Supplement to "Distributed Estimation and Inference for Semi-parametric Binary Response Models"

A Extensions: Data Heterogeneity and High-dimensional Settings

In this section, we discuss two natural extensions of proposed estimators `(Avg-SMSE)` and `(mSMSE)`.

First, an important question in distributed environments is the heterogeneity of datasets on different local machines. In Section A.1, we consider heterogeneous datasets with a shared parameter of interest $\boldsymbol{\beta}^\star$ and different distributions of covariates $(X, \boldsymbol{Z})$ across the machines. We show that `(mSMSE)` performs better than `(Avg-SMSE)` with a weaker condition. While handling heterogeneous data, the performance of the divide-and-conquer algorithm relies on the smallest sample size among the local machines, while the `(mSMSE)` method does not rely on such conditions.

We further consider a high-dimension extension in Section A.2 where the parameter of interest $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is a sparse vector with s non-zero elements and $s < n < p$. We modify `(mSMSE)` to adapt the idea of the Dantzig Selector (Candes and Tao, 2007) to reach the convergence rate $\sqrt{s}(\log p/n)^{\frac{\alpha}{2\alpha+1}}$ in a distributed environment, which is very close to the minimax optimal rate $(s \log p/n)^{\frac{\alpha}{2\alpha+1}}$ for the linear binary response model established by Feng et al. (2022) in a non-distributed environment. Compared to the low-dimensional settings above, the algorithm in this setting reduces the per-iteration communication cost to $p \times 1$ vectors.

A.1 Data Heterogeneity

Until now, we assumed homogeneity among the data stored on different machines, which means the distribution of $(x_i, \boldsymbol{z}_i, \epsilon_i)$ are the same for all i. It is of practical interest to consider the heterogeneous setting, since data on different machines may not be identically distributed. Therefore, we establish the theoretical results in the presence of heterogeneity in this section.

First, we remove the restriction that the sample size on each machine is the same. Denote the number of observations on the machine $\mathcal{H}_\ell$ to be m_ℓ, which satisfies $\sum_{\ell=1}^{L} m_\ell = n$. Then we have to modify Assumptions 2–4 for different distributions on different machines.

Assumption 1. *For X and $\boldsymbol{Z}$ on $\mathcal{H}_\ell$, define $\zeta := X + \boldsymbol{Z}^\top \boldsymbol{\beta}^*$, and assume that the conditional distribution density function of ζ, denoted by $\rho_\ell(\cdot \mid \boldsymbol{Z})$, is positive and bounded for almost every $\boldsymbol{Z}$. Further, for any integer $1 \le k \le \alpha$, assume that within a neighborhood of 0, $\rho_\ell^{(k)}(\cdot \mid \boldsymbol{Z})$ exists and is* uniformly *bounded for all ℓ and almost every $\boldsymbol{Z}$, i.e., $\exists M_{\rho,k}$ such that $\sup_{\zeta,\ell} \left|\rho_\ell^{(k)}(\zeta \mid \boldsymbol{Z})\right| \le M_{\rho,k}$.*

Assumption 2. *For $X, \boldsymbol{Z}, \epsilon$ on $\mathcal{H}_\ell$, let $F_\ell(\cdot \mid \boldsymbol{Z})$ denote the conditional distribution function of the noise ϵ, and assume that ϵ and X are independent given $\boldsymbol{Z}$. For any integer $1 \le k \le \alpha+1$, assume that $F_\ell^{(k)}(\cdot \mid \boldsymbol{Z})$ exists and is* uniformly *bounded within a neighborhood of 0 for all ℓ and almost every $\boldsymbol{Z}$, i.e., $\exists M_{F,k}$ such that $\sup_{\epsilon,\ell} \left|F_\ell^{(k)}(\epsilon \mid \boldsymbol{Z})\right| \le M_{F,k}$. Still, we assume $\operatorname{median}(\epsilon \mid \boldsymbol{Z}) = 0$ on each machine.*

Assumption 3. *There exist constants $c_1, c_2 > 0$ such that $c_1^{-1} < \Lambda_{\min}(V_\ell) < \Lambda_{\max}(V_\ell) < c_1$, $c_2^{-1} < \Lambda_{\min}(V_{s,\ell}) < \Lambda_{\max}(V_{s,\ell}) < c_2$, $\forall \ell$, where $V_\ell := 2\mathbb{E}_{\boldsymbol{Z}\in\mathcal{H}_\ell}\left(\rho_\ell(0 \mid \boldsymbol{Z}) F_\ell'(0 \mid \boldsymbol{Z}) \boldsymbol{Z}\boldsymbol{Z}^\top\right)$ and $V_{s,\ell} := \pi_V \mathbb{E}_{\boldsymbol{Z}\in\mathcal{H}_\ell}\left(\rho_\ell(0 \mid \boldsymbol{Z}) \boldsymbol{Z}\boldsymbol{Z}^\top\right)$.*

Assumptions 1–3 are parallel to Assumptions 2–4, requiring the uniform boundedness of the high-order derivatives and the eigenvalues of V_ℓ and $V_{s,\ell}$. Additionally, similar to (10), we define

$$U_\ell := \pi_U \mathbb{E}_{\boldsymbol{Z}\in\mathcal{H}_\ell}\Big(\sum_{k=1}^{\alpha} \frac{2(-1)^{k+1}}{k!\,(\alpha-k)!} F_\ell^{(k)}(0 \mid \boldsymbol{Z})\, \rho_\ell^{(\alpha-k)}(0 \mid \boldsymbol{Z})\, \boldsymbol{Z}\Big),$$

which is related to the bias of SMSE on each machine.

Under the modified assumptions, the data on each machine are no longer identically distributed, and therefore it is natural to allocate a different weight matrix W_ℓ to each machine, with $\sum_{\ell=1}^{L} W_\ell = I_{p\times p}$. Formally, the *weighted-Averaged SMSE* (`wAvg-SMSE`) is defined as follows:

$$\widehat{\boldsymbol{\beta}}_{(\texttt{wAvg-SMSE})} := \sum_{\ell=1}^{L} W_\ell \widehat{\boldsymbol{\beta}}_{\mathrm{SMSE},\ell}, \tag{1}$$

where $\widehat{\boldsymbol{\beta}}_{\mathrm{SMSE},\ell}$ is the SMSE on the ℓ-th machine that minimizes the objective function

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

For the multi-round method, we aim to apply the iterative smoothing to minimize a weighted sum of the objective functions on each machine, i.e., $\sum_{\ell=1}^{L} W_\ell F_{h,\ell}(\boldsymbol{\beta})$, which leads to updating the *weighted* mSMSE (`wmSMSE`) in the t-th step by

$$\widehat{\boldsymbol{\beta}}^{(t)}_{(\texttt{wmSMSE})} = \widehat{\boldsymbol{\beta}}^{(t-1)}_{(\texttt{wmSMSE})} - \Big(\sum_{\ell=1}^{L} W_\ell \nabla^2 F_{h,\ell}\big(\widehat{\boldsymbol{\beta}}^{(t-1)}_{(\texttt{wmSMSE})}\big)\Big)^{-1}\Big(\sum_{\ell=1}^{L} W_\ell \nabla F_{h,\ell}\big(\widehat{\boldsymbol{\beta}}^{(t-1)}_{(\texttt{wmSMSE})}\big)\Big). \tag{3}$$

A natural choice of weights is proportional to the local sample size, i.e., $W_\ell = \frac{m_\ell}{n} I_{p\times p}$. Using this weight, the variances in the asymptotic distribution in (9) and (14) become $\frac{1}{n}\sum_{\ell=1}^{L} m_\ell V_\ell^{-1} V_{s,\ell} V_\ell^{-1}$ and $\big(\frac{1}{n}\sum_{\ell=1}^{L} m_\ell V_\ell\big)^{-1}\big(\frac{1}{n}\sum_{\ell=1}^{L} m_\ell V_{s,\ell}\big)\big(\frac{1}{n}\sum_{\ell=1}^{L} m_\ell V_\ell\big)^{-1}$, respectively, which can be seen as a special case of Theorems A.1 and A.2 below. Nonetheless, such a choice is by no means optimal. We could further decrease both asymptotic variances by choosing a different weight matrix W_ℓ for each machine. To illustrate the choice of weights, we first derive the theoretical results for general weight matrices W_ℓ that satisfy the following restriction:

Assumption 4. *There exist constants $c_w, C_W > 0$ such that $c_W m_\ell/n \le \|W_\ell\|_2 \le C_W m_\ell/n$, with $n = \sum_{\ell=1}^{L} m_\ell$ and $\sum_{\ell=1}^{L} W_\ell = I_{p\times p}$.*

Assumption 4 requires that the 2-norm of W_ℓ is not too far away from m_ℓ/n, violation of which may lead to a low convergence rate. An extreme example is $W_1 = I_{p\times p}$ and $W_2 = \ldots = W_L = \mathbf{0}$, in which case only the data on a single machine will be used. Following the procedures in Sections 3.1 and 3.2, we establish the asymptotic normality for (`wAvg-SMSE`) and (`wmSMSE`) in Theorems A.1 and A.2, whose proof is given in Section C.4 in supplementary material.

Theorem A.1 (wAvg-SMSE)**.** *Suppose Assumptions 1 and 5–4 hold and the sample size of the smallest local batch $\min_\ell m_\ell \gtrsim n^{3/(2\alpha+1)}$. By taking $h = n^{-1/(2\alpha+1)}$, we have*

$$n^{\frac{\alpha}{2\alpha+1}}\Big(\widehat{\boldsymbol{\beta}}_{(\texttt{wAvg-SMSE})} - \boldsymbol{\beta}^*\Big) \xrightarrow{d} \mathcal{N}\Big(\sum_{\ell=1}^{L} W_\ell V_\ell^{-1} U_\ell,\ \sum_{\ell=1}^{L} \frac{n}{m_\ell} W_\ell V_\ell^{-1} V_{s,\ell} V_\ell^{-1} W_\ell^\top\Big). \tag{4}$$

Theorem A.2 (wmSMSE)**.** *Suppose Assumptions 1 and 5–4 hold and $\big\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\big\|_2 = O_{\mathbb{P}}(\delta_0)$. By taking $h_t = \max\big\{n^{-1/(2\alpha+1)}, \delta_0^{2^t/\alpha}\big\}$ at iteration $t = 1, 2, \ldots, T$, we have*

$$n^{\frac{\alpha}{2\alpha+1}}\Big(\widehat{\boldsymbol{\beta}}^{(T)}_{(\texttt{wmSMSE})} - \boldsymbol{\beta}^*\Big) \xrightarrow{d} \mathcal{N}\Big(\overline{V}_W^{-1}\overline{U}_W,\ \overline{V}_W^{-1}\sum_{\ell=1}^{L} \frac{n}{m_\ell} W_\ell V_{s,\ell} W_\ell^\top \overline{V}_W^{-\top}\Big), \tag{5}$$

for sufficiently large T, where $\overline{U}_W := \sum_{\ell=1}^{L} W_\ell U_\ell$, $\overline{V}_W := \sum_{\ell=1}^{L} W_\ell V_\ell$.

Assumptions 3 and 4 ensure the variance matrices in both (4) and (5) are finite. In particular, in the homogeneous setting, $W_\ell = (1/L)I_{p\times p}$, then (4) and (5) are identical to (9) and (14).

Remark 1. *In Theorem A.1, the condition* $\min_\ell m_\ell \gtrsim n^{3/(2\alpha+1)}$ *is placed to ensure* $\log m_\ell / \left(m_\ell h^3\right) = o(1)$ *for all* ℓ*, which is necessary to guarantee the convergence of* $\widehat{\boldsymbol{\beta}}_{\mathrm{SMSE},\ell}$ *to* $\boldsymbol{\beta}^*$ *on each machine. In the homogeneous setting, this is equivalent to the constraint* $L = o\big(m^{\frac{2}{3}(\alpha-1)}\big)$ *in Theorem 3.1. This condition requires the sample size of the smallest batch should increase at a certain rate as* $n \to \infty$*. On the other hand, for* `(wmSMSE)`*, there is no restriction on the smallest local sample size.*

Based on the above results, we are able to artificially choose the weight matrices $\{W_\ell\}$ to minimize the covariance matrices of the two methods in (4) and (5). By choosing $W_\ell^{*,(\texttt{wAvg-SMSE})} = \Big(\sum_{\ell=1}^L m_\ell V_\ell V_{s,\ell}^{-1} V_\ell\Big)^{-1} m_\ell V_\ell V_{s,\ell}^{-1} V_\ell$, both the trace and the Frobenius norm of the variance in (4) are minimized, and the corresponding minimum variance is

$$\Sigma^*_{(\texttt{wAvg-SMSE})} := n\Big(\sum_{\ell=1}^L m_\ell V_\ell V_{s,\ell}^{-1} V_\ell\Big)^{-1}. \tag{6}$$

For `(wmSMSE)`, if one chooses $W_\ell^{*,(\texttt{wmSMSE})} = \Big(\sum_{\ell=1}^L m_\ell V_\ell V_{s,\ell}^{-1}\Big)^{-1} m_\ell V_\ell V_{s,\ell}^{-1}$, the asymptotic variance in (5) will be the same as $\Sigma^*_{(\texttt{wAvg-SMSE})}$ in (6). Therefore, the multi-round method `(wmSMSE)` is at least as efficient as `(wAvg-SMSE)` by specifying certain weight matrices. Note that it is easy to verify the above optimal weights $W_\ell^{*,(\texttt{wAvg-SMSE})}$ and $W_\ell^{*,(\texttt{wmSMSE})}$ satisfy Assumption 4. The detailed derivation is given in Section C.4 in supplementary material.

A.2 High-dimensional Multi-round SMSE

In this section, we extend `(mSMSE)` to high-dimensional settings, where the dimension p is much larger than n. We assume that $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is a sparse vector with s non-zero elements. Recall (7),

$$\widehat{\boldsymbol{\beta}}^{(t)} = \widehat{\boldsymbol{\beta}}^{(t-1)} - \Big[\nabla^2 F_{h_t}\big(\widehat{\boldsymbol{\beta}}^{(t-1)}\big)\Big]^{-1} \nabla F_{h_t}\big(\widehat{\boldsymbol{\beta}}^{(t-1)}\big) = \widehat{\boldsymbol{\beta}}^{(t-1)} - \Big(V_n^{(t)}\Big)^{-1} U_n^{(t)}. \tag{7}$$

It is generally infeasible to compute the inverse of the Hessian matrix $V_n^{(t)} \in \mathbb{R}^{p\times p}$ in the high-dimensional case. Furthermore, it requires unacceptably high complexity to compute and communicate L high-dimensional matrices. To tackle these problems, we first note that (7) is equivalent to solving the following quadratic optimization problem:

$$\widehat{\boldsymbol{\beta}}^{(t)} := \operatorname*{argmin}_{\boldsymbol{\beta}} \quad \frac{1}{2}\boldsymbol{\beta}^\top V_n^{(t)} \boldsymbol{\beta} - \boldsymbol{\beta}^\top \Big(V_n^{(t)} \widehat{\boldsymbol{\beta}}^{(t-1)} - U_n^{(t)}\Big). \tag{8}$$

Due to high communication complexity, we only estimate the Hessian matrix $V_n^{(t)}$ using the samples on a single machine, e.g., $V_{m,1}^{(t)}$ on the first machine. Then (8) can be written as

$$\widehat{\boldsymbol{\beta}}^{(t)} := \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \quad \frac{1}{2}\boldsymbol{\beta}^\top V_{m,1}^{(t)}\boldsymbol{\beta} - \boldsymbol{\beta}^\top\left(V_{m,1}^{(t)}\widehat{\boldsymbol{\beta}}^{(t-1)} - U_n^{(t)}\right). \tag{9}$$

We adapt the idea of the Dantzig Selector proposed by Candes and Tao (2007), an ℓ_1-regularization approach known for estimating high-dimensional sparse parameters. Formally, in the t-th iteration, given $\widehat{\boldsymbol{\beta}}^{(t-1)}$, the bandwidth h_t and a regularization parameter $\lambda_n^{(t)}$, we compute $\widehat{\boldsymbol{\beta}}^{(t)}$ by

$$\widehat{\boldsymbol{\beta}}^{(t)} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{\|\boldsymbol{\beta}\|_1 : \left\|V_{m,1}^{(t)}\boldsymbol{\beta} - \left(V_{m,1}^{(t)}\widehat{\boldsymbol{\beta}}^{(t-1)} - U_n^{(t)}\right)\right\|_\infty \le \lambda_n^{(t)}\right\}. \tag{10}$$

Note that a feasible solution of (10) can be obtained by linear programming. A complete algorithm is presented in Algorithm 1. It is worthwhile to mention that in Algorithm 1, we do not directly compute the $p \times p$ matrix $V_{m,1}^{(t)}$, but instead compute the vector $V_{m,1}^{(t)}\widehat{\boldsymbol{\beta}}^{(t-1)}$ to save computation and storage cost.

Algorithm 1 High-dimensional Multi-round Maximum Score Estimator

Input: Datasets distributed on local machines $\{x_i, \boldsymbol{z}_i, y_i\}_{i\in\mathcal{H}_\ell}$ $(\ell = 1, 2, .., L)$, an initial estimator $\widehat{\boldsymbol{\beta}}^{(0)}$, the total number of iterations T, bandwidth sequence $\{h_t\}$ and parameters $\{\lambda_n^{(t)}\}$.

1: **for** $t = 1, 2, ..., T$ **do**
2: Send $\widehat{\boldsymbol{\beta}}^{(t-1)}$ to each machine;
3: **for** $\ell = 1, 2, ..., L$ **do**
4: Compute $U_{m,\ell}^{(t)}$ by equation (8) and send $U_{m,\ell}^{(t)}$ back to $\mathcal{H}_1$;
5: **end for**
6: Compute $U_n^{(t)} = \frac{1}{L}\sum_{\ell=1}^{L} U_{m,\ell}^{(t)}$ and $V_{m,1}^{(t)}\widehat{\boldsymbol{\beta}}^{(t-1)} = \frac{1}{mh_t^2}\sum_{i\in\mathcal{H}_1}(-y_i)H''\left(\frac{x_i+\boldsymbol{z}_i^\top\widehat{\boldsymbol{\beta}}^{(t-1)}}{h_t}\right)\left(\boldsymbol{z}_i^\top\widehat{\boldsymbol{\beta}}^{(t-1)}\right)\boldsymbol{z}_i$.
7: Obtain $\widehat{\boldsymbol{\beta}}^{(t)}$ by solving (10);
8: **end for**
9: Output $\widehat{\boldsymbol{\beta}}^{(T)}$.

Now we give the convergence rate of the estimator $\widehat{\boldsymbol{\beta}}^{(t)}$ at iteration t.

Theorem A.3. *Assume the assumptions in Theorem 3.4 hold.Further, assume that the dimension* $p = O(n^\nu)$ *for some* $\nu > 0$, *the sparsity* $s = O\left(m^{1/4}\right)$, *and the initial value*[1] $\widehat{\boldsymbol{\beta}}^{(0)}$ *satisfies* $\left\|\widehat{\boldsymbol{\beta}}^{(0)} -\right.$

[1]An initial estimator $\widehat{\boldsymbol{\beta}}^{(0)}$ can be obtained by existing high-dimensional MSE methods on a single machine in literature such as Mukherjee et al. (2019) and Feng et al. (2022).

$\boldsymbol{\beta}^*\big\|_2 = O_{\mathbb{P}}(\delta_{m,0})$ *and* $\big\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\big\|_1 = O_{\mathbb{P}}(\sqrt{s}\delta_{m,0})$, *for some* $\delta_{m,0} = o(1)$. *By choosing proper bandwidth* h_t, *parameter* $\lambda_n^{(t)}$ *and kernel function* $H(\cdot)$, *we can obtain that for* $1 \le t \le T$,

$$\big\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\big\|_2 = O_{\mathbb{P}}\Big(\sqrt{s}\Big(\frac{\log p}{n}\Big)^{\frac{\alpha}{2\alpha+1}} + (r_m)^t\delta_{m,0}\Big), \tag{11}$$

and $\big\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\big\|_1 = O_{\mathbb{P}}\big(\sqrt{s}\big\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\big\|_2\big)$, *where* r_m *is an infinitesimal quantity.*

Theorem A.3 summarizes the ℓ_2 and ℓ_1 error bounds of $\widehat{\boldsymbol{\beta}}^{(t)}$ in Algorithm 1. The details, including the explicit choice for h_t, $\lambda_n^{(t)}$ and $H(\cdot)$ and the formal definition of r_m, are relegated to Section B of supplementary material. The upper bound in (11) contains two terms. The second term comes from the error of the initial estimator, and it decreases exponentially as t increases. As the algorithm operates, this quantity finally gets dominated by the first term within at most $O(\log n)$ iterations. Furthermore, the conditions on m, n, and s in Theorem A.3 are placed to guarantee that the algorithm converges in finite iterations. The first term, $\sqrt{s}\,(\log p/n)^{\frac{\alpha}{2\alpha+1}}$, represents the statistical convergence rate of our proposed estimator, which is very close to the optimal rate that one can obtain without a distributed environment. Feng et al. (2022) recently established the minimax optimal rate $(s\log p/n)^{\frac{\alpha}{2\alpha+1}}$ of the maximum score estimator in high-dimensional settings. Compared to that, the established rate in Theorem A.3, $\sqrt{s}\,(\log p/n)^{\frac{\alpha}{2\alpha+1}}$, is slightly slower due to the different designs in the algorithms, with a difference of $s^{\frac{1}{4\alpha+2}}$. Since our proposed algorithm is designed based on the Dantzig Selector that directly controls the infinity norm of the gradient, which is different from the path-following algorithm used in Feng et al. (2022), their techniques cannot be directly applied to our proposed distributed estimator. It would be a potentially interesting future work to improve the estimator in distributed settings to match the optimal rate.

B Theoretical Results of the High-dimensional (mSMSE)

In this section, we give the complete theoretical analysis of the high-dimensional multi-round SMSE in Algorithm 1. First, we restate the conditions as the following two assumptions.

Assumption 5. *The initial value* $\widehat{\boldsymbol{\beta}}^{(0)}$ *satisfies*

$$\Big\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\Big\|_2 = O_{\mathbb{P}}\left(\delta_{m,0}\right), \quad \Big\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\Big\|_1 = O_{\mathbb{P}}\left(\sqrt{s}\delta_{m,0}\right), \tag{12}$$

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

Assumption 6. *The dimension* $p = O(n^{\nu})$ *for some* $\nu > 0$ *and the sparsity* $s = O(m^{r})$ *for some* $0 < r < 1/4$.

Assumption 5 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 could 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 6 restricts the dimension p and the sparsity s. The first constraint $p = O(n^{\nu})$ implies that $\log p = O(\log n)$, so we will simply use $\log p$ instead of $\log \max\{n, p\}$ in the following theorems. The sparsity condition $s = O(m^{r})$ is also necessary and standard to guarantee convergence. Under these assumptions, we first state the one-step improvement in the following theorem:

Theorem B.1. *Assume the assumptions in Theorem 3.4 and Assumptions 5, 6 hold. Further assume that* $\sqrt{s}\delta_{m,0} = O\left(h_1^{3/2}\right)$ *and* $\frac{s^2 \log m}{m h_1^3} = o(1)$. *By taking*

$$\lambda_n^{(1)} = C_0 \left(\sqrt{\frac{\log p}{n h_1}} + \sqrt{\frac{s \log p}{m h_1^3}} \delta_{m,0} + s\delta_{m,0}^2 + h_1^{\alpha} \right),$$

with sufficiently large constant C_0 *, we have*

$$\left\| \widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^* \right\|_2 = O_{\mathbb{P}}\left(\sqrt{s} \lambda_n^{(1)} \right) = O_{\mathbb{P}}\left[\sqrt{\frac{s \log p}{n h_1}} + \sqrt{\frac{s^2 \log p}{m h_1^3}} \delta_{m,0} + s^{3/2} \delta_{m,0}^2 + \sqrt{s} h_1^{\alpha} \right], \tag{13}$$

and

$$\left\| \widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^* \right\|_1 \leq 2\sqrt{s} \left\| \widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^* \right\|_2, \tag{14}$$

with probability tending to one.

The proof of Theorem B.1 is given in Section C.5. The additional assumption $\frac{s^2 \log p}{m h_1^3} = o(1)$ is necessary to guarantee the so-called restricted eigenvalue condition of $V_{m,1}^{(1)}$, which is a standard assumption to obtain the convergence rate of the Dantzig Selector in theory. Another assumption $\sqrt{s}\delta_{m,0} = O\left(h_1^{3/2}\right)$ is a technical condition to determine the dominant term in the convergence rate (See the proof for details). The convergence rate in (13) contains four terms. The sum of the second and the third terms can be rewritten as,

$$\left(\sqrt{\frac{s^2 \log p}{m h_1^3}} + s^{3/2} \delta_{m,0} \right) \delta_{m,0}, \tag{15}$$

which is related to the initial error $\delta_{m,0}$. If we further suppose that $s^{3/2}\delta_{m,0} = o(1)$, then (15) will be $o(\delta_{m,0})$, which can further be iteratively refined in the following iterations. The remaining terms (the first and the fourth terms) can be minimized by letting $\sqrt{\frac{s\log p}{nh_1}} \asymp \sqrt{s}h_1^{\alpha}$, leading to the optimal bandwidth $h^* := (\log p/n)^{1/(2\alpha+1)}$. Based on these results, we are ready to give the convergence rate of the estimator $\widehat{\boldsymbol{\beta}}^{(t)}$ at iteration t.

Theorem B.2. *For $t = 1, 2, \ldots, T$, define $h^* := (\log p/n)^{1/(2\alpha+1)}$, $r_m := \sqrt{\frac{s^2\log p}{m(h^*)^3}} + s^{3/2}\delta_{m,0}$ and*

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

Assume the assumptions in Theorem 3.4 and Assumptions 5,6 hold. Then there exists a constant α_0 such that, by choosing a kernel $H'(\cdot)$ with order $\alpha > \alpha_0$, bandwidth $h_t \equiv h^*$ for $t = 1, 2, \ldots, T$ and parameters

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

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

$$\left\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\right\|_2 = O_{\mathbb{P}}\left(\delta_{m,t}\right) \quad \text{and} \quad \left\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\right\|_1 \le 2\sqrt{s}\left\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\right\|_2, \tag{16}$$

with probability tending to one.

The proof of Theorem B.2 is given in Section C.5. The $\delta_{m,t}$ in Theorem B.2 denotes the upper bound on the ℓ_2-error of $\widehat{\boldsymbol{\beta}}^{(t)}$, which contains two terms. 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 $\widehat{\boldsymbol{\beta}}^{(t)}$ achieves the optimal rate after at most $O(\log n)$ iterations. Concretely, the number of required iterations is

$$\frac{\alpha\left[(\log n - \log\log n) - \frac{1}{2}\log s + \log\delta_{m,0}\right]}{(2\alpha+1)\left[-\log\left(r_m\right)\right]}, \tag{17}$$

which is larger than that in the low-dimensional case. Under the Assumption 6 and the assumption $m > n^c$ in Theorem 3.4, it's easy to see that (17) 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 constant certain α_0. See Remark 2 for explanation.

Remark 2. *To make sure that $r_m = \frac{s^2\log m}{m(h^*)^3} + s^{3/2}\delta_{m,0} = o(1)$ and $\sqrt{s}\delta_{m,0} = O\left((h^*)^{3/2}\right)$, we need to choose a kernel $H'(\cdot)$ with order α such that $\alpha > \alpha_0 := \max\left\{\frac{3}{2c\,(1-2r)} + \frac{r}{2(1-2r)}, \frac{3r}{2(1-4r)}\right\}$, where*

$c = (\log m)/(\log n) < 1$ *and* $r = (\log s)/(\log m) < 1/4$ *are supposed in Theorem 3.4 and Assumption 6. Here we plug in the rate* $\delta_{m,0} = (s \log p/m)^{\alpha/(2\alpha+1)}$ *obtained by the path-following algorithm by Feng et al. (2022). If the order of the kernel is not high enough (less than* α_0*), Algorithm 1 still works by choosing* $h_m = m^{-\frac{2\alpha(1-2r)}{3(2\alpha+1)}+\varepsilon}$ *for some small* $\varepsilon > 0$*, and the corresponding convergence rate will be* $\sqrt{s}h_m^{\alpha}$*.*

C Technical Proof of the Theoretical Results

C.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

$$\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^* = \left(V_{n,h_1}\left(\widehat{\boldsymbol{\beta}}^{(0)}\right)\right)^{-1} U_{n,h_1}\left(\widehat{\boldsymbol{\beta}}^{(0)}\right), \tag{18}$$

where

$$V_{n,h}(\boldsymbol{\beta}) = \nabla^2 F_h(\boldsymbol{\beta}) = \frac{1}{nh^2}\sum_{i=1}^{n}(-y_i)\, H''\left(\frac{x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}}{h}\right)\boldsymbol{z}_i\boldsymbol{z}_i^\top, \tag{19}$$

$$U_{n,h}(\boldsymbol{\beta}) = \frac{1}{nh^2}\sum_{i=1}^{n}(-y_i)\, H''\left(\frac{x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}}{h}\right)\boldsymbol{z}_i\boldsymbol{z}_i^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^*) - \frac{1}{nh}\sum_{i=1}^{n}(-y_i)\, H'\left(\frac{x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}}{h}\right)\boldsymbol{z}_i. \tag{20}$$

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

$$\left\|U_{n,h_1}\left(\widehat{\boldsymbol{\beta}}^{(0)}\right)\right\|_2 = O_{\mathbb{P}}\left(\delta_{m,0}^2 + h_1^{\alpha} + \sqrt{\frac{1}{nh_1}} + \delta_{m,0}\sqrt{\frac{\log n}{nh_1^3}}\right), \tag{21}$$

$$\left\|V_{n,h_1}\left(\widehat{\boldsymbol{\beta}}^{(0)}\right) - V\right\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{\log n}{nh_1^3}} + \delta_{m,0} + h_1\right), \tag{22}$$

and therefore

$$\left\|\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*\right\|_2 = O_{\mathbb{P}}\left(\delta_{m,0}^2 + h_1^{\alpha} + \sqrt{\frac{1}{nh_1}} + \delta_{m,0}\sqrt{\frac{\log n}{nh_1^3}}\right). \tag{23}$$

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. Also, note that the dimension p is fixed.

Proof of (21)

First prove (21). It suffices to show that

$$\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\leq\delta_{m,0}}\|U_{n,h}(\boldsymbol{\beta})\|_2=O_{\mathbb{P}}\left(\delta_{m,0}^2+h^{\alpha}+\sqrt{\frac{1}{nh}}+\delta_{m,0}\sqrt{\frac{\log n}{nh^3}}\right), \tag{24}$$

which implies (21) since $\left\|\widehat{\boldsymbol{\beta}}^{(0)}-\boldsymbol{\beta}^*\right\|_2\leq\delta_{m,0}$ 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 $\|\boldsymbol{v}-\boldsymbol{v}_{j_v}\|_2\leq 1/2$. Then we have

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

and thus

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

Therefore, to show (24), it suffices to show that

$$\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\leq\delta_{m,0}}\sup_{j_v\in[5^p]}\left|\boldsymbol{v}_{j_v}^\top U_{n,h}(\boldsymbol{\beta})\right|=O_{\mathbb{P}}\left(\sqrt{\frac{1}{nh}}+\delta_{m,0}\sqrt{\frac{\log n}{nh^3}}+\delta_{m,0}^2+h^{\alpha}\right). \tag{25}$$

From now on, we let $\boldsymbol{v}$ be an arbitrarily fixed vector in $\mathbb{S}^{p-1}$. Recall the definition

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

where

$$U_{h,i}(\boldsymbol{\beta}):=\frac{1}{h^2}(-y_i)H''\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}}{h}\right)\boldsymbol{z}_i\boldsymbol{z}_i^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^*)-\frac{1}{h}(-y_i)H'\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}}{h}\right)\boldsymbol{z}_i. \tag{26}$$

We have the following decomposition:

$$\begin{aligned}
&\boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta})\\
&=(1-\mathbb{E})\left[\boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta})-\boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta}^*)\right]+(1-\mathbb{E})\boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta}^*)+\mathbb{E}\boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta})\\
&=\frac{1}{n}\sum_{i=1}^n\phi_i^U(\boldsymbol{\beta})+(1-\mathbb{E})\boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta}^*)+\mathbb{E}\boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta})
\end{aligned} \tag{27}$$

where $\phi_i^U(\boldsymbol{\beta}) := (1-\mathbb{E})\left[\boldsymbol{v}^\top U_{h,i}(\boldsymbol{\beta}) - \boldsymbol{v}^\top U_{h,i}(\boldsymbol{\beta}^*)\right]$. We will separately bound the three terms in (27). through the following three steps.

Step 1

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

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \le \delta_{m,0}} \left| \frac{1}{n}\sum_{i=1}^n \phi_i^U(\boldsymbol{\beta}) \right| \le C_\phi \delta_{m,0}\sqrt{\frac{\log n}{nh^3}}, \tag{28}$$

with probability $1-2n^{-\gamma p}$.

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 could 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 $\left\|\boldsymbol{\beta}-\boldsymbol{\beta}_{j_\beta}\right\|_\infty \le \frac{2\delta_{m,0}}{n^\gamma}$.

By Assumption 5 which requires $\sup_i \|\boldsymbol{z}_i\|_\infty \le \overline{B}$, we have $\left|\boldsymbol{v}^\top \boldsymbol{z}_i\right| \le \overline{B}\|\boldsymbol{v}\|_1 \le \overline{B}\sqrt{p}\|\boldsymbol{v}\|_2$ for any $\boldsymbol{v} \in \mathbb{R}^p$. By Assumption 1, $H''(x)$, $H'(x)$ are both bounded and Lipschitz continuous, and thus we have

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

for some constant C depending on H but not on $\boldsymbol{\beta}$ or $\boldsymbol{\beta}_{j_\beta}$. Therefore,

$$\left|\phi_i^U(\boldsymbol{\beta}) - \phi_i^U\left(\boldsymbol{\beta}_{j_\beta}\right)\right| \le \left|(1-\mathbb{E})\left[\boldsymbol{v}^\top U_{h,i}(\boldsymbol{\beta}) - \boldsymbol{v}^\top U_{h,i}(\boldsymbol{\beta}_{j_\beta})\right]\right| \lesssim \frac{\overline{B}^3 p^{3/2}\delta_{m,0}^2}{n^\gamma h^3} + \frac{\overline{B}^2 p\delta_{m,0}}{n^\gamma h^2},$$

Choose γ to be sufficiently large such that

$$\frac{\overline{B}^3 p^{3/2}\delta_{m,0}^2}{n^\gamma h^3} + \frac{\overline{B}^2 p\delta_{m,0}}{n^\gamma h^2} \ll \delta_{m,0}\sqrt{\frac{\log n}{nh^3}},$$

and then we have

$$\begin{aligned}&\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\leq\delta_{m,0}}\left|\frac{1}{n}\sum_{i=1}^{n}\phi_i^U(\boldsymbol{\beta})\right|-\sup_{j_\beta\in[n^{\gamma p}]}\left|\frac{1}{n}\sum_{i=1}^{n}\phi_i^U(\boldsymbol{\beta}_{j_\beta})\right|\\&=o\left(\delta_{m,0}\sqrt{\frac{\log n}{nh^3}}\right).\end{aligned}\tag{29}$$

Now let $\boldsymbol{\beta}$ be a fixed vector in $\{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\leq\delta_{m,0}\}$. Again using $\left|\boldsymbol{v}^\top\boldsymbol{z}_i\right|\leq\overline{B}\sqrt{p}$, $\left|\boldsymbol{z}_i^\top\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\right|\leq\overline{B}\sqrt{p}\delta_{m,0}$ and the boundedness of $H'(x)$ and $H''(x)$, we have

$$\sup_i\left|\phi_i^U\left(\boldsymbol{\beta}\right)\right|=O\left(\frac{\overline{B}^2p\delta_{m,0}}{h^2}\right),\tag{30}$$

Recall that $\zeta=X+\boldsymbol{Z}^\top\boldsymbol{\beta}^*$ and $\rho\left(\cdot\,|\,\boldsymbol{Z}\right)$ denotes the density of ζ given $\boldsymbol{Z}$. Let $\mathbb{E}_{\cdot|\boldsymbol{Z}}$ denote the expectation conditional on $\boldsymbol{Z}$. By Assumption 2, $\rho\left(\cdot\,|\,\boldsymbol{Z}\right)$ is bounded in a neighborhood of zero, and hence we have, when n is large enough,

$$\begin{aligned}&\mathbb{E}_{\cdot|\boldsymbol{Z}}\left[H''\left(\frac{X+\boldsymbol{Z}^\top\boldsymbol{\beta}}{h}\right)\right]^2=\mathbb{E}_{\cdot|\boldsymbol{Z}}\left[H''\left(\frac{\boldsymbol{Z}^\top\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)+\zeta}{h}\right)\right]^2\\&=h\int_{-1}^{1}H''\left(\xi\right)^2\rho\left(\xi h-\boldsymbol{Z}^\top\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)\,|\,\boldsymbol{Z}\right)\mathrm{d}\xi\\&=O\left(h\right),\end{aligned}\tag{31}$$

where we use that $h=o(1)$ and $\boldsymbol{Z}^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\leq\overline{B}\sqrt{p}\delta_{m,0}=o(1)$. Similarly,

$$\begin{aligned}\mathbb{E}_{\cdot|\boldsymbol{Z}}\left[H'\left(\frac{X+\boldsymbol{Z}^\top\boldsymbol{\beta}}{h}\right)-H'\left(\frac{X+\boldsymbol{Z}^\top\boldsymbol{\beta}^*}{h}\right)\right]^2&=\mathbb{E}_{\cdot|\boldsymbol{Z}}\left[\frac{\boldsymbol{Z}^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^*)}{h}H''\left(\frac{X+\boldsymbol{Z}^\top\breve{\boldsymbol{\beta}}}{h}\right)\right]^2\\&=O\left(\frac{\overline{B}^2p\delta_{m,0}^2}{h}\right).\end{aligned}\tag{32}$$

By (31) and (32),

$$\begin{aligned}\mathrm{var}\left(\phi_i^U\left(\boldsymbol{\beta}\right)\right)&\leq\mathbb{E}\left(\boldsymbol{v}^\top U_{h,i}\left(\boldsymbol{\beta}\right)-\boldsymbol{v}^\top U_{h,i}(\boldsymbol{\beta}^*)\right)^2\\&=O\left(\frac{\overline{B}^4p^2\delta_{m,0}^2}{h^3}\right).\end{aligned}\tag{33}$$

Neither of the constant hidden in the big O notation of (30) or (33) depends on $\boldsymbol{\beta}$. By (30) and (33), we can apply Bernstein inequality to show that there exists a large enough $C_1>0$, which does not depend on $\boldsymbol{\beta}$, such that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\phi_i^U\left(\boldsymbol{\beta}\right)\right|>C_1\delta_{m,0}\sqrt{\frac{\log n}{nh^3}}\right)\leq2n^{-2\gamma p}.$$

The above inequality is true for any $\boldsymbol{\beta} \in \mathbb{R}^p$ that satisfies $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq \delta_{m,0}$. In particular, for any $j_\beta \in [n^{\gamma p}]$, with probability at least $1 - 2n^{-2\gamma p}$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} \phi_i^U \left(\boldsymbol{\beta}_{j_\beta} \right) \right| \leq C_1 \delta_{m,0} \sqrt{\frac{\log n}{nh^3}},$$

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

$$\sup_{j_\beta \in [n^{\gamma p}]} \left| \frac{1}{n} \sum_{i=1}^{n} \phi_i^U (\boldsymbol{\beta}_{j_\beta}) \right| \leq C_1 \delta_{m,0} \sqrt{\frac{\log n}{nh^3}}.$$

Combining with (29), we obtain (28).

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

$$(1 - \mathbb{E}) \boldsymbol{v}^\top U_{n,h} (\boldsymbol{\beta}^*) \geq C^* \sqrt{\frac{1}{nh}}, \tag{34}$$

with probability at least $1 - 2n^{-\gamma p}$, where γ has been specified in Step 1.

Note that $(1 - \mathbb{E}) \boldsymbol{v}^\top U_{n,h} (\boldsymbol{\beta}^*) = \frac{1}{n} \sum_{i=1}^{n} (1 - \mathbb{E}) \left[\boldsymbol{v}^\top U_{h,i} (\boldsymbol{\beta}^*) \right]$ 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 $\rho^{(1)}(\cdot \mid \boldsymbol{Z})$ is bounded uniformly for all $\boldsymbol{Z}$ in a neighborhood of 0, which implies that $\rho(t \mid Z) = \rho(0 \mid \boldsymbol{Z}) + O(t)$. The constant in $O(t)$ is the same for all t in the neighborhood. Therefore,

$$\begin{aligned}
&\mathbb{E}_{\cdot \mid \boldsymbol{Z}} \left[\boldsymbol{v}^\top U_{h,i} (\boldsymbol{\beta}^*) \right]^2 \\
=& \frac{\left(\boldsymbol{v}^\top \boldsymbol{Z} \right)^2}{h^2} \mathbb{E}_{\cdot \mid \boldsymbol{Z}} \left[H' \left(\frac{X + \boldsymbol{Z}^\top \boldsymbol{\beta}^*}{h} \right) \right]^2 \\
=& \int_{-1}^{1} \frac{\left(\boldsymbol{v}^\top \boldsymbol{Z} \right)^2}{h} \left[H' (\xi) \right]^2 \rho (\xi h \mid \boldsymbol{Z}) \, \mathrm{d}\xi \\
=& \frac{\left(\boldsymbol{v}^\top \boldsymbol{Z} \right)^2}{h} \int_{-1}^{1} \left[H' (\xi) \right]^2 \rho (0 \mid \boldsymbol{Z}) \, \mathrm{d}\xi + O \left((\boldsymbol{v}^\top \boldsymbol{Z})^2 \right).
\end{aligned} \tag{35}$$

Recall that $V_s := \pi_V \mathbb{E} \rho (0 \mid \boldsymbol{Z}) \boldsymbol{Z} \boldsymbol{Z}^\top$ and $\pi_V := \int_{-1}^{1} \left[H' (\xi) \right]^2 \mathrm{d}\xi$, which leads to

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

In Step 3, we also show that $\sup_{\boldsymbol{\beta}: \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq \delta_{m,0}} \mathbb{E} \left[\boldsymbol{v}^\top U_{h,i} (\boldsymbol{\beta}) \right] = o(1)$. Therefore,

$$\begin{aligned}
&\mathrm{var} \left[\sqrt{h} \boldsymbol{v}^\top U_{h,i} (\boldsymbol{\beta}^*) \right] \\
=& h \left\{ \mathbb{E} \left[\boldsymbol{v}^\top U_{h,i} (\boldsymbol{\beta}^*) \right]^2 - \left(\mathbb{E} \left[\boldsymbol{v}^\top U_{h,i} (\boldsymbol{\beta}^*) \right] \right)^2 \right\} \\
=& \boldsymbol{v}^\top V_s \boldsymbol{v} + o(1),
\end{aligned} \tag{36}$$

By CLT and Slusky's Theorem,

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

which implies (34).

Step 3

We will show that,

$$\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^{*}\|_{2}\leq\delta_{m,0}}\mathbb{E}\left[\boldsymbol{v}^{\top}U_{n,h}\left(\boldsymbol{\beta}\right)\right]\leq C_{E}(\delta_{m,0}^{2}+h^{\alpha}),\tag{37}$$

where C_E does not depend on $\boldsymbol{v}$.

By Assumption 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,

$$\left(2F\left(-t\,|\,\boldsymbol{Z}\right)-1\right)\rho\left(t\,|\,\boldsymbol{Z}\right)=\sum_{k=1}^{2\alpha+1}M_{k}\left(\boldsymbol{Z}\right)t^{k},\tag{38}$$

where $M_k(\boldsymbol{Z})$'s are constants depending on ρ, F, $\boldsymbol{Z}$, and t. Since $\rho^{(k)}(\cdot\,|\,\boldsymbol{Z})$ and $F^{(k)}(\cdot\,|\,\boldsymbol{Z})$ are bounded around 0 for all k and $\boldsymbol{Z}$, we know there exists a constant M such that $\sup_{k,\boldsymbol{Z},t}|M_k(\boldsymbol{Z})|\leq M$ for all t in a certain neighborhood of 0.

(In particular,

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

which will be used in the proof of Theorem 3.4.)

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

Now we are ready to compute the expectation of $\boldsymbol{v}^{\top}U_{n,h}(\boldsymbol{\beta})$. 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 $\zeta = X + \boldsymbol{Z}^\top \boldsymbol{\beta}^*$.

$$
\begin{aligned}
&\mathbb{E}_{\cdot|\boldsymbol{Z}}\left[\boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta})\right] \\
=&\boldsymbol{Z}^\top \boldsymbol{v} \cdot \mathbb{E}_{\cdot|\boldsymbol{Z}}\left[\frac{\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})}{h^2}\left[2\mathbb{I}\left(X + \boldsymbol{Z}^\top \boldsymbol{\beta}^* + \epsilon < 0\right) - 1\right] H''\left(\frac{X + \boldsymbol{Z}^\top \boldsymbol{\beta}}{h}\right)\right. \\
&\qquad \left. - \frac{1}{h}\left[2\mathbb{I}\left(X + \boldsymbol{Z}^\top \boldsymbol{\beta}^* + \epsilon < 0\right) - 1\right] H'\left(\frac{X + \boldsymbol{Z}^\top \boldsymbol{\beta}}{h}\right)\right] \\
=&\boldsymbol{Z}^\top \boldsymbol{v} \cdot \mathbb{E}_{\cdot|\boldsymbol{Z}}\left\{\left[2\mathbb{I}(\zeta + \epsilon < 0) - 1\right]\left[\frac{\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})}{h^2} H''\left(\frac{\zeta + \boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})}{h}\right) - \frac{1}{h} H'\left(\frac{\zeta + \boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})}{h}\right)\right]\right\} \\
=&\left(\boldsymbol{Z}^\top \boldsymbol{v}\right) \int_{-1}^{1} \left[2F\left(\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta}) - \xi h \,|\, \boldsymbol{Z}\right) - 1\right] \rho\left(\xi h - \boldsymbol{Z}^\top \Delta(\boldsymbol{\beta}) \mid \boldsymbol{Z}\right) \\
&\quad \cdot \left(\frac{\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})}{h} H''(\xi) - H'(\xi)\right) \mathrm{d}\xi \quad \left(\text{by changing variable } \xi = \frac{\zeta + \boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})}{h}\right) \\
=&\left(\boldsymbol{Z}^\top \boldsymbol{v}\right) \sum_{k=1}^{2\alpha+1} M_k(\boldsymbol{Z}) \int_{-1}^{1} \left(\xi h - \boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})\right)^k \left(\frac{\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})}{h} H''(\xi) - H'(\xi)\right) \mathrm{d}\xi.
\end{aligned} \tag{39}
$$

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

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

When $k = \alpha$,

$$
\begin{aligned}
&\int_{-1}^{1} \left(\xi h - \boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})\right)^\alpha \left(\frac{\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})}{h} H''(\xi) - H'(\xi)\right) \mathrm{d}\xi \\
=&\sum_{k'=0}^{\alpha} \binom{\alpha}{k'} h^{k'} \left(-\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})\right)^{\alpha-k'} \left[\left(\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})/h\right) \int_{-1}^{1} \xi^{k'} H''(\xi)\, \mathrm{d}\xi - \int_{-1}^{1} \xi^{k'} H'(\xi)\, \mathrm{d}\xi\right] \\
=&(\alpha-1)\left(-\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})\right)^\alpha - \pi_U h^\alpha,
\end{aligned} \tag{40}
$$

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 $\left|\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})\right| \le \overline{B}\sqrt{p}\delta_{m,0}$, $\delta_{m,0} = O(h)$ and that H', H'' are both bounded, we have

$$
\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \le \delta_{m,0}} \left|\int_{-1}^{1} \left(\xi h - \boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})\right)^k \left(\frac{\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})}{h} H''(\xi) - H'(\xi)\right) \mathrm{d}\xi\right| = O\left[h^{\alpha+1} + \delta_{m,0}^{\alpha+1}\right].
$$

Since $M_k(\boldsymbol{Z})$'s are uniformly bounded, we finally obtain

$$\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\le\delta_{m,0}}\left|\mathbb{E}\left[\boldsymbol{v}^\top U_{n,h}\left(\boldsymbol{\beta}\right)\right]\right| = O\left(\delta_{m,0}^2+h^\alpha\right), \tag{41}$$

which leads to (37).

Combining the three steps leads to, for any fixed vector $\boldsymbol{v}\in\mathbb{S}^{p-1}$, there exists a constant $C_{\boldsymbol{v}}$, with probability at least $1-4\left(n^{-\gamma}\right)^p$,

$$\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\le\delta_{m,0}}\left|\boldsymbol{v}^\top U_{n,h}\left(\boldsymbol{\beta}\right)\right| \le C_{\boldsymbol{v}}\left(\sqrt{\frac{1}{nh}}+\delta_{m,0}\sqrt{\frac{\log n}{nh^3}}+\delta_{m,0}^2+h^\alpha\right),$$

The constant $C_{\boldsymbol{v}}$ depends on $\boldsymbol{v}$. Recall in equation (25), we only consider finite number of $\boldsymbol{v}$'s. Take $C_{\sup}=\sup_{j\in[5^p]}C_{\boldsymbol{v}_j}$, and then we have

$$\sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\le\delta_{m,0}}\sup_{j_v\in[5^p]}\left|\boldsymbol{v}_{j_v}^\top U_{n,h}\left(\boldsymbol{\beta}\right)\right| \le C_{\sup}\left(\sqrt{\frac{1}{nh}}+\delta_{m,0}\sqrt{\frac{\log n}{nh^3}}+\delta_{m,0}^2+h^\alpha\right),$$

with probability at least $1-4(5n^{-\gamma})^p$. This proves (24) and completes the proof of (21).

Proof of (22)

The proof of (22) is similar. We will use the same $\boldsymbol{v}_{j_v}$ and $\boldsymbol{\beta}_{j_\beta}$ as before. 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, we only need 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 $\{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\le\delta_{m,0}\}$, there exists $j_\beta\in[n^{\gamma p}]$ such that $\left\|\boldsymbol{\beta}-\boldsymbol{\beta}_{j_\beta}\right\|_\infty\le\frac{2\delta_{m,0}}{n^\gamma}$. 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

$$V_{h,i}\left(\boldsymbol{\beta}\right):=\frac{1}{h^2}\left(-y_i\right)H''\left(\frac{x_i+\boldsymbol{z}_i^\top\boldsymbol{\beta}}{h}\right)\boldsymbol{z}_i\boldsymbol{z}_i^\top. \tag{42}$$

We have

$$\left|\boldsymbol{v}_{j_v}^\top V_{n,h}\left(\boldsymbol{\beta}\right)\boldsymbol{v}_{j_v}-\boldsymbol{v}_{j_v}^\top V_{n,h}\left(\boldsymbol{\beta}_{j_\beta}\right)\boldsymbol{v}_{j_v}\right| \le \sup_i\frac{\left(\boldsymbol{z}_i^\top\boldsymbol{v}_{j_v}\right)^2}{h^3}\boldsymbol{z}_i^\top\left(\boldsymbol{\beta}-\boldsymbol{\beta}_{j_\beta}\right)=O\left(\frac{\overline{B}^3p^{3/2}\delta_{m,0}}{n^\gamma h^3}\right).$$

By taking $\gamma>0$ large enough, it suffices to show that

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

In the following proof, we will again consider any fixed $\boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{v} \in \mathbb{R}^p$ that satisfy $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq \delta_{m,0}$ 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 before, but this time we only need the following expansion:

$$\rho\left(t \,|\, \boldsymbol{Z}\right) = \rho\left(0 \,|\, \boldsymbol{Z}\right) + O\left(t\right),$$

$$F\left(-t \,|\, \boldsymbol{Z}\right) = 1/2 - F'\left(0 \,|\, \boldsymbol{Z}\right) t + O\left(t^2\right).$$

Recall that $\int_{-1}^{1} xH''\left(x\right) \mathrm{d}x = -1$ and $\int_{-1}^{1} x^k H''\left(x\right) \mathrm{d}x = 0$ for $k = 0$ and $2 \leq k \leq \alpha$.

$$\begin{aligned}
&\mathbb{E}_{\cdot|\boldsymbol{Z}}\left(\boldsymbol{v}^\top V_{n,h}\left(\boldsymbol{\beta}\right)\boldsymbol{v}\right) \\
=&\frac{\left(\boldsymbol{v}^\top \boldsymbol{z}\right)^2}{h^2}\mathbb{E}_{\cdot|\boldsymbol{Z}}\left[2\mathbb{I}\left(\zeta + \epsilon < 0\right) - 1\right] H''\left(\frac{\boldsymbol{Z}^\top \Delta\left(\boldsymbol{\beta}\right) + \zeta}{h}\right) \\
=&\frac{\left(\boldsymbol{v}^\top \boldsymbol{Z}\right)^2}{h}\int_{-1}^{1}\left(2F\left(\boldsymbol{Z}^\top \Delta\left(\boldsymbol{\beta}\right) - \xi h \,|\, \boldsymbol{Z}\right) - 1\right)\rho\left(\xi h - \boldsymbol{Z}^\top \Delta\left(\boldsymbol{\beta}\right) \,|\, \boldsymbol{Z}\right) H''\left(\xi\right) \mathrm{d}\xi \\
=&-2\left(\boldsymbol{v}^\top \boldsymbol{Z}\right)^2 F'\left(0 \,|\, \boldsymbol{Z}\right)\rho\left(0 \,|\, \boldsymbol{Z}\right)\int_{-1}^{1} \xi H''\left(\xi\right) \mathrm{d}\xi + O\left(h + \left|\boldsymbol{Z}^\top \Delta\left(\boldsymbol{\beta}\right)\right|^2 / h\right) \\
=&2\left(\boldsymbol{v}^\top \boldsymbol{Z}\right)^2 F'\left(0 \,|\, \boldsymbol{Z}\right)\rho\left(0 \,|\, \boldsymbol{Z}\right) + O\left(h + \delta_{m,0}\right),
\end{aligned} \tag{43}$$

hence

$$\mathbb{E}\left(\boldsymbol{v}^\top \left[V_{n,h}\left(\boldsymbol{\beta}\right) - V\right]\boldsymbol{v}\right) = O\left(\delta_{m,0} + h\right). \tag{44}$$

Since $H''\left(x\right)$ is bounded and $\left|\boldsymbol{v}^\top \boldsymbol{z}_i\right| \leq \overline{B}\sqrt{p}$ (p is fixed), we have

$$\left|\boldsymbol{v}^\top V_{h,i}\left(\boldsymbol{\beta}\right)\boldsymbol{v}\right| = O\left(\frac{1}{h^2}\right). \tag{45}$$

Adding (31), we also have

$$\mathbb{E}\left(\boldsymbol{v}^\top V_{h,i}\left(\boldsymbol{\beta}\right)\boldsymbol{v}\right)^2 \leq \mathbb{E}\left[\frac{\left(\boldsymbol{v}^\top \boldsymbol{z}_i\right)^4}{h^4} O\left(h\right)\right] = O\left(\frac{1}{h^3}\right). \tag{46}$$

By (45) and (46), we could apply Bernstein inequality to show that

$$\mathbb{P}\left(\left|\left(1 - \mathbb{E}\right)\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{v}^\top V_{h,i}\left(\boldsymbol{\beta}\right)\boldsymbol{v}\right| > C_2\sqrt{\frac{p\log n}{nh^3}}\right) \leq 2n^{-2\gamma p},$$

for some constant $C_2 > 0$. By the same procedure in the proof of (21), we obtain, with probability at least $1 - 2\left(5n^{-\gamma}\right)^p$,

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

which completes the proof of (22).

Finally, (18), (21), and (22) directly leads to (11), given the assumption that $\Lambda_{\min}(V) > c_1^{-1}$ for some $c_1 > 0$ (Assumption 4).

□

Proof of Theorem 3.3

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

$$h_t = \max\{n^{-1/(2\alpha+1)}, m^{-2^t/(3\alpha)}\} \geq n^{-\frac{1}{2\alpha+1}},$$

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

We first show that, for any t,

$$\left\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\right\|_2 = O_{\mathbb{P}}\left(n^{-\alpha/(2\alpha+1)}\sqrt{\log n} + m^{-2^t/3}\right). \tag{47}$$

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

$$\left\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\right\|_2 = O_{\mathbb{P}}\left(\delta_{m,t-1}^2 + h_t^\alpha + \sqrt{\frac{1}{nh_t}} + \delta_{m,t-1}\sqrt{\frac{\log n}{nh_t^3}}\right), \tag{48}$$

which implies that

$$\left\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\right\|_2 = O_{\mathbb{P}}\left(\delta_{m,t-1}^2 + h_t^\alpha + \sqrt{\frac{\log n}{nh_t}}\right), \tag{49}$$

since $\delta_{m,t-1}\sqrt{\frac{\log n}{nh_t^3}} = O\big(\sqrt{\frac{\log n}{nh_t}}\big)$.

When $t = 1$, $\delta_{m,0} := m^{-1/3} \leq h_1 = m^{-\frac{2}{3\alpha}}$, which verifies $\delta_{m,0} = O(h_1)$. Then (47) holds since

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

Assume (47) holds for $t-1$, i.e.,

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

Then $\delta_{m,t-1} = n^{-\alpha/(2\alpha+1)}\sqrt{\log n} + m^{-2^{t-1}/3}$. Since $n^{-\alpha/(2\alpha+1)}\sqrt{\log n} \ll n^{-\frac{1}{2\alpha+1}} \leq h_t$ and $m^{-2^{t-1}/3} \leq m^{-\frac{2^t}{3\alpha}} \leq h_t$, it holds that $\delta_{m,t-1} = O(h_t)$. Then by (49),

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

since $h_t^\alpha = \max\left\{m^{-2^t/3}, n^{-\alpha/(2\alpha+1)}\right\}$ and $1/\sqrt{nh_t} \le n^{-\alpha/(2\alpha+1)}$. Therefore, we have proved that (47) holds for all t by induction.

To see (12), note that plugging $\delta_{m,t-1} = n^{-\alpha/(2\alpha+1)}\sqrt{\log n} + m^{-2^{t-1}/3}$ into (48) yields that

$$\left\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\right\|_2 = O_{\mathbb{P}}\left(\frac{\log n}{n^{2\alpha/(2\alpha+1)}} + m^{-2^t/3} + h_t^\alpha + \sqrt{\frac{1}{nh_t}} + \left(\frac{\sqrt{\log n}}{n^{\alpha/(2\alpha+1)}} + m^{-2^{t-1}/3}\right)\sqrt{\frac{\log n}{nh_t^3}}\right). \quad (50)$$

Since $\alpha > 1$,

$$\log n \cdot \sqrt{\frac{1}{nh_t^3}} \le \log n \cdot n^{-\frac{\alpha-1}{2\alpha+1}} = o(1).$$

Therefore, the rate in (50) is upper bounded by

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

which completes the proof. □

C.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 (20).

Lemma C.1. *Under Assumptions 1–5, if* $h = o(1)$, $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 = O_{\mathbb{P}}(\delta)$, $\delta = o\left(\max\{h^{\alpha/2}, h/\sqrt{\log n}\}\right)$ *then*

$$\mathbb{E}\left[U_{n,h}(\boldsymbol{\beta})\right] = Uh^\alpha + o_{\mathbb{P}}(h^\alpha), \quad (51)$$

and

$$(1-\mathbb{E})\sqrt{nh}U_{n,h}(\boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, V_s). \quad (52)$$

Proof of Lemma C.1. Recall that

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

and

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

where π_U is defined in Assumption 1. For any $\boldsymbol{v} \in \mathbb{S}^{p-1}$, the computation in (39) yields that

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{v}^\top U_{n,h}(\boldsymbol{\beta})\right] &= -\mathbb{E}\left[\left(\boldsymbol{Z}^\top \boldsymbol{v}\right)\pi_U M_\alpha(\boldsymbol{Z})\, h^\alpha\right] + O\left(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 + h^{\alpha+1}\right) \\ &= \boldsymbol{v}^\top U \cdot h^\alpha + o_{\mathbb{P}}(h^\alpha),\end{aligned}$$

where the big O notation hides a constant not depending on $\boldsymbol{\beta}$. This proves (51).

To show (52), further recall that in **Step 2** of the proof of (21), we show that

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

Furthermore, in **Step 1** of the same proof, we prove that

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

which yields (52) if $\delta = o\left(h/\sqrt{\log n}\right)$.

□

Proof of Theorem 3.4

We now restate Theorem 3.4 and give the proof using Lemma C.1.

Theorem 3.4. *Assume the local size $m > n^c$ for some constant $0 < c < 1$ and the assumptions in Theorem 3.3 hold. When T satisfies (13), we have*

$$n^{\frac{\alpha}{2\alpha+1}}\left(\widehat{\boldsymbol{\beta}}^{(T)}-\boldsymbol{\beta}^*\right) \xrightarrow{d} \mathcal{N}\Big(\lambda_h^{\frac{\alpha}{2\alpha+1}}V^{-1}U,\ \lambda_h^{\frac{-1}{2\alpha+1}}V^{-1}V_sV^{-1}\Big), \tag{53}$$

where U and V_s are defined in (10). The asymptotic mean squared error is given by

$$\mathbb{E}\left[\left(\widehat{\boldsymbol{\beta}}^{(T)}-\boldsymbol{\beta}^*\right)^\top\left(\widehat{\boldsymbol{\beta}}^{(T)}-\boldsymbol{\beta}^*\right)\right] = n^{-\frac{2\alpha}{2\alpha+1}}\cdot\operatorname{trace}\left[\lambda_h^{-\frac{1}{2\alpha+1}}V^{-1}V_sV^{-1}+\lambda_h^{-\frac{2\alpha}{2\alpha+1}}U^\top V^{-1}V^{-1}U\right],$$

by minimizing which we obtain the optimal λ_h^* as follows.

$$\lambda_h^* := \frac{\operatorname{trace}\left(V^{-1}V_sV^{-1}\right)}{2\alpha U^\top V^{-1}V^{-1}U}.$$

Proof. By Theorem 3.3, when

$$T > \log_2\left(\frac{6\alpha}{2\alpha+1}\cdot\frac{\log n-\log\lambda_h}{\log m}\right),$$

$h_T = \left(\frac{\lambda_h}{n}\right)^{\frac{1}{2\alpha+1}}$ and $m^{-2^{T-1}/3} \lesssim n^{-\alpha/(2\alpha+1)}$. Hence, $\left\|\widehat{\boldsymbol{\beta}}^{(T-1)}-\boldsymbol{\beta}^*\right\|_2 = O_{\mathbb{P}}\left(\sqrt{\log n}\cdot n^{-\frac{\alpha}{2\alpha+1}}\right) = o_{\mathbb{P}}\left(\max\{h_T^{\alpha/2}, h_T/\sqrt{\log n}\}\right)$, and thus the assumptions of Lemma C.1 hold. Since $\sqrt{nh_T} = \sqrt{\lambda_h}h_T^{-\alpha}$, we obtain by Lemma C.1 that

$$\sqrt{nh_T}\cdot U_{n,h_T}\left(\widehat{\boldsymbol{\beta}}^{(T-1)}\right) \xrightarrow{d} \mathcal{N}\left(\sqrt{\lambda_h}U, V_s\right),$$

and hence

$$\sqrt{nh_T}\cdot V^{-1}U_{n,h_T}\left(\widehat{\boldsymbol{\beta}}^{(T-1)}\right)\xrightarrow{d}\mathcal{N}\left(\sqrt{\lambda_h}V^{-1}U,V^{-1}V_sV\right).$$

By (22) and Assumption 4, $\left\|V_{n,h_T}^{-1}\left(\widehat{\boldsymbol{\beta}}^{(T-1)}\right)-V^{-1}\right\|_2=o_{\mathbb{P}}(1)$, which yields

$$\sqrt{nh_T}\cdot V_{n,h_T}^{-1}\left(\widehat{\boldsymbol{\beta}}^{(T-1)}\right)U_{n,h_T}\left(\widehat{\boldsymbol{\beta}}^{(T-1)}\right)\xrightarrow{d}\mathcal{N}\left(\sqrt{\lambda_h}V^{-1}U,V^{-1}V_sV^{-1}\right).$$

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

$$\begin{aligned}
&\mathbb{E}\left[\left(\widehat{\boldsymbol{\beta}}^{(T)}-\boldsymbol{\beta}^*\right)^\top\left(\widehat{\boldsymbol{\beta}}^{(T)}-\boldsymbol{\beta}^*\right)\right]\\
&=\mathbb{E}\mathrm{trace}\left[\left(\widehat{\boldsymbol{\beta}}^{(T)}-\boldsymbol{\beta}^*\right)^\top\left(\widehat{\boldsymbol{\beta}}^{(T)}-\boldsymbol{\beta}^*\right)\right]\\
&=\mathbb{E}\mathrm{trace}\left[\left(\widehat{\boldsymbol{\beta}}^{(T)}-\boldsymbol{\beta}^*\right)\left(\widehat{\boldsymbol{\beta}}^{(T)}-\boldsymbol{\beta}^*\right)^\top\right]\\
&\to n^{-\frac{2\alpha}{2\alpha+1}}\cdot\mathrm{trace}\left[\lambda_h^{-\frac{1}{2\alpha+1}}V^{-1}V_sV^{-1}+\lambda_h^{-\frac{2\alpha}{2\alpha+1}}U^\top V^{-1}V^{-1}U\right].
\end{aligned}$$

Then it's direct 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 C.1 in Section C.1 has already implies that $V_{n,h_T}\left(\widehat{\boldsymbol{\beta}}^{(T)}\right)\xrightarrow{p}V$, where $V_{n,h}(\boldsymbol{\beta})$ is defined in (19), 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 C.2. *Assume assumptions in Theorem 3.3 hold. Let $h_\kappa=n^{-\frac{\kappa}{2\alpha+1}}$ for some $0<\kappa<1$. When*

$$T\geq\log_2\left(\frac{6\alpha}{2\alpha+1}\cdot\frac{\log n-\log\lambda_h}{\log m}\right),$$

we have

$$\widehat{U}:=\frac{1}{nh_\kappa^{\alpha+1}}\sum_{i=1}^n y_iH'\left(\frac{x_i+\boldsymbol{z}_i^\top\widehat{\boldsymbol{\beta}}^{(T)}}{h_\kappa}\right)\boldsymbol{z}_i\xrightarrow{p}U,\tag{54}$$

$$\widehat{V}_s:=\frac{1}{nh_T}\sum_{i=1}^n\left[H'\left(\frac{x_i+\boldsymbol{z}_i^\top\widehat{\boldsymbol{\beta}}^{(T)}}{h_T}\right)\right]^2\boldsymbol{z}_i\boldsymbol{z}_i^\top\xrightarrow{p}V_s.\tag{55}$$

Proof. When $t>\log_2\left(\frac{3\alpha}{2\alpha+1}\cdot\frac{\log n-\log\lambda_h}{\log m}\right)$, we have

$$n^{-\frac{\kappa}{2\alpha+1}}=h_\kappa\gg h_t=\left(\frac{\lambda_h}{n}\right)^{\frac{1}{2\alpha+1}},$$

and

$$\left\|\widehat{\boldsymbol{\beta}}^{(t)}-\boldsymbol{\beta}^{*}\right\|_{2}=O_{\mathbb{P}}\left(\sqrt{\log n}\cdot n^{-\frac{\alpha}{2\alpha+1}}\right)=o_{\mathbb{P}}\left(h_{\kappa}^{\alpha}\right).$$

Replacing h_t with h^* in the proof of Theorem 3.4 leads to

$$\mathbb{E}\boldsymbol{v}^{\top}U_{n,h_{\kappa}}\left(\widehat{\boldsymbol{\beta}}^{(t)}\right)=\boldsymbol{v}^{\top}Uh_{\kappa}^{\alpha}+o_{\mathbb{P}}\left(h_{\kappa}^{\alpha}\right),$$

and

$$\sqrt{nh_{\kappa}}\left(1-\mathbb{E}\right)\boldsymbol{v}^{\top}U_{n,h_{\kappa}}\left(\widehat{\boldsymbol{\beta}}^{(t)}\right)\xrightarrow{d}\mathcal{N}\left(0,\boldsymbol{v}^{\top}V_{s}\boldsymbol{v}\right),$$

for any $\boldsymbol{v}\in\mathbb{S}^{p}$. Also, $V_{n,h_{\kappa}}\left(\widehat{\boldsymbol{\beta}}^{(t)}\right)\xrightarrow{p}V$ by Proposition 3.2 . Note that

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

and $\left(h_{\kappa}\right)^{-\alpha}\ll\sqrt{nh_{\kappa}}$. Then

$$\begin{aligned}\boldsymbol{v}^{\top}\widehat{U}&=\left(h_{\kappa}\right)^{-\alpha}\left[\mathbb{E}\boldsymbol{v}^{\top}U_{n,h_{\kappa}}\left(\widehat{\boldsymbol{\beta}}^{(t)}\right)+\left(1-\mathbb{E}\right)\boldsymbol{v}^{\top}U_{n,h_{\kappa}}\left(\widehat{\boldsymbol{\beta}}^{(t)}\right)-\boldsymbol{v}^{\top}V_{n,h_{\kappa}}\left(\widehat{\boldsymbol{\beta}}^{(t)}\right)\left(\widehat{\boldsymbol{\beta}}^{(t)}-\boldsymbol{\beta}^{*}\right)\right]\\&=\boldsymbol{v}^{\top}U+o_{\mathbb{P}}\left(1\right),\end{aligned}$$

which yields (54).

To prove (55), recall

$$\widehat{V}_{s}:=\frac{1}{nh_{t}}\sum_{i=1}^{n}\left[H'\left(\frac{x_{i}+\boldsymbol{z}_{i}^{\top}\widehat{\boldsymbol{\beta}}^{(t)}}{h_{t}}\right)\right]^{2}\boldsymbol{z}_{i}\boldsymbol{z}_{i}^{\top}.$$

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

$$\begin{aligned}\mathbb{E}\boldsymbol{v}^{\top}\widehat{V}_{s}\boldsymbol{v}&=\mathbb{E}\left(\mathbb{E}_{\cdot\,|\,\boldsymbol{Z}}\boldsymbol{v}^{\top}\widehat{V}_{s}\boldsymbol{v}\right)\\&=\mathbb{E}\left[\left(\boldsymbol{Z}^{\top}\boldsymbol{v}\right)^{2}\int_{-1}^{1}\left[H'\left(\xi\right)\right]^{2}\rho\left(\xi h_{t}-\boldsymbol{Z}^{\top}\left(\widehat{\boldsymbol{\beta}}^{(t)}-\boldsymbol{\beta}^{*}\right)\,|\,\boldsymbol{Z}\right)\,\mathrm{d}\xi\right]\\&=\mathbb{E}\left(\boldsymbol{Z}^{\top}\boldsymbol{v}\right)^{2}\pi_{V}\rho\left(0\,|\,\boldsymbol{Z}\right)+O\left(h_{t}+\left\|\widehat{\boldsymbol{\beta}}^{(t)}-\boldsymbol{\beta}^{*}\right\|_{2}\right)\\&=\boldsymbol{v}^{\top}V_{s}\boldsymbol{v}+o\left(1\right).\end{aligned}$$

Also, since $H'\left(x\right)$ and $\left\|\boldsymbol{z}_{i}\right\|_{\infty}$ are both bounded, $\operatorname{var}\left(\boldsymbol{v}^{\top}\widehat{V}_{s}\boldsymbol{v}\right)=O\left(\frac{1}{nh_{t}^{2}}\right)=o\left(1\right)$. By Chebyshev's inequality, (55) is proved.

□

C.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

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

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

where $U_{h,i}(\boldsymbol{\beta})$ and $V_{h,i}(\boldsymbol{\beta})$ are defined in (26) and (42). 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 $\nabla_{\boldsymbol{\beta}}F_{h,\ell}(\boldsymbol{\beta}^*) = -U_{m,\ell,h}(\boldsymbol{\beta}^*)$. Horowitz (1992) showed that

$$\left\|\widehat{\boldsymbol{\beta}}_{\mathrm{SMSE},\ell} - \boldsymbol{\beta}^*\right\|_2 = O_{\mathbb{P}}\left(h^\alpha + \frac{1}{\sqrt{mh}}\right),$$

and thus by the proof of (22), if $\log m/mh^3 \to 0$, we have $\left\|V^{-1} - V^{-1}_{m,\ell,h}\left(\breve{\boldsymbol{\beta}}_\ell\right)\right\|_2 = o(1)$ with probability at least $1 - 2(5m^{-\gamma})^p$, where γ is a large positive constant. Hence, we have

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

Therefore,

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

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

By Lemma C.1,

$$\mathbb{E}U_{n,h}(\boldsymbol{\beta}^*) = Uh^\alpha + o_{\mathbb{P}}(h^\alpha),$$

and

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

If $L = o\left(m^{\frac{2}{3}(\alpha-1)}/\left(\log m\right)^{\frac{2\alpha+1}{3}}\right)$ and $h = \left(\frac{\lambda_h}{n}\right)^{\frac{1}{2\alpha+1}}$, we have $\frac{\log m}{mh^3} = o\left(1\right)$ and $\sqrt{\lambda_h}h^{-\alpha} = \sqrt{mLh}$, and thus

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

which already 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

$$h^{-\alpha}\left(\widehat{\boldsymbol{\beta}}_{(\texttt{Avg-SMSE})} - \boldsymbol{\beta}^* - V^{-1}\mathbb{E}U_{n,h}\left(\boldsymbol{\beta}^*\right)\right) = o_{\mathbb{P}}\left(1\right).$$

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

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

which yields

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

□

C.4 Proof of the Results for (wAvg-SMSE) and (wmSMSE)

Proof of Theorem A.1

Proof. Analogous to the proof of Theorem 3.1,

$$\widehat{\boldsymbol{\beta}}_{(\texttt{wAvg-SMSE})} - \boldsymbol{\beta}^* = \sum_{\ell=1}^{L} W_\ell V_\ell^{-1} U_{m_\ell,\ell,h_\ell}(\boldsymbol{\beta}^*) + o_{\mathbb{P}}(1)\sum_{\ell=1}^{L} W_\ell U_{m_\ell,\ell,h_\ell}. \tag{56}$$

By the uniformness, the bias satisfies

$$\mathbb{E}\left[\sum_{\ell=1}^{L} W_\ell V_\ell^{-1} U_{m_\ell,\ell,h_\ell}(\boldsymbol{\beta}^*)\right] = \sum_{\ell=1}^{L} W_\ell V_\ell^{-1} U_\ell h^\alpha + o\left(h^\alpha\right) = B_{(\texttt{wAvg-SMSE})} + o\left(h^\alpha\right), \tag{57}$$

where $B_{(\texttt{wAvg-SMSE})} := \sum_{\ell=1}^{L} W_\ell V_\ell^{-1} U_\ell h^\alpha$. For any $\boldsymbol{v}$ satisfying $\|\boldsymbol{v}\|_2 = 1$, define $\Phi_i := (1 - \mathbb{E})m_\ell^{-1}\boldsymbol{v}^\top W_\ell V_\ell^{-1} U_{h_\ell,i}(\boldsymbol{\beta}^*)$ where $i \in \mathcal{H}_\ell$. (Note that the definitions of Φ_i are different in the proof of different theorems, but they are parallel.) Then

$$(1-\mathbb{E})\left[\sum_{\ell=1}^{L} \boldsymbol{v}^\top W_\ell V_\ell^{-1} U_{m_\ell,\ell,h_\ell}(\boldsymbol{\beta}^*)\right] = (1-\mathbb{E})\left[\sum_{\ell=1}^{L}\sum_{i\in\mathcal{H}_\ell} m_\ell^{-1}\boldsymbol{v}^\top W_\ell V_\ell^{-1} U_{h_\ell,i}(\boldsymbol{\beta}^*)\right] = \sum_{i=1}^{n}\Phi_i. \tag{58}$$

Recall that by (30) and (36),

$$\sup_i |\Phi_i| = O\left(\max_\ell \frac{\|W_\ell\|_2}{m_\ell h}\right),$$

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

The Lindeberg's condition is satisfied, since

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

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

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

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.

$$\begin{aligned}
&\left(\boldsymbol{v}^\top\Sigma_{(\texttt{wAvg-SMSE})}\boldsymbol{v}\right)^{-1/2}\boldsymbol{v}^\top\left(\widehat{\boldsymbol{\beta}}_{(\texttt{wAvg-SMSE})}-\boldsymbol{\beta}^*-B_{(\texttt{wAvg-SMSE})}\right)\\
=&\left(\boldsymbol{v}^\top\Sigma_{(\texttt{wAvg-SMSE})}\boldsymbol{v}\right)^{-1/2}\sum_{i=1}^n\Phi_i+\left(\boldsymbol{v}^\top\Sigma_{(\texttt{wAvg-SMSE})}\boldsymbol{v}\right)^{-1/2}o(h^\alpha)\\
&+o_{\mathbb{P}}(1)\left(\boldsymbol{v}^\top\Sigma_{(\texttt{wAvg-SMSE})}\boldsymbol{v}\right)^{-1/2}\sum_{\ell=1}^L W_\ell U_{m_\ell,\ell,h_\ell}(\boldsymbol{\beta}^*).
\end{aligned} \tag{59}$$

When $\left(\boldsymbol{v}^\top\Sigma_{(\texttt{wAvg-SMSE})}\boldsymbol{v}\right)^{1/2}\asymp h^\alpha$, the second term on the RHS of (59) 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 also asymptotically normal by repeating the previous procedure. 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. In particular, when Assumption 4 is satisfied, the matrix $\sum_{\ell=1}^L\frac{n}{m_\ell}W_\ell V_\ell^{-1}V_{s,\ell}V_\ell^{-1}W_\ell^\top$ is finite and $\left(\boldsymbol{v}^\top\Sigma_{(\texttt{wAvg-SMSE})}\boldsymbol{v}\right)^{1/2}\asymp 1/\sqrt{nh}$. Then the above results directly lead to Theorem A.1.

Additionally, we show that the minimum of the asymptotic variance in (4),

$$\sum_{\ell=1}^L\frac{n}{m_\ell}W_\ell V_\ell^{-1}V_{s,\ell}V_\ell^{-1}W_\ell^\top,$$

is minimized at

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

in the sense of both trace and Frobenius norm. For trace, we want to solve the following optimization problem:

$$\min_{W_\ell} \operatorname{trace}\left(\sum_{\ell=1}^{L} \frac{n}{m_\ell} W_\ell V_\ell^{-1} V_{s,\ell} V_\ell^{-1} W_\ell^\top\right), \text{ s.t.} \sum_{\ell=1}^{L} W_\ell = I_{p\times p}.$$

The Lagrangian is

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

By taking derivative w.r.t W_ℓ and letting the derivative be zero, we obtain $2\frac{n}{m_\ell} W_\ell V_\ell^{-1} V_{s,\ell} V_\ell^{-1} + \Lambda = \mathbf{0}$ or $W_\ell \propto \left(m_\ell^{-1} V_\ell^{-1} V_{s,\ell} V_\ell^{-1}\right)^{-1} = m_\ell V_\ell V_{s,\ell}^{-1} V_\ell$. By the constraint that $\sum_{\ell=1}^{L} W_\ell = I_{p\times p}$, we find the minimizer $W_\ell^{*,(\texttt{wAvg}-\texttt{SMSE})}$ defined above. The proof for Frobenius norm is similar.

□

Proof of Theorem A.2

Proof. For weighted (mSMSE), we have

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

Define the weighted sample size $\overline{n}_W := 1/\left(\sum_{\ell=1}^{L} \frac{\|W_\ell\|_2^2}{m_\ell}\right)$. We will show the one-step error under the assumptions in Proposition 3.2 (replacing n by $\overline{n}_W$),

$$\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^* = O_{\mathbb{P}}\left(\sqrt{\frac{\log \overline{n}_W}{\overline{n}_W h}} + \delta_{m,0}^2 + h^\alpha\right). \tag{60}$$

The proof is parallel to the proof of Proposition 3.2 , and we will only show the difference in the remaining part.

By the uniformness in the assumptions, similar to (39), we can obtain

$$\mathbb{E}\left[\boldsymbol{v}^\top\left(\sum_{\ell=1}^{L}\sum_{i\in\mathcal{H}_\ell} m_\ell^{-1} W_\ell U_{h,i}\left(\widehat{\boldsymbol{\beta}}^{(0)}\right)\right)\right] = O\left(\delta_{m,0}^2 + h^\alpha\right).$$

Let $\Phi_i = m_\ell^{-1}\boldsymbol{v}^\top W_\ell U_{h,i}\left(\widehat{\boldsymbol{\beta}}^{(0)}\right)$ with $i \in \mathcal{H}_\ell$. By (30) and (36),

$$\sup_i |\Phi_i| = O\left(\max_\ell \frac{\|W_\ell\|_2}{m_\ell h}\right),$$

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

By Bernstein's inequality, for any γ, there exists a constant C_W such that

$$\mathbb{P}\left(\sum_i \Phi_i \geq C_W \sqrt{\frac{p \log \overline{n}_W}{\overline{n}_W h}}\right) \leq 2(\overline{n}_W)^{-2\gamma p}.$$

Then following the same procedure as the proof of Proposition 3.2 , we can obtain that

$$\left\| \sum_{\ell=1}^{L} \sum_{i \in \mathcal{H}_\ell} m_\ell^{-1} W_\ell U_{h,i} \left(\widehat{\boldsymbol{\beta}}^{(0)}\right) \right\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{\log \overline{n}_W}{\overline{n}_W h}} + \delta_{m,0}^2 + h^\alpha\right).$$

Similarly, we can also obtain the convergence of the Hessian matrix in the same way, given by

$$\left\| \sum_{\ell=1}^{L} \sum_{i \in \mathcal{H}_\ell} m_\ell^{-1} W_\ell V_{h,i} \left(\widehat{\boldsymbol{\beta}}^{(0)}\right) - \overline{V}_W \right\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{\log \overline{n}_W}{\overline{n}_W h^3}} + \delta_{m,0} + h\right).$$

Hence, (60) holds, and we use the same strategy to choose the bandwidth h_t as in Theorem 3.3. When t is large enough, $h \asymp \overline{n}_W^{-1/(2\alpha+1)}$ and the error $\delta_{t-1} = \left\|\widehat{\boldsymbol{\beta}}^{(t-1)}_{(\texttt{wmSMSE})} - \boldsymbol{\beta}^*\right\|_2 = O\left(\overline{n}_W^{-\alpha/(2\alpha+1)} \log \overline{n}_W\right)$, which satisfies $\delta_{t-1}^2 = o(h^\alpha)$. Therefore, similar to the proof of (51), we have

$$\mathbb{E}\left(\sum_{\ell=1}^{L} \sum_{i \in \mathcal{H}_\ell} m_\ell^{-1} W_\ell U_{h,i} \left(\widehat{\boldsymbol{\beta}}^{(t-1)}\right)\right) = \overline{U}_W h^\alpha + o\left(h^\alpha\right).$$

Also, similar to the proof of Theorem A.1, when $h^\alpha \asymp \left(h^{-1} \sum_{\ell=1}^{L} m_\ell^{-1} W_\ell V_{s,\ell} W_\ell^\top\right)^{1/2}$, which is equivalent to $h \asymp \overline{n}_W^{-1/(2\alpha+1)}$, we have

$$\left(h^{-1} \sum_{\ell=1}^{L} m_\ell^{-1} W_\ell V_{s,\ell} W_\ell^\top\right)^{-1/2} \left(\sum_{\ell=1}^{L} \sum_{i \in \mathcal{H}_\ell} m_\ell^{-1} W_\ell U_{h,i} \left(\widehat{\boldsymbol{\beta}}^{(t-1)}\right) - \overline{U}_W h^\alpha\right) \xrightarrow{d} \mathcal{N}\left(0, I_{p\times p}\right).$$

In particular, when Assumption 4 is satisfied, $\overline{n}_W \asymp n$, and $n \sum_{\ell=1}^{L} m_\ell^{-1} W_\ell V_{s,\ell} W_\ell^\top$ is a finite matrix. Then we obtain

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

Combining with

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

and

$$\left\| \sum_{\ell=1}^{L} \sum_{i \in \mathcal{H}_\ell} m_\ell^{-1} W_\ell V_{h,i} \left(\widehat{\boldsymbol{\beta}}^{(t-1)} \right) - \overline{V}_W \right\|_2 = O_{\mathbb{P}} \left(\sqrt{\frac{\log \overline{n}_W}{\overline{n}_W h^3}} + \delta_{m,0} + h_t \right),$$

by Slusky's Theorem, (5) is true.

□

C.5 Proof of the Results for the High-dimensional (mSMSE)

Before starting the proof, we first formalize our notation. Define

$$V_n(\boldsymbol{\beta}) = \frac{1}{nh^2} \sum_{i=1}^{n} (-y_i) H'' \left(\frac{x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}}{h} \right) \boldsymbol{z}_i \boldsymbol{z}_i^\top. \tag{61}$$

$$V_{m,\ell}(\boldsymbol{\beta}) = \frac{1}{nh^2} \sum_{i \in \mathcal{H}_\ell} (-y_i) H'' \left(\frac{x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}}{h} \right) \boldsymbol{z}_i \boldsymbol{z}_i^\top. \tag{62}$$

$$U_n(\boldsymbol{\beta}) = \frac{1}{nh} \sum_{i=1}^{n} (-y_i) H' \left(\frac{x_i + \boldsymbol{z}_i^\top \boldsymbol{\beta}}{h} \right) \boldsymbol{z}_i. \tag{63}$$

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 C.3. *Assume Assumptions 1–5, 5 and 6 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 5) and* $h = o(1)$, *we have the following results:*

$$\left\| (1 - \mathbb{E}) \left[V_n\left(\widehat{\boldsymbol{\beta}}^{(0)}\right) - V \right] \right\|_{\max} = O_{\mathbb{P}} \left(\sqrt{\frac{\log p}{nh^3}} \right), \tag{64}$$

$$\sup_{\|\boldsymbol{v}_1\|_2 = \|\boldsymbol{v}_2\|_2 = 1} \boldsymbol{v}_1^\top \left(\mathbb{E}\left[V_n\left(\widehat{\boldsymbol{\beta}}^{(0)}\right) \right] - V \right) \boldsymbol{v}_2 = O_{\mathbb{P}} \left(\sqrt{s}\delta_{m,0} + h^\alpha \right), \tag{65}$$

$$\left\| (1 - \mathbb{E}) \left[V_{m,1}\left(\widehat{\boldsymbol{\beta}}^{(0)}\right) - V \right] \right\|_{\max} = O_{\mathbb{P}} \left(\sqrt{\frac{\log p}{mh^3}} \right), \tag{66}$$

$$\sup_{\|\boldsymbol{v}_1\|_2 = \|\boldsymbol{v}_2\|_2 = 1} \boldsymbol{v}_1^\top \left(\mathbb{E}\left[V_{m,1}\left(\widehat{\boldsymbol{\beta}}^{(0)}\right) \right] - V \right) \boldsymbol{v}_2 = O_{\mathbb{P}} \left(\sqrt{s}\delta_{m,0} + h^\alpha \right). \tag{67}$$

Additionally, define

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

and then we have

$$\left\|(1-\mathbb{E})\Psi_n\big(\widehat{\boldsymbol{\beta}}^{(0)}\big)\right\|_\infty = O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{nh}}\right), \tag{68}$$

and

$$\sup_{\|\boldsymbol{v}\|_2=1}\left|\boldsymbol{v}^\top\mathbb{E}\left[\Psi_n\big(\widehat{\boldsymbol{\beta}}^{(0)}\big)\right]\right| = O_{\mathbb{P}}\left(s\delta_{m,0}^2 + h^\alpha\right). \tag{69}$$

Proof. **Proof of (64):**

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

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

By Assumption 5, there exists constant C_{ℓ_1}, C_{ℓ_2} such that $\mathbb{P}\left(\boldsymbol{\beta}^{(0)}\in\Theta\right)\to 1$, 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 $\boldsymbol{\beta}^{(0)}\in\Theta$ in the following proof.

For each $(j_1,j_2)\in\{1,..,p\}\times\{1,..,p\}$, define

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

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

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

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

$$\begin{aligned}
&\left\|(1-\mathbb{E})\left[V_n\left(\widehat{\boldsymbol{\beta}}^{(0)}\right) - V\right]\right\|_{\max}\\
&= \sup_{j_1,j_2}\left|(1-\mathbb{E})\left[V_{n,j_1,j_2}\left(\widehat{\boldsymbol{\beta}}^{(0)}\right) - V_{j_1,j_2}\right]\right|\\
&\le \sup_{j_1,j_2}\Phi^V_{n,j_1,j_2} + \sup_{j_1,j_2}\left|(1-\mathbb{E})\,V_{n,j_1,j_2}\left(\boldsymbol{\beta}^*\right)\right|,
\end{aligned} \tag{70}$$

we break the proof of (64) into two steps, separately controlling the two terms in the last line of (70).

Step 1:

$$\sup_{j_1,j_2}\Phi^V_{n,j_1,j_2}=O_{\mathbb{P}}\left(\delta_{m,0}\sqrt{\frac{s\log p}{nh^6}}\right)\underline{\underline{\sqrt{s}\delta_{m,0}=O\left(h^{3/2}\right)}}O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{nh^3}}\right). \tag{71}$$

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 $\boldsymbol{z}_i^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\le\overline{B}\left\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\right\|_1\le\overline{B}\sqrt{s}\delta_{m,0}$. Since $H''(x)$ is Lipschitz, we have

$$\overline{\phi}:=\sup_{i,j_1,j_2}\sup_{\boldsymbol{\beta}\in\Theta}\left|\phi^V_{ij_1j_2}(\boldsymbol{\beta})\right|=O\left(\frac{\sqrt{s}\delta_{m,0}}{h^3}\right). \tag{72}$$

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

$$\begin{aligned}&\sup_i\sup_{\boldsymbol{\beta}\in\Theta}\mathbb{E}_{\cdot|\boldsymbol{Z}}\left[H''\left(\frac{X+\boldsymbol{Z}^\top\boldsymbol{\beta}}{h}\right)-H''\left(\frac{X+\boldsymbol{Z}^\top\boldsymbol{\beta}^*}{h}\right)\right]^2\\&=\sup_i\sup_{\boldsymbol{\beta}\in\Theta}h\int_{-1}^1\left[H''\left(\xi+\frac{\boldsymbol{Z}^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^*)}{h}\right)-H''(\xi)\right]^2\rho(h\xi\,|\,\boldsymbol{Z})\,\mathrm{d}\xi\\&=O\left(s\delta^2_{m,0}/h\right),\end{aligned} \tag{73}$$

which implies that

$$\sup_{i,j_1,j_2}\sup_{\boldsymbol{\beta}\in\Theta}\mathbb{E}\left[\left|\phi^V_{ij_1j_2}(\boldsymbol{\beta})\right|^2\right]=O\left(s\delta^2_{m,0}/h^5\right). \tag{74}$$

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

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

Further, as

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

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

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

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

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

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 6 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

$$\mathbb{E}\Phi^V_{n,j_1,j_2} = O\left(\delta_{m,0}\sqrt{\frac{s \log p}{nh^6}}\right). \tag{75}$$

To show (71), we use Theorem 7.3 in Bousquet (2003), which is restated as the following Lemma C.4.

Lemma C.4 (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 C.4 requires $\mathcal{F}$ to be countable. We first apply Lemma C.4 to prove (71) 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 (72), (74) and (75), we have $f(\boldsymbol{z}_i) \leq 1$,

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

and

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

By Lemma C.4, for all $x > 0$, with probability $1 - e^{-x}$,

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

By taking $x = 3\log\max\{n,p\}$, plugging (72) 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 $1 - 1/\max\{n,p\}^3$,

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

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

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

By the continuity of $\phi_{ij_1j_2}^V(\boldsymbol{\beta})$,

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

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

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

with probability at least $1 - 2/\max\{n,p\}$, which completes the proof of (71).

Step 2:

$$\sup_{j_1,j_2}\left|(1-\mathbb{E})V_{n,j_1,j_2}(\boldsymbol{\beta}^*)\right| = O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{nh^3}}\right). \tag{76}$$

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

$$\mathbb{E}_{\cdot\,|\,\boldsymbol{Z}}\left[H''\left(\frac{X+\boldsymbol{Z}^\top\boldsymbol{\beta}^*}{h}\right)\right]^2 = h\int_{-1}^{1}\left[H''(\xi)\right]^2\rho\left(\xi h\,|\,\boldsymbol{Z}\right)\mathrm{d}\xi = O(h). \tag{77}$$

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 large enough C to make

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

This implies that

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

which proves (76). Together with (70) and (71), we conclude the proof of (64).

Proof of (65):

Recall that, by Equation (38), for almost every $\boldsymbol{Z}$,

$$\begin{aligned}&\left(2F\left(-t\,|\,\boldsymbol{Z}\right)-1\right)\rho\left(t\,|\,\boldsymbol{Z}\right)\\ &= 2F^{(1)}\left(0\,|\,\boldsymbol{Z}\right)\rho\left(0\,|\,\boldsymbol{Z}\right)t + \sum_{k=2}^{2\alpha+1}M_k(\boldsymbol{Z})\,t^k,\end{aligned} \tag{78}$$

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 $\sup_k|M_k(\boldsymbol{Z})| \leq M$ 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 $\boldsymbol{\beta} \in \Theta$,

$$
\begin{aligned}
&\mathbb{E}_{\cdot|\boldsymbol{Z}}\left[\frac{\left(\boldsymbol{v}_1^\top \boldsymbol{Z}\right)\left(\boldsymbol{v}_2^\top \boldsymbol{Z}\right)}{h^2}(-y) H''\left(\frac{X + \boldsymbol{Z}^\top \boldsymbol{\beta}}{h}\right)\right] \\
=&\frac{\left(\boldsymbol{v}_1^\top \boldsymbol{Z}\right)\left(\boldsymbol{v}_2^\top \boldsymbol{Z}\right)}{h^2}\mathbb{E}_{\cdot\,|\,\boldsymbol{Z}}\left[2\mathbb{I}(\boldsymbol{Z} + \epsilon < 0) - 1\right] H''\left(\frac{\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta}) + \zeta}{h}\right) \\
=&\frac{\left(\boldsymbol{v}_1^\top \boldsymbol{Z}\right)\left(\boldsymbol{v}_2^\top \boldsymbol{Z}\right)}{h}\int_{-1}^{1}\left[2F\left(\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta}) - \xi h \,|\, \boldsymbol{Z}\right) - 1\right]\rho\left(\xi h - \boldsymbol{Z}^\top \Delta(\boldsymbol{\beta}) \,|\, \boldsymbol{Z}\right) H''(\xi)\,\mathrm{d}\xi \qquad (79) \\
=&\frac{\left(\boldsymbol{v}_1^\top \boldsymbol{Z}\right)\left(\boldsymbol{v}_2^\top \boldsymbol{Z}\right)}{h} 2F^{(1)}(0 \,|\, \boldsymbol{Z})\,\rho(0 \,|\, \boldsymbol{Z})\int_{-1}^{1}\left(\xi h - \boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})\right) H''(\xi)\,\mathrm{d}\xi \\
&+ \frac{\left(\boldsymbol{v}_1^\top \boldsymbol{Z}\right)\left(\boldsymbol{v}_2^\top \boldsymbol{Z}\right)}{h}\cdot\sum_{k=2}^{2\alpha+1} M_k(\boldsymbol{Z})\int_{-1}^{1}\left(\xi h - \boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})\right)^k H''(\xi)\,\mathrm{d}\xi
\end{aligned}
$$

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

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

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

$$
\begin{aligned}
\left|\int_{-1}^{1}\left(\xi h - \boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})\right)^k H''(\xi)\,\mathrm{d}\xi\right| &\le \int_{-1}^{1} 2^{k-1}\left(|\xi h|^k + \left|\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})\right|^k\right)\left|H''(\xi)\right|\,\mathrm{d}\xi \\
&\le 2^{2\alpha}\sup_x\left|H''(x)\right|\left[h^{\alpha+1} + \left|\boldsymbol{Z}^\top \Delta(\boldsymbol{\beta})\right|^k\right] \\
&\le 2^{2\alpha}\left(1 + \overline{B}^k\right)\sup_x\left|H''(x)\right| h^{\alpha+1}.
\end{aligned}
$$

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

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

where

$$C_{EV} = M\left((\alpha-1)\left[\max\left\{1,\overline{B}\right\}\right]^{\alpha-1} + (\alpha+1)\,2^{2\alpha}\left(1+\left[\max\left\{1,\overline{B}\right\}\right]^{2\alpha+1}\right)\sup_x\left|H''(x)\right|\right)$$

is a constant not depending on $\boldsymbol{\beta}$ and $\boldsymbol{Z}$. Therefore,

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\Theta}\sup_{\|\boldsymbol{v}_1\|_2=\|\boldsymbol{v}_2\|_2=1}\boldsymbol{v}_1^\top\left(\mathbb{E}\left[V_n(\boldsymbol{\beta})\right]-V\right)\boldsymbol{v}_2 \\
&=\sup_{\boldsymbol{\beta}\in\Theta}\sup_{\|\boldsymbol{v}_1\|_2=\|\boldsymbol{v}_2\|_2=1}\mathbb{E}\left[\frac{\left(\boldsymbol{v}_1^\top\boldsymbol{Z}\right)\left(\boldsymbol{v}_2^\top\boldsymbol{Z}\right)}{h}(-y)\,H''\left(\frac{\boldsymbol{Z}^\top\boldsymbol{\beta}}{h}\right)-\frac{\left(\boldsymbol{v}_1^\top\boldsymbol{Z}\right)\left(\boldsymbol{v}_2^\top\boldsymbol{Z}\right)}{h}2F^{(1)}(0\,|\,\boldsymbol{Z})\,\rho(0\,|\,\boldsymbol{Z})\right] \\
&\lesssim \sqrt{s}\delta_{m,0}+h^\alpha,
\end{aligned} \tag{80}$$

which completes the proof of (65).

Proof of (66) and (67):

Equation (66) and (67) can be shown in the same way as above by replacing all the n with m.

Proof of (68):

The proof of (68) is analogous to that of (64). We will omit some details since they are the same. For each $j\in\{1,...,p\}$, define

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

Then we have

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

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

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

and

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

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

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

(Details are the same as the proof of (75), while the only difference is that $\phi_{ij}^U\left(\boldsymbol{\beta}\right)$ is a Lipschitz function of $\boldsymbol{z}_i^\top\left(\boldsymbol{\beta}-\boldsymbol{\beta}^*\right)$ with Lipschitz constant $\asymp 1/h^2$, instead of $1/h^3$ for $\phi_{ij}^V\left(\boldsymbol{\beta}\right)$.)

Using Lemma C.4 again, we can show that

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

Similar to the proof of (76), we have

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

and

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

By Bernstein's inequality,

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

Proof of (69):

Recall equation (38) and $\pi_U = \int_{-1}^1 x^\alpha H'\left(x\right)\mathrm{d}x \neq 0$. For any $\boldsymbol{v}\in\mathbb{R}^p$ that satisfies $\left\|\boldsymbol{v}\right\|_2 = 1$

and $\boldsymbol{\beta} \in \Theta$, we have

$$
\begin{aligned}
&\mathbb{E}_{\cdot|\boldsymbol{Z}} \boldsymbol{v}^{\top} \Psi_n(\boldsymbol{\beta}) \\
=&(\boldsymbol{v}^{\top}\boldsymbol{Z}) \cdot \mathbb{E}_{\cdot|\boldsymbol{Z}} \left[\frac{\boldsymbol{Z}^{\top}\Delta(\boldsymbol{\beta})}{h^2} [2\mathbb{I}(\zeta+\epsilon<0)-1] H''\left(\frac{X+\boldsymbol{Z}^{\top}\boldsymbol{\beta}}{h}\right) - \frac{1}{h}[2\mathbb{I}(\zeta+\epsilon<0)-1] H'\left(\frac{X+\boldsymbol{Z}^{\top}\boldsymbol{\beta}}{h}\right) \right] \\
=&(\boldsymbol{v}^{\top}\boldsymbol{Z}) \int_{-1}^{1} \left[2F\left(\boldsymbol{Z}^{\top}\Delta(\boldsymbol{\beta}) - \xi h \,|\, \boldsymbol{Z}\right) - 1 \right] \rho\left(\xi h - \boldsymbol{Z}^{\top}\Delta(\boldsymbol{\beta}) \,|\, \boldsymbol{Z}\right) \\
&\cdot \left(\frac{\boldsymbol{Z}^{\top}\Delta(\boldsymbol{\beta})}{h} H''(\xi) - H'(\xi) \right) \mathrm{d}\xi \\
=&(\boldsymbol{v}^{\top}\boldsymbol{Z}) \sum_{k=1}^{2\alpha+1} M_k(\boldsymbol{Z}) \int_{-1}^{1} \left(\xi h - \boldsymbol{Z}^{\top}\Delta(\boldsymbol{\beta})\right)^k \left(\frac{\boldsymbol{Z}^{\top}\Delta(\boldsymbol{\beta})}{h} H''(\xi) - H'(\xi) \right) \mathrm{d}\xi
\end{aligned}
\tag{81}
$$

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

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

For $k = \alpha$,

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

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

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

Therefore, $\mathbb{E}\left[\boldsymbol{v}^{\top}\Psi_n(\boldsymbol{\beta})\right] \lesssim \mathbb{E}(\boldsymbol{v}^{\top}\boldsymbol{Z})\left(s\delta_{m,0}^2 + h^{\alpha}\right) \lesssim s\delta_{m,0}^2 + h^{\alpha}$, which completes the proof of (69).

□

Proof of Theorem B.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 1,

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

Let $\boldsymbol{\delta} := \widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*$.

Using Lemma C.3, with probability tending to 1, we have

$$\begin{aligned}
\left\| V_{m,1}\boldsymbol{\beta}^* - \left(V_{m,1}\widehat{\boldsymbol{\beta}}^{(0)} - U_n \right) \right\|_\infty &\le \left\| U_n - V_n\left(\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\right) \right\|_\infty + \left\| \left(V_{m,1} - V_n\right)\left(\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\right) \right\|_\infty \\
&\le \left\| U_n - V_n\left(\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\right) \right\|_\infty + \|V_{m,1} - V_n\|_{\max} \left\| \widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^* \right\|_1 \\
&= O\left(\sqrt{\frac{\log p}{nh}} + \sqrt{\frac{s\log p}{mh^3}}\delta_{m,0} + s\delta_{m,0}^2 + h^\alpha \right) \\
&\le \lambda_n,
\end{aligned} \tag{82}$$

when C_λ is large enough. By the definition,

$$\left\| V_{m,1}\widehat{\boldsymbol{\beta}}^{(1)} - \left(V_{m,1}\widehat{\boldsymbol{\beta}}^{(0)} - U_n \right) \right\|_\infty \le \lambda_n. \tag{83}$$

Hence, with probability tending to 1,

$$\left\| V_{m,1}\left(\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*\right) \right\|_\infty \le 2\lambda_n. \tag{84}$$

From (82) and the optimality of $\widehat{\boldsymbol{\beta}}^{(1)}$, we have $\left\|\widehat{\boldsymbol{\beta}}^{(1)}\right\|_1 \le \|\boldsymbol{\beta}^*\|_1$, 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,

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

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, $\|\boldsymbol{\delta}\|_1 \le 2\sqrt{s}\,\|\boldsymbol{\delta}\|_2$ and $\|V_{m,1}\boldsymbol{\delta}\|_\infty \le 2\lambda_{m,0}$. Therefore,

$$\begin{aligned}
\boldsymbol{\delta}^\top V_{m,1}\boldsymbol{\delta} &= \boldsymbol{\delta}^\top V\boldsymbol{\delta} + \boldsymbol{\delta}^\top\left(V_{m,1} - \mathbb{E}\left[V_{m,1}\right]\right)\boldsymbol{\delta} + \boldsymbol{\delta}^\top\left(\mathbb{E}\left[V_{m,1}\right] - V\right)\boldsymbol{\delta} \\
&\ge \Lambda_{\min}(V)\,\|\boldsymbol{\delta}\|_2^2 - \|(1-\mathbb{E})V_{m,1}\|_{\max}\,\|\boldsymbol{\delta}\|_1^2 + \boldsymbol{\delta}^\top\left(\mathbb{E}\left[V_{m,1}\right] - V\right)\boldsymbol{\delta} \\
&\ge \Lambda_{\min}(V)\,\|\boldsymbol{\delta}\|_2^2 - s\,\|(1-\mathbb{E})V_{m,1}\|_{\max}\,\|\boldsymbol{\delta}\|_2^2 - \left|\boldsymbol{\delta}^\top\left(\mathbb{E}\left[V_{m,1}\right] - V\right)\boldsymbol{\delta}\right|.
\end{aligned} \tag{85}$$

By (71) and (76),

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

By (67),

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

Therefore, (85) leads to

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

with probability tending to one. On the other hand,

$$\boldsymbol{\delta}^\top V_{m,1} \boldsymbol{\delta} \leq \|\boldsymbol{\delta}\|_1 \|V_{m,1}\boldsymbol{\delta}\|_\infty = O_{\mathbb{P}}\left(\sqrt{s}\lambda_n \|\boldsymbol{\delta}\|_2 \right),$$

Combining the two inequalities above, we finally get $\|\boldsymbol{\delta}\|_2 = O_{\mathbb{P}}(\sqrt{s}\lambda_n)$, which completes the proof. □

Proof of Theorem B.2

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

We will show Theorem B.2 by induction. Assume Theorem B.2 is true for t. Our assumption $\alpha > \alpha_0$ also ensures 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 $s^{3/2}\delta_{m,t} = o(1), s\delta_{m,t} = O\left(h^{3/2}\right)$. Then by Theorem B.1, by taking

$$\lambda_n^{(t+1)} = C_\lambda \left[\left(\frac{\log p}{n} \right)^{\frac{\alpha}{2\alpha+1}} + \left(\sqrt{\frac{s \log p}{mh^3}} + s\delta_{m,t} \right) \delta_{m,t} \right],$$

we have

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

and

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

with probability tending to 1. This completes the proof.

□

D 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 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 7. *Assume that there exists a neighborhood of 0 such that, for all $(\boldsymbol{\beta}^*, \rho, F) \in \Theta$ and all integers $1 \le k \le \alpha$, the k-th order derivative of $\rho(\cdot \mid \boldsymbol{Z})$ exists in this neighborhood 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 8. *Assume that ϵ and X are independent given $\boldsymbol{Z}$, and there exists a neighborhood of 0 such that, for all $(\boldsymbol{\beta}^*, \rho, F) \in \Theta$ and all integers $1 \le k \le \alpha + 1$, the k-th order derivative of $F(\cdot \mid \boldsymbol{Z})$ exists in this neighborhood for almost every $\boldsymbol{Z}$. Furthermore, there exists a constant $C_{\Theta,2} > 0$ such that $\sup_{\epsilon,\boldsymbol{Z},k} \left|F^{(k)}(\epsilon \mid \boldsymbol{Z})\right| < C_{\Theta,2}$.*

Assumption 9. *Assume that there exists a constant $c_\Theta > 0$ such that, for all $(\boldsymbol{\beta}^*, \rho, F) \in \Theta$, the matrix $V = 2\mathbb{E}\left[\rho(0 \mid \boldsymbol{Z}) F'(0 \mid \boldsymbol{Z}) \boldsymbol{Z}\boldsymbol{Z}^\top\right]$ satisfies $c_\Theta^{-1} < \Lambda_{\min}(V) < \Lambda_{\max}(V) < c_\Theta$, where $\Lambda_{\min}$ ($\Lambda_{\max}$) denotes the minimum (maximum) eigenvalue of V.*

Assumptions 7–9 for the distribution class Θ are parallel to Assumptions 2–4 for a fixed distribution. Assumptions 7–8 require α-order smoothness on a fixed neighborhood 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 9 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 D.1. *Assume Assumptions 1, 5, 7, 8 and 9 hold, and $\sup_{(\boldsymbol{\beta}^*,\rho,F)\in\Theta} \left\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\right\|_2 = O_{\mathbb{P}}(m^{-1/3})$. By choosing $h_t = \max\left\{(\lambda_h/n)^{\frac{1}{2\alpha+1}}, m^{-\frac{2^t}{3\alpha}}\right\}$ at iteration $t = 1, 2, \ldots, T$, when T satis-*

Table 1: The coverage rates (nominal 95%) of `(mSMSE)` in the first four iterations with different values of λ_h and $\log_m(n)$. The dimension $p = 1$ and the noise is homoscedastic normal.

λ_h	$\log_m(n)$	$t=1$	$t=2$	$t=3$	$t=4$	λ_h	$\log_m(n)$	$t=1$	$t=2$	$t=3$	$t=4$
	1.35	0.90	0.95	0.96	0.96		1.35	0.92	0.93	0.93	0.93
1	1.55	0.88	0.96	0.96	0.96	10	1.55	0.91	0.95	0.95	0.95
	1.75	0.83	0.95	0.95	0.95		1.75	0.89	0.93	0.93	0.93
	1.35	0.92	0.93	0.93	0.93		1.35	0.91	0.94	0.94	0.94
20	1.55	0.90	0.96	0.96	0.96	$\widehat{\lambda_h^*}$	1.55	0.93	0.95	0.95	0.95
	1.75	0.89	0.94	0.94	0.94		1.75	0.84	0.92	0.94	0.94

fies (13)*, we have:* $\forall \varepsilon > 0$, $\exists M_\varepsilon, N_\varepsilon$*, such that* $\forall n \geq N_\varepsilon$*, it holds that*

$$\sup_{(\boldsymbol{\beta}^*, \rho, F) \in \Theta} \mathbb{P}\left(\left\| \widehat{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^* \right\|_2 > M_\varepsilon n^{-\frac{\alpha}{2\alpha+1}} \right) < \varepsilon. \tag{86}$$

Note that (86) is equivalent to $\left\| \widehat{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^* \right\|_2 = O_{\mathbb{P}}(n^{-\frac{\alpha}{2\alpha+1}})$ if Θ only contains a single distribution. The proof of Theorem D.1 is almost the same as the proof of Proposition C.1 and Theorem 3.3, by noting that, for all distributions in Θ, the Taylor's expansion in (38) and the computation related to the bias in (39) are always correct. Moreover, the constant in the big O notation in (41) is uniform for all distributions in Θ, which is guaranteed by Assumptions 7 and 8.

E Additional Results in Simulations

E.1 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 3.3. An expression of the optimal value of λ_h^* is obtained in (15) to minimize 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 1 to 17 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 20, and report the coverage rates of the first four iterations of `(mSMSE)` in Table 1, with different λ_h and $\log_m(n)$. In the first iteration, it seems that larger λ_h's lead to higher coverage rates, which may indicate that a larger bandwidth improves the initial estimator more aggressively at the beginning of the

algorithm. However, after the (mSMSE) converges in two or three iterations, all estimators achieve near-nominal coverage rates, no matter what value λ_h is. Therefore, (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.

Table 2: The cpu times (in seconds) that different methods take to compute the estimator, with $p = 10$, $\log_m(n)$ from 1.35 to 1.75 and two types of noise.

Noise Type	$\log_m(n)$	(mSMSE) $t = 2$	(mSMSE) $t = 3$	(Avg-SMSE)	pooled-SMSE
Homoscedastic Normal	1.35	0.364	0.485	0.307	0.559
	1.45	0.365	0.492	0.328	1.356
	1.55	0.426	0.605	0.370	3.260
	1.65	0.443	0.627	0.399	9.267
	1.75	0.474	0.665	0.459	23.208
Homoscedastic Uniform	1.35	0.343	0.474	0.310	0.728
	1.45	0.355	0.492	0.330	1.749
	1.55	0.418	0.598	0.370	4.202
	1.65	0.431	0.616	0.400	11.893
	1.75	0.466	0.659	0.464	29.909
Heteroscedastic Normal	1.35	0.317	0.438	0.316	0.555
	1.45	0.379	0.503	0.334	1.322
	1.55	0.433	0.608	0.357	3.228
	1.65	0.452	0.634	0.396	9.132
	1.75	0.483	0.672	0.458	22.369

E.2 Time Complexity

In this section, we compare the computational complexity of each method. The average CPU times that each method takes when $p = 10$ are reported in Table 2. 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 2, 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.

E.3 Results for Other Types of Noise

Table 3: 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 1.35 to 1.75 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)`		
1.35	0.02	0.70	0.88	0.31	0.60	0.91	0.04	0.44	0.92
1.45	−0.14	0.48	0.88	0.17	0.32	0.94	−0.30	0.21	0.92
1.55	−0.16	0.26	0.84	0.11	0.12	0.92	−0.76	0.09	0.62
1.65	−0.18	0.16	0.80	0.09	0.04	0.94	−1.29	0.04	0.03
1.75	−0.12	0.08	0.78	0.08	0.03	0.94	−1.85	0.01	0.00
	`(mSMSE)` $t = 2$			`(Avg-MSE)`			pooled-SMSE		
1.35	0.32	0.61	0.9	−0.33	0.84	0.90	0.30	0.60	0.90
1.45	0.17	0.34	0.93	−0.38	0.22	0.92	0.15	0.32	0.94
1.55	0.12	0.14	0.93	−0.42	0.11	0.86	0.10	0.12	0.93
1.65	0.09	0.05	0.94	−0.36	0.07	0.80	0.07	0.04	0.96
1.75	0.08	0.03	0.94	−0.37	0.02	0.65	0.06	0.03	0.94

In this section, we report the bias, variance and coverage rates in Tables 3–6 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. These findings are all consistent with the results for the homoscedastic normal noise.

Table 4: 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 1.35 to 1.75 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)`		
1.35	−1.84	20.99	0.73	0.64	8.05	0.85	−2.30	2.63	0.91
1.45	−2.14	12.39	0.62	0.29	3.40	0.88	−4.44	1.57	0.55
1.55	−1.92	6.83	0.59	0.20	1.40	0.93	−6.13	0.74	0.02
1.65	−1.81	3.93	0.49	0.10	0.59	0.92	−7.18	0.33	0.00
1.75	−1.87	2.27	0.37	0.11	0.26	0.93	−7.97	0.10	0.00
	`(mSMSE)` $t = 2$			`(mSMSE)` $t = 4$			pooled-SMSE		
1.35	−0.12	11.71	0.78	0.55	6.27	0.84	0.62	4.76	0.93
1.45	0.52	5.06	0.86	0.62	2.65	0.89	0.29	2.63	0.93
1.55	0.12	2.03	0.85	0.27	1.26	0.94	0.08	1.24	0.94
1.65	0.06	0.95	0.86	0.17	0.58	0.92	0.07	0.56	0.92
1.75	0.11	0.31	0.89	0.11	0.26	0.94	0.05	0.26	0.93

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 1.35 to 1.75 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)`		
1.35	0.06	0.30	0.92	0.15	0.27	0.96	0.10	0.24	0.94
1.45	0.00	0.17	0.90	0.11	0.16	0.94	−0.03	0.11	0.94
1.55	−0.06	0.09	0.88	0.05	0.06	0.94	−0.29	0.07	0.86
1.65	−0.04	0.04	0.90	0.04	0.03	0.98	−0.60	0.03	0.30
1.75	−0.05	0.02	0.88	0.03	0.01	0.96	−0.96	0.01	0.00
	`(mSMSE)` $t = 2$			`(Avg-MSE)`			pooled-SMSE		
1.35	0.14	0.26	0.96	−0.40	0.51	0.89	0.14	0.26	0.96
1.45	0.11	0.16	0.94	−0.36	0.22	0.88	0.10	0.16	0.94
1.55	0.05	0.06	0.94	−0.40	0.12	0.82	0.04	0.06	0.94
1.65	0.04	0.03	0.98	−0.42	0.03	0.69	0.03	0.03	0.97
1.75	0.04	0.01	0.96	−0.41	0.01	0.40	0.02	0.01	0.96

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 1.35 to 1.75 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)`		
1.35	−0.67	19.36	0.82	0.74	4.86	0.97	−0.24	4.08	0.96
1.45	−1.25	12.14	0.72	0.36	2.20	0.96	−1.46	1.88	0.84
1.55	−1.39	11.72	0.63	0.18	1.22	0.94	−2.47	0.93	0.34
1.65	−1.29	6.05	0.56	0.17	0.64	0.95	−3.28	0.38	0.00
1.75	−1.18	2.91	0.49	0.14	0.35	0.92	−3.84	0.16	0.00
	`(mSMSE)` $t = 2$			`(mSMSE)` $t = 4$			pooled-SMSE		
1.35	0.78	6.04	0.96	0.79	4.61	0.97	0.69	4.59	0.97
1.45	0.31	2.75	0.93	0.36	2.17	0.96	0.27	2.15	0.95
1.55	0.18	1.63	0.91	0.19	1.20	0.94	0.10	1.20	0.94
1.65	0.13	0.74	0.92	0.17	0.63	0.95	0.09	0.63	0.95
1.75	0.12	0.39	0.90	0.14	0.35	0.93	0.05	0.35	0.92

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

Candes, E. and T. Tao (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics 35*(6), 2313–2351.

Feng, H., Y. Ning, and J. Zhao (2022). Nonregular and minimax estimation of individualized thresholds in high dimension with binary responses. *Annals of Statistics*, To appear.

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

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

Mukherjee, D., M. Banerjee, and Y. Ritov (2019). Non-standard asymptotics in high dimensions: Manski's maximum score estimator revisited. *arXiv preprint arXiv:1903.10063*.