| \widehat{\boldsymbol{\beta}}_{\widehat{h}^{(1)}}^{(1)} - \widehat{\boldsymbol{\beta}}_{h^{**}}^{(1)} \right\|_2 \lesssim \sqrt{s} \lambda_{n,h^{**}}^{(1)} \asymp \delta_{m,1}, \left\| \widehat{\boldsymbol{\beta}}_{\widehat{h}^{(1)}}^{(1)} - \widehat{\boldsymbol{\beta}}_{h^{**}}^{(1)} \right\|_1 \lesssim s \lambda_{n,h^{**}}^{(1)} \asymp \sqrt{s} \delta_{m,1},$$

since $h^{**} \asymp h^*$. Together with $\left\| \widehat{\boldsymbol{\beta}}_{h^{**}}^{(1)} - \boldsymbol{\beta}^* \right\|_2 \lesssim \sqrt{s} \lambda_{n,h^{**}}^{(1)}$ and $\left\| \widehat{\boldsymbol{\beta}}_{h^{**}}^{(1)} - \boldsymbol{\beta}^* \right\|_1 \lesssim s \lambda_{n,h^{**}}^{(1)}$, we complete the proof of (36) for $t = 1$. The proof for $t \geq 2$ is straightforward by combining the proof for Theorem A.1 and that for $t = 1$. □

APPENDIX C: DISCUSSIONS ON THE SUPER-EFFICIENCY PHENOMENON

In this section, we show that our estimator $\widehat{\boldsymbol{\beta}}^{(T)}$ achieves the same asymptotic performance over a class of underlying distributions under certain uniform assumptions. In model (1), for any $\boldsymbol{\beta}^*$, denote the density function of $\zeta := X + \boldsymbol{Z}^\top \boldsymbol{\beta}^*$ conditional on $\boldsymbol{Z}$ by $\rho(\cdot \mid \boldsymbol{Z})$ and the cumulative distribution function of ϵ conditional on $\boldsymbol{Z}$ by $F(\cdot \mid \boldsymbol{Z})$. We define the distribution class Θ to be the set of tuples $(\boldsymbol{\beta}^*, \rho, F)$ that satisfy the following assumptions:

ASSUMPTION 14. *Assume that for all $(\boldsymbol{\beta}^*, \rho, F) \in \Theta$ and all integers $1 \leq k \leq \alpha$, the k-th order derivative of $\rho(\cdot \mid \boldsymbol{Z})$ exists for almost every $\boldsymbol{Z}$. Furthermore, there exists a constant $C_{\Theta,1} > 0$ such that $\sup_{\zeta,\boldsymbol{Z},k} \left| \rho^{(k)}(\zeta \mid \boldsymbol{Z}) \right| < C_{\Theta,1}$.*

ASSUMPTION 15. *Assume that ϵ and X are independent given $\boldsymbol{Z}$, and for all $(\boldsymbol{\beta}^*, \rho, F) \in \Theta$ and all integers $1 \leq k \leq \alpha + 1$, the k-th order derivative of $F(\cdot \mid \boldsymbol{Z})$ exists for almost every $\boldsymbol{Z}$. Furthermore, there exists a constant $C_{\Theta,2} > 0$ such that $\sup_{\epsilon,\boldsymbol{Z},k} \left| F^{(k)}(\epsilon \mid \boldsymbol{Z}) \right| < C_{\Theta,2}$.*

TABLE 1

The bias, variance and coverage rates of `(mSMSE)` *($t = 1, 2, 3$),* `(Avg-MSE)`, `(Avg-SMSE)` *and pooled-SMSE, with $p = 1$, $\log_m(n)$ from 1.5 to 1.9 and homoscedastic normal noise.*

$\log_m(n)$	Bias ($\times 10^{-2}$)	Variance ($\times 10^{-4}$)	Coverage Rate	Bias ($\times 10^{-2}$)	Variance ($\times 10^{-4}$)	Coverage Rate	Bias ($\times 10^{-2}$)	Variance ($\times 10^{-4}$)	Coverage Rate
	`(mSMSE)` $t = 1$			`(mSMSE)` $t = 3$			`(Avg-SMSE)`		
1.5	−0.19	2.52	0.92	0.21	1.29	0.93	−0.20	1.18	0.95
1.6	−0.47	3.92	0.85	0.07	0.88	0.92	−0.57	0.71	0.86
1.7	−0.45	2.11	0.84	0.03	0.38	0.95	−1.02	0.32	0.60
1.8	−0.23	1.45	0.86	0.12	0.27	0.96	−1.46	0.21	0.13
1.9	−0.32	1.91	0.83	0.04	0.12	0.96	−1.95	0.08	0.00
	`(mSMSE)` $t = 2$			`(Avg-MSE)`			pooled-SMSE		
1.5	0.19	1.32	0.92	−1.32	1.70	0.84	0.20	1.29	0.93
1.6	0.10	0.97	0.85	−1.39	1.04	0.69	0.07	0.87	0.93
1.7	0.03	0.41	0.84	−1.35	0.47	0.48	0.02	0.38	0.95
1.8	0.11	0.30	0.86	−1.26	0.25	0.28	0.11	0.27	0.96
1.9	0.02	0.13	0.83	−1.32	0.09	0.01	0.03	0.12	0.96

ASSUMPTION 16. *Assume that there exists a constant $c_\Theta > 1$ such that, for all $(\boldsymbol{\beta}^*, \rho, F) \in \Theta$, the matrices $V = 2\mathbb{E}\left[\rho(0 \mid \boldsymbol{Z}) F'(0 \mid \boldsymbol{Z}) \boldsymbol{Z}\boldsymbol{Z}^\top\right]$ and $V_s = \pi_V \mathbb{E}\left[\rho(0 \mid \boldsymbol{Z}) \boldsymbol{Z}\boldsymbol{Z}^\top\right]$ satisfy $c_\Theta^{-1} < \Lambda_{\min}(V) < \Lambda_{\max}(V) < c_\Theta$, $c_\Theta^{-1} < \Lambda_{\min}(V_s) < \Lambda_{\max}(V_s) < c_\Theta$, where $\Lambda_{\min}$ ($\Lambda_{\max}$) denotes the minimum (maximum) eigenvalue.*

Assumptions 14–16 for the distribution class Θ are parallel to Assumptions 2–4 for a fixed distribution. Assumptions 14–15 require α-order smoothness for all ρ and F, which ensures that the Taylor's expansion in the technical proof always hold when n is sufficiently large. Furthermore, the constants $C_{\Theta,1}$ and $C_{\Theta,2}$ provide uniform upper bounds for the derivatives of ρ and F over Θ. Similarly, Assumption 16 ensures that the population Hessian matrix is always positive semi-definite with eigenvalues uniformly bounded away from 0 and ∞. Under these assumptions, replicating the analysis of `(mSMSE)` in Section 3 leads to the following result:

THEOREM C.1. *Assume Assumptions 1, 5, 14, 15 and 16 hold, and there exists a constant $0 < c_2 < 1$ such that $p = O(m^{c_2})$. Further assume that $\sup_{(\boldsymbol{\beta}^*,\rho,F)\in\Theta} \left\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\right\|_2 = O_{\mathbb{P}}((p/m)^{1/3})$. When T satisfies (13), by choosing $h_t = \max\left\{(p/n)^{\frac{1}{2\alpha+1}}, (p/m)^{\frac{2^t}{3\alpha}}\right\}$ at iteration $t = 1, 2, \ldots, T$, we have: $\forall \varepsilon > 0$, $\exists M_\varepsilon, N_\varepsilon$, such that $\forall n \geq N_\varepsilon$, it holds that*

$$\sup_{(\boldsymbol{\beta}^*,\rho,F)\in\Theta} \mathbb{P}\left(\left\|\widehat{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\right\|_2 > M_\varepsilon (p/n)^{\frac{\alpha}{2\alpha+1}}\right) < \varepsilon. \tag{119}$$

Note that (119) is equivalent to $\left\|\widehat{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\right\|_2 = O_{\mathbb{P}}((p/n)^{\frac{\alpha}{2\alpha+1}})$ if Θ only contains a single distribution. The proof of Theorem C.1 is almost the same as the proof of Proposition B.1 and Theorem 3.3, by noting that, for all distributions in Θ, the Taylor's expansion in (55) and the computation related to the bias in (56) are always correct. Moreover, the constant in the big O notation in (58) is uniform for all distributions in Θ, which is guaranteed by Assumptions 14 and 15.

APPENDIX D: ADDITIONAL RESULTS IN SIMULATIONS

D.1. Bias and Variance. In this section, we report the bias and the variance of the estimators with different $\log_m(n)$ in Tables 1 and 2. In the tables, when $t \geq 3$, both the bias and the variance of our proposed multiround method generally decrease as n increases, and they are close to the bias and the variance of the pooled-SMSE. This is consistent with our theoretical analysis in Section 3.2, where we establish that the bias and the variance of `(mSMSE)`

TABLE 2

The bias, variance and coverage rates of (mSMSE) *($t = 1, 2, 3, 4$),* (Avg-SMSE) *and pooled-SMSE, with $p = 10$, $\log_m(n)$ from 1.5 to 1.9 and homoscedastic normal noise.*

$\log_m(n)$	Bias ($\times 10^{-2}$)	Variance ($\times 10^{-4}$)	Coverage Rate	Bias ($\times 10^{-2}$)	Variance ($\times 10^{-4}$)	Coverage Rate	Bias ($\times 10^{-2}$)	Variance ($\times 10^{-4}$)	Coverage Rate
	(mSMSE) $t = 1$			(mSMSE) $t = 3$			(Avg-SMSE)		
1.5	−7.00	114.33	0.66	0.54	14.50	0.93	−0.39	7.94	0.95
1.6	−9.33	119.20	0.44	0.69	8.43	0.93	−1.77	4.43	0.89
1.7	−9.68	130.22	0.34	0.39	5.97	0.95	−2.94	2.08	0.57
1.8	−10.71	140.50	0.23	0.32	3.63	0.90	−3.97	1.19	0.10
1.9	−11.56	151.23	0.12	0.12	1.39	0.94	−4.73	0.65	0.01
	(mSMSE) $t = 2$			(mSMSE) $t = 4$			pooled-SMSE		
1.5	0.64	29.11	0.86	1.22	9.83	0.95	1.21	9.74	0.95
1.6	0.65	16.98	0.84	0.94	5.55	0.94	0.97	5.51	0.95
1.7	0.22	12.94	0.80	0.76	2.56	0.98	0.77	2.56	0.98
1.8	0.47	7.67	0.80	0.60	1.92	0.93	0.60	1.91	0.93
1.9	−0.11	4.69	0.74	0.27	0.89	0.97	0.24	0.88	0.97

TABLE 3

The coverage rates (nominal 95%) of (mSMSE) *with different values of λ_h and $\log_m(n)$. The noise is homoscedastic normal.*

λ_h	$\log_m(n)$	$p = 1$	$p = 10$	λ_h	$\log_m(n)$	$p = 1$	$p = 10$
	1.5	0.94	0.93		1.5	0.95	0.92
1	1.7	0.93	0.92	10	1.7	0.94	0.94
	1.9	0.96	0.94		1.9	0.96	0.92
	1.5	0.93	0.91		1.5	0.94	0.93
30	1.7	0.95	0.95	$\widehat{\lambda_h^*}$	1.7	0.94	0.94
	1.9	0.94	0.95		1.9	0.96	0.92

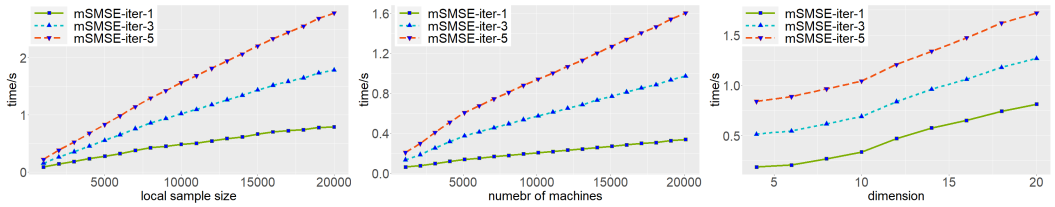
TABLE 4

The CPU time (in seconds) that different methods take to compute the estimator, with $p = 10$, $\log_m(n)$ from 1.5 to 1.9, and homoscedastic normal noise.

$\log_m(n)$	(mSMSE) $t = 2$	(mSMSE) $t = 3$	(Avg-SMSE)	pooled-SMSE
1.5	0.086	0.123	0.121	0.264
1.6	0.092	0.128	0.127	0.511
1.7	0.106	0.152	0.146	1.025
1.8	0.139	0.200	0.148	1.847
1.9	0.195	0.279	0.156	3.782

are both of the rate $n^{-\alpha/(2\alpha+1)}$. On the contrary, the biases of the averaging methods are much larger than the bias of our method and stay large as n increases, as the bias cannot be reduced by averaging in a distributed environment. Note that the bias of (Avg-SMSE) is high since the necessary condition $L = \left(m^{\frac{2(\alpha-1)}{3}}/(p\log m)^{\frac{2\alpha+1}{3}}\right)$ in Theorem 3.1 is violated. While the bias stays large, the variance decreases as n increases, and therefore we observe the failure of inference when n is large for (Avg-MSE) and (Avg-SMSE).

D.2. Sensitivity Analysis. In this section, we use numerical experiments to show the sensitivity of the constant λ_h in the bandwidth h_t in Theorem 3.4. An expression of the optimal value λ_h^* is given in (15) by minimizing the asymptotic mean squared error. We estimate λ_h^* using $\widehat{U}$, $\widehat{V}$, and $\widehat{V}_s$. Under our experiment settings, the estimated constant $\widehat{\lambda_h^*}$ ranges from 10 to 25 in practice. To study the effect of λ_h on the validity of inference, we choose a wider range for λ_h, from 1 to 30, and report the coverage rates of (mSMSE) in Table 3 for $p = 1$ and 10, with different λ_h and $\log_m(n)$. In summary, (mSMSE) generally allows arbitrary choices of λ_h in a wide range, which suggests that our proposed (mSMSE) algorithm is robust with respect to λ_h.



(a) $p = 10$, $L = 50$ (b) $p = 10$, $m = 100$ (c) $m = 5000$, $L = 500$

Fig 1: The CPU time (in seconds) of `(mSMSE)` as we increase m, L, p in subfigures (a), (b), (c), respectively.

D.3. Time Complexity. In this section, we compare the computational complexity of each method. The average CPU time that each method takes when $p = 10$ is reported in Table 4. The computation time is recorded in a simulated distributed environment on a RedHat Enterprise Linux cluster containing 524 Lenovo SD650 nodes interconnected by high-speed networks. On each computer node, two Intel Xeon Platinum 8268 24C 205W 2.9GHz Processors are equipped with 48 processing cores.

In Table 4, we first notice that the speed of `(mSMSE)` is much faster than the pooled estimator, and the discrepancy greatly increases when n gets larger. Second, the computation time of `(mSMSE)` is comparable to `(Avg-SMSE)`. This result may seem counterintuitive since `(mSMSE)` still requires running an SMSE on the first machine for the initial estimator. However, since the computation time of `(Avg-SMSE)` is mainly determined by the maximum computation time of the L local machines, `(Avg-SMSE)` greatly suffers from the computational performance of the "worst" machine, especially when the number of machines is large. On the other hand, `(mSMSE)` only runs SMSE on one machine and therefore achieves comparable computation time in the experiments.

Now we focus on our `(mSMSE)` and demonstrate the computational complexity as we increase the local sample size m, number of machines L, and dimension p. The computational cost of Algorithm 1 is composed of four parts: (a) computing the initial estimator in step 1, whose time complexity depends on what initial is used, (b) computing the local gradient and Hessian in step 5, whose time complexity is $O(mp^2)$ per iteration, (c) computing global gradient and Hessian in step 8, whose time complexity is $O(Lp^2)$ per iteration, and (d) the Newton's update in step 9, whose time complexity is $O(p^3)$ per iteration. We note that p is smaller than m, and therefore the computation cost of `(mSMSE)` is dominated by (b) and (c), which is scalable with m, L, and p^2. To verify the computational scalability, we run numerical simulations for increasing m, L, and p, and present the computation time of `(mSMSE)` with one, three, and five iterations in Figure 1, which verifies that the computation time increases linearly in m and L and superlinearly in p.

D.4. Results for Non-Gaussian Noises. In this section, we report the bias, variance and coverage rates in Tables 5–8 for the other two noise types, i.e., the homoscedastic uniform and heteroscedastic normal noise, with $p = 1$ and 10. From these tables, we can still see the failure of inference of `(Avg-MSE)` and `(Avg-SMSE)` when $\log_m(n)$ is large, while the `(mSMSE)` method with $t \geq 3$ achieves near-nominal coverage rates no matter how large $\log_m(n)$ is. In addition, we also report the time cost of each method in Table 9, which shows that the computational time of `(mSMSE)` is comparable to `(Avg-SMSE)`. These findings are all consistent with the results for the homoscedastic normal noise.

TABLE 5

The bias, variance and coverage rates of (mSMSE) *($t = 1, 2, 3$),* (Avg-MSE), (Avg-SMSE) *and pooled-SMSE, with $p = 1$, $\log_m(n)$ from 1.5 to 1.9 and homoscedastic uniform noise.*

$\log_m(n)$	Bias ($\times 10^{-2}$)	Variance ($\times 10^{-4}$)	Coverage Rate	Bias ($\times 10^{-2}$)	Variance ($\times 10^{-4}$)	Coverage Rate	Bias ($\times 10^{-2}$)	Variance ($\times 10^{-4}$)	Coverage Rate
	(mSMSE) $t = 1$			(mSMSE) $t = 3$			(Avg-SMSE)		
1.5	−0.17	5.16	0.90	0.17	2.37	0.93	−0.89	1.53	0.95
1.6	−0.40	5.91	0.81	0.21	1.42	0.90	−1.38	1.02	0.74
1.7	−0.69	7.06	0.84	0.09	0.80	0.94	−2.11	0.48	0.23
1.8	−0.27	2.13	0.86	0.15	0.41	0.94	−2.78	0.23	0.00
1.9	−0.14	0.83	0.86	0.08	0.22	0.97	−3.42	0.13	0.00
	(mSMSE) $t = 2$			(Avg-MSE)			pooled-SMSE		
1.5	0.15	2.63	0.90	−1.37	2.70	0.81	0.15	2.35	0.93
1.6	0.19	1.68	0.81	−1.26	1.27	0.78	0.20	1.41	0.90
1.7	0.10	1.08	0.84	−1.40	0.59	0.57	0.08	0.80	0.94
1.8	0.13	0.43	0.86	−1.25	0.33	0.41	0.13	0.41	0.94
1.9	0.09	0.23	0.86	−1.29	0.21	0.16	0.07	0.22	0.97

TABLE 6

The bias, variance and coverage rates of (mSMSE) *($t = 1, 2, 3, 4$),* (Avg-SMSE) *and pooled-SMSE, with $p = 10$, $\log_m(n)$ from 1.5 to 1.9 and homoscedastic uniform noise.*

$\log_m(n)$	Bias ($\times 10^{-2}$)	Variance ($\times 10^{-4}$)	Coverage Rate	Bias ($\times 10^{-2}$)	Variance ($\times 10^{-4}$)	Coverage Rate	Bias ($\times 10^{-2}$)	Variance ($\times 10^{-4}$)	Coverage Rate
	(mSMSE) $t = 1$			(mSMSE) $t = 3$			(Avg-SMSE)		
1.5	−7.75	129.45	0.65	0.52	23.40	0.91	−2.00	12.00	0.97
1.6	−10.50	211.98	0.49	0.12	15.64	0.91	−3.84	4.68	0.89
1.7	−8.79	120.42	0.43	0.43	8.29	0.92	−5.34	2.79	0.37
1.8	−10.10	196.01	0.29	0.32	3.91	0.92	−6.38	1.15	0.02
1.9	−8.48	96.22	0.23	0.39	2.69	0.92	−7.18	0.71	0.00
	(mSMSE) $t = 2$			(mSMSE) $t = 4$			pooled-SMSE		
1.5	0.02	44.75	0.82	1.07	19.06	0.91	1.03	18.97	0.92
1.6	−1.73	36.80	0.77	0.50	9.83	0.94	0.52	9.75	0.95
1.7	−1.08	21.49	0.78	0.58	5.94	0.95	0.54	5.89	0.95
1.8	−0.56	12.31	0.77	0.53	3.37	0.92	0.43	3.37	0.93
1.9	−0.29	8.00	0.80	0.47	2.01	0.94	0.39	2.00	0.95

TABLE 7

The bias, variance and coverage rates of (mSMSE) *($t = 1, 2, 3$),* (Avg-MSE), (Avg-SMSE) *and pooled-SMSE, with $p = 1$, $\log_m(n)$ from 1.5 to 1.9 and heteroscedastic normal noise.*

$\log_m(n)$	Bias ($\times 10^{-2}$)	Variance ($\times 10^{-4}$)	Coverage Rate	Bias ($\times 10^{-2}$)	Variance ($\times 10^{-4}$)	Coverage Rate	Bias ($\times 10^{-2}$)	Variance ($\times 10^{-4}$)	Coverage Rate
	(mSMSE) $t = 1$			(mSMSE) $t = 3$			(Avg-SMSE)		
1.5	0.02	2.11	0.90	0.35	1.41	0.92	−0.08	1.25	0.92
1.6	−0.25	3.23	0.86	0.19	0.70	0.94	−0.54	0.75	0.84
1.7	−0.25	1.90	0.84	0.07	0.39	0.95	−1.01	0.30	0.56
1.8	−0.06	0.58	0.85	0.11	0.27	0.90	−1.46	0.16	0.08
1.9	−0.22	0.70	0.78	0.06	0.15	0.92	−1.80	0.08	0.00
	(mSMSE) $t = 2$			(Avg-MSE)			pooled-SMSE		
1.5	0.34	1.42	0.90	−1.25	1.95	0.80	0.34	1.41	0.92
1.6	0.19	0.80	0.86	−1.41	1.09	0.64	0.19	0.69	0.94
1.7	0.05	0.41	0.84	−1.44	0.43	0.42	0.06	0.39	0.95
1.8	0.11	0.27	0.85	−1.46	0.23	0.15	0.10	0.27	0.90
1.9	0.06	0.16	0.78	−1.40	0.11	0.00	0.05	0.15	0.93

D.5. Results for $p = 20$. In this section, we report the performance of (Avg-SMSE) and (mSMSE) under local size $m = 2000$ and dimension $p = 20$. Figure 2 presents the coverage rates as a function of $\log_m(n)$ with $p = 20$. Our proposed (mSMSE), as well as the pooled estimator, achieves a high coverage rate around 95% no matter how large $\log_m(n)$ is, while the averaging methods both fail when $\log_m(n)$ is large. Table 10 reports the bias and the variance of the estimators with different $\log_m(n)$. When $t = 4$, both the bias and the variance of our proposed multiround method generally decrease as n increases, and they are close to the bias and the variance of the pooled-SMSE. These findings are all consistent with the results for $p = 1$ and $p = 10$.

TABLE 8

The bias, variance and coverage rates of `(mSMSE)` *($t = 1, 2, 3, 4$),* `(Avg-SMSE)` *and pooled-SMSE, with $p = 10$, $\log_m(n)$ from 1.5 to 1.9 and heteroscedastic normal noise.*

$\log_m(n)$	Bias ($\times 10^{-2}$)	Variance ($\times 10^{-4}$)	Coverage Rate	Bias ($\times 10^{-2}$)	Variance ($\times 10^{-4}$)	Coverage Rate	Bias ($\times 10^{-2}$)	Variance ($\times 10^{-4}$)	Coverage Rate
	`(mSMSE)` $t=1$			`(mSMSE)` $t=3$			`(Avg-SMSE)`		
1.5	−5.32	51.33	0.54	0.67	6.80	0.91	1.35	4.11	0.94
1.6	−5.42	54.03	0.42	0.42	5.15	0.92	0.35	2.25	0.94
1.7	−4.95	41.12	0.38	0.34	2.90	0.90	−0.34	1.40	0.90
1.8	−5.17	61.55	0.34	−0.07	5.18	0.88	−0.92	0.61	0.74
1.9	−6.65	75.74	0.22	−0.27	3.92	0.87	−1.35	0.30	0.28
	`(mSMSE)` $t=2$			`(mSMSE)` $t=4$			pooled-SMSE		
1.5	1.67	9.03	0.87	1.12	3.53	0.94	1.12	3.53	0.94
1.6	1.17	6.30	0.83	0.82	1.92	0.94	0.82	1.92	0.94
1.7	1.12	7.13	0.74	0.70	1.17	0.93	0.69	1.17	0.94
1.8	1.21	8.72	0.77	0.50	0.65	0.92	0.45	0.65	0.94
1.9	1.28	8.44	0.68	0.29	0.36	0.94	0.28	0.36	0.95

TABLE 9

The cpu times (in seconds) that different methods take to compute the estimator, with $p = 10$, $\log_m(n)$ from 1.5 to 1.9 and two types of noise.

Noise Type	$\log_m(n)$	`(mSMSE)` $t=2$	`(mSMSE)` $t=3$	`(Avg-SMSE)`	pooled-SMSE
Homoscedastic Uniform	1.5	0.091	0.126	0.111	0.323
	1.6	0.094	0.133	0.125	0.634
	1.7	0.109	0.154	0.145	1.276
	1.8	0.141	0.202	0.148	2.359
	1.9	0.196	0.282	0.158	4.784
Heteroscedastic Normal	1.5	0.088	0.122	0.113	0.984
	1.6	0.090	0.126	0.132	2.157
	1.7	0.106	0.156	0.144	4.768
	1.8	0.139	0.198	0.145	10.688
	1.9	0.190	0.276	0.156	22.036

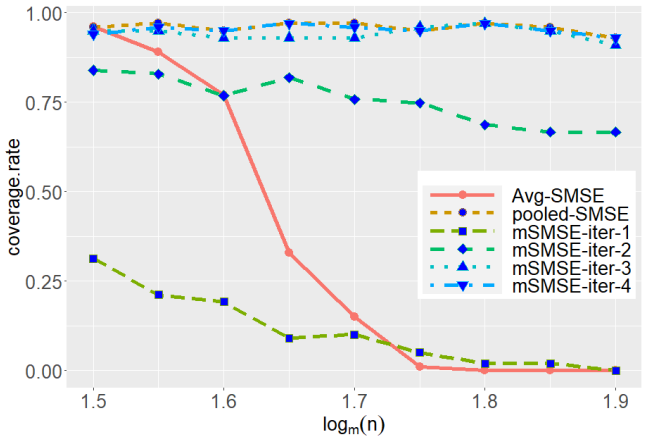


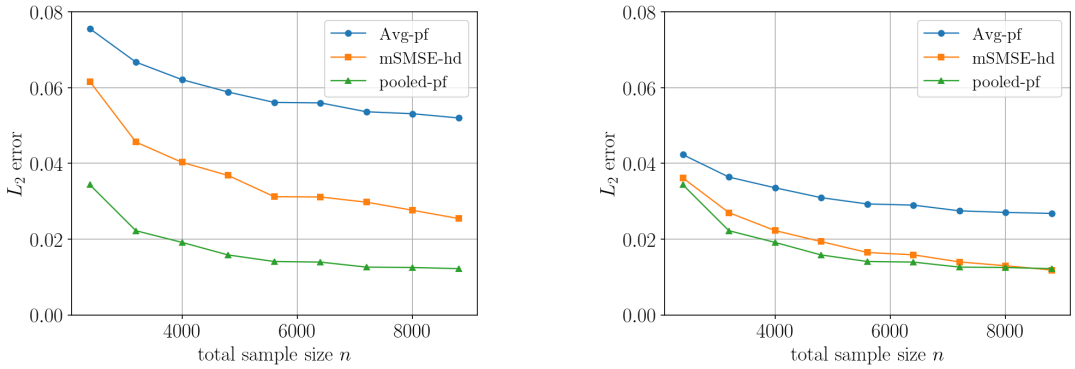
Fig 2: Coverage rates for different methods with $p = 20$ and homoscedastic normal noise.

D.6. Results for High-Dimensional Simulations. In this section, we present the performance of Algorithm 3 using high-dimensional simulations. Following the settings in Feng et al. (2022), we first generate the covariates $(x_i, \boldsymbol{z}_i) \sim \mathcal{N}(0, \Sigma_{0.5})$, where $\Sigma_{0.5}$ is an AR(1) covariance matrix with correlation coefficient 0.5. The parameter of interest $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is set to be $(\underbrace{1/\sqrt{s}, \ldots, 1/\sqrt{s}}_{s \text{ entries}}, 0, \ldots, 0)^\top$, where we fix the dimension $p = 500$ and the sparsity $s = 10$. The responses are generated by $y_i = \text{sign}\big(x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}^* + \epsilon_i\big)$ for $i = 1, 2, \ldots, n$, with $\epsilon_i \sim \mathcal{N}(0, (0.5)^2)$. The n observations are then evenly divided into $L = n/m$ subsets, where we choose the local sample size $m \in \{400, 800\}$, and vary the total sample size

TABLE 10

The bias, variance and coverage rates of `(mSMSE)` *($t = 1, 2, 3, 4$),* `(Avg-SMSE)` *and pooled-SMSE, with $p = 20$, $\log_m(n)$ from 1.5 to 1.9 and homoscedastic normal noise.*

$\log_m(n)$	Bias ($\times 10^{-2}$)	Variance ($\times 10^{-4}$)	Coverage Rate	Bias ($\times 10^{-2}$)	Variance ($\times 10^{-4}$)	Coverage Rate	Bias ($\times 10^{-2}$)	Variance ($\times 10^{-4}$)	Coverage Rate
	`(mSMSE)` $t=1$			`(mSMSE)` $t=3$			`(Avg-SMSE)`		
1.5	−14.78	215.96	0.31	0.97	15.58	0.96	−0.58	5.49	0.96
1.6	−13.39	151.88	0.20	0.71	4.83	0.93	−2.41	2.42	0.77
1.7	−14.47	138.67	0.10	0.45	5.73	0.93	−3.83	1.38	0.15
1.8	−15.57	142.76	0.02	0.49	1.08	0.97	−5.13	0.72	0.00
1.9	−16.12	125.10	0.00	0.45	1.33	0.91	−5.71	0.30	0.00
	`(mSMSE)` $t=2$			`(mSMSE)` $t=4$			pooled-SMSE		
1.5	1.22	18.83	0.84	1.77	6.55	0.95	1.79	6.49	0.96
1.6	0.73	17.80	0.77	1.11	3.53	0.95	1.08	3.52	0.95
1.7	−0.07	8.09	0.76	0.92	1.97	0.96	0.86	1.94	0.97
1.8	−0.53	3.63	0.68	0.61	0.95	0.97	0.54	0.95	0.97
1.9	−0.25	3.80	0.67	0.58	0.76	0.93	0.50	0.76	0.93



(a) $m = 400$ (b) $m = 800$

Fig 3: The L_2 estimation errors of different methods in the high-dimensional setting.

$n \in [2400, 8800]$. On the simulated datasets, we compare the L_2 estimation errors of the following three algorithms:

(1) "mSMSE-hd": our proposed high-dimensional mSMSE algorithm in Algorithm 3;
(2) "Avg-pf": the Averaged Divide-and-Conquer path-following algorithm, which applies the path-following algorithm proposed in Feng et al. (2022) on each subset and aggregates the local estimators by averaging;
(3) "pooled-pf": the path-following algorithm using the entire dataset.

For the smoothing kernel in Algorithm 3, we use a higher-order biweight kernel with $\alpha = 6$:

$$H'(x) = \frac{4725}{2048}(1 - x^2)^2\Big(1 - \frac{22}{3}x^2 + \frac{143}{15}x^4\Big)\mathbb{I}(x \le 1).$$

Moreover, in each iteration, the bandwidth h_t and the penalty parameter $\lambda_n^{(t)}$ are chosen through cross-validation. Concretely, we use the score function defined in (2) as a measure and, among a grid of possible values, select the combination $(h_t, \lambda_n^{(t)})$ that achieves the highest cross-validated score. The same strategy is also used for tuning the other two algorithms.

Figure 3 presents the L_2 estimation errors of the three algorithms averaged over 500 independent runs. Clearly, the proposed high-dimensional mSMSE algorithm ("mSMSE-hd") always achieves lower estimation error compared to the Averaged Divide-and-Conquer algorithm ("Avg-pf"), and their difference increases as n grows. In Figure 3b, we also see that the

TABLE 11

The bias, variance and coverage rates of `(Avg−SMSE)`*, with bandwidth chosen by 5-fold cross-validation,* $p = 10$*,* $\log_m(n)$ *from* 1.4 *to* 2.0*, and homoscedastic normal noise.*

$\log_m(n)$	bias ($\times 10^{-2}$)	variance ($\times 10^{-4}$)	coverage rate
1.4	−2.80	34.72	0.95
1.5	−2.89	38.53	0.89
1.6	−3.25	29.01	0.76
1.7	−3.94	15.31	0.57
1.8	−3.18	16.20	0.53
1.9	−1.97	18.70	0.52
2.0	−0.59	21.21	0.45

L_2 errors of "mSMSE-hd" get very close to those of the oracle, the path-following algorithm applied to the pooled data ("pooled-pf").

D.7. More discussions on choosing bandwidth for `(Avg−SMSE)`**.** In this section, we provide more discussions on the choice of bandwidth h for the `(Avg−SMSE)` algorithm. Theorem 3.1 for `(Avg−SMSE)` shows that, under a constraint $L = o\big(m^{\frac{2}{3}(\alpha-1)}/(p\log m)^{\frac{2\alpha+1}{3}}\big)$, `(Avg−SMSE)` achieves the optimal convergence rate $(p/n)^{\alpha/(2\alpha+1)}$. This constraint comes from a condition $\frac{p\log m}{mh^3} = o(1)$, which is necessary to ensure the convergence of the empirical Hessian to the population Hessian of the smoothed objective. Aiming to obtain the optimal convergence rate $(p/n)^{\alpha/(2\alpha+1)}$, one would like to choose the bandwidth $h \asymp h^* = (p/n)^{\frac{1}{2\alpha+1}}$, and the constraint $L = o\big(m^{\frac{2}{3}(\alpha-1)}/(p\log m)^{\frac{2\alpha+1}{3}}\big)$ follows from plugging $h^* = (p/n)^{\frac{1}{2\alpha+1}}$ into $\frac{p\log m}{m(h^*)^3} = o(1)$. This constraint translates to the total sample size $n = mL = o(m^{\frac{2\alpha+1}{3}}/(p\log m)^{\frac{2\alpha+1}{3}})$.

However, in the simulation, we increase the total sample size n beyond this constraint, and therefore the constraint is violated. If we look at the theoretical analysis in this case, there would be an additional bias term of the order

$$O\Big(\sqrt{\frac{p\log m}{m(h^*)^3}}\Big) = O\Big(\sqrt{\frac{n^{\frac{3}{2\alpha+1}}p^{\frac{2\alpha-2}{2\alpha+1}}\log m}{m}}\Big),$$

which increases in n and invalidates the asymptotic normality established in (9). This is reflected in Tables 1 and 2 in Section D.1, where we can see that, as $\log_m(n)$ increases, the bias of `(Avg−SMSE)` becomes larger accompanied by a reduction in the coverage rate. This phenomenon calls for the need to propose a multiround procedure `(mSMSE)` which can completely remove this constraint on L.

Additionally, as discussed in Remark 2, `(Avg−SMSE)` still works for bandwidth $h > h^*$ at the sacrifice of the convergence rate. In particular, when the constraint $\frac{p\log m}{m(h^*)^3} = o(1)$ is violated, we can alternatively select a bandwidth $h > h^*$ such that $\frac{p\log m}{mh^3} = o(1)$, which results in a slower (sub-optimal) convergence rate. In addition, no asymptotic distribution is provided in theory for this sub-optimal bandwidth since the bias term will dominate the error.

In addition to Tables 1 and 2, we provide additional simulations to confirm this phenomenon numerically in Table 11, where we use a scale constant c_h to fine-tune the bandwidth $h = c_h(p/n)^{1/(2\alpha+1)}$. The constant c_h is determined by a five-fold cross-validation (CV). As compared to the results for fixed c_h in Table 2, when $\log_m(n)$ is small, the performance of `(Avg−SMSE)` with fine-tuned c_h (via CV) is comparable to that with fixed c_h. As $\log_m(n)$ goes larger, the fine-tuned c_h leads to smaller bias than the fixed c_h. Nonetheless, the convergence rates still fall down as $\log_m(n)$ goes larger, and moreover, the variances are significantly larger than those with fixed c_h. Therefore, there is no perfect way to choose

bandwidth when the constraint $\frac{p \log m}{m(h^*)^3} = o(1)$ is violated, which necessitates the development of `(mSMSE)`.

REFERENCES

Bousquet, O. (2003). Concentration inequalities for sub-additive functions using the entropy method. In *Stochastic Inequalities and Applications*, pp. 213–247. Springer.

Cai, T. and W. Liu (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association 106*(494), 672–684.

Cai, T. T., C.-H. Zhang, and H. H. Zhou (2010). Optimal rates of convergence for covariance matrix estimation. *Annals of Statistics 38*(4), 2118–2144.

Feng, H., Y. Ning, and J. Zhao (2022). Nonregular and minimax estimation of individualized thresholds in high dimension with binary responses. *Annals of Statistics 50*(4), 2284–2305.

Horowitz, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica 60*(3), 505–531.

Lepskii, O. (1991). On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability & Its Applications 35*(3), 454–466.

Luo, J., Q. Sun, and W.-X. Zhou (2022). Distributed adaptive Huber regression. *Computational Statistics & Data Analysis 169*, 107419.