The Outer Product of Gradient Estimation in High Dimensions and its Application

Technical Proofs

Zhibo Cai, Yingcun Xia and Weiqiang Hang

National University of Singapore

S.1. Justification of Assumption (A1)

As stated in the discussion of Appendix A.1, the following proposition and remarks show that the condition in (A1) is satisfied in many cases.

Proposition 1. *In model (2.1), suppose* $Y = m(\mathbf{X}) + \varepsilon = g(B_0^\top \mathbf{X}) + \varepsilon$ *where* $\varepsilon \perp \mathbf{X}$*, and* $\nabla g(\cdot)$ *is bounded, and* $B_0 : p \times d$ *is the dimension reduction directions. Assuming that* $\mathbf{X}$ *has a compact support and it is block-wise independence in the sense that* $X^{[i]}$ *and* $X^{[j]}$ *are independent when* $|i-j| > T$ *for some positive integer* T*. Suppose the distance variances* $dVar(X^{[j]})$ *are bounded away from 0, i.e.* $\min\{dVar(X^{[1]}), \dots, dVar(X^{[p]})\} > c$ *for some positive constant* c*. Then, there exists* γ *such that*

$$\alpha_1 + \alpha_2 + \cdots + \alpha_p \leq \gamma,$$

where $\gamma = \gamma(d)$ *is only related to* d*.*

Proof of Proposition 1. For notation simplicity, we consider $X^{[T+1]}$ as an example. Let $\mathbf{X}_1^{[T+1]} = \big(X^{[1]}, \dots, X^{[2T+1]}\big)$ and $\mathbf{X}_2^{[T+1]} = \big(X^{[2T+2]}, \dots, X^{[p]}\big)$, the Taylor's series of $m(\mathbf{X})$ with a mean-value form remainder is

$$m(\mathbf{X}) = \mathbf{X}_1^{[T+1]} \frac{\partial m(\mathbf{x})}{\partial \mathbf{x}_1^{[T+1]}}\bigg|_{\mathbf{x}=(\tilde{\mathbf{X}}_1^{[T+1]}, \mathbf{X}_2^{[T+1]})} + m(\mathbf{0}, \mathbf{X}_2^{[T+1]}) := Z_1 + Z_2, \tag{S.1}$$

where $\tilde{\mathbf{X}}_1^{[T+1]}$ is a point between $\mathbf{0}$ and $\mathbf{X}_1^{[T+1]}$. It is obvious that Z_2 is independent with $X^{[T+1]}$

by the block-wise independence condition. Since $\mathbf{X}$ has a compact support, and $\nabla g(\cdot)$ is bounded,

$$\begin{aligned}|Z_1| &\le \sum_{i=1}^{2T+1} \big|X^{[i]}\beta^{[i]}\nabla g\big(B_0^\top(\tilde{\mathbf{X}}_1^{[T+1]}, \mathbf{X}_2^{[T+1]})^\top\big)\big| \\ &\le \sum_{i=1}^{2T+1} \big|X^{[i]}\big|\|\beta^{[i]}\|\|\nabla g\big(B_0^\top(\tilde{\mathbf{X}}_1^{[T+1]}, \mathbf{X}_2^{[T+1]})^\top\big)\| \\ &\le A_1A_2\sum_{i=1}^{2T+1}\|\beta^{[i]}\|, \quad \text{almost surely}\end{aligned} \tag{S.2}$$

where $A_1 = \sup_{\mathbf{v}\in\mathbb{R}^q}\|\nabla g(\mathbf{v})\|$ and $\sup_i |X^{[i]}| \le A_2$ almost surely.

By the definition of R_{T+1} given in (2.5), we can denote it by $R_{T+1} = Y - c_{T+1}\cdot X^{[T+1]}$, where $c_{T+1} = Cov(X^{[T+1]}, Y)/\sqrt{Var(X^{[T+1]})}$. We have

$$\begin{aligned}Cov(X^{[T+1]}, Y) &= Cov(X^{[T+1]}, m(\mathbf{X}) + \epsilon) = Cov(X^{[T+1]}, Z_1 + Z_2 + \epsilon) \\ &= Cov(X^{[T+1]}, Z_1) = \mathbf{E}(X^{[T+1]}\cdot Z_1) - \mathbf{E}(X^{[T+1]})\cdot\mathbf{E}(Z_1).\end{aligned}$$

Therefore, by (S.2),

$$\big|Cov(X^{[T+1]}, Y)\big| \le \big|\mathbf{E}(X^{[T+1]}\cdot Z_1)\big| + \big|\mathbf{E}(X^{[T+1]})\cdot\mathbf{E}(Z_1)\big| \le 2A_1A_2^2\sum_{i=1}^{2T+1}\|\beta^{[i]}\|.$$

We can conclude by the conditions that

$$c_{T+1} \le A_3\sum_{i=1}^{2T+1}\|\beta^{[i]}\|. \tag{S.3}$$

Denote the characteristic function of any random variable U by $\varphi_U(t) = \mathbf{E}\big[e^{\sqrt{-1}tU}\big]$, then

$$\begin{aligned}&\big|\varphi_{X^{[T+1]}, Y-c_{T+1}\cdot X^{[T+1]}}(s,t) - \varphi_{X^{[T+1]}}(s)\varphi_{Y-c_{T+1}\cdot X^{[T+1]}}(t)\big|^2 \\ &= \big|\varphi_{X^{[T+1]}, Z_1+Z_2+\varepsilon-c_{T+1}\cdot X^{[T+1]}}(s,t) - \varphi_{X^{[T+1]}}(s)\varphi_{Z_1+Z_2+\varepsilon-c_{T+1}\cdot X^{[T+1]}}(t)\big|^2 \\ &= \big|\varphi_\varepsilon(t)\big|^2\big|\varphi_{X^{[T+1]}, Z_1+Z_2-c_{T+1}\cdot X^{[T+1]}}(s,t) - \varphi_{X^{[T+1]}}(s)\varphi_{Z_1+Z_2-c_{T+1}\cdot X^{[T+1]}}(t)\big|^2 \\ &\le \big|\varphi_{X^{[T+1]}, Z_1+Z_2-c_{T+1}\cdot X^{[T+1]}}(s,t) - \varphi_{X^{[T+1]}}(s)\varphi_{Z_1+Z_2-c_{T+1}\cdot X^{[T+1]}}(t)\big|^2 \\ &\le 2\big|\varphi_{X^{[T+1]}, Z_1+Z_2-c_{T+1}\cdot X^{[T+1]}}(s,t) - \varphi_{X^{[T+1]}, Z_2}(s,t)\big|^2 \\ &\quad + 2\big|\varphi_{X^{[T+1]}, Z_2}(s,t) - \varphi_{X^{[T+1]}}(s)\varphi_{Z_1+Z_2-c_{T+1}\cdot X^{[T+1]}}(t)\big|^2 \\ &= 2\big|\mathbf{E}\big[e^{\sqrt{-1}(sX^{[T+1]}+tZ_2)}(e^{\sqrt{-1}t(Z_1-c_{T+1}\cdot X^{[T+1]})} - 1)\big]\big|^2 \\ &\quad + 2\big|\varphi_{X^{[T+1]}}(s)\big|^2\big|\mathbf{E}\big[e^{\sqrt{-1}tZ_2}(e^{\sqrt{-1}t(Z_1-c_{T+1}\cdot X^{[T+1]})} - 1)\big]\big|^2 \\ &\le 2\mathbf{E}\big|e^{\sqrt{-1}(sX^{[T+1]}+tZ_2)}(e^{\sqrt{-1}t(Z_1-c_{T+1}\cdot X^{[T+1]})} - 1)\big|^2 \\ &\quad + 2\big|\varphi_{X^{[T+1]}}(s)\big|^2\mathbf{E}\big|e^{\sqrt{-1}tZ_2}\big|^2\mathbf{E}\big|e^{\sqrt{-1}t(Z_1-c_{T+1}\cdot X^{[T+1]})} - 1\big|^2 \\ &= 2\mathbf{E}\big|e^{\sqrt{-1}t(Z_1-c_{T+1}\cdot X^{[T+1]})} - 1\big|^2 + 2\big|\varphi_{X^{[T+1]}}(s)\big|^2\mathbf{E}\big|e^{\sqrt{-1}t(Z_1-c_{T+1}\cdot X^{[T+1]})} - 1\big|^2.\end{aligned} \tag{S.4}$$

By (S.2) and (S.3), there exists a constant $A_4 > 0$ such that

$$\left|Z_1 - c_{T+1} \cdot X^{[T+1]}\right| \le A_4 \sum_{i=1}^{2T+1} \|\beta^{[i]}\|. \tag{S.5}$$

Then, using (S.4) and (S.5) and the definition of distance covariance (Székely et al., 2007), there is a constant A_5 such that

$$dCov^2(R_{T+1}, X^{[T+1]}) \le A_5 \bigg(\sum_{i=1}^{2T+1} \|\beta^{[i]}\| \bigg)^2.$$

By the condition of distance variance, we have

$$dCor^2(R_{T+1}, X^{[T+1]}) \le A_6 \bigg(\sum_{i=1}^{2T+1} \|\beta^{[i]}\| \bigg)^2 \le A_6 \cdot (2T+1) \sum_{i=1}^{2T+1} \|\beta^{[i]}\|^2,$$

for some positive constant A_6. Moreover, note that all the constants mentioned above are uniform for $j = 1, 2, \ldots, p$, we have

$$\begin{aligned}
\alpha_1 + \alpha_2 + \cdots + \alpha_p &= \sum_{j=1}^{p} dCor^2(X^{[j]}, R_j) \\
&\le \sum_{j=1}^{p} A_6(2T+1) \sum_{i=j-T}^{j+T} \|\beta^{[i]}\|^2 \\
&= A_6(2T+1) \sum_{i=j-T}^{j+T} \sum_{j=1}^{p} \|\beta^{[i]}\|^2 \\
&= A_6(2T+1)^2 d.
\end{aligned}$$

Therefore, the statement follows by selecting $\gamma_0 = A_6(2T+1)^2 d$. $\square$

Remark 1. The block-wise independence can also be extended to serial dependence with appropriate decreasing rate as commonly used in time series analysis, such as α-mixing assumption; see Fan and Yao (2008). In addition, γ can be proportional to d under these conditions. Thus, $d = o(\log(n))$ by condition (A1), which means that the reduced dimension d is also able to diverge but should has a sub-logarithm rate.

S.2. Proofs of Theorems 1 and Theorem 2

We first consider the partial derivatives of regression function $m(\cdot)$. Since the first and second order partial derivatives of function $g(\cdot)$ are bounded under Assumption (A2), the partial derivatives of $m(\cdot)$ will be controlled by the coefficients in B_0.

Lemma 1. *Under conditions (A2), we have for $1 \le j, k \le p$,*

$$\sup_{\mathbf{x} \in \mathbb{R}^p} \left| \frac{\partial m(\mathbf{x})}{\partial x^{[j]}} \right| = O\big(\|\beta^{[j]}\|\big),$$
$$\sup_{\mathbf{x} \in \mathbb{R}^p} \left| \frac{\partial^2 m(\mathbf{x})}{\partial x^{[j]} \partial x^{[k]}} \right| = O\big(\|\beta^{[j]}\| \cdot \|\beta^{[k]}\|\big).$$

Proof of Lemma 1. The first order partial derivative with respect to $x^{[j]}$ is

$$\sup_{\mathbf{x} \in \mathbb{R}^p} \left| \frac{\partial m(\mathbf{x})}{\partial x^{[j]}} \right| = \sup_{\mathbf{x} \in \mathbb{R}^p} \big| [\beta^{[j]}]^\top \nabla g(B_0^\top \mathbf{x}) \big| \le \sup_{\mathbf{x} \in \mathbb{R}^p} \|\beta^{[j]}\| \|\nabla g(B_0^\top \mathbf{x})\| = O\big(\|\beta^{[j]}\|\big), \tag{S.6}$$

where $\nabla g(B_0^\top \mathbf{x})$ is the gradient of $g(\cdot)$ at point $B_0^\top \mathbf{x}$. In (S.6), the inequality holds by Cauchy-Schwartz inequality, and the last equation holds by Condition (A2). Similarly,

$$\begin{aligned} \sup_{\mathbf{x} \in \mathbb{R}^p} \left| \frac{\partial^2 m(\mathbf{x})}{\partial x^{[j]} \partial x^{[k]}} \right| &= \sup_{\mathbf{x} \in \mathbb{R}^p} \big| [\beta^{[j]}]^\top \mathcal{H}_g(B_0^\top \mathbf{x}) \beta^{[k]} \big| \\ &\le \sup_{\mathbf{x} \in \mathbb{R}^p} \|\beta^{[j]}\| \|\mathcal{H}_g(B_0^\top \mathbf{x}) \beta^{[k]}\| \le \|\beta^{[j]}\| \|\beta^{[k]}\| \sup_{\mathbf{x} \in \mathbb{R}^p} \|\mathcal{H}_g(B_0^\top \mathbf{x})\| \\ &= O\big(\|\beta^{[j]}\| \cdot \|\beta^{[k]}\|\big), \end{aligned} \tag{S.7}$$

where $\mathcal{H}_g(B_0^\top \mathbf{x})$ is the Hessian matrix of $g(\cdot)$ at point $B_0^\top \mathbf{x}$, and $\|\mathcal{H}_g(B_0^\top \mathbf{x})\|$ represents its largest eigenvalue. In (S.7), the first inequality holds by Cauchy-Schwartz inequality, while the second inequality holds because of the property of eigenvalue. $\square$

For ease of exposition, we introduce the following notations. A local approximation of $m(\mathbf{z})$ by a polynomial of total order r is given as

$$m(\mathbf{z}) \approx \sum_{0 \le |\mathbf{k}| \le r} \frac{1}{\mathbf{k}!} (D^{\mathbf{k}} m)(\mathbf{x}) (\mathbf{z} - \mathbf{x})^{\mathbf{k}}, \tag{S.8}$$

where

$$\mathbf{k} = \big(k^{[1]}, \dots, k^{[p]}\big), \mathbf{k}! = k^{[1]}! \times \cdots \times k^{[p]}!, \ \big|\mathbf{k}\big| = \sum_{j=1}^{p} k^{[j]};$$

$$\mathbf{x}^{\mathbf{k}} = (x^{[1]})^{k^{[1]}} \times \cdots \times (x^{[p]})^{k^{[p]}}, \ \sum_{0 \le |\mathbf{k}| \le r} = \sum_{j=0}^{r} \sum_{\substack{k^{[1]}=0 \\ k^{[1]} + \cdots + k^{[p]} = j}}^{j} \cdots \sum_{k^{[p]}=0}^{j};$$

and

$$(D^{\mathbf{k}} m)(\mathbf{x}) = \frac{\partial^{\mathbf{k}} m(\mathbf{y})}{\partial (y^{[1]})^{k^{[1]}} \cdots \partial (y^{[p]})^{k^{[p]}}} \bigg|_{\mathbf{y} = \mathbf{x}}.$$

By linear regression and distance correlation estimation (Székely et al., 2007), $\hat{\boldsymbol{\alpha}} \to \boldsymbol{\alpha}$ and the rate of convergence is $O_p\big(n^{-1/2}\big)$. Then, by condition (A3)

$$\begin{aligned} \big|K_h\big(\mathbf{x};\hat{\boldsymbol{\alpha}}\big)-K_h\big(\mathbf{x};\boldsymbol{\alpha}\big)\big| &\le C\cdot\left\|\big(\frac{x^{[1]}}{h^{\hat{\alpha}_1}},\ldots,\frac{x^{[p]}}{h^{\hat{\alpha}_p}}\big)-\big(\frac{x^{[1]}}{h^{\alpha_1}},\ldots,\frac{x^{[p]}}{h^{\alpha_p}}\big)\right\| \\ &= C\cdot\left\|\big(\frac{h^{\alpha_1-\hat{\alpha}_1}-1}{h^{\alpha_1}}x^{[1]},\ldots,\frac{h^{\alpha_p-\hat{\alpha}_p}-1}{h^{\alpha_p}}x^{[p]}\big)\right\| \\ &= O_p\big(\frac{-\log(h)}{h\sqrt{n}}\cdot\|\mathbf{x}\|\big), \end{aligned} \tag{S.9}$$

where the last equation holds because of $\hat{\alpha}_j-\alpha_j=O_p(n^{-1/2})$ and $h^{\alpha_j}\ge h$ for all j. The estimation problem can be written as minimizing

$$\sum_{i=1}^{n}\Big[Y_i-\sum_{0\le|\mathbf{k}|\le 1}b_{\mathbf{k}}(\mathbf{x})(\mathbf{X}_i-\mathbf{x})^{\mathbf{k}}\Big]^2K_h\big(\mathbf{X}_i-\mathbf{x};\hat{\boldsymbol{\alpha}}\big) \tag{S.10}$$

with respect to $b_{\mathbf{k}}(\mathbf{x})$. Denote the minimizer of (S.10) by $\hat{b}_{\mathbf{k}}(\mathbf{x})$, then we have estimation $(\widehat{D^{\mathbf{k}}m})(\mathbf{x}) = \mathbf{k}!\hat{b}_{\mathbf{k}}(\mathbf{x})$. The minimization of (S.10) leads to the set of equations

$$t_{\mathbf{j}}(\mathbf{x})=\sum_{0\le|\mathbf{k}|\le 1}h^{\mathbf{k}\cdot\boldsymbol{\alpha}}\hat{b}_{\mathbf{k}}(\mathbf{x})s_{\mathbf{j}+\mathbf{k}}(\mathbf{x}),\quad 0\le\big|\mathbf{j}\big|\le 1, \tag{S.11}$$

where

$$\begin{aligned} t_{\mathbf{j}}(\mathbf{x}) &= \frac{1}{n}\sum_{i=1}^{n}Y_i[\mathbf{Z}_i(h;\boldsymbol{\alpha})-\mathbf{z}(h;\boldsymbol{\alpha})]^{\mathbf{j}}K_h\big(\mathbf{X}_i-\mathbf{x};\hat{\boldsymbol{\alpha}}\big), \\ s_{\mathbf{j}}(\mathbf{x}) &= \frac{1}{n}\sum_{i=1}^{n}[\mathbf{Z}_i(h;\boldsymbol{\alpha})-\mathbf{z}(h;\boldsymbol{\alpha})]^{\mathbf{j}}K_h\big(\mathbf{X}_i-\mathbf{x};\hat{\boldsymbol{\alpha}}\big), \end{aligned} \tag{S.12}$$

with

$$\mathbf{z}(h;\boldsymbol{\alpha})=\big(\frac{x^{[1]}}{h^{\alpha_1}},\ldots,\frac{x^{[p]}}{h^{\alpha_p}}\big). \tag{S.13}$$

Define $\tau(\mathbf{x})=\big(\tau_0(\mathbf{x}),\ldots,\tau_p(\mathbf{x})\big)^\top$, where $\tau_0(\mathbf{x})=t_{(0,\ldots,0)}(\mathbf{x}),\tau_1(\mathbf{x})=t_{(1,\ldots,0)}(\mathbf{x}),\ldots,\tau_p(\mathbf{x})=t_{(0,\ldots,1)}(\mathbf{x})$. Arranging $h^{\mathbf{k}\cdot\boldsymbol{\alpha}}\hat{b}_{\mathbf{k}}(\mathbf{x}), 0\le|\mathbf{k}|\le 1$ in the same order, we can obtain $\hat{\boldsymbol{\theta}}$ as an estimator of column vector $\boldsymbol{\theta}(\mathbf{x})=\big(\theta_0(\mathbf{x}),\ldots,\theta_p(\mathbf{x})\big)^\top := \big(m(\mathbf{x}),h^{\alpha_1}m^{[1]}(\mathbf{x}),\ldots,h^{\alpha_p}m^{[p]}(\mathbf{x})\big)^\top$. Then, let $\mathbf{S}(\mathbf{x})$ be a $(p+1)\times(p+1)$ matrix, where the (k,l) entry is $s_{\mathbf{j}}(\mathbf{x})$ with $(k-1)$-th and $(l-1)$-th elements in $\mathbf{j}$ are 1 for $1\le k,l\le p$ and $k\ne l$. Other entries in $\mathbf{S}(\mathbf{x})$ can be obtained similarly. Thus, the set of equations in (S.11) can be written in matrix as

$$\tau(\mathbf{x})=\mathbf{S}(\mathbf{x})\hat{\theta}(\mathbf{x}).$$

Since $\mathbf{S}(\mathbf{x})$ is positive semi-definite when $K(\cdot) > 0$, we can henceforth assume the matrix is invertible and write

$$\hat{\theta}(\mathbf{x}) = \mathbf{S}^{-1}(\mathbf{x})\tau(\mathbf{x}),$$

as the solution of the set of equations (S.11).

A fundamental decomposition for the error $\hat{\theta}(\mathbf{x}) - \theta(\mathbf{x})$ is provided next. Firstly, let

$$t^*_{\mathbf{j}}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}[Y_i - m(\mathbf{X}_i)][\mathbf{Z}_i(h;\boldsymbol{\alpha}) - \mathbf{z}(h;\boldsymbol{\alpha})]^{\mathbf{j}} K_h\big(\mathbf{X}_i - \mathbf{x};\hat{\boldsymbol{\alpha}}\big), \tag{S.14}$$

we have

$$t_{\mathbf{j}}(\mathbf{x}) - t^*_{\mathbf{j}}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} m(\mathbf{X}_i)[\mathbf{Z}_i(h;\boldsymbol{\alpha}) - \mathbf{z}(h;\boldsymbol{\alpha})]^{\mathbf{j}} K_h\big(\mathbf{X}_i - \mathbf{x};\hat{\boldsymbol{\alpha}}\big). \tag{S.15}$$

The Taylor series of $m(\mathbf{X}_i)$ at point $\mathbf{x}$ with a mean-value form of remainder is

$$m(\mathbf{X}_i) = \sum_{0\le|\mathbf{k}|\le 1}\frac{1}{\mathbf{k}!}(D^{\mathbf{k}}m)(\mathbf{x})(\mathbf{X}_i - \mathbf{x})^{\mathbf{k}} + \sum_{|\mathbf{k}|=2}\frac{1}{\mathbf{k}!}(D^{\mathbf{k}}m)(\tilde{\mathbf{x}}_i)(\mathbf{X}_i - \mathbf{x})^{\mathbf{k}}, \tag{S.16}$$

where $\tilde{\mathbf{x}}_i$ is a point between $\mathbf{x}$ and $\mathbf{X}_i$. Substituting (S.16) and (S.12) to (S.15), we find

$$t_{\mathbf{j}}(\mathbf{x}) - t^*_{\mathbf{j}}(\mathbf{x}) = \sum_{0\le|\mathbf{k}|\le 1}\frac{1}{\mathbf{k}!}h^{\mathbf{k}\cdot\boldsymbol{\alpha}}(D^{\mathbf{k}}m)(\mathbf{x})s_{\mathbf{j}+\mathbf{k}}(\mathbf{x}) + e_{\mathbf{j}}(\mathbf{x}),$$

where

$$e_{\mathbf{j}}(\mathbf{x}) = \frac{1}{n}\sum_{|\mathbf{k}|=2}\frac{h^{\mathbf{k}\cdot\boldsymbol{\alpha}}}{\mathbf{k}!}\sum_{i=1}^{n}(D^{\mathbf{k}}m)(\tilde{\mathbf{x}}_i)[\mathbf{Z}_i(h;\boldsymbol{\alpha}) - \mathbf{z}(h;\boldsymbol{\alpha})]^{\mathbf{k}+\mathbf{j}} K_h\big(\mathbf{X}_i - \mathbf{x};\hat{\boldsymbol{\alpha}}\big). \tag{S.17}$$

By (S.11) and $(D^{\mathbf{k}}m)(\mathbf{x}) = \mathbf{k}!b_{\mathbf{k}}(\mathbf{x})$, we obtain

$$t^*_{\mathbf{j}}(\mathbf{x}) = \sum_{0\le|\mathbf{k}|\le 1} h^{\mathbf{k}\cdot\boldsymbol{\alpha}}[\hat{b}_{\mathbf{k}}(\mathbf{x}) - b_{\mathbf{k}}(\mathbf{x})]s_{\mathbf{j}+\mathbf{k}}(\mathbf{x}) - e_{\mathbf{j}}(\mathbf{x}). \tag{S.18}$$

For $0 \le |\mathbf{j}| \le 1$, using the same arrangement as for $\tau(\mathbf{x})$, we can define the $(p+1)$ column vector $\tau^*(\mathbf{x})$ and $\mathbf{e}(\mathbf{x})$ as follows

$$\tau^*(\mathbf{x}) = \begin{bmatrix} t^*_{(0,\ldots,0)}(\mathbf{x}) \\ t^*_{(1,\ldots,0)}(\mathbf{x}) \\ \vdots \\ t^*_{(0,\ldots,1)}(\mathbf{x}) \end{bmatrix}, \quad \mathbf{e}(\mathbf{x}) = \begin{bmatrix} e_{(0,\ldots,0)}(\mathbf{x}) \\ e_{(1,\ldots,0)}(\mathbf{x}) \\ \vdots \\ e_{(0,\ldots,1)}(\mathbf{x}) \end{bmatrix}.$$

The vector form of (S.18) is

$$\tau^*(\mathbf{x}) = \mathbf{S}(\mathbf{x})\big(\hat{\theta}(\mathbf{x}) - \theta(\mathbf{x})\big) - \mathbf{e}(\mathbf{x}).$$

Thus,

$$\hat{\theta}(\mathbf{x}) - \theta(\mathbf{x}) = \mathbf{S}^{-1}(\mathbf{x})\tau^*(\mathbf{x}) + \mathbf{S}^{-1}(\mathbf{x})\mathbf{e}(\mathbf{x}). \tag{S.19}$$

In the following proof, we use C to represent any positive constants that may be different from case to case.

Lemma 2. *Let D be any compact subset of $\mathbb{R}^p$ and conditions (A1) and (A3) hold. Assume $h = h_n \to 0$ and $p_n \log(n)/(nh_n^{|\boldsymbol{\alpha}|}) \to \infty$ as $n \to \infty$. Then, for each $\mathbf{j}$ with $0 \le |\mathbf{j}| \le 3$,*

$$\sup_{\mathbf{x}\in D} \big|s_{\mathbf{j}}(\mathbf{x}) - \mathbf{E}[s_{\mathbf{j}}(\mathbf{x})]\big| = O\bigg(\big(\frac{p_n \log(n)}{nh_n^{|\boldsymbol{\alpha}|}}\big)^{1/2}\bigg) \quad a.s.$$

Proof of Lemma 2. By the condition of random vector $\mathbf{X}$ in (A1), we have $f_{\mathbf{X}}(\mathbf{x}) < C_1$ on the support. By condition (A1), (A3) and equation (S.9),

$$\begin{aligned}
\big|\mathbf{E}[s_{\mathbf{j}}(\mathbf{x})]\big| &= \int_{\mathbb{R}^p} \big(\frac{u^{[1]} - x^{[1]}}{h^{\alpha_1}}, \cdots, \frac{u^{[p]} - x^{[p]}}{h^{\alpha_p}}\big)^{\mathbf{j}} K_h(\mathbf{u} - \mathbf{x}; \boldsymbol{\alpha}) f_{\mathbf{X}}(\mathbf{u}) d\mathbf{u} + O(\frac{-\log(h)}{h\sqrt{n}}) \\
&\le C_1 \int_{\mathbb{R}^p} \big(\frac{u^{[1]} - x^{[1]}}{h^{\alpha_1}}, \cdots, \frac{u^{[p]} - x^{[p]}}{h^{\alpha_p}}\big)^{\mathbf{j}} K_h(\mathbf{u} - \mathbf{x}; \boldsymbol{\alpha}) d\mathbf{u} + O(\frac{-\log(h)}{h\sqrt{n}}) \\
&= C_1 \int_{\mathbb{R}^p} \mathbf{t}^{\mathbf{j}} K(\mathbf{t}) d\mathbf{t} + O(\frac{-\log(h)}{h\sqrt{n}}) < \infty.
\end{aligned} \tag{S.20}$$

Thus, we have $\big|\mathbf{E}[s_{\mathbf{j}}(\mathbf{x})]\big| = O(1)$. Next, by (S.9), (S.20) and condition (A3)

$$\begin{aligned}
& nh_n^{|\boldsymbol{\alpha}|} Var[s_{\mathbf{j}}(\mathbf{x})] = h_n^{|\boldsymbol{\alpha}|} Var\bigg(\big[\mathbf{Z}_1(h_n; \boldsymbol{\alpha}) - \mathbf{z}(h_n; \boldsymbol{\alpha})\big]^{\mathbf{j}} K_{h_n}\big(\mathbf{X}_1 - \mathbf{x}; \boldsymbol{\alpha}\big)\bigg) + o\big(h_n^{|\boldsymbol{\alpha}|}\big) \\
&= h_n^{|\boldsymbol{\alpha}|} \mathbf{E}\bigg(\big[\mathbf{Z}_1(h_n; \boldsymbol{\alpha}) - \mathbf{z}(h_n; \boldsymbol{\alpha})\big]^{2\mathbf{j}} K_{h_n}^2\big(\mathbf{X}_1 - \mathbf{x}; \boldsymbol{\alpha}\big)\bigg) - h_n^{|\boldsymbol{\alpha}|}\big(\mathbf{E}[s_{\mathbf{j}}(\mathbf{x})]\big)^2 + o\big(h_n^{|\boldsymbol{\alpha}|}\big) \\
&= \int_{\mathbb{R}^p} \big(\frac{u^{[1]} - x^{[1]}}{h_n^{\alpha_1}}, \cdots, \frac{u^{[p]} - x^{[p]}}{h_n^{\alpha_p}}\big)^{2\mathbf{j}} \big[\frac{1}{h_n^{|\boldsymbol{\alpha}|}} K^2\big(\frac{u^{[1]} - x^{[1]}}{h_n^{\alpha_1}}, \cdots, \frac{u^{[p]} - x^{[p]}}{h_n^{\alpha_p}}\big)\big] f_{\mathbf{X}}(\mathbf{u}) d\mathbf{u} + O(h_n^{|\boldsymbol{\alpha}|}) \\
&= O(1) + O(h_n^{|\boldsymbol{\alpha}|}) = O(1).
\end{aligned} \tag{S.21}$$

Let $L = L(n) = \bigg\lceil \big(\frac{n}{h_n^{|\boldsymbol{\alpha}|+2|\mathbf{j}|+2}\log(n)}\big)^{1/2} \bigg\rceil^p$, where $\lceil \cdot \rceil$ represents ceiling function. For computation simplicity, assume $\big(\frac{n}{h_n^{|\boldsymbol{\alpha}|+2|\mathbf{j}|+2}\log(n)}\big)^{1/2}$ is a positive integer, and $L = L(n) = \big[\frac{n}{h_n^{|\boldsymbol{\alpha}|+2|\mathbf{j}|+2}\log(n)}\big]^{p/2}$. Since D is compact, it can be covered by $L(n)$ cubes $I_k = I_{n,k}$ centered at $\mathbf{x}_k$ with side length ℓ_n for $k = 1, \dots, L(n)$. Clearly, $\ell_n \le C_2/L^{1/p}(n)$ for some positive constant C_2. Then, we can write

$$\begin{aligned}
& \sup_{\mathbf{x}\in D} \big|s_{\mathbf{j}}(\mathbf{x}) - \mathbf{E}[s_{\mathbf{j}}(\mathbf{x})]\big| = \max_{1\le k\le L(n)} \sup_{\mathbf{x}\in D\cap I_k} \big|s_{\mathbf{j}}(\mathbf{x}) - \mathbf{E}[s_{\mathbf{j}}(\mathbf{x})]\big| \\
&\le \max_{1\le k\le L(n)} \sup_{\mathbf{x}\in D\cap I_k} \big|s_{\mathbf{j}}(\mathbf{x}) - s_{\mathbf{j}}(\mathbf{x}_k)\big| + \max_{1\le k\le L(n)} \big|s_{\mathbf{j}}(\mathbf{x}_k) - \mathbf{E}[s_{\mathbf{j}}(\mathbf{x}_k)]\big| \\
&\quad + \max_{1\le k\le L(n)} \sup_{\mathbf{x}\in D\cap I_k} \big|\mathbf{E}[s_{\mathbf{j}}(\mathbf{x}_k)] - \mathbf{E}[s_{\mathbf{j}}(\mathbf{x})]\big| \\
&= I + II + III.
\end{aligned}$$

Since both $\mathbf{X}$ and $K(\cdot)$ have compact support in $\mathbb{R}^p$, by (S.12), (S.13), condition (A1) and (A3),

$$
\begin{aligned}
&\big|s_{\mathbf{j}}(\mathbf{x})-s_{\mathbf{j}}(\mathbf{x}_k)\big|\\
=&\frac{1}{nh_n^{|\boldsymbol{\alpha}|}}\Big|\sum_{i=1}^{n}[\mathbf{Z}_i(h;\boldsymbol{\alpha})-\mathbf{z}(h;\boldsymbol{\alpha})]^{\mathbf{j}}K_h\big(\mathbf{X}_i-\mathbf{x};\hat{\boldsymbol{\alpha}}\big)-[\mathbf{Z}_i(h;\boldsymbol{\alpha})-\mathbf{z}_k(h;\boldsymbol{\alpha})]^{\mathbf{j}}K_h\big(\mathbf{X}_i-\mathbf{x}_k;\hat{\boldsymbol{\alpha}}\big)\Big|\\
\leq&\frac{1}{nh_n^{|\boldsymbol{\alpha}|}}\sum_{i=1}^{n}\Big|[\mathbf{Z}_i(h;\boldsymbol{\alpha})-\mathbf{z}(h;\boldsymbol{\alpha})]^{\mathbf{j}}K_h\big(\mathbf{X}_i-\mathbf{x};\hat{\boldsymbol{\alpha}}\big)-[\mathbf{Z}_i(h;\boldsymbol{\alpha})-\mathbf{z}_k(h;\boldsymbol{\alpha})]^{\mathbf{j}}K_h\big(\mathbf{X}_i-\mathbf{x};\hat{\boldsymbol{\alpha}}\big)\Big|\\
&\qquad+\Big|[\mathbf{Z}_i(h;\boldsymbol{\alpha})-\mathbf{z}_k(h;\boldsymbol{\alpha})]^{\mathbf{j}}K_h\big(\mathbf{X}_i-\mathbf{x};\hat{\boldsymbol{\alpha}}\big)-[\mathbf{Z}_i(h;\boldsymbol{\alpha})-\mathbf{z}_k(h;\boldsymbol{\alpha})]^{\mathbf{j}}K_h\big(\mathbf{X}_i-\mathbf{x}_k;\hat{\boldsymbol{\alpha}}\big)\Big|\\
\leq&\frac{1}{nh_n^{|\boldsymbol{\alpha}|}}\sum_{i=1}^{n}K_h\big(\mathbf{X}_i-\mathbf{x}_k;\hat{\boldsymbol{\alpha}}\big)\Big|[\mathbf{Z}_i(h;\boldsymbol{\alpha})-\mathbf{z}(h;\boldsymbol{\alpha})]^{\mathbf{j}}-[\mathbf{Z}_i(h;\boldsymbol{\alpha})-\mathbf{z}_k(h;\boldsymbol{\alpha})]^{\mathbf{j}}\Big|\\
&\qquad+\Big|[\mathbf{Z}_i(h;\boldsymbol{\alpha})-\mathbf{z}_k(h;\boldsymbol{\alpha})]^{\mathbf{j}}\Big|\cdot C\Big\|[\mathbf{Z}_i(h;\hat{\boldsymbol{\alpha}})-\mathbf{z}(h;\hat{\boldsymbol{\alpha}})]-[\mathbf{Z}_i(h;\hat{\boldsymbol{\alpha}})-\mathbf{z}_k(h;\hat{\boldsymbol{\alpha}})]\Big\|
\end{aligned}
\tag{S.22}
$$

By the definition of I_k, we have

$$
\sup_{\mathbf{x}\in D\cap I_k}\Big|[\mathbf{Z}_i(h;\boldsymbol{\alpha})-\mathbf{z}(h;\boldsymbol{\alpha})]^{\mathbf{j}}-[\mathbf{Z}_i(h;\boldsymbol{\alpha})-\mathbf{z}_k(h;\boldsymbol{\alpha})]^{\mathbf{j}}\Big|\leq h_n^{-|\mathbf{j}|}\ell_n,
$$

and

$$
\sup_{\mathbf{x}\in D\cap I_k}\Big\|[\mathbf{Z}_i(h;\hat{\boldsymbol{\alpha}})-\mathbf{z}(h;\hat{\boldsymbol{\alpha}})]-[\mathbf{Z}_i(h;\hat{\boldsymbol{\alpha}})-\mathbf{z}_k(h;\hat{\boldsymbol{\alpha}})]\Big\|\leq h_n^{-1}\sqrt{p_n}\ell_n.
$$

Since kernel function $K_h(\cdot)$ is bounded and $\mathbf{X}$ has a compact support, we can substitute two previous inequalities to (S.22) to get

$$
\begin{aligned}
I&\leq\frac{1}{nh_n^{|\boldsymbol{\alpha}|}}\sum_{i=1}^{n}\big\{C_3h_n^{-|\mathbf{j}|}\ell_n+h_n^{-|\mathbf{j}|}C_4^{|\mathbf{j}|}\cdot Ch_n^{-1}\sqrt{p_n}\ell_n\big\}\\
&=O\Big(\frac{p_n^{1/2}\ell_n}{h_n^{|\boldsymbol{\alpha}|+|\mathbf{j}|+1}}\Big)=O\Big(\big[\frac{p_n\log(n)}{nh_n^{|\boldsymbol{\alpha}|}}\big]^{1/2}\Big)\quad a.s.
\end{aligned}
\tag{S.23}
$$

From (S.23) we can immediately get

$$
III=O\Big(\big[\frac{p_n\log(n)}{nh_n^{|\boldsymbol{\alpha}|}}\big]^{1/2}\Big)\quad a.s.
\tag{S.24}
$$

The remaining task is to show $II=O\Big(\big[\frac{p_n\log(n)}{nh_n^{|\boldsymbol{\alpha}|}}\big]^{1/2}\Big)$ almost surely. Write

$$
s_{\mathbf{j}}(\mathbf{x})-\mathbf{E}[s_{\mathbf{j}}(\mathbf{x})]:=\sum_{i=1}^{n}V_{\mathbf{j},i}(\mathbf{x}),
$$

where

$$
V_{\mathbf{j},i}(\mathbf{x})=\frac{1}{nh_n^{|\boldsymbol{\alpha}|}}\big\{[\mathbf{Z}_i(h;\boldsymbol{\alpha})-\mathbf{z}(h;\boldsymbol{\alpha})]^{\mathbf{j}}K_h\big(\mathbf{X}_i-\mathbf{x};\hat{\boldsymbol{\alpha}}\big)-\mathbf{E}\big[[\mathbf{Z}_i(h;\boldsymbol{\alpha})-\mathbf{z}(h;\boldsymbol{\alpha})]^{\mathbf{j}}K_h\big(\mathbf{X}_i-\mathbf{x};\hat{\boldsymbol{\alpha}}\big)\big]\big\}.
\tag{S.25}
$$

Then for each $\eta > 0$,

$$P\big(II > \eta\big) \leq L(n) \max_{1\leq k\leq L(n)} P\big(\big|s_{\mathbf{j}}(\mathbf{x}_k) - \mathbf{E}[s_{\mathbf{j}}(\mathbf{x}_k)]\big| > \eta\big).$$

By assumption (A1) and (A3), let

$$[\mathbf{Z}_i(h;\boldsymbol{\alpha}) - \mathbf{z}(h;\boldsymbol{\alpha})]^{\mathbf{j}} K_h\big(\mathbf{X}_i - \mathbf{x};\hat{\boldsymbol{\alpha}}\big) \leq A_1 \quad a.s.$$

for $\mathbf{j}$ with $0 \leq |\mathbf{j}| \leq 4$. We have by (S.25),

$$\big|V_{\mathbf{j},i}(\mathbf{x})\big| \leq \frac{2A_1}{nh_n^{|\boldsymbol{\alpha}|}} \quad a.s. \quad i = 1,\ldots,n.$$

Define

$$\lambda_n = \frac{1}{4A_1}[np_n h_n^{|\boldsymbol{\alpha}|}\log(n)]^{1/2},$$

then by the restriction of h_n, for large enough n,

$$\lambda_n\big|V_{\mathbf{j},i}(\mathbf{x})\big| \leq \frac{1}{2}, \quad i = 1,\ldots,n.$$

Thus, $\exp\{\pm\lambda_n V_{\mathbf{j},i}(\mathbf{x})\} \leq 1 \pm \lambda_n V_{\mathbf{j},i}(\mathbf{x}) + \lambda_n^2 V_{\mathbf{j},i}^2(\mathbf{x})$ because $e^t \leq 1 + t + t^2$ for $|t| \leq 1/2$. Based on this inequality, we have

$$\mathbf{E}\Big[e^{\pm\lambda_n V_{\mathbf{j},i}(\mathbf{x})}\Big] \leq 1 + \lambda_n^2\mathbf{E}[V_{\mathbf{j},i}^2(\mathbf{x})] \leq e^{\lambda_n^2\mathbf{E}[V_{\mathbf{j},i}^2(\mathbf{x})]}. \tag{S.26}$$

By (S.26), Markov's inequality and the independence of $\{V_{\mathbf{j},i}\}_{i=1}^n$,

$$\begin{aligned} P\big(\big|s_{\mathbf{j}}(\mathbf{x}_k) - \mathbf{E}[s_{\mathbf{j}}(\mathbf{x}_k)]\big| > \eta\big) &\leq \frac{\mathbf{E}\big[e^{\lambda_n\sum_{i=1}^n V_{\mathbf{j},i}(\mathbf{x}_k)}\big] + \mathbf{E}\big[e^{-\lambda_n\sum_{i=1}^n V_{\mathbf{j},i}(\mathbf{x}_k)}\big]}{e^{\lambda_n\eta}} \\ &\leq 2e^{-\lambda_n\eta}\Big\{e^{\lambda_n^2\sum_{i=1}^n\mathbf{E}[V_{\mathbf{j},i}^2(\mathbf{x}_k)]}\Big\} \\ &= 2e^{-\lambda_n\eta}\Big\{e^{\lambda_n^2 Var[s_{\mathbf{j}}(\mathbf{x}_k)]}\Big\}. \end{aligned} \tag{S.27}$$

Denote the upper bound on $nh_n^{|\boldsymbol{\alpha}|}Var[s_{\mathbf{j}}(\mathbf{x})]$ by constant A_2, then by (S.21) and (S.27),

$$\max_{1\leq k\leq L(n)} P\big(\big|s_{\mathbf{j}}(\mathbf{x}_k) - \mathbf{E}[s_{\mathbf{j}}(\mathbf{x}_k)]\big| > \eta\big) \leq 2\exp\Big\{-\lambda_n\eta + \frac{\lambda_n^2 A_2}{nh_n^{|\boldsymbol{\alpha}|}}\Big\}.$$

Let $\eta = \eta_n = A_3\big[p_n\log(n)/(nh_n^{|\boldsymbol{\alpha}|})\big]^{1/2}$, we have

$$P\big(II > \eta_n\big) \leq L(n)\exp\Big\{\Big(-\frac{A_3}{4A_1} + \frac{A_2}{16A_1^2}\Big)p_n\log(n)\Big\} = L(n)n^{-ap_n}, \tag{S.28}$$

where $a = \frac{A_3}{4A_1} - \frac{A_2}{16A_1^2}$. By selecting a large enough A_3, we can ensure $L(n)n^{-ap_n}$ is summable. Then, it follows by (S.28) and the Borel-Cantelli lemma that

$$II = O(\eta_n) = O\left(\left[p_n \frac{\log(n)}{nh_n^{|\boldsymbol{\alpha}|}}\right]^{1/2}\right). \tag{S.29}$$

Consequently, Lemma 2 follows from (S.23), (S.24) and (S.28) . □

Then, the strong consistency of matrices $\mathbf{S}$ can be obtained.

Lemma 3. *Under the same conditions as in Lemma 2, we have, uniformly in* $\mathbf{x} \in D$,

$$\mathbf{S}(\mathbf{x}) \to \mathbf{E}(\mathbf{S}(\mathbf{x})), \quad a.s. \quad as \quad n \to \infty.$$

We next consider the uniform strong consistency of the error term $e_{\mathbf{j}}(\mathbf{x})$ given in (S.17).

Lemma 4. *Let D be any compact subset of $\mathbb{R}^p$. Let condition (A1) - (A4) hold, we have for each* $\mathbf{j}$ *with* $0 \le |\mathbf{j}| \le 1$,

$$\sup_{\mathbf{x}\in D} \left|e_{\mathbf{j}}(\mathbf{x}) - \mathbf{E}[e_{\mathbf{j}}(\mathbf{x})]\right| = O\left(\omega_n^2 \left[\frac{p_n \log(n)}{nh_n^{|\boldsymbol{\alpha}|}}\right]^{1/2}\right) \quad a.s.$$

$$\sup_{\mathbf{x}\in D} \left|e_{\mathbf{j}}(\mathbf{x})\right| = O(\omega_n^2) \quad a.s. \tag{S.30}$$

Proof of Lemma 4. For notation simplicity, we can write

$$e_{\mathbf{j}}(\mathbf{x}) = \sum_{|\mathbf{k}|=2} G_{n,\mathbf{k}+\mathbf{j}}(\mathbf{x}), \quad 0 \le |\mathbf{j}| \le 1,$$

where

$$G_{n,\mathbf{k}+\mathbf{j}}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{h^{\mathbf{k}\cdot\boldsymbol{\alpha}}}{\mathbf{k}!}(D^{\mathbf{k}}m)(\tilde{\mathbf{x}}_i)[\mathbf{Z}_i(h;\boldsymbol{\alpha}) - \mathbf{z}(h;\boldsymbol{\alpha})]^{\mathbf{k}+\mathbf{j}} K_h(\mathbf{X}_i - \mathbf{x};\hat{\boldsymbol{\alpha}}).$$

For $|\mathbf{k}| = 2$, by the definition of $(D^{\mathbf{k}}m)$, there are $j, k \in \{1, \ldots, p\}$ such that

$$h_n^{\mathbf{k}\cdot\boldsymbol{\alpha}}(D^{\mathbf{k}}m)(\mathbf{x}) = h_n^{\alpha_j+\alpha_k}\frac{\partial^2 m(\mathbf{x})}{\partial x^{[j]}\partial x^{[k]}}.$$

If $\alpha_j = 0$ or $\alpha_k = 0$, then $h_n^{\mathbf{k}\alpha}(D^{\mathbf{k}}m)(\mathbf{x}) = 0$; otherwise, by Lemma 1,

$$\sup_{\mathbf{x}\in\mathbb{R}^p} \left|h_n^{\mathbf{k}\cdot\boldsymbol{\alpha}}(D^{\mathbf{k}}m)(\mathbf{x})\right| \le C h_n^{\alpha_j+\alpha_k} \cdot \|\beta^{[j]}\| \cdot \|\beta^{[k]}\|,$$

for some constant C and specific pair $(j,k) \in \{1,\ldots,p\}^2$. By (S.9), since $\mathbf{k}! \ge 1$,

$$\left|\mathbf{E}[G_{n,\mathbf{k}+\mathbf{j}}(\mathbf{x})]\right| \le C h_n^{\alpha_j+\alpha_k} \cdot \|\beta^{[j]}\| \cdot \|\beta^{[k]}\| \cdot \left|\mathbf{E}[s_{\mathbf{k}+\mathbf{j}}(\mathbf{x})]\right| = C h_n^{\alpha_j+\alpha_k} \cdot \|\beta^{[j]}\| \cdot \|\beta^{[k]}\|.$$

Thus,

$$\begin{aligned}\sup_{\mathbf{x}\in\mathbb{R}^p}\left|\mathbf{E}[e_{\mathbf{j}}(\mathbf{x})]\right| &\leq \sup_{\mathbf{x}\in\mathbb{R}^p}\sum_{|\mathbf{k}|=2}\left|\mathbf{E}[G_{n,\mathbf{j}+\mathbf{k}}(\mathbf{x})]\right| \\ &\leq \sum_{\alpha_j\neq 0}\sum_{\alpha_k\neq 0} Ch_n^{\alpha_j+\alpha_k}\cdot\|\beta^{[j]}\|\cdot\|\beta^{[k]}\| \\ &= C\cdot\left(\sum_{\alpha_j\neq 0} h_n^{\alpha_j}\|\beta^{[j]}\|\right)^2 = O(\omega_n^2).\end{aligned}\tag{S.31}$$

Similarly,

$$\sup_{\mathbf{x}\in D}\left|G_{n,\mathbf{k}+\mathbf{j}}(\mathbf{x})-\mathbf{E}[G_{n,\mathbf{k}+\mathbf{j}}(\mathbf{x})]\right| \leq Ch_n^{\alpha_j+\alpha_k}\cdot\|\beta^{[j]}\|\cdot\|\beta^{[k]}\|\cdot\sup_{\mathbf{x}\in D}\left|s_{\mathbf{k}+\mathbf{j}}(\mathbf{x})-\mathbf{E}[s_{\mathbf{k}+\mathbf{j}}(\mathbf{x})]\right|,$$

for some positive constant C. Then,

$$\begin{aligned}\sup_{\mathbf{x}\in D}\left|e_{\mathbf{j}}(\mathbf{x})-\mathbf{E}[e_{\mathbf{j}}(\mathbf{x})]\right| &\leq \sum_{|\mathbf{k}|=2}\sup_{\mathbf{x}\in D}\left|G_{n,\mathbf{k}+\mathbf{j}}(\mathbf{x})-\mathbf{E}[G_{n,\mathbf{k}+\mathbf{j}}(\mathbf{x})]\right| \\ &\leq \sum_{\alpha_j\neq 0}\sum_{\alpha_k\neq 0} Ch_n^{\alpha_j+\alpha_k}\cdot\|\beta^{[j]}\|\cdot\|\beta^{[k]}\|\cdot O\left(\left[\frac{p_n\log(n)}{nh_n^{|\boldsymbol{\alpha}|}}\right]^{1/2}\right), \\ &= O\left(\omega_n^2\left[\frac{p_n\log(n)}{nh_n^{|\boldsymbol{\alpha}|}}\right]^{1/2}\right) \quad a.s.\end{aligned}\tag{S.32}$$

Obviously, (S.30) directly follows by (S.31) and (S.32). Similar results for vector $\mathbf{e}(\mathbf{x})$ can also be obtained. $\square$

Using similar methods as in Lemma 2, we can get asymptotic result for $t^*_{\mathbf{j}}(\mathbf{x})$.

Lemma 5. *Let D be any compact subset of $\mathbb{R}^d$ and conditions (A1) - (A4) hold. Let Y be bounded almost surely. For each $\mathbf{j}$ with $0\leq|\mathbf{j}|\leq 1$,*

$$\sup_{\mathbf{x}\in D}\left|t^*_{\mathbf{j}}(\mathbf{x})\right| = O\left(\left[\frac{p_n\log(n)}{nh_n^{|\boldsymbol{\alpha}|}}\right]^{1/2}\right) \quad a.s.$$

Proof of Lemma 5. Using the same definition of I_k's, we have

$$\begin{aligned}\sup_{\mathbf{x}\in D}\left|t^*_{\mathbf{j}}(\mathbf{x})\right| &= \max_{1\leq k\leq L(n)}\sup_{\mathbf{x}\in D\cap I_k}\left|t^*_{\mathbf{j}}(\mathbf{x})\right| \\ &\leq \max_{1\leq k\leq L(n)}\sup_{\mathbf{x}\in D\cap I_k}\left|t^*_{\mathbf{j}}(\mathbf{x})-t^*_{\mathbf{j}}(\mathbf{x}_k)\right| + \max_{1\leq k\leq L(n)}\left|t^*_{\mathbf{j}}(\mathbf{x}_k)\right| \\ &= I + II.\end{aligned}$$

Now by equality (S.14), since Y is almost surely bounded and $\mathbf{X}$ has a compact support, we have almost surely

$$\begin{aligned}
&\big|t^*_{\mathbf{j}}(\mathbf{x}) - t^*_{\mathbf{j}}(\mathbf{x}_k)\big| \\
\le\ & \frac{1}{n}\sum_{i=1}^{n}\bigg|[Y_i - m(\mathbf{X}_i)]\cdot\frac{1}{h_n^{|\boldsymbol{\alpha}|}}\cdot \\
&\quad \big\{[\mathbf{Z}_i(h;\boldsymbol{\alpha}) - \mathbf{z}(h;\boldsymbol{\alpha})]^{\mathbf{j}} K_h\big(\mathbf{X}_i - \mathbf{x};\hat{\boldsymbol{\alpha}}\big) - [\mathbf{Z}_i(h;\boldsymbol{\alpha}) - \mathbf{z}_k(h;\boldsymbol{\alpha})]^{\mathbf{j}} K_h\big(\mathbf{X}_i - \mathbf{x}_k;\hat{\boldsymbol{\alpha}}\big)\big\}\bigg| \\
\le\ & \frac{C}{nh_n^{\gamma}}\sum_{i=1}^{n}\bigg|[\mathbf{Z}_i(h;\boldsymbol{\alpha}) - \mathbf{z}(h;\boldsymbol{\alpha})]^{\mathbf{j}} K_h\big(\mathbf{X}_i - \mathbf{x};\hat{\boldsymbol{\alpha}}\big) - [\mathbf{Z}_i(h;\boldsymbol{\alpha}) - \mathbf{z}_k(h;\boldsymbol{\alpha})]^{\mathbf{j}} K_h\big(\mathbf{X}_i - \mathbf{x}_k;\hat{\boldsymbol{\alpha}}\big)\bigg|,
\end{aligned}$$

for some constant $C > 0$. Follow the same steps used in (S.22) and (S.23), we have

$$I \le \frac{C}{h_n^{\gamma+1}}\cdot\frac{p_n^{1/2}}{L^{1/p}(n)} = O\bigg(\big[\frac{p_n\log(n)}{nh_n^{|\boldsymbol{\alpha}|}}\big]^{1/2}\bigg) \quad a.s.$$

For II, we can follow the same steps used for (S.29) in the proof of Lemma 2 to get

$$II = O\bigg(\big[\frac{p_n\log(n)}{nh_n^{|\boldsymbol{\alpha}|}}\big]^{1/2}\bigg) \quad a.s.$$

Thus, the statement follows. □

Then, we can prove Theorem 1 by combining the results of previous lemmas.

Proof of Theorem 1. By condition (A5) and Lemma 3, suppose $\lambda_{min}(\mathbf{S}(\mathbf{x})) > c > 0$ for large enough n, where λ_{min} represents the smallest eigenvalue. Thus, $\lambda_{max}(\mathbf{S}^{-1}(\mathbf{x})) < c^{-1} < \infty$, where λ_{max} is the largest eigenvalue. Then, we have by Lemma 5 that

$$\sup_{\mathbf{x}\in D}\big|\mathbf{S}^{-1}(\mathbf{x})\tau^*(\mathbf{x})\big|_{max} \le \sup_{\mathbf{x}\in D} c^{-1}\big|\tau^*(\mathbf{x})\big|_{max} = O\bigg(\big[\frac{p_n\log(n)}{nh_n^{|\boldsymbol{\alpha}|}}\big]^{1/2}\bigg) \quad a.s.$$

Here, $|\mathbf{u}|_{max}$ denote the largest absolute element in a vector $\mathbf{u}$. Similarly, by Lemma 4 and condition (A5),

$$\sup_{\mathbf{x}\in D}\big|\mathbf{S}^{-1}(\mathbf{x})\mathbf{e}(\mathbf{x})\big| = O(\omega_n^2) \quad a.s.$$

Therefore, the result follows after dividing both sides by $h_n^{\alpha_j}$. □

Before the proof of Theorem 2, we propose the following lemma firstly.

Lemma 6. *Let M and N be two symmetric $p\times p$ matrices with eigen-decomposition*

$$M = \sum_{i=1}^{p}\lambda_i\mathbf{v}_i\mathbf{v}_i^{\top}, \quad \lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p,$$

and

$$N = \sum_{i=1}^{p} \delta_i \mathbf{u}_i \mathbf{u}_i^\top, \quad \delta_1 \geq \delta_2 \geq \cdots \geq \delta_p,$$

where λ's and δ's are eigenvalues of M and N, $\mathbf{v}$'s and $\mathbf{u}$'s are orthogonal unit eigenvectors correspondingly. Furthermore, let

$$\lambda_{n_{j-1}+1} = \cdots = \lambda_{n_j} = \tilde{\lambda}_j, \quad n_0 = 0 < n_1 < \cdots < n_s = p, \quad j = 1, \ldots, s,$$

such that

$$\tilde{\lambda}_1 > \tilde{\lambda}_2 > \cdots > \tilde{\lambda}_s \geq 0.$$

Suppose $\tilde{\lambda}_{s-1} - \tilde{\lambda}_s > c > 0$ and $M - N = \mathbf{O}(\alpha)$, where $\mathbf{O}(\alpha)$ represents any matrix that each entry is of order $O(\alpha)$ for simplicity. Then,
(i) $|\lambda_i - \delta_i| = O(p\alpha)$, for $i = 1, \ldots, p$;
(ii) $\big|\sum_{i=n_{j-1}+1}^{n_j} \mathbf{u}_i \mathbf{u}_i^\top - \sum_{i=n_{j-1}+1}^{n_j} \mathbf{v}_i \mathbf{v}_i^\top\big| = O\big(p\alpha\big)$ for $j = 1, \ldots, s$.

Proof of Lemma 6. By von Neumanna's ineqaulity and the property of trace, we have

$$|\lambda_i - \delta_i| = \sqrt{(\lambda_i - \delta_i)^2} \leq \sqrt{\sum_{i=1}^{p} (\lambda_i - \delta_i)^2} \leq \sqrt{tr[(M-N)^2]} = O(p\alpha),$$

which shows (i).

Let $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_p)$ and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_p)$. By the definitions of u_i's and v_i's, the eigenvalues of $\mathbf{U}$ and $\mathbf{V}$ are either 1 or -1. Therefore,

$$\begin{aligned} N &= \sum_{i=1}^{p} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top + \sum_{i=1}^{p} (\delta_i - \lambda_i) \mathbf{u}_i \mathbf{u}_i^\top \\ &= \sum_{j=1}^{s} \tilde{\lambda}_j \sum_{i \in L_j} \mathbf{u}_i \mathbf{u}_i^\top + \mathbf{U}\ diag(\delta_1 - \lambda_1, \delta_2 - \lambda_2, \ldots, \delta_p - \lambda_p)\ \mathbf{U}^\top \\ &= N' + \mathbf{O}\big(p\alpha\big), \end{aligned}$$

where $L_j = (n_{j-1}+1, \ldots, n_j)$. Then, it is obvious that $M - N' = \mathbf{O}\big(p\alpha\big)$.

When $s = 1$, (ii) is trivial. Assume (ii) is true for $s = t$, when $s = t+1$,

$$\sum_{j=1}^{t} (\tilde{\lambda}_j - \tilde{\lambda}_{t+1}) \sum_{i \in L_j} \mathbf{u}_i \mathbf{u}_i^\top = \sum_{j=1}^{t} (\tilde{\lambda}_j - \tilde{\lambda}_{t+1}) \sum_{i \in L_j} \mathbf{v}_i \mathbf{v}_i^\top + \mathbf{O}(p\alpha), \tag{S.33}$$

where the LHS equals to $[N' - \mathbf{I}_p]$ and the RHS is $[M - \mathbf{I}_p + \mathbf{O}(p\alpha)]$. Multiply from right by $v_k, k \in L_{t+1}$ on both sides of (S.33), we have

$$\sum_{j=1}^{t} (\tilde{\lambda}_j - \tilde{\lambda}_{t+1}) \sum_{i \in L_j} \mathbf{u}_i (\mathbf{u}_i^\top \mathbf{v}_k) = \mathbf{O}(p\alpha),$$

which implies that $\mathbf{u}_i^\top \mathbf{v}_k = O(p\alpha)$ for $1 \le i \le n_t$ and $k \in L_{t+1}$. Thus, we have

$$\mathbf{U}_1^\top \mathbf{V}_2 = \mathbf{O}(p\alpha), \quad \mathbf{V}_1^\top \mathbf{U}_2 = \mathbf{O}(p\alpha), \tag{S.34}$$

where

$$\mathbf{U}_1 = (\mathbf{u}_1, \ldots, \mathbf{u}_{n_t}), \mathbf{U}_2 = (\mathbf{u}_{n_t+1}, \ldots, \mathbf{u}_{n_{t+1}}), \mathbf{V}_1 = (\mathbf{v}_1, \ldots, \mathbf{v}_{n_t}), \mathbf{V}_2 = (\mathbf{v}_{n_t+1}, \ldots, \mathbf{v}_{n_{t+1}}).$$

By the property of singular value, the largest singular value of $\mathbf{V}_1, \mathbf{V}_2$ and $\mathbf{U}_1, \mathbf{U}_2$ are not larger than 1. By (S.34) and note that $n_{t+1} = p$,

$$\mathbf{V}_2^\top \mathbf{U}_2 \mathbf{U}_2^\top \mathbf{V}_2 = \mathbf{V}_2^\top (\mathbf{I}_p - \mathbf{U}_1 \mathbf{U}_1^\top) \mathbf{V}_2 = \mathbf{V}_2^\top \mathbf{V}_2 + \mathbf{O}(p\alpha) = \mathbf{I}_{p-n_t} + \mathbf{O}(p\alpha). \tag{S.35}$$

Let $\mathbf{U}_2 = \mathbf{V}_1 \mathbf{G}_1 + \mathbf{V}_2 \mathbf{G}_2$, where $\mathbf{G}_1 : n_t \times (p - n_t)$ and $\mathbf{G}_2 : (p - n_t) \times (p - n_t)$. By (S.33) and (S.34),

$$\mathbf{U}_2 = \mathbf{V}_2 \mathbf{G}_2 + \mathbf{V}_1 (\mathbf{G}_1 + \mathbf{V}_1^\top \mathbf{V}_2 \mathbf{G}_2) = \mathbf{V}_2 \mathbf{G}_2 + \mathbf{V}_1 (\mathbf{V}_1^\top \mathbf{U}_2) = \mathbf{V}_2 \mathbf{G}_2 + \mathbf{O}(p\alpha). \tag{S.36}$$

Then, by (S.35) and (S.36)

$$\mathbf{G}_2 \mathbf{G}_2^\top = \mathbf{V}_2^\top \mathbf{V}_2 \mathbf{G}_2 \mathbf{G}_2^\top \mathbf{V}_2^\top \mathbf{V}_2^\top = \mathbf{V}_2^\top \mathbf{U}_2 \mathbf{U}_2^\top \mathbf{V}_2^\top + \mathbf{O}(p\alpha) = \mathbf{I}_{p-n_t} + \mathbf{O}(p\alpha). \tag{S.37}$$

From (S.36) and (S.37), it follows that

$$\sum_{j \in L_{t+1}} \mathbf{u}_i \mathbf{u}_i^\top = \mathbf{U}_2 \mathbf{U}_2^\top = \mathbf{V}_2 \mathbf{V}_2^\top + \mathbf{O}(n_t \alpha) = \sum_{j \in L_{t+1}} \mathbf{v}_i \mathbf{v}_i^\top + \mathbf{O}(p\alpha), \tag{S.38}$$

and that

$$\sum_{j=1}^{t-1} \tilde{\lambda}_j \sum_{i \in L_j} \mathbf{u}_i \mathbf{u}_i^\top + \tilde{\lambda}_t \sum_{i \in L_t \cup L_{t+1}} \mathbf{u}_i \mathbf{u}_i^\top = \sum_{j=1}^{t-1} \tilde{\lambda}_j \sum_{i \in L_j} \mathbf{v}_i \mathbf{v}_i^\top + \tilde{\lambda}_t \sum_{i \in L_t \cup L_{t+1}} \mathbf{v}_i \mathbf{v}_i^\top + \mathbf{O}(p\alpha).$$

By induction,

$$\sum_{i \in L_j} \mathbf{u}_i \mathbf{u}_i^\top = \sum_{i \in L_j} \mathbf{v}_i \mathbf{v}_i^\top + \mathbf{O}(p\alpha), \quad j = 1, \ldots, t-1, \tag{S.39}$$

and

$$\sum_{i \in L_t \cup L_{t+1}} \mathbf{u}_i \mathbf{u}_i^\top = \sum_{i \in L_t \cup L_{t+1}} \mathbf{v}_i \mathbf{v}_i^\top + \mathbf{O}(p\alpha). \tag{S.40}$$

Therefore, (ii) is true for $s = t + 1$ by (S.38) - (S.40). $\square$

PROOF OF THEOREM 2. By condition (A1), we can denote the support of $\mathbf{X}$ by D, which is a compact set in $\mathbb{R}^p$. Then, for every $\mathbf{x} \in D$,

$$\hat{\mathbf{b}}(\mathbf{x}) \overset{a.s.}{=} \mathbf{b}(\mathbf{x}) + \Delta\mathbf{b}_n(\mathbf{x}) = B_0 \nabla g(B_0^\top \mathbf{x}) + \Delta\mathbf{b}_n(\mathbf{x}), \tag{S.41}$$

where $\Delta\mathbf{b}_n(\mathbf{x}) = \big(\mathbf{b}_n^{[1]}(\mathbf{x}), \ldots, \mathbf{b}_n^{[p]}(\mathbf{x})\big)^\top$ is a p-dimensional vector. By high-dimensional linear regression, it is easy to know that $\mathbf{b}_n^{[j]}(\mathbf{x}) = O_p\big(\sqrt{p_n/n}\big)$ when $\alpha_j = 0$. Otherwise, by Theorem 1, we have $\mathbf{b}_n^{[j]}(\mathbf{x}) = O(c_n^{[j]})$ almost surely with $c_n^{[j]} = \big(\frac{p_n \log(n)}{n h_n^{|\boldsymbol{\alpha}|+2\alpha_j}}\big)^{1/2} + \omega_n^2 / h_n^{\alpha_j}$. Since $\sqrt{p_n/n} = o(c_n^{[j]})$ for all j's, it is obvious that $\mathbf{b}_n^{[j]}(\mathbf{x}) = O_p(c_n^{[j]})$.

Let $(B_0, \tilde{B}_0)$ be a $p \times p$ orthogonal matrix, we can write

$$\hat{\mathbf{b}}_j := (B_0, \tilde{B}_0) \begin{pmatrix} \nabla g(B_0^\top \mathbf{X}_j) + B_0^\top \Delta\mathbf{b}_n(\mathbf{X}_j) \\ \tilde{B}_0^\top \Delta\mathbf{b}_n(\mathbf{X}_j) \end{pmatrix},$$

and

$$\hat{\Sigma} := \frac{1}{n} \sum_{j=1}^n \hat{\mathbf{b}}_j \hat{\mathbf{b}}_j^\top = (B_0, \tilde{B}_0) G_n(p_n, h_n) (B_0, \tilde{B}_0)^\top. \tag{S.42}$$

In the previous equality, $G_n(p_n, h_n)$ is a $p \times p$ matrix defined as

$$\begin{aligned} G(p_n, h_n) &:= \frac{1}{n} \sum_{j=1}^n \begin{pmatrix} \nabla g(B_0^\top \mathbf{X}_j) + B_0^\top \Delta\mathbf{b}_n(\mathbf{X}_j) \\ \tilde{B}_0^\top \Delta\mathbf{b}_n(\mathbf{X}_j) \end{pmatrix} \begin{pmatrix} \nabla g(B_0^\top \mathbf{X}_j) + B_0^\top \Delta\mathbf{b}_n(\mathbf{X}_j) \\ \tilde{B}_0^\top \Delta\mathbf{b}_n(\mathbf{X}_j) \end{pmatrix}^\top \\ &= \begin{pmatrix} \Lambda_n^{(1)} & \Lambda_n^{(2)} \\ \Lambda_n^{(3)} & \Lambda_n^{(4)} \end{pmatrix} \end{aligned}$$

where

$$\begin{aligned} \Lambda_n^{(1)} &= \frac{1}{n} \sum_{j=1}^n \big\{ \nabla g(B_0^\top \mathbf{X}_j) \nabla^\top g(B_0^\top \mathbf{X}_j) \\ &\qquad + 2 B_0^\top \Delta\mathbf{b}_n(\mathbf{X}_j) \nabla^\top g(B_0^\top \mathbf{X}_j) + B_0^\top \Delta\mathbf{b}_n(\mathbf{X}_j) [\Delta\mathbf{b}_n(\mathbf{X}_j)]^\top B_0 \big\} \\ \Lambda_n^{(3)} &= \big(\Lambda_n^{(2)}\big)^\top = \frac{1}{n} \sum_{j=1}^n \tilde{B}_0^\top \Delta\mathbf{b}_n(\mathbf{X}_j) \nabla^\top g(B_0^\top \mathbf{X}_j) \\ \Lambda_n^{(4)} &= \frac{1}{n} \sum_{j=1}^n \tilde{B}_0^\top \Delta\mathbf{b}_n(\mathbf{X}_j) [\Delta\mathbf{b}_n(\mathbf{X}_j)]^\top \tilde{B}_0. \end{aligned}$$

By condition (A6), $\nabla g(B_0^\top \mathbf{x})$ is bounded for all possible $\mathbf{x}$. For a p-dimensional unit vector β,

$$\beta^\top \Delta\mathbf{b}_n(\mathbf{x}) \le \|\beta\| \cdot \|\Delta\mathbf{b}_n(\mathbf{x})\| \le \Big[\sum_{\alpha_j \neq 0} (c_n^{[j]})^2 + \sum_{\alpha_j = 0} p_n/n \Big]^{1/2} := \sigma_n \quad \text{in probability.}$$

Therefore,

$$B_0^\top \Delta\mathbf{b}_n(\mathbf{x}) \nabla^\top g(B_0^\top \mathbf{x}) = \mathcal{E}\big(\sigma_n\big),$$

where $\mathcal{E}(\sigma_n)$ represents matrix that each entry is of order $O_p(\sigma_n)$. And similarly,

$$B_0^\top \Delta \mathbf{b}_n(\mathbf{X}_j)[\Delta \mathbf{b}_n(\mathbf{X}_j)]^\top B_0 = \mathcal{E}\big(\sigma_n^2\big).$$

By the central limit theorem, one is easy to derive that

$$\frac{1}{n}\sum_{j=1}^n \nabla g(B_0^\top \mathbf{X}_j)\nabla^\top g(B_0^\top \mathbf{X}_j) = \int_{\mathbb{R}^p} \nabla g(B_0^\top \mathbf{x})\nabla^\top g(B_0^\top \mathbf{x}) f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} + \mathcal{E}\big(1/\sqrt{n}\big).$$

Consequently, we have

$$\Lambda_n^{(1)} = \int_{\mathbb{R}^p} \nabla g(B_0^\top \mathbf{x})\nabla^\top g(B_0^\top \mathbf{x}) f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} + \mathcal{E}\big(\sigma_n\big) := \Lambda_\infty^{(1)} + \mathcal{E}\big(\sigma_n + 1/\sqrt{n}\big).$$

By (i) of Lemma 6, it can be known that the eigenvalues of $\Lambda_n^{(1)}$ is asymptotically converge to the eigenvalues of $\Lambda_\infty^{(1)}$ in probability with order $O\big(d\cdot(\sigma_n + 1/\sqrt{n})\big)$.

Next,

$$\Lambda_n^{(4)} = \tilde{B}_0^\top \int_{\mathbb{R}^p} \Delta \mathbf{b}(\mathbf{x})\Delta \mathbf{b}^\top(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})d\mathbf{x}\tilde{B}_0 + \mathcal{E}\big(\sqrt{1/n}\big) := \Lambda_\infty^{(4)} + \mathcal{E}\big(\sqrt{1/n}\big).$$

Since the eigenvalue of $\Delta\mathbf{b}(\mathbf{x})\Delta\mathbf{b}^\top(\mathbf{x})$ is either 0 or $\|\Delta\mathbf{b}(\mathbf{x})\|^2 = O_p\big(\sigma_n^2\big)$, the eigenvalues of matrix $\Lambda_\infty^{(4)}$ have order $O\big(\sigma_n^2\big)$. By (i) of Lemma 6 again, the eigenvalues of matrix $\Lambda_\infty^{(4)}$ have order $O\big(\sigma_n^2 + (p-d)/\sqrt{n}\big)$.

Let $\lambda_1 \geq \cdots \geq \lambda_p$ be the eigenvalues of $\hat{\Sigma}$ and $\hat{\beta}_1, \ldots, \hat{\beta}_p$ be their corresponding unit orthogonal eigenvectors. By the Eigenvalue Interlacing Theorem and property of p_n, d_n and h_n in the assumptions, we have $\min\{\lambda_1, \ldots, \lambda_d\} > c > 0$ and $\max\{\lambda_{d+1}, \ldots, \lambda_p\} = O\big(\sigma_n^2 + (p-d)/\sqrt{n}\big) = o(1)$. Therefore, the "top-d" eigenvalues can be distinguish from others asymptotically.

Similar to $\Lambda_n^{(1)}$, it can be shown that $\Lambda_n^{(2)} = \big(\Lambda_n^{(3)}\big)^\top = \mathcal{E}\big(\sigma_n + 1/\sqrt{n}\big)$. It is noteworthy to mention that, by the definition of $(B_0, \tilde{B}_0)$, the norm of each column or row vector has order 1. Then, by (S.42), we have in probability

$$\hat{\Sigma} = B_0\Lambda_n^{(1)}B_0^\top + \mathcal{E}\big(\sigma_n + 1/\sqrt{n}\big).$$

Let $\hat{B} = \big(\hat{\beta}_1, \ldots, \hat{\beta}_d\big)$, using (ii) of Lemma 6, we can get

$$\hat{B}\hat{B}^\top - B_0^\top B_0^\top = \mathcal{E}(d_n p_n \sigma_n + d_n p_n/\sqrt{n}) \quad \text{in probability.} \tag{S.43}$$

Therefore, by assumption (A7),

$$\big|\hat{B}\hat{B}^\top - B_0^\top B_0^\top\big| \xrightarrow{P} 0 \quad \text{as } n \to \infty.$$

This completes the proof. $\square$

Remark 2. *Note that we allow $d = d_n \to \infty$ providing that (S.43) converges. Considering the requirement on $d = d_n$ given in Remark 1, the estimator of CMS is also consistent when $d_n = o(\log(n))$.*

S.3. Proofs of Theorem 3 and Theorem 4

We start with the following Lemma.

Lemma 7. *Suppose conditions in Theorem 3 hold, we have*

$$CV(d,k) = \zeta_0(d) + \left\{\zeta_1(d)k^{-1} + \zeta_2(d)(k/n)^{\frac{4}{d}}\right\}\{1 + o_p(1)\},$$

where $\zeta_0(d) = \mathbf{E}\big|Y - \mathbf{1}(p_d(\mathbf{U}) > 1/2)\big|$. $\zeta_1(d)$ and $\zeta_2(d)$ are two non-negative values given in Appendix.

For the constants in the lemma, let

$$a(\mathbf{u}) = \frac{\sum_{s=1}^{d} c_{s,d}\{p_d^{(s)}(\mathbf{u})f_d^{(s)}(\mathbf{u}) + (1/2)p_d^{(ss)}(\mathbf{u})f_d(\mathbf{u})\}}{a_d^{1+2/d} f_d(\mathbf{u})^{1+2/d}},$$

where $c_{s,d} = \int_{v:\|v\|\leqslant 1} v_s^2 dv$ with v_s being the s-th element of vector v. Then,

$$\zeta_1(d) = \int_\Omega \frac{f_d(\mathbf{u}_0)}{4\|\dot{p}_d(\mathbf{u}_0)\|} dVol^{d-1}(\mathbf{u}_0) \quad \text{and}$$
$$\zeta_2(d) = \int_\Omega \frac{f_d(\mathbf{u}_0)}{\|\dot{p}_d(\mathbf{u}_0)\|} a(\mathbf{u}_0)^2 dVol^{d-1}(\mathbf{u}_0),$$

where Vol^{d-1} denotes the natural $(d-1)$-dimensional volume where Ω inherits as a subset of $\mathbb{R}^d$. It is obvious that $\zeta_1 > 0$ and $\zeta_2 \geqslant 0$, while the equality holds if and only if $a(\mathbf{u}) = 0$ for all $u \in \Omega$.

Proof of Lemma 7. For notation simplicity, let $\mathcal{T}_n = \{(\mathbf{U}_i, Y_i), i = 1, \ldots, n\} = \mathcal{S}_n^{PD}$, where $\mathbf{U}_i = (PD)^\top \mathbf{X}_i$. Since $Y \in \{0, 1\}$, it is easy to obtain that

$$\begin{aligned} CV(d,k) &= \frac{1}{n}\sum_{i=1}^{n} \big|Y_i - \mathbf{1}(\hat{p}_{d,\setminus i}(\mathbf{U}_i) > \frac{1}{2})\big| \\ &= \frac{1}{n}\sum_{i=1}^{n} \left\{\big|Y_i - \mathbf{1}(\hat{p}_{d,\setminus i}(\mathbf{U}_i) > \frac{1}{2})\big| - \big|Y_i - \mathbf{1}(p_d(\mathbf{U}_i) > \frac{1}{2})\big|\right\} + \frac{1}{n}\sum_{i=1}^{n} \big|Y_i - \mathbf{1}(p_d(\mathbf{U}_i) > \frac{1}{2})\big| \end{aligned}$$

where $\hat{p}_{d,\setminus i}(\mathbf{U}_i)$ is the kNN estimation of $\mathbf{P}(Y_i = 1|\mathbf{U}_i)$ based on delete-one-observation $\mathcal{T}_n \setminus (\mathbf{U}_i, Y_i)$.

Let $\mathcal{R}_{d,n-1}^{kNN} = \mathbf{E}\big|Y_i - \mathbf{1}(\hat{p}_{d,\setminus i}(\mathbf{U}_i) > \frac{1}{2})\big|$, where the expectation is computed with respect to $(\mathbf{U}_i, Y_i)_{i=1}^n$. We will first show that

$$\frac{1}{n}\sum_{i} \left\{\big|Y_i - \mathbf{1}(\hat{p}_{d,\setminus i}(\mathbf{U}_i) > \frac{1}{2})\big| - \big|Y_i - \mathbf{1}(p_d(\mathbf{U}_i) > \frac{1}{2})\big|\right\} = \mathcal{R}_{d,n-1}^{kNN} - \zeta_0(d) + o_p\big(k^{-1} + (k/n)^{\frac{4}{d}}\big). \quad \text{(S.44)}$$

For notation simplicity, denote $\left|Y_i - \mathbf{1}(\hat{p}_{d,\setminus i}(\mathbf{U}_i) > \frac{1}{2})\right|$ and $\left|Y_i - \mathbf{1}(p_d(\mathbf{U}_i) > \frac{1}{2})\right|$ by $\hat{\xi}_i(d)$ and $\xi_i(d)$ respectively. Thus, we can compute the expectation

$$\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\xi}_i(d) - \xi_i(d)\right\}\right] = \mathcal{R}_{d,n-1}^{kNN} - \zeta_0(d),$$

and the variance

$$\begin{aligned}
&\mathbf{Var}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\xi}_i(d) - \xi_i(d)\right\}\right] = n^{-2}\mathbf{E}\left[\sum_{i=1}^{n}\left\{\hat{\xi}_i(d) - \xi_i(d)\right\}\right]^2 - \left(\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\xi}_i(d) - \xi_i(d)\right\}\right]\right)^2 \\
&= n^{-2}\mathbf{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\hat{\xi}_i(d) - \xi_i(d)\right\}\left\{\hat{\xi}_j(d) - \xi_j(d)\right\}\right] - \left[\mathcal{R}_{d,n-1}^{kNN} - \zeta_0(d)\right]^2.
\end{aligned}$$

Consider the first term on the RHS, since $(\mathbf{U}_i, Y_i)_{i=1}^n$ are independent and identically distributed,

$$\begin{aligned}
&n^{-2}\mathbf{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\hat{\xi}_i(d) - \xi_i(d)\right\}\left\{\hat{\xi}_j(d) - \xi_j(d)\right\}\right] = \mathbf{E}\left[\left\{\hat{\xi}_i(d) - \xi_i(d)\right\}\left\{\hat{\xi}_j(d) - \xi_j(d)\right\}\right] \\
&= \mathbf{E}\left[\left\{\hat{\xi}_i(d) - \xi_i(d)\right\}\left\{\hat{\xi}_j(d) - \xi_j(d)\right\} \cdot \mathbf{1}\left\{\|\mathbf{U}_i - \mathbf{U}_j\| > r(\mathbf{U}_i, k) + r(\mathbf{U}_j, k)\right\}\right] \\
&\quad + \mathbf{E}\left[\left\{\hat{\xi}_i(d) - \xi_i(d)\right\}\left\{\hat{\xi}_j(d) - \xi_j(d)\right\} \cdot \mathbf{1}\left\{\|\mathbf{U}_i - \mathbf{U}_j\| \leqslant r(\mathbf{U}_i, k) + r(\mathbf{U}_j, k)\right\}\right] \\
&= I + II,
\end{aligned} \tag{S.45}$$

where $r(u, k) := \|\mathbf{U}_{(k)} - u\|$ is the distance between u and its k-th nearest neighbor.

When the distance $\|\mathbf{U}_i - \mathbf{U}_j\| > r(\mathbf{U}_i, k) + r(\mathbf{U}_j, k)$, $\hat{p}_{d,\setminus i}(\mathbf{U}_i)$ and $\hat{p}_{d,\setminus j}(U_j)$ are independent, since there is no $\mathbf{U}_t$ such that both $\|\mathbf{U}_t - \mathbf{U}_i\| \leqslant r(\mathbf{U}_i, k)$ and $\|\mathbf{U}_t - \mathbf{U}_j\| \leqslant r(\mathbf{U}_j, k)$ are satisfied. Thus, $\left\{\hat{\xi}_i(d) - \xi_i(d)\right\}$ and $\left\{\hat{\xi}_j(d) - \xi_j(d)\right\}$ are independent under this condition. Then,

$$I \leqslant \mathbf{E}\left\{\hat{\xi}_i(d) - \xi_i(d)\right\}\mathbf{E}\left\{\hat{\xi}_j(d) - \xi_j(d)\right\} = \left[\mathcal{R}_{d,n-1}^{kNN} - \zeta_0(d)\right]^2. \tag{S.46}$$

For term II, we can assume $r(\mathbf{U}_i, k) \geqslant r(\mathbf{U}_j, k)$ without loss of generality. Since $\hat{\xi}_i(d)$ and $\xi_i(d)$ can only be 0 or 1, $\left\{\hat{\xi}_i(d) - \xi_i(d)\right\}\left\{\hat{\xi}_j(d) - \xi_j(d)\right\} \leqslant \left|\hat{\xi}_i(d) - \xi_i(d)\right|$. Then,

$$\begin{aligned}
II &\leqslant \mathbf{E}\left[\left|\hat{\xi}_i(d) - \xi_i(d)\right| \cdot \mathbf{1}\left\{\|\mathbf{U}_i - \mathbf{U}_j\| \leqslant r(\mathbf{U}_i, k) + r(\mathbf{U}_j, k)\right\}\right] \\
&\leqslant \mathbf{E}\left[\left|\left|Y_i - \mathbf{1}(\hat{p}_{d,\setminus i}(\mathbf{U}_i) > \frac{1}{2})\right| - \left|Y_i - \mathbf{1}(p_d(\mathbf{U}_i) > \frac{1}{2})\right|\right| \cdot \mathbf{1}\left\{\|\mathbf{U}_i - \mathbf{U}_j\| \leqslant 2r(\mathbf{U}_i, k)\right\}\right] \\
&= \mathbf{E}\left[\left|\mathbf{1}(\hat{p}_{d,\setminus i}(\mathbf{U}_i) > \frac{1}{2}) - \mathbf{1}(p_d(\mathbf{U}_i) > \frac{1}{2})\right| \cdot \mathbf{1}\left\{\|\mathbf{U}_i - \mathbf{U}_j\| \leqslant 2r(\mathbf{U}_i, k)\right\}\right].
\end{aligned}$$

By the boundedness of $\dot{p}_d(\mathbf{u})$ in (B2) and the boundedness of d, $p_d(\mathbf{u})$ has Lipschitz continuity condition. By (B3) one can prove that $f(x) \geq f_{min} > 0$ for some constant f_{min}. According to the proofs in Chaudhuri and Dasgupta (2014) (Lemma 2, Theorem 3 and Theorem 5), let $\partial_\Delta = \big\{\mathbf{u} \in W \big| |p_d(\mathbf{u}) - \frac{1}{2}| \leqslant \Delta\big\}$, where $\Delta = \sqrt{\frac{1}{k}\log\frac{2}{\delta} + A_1\big(\frac{k}{2n}\big)^{1/d}}$ for any $\delta > 0$. Then

$$\big|\mathbf{1}(\hat{p}_{d,\setminus i}(\mathbf{U}_i) > \frac{1}{2}) - \mathbf{1}(p_d(\mathbf{U}_i) > \frac{1}{2})\big| \leqslant \mathbf{1}(\mathbf{U}_i \in \partial_\Delta) + \mathbf{1}\big((\mathbf{U}_1, \ldots, \mathbf{U}_n) \in \Phi_1\big) + \mathbf{1}((\mathbf{U}_1, \ldots, \mathbf{U}_n) \in \Phi_2),$$

where Φ_1 and Φ_2 are small sets such that $\mathbf{E}\big[\mathbf{1}\big((\mathbf{U}_1, \ldots, \mathbf{U}_n) \in \Phi_r\big)\big] \leqslant \delta^2$ for $r = 1, 2$. Therefore, we have

$$\begin{aligned}
&\mathbf{E}\bigg[\big|\mathbf{1}(\hat{p}_{d,\setminus i}(\mathbf{U}_i) > \frac{1}{2}) - \mathbf{1}(p_d(\mathbf{U}_i) > \frac{1}{2})\big| \cdot \mathbf{1}\big\{\|\mathbf{U}_i - \mathbf{U}_j\| \leqslant 2r(\mathbf{U}_i, k)\big\}\bigg] \\
\leqslant\ &\mathbf{E}\bigg[\mathbf{1}(\mathbf{U}_i \in \partial_\Delta) \cdot \mathbf{1}\big\{\|\mathbf{U}_i - \mathbf{U}_j\| \leqslant 2r(\mathbf{U}_i, k)\big\}\bigg] \\
&+ \mathbf{E}\bigg[\mathbf{1}\big((\mathbf{U}_1, \ldots, \mathbf{U}_n) \in \Phi_1\big)\bigg] + \mathbf{E}\bigg[\mathbf{1}\big((\mathbf{U}_1, \ldots, \mathbf{U}_n) \in \Phi_2\big)\bigg] \\
=\ &\mathbf{E}\bigg[\mathbf{1}(\mathbf{U}_i \in \partial_\Delta) \cdot \mathbf{E}\big[\mathbf{1}\big\{\|\mathbf{U}_i - \mathbf{U}_j\| \leqslant 2r(\mathbf{U}_i, k)\big\}\big|\mathbf{U}_i\big]\bigg] + 2\delta^2
\end{aligned}$$

Next, we derive the property of $r(\mathbf{U}, k)$. By Lemma 6.4 in Györfi et al. (2006), we have

$$\mathbf{E}[r(\mathbf{U}, 1)^2] \leq \tilde{c} n^{-2/d}.$$

Split $\mathbf{U}_1, \ldots, \mathbf{U}_n$ into $k+1$ segments such that the first k of them have $\lfloor \frac{n}{k} \rfloor$ elements and rest in the last segment. Let $r_j(U, 1)$ be the distance from $\mathbf{U}$ to the nearest point in j-th segment, then

$$\mathbf{E}[r(\mathbf{U}, k)^2] \leq \max_{j \in \{1, \ldots, k\}} \mathbf{E}[r_j(\mathbf{U}, 1)^2] \leq \tilde{c} \lfloor \frac{n}{k} \rfloor^{-2/d}.$$

For any $\epsilon > 0$, let $M = \sqrt{\tilde{c}/2\epsilon}$, for all $n \in \mathbb{N}$

$$\mathbf{P}\bigg(\frac{r(\mathbf{U}, k)}{(k/n)^{\frac{1}{d}}} > M\bigg) \leq \frac{\mathbf{E}[r(\mathbf{U}, k)^2]/(k/n)^{\frac{2}{d}}}{M^2} \leq \epsilon.$$

Thus, $r(\mathbf{U}, k) = O_p((k/n)^{\frac{1}{d}})$ for all $\mathbf{U} = (PD_0)^\top \mathbf{X}$.

By (B3),

$$\mathbf{E}\big[\mathbf{1}\big\{\|\mathbf{U}_i - \mathbf{U}_j\| \leqslant 2r(\mathbf{U}_i, k)\big\}\big|\mathbf{U}_i\big] = F_d\big(B_{2r(\mathbf{U}_i,k)}(\mathbf{U}_i)|\mathbf{U}_i\big) \to C a_d 2^d r(\mathbf{U}_i, k)^d,$$

for some positive constant C. In the following proof, C will always denote positive constant but may be different in different places. Since d is bounded,

$$\mathbf{E}\big[\mathbf{1}\big\{\|\mathbf{U}_i - \mathbf{U}_j\| \leqslant 2r(\mathbf{U}_i, k)\big\}\big|\mathbf{U}_i\big] \leq A_2 r(\mathbf{U}_i, k)^d = O_p(k/n). \tag{S.47}$$

Then, using equation (2.1) in Samworth et al. (2012) which can be derived from (B4), we have

$$\begin{aligned}\mathbf{E}\big[\mathbf{1}(\mathbf{U}_i\in\partial_\Delta)\big]&=F_d\bigg(\bigg\{\mathbf{u}\in W\bigg|\big|p_d(\mathbf{u})-\frac{1}{2}\big|\leqslant\sqrt{\frac{1}{k}\log\frac{2}{\delta}}+A_1\Big(\frac{k}{2n}\Big)^{1/d}\bigg\}\bigg)\\&=O\bigg(\sqrt{\frac{1}{k}\log\frac{2}{\delta}}+A_1\Big(\frac{k}{2n}\Big)^{1/d}\bigg).\end{aligned}\tag{S.48}$$

Let $\delta=\frac{1}{k^2}$, by (S.47) and (S.48)

$$\begin{aligned}&\mathbf{E}\Big[\mathbf{1}(\mathbf{U}_i\in\partial_\Delta)\cdot\mathbf{E}\big[\mathbf{1}\big\{\|\mathbf{U}_i-\mathbf{U}_j\|\leqslant 2r(\mathbf{U}_i,k)\big\}\big|\mathbf{U}_i\big]\Big]\\&=\mathbf{E}\Big[\mathbf{1}(\mathbf{U}_i\in\partial_\Delta)\cdot O_p(k/n)\Big]=O\bigg(\frac{k^{1/2}\log k}{n}+\Big(\frac{k}{2n}\Big)^{\frac{d+1}{d}}\bigg).\end{aligned}$$

Thus, it follows that

$$II\leqslant O\bigg(\frac{k^{1/2}\log k}{n}+\Big(\frac{k}{2n}\Big)^{\frac{d+1}{d}}\bigg)+o\big((1/k)^2\big)\tag{S.49}$$

For a large enough n, substitute (S.46) and (S.49) into equation (S.45),

$$\begin{aligned}&n^{-2}\mathbf{E}\Big[\sum_{i=1}^n\sum_{j=1}^n\big\{\hat{\xi}_i(d)-\xi_i(d)\big\}\big\{\hat{\xi}_j(d)-\xi_j(d)\big\}\Big]\\&\leqslant\big[\mathcal{R}_d(C_{n-1}^{kNN})-\zeta_0(d)\big]^2+O\bigg(\frac{k^{1/2}\log k}{n}+\Big(\frac{k}{2n}\Big)^{\frac{d+1}{d}}\bigg)+o\big((1/k)^2\big).\end{aligned}$$

Consequently,

$$\mathbf{Var}\Big[\frac{1}{n}\sum_{i=1}^n\big\{\hat{\xi}_i(d)-\xi_i(d)\big\}\Big]=O\bigg(\frac{k^{1/2}\log k}{n}+\Big(\frac{k}{2n}\Big)^{\frac{d+1}{d}}\bigg)+o\big((1/k)^2\big).$$

It is obvious that $O\big(\frac{k^{1/2}\log k}{n}\big)=o\big((1/k)^2\big)$ when $k=o(n^{2/5})$, $O\big(\big(\frac{k}{2n}\big)^{\frac{d+1}{d}}\big)=o\big(\big(\frac{k}{n}\big)^{\frac{8}{d}}\big)$ when $d>7$ and $O\big(\big(\frac{k}{2n}\big)^{\frac{d+1}{d}}\big)=o\big((1/k)^2\big)$ when $d\leqslant 7$ and $k=o(n^{4/11})$. In conclusion, by (B5) and Chebyshev's inequality, for every $\epsilon>0$,

$$\begin{aligned}&\mathbf{P}\bigg(\frac{\big|n^{-1}\sum_{i=1}^n\{\hat{\xi}_i-\xi_i\}-\big[\mathcal{R}_d(C_{n-1}^{kNN})-\alpha(d)\big]\big|}{k^{-1}+(k/n)^{4/d}}>\epsilon\bigg)\leqslant\epsilon^{-2}\mathbf{Var}\bigg(\frac{n^{-1}\sum_{i=1}^n\{\hat{\xi}_i-\xi_i\}}{k^{-1}+(k/n)^{4/d}}\bigg)\\&=\epsilon^{-2}[k^{-1}+(k/n)^{4/d}]^{-2}\cdot\mathbf{Var}\Big[\frac{1}{n}\sum_{i=1}^n\big\{\hat{\xi}_i(d)-\xi_i(d)\big\}\Big]\\&=\epsilon^{-2}[k^{-1}+(k/n)^{4/d}]^{-2}\cdot o\big((1/k)^2+(k/n)^{8/d}\big)\to 0,\quad\text{as } n\to\infty.\end{aligned}$$

Thus, equation (S.44) is shown.

Using Theorem 1 in Samworth et al. (2012) with $w_{ni} = k^{-1}$ for $i \in \{1, \ldots, k\}$,

$$\begin{aligned} \mathcal{R}_{d,n-1}^{kNN} - \zeta_0(d) &= \left\{\zeta_1(d)k^{-1} + \zeta_2(d)\left(\frac{k}{n-1}\right)^{\frac{4}{d}}\right\}\{1 + o_p(1)\} \\ &= \left\{\zeta_1(d)k^{-1} + \zeta_2(d)(k/n)^{\frac{4}{d}}\right\}\{1 + o_p(1)\}. \end{aligned} \tag{S.50}$$

Substitute (S.50) into (S.44), we have

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}\left\{\left|Y_i - \mathbf{1}(\hat{p}_{d,\setminus i}(\mathbf{U}_i) > \frac{1}{2})\right| - \left|Y_i - \mathbf{1}(p_d(\mathbf{U}_i) > \frac{1}{2})\right|\right\} \\ =& \left\{\zeta_1(d)k^{-1} + \zeta_2(d)(k/n)^{\frac{4}{d}}\right\}\{1 + o_p(1)\} + o_p\big(k^{-1} + (k/n)^{\frac{4}{d}}\big) \\ =& \left\{\zeta_1(d)k^{-1} + \zeta_2(d)(k/n)^{\frac{4}{d}}\right\}\{1 + o_p(1)\} \end{aligned} \tag{S.51}$$

In addition, by the central limit theorem, we have

$$\frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \mathbf{1}(p_d(\mathbf{U}_i) > \frac{1}{2})\right| = \zeta_0(d) + O_p(n^{-1/2}) = \zeta_0(d) + o_p\big(k^{-1}\big). \tag{S.52}$$

Thus, substitute (S.51) and (S.52) to (0.3),

$$CV(d,k) = \zeta_0(d) + \left\{\zeta_1(d)k^{-1} + \zeta_2(d)\left(\frac{k}{n}\right)^{\frac{4}{d}}\right\}\{1 + o_p(1)\},$$

which completes the proof. □

Proof of Theorem 3. By Lemma 7, the cross-validation (or prediction error) $CV(d,k)$ is bigger than the first term $\zeta_0(d)$, which is the smallest risk one might be attained by any classifier based on $\mathcal{S}_n^{PD}$. Since HOPG method can order the projected directions in order with importance, we can suppose Y only depends on the first d_0 directions of $\mathbf{Z} = (z_1, \ldots, z_m)$, i.e. $Y|z_1, \ldots, z_m$ and $Y|z_1, \ldots, z_{d_0}$ have same distribution. Assuming d_0 is the smallest true dimension, we first show that

$$\begin{aligned} \zeta_0(d_0) < \zeta_0(d) \quad &\text{for} \quad 0 \leqslant d < d_0, \\ \zeta_0(d_0) = \zeta_0(d) \quad &\text{for} \quad d_0 \leqslant d \leqslant m. \end{aligned} \tag{S.53}$$

By the definition of $\zeta_0(d)$, for any d-dimensional classifier C_d,

$$\zeta_0(d) = \mathbf{P}\big\{Y \neq \mathbf{1}\big(p_d(z_1, \ldots, z_d) > \frac{1}{2}\big)\big\} = \min_{C_d}\mathbf{P}\big(Y \neq C_d(z_1, \ldots, z_d)\big).$$

Hence, it is obvious that $\zeta_0(d) \leqslant \zeta_0(d-1)$. Since class label is either 0 or 1, we have

$$
\begin{aligned}
&\zeta_0(d-1) - \zeta_0(d) \\
= &\mathbf{E}\Big\{\mathbf{E}\Big[\mathbf{1}\Big(Y \neq \mathbf{1}\big(p_{d-1}(z_1,\ldots,z_{d-1}) > \frac{1}{2}\big) - \mathbf{1}\Big(Y \neq \mathbf{1}\big(p_d(z_1,\ldots,z_d) > \frac{1}{2}\big)\Big|(z_1,\ldots,z_d)\Big]\Big\} \\
= &\mathbf{E}\Big[p_d(z_1,\ldots,z_d)\mathbf{1}\big(p_{d-1}(z_1,\ldots,z_{d-1}) \leqslant \frac{1}{2}\big) + \big(1-p_d(z_1,\ldots,z_d)\big)\mathbf{1}\big(p_{d-1}(z_1,\ldots,z_{d-1}) > \frac{1}{2}\big) \\
&\quad - p_d(z_1,\ldots,z_d)\mathbf{1}\big(p_d(z_1,\ldots,z_d) \leqslant \frac{1}{2}\big) - \big(1-p_d(z_1,\ldots,z_d)\big)\mathbf{1}\big(p_d(z_1,\ldots,z_d) > \frac{1}{2}\big)\Big] \\
= &\mathbf{E}\Big[\big|2p_d(z_1,\ldots,z_d) - 1\big|\mathbf{1}\Big\{\mathbf{1}\big(p_d(z_1,\ldots,z_d) > \frac{1}{2}\big) \neq \mathbf{1}\big(p_{d-1}(z_1,\ldots,z_{d-1}) > \frac{1}{2}\big)\Big\}\Big],
\end{aligned}
$$

where $p_{d-1}(z_1,\ldots,z_{d-1}) = \mathbf{E}\big[\mathbf{1}\big(Y=1\big)\big|(z_1,\ldots,z_{d-1})\big]$. By assumption (B4), we can get $\mathbf{P}\big(p_d(z_1,\ldots,z_d) = \frac{1}{2}\big) = 0$ (it can be derived from equation (2.1) in Samworth et al. (2012)). Therefore, if $\zeta_0(d-1) = \zeta_0(d)$, $\mathbf{1}\big(p_d(z_1,\ldots,z_d) > \frac{1}{2}\big) = \mathbf{1}\big(p_{d-1}(z_1,\ldots,z_{d-1}) > \frac{1}{2}\big)$ almost surely. Specifically, if $\zeta_0(d_0) = \zeta_0(d_0-1)$, we have almost surely

$$\mathbf{1}\big(p_{d_0}(z_1,\ldots,z_{d_0}) > \frac{1}{2}\big) = \mathbf{1}\big(p_{d_0-1}(z_1,\ldots,z_{d_0-1}) > \frac{1}{2}\big).$$

This contradicts the definition of d_0 (smallest true dimension). So $\zeta_0(d_0) < \zeta_0(d_0-1) \leqslant \zeta_0(d)$ for $d < d_0$.

Obviously, because class Y only depends on the first $d_0(< m)$ features of Z, for $d > d_0$,

$$p_d(z_1,\ldots,z_d) = p_{d-1}(z_1,\ldots,z_{d-1}) \quad \text{a.s.}$$

which leads to $\zeta_0(d_0) = \zeta_0(d_0+1) = \cdots = \zeta_0(m)$. Hence, we can get (S.53).

Then, using the the asymptotic expansion in Lemma 7, Theorem 4 can be shown by the following two parts,

(a) for $1 \leqslant d < d_0$, $\lim_{n\to\infty}\{P(\hat{d} = d)\} = 0$;

(b) for $d_0 < d \leqslant m$, $\lim_{n\to\infty}\{P(\hat{d} = d)\} = 0$.

According to the proof of Lemma 7, it is obvious that $CV(d,k) - \zeta_0(d) = o_p(1)$ as $n \to \infty$. Thus, for every k and $1 \leqslant d < d_0$, as $n \to \infty$,

$$\frac{CV(d,k)}{CV(d_0,k)} \to \frac{\zeta_0(d) + o_p(1)}{\zeta_0(d_0) + o_p(1)} > 1, \quad \text{in probability.}$$

Then, for every k,

$$
\begin{aligned}
&\mathbf{P}\{CV(d,k) \leqslant CV(d',k), 1 \leqslant d' \leqslant m\} \leqslant \mathbf{P}\{CV(d,k) \leqslant CV(d_0,k)\} \\
= &\mathbf{P}\left(\frac{CV(d,k)}{CV(d_0,k)} \leqslant 1\right) = \mathbf{P}\left(\frac{\zeta_0(d) + o_p(1)}{\zeta_0(d_0) + o_p(1)} \leqslant 1\right) \to 0 \quad \text{as } n \to \infty.
\end{aligned}
$$

Therefore, it follows that

$$\lim_{n\to\infty}\{\mathbf{P}(\hat{d}=d)\}=0 \quad \text{for } 0\leqslant d<d_0,$$

which proves part (a).

By Theorem 3, the optimal choice of k can be derived as

$$k_{opt}=\left\lfloor \left(\frac{\zeta_1(d)}{\zeta_2(d)}\times\frac{d}{4}\right)^{d/(d+4)} n^{4/(d+4)}\right\rfloor.$$

It is obvious that k_{opt} satisfies the restriction of k in (A5) when $d>7$. Then, we have

$$\begin{aligned}
CV(d,k_{opt}) &= \zeta_0(d)+\left\{\zeta_1(d)\left(\frac{\zeta_1(d)}{\zeta_2(d)}\cdot\frac{d}{4}\right)^{\frac{-d}{d+4}} n^{\frac{-4}{d+4}}+\zeta_2(d)\left(\frac{\zeta_1(d)}{\zeta_2(d)}\cdot\frac{d}{4}\right)^{\frac{4}{d+4}} n^{\frac{-4}{d+4}}\right\}\{1+o_p(1)\}\\
&= \zeta_0(d)+\left\{\left[\left(\frac{d}{4}\right)^{\frac{-d}{d+4}}+\left(\frac{d}{4}\right)^{\frac{4}{d+4}}\right]\zeta_1(d)^{\frac{4}{d+4}}\zeta_2(d)^{\frac{d}{d+4}}n^{\frac{-4}{d+4}}\right\}\{1+o_p(1)\}\\
&= \zeta_0(d)+\beta(d)n^{\frac{-4}{d+4}}+o_p(n^{\frac{-4}{d+4}}),
\end{aligned} \tag{S.54}$$

where $\beta(d)$ is a constant depending on d.

For part (b), let $d>d_0$, it follows from (S.53) that $\zeta_0(d)=\zeta_0(d_0)$. Since, d is bounded, we can assume M is its upper bound. When $7<d_0<d\leqslant\min\{m,M\}$, by equation (S.54) and $\zeta_0(d)=\zeta_0(d_0)$, we have

$$\begin{aligned}
\min_k CV(d,k)-\min_k CV(d_0,k) &= \left\{\beta(d)n^{\frac{-4}{d+4}}+o_p(n^{\frac{-4}{d+4}})\right\}-\left\{\beta(d_0)n^{\frac{-4}{d_0+4}}+o_p(n^{\frac{-4}{d_0+4}})\right\}\\
&= n^{\frac{-4}{d+4}}\left\{\beta(d)-\beta(d_0)n^{\frac{4}{d+4}-\frac{4}{d_0+4}}\right\}+o_p(n^{\frac{-4}{d+4}})\\
&\sim n^{\frac{-4}{d+4}}\beta(d)+o_p(n^{\frac{-4}{d+4}}) \quad \text{as } n\to\infty.
\end{aligned}$$

Since $\beta(d)>0$, for $7<d_0<d\leqslant\min\{m,M\}$,

$$\begin{aligned}
\lim_{n\to\infty}\{\mathbf{P}(\hat{d}=d)\} &= \lim_{n\to\infty}\mathbf{P}\{\min_k CV(d,k)\leqslant\min_k CV(d',k), 1\leqslant d'\leqslant p\}\\
&\leqslant \lim_{n\to\infty}\mathbf{P}\{\min_k CV(d,k)\leqslant\min_k CV(d_0,k)\}\\
&= \lim_{n\to\infty}\mathbf{P}\{n^{\frac{4}{d+4}}[\min_k CV(d,k)-\min_k CV(d_0,k)]\leqslant 0\}\\
&= \mathbf{P}\{\beta(d)+o_p(1)\leqslant 0\}=0.
\end{aligned}$$

When $d_0<d\leqslant 7$, for ever k,

$$\frac{CV(d,k)}{CV(d_0,k)}\to\frac{\zeta_0(d)+o_p(1)}{\zeta_0(d_0)+o_p(1)}\to_p 1, \quad \text{as } n\to\infty.$$

This formula means that $CV(d,k)=CV(d_0,k)$ in probability as $n\to\infty$. Since $\hat{d}$ is the smallest minimizer of $CV(d,k)$, $\lim_{n\to\infty}\{\mathbf{P}(\hat{d}=d)\}=0$ in this situation. Hence, we complete the proof of part (b).

Consequently,

$$\lim_{n\to\infty}\{\mathbf{P}(\hat{d}=d_0)\} = \lim_{n\to\infty}\{1-\mathbf{P}(\hat{d}\neq d_0)\}$$
$$= \lim_{n\to\infty}\{1-\sum_{1\leqslant d<d_0}\mathbf{P}(\hat{d}=d)-\sum_{d_0<d\leqslant\min\{m,M\}}\mathbf{P}(\hat{d}=d)\}=1.$$

We complete the proof. $\square$

Proof of Theorem 4. First of all, using HOPG estimation, we have by Theorem 2

$$\big|\widehat{D}\widehat{D}^\top - D_0D_0^\top\big| \to 0, \quad \text{in probability.}$$

Then,

$$\begin{aligned}
&\big|\mathbf{E}\big(Y|D_0^\top P^\top\mathbf{X}=D_0^\top P^\top\mathbf{x}\big)-\mathbf{E}\big(Y|\widehat{D}^\top P^\top\mathbf{X}=\widehat{D}^\top P^\top\mathbf{x}\big)\big|\\
&=\big|\mathbf{E}\big(Y|D_0D_0^\top P^\top\mathbf{X}=D_0D_0^\top P^\top\mathbf{x}\big)-\mathbf{E}\big(Y|\widehat{D}\widehat{D}^\top P^\top\mathbf{X}=\widehat{D}\widehat{D}^\top P^\top\mathbf{x}\big)\big|\\
&\leq\big|\mathbf{E}\big(Y|D_0D_0^\top P^\top\mathbf{X}=D_0D_0^\top P^\top\mathbf{x}\big)-\mathbf{E}\big(Y|D_0D_0^\top P^\top\mathbf{X}=\widehat{D}\widehat{D}^\top P^\top\mathbf{x}\big)\big|\\
&\quad+\big|\mathbf{E}\big(Y|D_0D_0^\top P^\top\mathbf{X}=\widehat{D}\widehat{D}^\top P^\top\mathbf{x}\big)-\mathbf{E}\big(Y|\widehat{D}\widehat{D}^\top P^\top\mathbf{X}=\widehat{D}\widehat{D}^\top P^\top\mathbf{x}\big)\big|\\
&\to 0.
\end{aligned} \tag{S.55}$$

By definition of D_0 and $Y\in\{0,1\}$,

$$p(\mathbf{x})=\mathbf{P}\big(Y=1|P^\top\mathbf{X}=P^\top\mathbf{x}\big)=\mathbf{E}\big(Y|P^\top\mathbf{X}=P^\top\mathbf{x}\big)=\mathbf{E}\big(Y|D_0^\top P^\top\mathbf{X}=D_0^\top P^\top\mathbf{x}\big). \tag{S.56}$$

In addition, by the consistency of kNN regression (e.g. Devroye et al. (1994)),

$$\hat{p}(\mathbf{x})\to\mathbf{P}\big(Y=1|\widehat{D}^\top P^\top\mathbf{X}=\widehat{D}^\top P^\top\mathbf{x}\big)=\mathbf{E}\big(Y|\widehat{D}^\top P^\top\mathbf{X}=\widehat{D}^\top P^\top\mathbf{x}\big)\quad\text{a.s.} \tag{S.57}$$

Combine (S.55), (S.56) and (S.57), we have

$$\lim_{n\to\infty}\hat{p}(\mathbf{x})\to p(\mathbf{x})\quad\text{in probability.}$$

We complete the proof. $\square$

In the multi-categorical cases, Theorem 3 and Theorem 4 hold with assumptions (B1), (B2'), (B3), (B4'), (B5) and (B6) given in the Appendix. Lemma 7 is still true by replacing constants ζ_1 and ζ_2 by $\tilde{\zeta}_1=\sum_{\ell_1\neq\ell_2}\zeta_{1,\ell_1,\ell_2}$ and $\tilde{\zeta}_2=\sum_{\ell_1\neq\ell_2}\zeta_{2,\ell_1,\ell_2}$ (c.f. Samworth et al. (2012)). The definitions of ζ_{1,ℓ_1,ℓ_2} and ζ_{2,ℓ_1,ℓ_2} are

$$\tilde{\zeta}_{1,\ell_1,\ell_2}(d)=\int_{\Omega_{\ell_1,\ell_2}}\frac{f_d(\mathbf{u}_0)}{p_d^{\ell_1,\ell_2}(\mathbf{u}_0)(1-p_d^{\ell_1,\ell_2}(\mathbf{u}_0))\|\dot{p}_d^{\ell_1,\ell_2}(\mathbf{u}_0)\|}dVol^{d-1}(\mathbf{u}_0)$$

and

$$\tilde{\zeta}_{2,\ell_1,\ell_2}(d) = \int_{\Omega_{\ell_1,\ell_2}} \frac{f_d(\mathbf{u}_0)}{\|\dot{p}_d^{\ell_1,\ell_2}(\mathbf{u}_0)\|} a_{\ell_1,\ell_2}(\mathbf{u}_0)^2 dVol^{d-1}(\mathbf{u}_0),$$

where $p_d^{\ell_1,\ell_2}(\mathbf{u}_0)$ denotes the common value that $p_d^{\ell_1}$ and $p_d^{\ell_2}$ take at $\mathbf{u}_0 \in \Omega_{\ell_1,\ell_2}$, and $a_{\ell_1,\ell_2}(\cdot)$ can be obtained by changing $p_d(\cdot)$ to $p_d^{\ell_1,\ell_2}(\cdot)$ in the definition of $a(\cdot)$.

References

Chaudhuri, K. and Dasgupta, S. (2014). Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 3437–3445.

Devroye, L., Gyorfi, L., Krzyzak, A., and Lugosi, G. (1994). On the strong universal consistency of nearest neighbor regression function estimates. *The Annals of Statistics*, pages 1371–1385.

Fan, J. and Yao, Q. (2008). *Nonlinear time series: nonparametric and parametric methods*. Springer Science & Business Media.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

Samworth, R. J. et al. (2012). Optimal weighted nearest neighbour classifiers. *The Annals of Statistics*, 40(5):2733–2763.

Székely, G. J., Rizzo, M. L., Bakirov, N. K., et al. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794.