# Kernel Ridge Density Estimation in Smoothing Spline ANOVA Models: a Random Sketching Approach

October 18, 2023

## 1. Introduction

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in [0,1]^r \stackrel{\text{def}}{=} \mathcal{X}$ be i.i.d. copies of a random vector $\boldsymbol{X}$. We assume that the the density of $\boldsymbol{X}$ is $p_0(\boldsymbol{x}) = e^{f_0(\boldsymbol{x})}$, i.e., for any Lebsgure measurable set $A \subset \mathcal{X}$, it holds that

$$\mathbb{P}(\boldsymbol{X} \in A) = \int_A p_0(\boldsymbol{x}) d\boldsymbol{x} = \int_A e^{f_0(\boldsymbol{x})} d\boldsymbol{x}.$$

.

Let us assume $f_0 \in \mathcal{H}$ for some reproducing kernel Hilbert space $\mathcal{H}$ with reproducing kernel $R(\cdot,\cdot)$ and norm $\|\cdot\|_{\mathcal{H}}$. The following kernel ridge density estimator is widely used in the literature:

$$\widehat{f}_{KRD} = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \left\{ L_{n,\lambda}(f) := -\frac{1}{n} \sum_{i=1}^n f(\boldsymbol{X}_i) + \int_{\mathcal{X}} e^{f(\boldsymbol{x})} d\boldsymbol{x} + \lambda \|f\|_{\mathcal{H}}^2 \right\}.$$

By Riesz representation theorem, the analytic expression of the KRD estimator is $\widehat{f}_{KRD}(\boldsymbol{x}) = \widehat{\boldsymbol{c}}_{KRD}^\top \boldsymbol{\Psi}(\boldsymbol{x})$, where

$$\widehat{\boldsymbol{c}}_{KRD} = \underset{\boldsymbol{c} \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ -\frac{1}{n} \mathbf{1}^\top \boldsymbol{R}\boldsymbol{c} + \int_{\mathcal{X}} e^{\boldsymbol{c}^\top \boldsymbol{\Psi}(\boldsymbol{x})} d\boldsymbol{x} + \lambda \boldsymbol{c}^\top \boldsymbol{R}\boldsymbol{c} \right\},$$

and $\boldsymbol{\Psi}(\boldsymbol{x}) = (R(\boldsymbol{X}_1, \boldsymbol{x}), \ldots, R(\boldsymbol{X}_n, \boldsymbol{x}))^\top \in \mathbb{R}^n$, $\boldsymbol{R} = [R(\boldsymbol{X}_i, \boldsymbol{X}_j)] \in \mathbb{R}^{n \times n}$.

To reduce computation, we consider a random sketching approach. Let $S \in \mathbb{R}^{p \times n}$ be a random sketching matrix with $p << n$. We consider the following random sketching estimator:

$$\widehat{f} = \underset{f \in \mathcal{H}_S}{\operatorname{argmin}} L_{n,\lambda}(f),$$

where $\mathcal{H}_S = \{f \in \mathcal{H} : f(\boldsymbol{x}) = \boldsymbol{c}^\top S \boldsymbol{\Psi}(\boldsymbol{x}) \text{ for } \boldsymbol{c} \in \mathbb{R}^p\}$. Hence it follows that $\widehat{f}(\boldsymbol{x}) = \widehat{\boldsymbol{c}}^\top S \boldsymbol{\Psi}(\boldsymbol{x})$, where

$$\widehat{\boldsymbol{c}} = \underset{\boldsymbol{c} \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ -\frac{1}{n} \mathbf{1}^\top R S^\top \boldsymbol{c} + \int_{\mathcal{X}} e^{\boldsymbol{c}^\top S \boldsymbol{\Psi}(\boldsymbol{x})} d\boldsymbol{x} + \lambda \boldsymbol{c}^\top S R S^\top \boldsymbol{c} \right\}.$$

Compare with the classical KRD, the random sketched estimator only requires estimating a $p$-dimensional vector.

## 2. Upper Bound

**Assumption A.** *d*

**Lemma 1.** *If $\pi_{\boldsymbol{X}}$ is the uniform distribution over $\mathcal{X}$, then*

$$\psi_{\boldsymbol{i}}(\boldsymbol{x}) = \varphi_{i_1}(x_1)\varphi_{i_2}(x_2)\dots\varphi_{i_r}(x_r), \tag{2.1}$$

*where $\varphi_1(x) = 1, \varphi_{2i}(x) = \sqrt{2}\sin(2i\pi x), \varphi_{2i+1}(x) = \sqrt{2}\cos(2i\pi x)$ for $i \in \mathbb{N}$.*

**Lemma 2.** *Under Assumption xxx, the following statements hold for some constant $C > 0$.*

*(i)* $\|\mathcal{W}_\lambda f\|_{L_2} \leqslant \|\mathcal{W}_\lambda f\|_\lambda \leqslant C\lambda^{1/2}\|f\|_{\mathcal{H}}$;

*(ii)* $\sup_{\boldsymbol{x},\boldsymbol{x}'\in\mathcal{X}} |K(\boldsymbol{x},\boldsymbol{x}')| \leqslant Ch^{-1}$;

*(iii)* $\|f\|_{\sup} \leqslant Ch^{-1/2}\|f\|_\lambda$, $\|f\|_{\sup} \leqslant C\|f\|_{\mathcal{H}}$;

*(iv)* $\|\mathcal{W}_\lambda f\|_{L_2} \leqslant \|f\|_{L_2}$;

*Proof.* (i) By definition, we have

$$\|\mathcal{W}_\lambda f\| = \sup_{\|g\|=1} \langle \mathcal{W}_\lambda f, g\rangle = \sup_{\|g\|=1} \lambda\langle f, g\rangle_{\mathcal{H}} \leqslant \lambda\|f\|_{\mathcal{H}} \sup_{\|g\|=1} \|g\|_{\mathcal{H}}.$$

Since $\|g\|^2 = V(g,g) + \lambda\|g\|^2_{\mathcal{H}}$, we see that $\lambda^{1/2}\|g\|_{\mathcal{H}} \leqslant \|g\| = 1$, which completes the proof.

(ii) By Assumption xxx, we have

$$|K(\boldsymbol{x},\boldsymbol{x})| = \left|\sum_{\boldsymbol{i}\in\mathbb{I}} \frac{\psi_{\boldsymbol{i}}(\boldsymbol{x})\psi'_{\boldsymbol{i}}(\boldsymbol{x})}{1+\lambda/\rho_{\boldsymbol{i}}}\right| \leqslant \sum_{\boldsymbol{i}\in\mathbb{I}} \frac{C}{1+\lambda/\rho_{\boldsymbol{i}}} = Ch^{-1}.$$

(iii) Using statement (ii), we show that $|f(\boldsymbol{x})| = |\langle f, K_{\boldsymbol{x}}\rangle| \leqslant \|f\|\|K_{\boldsymbol{x}}\| = \|f\|\sqrt{K(\boldsymbol{x},\boldsymbol{x})} \leqslant C\|f\|h^{-1/2}$. Similarly, we have $|f(\boldsymbol{x})| = |\langle f, R_{\boldsymbol{x}}\rangle_{\mathcal{H}}| \leqslant \|f\|_{\mathcal{H}}\|R_{\boldsymbol{x}}\|_{\mathcal{H}} = \|f\|_{\mathcal{H}}\sqrt{R(\boldsymbol{x},\boldsymbol{x})} \leqslant C\|f\|_{\mathcal{H}}$.

(iv) For any $f \in \mathcal{H}$, it admits a series expansion $f = \sum_{\boldsymbol{i}\in\mathbb{I}} c_{\boldsymbol{i}}\psi_{\boldsymbol{i}}$ with $c_{\boldsymbol{i}} = V(f,\psi_{\boldsymbol{i}})$. Since $\mathcal{W}_\lambda\psi_{\boldsymbol{i}} = \lambda\psi_{\boldsymbol{i}}/(\lambda+\rho_{\boldsymbol{i}})$, we show that $\mathcal{W}_\lambda f = \sum_{\boldsymbol{i}\in\mathbb{I}} c_{\boldsymbol{i}}\lambda\psi_{\boldsymbol{i}}/(\lambda+\rho_{\boldsymbol{i}})$ and

$$V(\mathcal{W}_\lambda f, \mathcal{W}_\lambda f) = \sum_{\boldsymbol{i}\in\mathbb{I}} \frac{\lambda^2 c_{\boldsymbol{i}}^2}{(\lambda+\rho_{\boldsymbol{i}})^2} \leqslant \sum_{\boldsymbol{i}\in\mathbb{I}} c_{\boldsymbol{i}}^2 = V(f,f).$$

□

**Theorem 1.** *Upper bound*

*Proof.* The result will be proved by contradiction. Let us assume that for some $\epsilon, B_\epsilon > 0$, it holds that $\mathbb{P}(E_{n,B}) \geqslant \epsilon$ for all $B \geqslant B_\epsilon$. Here $E_{n,B} = \{\|\widehat{f} - \widehat{f}_*\|_\lambda \geqslant B\delta_n, \|\widehat{f}_* - f_0\|_\lambda \leqslant B\delta_n/K\}$ is an event for some $0 < K < B$. W.L.O.G, we can assume $B_\epsilon \geqslant 1$.

On event $E_{n,B}$, the definition of $\widehat{f}$ implies that

$$\inf_{f\in\mathcal{H}_S:\|f-\widehat{f}_*\|_\lambda\geqslant B\delta_n} L_{n,\lambda}(f) - L_{n,\lambda}(\widehat{f}_*) \leqslant 0.$$

By convexity of $f \to L_{n,\lambda}(f)$, it holds that

$$\inf_{f \in \mathcal{H}_S : \|f - \widehat{f}_*\|_\lambda = B\delta_n} L_{n,\lambda}(f) - L_{n,\lambda}(\widehat{f}_*) \leqslant 0.$$

Hence, there is a sequence $f_n \in \mathcal{H}_S$ such that $\|f_n - \widehat{f}_*\|_\lambda = B\delta_n$ and $L_{n,\lambda}(f_n) - L_{n,\lambda}(\widehat{f}_*) \leqslant 0$. Let $g_n = f_n - f_0$, and it follows from triangular inequality that

$$B(1 - 1/K)\delta_n \leqslant \|g\|_\lambda \leqslant 2B\delta_n.$$

As a consequence, it holds on event $E_{n,B}$ that

$$L_{n,\lambda}(f_0 + g_n) - L_{n,\lambda}(f_0) \leqslant L_{n,\lambda}(\widehat{f}_*) - L_{n,\lambda}(f_0).$$

By direct examination, it follows that

$$\begin{aligned}
& L_{n,\lambda}(f_0 + g_n) - L_{n,\lambda}(f_0) \\
=& - \mathbb{P}_n g_n + \int_{\mathcal{X}} e^{f_0(\boldsymbol{x})} \left\{ e^{g_n(\boldsymbol{x})} - 1 \right\} d\boldsymbol{x} + \lambda \|f_0 + g_n\|_{\mathcal{H}}^2 - \lambda \|f_0\|_{\mathcal{H}}^2 \\
=& - \mathbb{P}_n g_n + \int_{\mathcal{X}} e^{f_0(\boldsymbol{x})} \left\{ e^{g_n(\boldsymbol{x})} - 1 \right\} d\boldsymbol{x} + \lambda \|g_n\|_{\mathcal{H}}^2 + 2\lambda \langle f_0, g_n \rangle_{\mathcal{H}} \\
\overset{\text{(i)}}{\geqslant}& - \kappa_n \|g_n\|_\lambda - \mathbb{P} g_n + \int_{\mathcal{X}} e^{f_0(\boldsymbol{x})} \left\{ e^{g_n(\boldsymbol{x})} - 1 \right\} d\boldsymbol{x} + \lambda \|g_n\|_{\mathcal{H}}^2 + 2\lambda \langle f_0, g_n \rangle_{\mathcal{H}} \\
=& - \kappa_n \|g_n\|_\lambda + \int_{\mathcal{X}} e^{f_0(\boldsymbol{x})} \left\{ e^{g_n(\boldsymbol{x})} - 1 - g_n(\boldsymbol{x}) - \frac{1}{2} g_n^2(\boldsymbol{x}) \right\} d\boldsymbol{x} + \lambda \|g_n\|_{\mathcal{H}}^2 + 2\lambda \langle f_0, g_n \rangle_{\mathcal{H}} + \frac{1}{2} \|g_n\|_{L_2}^2 \\
\overset{\text{(ii)}}{\geqslant}& - \kappa_n \|g_n\|_\lambda - \frac{1}{4} \|g_n\|_{L_2}^2 + \lambda \|g_n\|_{\mathcal{H}}^2 + 2\lambda \langle f_0, g_n \rangle_{\mathcal{H}} + \frac{1}{2} \|g_n\|_{L_2}^2 \\
=& - \kappa_n \|g_n\|_\lambda + \frac{1}{4} \|g_n\|_{L_2}^2 + \lambda \|g_n\|_{\mathcal{H}}^2 + 2\lambda \langle f_0, g_n \rangle_{\mathcal{H}} \\
\overset{\text{(iii)}}{\geqslant}& - \kappa_n \|g_n\|_\lambda + \frac{1}{4} \|g_n\|_\lambda^2 - 2\lambda \|f_0\|_{\mathcal{H}} \|g_n\|_{\mathcal{H}} \\
\overset{\text{(iv)}}{\geqslant}& - \kappa_n \|g_n\|_\lambda + \frac{1}{4} \|g_n\|_\lambda^2 - 2\lambda^{1/2} \|f_0\|_{\mathcal{H}} \|g_n\|_\lambda,
\end{aligned}$$

where xxx. By similar argument, we have

$$\begin{aligned}
L_{n,\lambda}(\widehat{f}_*) - L_{n,\lambda}(f_0) &\leqslant \kappa_n \|\widehat{f}_* - f_0\|_\lambda + \|\widehat{f}_* - f_0\|_\lambda^2 + 2\lambda \|f_0\|_{\mathcal{H}} \|\widehat{f}_* - f_0\|_{\mathcal{H}} \\
&\overset{\text{(i)}}{\leqslant} B\kappa_n \delta_n / K + B^2 \delta_n^2 / K^2 + 2\lambda^{1/2} \|f_0\|_{\mathcal{H}} \delta_n / K.
\end{aligned}$$

Here xxx.

Combining the above two displays, we have

$$\frac{1}{4} \|g_n\|_\lambda^2 \leqslant B\kappa_n \delta_n / K + B^2 \delta_n^2 / K^2 + 2\lambda^{1/2} \|f_0\|_{\mathcal{H}} \delta_n / K + \kappa_n \|g_n\|_\lambda + 2\lambda^{1/2} \|f_0\|_{\mathcal{H}} \|g_n\|_\lambda.$$

Since $x^2 \leqslant A + Bx$ implies $x \leqslant \sqrt{2A} + 2B \leqslant 2\sqrt{A} + 2B$, the preceding leads to

$$\begin{aligned}
\|g_n\|_\lambda &\leqslant 4\sqrt{B\kappa_n\delta_n/K + B^2\delta_n^2/K^2 + 2\lambda^{1/2}\|f_0\|_{\mathcal{H}}\delta_n/K} + 2(\kappa_n + 2\lambda^{1/2}\|f_0\|_{\mathcal{H}}) \\
&\overset{\text{(i)}}{\leqslant} 4\sqrt{B^2\kappa_n\delta_n/K^2 + B^2\delta_n^2/K^2 + 2B^2\lambda^{1/2}\|f_0\|_{\mathcal{H}}\delta_n/K^2} + 2(\kappa_n + 2\lambda^{1/2}\|f_0\|_{\mathcal{H}}) \\
&\leqslant 4BK^{-1}\sqrt{\kappa_n^2 + \delta_n^2 + \delta_n^2 + \|f_0\|_{\mathcal{H}}\lambda + \|f_0\|_{\mathcal{H}}\delta_n^2} + 4BK^{-1}(\kappa_n + \lambda^{1/2}\|f_0\|_{\mathcal{H}}) \\
&\leqslant 4BK^{-1}(2\delta_n + \|f_0\|_{\mathcal{H}}^{1/2}\delta_n + 2\kappa_n + \|f_0\|_{\mathcal{H}}\lambda^{1/2} + \|f_0\|_{\mathcal{H}}^{1/2}\lambda^{1/2}) \\
&\leqslant (8 + \|f_0\|_{\mathcal{H}} + \|f_0\|_{\mathcal{H}}^{1/2})BK^{-1}(\delta_n + \kappa_n + \lambda^{1/2}) \\
&\overset{\text{(ii)}}{\leqslant} 2(8 + \|f_0\|_{\mathcal{H}} + \|f_0\|_{\mathcal{H}}^{1/2})BK^{-1}\delta_n,
\end{aligned}$$

where xxx. Noting that $\|g_n\|_\lambda \geqslant B(1-1/K)\delta_n$ holds on event $E_{n,B}$, we conclude that

$$\mathbb{P}\left(B(1-1/K)\delta_n \leqslant \|g_n\|_\lambda \leqslant 2(8 + \|f_0\|_{\mathcal{H}} + \|f_0\|_{\mathcal{H}}^{1/2})BK^{-1}\delta_n\right) \geqslant \mathbb{P}(E_{n,B}) \geqslant \epsilon,$$

for all $B \geqslant B_\epsilon$. Now, we can choose $K$ such that

$$K - 1 > 2(8 + \|f_0\|_{\mathcal{H}} + \|f_0\|_{\mathcal{H}}^{1/2}).$$

Hence, xxx implies that $0 \geqslant \epsilon$, which is a contradiction.

$$\begin{aligned}
0 &\leqslant L_{n,\lambda}(\widehat{f}_*) - L_{n,\lambda}(f_0 + g_n) \\
&= L_{n,\lambda}(\widehat{f}_*) - L_{n,\lambda}(f_0 + g_n) \\
&= \mathbb{P}_n\widehat{f} - \mathbb{P}_n\widehat{f}_* + \int_{\mathcal{X}} e^{\widehat{f}_*(\boldsymbol{x})}d\boldsymbol{x} - \int_{\mathcal{X}} e^{\widehat{f}(\boldsymbol{x})}d\boldsymbol{x} + \lambda\|\widehat{f}_*\|_{\mathcal{H}}^2 - \lambda\|\widehat{f}\|_{\mathcal{H}}^2 \\
&\leqslant -\kappa_n\|\widehat{f} - f_0\|_{L_2} + \kappa_n\|\widehat{f}_* - f_0\|_{L_2} + \mathbb{P}(\widehat{f} - \mathbb{P}\widehat{f}_* \\
&\quad + \int_{\mathcal{X}} e^{\widehat{f}_*(\boldsymbol{x})}d\boldsymbol{x} - \int_{\mathcal{X}} e^{\widehat{f}(\boldsymbol{x})}d\boldsymbol{x} + \lambda\|\widehat{f}_*\|_{\mathcal{H}}^2 - \lambda\|\widehat{f}\|_{\mathcal{H}}^2 \\
&\leqslant C\kappa_n\|\widehat{f}\|_{\mathcal{H}} + C\kappa_n\|\widehat{f}_*\|_{\mathcal{H}} + \mathbb{P}\widehat{f} - \mathbb{P}\widehat{f}_* + \int_{\mathcal{X}} e^{\widehat{f}_*(\boldsymbol{x})}d\boldsymbol{x} - \int_{\mathcal{X}} e^{\widehat{f}(\boldsymbol{x})}d\boldsymbol{x} + \lambda\|\widehat{f}_*\|_{\mathcal{H}}^2 - \lambda\|\widehat{f}\|_{\mathcal{H}}^2
\end{aligned}$$

$$\begin{aligned}
&-\mathbb{P}(\widehat{f} - f_0) + \int_{\mathcal{X}} e^{f_0(\boldsymbol{x})}\left\{e^{\widehat{f}(\boldsymbol{x}) - f_0(\boldsymbol{x})} - 1\right\} + \lambda\|\widehat{f}\|_{\mathcal{H}}^2 - C\kappa_n\|\widehat{f}\|_{\mathcal{H}} \\
&\leqslant -\mathbb{P}(\widehat{f}_* - f_0) + \int_{\mathcal{X}} e^{f_0(\boldsymbol{x})}\left\{e^{\widehat{f}_*(\boldsymbol{x}) - f_0(\boldsymbol{x})} - 1\right\} + \lambda\|\widehat{f}_*\|_{\mathcal{H}}^2 + C\kappa_n\|\widehat{f}_*\|_{\mathcal{H}} \\
&\leqslant C\|\widehat{f}_* - f_0\|_{L_2}^2 + \lambda\|\widehat{f}_*\|_{\mathcal{H}}^2 + C\kappa_n\|\widehat{f}_*\|_{\mathcal{H}} \\
&\leqslant C(\delta_n^2 + \lambda + \kappa_n).
\end{aligned}$$

$$\delta_n = n^{-\frac{2m}{2m+1}} +$$

Hence, it holds that

$$\|\widehat{f}\|_{\mathcal{H}}^2 \leqslant C(\delta_n^2/\lambda + 1 + \kappa_n/\lambda + \kappa_n^2/\lambda^2) = O_P(1).$$

$$\int_{\mathcal{X}} e^{f(\boldsymbol{x})} d\boldsymbol{x} - \int_{\mathcal{X}} e^{f_0(\boldsymbol{x})} d\boldsymbol{x}$$

□

**Lemma 3.**

*Proof.*

$$-\mathbb{E}\{g(\boldsymbol{X})\} - 1 + \int_{\mathcal{X}} e^{f_0(\boldsymbol{x})+g(\boldsymbol{x})} d\boldsymbol{x} = \int_{\mathcal{X}} e^{f_0(\boldsymbol{x})} \left\{ e^{g(\boldsymbol{x})} - 1 - g(\boldsymbol{x}) \right\} d\boldsymbol{x} \geqslant 0.$$

□

**Lemma 4.** $\|f\|_{\sup}^2 \leqslant Ch^{-1}(\|f\|_{L_2}^2 + \lambda\|f\|_{\mathcal{H}}^2)$

*Proof.* d □

**Lemma 5.** *For any $t > 0$, it holds that $e^{-t}x^2/2 \leqslant e^x - 1 - xe^t x^2/2$ when $x \in [-t, t]$*

*Proof.* We only prove the lower bound as the upper bound can be proved similarly. Let $g(x) = e^x - 1 - x - e^{-t}x^2/2$, and it holds that

$$g'(x) = e^x - 1 - e^{-t}x.$$

When $x \in [0, t]$, the inequality $e^x - 1 \geqslant x \geqslant e^{-t}x$ implies $g'(x) \geqslant 0$. When $x \in [-t, 0]$, mean value theorem implies that

$$1 - e^x = e^0 - e^x = -xe^{sx} \geqslant -xe^{-t},$$

where $s \in [0, 1]$. Therefore, we show that $g'(x) \leqslant 0$ when $x \in [-t, 0]$. Therefore, it follows that $g(x) \geqslant 0$ for all $x \in [-t, t]$ □

## 3. Some Lemmas

**Lemma 6.**

*Proof.* When $r = 1$, $\{\phi_i, i \geqslant 1\}$ is the Fourier basis of $\mathbb{S}_m$, and $\langle \phi_k, \phi_s \rangle_{L_2} = \delta_{ks}$. Here, we use the fact that $\pi_{\boldsymbol{X}}$ is the uniform density. Direct examination leads to

$$\langle \phi_k, \phi_k \rangle_{\mathbb{S}_m} \quad = \quad \begin{cases} 1 & \text{if } k = 1, \\ (2i\pi)^{2m} & \text{if } k = 2i, 2i-1 \text{ with } i \geqslant 1, \end{cases}$$

and $\langle \phi_k, \phi_s \rangle_{\mathbb{S}_m} = 0$ if $k \neq s$. Hence, we conclude that $\langle \phi_k, \phi_k \rangle_{\mathbb{S}_m} \asymp k^{2m}$.

When $r > 1$, the definition of tensor product space implies $\{\psi_{\boldsymbol{i}}, i \in \mathbb{I}_q\}$ is a basis of $\mathcal{H}$ under $\langle \cdot, \cdot \rangle_{L_2}$. Moreover, since $\pi_{\boldsymbol{X}}$ is the uniform density, using the result of $r = 1$, we can verify $\langle \psi_{\boldsymbol{i}}, \psi_{\boldsymbol{j}} \rangle_{L_2} = \delta_{\boldsymbol{ij}}$ and $\langle \psi_{\boldsymbol{i}}, \psi_{\boldsymbol{j}} \rangle_{\mathcal{H}} = \rho_{\boldsymbol{i}} \delta_{\boldsymbol{ij}}$ with $\rho_{\boldsymbol{i}} \asymp \boldsymbol{i}^{2m}$. $\square$

$$\langle f, g \rangle = V(f, g) + \lambda \langle f, g \rangle_{\mathcal{H}}, \quad \|f\|^2 = \langle f, f \rangle \tag{3.1}$$

### 3.1. Modeling Continuous Variables

$$H(x, y)$$

### 3.2. Modeling Discrete Variables

$$D(x, y) = I(x = y) = 1/S_i + \{I(x = y) - 1/S_i\}$$

$$\langle f, g \rangle_{\mathcal{H}_i} = \sum_{k=1}^{S_i} f(k) g(k)$$

$$\langle f, D_x \rangle_{\mathcal{H}_i} = \sum_{k=1}^{S_i} f(k) D_x(k) = f(x)$$

$$e_{i,1}(x) = 1,$$

### 3.3. Tensor Product Space

$$R^{(q)}(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{r} H(x_i, y_i) \times \prod_{i=r+1}^{r+d} D(x_i, y_i).$$

$$\mathbb{I} = \{\boldsymbol{i} = (i_1, \ldots, i_{r+d}) : i_1, \ldots, i_r \in \mathbb{N}_+, i_k \in [S_k] \text{ for } k = r+1, \ldots, r+d\}$$

$$\begin{aligned}
\mathbb{I}_{cc} &= \{\boldsymbol{i} \in \mathbb{I} : i_1 = 1 \text{ or } i_2 = 1\}, \\
\mathbb{I}_{cd} &= \{\boldsymbol{i} \in \mathbb{I} : i_1 = 1 \text{ or } i_{r+1} = 1\}, \\
\mathbb{I}_{dd} &= \{\boldsymbol{i} \in \mathbb{I} : i_{r+1} = 1 \text{ or } i_{r+2} = 1\},
\end{aligned}$$

$$\rho_{\boldsymbol{i}} \asymp i_1^{-2m} \dots i_r^{-2m}$$

$$\psi_{\boldsymbol{i}}(\boldsymbol{x}) = \phi_{i_1}(x_1) \dots \phi_{i_r}(x_r) e_{r+1,i_{r+1}}(x_{r+1}) \dots e_{r+d,i_{r+d}}(x_{r+d})$$

$$\begin{aligned}
\langle f, g \rangle_{L_2} &= \int_{\mathcal{X}} f(\boldsymbol{x}) g(\boldsymbol{x}) e^{f_0(\boldsymbol{x})} d\boldsymbol{x}, \\
V(f,g) &= \int_{\mathcal{X}} f(\boldsymbol{x}) g(\boldsymbol{x}) d\boldsymbol{x}, \\
\langle f, g \rangle &= V(f,g) + \lambda \langle f, g \rangle_{\mathcal{H}}
\end{aligned}$$

Let $\mathcal{D}_i = \{1, \dots, S_i\}$, let us define

$$\begin{aligned}
\widehat{f} &= \operatorname*{argmin}_{f \in \mathcal{H}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{X}_i) + \int_{\mathcal{X}} e^{f(\boldsymbol{x})} d\boldsymbol{x} + \lambda \|f\|_{\mathcal{H}}^2 \right\}, \\
\widehat{f}^* &= \operatorname*{argmin}_{f \in \mathcal{H}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} W_i f(\boldsymbol{X}_i) + \int_{\mathcal{X}} e^{f(\boldsymbol{x})} d\boldsymbol{x} + \lambda \|f\|_{\mathcal{H}}^2 \right\},
\end{aligned}$$

$$\begin{aligned}
L_{n,\lambda}(f) &= -\frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{X}_i) + \int_{\mathcal{X}} e^{f(\boldsymbol{x})} d\boldsymbol{x} + \lambda \|f\|_{\mathcal{H}}^2, \\
L_{n,\lambda}^*(f) &= -\frac{1}{n} \sum_{i=1}^{n} W_i f(\boldsymbol{X}_i) + \int_{\mathcal{X}} e^{f(\boldsymbol{x})} d\boldsymbol{x} + \lambda \|f\|_{\mathcal{H}}^2.
\end{aligned}$$

$$\kappa_n \stackrel{\text{def}}{=} \sup_{f \in \mathcal{F}_1} |(\mathbb{P}_n - \mathbb{P})(f)| = O_P\left(\frac{1}{\sqrt{nh}}\right). \tag{3.2}$$

**Assumption B.** *(i)* $f_0 \in \mathcal{H}$ *and* $\|f_0\|_{\mathcal{H}} < \infty$.

## 4. Some Lemmas

**Lemma 7.** *If $2mk > 1$ and $\lambda \to 0$, then it follows that*

$$\begin{aligned}
\sum_{\boldsymbol{i} \in \mathbb{I}} \frac{1}{(1+\lambda i_1^{2m} \dots i_r^{2m})^k} &\asymp \lambda^{-\frac{1}{2m}} [\log(1/\lambda)]^{r-1}, \\
\sum_{\boldsymbol{i} \in \mathbb{I}_{cc}} \frac{1}{(1+\lambda i_1^{2m} \dots i_r^{2m})^k} &\asymp \lambda^{-\frac{1}{2m}} [\log(1/\lambda)]^{r-2}, \\
\sum_{\boldsymbol{i} \in \mathbb{I}_{cd}} \frac{1}{(1+\lambda i_1^{2m} \dots i_r^{2m})^k} &\asymp \lambda^{-\frac{1}{2m}} [\log(1/\lambda)]^{r-1}, \\
\sum_{\boldsymbol{i} \in \mathbb{I}_{dd}} \frac{1}{(1+\lambda i_1^{2m} \dots i_r^{2m})^k} &\asymp \lambda^{-\frac{1}{2m}} [\log(1/\lambda)]^{r-1}.
\end{aligned}$$

*Proof.* This is xxx. □

## 5. Lower Bound

**Lemma 8.** *There is a constant $c_{m,r} > 0$ depending on $m, r$ such that if $\|f\|_{\mathcal{H}} \leqslant c_{m,r}$, then it holds that $\log(1+f) \in \mathcal{H}$ and $\|\log(1+f)\|_{\mathcal{H}} \geqslant \|f\|_{\mathcal{H}}/4$.*

*Proof.* If $f = 0$, the statements hold trivially. It suffices to consider the case when $f \neq 0$.

By Lemma 2.2 in Lin (2000), there is a constant $C_{m,r} \geqslant 1$ depending on $m, r$ such that

$$\|f^2\|_{\mathcal{H}} \leqslant C_{m,r} \|f\|_{\mathcal{H}}^2.$$

Let $c_{m,r} > 0$ be a small constant such that $C_{m,r} c_{m,r} \leqslant 1/3$, and the above inequality implies that

$$\|f^k\|_{\mathcal{H}} \leqslant C_{m,r}^{k-1} \|f\|_{\mathcal{H}}^k \leqslant C_{m,r}^{k-1} c_{m,r}^k \leqslant C_{m,r}^k c_{m,r}^k \leqslant 3^{-k}.$$

Let us define a sequence of functions

$$g_n(\boldsymbol{x}) = \sum_{k=0}^{n} \frac{(-1)^k [f(\boldsymbol{x})]^{k+1}}{k+1}.$$

By direct examination, it holds that

$$\|g_n - g_{n+s}\|_{\mathcal{H}} \leqslant \sum_{k=n+1}^{n+s} \frac{3^{-(k+1)}}{k+1} \leqslant \sum_{k=n+1}^{\infty} 3^{-(k+1)} \leqslant 3^{-(n+1)} \to 0 \quad \text{as } n, s \to \infty.$$

Therefore, we show that $g_n$ is a Cauchy sequence in $\mathcal{H}$. Since $\mathcal{H}$ is complete, there is a limit $g \in \mathcal{H}$ such that $\|g_n - g\|_{\mathcal{H}} \to 0$. In addition, Lemma 2(iii) implies that $\|g_n - g\|_{\sup} \to 0$.

Finally, let us verify $g = f$. Since Lemma 2(iii) implies that $\|f\|_{\sup} \leqslant C\|f\|_{\mathcal{H}} \leqslant C c_{m,k}$ for some $C > 0$ depending on $m, r$, we can choose $c_{m,k} > 0$ small enough such that $\|f\|_{\sup} \leqslant 1/2$. Hence, it follows from Taylor's expansion of $\log(1+x)$ that $\|g_n - \log(1+f)\|_{\sup} \to 0$. By the uniqueness of limit in supremum norm, we conclude that $g = \log(1+f) \in \mathcal{H}$, which is the first statement.

For the second statement, the Taylor's expansion of $\log(1-x)$ implies that

$$\begin{aligned}\|g_n - f\|_{\mathcal{H}} = \left\| \sum_{k=1}^{n} \frac{(-1)^k f^{k+1}}{k+1} \right\|_{\mathcal{H}} &\leqslant \sum_{k=1}^{n} \frac{1}{k+1} C_{m,r}^{k} \|f\|_{\mathcal{H}}^{k+1} \\ &= C_{m,r}^{-1} \sum_{k=0}^{n} \frac{1}{k+1} C_{m,r}^{k+1} \|f\|_{\mathcal{H}}^{k+1} - \|f\|_{\mathcal{H}} \\ &\leqslant C_{m,r}^{-1} \sum_{k=0}^{\infty} \frac{1}{k+1} C_{m,r}^{k+1} \|f\|_{\mathcal{H}}^{k+1} - \|f\|_{\mathcal{H}} \\ &= -C_{m,r}^{-1} \log\left(1 - C_{m,r}\|f\|_{\mathcal{H}}\right) - \|f\|_{\mathcal{H}}.\end{aligned}$$

Combining the above inequality and triangular inequality, it holds that

$$\begin{aligned}\|g_n\|_{\mathcal{H}} \geqslant \|f\|_{\mathcal{H}} - \|g_n - f\|_{\mathcal{H}} &\geqslant 2\|f\|_{\mathcal{H}} + C_{m,r}^{-1} \log\left(1 - C_{m,r}\|f\|_{\mathcal{H}}\right) \\ &\overset{\text{(i)}}{\geqslant} 2\|f\|_{\mathcal{H}} - \frac{3}{2}\|f\|_{\mathcal{H}} = \frac{1}{2}\|f\|_{\mathcal{H}},\end{aligned}$$

where (i) is is due to $\log(1-x) \geqslant -3x/2$ when $x \in [0, 1/3]$ and the fact that $C_{m,r}\|f\|_{\mathcal{H}} \leqslant 1/3$. Since $\|g_n - g\|_{\mathcal{H}} \to 0$ and $\|f\|_{\mathcal{H}} > 0$, we conclude that $\|g\|_{\mathcal{H}} \geqslant \|f\|_{\mathcal{H}}/4$. □

**Lemma 9.** *Assume $f, g \in \mathcal{H}$ such that $\|f\|_{\sup} \leqslant C$ and $\|g\|_{\sup} \leqslant C$, then it holds that*

$$e^{-2C} V(f-g, f-g) \leqslant V(e^f - e^g, e^f - e^g) \leqslant e^{2C} V(f-g, f-g).$$

*Proof.* By Taylor expansion, it holds that

$$|e^{f(\boldsymbol{x})} - e^{g(\boldsymbol{x})}| = e^{u} |f(\boldsymbol{x}) - g(\boldsymbol{x})| \leqslant e^{C} |f(\boldsymbol{x}) - g(\boldsymbol{x})|,$$

where $u = u(\boldsymbol{x})$ is a value between $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$. Therefore, we have

$$V(e^f - e^g, e^f - e^g) = \int_{\mathcal{X}} |e^{f(\boldsymbol{x})} - e^{g(\boldsymbol{x})}|^2 d\boldsymbol{x} \leqslant e^{2C} \int_{\mathcal{X}} |f(\boldsymbol{x}) - g(\boldsymbol{x})|^2 d\boldsymbol{x} = e^{2C} V(f-g, f-g),$$

which is the upper bound. The lower bound can be proved similarly. □

**Lemma 10.** *Suppose that $f \in \mathcal{H}$ satisfies $\int_{\mathcal{X}} f(\boldsymbol{x}) d\boldsymbol{x} = 0$. Let $w_f$ be a normalizing constant such that $\int_{\mathcal{X}} e^{f(\boldsymbol{x}) + w_f} d\boldsymbol{x} = 1$. There is a universal constant $B \in (0, 1]$ such that for all $f$ with $\|f\|_{\sup} \leqslant B$, the following statements hold:*

(i). $|w_f| \leqslant 2V(f, f)$;
(ii). $\left| e^{f(\boldsymbol{x}) + w_f} - 1 - \big(f(\boldsymbol{x}) + w_f\big) \right| \leqslant |f(\boldsymbol{x}) + w_f|^2$ *for all* $\boldsymbol{x} \in \mathcal{X}$.

*Proof.* Noting that

$$\lim_{x \to 0} \frac{e^x - 1 - x}{x^2} = \frac{1}{2}, \quad \lim_{x \to 0} \frac{\log(1+x)}{x} = 1, \quad \lim_{x \to 0} \frac{\log(1-x)}{x} = -1$$

there is a $B \in (0,1]$ such that

$$|e^x - 1 - x| \leqslant x^2, \quad \text{for all } |x| \leqslant 3B,$$
$$\frac{1}{2}x \leqslant \log(1+x) \leqslant 2x, \quad \text{for all } 0 \leqslant x \leqslant 3B,$$
$$-2x \leqslant \log(1-x) \leqslant -\frac{1}{2}x, \quad \text{for all } 0 \leqslant x \leqslant 3B.$$

Hence, if $\|f\|_{\sup} \leqslant B$, it holds that

$$|e^{-w_f} - 1| = \left|\int_{\mathcal{X}} e^{f(\boldsymbol{x})}dx - 1\right| = \left|\int_{\mathcal{X}} e^{f(\boldsymbol{x})}dx - \int_{\mathcal{X}} \big(1 + f(\boldsymbol{x})\big)\, d\boldsymbol{x}\right| \leqslant \int_{\mathcal{X}} f^2(\boldsymbol{x})d\boldsymbol{x} = V(f,f).$$

which further leads to $1 - V(f,f) \leqslant e^{-w_f} \leqslant 1 + V(f,f)$. Taking logarithm and using the fact that $V(f,f) \leqslant B^2 \leqslant B$, we see that

$$-2V(f,f) \leqslant -\log\big(1 + V(f,f)\big) \leqslant w_f \leqslant -\log\big(1 - V(f,f)\big) \leqslant 2V(f,f),$$

which is the first statement.

Moreover, the above inequality implies that

$$|f(\boldsymbol{x}) + w_f| \leqslant \|f\|_{\sup} + |w_f| \leqslant \delta + 2V(f,f) \leqslant B + 2B^2 \leqslant 3B.$$

Hence, we have

$$\left|e^{f(\boldsymbol{x})+w_f} - 1 = \big(f(\boldsymbol{x}) + w_f\big)\right| \leqslant |f(\boldsymbol{x}) + w_f|^2,$$

for all $\boldsymbol{x} \in \mathcal{X}$, which proves the second statement. $\square$

**Lemma 11.**

$$|\{\boldsymbol{i} \in \mathbb{I} : i_1 \ldots i_{r+d} \leqslant C\}| \asymp [\log(C)]^{r-1}C,$$
$$|\{\boldsymbol{i} \in \mathbb{I}_{cc} : i_1 \ldots i_{r+d} \leqslant C\}| \asymp [\log(C)]^{r-1}C,$$
$$|\{\boldsymbol{i} \in \mathbb{I}_{cd} : i_1 \ldots i_{r+d} \leqslant C\}| \asymp [\log(C)]^{r-1}C,$$
$$|\{\boldsymbol{i} \in \mathbb{I}_{dd} : i_1 \ldots i_{r+d} \leqslant C\}| \asymp [\log(C)]^{r-1}C,$$

*Proof.* This is xxx $\square$

**Lemma 12.** *For $(a_1, \ldots, a_r)^\top \in \mathbb{R}^r$, it follows that*

$$\int_{\substack{y_1 \ldots y_r \leqslant C \\ y_1, \ldots, y_r \geqslant 1}} y_1^{a_1} \ldots y_r^{a_r} dy_1 \ldots dy_r \asymp [\log(C)]^{N_{\max}-1} C^{a_{\max}+1}.$$

*Here $a_{\max} = \max_{1 \leqslant i \leqslant r} a_i$ and $N_{\max} = \sum_{i=1}^r I(a_i = a_{\max})$.*

*Proof.* Let $b_1 < b_2 < \ldots < b_p$ be the unique values among $a_1, \ldots, a_r$. For simplicity, we assume that $p = 3$. The proof of $p \neq 3$ can be done similarly. Due to the symmetry, we always can relabel the indexes so that $a_1 = \ldots = a_{s_1} = b_1$, $a_{s_1+1} = \ldots = a_{s_2} = b_2$, and $a_{s_2+1} = \ldots = a_{s_3} = b_3$, where $s_1$, $s_2 - s_1$, $s_3 - s_2$ are the numbers of $a_i$'s that equals $b_1, b_2$, and $b_3$, respectively. In particular, we have $a_{\max} = b_3$, $r = s_3$, $N_{\max} = s_3 - s_2$ when $p = 3$. Let $I$ be the desired integral, and direct examination leads to

$$\begin{aligned} I &= \int_{\substack{y_1 \ldots y_{s_3} \leqslant C \\ y_1, \ldots, y_r \geqslant 1}} y_1^{b_1} \ldots y_{s_1}^{b_1} y_{s_1+1}^{b_2} \ldots y_{s_2}^{b_2} y_{s_2+1}^{b_3} \ldots y_{s_3}^{b_3} \\ &\asymp \int_1^C \int_1^{z_{s_3}} \ldots \int_1^{z_2} z_1^{-1} \ldots z_{s_1-1}^{-1} z_{s_1}^{b_1-b_2-1} z_{s_1+1} \ldots z_{s_2-1}^{-1} z_{s_2}^{b_2-b_3-1} z_{s_2+1}^{-1} \ldots z_{s_3-1}^{-1} z_{s_3}^{b_3}. \end{aligned}$$

Using the fact that $b_i < b_{i+1}$, similar argument as in the proof of Lemma 7, we can show that

$$I \asymp \int_1^C [\log(z_{s_3})]^{s_3-s_2-1} z_{s_3}^{b_3} dz_{s_3} \asymp [\log(C)]^{s_3-s_2-1} C^{b_3+1}.$$

Since $N_{\max} = s_3 - s_2$ and $a_{\max} = b_3$, the result follows. $\square$

**Lemma 13.** *For $(a_1, \ldots, a_r)^\top \in (0, \infty)^r$ and $(b_1, \ldots, b_r)^\top \in \mathbb{R}^r$, we have*

$$\int_{\substack{x_1^{a_1} \ldots x_r^{a_r} \leqslant C \\ x_1, \ldots, x_r \geqslant 1}} x_1^{b_1} \ldots x_r^{b_r} dx_1 \ldots dx_r \asymp [\log(C)]^{N_* - 1} C^{\frac{b_* + 1}{a_*}}.$$

*Here $(a_*, b_*)$ satisfies $(b_* + 1)/a_* = \max_{1 \leqslant i \leqslant r}(b_i + 1)/a_i$ and $N_* = \sum_{i=1}^r I(a_i = a_*, b_i = b_*)$.*

*Proof.* Change of variable leads to

$$\int_{\substack{x_1^{a_1} \ldots x_r^{a_r} \leqslant C \\ x_1, \ldots, x_r \geqslant 1}} x_1^{b_1} \ldots x_r^{b_r} \asymp \int_{\substack{y_1 \ldots y_r \leqslant C \\ y_1, \ldots, y_r \geqslant 1}} y_1^{\frac{b_1+1}{a_1} - 1} \ldots y_1^{\frac{b_r+1}{a_r} - 1}.$$

Using Lemma 12, we complete the proof. $\square$

**Lemma 14.** *For any $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r)^\top \in \mathbb{B}_q$, let $a_i = (m - \beta_i)/m$ for $i = 1, \ldots, r$. Moreover, define $\mathbb{J} = \{\boldsymbol{i} \in \mathbb{I} : i_1^{a_1} \ldots i_r^{a_r} \leqslant C\}$. Then it follows that*

$$\sum_{\boldsymbol{i} \in \mathbb{J}} i_1^{2(m-\beta_1)} \ldots i_r^{2(m-\beta_r)} \asymp [\log(C)]^{N_{\max} \wedge q - 1} C^{2m + \frac{m}{m - \beta_{\max}}}.$$

*Here $\beta_{\max} = \max_{1 \leqslant i \leqslant r} \beta_i$ and $N_{\max} = \sum_{i=1}^r I(\beta_i = \beta_{\max})$.*

*Proof.* For any $A \subseteq \{1, \ldots, r\}$, we define $\mathbb{J}_A = \{\boldsymbol{i} \in \mathbb{J} : i_k > 1$ for all $k \in A$ and $i_k = 1$ for all $k \notin A\}$. By the definition, it follows that

$$\sum_{\boldsymbol{i} \in \mathbb{J}} i_1^{2(m-\beta_1)} \ldots i_r^{2(m-\beta_r)} \leqslant \sum_{A : |A| \leqslant q} \sum_{\boldsymbol{i} \in \mathbb{J}_A} i_1^{2(m-\beta_1)} \ldots i_r^{2(m-\beta_r)}.$$

For any $A = \{k_1, \ldots, k_s\}$ with $s = |A| \leqslant q$, it follows that

$$\begin{aligned}\sum_{\boldsymbol{i} \in \mathbb{J}_A} i_1^{2(m-\beta_1)} \ldots i_r^{2(m-\beta_r)} &= \sum_{\substack{i_{k_1}^{a_{k_1}} \ldots i_{k_s}^{a_{k_s}} \leqslant C \\ i_{k_1}, \ldots, i_{k_s} \geqslant 1}} i_{k_1}^{2(m-\beta_{k_1})} \ldots i_{k_s}^{2(m-\beta_{k_s})} \\ &\overset{\text{(i)}}{\lesssim} [\log(C)]^{N_{\max} \wedge q - 1} C^{2m + \frac{m}{m - \beta_{\max}}}.\end{aligned}$$

Here (i) is due to integration approximation and Lemma 13. Hence, we show that

$$\sum_{\boldsymbol{i} \in \mathbb{J}} i_1^{2(m-\beta_1)} \ldots i_r^{2(m-\beta_r)} \lesssim [\log(C)]^{N_{\max} \wedge q - 1} C^{2m + \frac{m}{m - \beta_{\max}}},$$

which is the upper bound.

To establish the lower bound, noting that $\boldsymbol{\beta} \in \mathbb{B}_q$, we may assume $\beta_1, \ldots, \beta_q \geqslant 0$ and $\beta_{q+1} = \ldots = \beta_r = 0$ for simplicity. Let $A_0 = \{1, \ldots, q\}$, then it follows that

$$\begin{aligned}\sum_{\boldsymbol{i} \in \mathbb{J}} i_1^{2(m-\beta_1)} \ldots i_r^{2(m-\beta_r)} &\gtrsim \sum_{\boldsymbol{i} \in \mathbb{J}_{A_0}} i_1^{2(m-\beta_1)} \ldots i_r^{2(m-\beta_r)} \\ &= \sum_{\substack{i_1^{a_1} \ldots i_q^{a_q} \leqslant C \\ i_1, \ldots, i_q \geqslant 1}} i_1^{2(m-\beta_1)} \ldots i_q^{2(m-\beta_q)} \\ &\overset{\text{(i)}}{\asymp} [\log(C)]^{N_{\max} \wedge q - 1} C^{2m + \frac{m}{m - \beta_{\max}}},\end{aligned}$$

where (i) is due to Lemma 13 and integration approximation. Hence, we prove the lower bound. Finally, the upper bound and lower bound together lead to the desired result. $\square$

**Lemma 15.** *Fro any $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r)^\top \in \mathbb{B}_q$, let $a_i = (m - \beta_i)/m$ for $i = 1, \ldots, r$. Moreover, define $\mathbb{J} = \{\boldsymbol{i} \in \mathbb{I} : i_1^{a_1} \ldots i_r^{a_r} \leqslant C\}$. Then it follows that*

$$\sum_{\boldsymbol{i} \in \mathbb{J}} i_1^{-2\beta_1} \ldots i_r^{-2\beta_r} \asymp [\log(C)]^{N_{\min} \wedge q - 1} C^{\frac{m(1 - 2\beta_{\min})}{(m - \beta_{\min})}}.$$

*Here $\beta_{\min} = \min_{1 \leqslant i \leqslant r} \beta_i$ and $N_{\min} = \sum_{i=1}^r I(\beta_i = \beta_{\min})$.*

*Proof.* The proof is similar to that of Lemma 14. Hence, we omit it. $\square$

**Theorem 2.** *Let $\Omega = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leqslant 1, \int_{\mathcal{X}} e^{f(\boldsymbol{x})} d\boldsymbol{x} = 1\}$, there is a constant $C > 0$ free of $n$ such that*

$$\inf_{\widehat{f}} \sup_{f \in \Omega} \mathbb{E}_f \left\{ \int_{\mathcal{X}} \left| \widehat{f}(\boldsymbol{x}) - f(\boldsymbol{x}) \right|^2 d\boldsymbol{x} \right\} \geqslant C \left( \frac{n}{[\log(n)]^{q-1}} \right)^{-\frac{2m}{2m+1}}.$$

*Here the infimum is taking over all estimators based on $n$ i.i.d. observations, and $\mathbb{E}_f$ is to indicate that the expectation is with respect to observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ generated from the density $p_f = e^f$.*

***Proof of Theorem 2.*** Let us define

$$\mathbb{J} = \left\{ \boldsymbol{i} \in \mathbb{I}_r : i_1^a \dots i_q^a \leqslant N, i_1, \dots, i_q \geqslant 1, i_{q+1} = \dots = i_r = 1, \min\{i_1, \dots, i_q\} > 1 \right\}$$

and $d = \sum_{\boldsymbol{i} \in \mathbb{J}} i_1^{2m} \dots i_q^{2m}$. Here $N \asymp \left(n[\log(n)]^{1-q}\right)^{1/(2m+1)}$ is an integer. By Lemmas 11 and 14, we have

$$|\mathbb{J}| \asymp [\log(N)]^{q-1} N, \quad d \asymp [\log(N)]^{q-1} N^{2m+1}. \tag{5.1}$$

For any binary sequence $\boldsymbol{b} = \{b_{\boldsymbol{i}} : \boldsymbol{i} \in \mathbb{J}\} \in \{0,1\}^{|\mathbb{J}|}$ and constant $c > 0$ small enough, let us define

$$p_{\boldsymbol{b}}(\boldsymbol{x}) = \frac{c}{\sqrt{d}} \sum_{\boldsymbol{i} \in \mathbb{J}} b_{\boldsymbol{i}} \psi_{\boldsymbol{i}}(\boldsymbol{x}) + 1, \quad f_{\boldsymbol{b}}(\boldsymbol{x}) = \log\left(p_{\boldsymbol{b}}(\boldsymbol{x})\right),$$

which corresponds to density and log density. It can be verified that $\int_{\mathcal{X}} p_{\boldsymbol{b}}(\boldsymbol{x}) d\boldsymbol{x} = 1$. Moreover, for all $\boldsymbol{b} \in \{0,1\}^{|\mathbb{J}|}$, it follows that

$$\|p_{\boldsymbol{b}} - 1\|_{\mathcal{H}}^2 \asymp \frac{c^2}{d} \sum_{\boldsymbol{i} \in \mathbb{J}} b_{\boldsymbol{i}}^2 \rho_{\boldsymbol{i}}^{-1} \lesssim \frac{c^2}{d} \sum_{\boldsymbol{i} \in \mathbb{J}} i_1^{2m} \dots i_q^{2m} \overset{\text{(i)}}{=} c^2,$$

where (i) comes from the definition of $d$. Hence, if $c > 0$ is small enough, we have $\|p_{\boldsymbol{b}} - 1\|_{\sup} \leqslant 1/2$, which further leads to

$$1/2 \leqslant p_{\boldsymbol{b}}(\boldsymbol{x}) \leqslant 2, \quad \text{for all } \boldsymbol{x} \in \mathcal{X}. \tag{5.2}$$

By Lemma 8, we can choose $c > 0$ small enough such that

$$f_{\boldsymbol{b}} = \log(p_{\boldsymbol{b}}) = \log\left(1 + (p_{\boldsymbol{b}} - 1)\right) \in \mathcal{H}, \text{ for all } \boldsymbol{b} \in \{0,1\}^{|\mathbb{J}|}.$$

Furthermore, Varshamov-Gilbert bound (Lemma 2.9 in Tsybakov, 2008) implies that there is a collection $\mathcal{B} \subseteq \{0,1\}^{|\mathbb{J}|}$ such that $\boldsymbol{b}_0 = (0, 0, \dots, 0) \in \mathcal{B}$, $|\mathcal{B}| \geqslant 2^{|\mathbb{J}|/8}$ and $\sum_{\boldsymbol{i} \in \mathbb{J}} (b_{\boldsymbol{i}} - \widetilde{b}_{\boldsymbol{i}})^2 \geqslant |\mathbb{J}|/8$ for any different $\boldsymbol{b}, \widetilde{\boldsymbol{b}} \in \mathcal{B}$. By Taylor's theorem, we see that

$$\begin{aligned}
|f_{\boldsymbol{b}}(\boldsymbol{x}) - f_{\widetilde{\boldsymbol{b}}}(\boldsymbol{x})| &= \left| \log\left(p_{\boldsymbol{b}}(\boldsymbol{x})\right) - \log\left(p_{\widetilde{\boldsymbol{b}}}(\boldsymbol{x})\right) \right| \\
&= \frac{1}{\left| s p_{\boldsymbol{b}}(\boldsymbol{x}) + (1-s) p_{\widetilde{\boldsymbol{b}}}(\boldsymbol{x}) \right|} \left| p_{\boldsymbol{b}}(\boldsymbol{x}) - p_{\widetilde{\boldsymbol{b}}}(\boldsymbol{x}) \right| \overset{\text{(I)}}{\geqslant} \frac{1}{2} \left| p_{\boldsymbol{b}}(\boldsymbol{x}) - p_{\widetilde{\boldsymbol{b}}}(\boldsymbol{x}) \right|,
\end{aligned}$$

where $s \in [0,1]$, and (i) is due to (5.2). Hence, it follows that

$$\begin{aligned}
\int_{\mathcal{X}} \left| f_{\boldsymbol{b}}(\boldsymbol{x}) - f_{\widetilde{\boldsymbol{b}}}(\boldsymbol{x}) \right|^2 d\boldsymbol{x} &\geqslant \frac{1}{4} \int_{\mathcal{X}} \left| p_{\boldsymbol{b}}(\boldsymbol{x}) - p_{\widetilde{\boldsymbol{b}}}(\boldsymbol{x}) \right|^2 d\boldsymbol{x} \\
&= \frac{c^2}{4d} \int_{\mathcal{X}} \left| \sum_{\boldsymbol{i} \in \mathbb{J}} (b_{\boldsymbol{i}} - \widetilde{b}_{\boldsymbol{i}}) \psi_{\boldsymbol{i}}(\boldsymbol{x}) \right|^2 d\boldsymbol{x} \\
&= \frac{c^2}{4d} \sum_{\boldsymbol{i} \in \mathbb{J}} (b_{\boldsymbol{i}} - \widetilde{b}_{\boldsymbol{i}})^2 \gtrsim \frac{|\mathbb{J}|}{d} \overset{\text{(i)}}{\asymp} N^{-2m},
\end{aligned} \tag{5.3}$$

for all different $\boldsymbol{b}, \widetilde{\boldsymbol{b}} \in \mathcal{B}$. Here (i) comes from (5.1).

Similarly, we can show that

$$
\begin{aligned}
KL(p_{\boldsymbol{b}}, p_{\boldsymbol{b}_0}) &= \int_{\mathcal{X}} p_{\boldsymbol{b}}(\boldsymbol{x}) \log\left(\frac{p_{\boldsymbol{b}}(\boldsymbol{x})}{p_{\boldsymbol{b}_0}(\boldsymbol{x})}\right) d\boldsymbol{x} \\
&= \int_{\mathcal{X}} p_{\boldsymbol{b}}(\boldsymbol{x}) \log\big(p_{\boldsymbol{b}}(\boldsymbol{x})\big)\, d\boldsymbol{x} \\
&\leqslant \int_{\mathcal{X}} p_{\boldsymbol{b}}(\boldsymbol{x}) \big(p_{\boldsymbol{b}}(\boldsymbol{x}) - 1\big)\, d\boldsymbol{x} \\
&= \int_{\mathcal{X}} \big(p_{\boldsymbol{b}}(\boldsymbol{x}) - 1\big)^2\, d\boldsymbol{x} \\
&= \frac{c^2}{d} \sum_{\boldsymbol{i} \in \mathbb{J}} b_{\boldsymbol{i}} \leqslant \frac{c^2}{d} |\mathbb{J}| \overset{\text{(i)}}{\asymp} \frac{c^2}{d} [\log(N)]^{q-1} N \asymp N^{-2m}.
\end{aligned} \tag{5.4}
$$

Here (i) comes from (5.1). By the choice of $N$, we have

$$
KL(p_{\boldsymbol{b}}, p_{\boldsymbol{b}_0}) \lesssim N^{-2m} \asymp \frac{[\log(N)]^{q-1} N}{n} \asymp \frac{\log(|\mathcal{B}|)}{n} \asymp \frac{|\mathbb{J}|}{n}. \tag{5.5}
$$

Combining Fano's Lemma (Lemma 2.10 in Tsybakov, 2008) with (5.3)-(5.5), we show that the lower bound is $N^{-2m}$, which completes the proof after substituting the value of $N$. ☐

**Theorem 3.** *Let $\Omega = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leqslant 1, \int_{\mathcal{X}} e^{f(\boldsymbol{x})} d\boldsymbol{x} = 1\}$, there is a constant $C > 0$ free of $n$ such that*

$$
\inf_{\widehat{f}} \sup_{f \in \Omega} \mathbb{E}_f \left\{ \int_{\mathcal{X}} \left| \partial^{\boldsymbol{\beta}} \widehat{f}(\boldsymbol{x}) - \partial^{\boldsymbol{\beta}} f(\boldsymbol{x}) \right|^2 d\boldsymbol{x} \right\} \geqslant C \left( \frac{n}{[\log(n)]^{q-1}} \right)^{-\frac{2(m-\beta)}{2m+1}}.
$$

*Here the infimum is taking over all estimators based on $n$ i.i.d. observations, and $\mathbb{E}_f$ is to indicate that the expectation is with respect to observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ generated from the density $p_f = e^f$.*

***Proof of Theorem 3.*** Since $\boldsymbol{\beta} \in \mathbb{B}_q$ with $\beta_1, \ldots, \beta_r \in \{0, \beta\}$ and $\sum_{i=1}^r I(\beta_i > 0) = q$, by symmetry, we can assume $\beta_1 = \ldots = \beta_q = \beta$ and $\beta_{q+1} = \ldots = \beta_r = 0$.

Let us define $a = (m-\beta)/m$, $d = \sum_{\boldsymbol{i} \in \mathbb{J}} i_1^{2(m-\beta)} \ldots i_q^{2(m-\beta)}$, and

$$
\mathbb{J} = \{\boldsymbol{i} \in \mathbb{I}_r : i_1^a \ldots i_q^a \leqslant N, i_1, \ldots, i_q \geqslant 1, i_{q+1} = \ldots = i_r = 1, \boldsymbol{i} \neq (1, 1, \ldots, 1)\}.
$$

Here $N \asymp \big(n[\log(n)]^{1-q}\big)^{(m-\beta)/(2m^2+m)}$ is an integer. By Lemmas 11 and 14, we have

$$
|\mathbb{J}| \asymp [\log(N)]^{q-1} N^{\frac{m}{m-\beta}}, \quad d \asymp [\log(N)]^{q-1} N^{2m + \frac{m}{m-\beta}}. \tag{5.6}
$$

For any binary sequence $\boldsymbol{b} = \{b_{\boldsymbol{i}} : \boldsymbol{i} \in \mathbb{J}\} \in \{0, 1\}^{|\mathbb{J}|}$ and constant $c > 0$ small enough, let us define

$$
f_{\boldsymbol{b}}(\boldsymbol{x}) = \frac{c}{\sqrt{d}} \sum_{\boldsymbol{i} \in \mathbb{J}} b_{\boldsymbol{i}} i_1^{-\beta} \ldots i_q^{-\beta} \psi_{\boldsymbol{i}}(\boldsymbol{x}), \quad p_{\boldsymbol{b}}(\boldsymbol{x}) = e^{f_{\boldsymbol{b}}(\boldsymbol{x}) + w_{\boldsymbol{b}}},
$$

where $w_{\boldsymbol{b}}$ is a normalizing constant such that $\int_{\mathcal{X}} e^{f_{\boldsymbol{b}}(\boldsymbol{x}) + w_{\boldsymbol{b}}} d\boldsymbol{x} = 1$.

It can be verified that $\int_{\mathcal{X}} p_{\boldsymbol{b}}(\boldsymbol{x})d\boldsymbol{x} = 1$. Moreover, for all $\boldsymbol{b} \in \{0,1\}^{|\mathbb{J}|}$, it follows that

$$\|f_{\boldsymbol{b}}\|_{\sup}^2 \overset{\text{(i)}}{\lesssim} \|f_{\boldsymbol{b}}\|_{\mathcal{H}}^2 \asymp \frac{c^2}{d}\sum_{\boldsymbol{i}\in\mathbb{J}} b_{\boldsymbol{i}}^2 i_1^{-2\beta}\dots i_q^{-2\beta}\rho_{\boldsymbol{i}}^{-1} \lesssim \frac{c^2}{d}\sum_{\boldsymbol{i}\in\mathbb{J}} i_1^{2(m-\beta)}\dots i_q^{2(m-\beta)} \overset{\text{(ii)}}{=} c^2,$$

where (i) is due to Lemma 2(iii), and (ii) comes from the definition of $d$. Hence, if $c > 0$ is small enough, Lemma 10 implies the following statements:

$$|w_{\boldsymbol{b}}| \leqslant 2V(f_{\boldsymbol{b}}, f_{\boldsymbol{b}}), \quad \left|e^{f_{\boldsymbol{b}}(\boldsymbol{x})+w_{\boldsymbol{b}}} - 1 - \big(f_{\boldsymbol{b}}(\boldsymbol{x}) + w_{\boldsymbol{b}}\big)\right| \leqslant |f_{\boldsymbol{b}}(\boldsymbol{x}) + w_{\boldsymbol{b}}|^2 \text{ for all } \boldsymbol{x} \in \mathcal{X}. \tag{5.7}$$

Furthermore, Varshamov-Gilbert bound (Lemma 2.9 in Tsybakov, 2008) implies that there is a collection $\mathcal{B} \subseteq \{0,1\}^{|\mathbb{J}|}$ such that $\boldsymbol{b}_0 = (0,0,\dots,0) \in \mathcal{B}$, $|\mathcal{B}| \geqslant 2^{|\mathbb{J}|/8}$ and $\sum_{\boldsymbol{i}\in\mathbb{J}}(b_{\boldsymbol{i}} - \widetilde{b}_{\boldsymbol{i}})^2 \geqslant |\mathbb{J}|/8$ for any different $\boldsymbol{b}, \widetilde{\boldsymbol{b}} \in \mathcal{B}$. Hence, we have

$$\begin{aligned}\int_{\mathcal{X}} \left|\partial^{\boldsymbol{\beta}} f_{\boldsymbol{b}}(\boldsymbol{x}) - \partial^{\boldsymbol{\beta}} f_{\widetilde{\boldsymbol{b}}}(\boldsymbol{x})\right|^2 d\boldsymbol{x} &= \frac{c^2}{d}\int_{\mathcal{X}} \left|\sum_{\boldsymbol{i}\in\mathbb{J}}(b_{\boldsymbol{i}} - \widetilde{b}_{\boldsymbol{i}}) i_1^{-\beta}\dots i_q^{-\beta}\partial^{\boldsymbol{\beta}}\psi_{\boldsymbol{i}}(\boldsymbol{x})\right|^2 d\boldsymbol{x} \\ &\overset{\text{(i)}}{\asymp} \frac{c^2}{d}\sum_{\boldsymbol{i}\in\mathbb{J}}(b_{\boldsymbol{i}} - \widetilde{b}_{\boldsymbol{i}})^2 \gtrsim \frac{|\mathbb{J}|}{d} \overset{\text{(ii)}}{\asymp} N^{-2m},\end{aligned} \tag{5.8}$$

for all different $\boldsymbol{b}, \widetilde{\boldsymbol{b}} \in \mathcal{B}$. Here (i) is due to xxx, and (ii) comes from (5.6).

By direct examination, we have

$$\begin{aligned}KL(p_{\boldsymbol{b}}, p_{\boldsymbol{b}_0}) &= \int_{\mathcal{X}} p_{\boldsymbol{b}}(\boldsymbol{x}) \log\left(\frac{p_{\boldsymbol{b}}(\boldsymbol{x})}{p_{\boldsymbol{b}_0}(\boldsymbol{x})}\right) d\boldsymbol{x} \\ &= \int_{\mathcal{X}} e^{f_{\boldsymbol{b}}(\boldsymbol{x})+w_{\boldsymbol{b}}}\big(f_{\boldsymbol{b}}(\boldsymbol{x}) + w_{\boldsymbol{b}}\big)\, d\boldsymbol{x} = A_{\boldsymbol{b}} + B_{\boldsymbol{b}},\end{aligned}$$

where

$$\begin{aligned}A_{\boldsymbol{b}} &= \int_{\mathcal{X}} \left\{e^{f_{\boldsymbol{b}}(\boldsymbol{x})+w_{\boldsymbol{b}}} - 1 - \big(f_{\boldsymbol{b}}(\boldsymbol{x}) + w_{\boldsymbol{b}}\big)\right\}\big(f_{\boldsymbol{b}}(\boldsymbol{x}) + w_{\boldsymbol{b}}\big)\, d\boldsymbol{x}, \\ B_{\boldsymbol{b}} &= \int_{\mathcal{X}} \left\{1 + \big(f_{\boldsymbol{b}}(\boldsymbol{x}) + w_{\boldsymbol{b}}\big)\right\}\big(f_{\boldsymbol{b}}(\boldsymbol{x}) + w_{\boldsymbol{b}}\big)\, d\boldsymbol{x}.\end{aligned}$$

Using (5.7), it holds that

$$\begin{aligned}|A_{\boldsymbol{b}}| &\leqslant \int_{\mathcal{X}} \left|e^{f_{\boldsymbol{b}}(\boldsymbol{x})+w_{\boldsymbol{b}}} - 1 - \big(f_{\boldsymbol{b}}(\boldsymbol{x}) + w_{\boldsymbol{b}}\big)\right| \left|f_{\boldsymbol{b}}(\boldsymbol{x}) + w_{\boldsymbol{b}}\right| d\boldsymbol{x}, \\ &\overset{\text{(i)}}{\leqslant} \big(\|f_{\boldsymbol{b}}\|_{\sup} + |w_{\boldsymbol{b}}|\big)\left\{\int_{\mathcal{X}} \big(f_{\boldsymbol{b}}(\boldsymbol{x}) + w_{\boldsymbol{b}}\big)^2 d\boldsymbol{x}\right\} \\ &\overset{\text{(ii)}}{\lesssim} \int_{\mathcal{X}} f_{\boldsymbol{b}}^2(\boldsymbol{x})d\boldsymbol{x} + w_{\boldsymbol{b}}^2 \lesssim \int_{\mathcal{X}} f_{\boldsymbol{b}}^2(\boldsymbol{x})d\boldsymbol{x},\end{aligned}$$

where (i), (ii) and (iii) are to (5.7) and the fact that $V(f_{\boldsymbol{b}}, f_{\boldsymbol{b}}) \leqslant \|f_{\boldsymbol{b}}\|_{\mathcal{H}}^2 \lesssim c^2$. Using similar arguments, we can show that

$$|B_{\boldsymbol{b}}| \overset{\text{(i)}}{=} \left|w_{\boldsymbol{b}} + \int_{\mathcal{X}} \big(f_{\boldsymbol{b}}(\boldsymbol{x}) + w_{\boldsymbol{b}}\big)^2 d\boldsymbol{x}\right| \lesssim \int_{\mathcal{X}} f_{\boldsymbol{b}}^2(\boldsymbol{x})d\boldsymbol{x},$$

where (i) uses the fact that $\int_{\mathcal{X}} f_{\boldsymbol{b}}(\boldsymbol{x})dx = 0$. Combining the above three inequalities, we conclude that

$$
\begin{aligned}
KL(p_{\boldsymbol{b}}, p_{\boldsymbol{b}_0}) &\lesssim \int_{\mathcal{X}} f_{\boldsymbol{b}}^2(\boldsymbol{x})d\boldsymbol{x} \\
&= \frac{c^2}{d}\int_{\mathcal{X}} \left| \sum_{\boldsymbol{i}\in\mathbb{J}} b_{\boldsymbol{i}} i_1^{-\beta}\dots i_q^{-\beta}\psi_{\boldsymbol{i}}(\boldsymbol{x}) \right|^2 d\boldsymbol{x} \\
&= \frac{c^2}{d}\sum_{\boldsymbol{i}\in\mathbb{J}} b_{\boldsymbol{i}} i_1^{-2\beta}\dots i_q^{-2\beta} \\
&\leqslant \frac{c^2}{d}\sum_{\boldsymbol{i}\in\mathbb{J}} i_1^{-2\beta}\dots i_q^{-2\beta} \overset{\text{(i)}}{\asymp} \frac{c^2}{d}[\log(N)]^{q-1}N^{\frac{(1-2\beta)m}{m-\beta}} \overset{\text{(ii)}}{\asymp} N^{-2m-\frac{2m\beta}{m-\beta}}.
\end{aligned} \tag{5.9}
$$

Here (i) is due to Lemma 15, and (ii) follows from the definition of $d$. By the choice of $N$, we have

$$
KL(p_{\boldsymbol{b}}, p_{\boldsymbol{b}_0}) \lesssim N^{-2m-\frac{2m\beta}{m-\beta}} \lesssim \frac{[\log(N)]^{q-1}N^{\frac{m}{m-\beta}}}{n} \asymp \frac{\log(|\mathcal{B}|)}{n} \asymp \frac{|\mathbb{J}|}{n}. \tag{5.10}
$$

Combining Fano's Lemma (Lemma 2.10 in Tsybakov, 2008) with (5.8)-(5.10), we show that the lower bound is $N^{-2m}$, which completes the proof after substituting the value of $N$. □

## 6. Density Estimation

$$
L_{n,\lambda}(f) = -\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{X}_i) + \int_{\mathcal{X}} e^{f(\boldsymbol{x})}d\boldsymbol{x} + \lambda\|f\|_{\mathcal{H}}^2.
$$

$$
\begin{aligned}
DL^*_{n,\lambda}(f)g &= -\frac{1}{n}\sum_{i=1}^{n} W_i g(\boldsymbol{X}_i) + \int_{\mathcal{X}} e^{f(\boldsymbol{x})} g(\boldsymbol{x}) d\boldsymbol{x} + \langle \mathcal{W}_\lambda f, g\rangle \\
&= \langle -\frac{1}{n}\sum_{i=1}^{n} W_i K_{\boldsymbol{X}_i} + u_f + \mathcal{W}_\lambda f, g\rangle \\
&\stackrel{\text{def}}{=} \langle S_{n,\lambda}(f), g\rangle, \\
D^2 L^*_{n,\lambda}(f) g_1 g_2 &= \int_{\mathcal{X}} e^{f(\boldsymbol{x})} g_1(\boldsymbol{x}) g_2(\boldsymbol{x}) d\boldsymbol{x} + \langle \mathcal{W}_\lambda g_1, g_2\rangle, \\
S_{n,\lambda}(f) &= -\frac{1}{n}\sum_{i=1}^{n} W_i K_{\boldsymbol{X}_i} + u_f + \mathcal{W}_\lambda f, \text{ where } \langle u_f, g\rangle = \int_{\mathcal{X}} e^{f(\boldsymbol{x})} g(\boldsymbol{x}) d\boldsymbol{x}, \\
DS_{n,\lambda}(f) g_1 g_2 &= \int_{\mathcal{X}} e^{f(\boldsymbol{x})} g_1(\boldsymbol{x}) g_2(\boldsymbol{x}) d\boldsymbol{x} + \langle \mathcal{W}_\lambda g_1, g_2\rangle, \\
S_\lambda &= \mathbb{E}(S_{n,\lambda}), \\
\langle S_\lambda(f), g\rangle &= -\mathbb{E}\{g(\boldsymbol{X})\} + \langle h_f + \mathcal{W}_\lambda f, g\rangle, \\
\langle S_\lambda(f_0), g\rangle &= -\mathbb{E}\{g(\boldsymbol{X})\} + \langle h_{f_0} + \mathcal{W}_\lambda f_0, g\rangle \\
&= -\mathbb{E}\{g(\boldsymbol{X})\} + \int_{\mathcal{X}} e^{f_0(\boldsymbol{x})} g(\boldsymbol{x}) d\boldsymbol{x} + \langle \mathcal{W}_\lambda f_0, g\rangle \\
&= \langle \mathcal{W}_\lambda f_0, g\rangle, \\
S_\lambda(f_0) &= \mathcal{W}_\lambda f_0.
\end{aligned}
$$

$$
\frac{\|S_{n,\lambda}(f+g) - S_{n,\lambda}(f) - h\|}{\|g\|} =
$$

$$
\begin{aligned}
\|S_{n,\lambda}(f+g) - S_{n,\lambda}(f) - Ag\| &= \sup_{\|u\|=1} \langle S_{n,\lambda}(f+g) - S_{n,\lambda}(f) - Ag, u\rangle \\
&= \sup_{\|u\|=1} \langle h_{f+g} - h_f + \mathcal{W}_\lambda g - Ag, u\rangle
\end{aligned}
$$

$$
\langle Bg, u\rangle = \int_{\mathcal{X}} e^{f(\boldsymbol{x})} g(\boldsymbol{x}) u(\boldsymbol{x}) dx
$$

$$
\begin{aligned}
\left|\langle h_{f+g} - h_f - Bg, u\rangle\right| &= \left|\int_{\mathcal{X}} \left(e^{f(\boldsymbol{x})+g(\boldsymbol{x})} - e^{f(\boldsymbol{x})} - e^{f(\boldsymbol{x})}g(\boldsymbol{x})\right) u(\boldsymbol{x})d\boldsymbol{x}\right| \\
&\leqslant \sqrt{\int_{\mathcal{X}} \left(e^{f(\boldsymbol{x})+g(\boldsymbol{x})} - e^{f(\boldsymbol{x})} - e^{f(\boldsymbol{x})}g(\boldsymbol{x})\right)^2 d\boldsymbol{x}}\sqrt{\int_{\mathcal{X}} u^2(\boldsymbol{x})d\boldsymbol{x}} \\
&\leqslant \sqrt{\int_{\mathcal{X}} \left(e^{f(\boldsymbol{x})+g(\boldsymbol{x})} - e^{f(\boldsymbol{x})} - e^{f(\boldsymbol{x})}g(\boldsymbol{x})\right)^2 d\boldsymbol{x}} \times \|u\| \\
&\leqslant \sqrt{\int_{\mathcal{X}} \left(e^{f(\boldsymbol{x})+g(\boldsymbol{x})} - e^{f(\boldsymbol{x})} - e^{f(\boldsymbol{x})}g(\boldsymbol{x})\right)^2 d\boldsymbol{x}}.
\end{aligned}
$$

Since $\|g\| \to 0$ implies $\|g\|_{\sup} \to 0$, it holds that

$$
\lim_{\|g\|\to 0} \sup_{\boldsymbol{x}\in\mathcal{X}} \left|\frac{e^{g(\boldsymbol{x})} - 1 - g(\boldsymbol{x})}{g(\boldsymbol{x})}\right| = 0.
$$

Hence, we show that

$$
\begin{aligned}
\int_{\mathcal{X}} \left(e^{f(\boldsymbol{x})+g(\boldsymbol{x})} - e^{f(\boldsymbol{x})} - e^{f(\boldsymbol{x})}g(\boldsymbol{x})\right)^2 d\boldsymbol{x} &= \int_{\mathcal{X}} e^{2f(\boldsymbol{x})} \left(\frac{e^{g(\boldsymbol{x})} - 1 - g(\boldsymbol{x})}{g(\boldsymbol{x})}\right)^2 g^2(\boldsymbol{x})d\boldsymbol{x} \\
&\leqslant e^{2\|f\|_{\sup}} \sup_{\boldsymbol{x}\in\mathcal{X}} \left|\frac{e^{g(\boldsymbol{x})} - 1 - g(\boldsymbol{x})}{g(\boldsymbol{x})}\right| \int_{\mathcal{X}} g^2(\boldsymbol{x})d\boldsymbol{x} \\
&\leqslant e^{2\|f\|_{\sup}} \sup_{\boldsymbol{x}\in\mathcal{X}} \left|\frac{e^{g(\boldsymbol{x})} - 1 - g(\boldsymbol{x})}{g(\boldsymbol{x})}\right| \|g\|^2.
\end{aligned}
$$

By xxx, we see that

$$
\lim_{\|g\|\to 0} \sup_{\|u\|=1} \frac{\left|\langle h_{f+g} - h_f - Bg, u\rangle\right|}{\|g\|} \leqslant e^{\|f\|_{\sup}} \lim_{\|g\|\to 0} \sqrt{\sup_{\boldsymbol{x}\in\mathcal{X}} \left|\frac{e^{g(\boldsymbol{x})} - 1 - g(\boldsymbol{x})}{g(\boldsymbol{x})}\right|} = 0.
$$

We show that

$$
\begin{aligned}
DS_{n,\lambda}(f)g_1g_2 &= \int_{\mathcal{X}} e^{f(\boldsymbol{x})} g_1(\boldsymbol{x})g_2(\boldsymbol{x})d\boldsymbol{x} + \langle \mathcal{W}_\lambda g_1, g_2\rangle, \\
DS_\lambda(f)g_1g_2 &= DS_{n,\lambda}(f)g_1g_2.
\end{aligned}
$$

**Lemma 16.** *Suppose that* $\lim_{n\to\infty} \|g_n\|_{\sup} = 0$ *and* $\|f\|_{\sup} < \infty$*, then it holds that*

$$
\left|\int_{\mathcal{X}} e^{f(\boldsymbol{x})} \left(e^{g_n(\boldsymbol{x})} - 1 - g_n(\boldsymbol{x})\right) d\boldsymbol{x} - \frac{1}{2}\int_{\mathcal{X}} e^{f(\boldsymbol{x})} g_n^2(\boldsymbol{x})d\boldsymbol{x}\right| \leqslant c_n \int_{\mathcal{X}} e^{f(\boldsymbol{x})} g_n^2(\boldsymbol{x})d\boldsymbol{x},
$$

*where*

$$
c_n = \sup_{\boldsymbol{x}:g_n(\boldsymbol{x})\neq 0} \left|\frac{e^{g_n(\boldsymbol{x})} - 1 - g_n(\boldsymbol{x}) - \frac{1}{2}g_n^2(\boldsymbol{x})}{g_n^2(\boldsymbol{x})}\right| \to 0.
$$

*Proof.* Since $\|g_n\|_{\sup} \to 0$, L'Hopital's rule implies that

$$\lim_{n\to\infty} \sup_{\boldsymbol{x}: g_n(\boldsymbol{x})\neq 0} \left| \frac{e^{g_n(\boldsymbol{x})} - 1 - g_n(\boldsymbol{x}) - \frac{1}{2}g_n^2(\boldsymbol{x})}{g_n^2(\boldsymbol{x})} \right| = 0.$$

Hence, it holds that

$$\begin{aligned}
&\int_{\mathcal{X}} e^{f(\boldsymbol{x})} \left( e^{g_n(\boldsymbol{x})} - 1 - g_n(\boldsymbol{x}) - \frac{1}{2}g_n^2(\boldsymbol{x}) \right) d\boldsymbol{x} \\
\leqslant &\int_{\mathcal{X}} e^{f(\boldsymbol{x})} \left| e^{g_n(\boldsymbol{x})} - 1 - g_n(\boldsymbol{x}) - \frac{1}{2}g_n^2(\boldsymbol{x}) \right| d\boldsymbol{x} \\
= &\int_{\boldsymbol{x}: g_n(\boldsymbol{x})\neq 0} e^{f(\boldsymbol{x})} \left| \frac{e^{g_n(\boldsymbol{x})} - 1 - g_n(\boldsymbol{x}) - \frac{1}{2}g_n^2(\boldsymbol{x})}{g_n^2(\boldsymbol{x})} \right| g_n^2(\boldsymbol{x}) d\boldsymbol{x} \\
\leqslant &\sup_{\boldsymbol{x}: g_n(\boldsymbol{x})\neq 0} \left| \frac{e^{g_n(\boldsymbol{x})} - 1 - g_n(\boldsymbol{x}) - \frac{1}{2}g_n^2(\boldsymbol{x})}{g_n^2(\boldsymbol{x})} \right| \int_{\mathcal{X}} e^{f(\boldsymbol{x})} g_n^2(\boldsymbol{x}) d\boldsymbol{x}.
\end{aligned}$$

□

**Lemma 17.** *Suppose that* $\lim_{n\to\infty} \|g_n\|_{\sup} = 0$ *and* $\|f\|_{\sup} < \infty$*, then it holds that*

$$\left| \int_{\mathcal{X}} e^{f(\boldsymbol{x})} \left( e^{g_n(\boldsymbol{x})} - 1 \right) g_n(\boldsymbol{x}) d\boldsymbol{x} - \int_{\mathcal{X}} e^{f(\boldsymbol{x})} g_n^2(\boldsymbol{x}) d\boldsymbol{x} \right| \leqslant c_n \int_{\mathcal{X}} e^{f(\boldsymbol{x})} g_n^2(\boldsymbol{x}) d\boldsymbol{x},$$

*where*

$$c_n = \sup_{\boldsymbol{x}: g_n(\boldsymbol{x})\neq 0} \left| \frac{e^{g_n(\boldsymbol{x})} - 1 - g_n(\boldsymbol{x})}{g_n(\boldsymbol{x})} \right| \to 0.$$

*Proof.* The proof is similar to that of Lemma 16, and we omit it. □

**Theorem 4.** *Under xxx, if* $nh^2 \to \infty$ *and* $\lambda \to 0$*, then it holds that*

$$\|\widehat{f} - f_0\|^2 = O_P\left( \lambda + \frac{1}{nh} \right).$$

*Proof.* The result will be proved by contradiction. Let us assume that for some $\delta, B_\delta > 0$, it holds that $\mathbb{P}(E_{n,B}) \geqslant \delta$ for all $B \geqslant B_\delta$. Here $E_{n,B} = \{\|\widehat{f} - f_0\| \geqslant B(\kappa_n + \lambda^{1/2})\}$ is an event.

On event $E_{n,B}$, the definition of $\widehat{f}$ implies that

$$\inf_{f: \|f - f_0\| \geqslant B(\kappa_n + \lambda^{1/2})} L_{n,\lambda}(f) - L_{n,\lambda}(f_0) < 0.$$

By convexity of $f \to L_{n,\lambda}(f)$, it holds that

$$\inf_{f: \|f - f_0\| = B(\kappa_n + \lambda^{1/2})} L_{n,\lambda}(f) - L_{n,\lambda}(f_0) < 0.$$

This implies that there is a sequence $g_n \in \mathcal{H}$ such that $\|g_n\| = B(\kappa_n + \lambda^{1/2})$ and $0 > L_{n,\lambda}(f_0 + g_n) - L_{n,\lambda}(f_0)$. As a consequence, it holds on event $E_{n,B}$ that

$$
\begin{aligned}
0 &> L_{n,\lambda}(f_0 + g_n) - L_{n,\lambda}(f_0) \\
&= -\frac{1}{n}\sum_{i=1}^{n} g_n(\boldsymbol{X}_i) + \int_{\mathcal{X}} \left(e^{f_0(\boldsymbol{x})+g_n(\boldsymbol{x})} - e^{f_0(\boldsymbol{x})}\right) d\boldsymbol{x} + \lambda\|f_0 + g_n\|_{\mathcal{H}}^2 - \lambda\|f_0\|_{\mathcal{H}}^2 \\
&= -\frac{1}{n}\sum_{i=1}^{n} g_n(\boldsymbol{X}_i) + \int_{\mathcal{X}} e^{f_0(\boldsymbol{x})}\left(e^{g_n(\boldsymbol{x})} - 1\right) d\boldsymbol{x} + \lambda\|g_n\|_{\mathcal{H}}^2 + 2\lambda\langle f_0, g_n\rangle_{\mathcal{H}} \\
&= -\mathbb{P}_n g_n + \mathbb{P} g_n + \int_{\mathcal{X}} e^{f_0(\boldsymbol{x})}\left(e^{g_n(\boldsymbol{x})} - 1 - g_n(\boldsymbol{x})\right) d\boldsymbol{x} + \lambda\|g_n\|_{\mathcal{H}}^2 + 2\lambda\langle f_0, g_n\rangle_{\mathcal{H}} \\
&= -(\mathbb{P}_n - \mathbb{P})g_n + \int_{\mathcal{X}} e^{f_0(\boldsymbol{x})}\left(e^{g_n(\boldsymbol{x})} - 1 - g_n(\boldsymbol{x}) - \frac{1}{2}g_n^2(\boldsymbol{x})\right) d\boldsymbol{x} + \frac{1}{2}\langle g_n, g_n\rangle_{L_2} \\
&\quad + \lambda\|g_n\|_{\mathcal{H}}^2 + 2\lambda\langle f_0, g_n\rangle_{\mathcal{H}}. \qquad (6.1)
\end{aligned}
$$

Noting that $\|g_n\|_{\sup} \leqslant CBh^{-1/2}(\kappa_n + \lambda^{1/2}) = o_P(1)$ for some $C > 0$ due to Lemma 2(iii), it follows from that

$$
\begin{aligned}
\frac{1}{2}\langle g_n, g_n\rangle_{L_2} + \lambda\|g_n\|_{\mathcal{H}}^2 &\overset{\text{(i)}}{\leqslant} \kappa_n\|g_n\| + c_n\langle g_n, g_n\rangle_{L_2} + 2\lambda\|f_0\|_{\mathcal{H}}\|g_n\|_{\mathcal{H}} \\
&\overset{\text{(ii)}}{\leqslant} \kappa_n\|g_n\| + c_n\langle g_n, g_n\rangle_{L_2} + C\lambda^{1/2}\|g_n\|. \qquad (6.2)
\end{aligned}
$$

Here (i) makes use of Lemma 16, (3.2), (6.1), and Cauchy–Schwarz inequality, and (ii) is due to the definition of $\|\cdot\|$ in (3.1). Since $c_n = o_P(1)$ by Lemma 16, we can assume $|c_n| \leqslant 1/4$. Therefore, the above inequality implies that the following holds on event $E_{n,B}$ :

$$
\begin{aligned}
\frac{1}{4}\|g_n\|^2 &\overset{\text{(i)}}{=} \frac{1}{4}V(g_n, g_n) + \frac{1}{4}\lambda\|g_n\|_{\mathcal{H}}^2 \\
&\overset{\text{(ii)}}{\leqslant} (1/2 - |c_n|)V(g_n, g_n) + \frac{1}{4}\lambda\|g_n\|_{\mathcal{H}}^2 \\
&\overset{\text{(iii)}}{\leqslant} C(1/2 - |c_n|)\langle g_n, g_n\rangle_{L_2} + \frac{C}{4}\lambda\|g_n\|_{\mathcal{H}}^2 \overset{\text{(iv)}}{\leqslant} C\kappa_n\|g_n\| + C^2\lambda^{1/2}\|g_n\|,
\end{aligned}
$$

where (i) is due to the definition of $\|\cdot\|$ in (3.1), (ii) follows since $|c_n| \leqslant 1/4$, (iii) makes use of Assumption B, and (iv) is from (6.2). Therefore, the above inequality implies the following holds on event $E_{n,B}$:

$$
\|g_n\| \leqslant (4C + C^2)(\kappa_n + \lambda^{1/2}).
$$

However, since $\|g_n\| = B(\kappa_n + \lambda^{1/2})$ on event $E_{n,B}$, it implies that

$$
0 < \delta \leqslant \mathbb{P}(E_{n,B}) \leqslant \mathbb{P}\left(\|g_n\| \leqslant (4C + C^2)(\kappa_n + \lambda^{1/2})\right) = \mathbb{P}(B < 4C + C^2).
$$

The above inequality holds for all $B \geqslant B_\delta$, which is a contradiction. $\square$

### *6.1. Derivative Estimation*

**Lemma 18.** *If $m > \beta_{\max}$, then $V(\partial^{\boldsymbol{\beta}} f, \partial^{\boldsymbol{\beta}} f) \leqslant \lambda^{-\frac{\beta_{\max}}{m}} \|f\|^2$ for all $f \in \mathcal{H}$.*

*Proof.* Let $\psi_{\boldsymbol{i}}(\boldsymbol{x}) = \phi_{i_1}(x_1)\ldots\phi_{i_r}(x_r)$, where $\phi_i's$ are the Fourier basis functions in (2.1). For any $f \in \mathcal{H}$, it follows that $f = \sum_{\boldsymbol{i}\in\mathbb{I}_q} c_{\boldsymbol{i}}\psi_{\boldsymbol{i}}$ for some sequence $c_{\boldsymbol{i}}$'s. Therefore, it follows that

$$V(\partial^{\boldsymbol{\beta}} f, \partial^{\boldsymbol{\beta}} f) = V\left(\sum_{\boldsymbol{i}\in\mathbb{I}_q} c_{\boldsymbol{i}}\partial^{\boldsymbol{\beta}}\psi_{\boldsymbol{i}}, \sum_{\boldsymbol{i}\in\mathbb{I}_q} c_{\boldsymbol{i}}\partial^{\boldsymbol{\beta}}\psi_{\boldsymbol{i}}\right) \overset{\text{(ii)}}{=} \sum_{\boldsymbol{i}\in\mathbb{I}_q} c_{\boldsymbol{i}}^2 V\left(\partial^{\boldsymbol{\beta}}\psi_{\boldsymbol{i}}, \partial^{\boldsymbol{\beta}}\psi_{\boldsymbol{i}}\right) \overset{\text{(ii)}}{\asymp} \sum_{\boldsymbol{i}\in\mathbb{I}_q} c_{\boldsymbol{i}}^2 i_1^{2\beta_1}\ldots i_r^{2\beta_r}.$$

Here (i) is due to the fact that $\partial^{\boldsymbol{\beta}}\psi_{\boldsymbol{i}}$'s are orthogonal under $V(\cdot,\cdot)$, and (ii) follows from xxx. By Lemma xxx, it holds that

$$\|f\|^2 = \sum_{\boldsymbol{i}\in\mathbb{I}_q} c_{\boldsymbol{i}}^2(1+\lambda/\rho_{\boldsymbol{i}}) \asymp \sum_{\boldsymbol{i}\in\mathbb{I}_q} c_{\boldsymbol{i}}^2(1+\lambda i_1^{2m}\ldots i_r^{2m}).$$

The desired result will follow if we prove the following inequality

$$\lambda^{\frac{\beta_{\max}}{m}} i_1^{2\beta_1}\ldots i_r^{2\beta_r} \lesssim 1+\lambda i_1^{2m}\ldots i_r^{2m} \tag{6.3}$$

for all $\boldsymbol{i} = (i_1,\ldots,i_r)^\top \in \mathbb{I}_q$. If $\lambda^{\frac{\beta_{\max}}{m}} i_1^{2\beta_1}\ldots i_r^{2\beta_r} \leqslant 1$, then (6.3) holds. If $\lambda^{\frac{\beta_{\max}}{m}} i_1^{2\beta_1}\ldots i_r^{2\beta_r} > 1$, then we have

$$\lambda^{\frac{\beta_{\max}}{m}} i_1^{2\beta_1}\ldots i_r^{2\beta_r} \overset{\text{(i)}}{\leqslant} (\lambda^{\frac{\beta_{\max}}{m}} i_1^{2\beta_1}\ldots i_r^{2\beta_r})^{\frac{m}{\beta_{\max}}} = \lambda i_1^{\frac{2m\beta_1}{\beta_{\max}}}\ldots i_r^{\frac{2m\beta_r}{\beta_{\max}}} \leqslant \lambda i_1^{2m}\ldots i_r^{2m},$$

where (i) is due to $m > \beta_{\max}$. Therefore, we verify (6.3). ☐

**Theorem 5.** *Under Assumptions xxx, it holds that*

$$\int_{\mathcal{X}} \left(\partial^\beta \widehat{f}(\boldsymbol{x}) - \partial^\beta f_0(\boldsymbol{x})\right)^2 d\boldsymbol{x} = O_P\left(\lambda^{1-\frac{\beta_{\max}}{m}} + n^{-1}\lambda^{-\frac{1+2\beta_{\max}}{2m}}[\log(n)]^{q-1}\right).$$

*As a consequence, if $\lambda \asymp \left(n[\log(n)]^{1-q}\right)^{-2m/(2m+1)}$, it follows that*

$$\int_{\mathcal{X}} \left(\partial^\beta \widehat{f}(\boldsymbol{x}) - \partial^\beta f_0(\boldsymbol{x})\right)^2 d\boldsymbol{x} = O_P\left\{\left(\frac{n}{[\log(n)]^{q-1}}\right)^{-\frac{2(m-\beta_{\max})}{2m+1}}\right\}.$$

*Proof.* Combining Theorem 4 and Lemma 18, we conclude that

$$V(\partial^\beta\widehat{f} - \partial^\beta f_0, \partial^\beta\widehat{f} - \partial^\beta f_0) \leqslant \lambda^{-\beta_{\max}/m}\|\widehat{p} - p_0\|^2 = O_P\left(\lambda^{1-\beta_{\max}/m} + \frac{1}{nh\lambda^{\beta_{\max}/m}}\right).$$

Using the above inequality and the fact that $h^{-1} \asymp \lambda^{-\frac{1}{2m}}[\log(1/\lambda)]^{q-1}$ from Lemma 7, we complete the proof. ☐

## 7. Uniform Convergence

Let $\xi_1, \ldots, \xi_n \in \mathbb{R}^d$ be a sequence of i.i.d. random vectors, and Let $\mathcal{F}$ be a class of functions from $\mathbb{R}^d$ to $\mathbb{R}$. The Rademacher complexity of $\mathcal{F}$ is defined as

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}\left\{\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} r_i f(\xi_i)\right\},$$

where $r_1, \ldots, r_n$ is a sequence of i.i.d. Rademacher random variables. For simplicity, let us define

$$\mathbb{P}(f) = \mathbb{E}\{f(\xi_1)\}, \quad \mathbb{P}_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(\xi_i).$$

**Lemma 19.** *Let $\mathcal{F}$ be a class of functions, then it holds that $\mathbb{E}(\sup_{f\in\mathcal{F}} |(\mathbb{P}-\mathbb{P}_n)(f)|) \leqslant 4\mathcal{R}_n(\mathcal{F})$.*

*Proof.* Let $\xi_1', \ldots, \xi_n'$ be an independent sample from $\xi_1, \ldots, \xi_n$. Using Jensen's inequality and the standard symmetrization trick in empirical process, it follows that

$$\begin{aligned}
\mathbb{E}\left(\sup_{f\in\mathcal{F}}(\mathbb{P}-\mathbb{P}_n)(f)\right) &= \mathbb{E}\left(\sup_{f\in\mathcal{F}}[\mathbb{P}(f)-\mathbb{P}_n(f)]\right) \\
&= \mathbb{E}\left(\sup_{f\in\mathcal{F}}\left[\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} f(\xi_i')\right) - \frac{1}{n}\sum_{i=1}^{n} f(\xi_i)\right]\right) \\
&\leqslant \mathbb{E}\left(\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^{n}[f(\xi_i') - f(\xi_i)]\right]\right) \\
&= \mathbb{E}\left(\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^{n} r_i[f(\xi_i') - f(\xi_i)]\right]\right) \\
&\leqslant 2\mathbb{E}\left(\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n} r_i f(\xi_i)\right) = 2\mathcal{R}_n(\mathcal{F}),
\end{aligned}$$

where $r_1, \ldots, r_n$ are Rademacher random variables. Similarly, we can show that

$$\begin{aligned}
\mathbb{E}\left(\sup_{f\in-\mathcal{F}}(\mathbb{P}-\mathbb{P}_n)(f)\right) \leqslant 2\mathbb{E}\left(\sup_{f\in-\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n} r_i f(\xi_i)\right) &= 2\mathbb{E}\left(\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n} -r_i f(\xi_i)\right) \\
&\overset{\text{(i)}}{=} 2\mathbb{E}\left(\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n} r_i f(\xi_i)\right) = 2\mathcal{R}_n(\mathcal{F}),
\end{aligned}$$

where (i) holds as $r_i$ and $-r_i$ have the same distribution. Noting that

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}}|(\mathbb{P}-\mathbb{P}_n)(f)|\right) \leqslant \mathbb{E}\left(\sup_{f\in\mathcal{F}}(\mathbb{P}-\mathbb{P}_n)(f)\right) + \mathbb{E}\left(\sup_{f\in-\mathcal{F}}(\mathbb{P}-\mathbb{P}_n)(f)\right),$$

the proof is completed by combining the above three inequalities. □

**Lemma 20.**

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}_M}|(\mathbb{P}_n-\mathbb{P})(f)|\right) \leqslant \frac{CM}{\sqrt{nh}}.$$

*where $\mathcal{F}_M=\{f: f\in\mathcal{H},\|f\|\leqslant M\}$, and $C>0$ is a universal constant.*

*Proof.* We use constant $C>0$ to denote a universal constant. Let $\zeta_i=\boldsymbol{X}_i$, and direct examination implies that

$$\begin{aligned}\mathcal{R}_n(\mathcal{F}_M) &= \mathbb{E}\left(\sup_{f\in\mathcal{F}_M}\frac{1}{n}\sum_{i=1}^n r_i f(\boldsymbol{X}_i)\right)\\ &= \mathbb{E}\left(\sup_{f\in\mathcal{F}_M}\frac{1}{n}\langle f,\sum_{i=1}^n r_i K_{\boldsymbol{X}_i}\rangle\right)\\ &\leqslant \frac{1}{n}\sup_{f\in\mathcal{F}_M}\|f\|\mathbb{E}\left(\left\|\sum_{i=1}^n r_i K_{\boldsymbol{X}_i}\right\|\right)\\ &\leqslant \frac{M}{n}\sqrt{\mathbb{E}\left(\left\|\sum_{i=1}^n r_i K_{\boldsymbol{X}_i}\right\|^2\right)}\\ &= \frac{M}{n}\sqrt{\sum_{i=1}^n\mathbb{E}\{K(\boldsymbol{X}_i,\boldsymbol{X}_i)\}} = M\sqrt{\frac{\mathbb{E}\{K(\boldsymbol{X}_1,\boldsymbol{X}_1)\}}{n}} \overset{\text{(i)}}{\leqslant} \frac{CM}{\sqrt{nh}},\end{aligned} \tag{7.1}$$

where (i) follows from Lemma 2. Finally, using Lemma 19 and (7.1), it follows that

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}_M}|(\mathbb{P}-\mathbb{P}_n)(f)|\right)\leqslant\frac{CM}{\sqrt{nh}},$$

which completes the proof. □

***Proof of (3.2).*** These are direct consequences of Lemma 20 with $M=1$. □

**Lemma 21.** *Suppose that $Y_n\in[0,B_1]$ and $\liminf_{n\to\infty}\mathbb{E}(Y_n)=B_2$ for some $B_1,B_2>0$, then there is a constant $\delta>0$ such that $\liminf_{n\to\infty}\mathbb{P}(Y_n\geqslant\delta)\geqslant\delta$.*

*Proof.* Assume the statement is false. Then there is a sequence $\delta_k\to 0$ such that $\liminf_{n\to\infty}\mathbb{P}(Y_n\geqslant\delta_k)<\delta_k$. Hence, it holds that

$$\mathbb{E}(Y_n)=\mathbb{E}\left(Y_nI(Y_n\geqslant\delta_k)\right)+\mathbb{E}\left(Y_nI(Y_n<\delta_k)\right)\leqslant B_1\mathbb{P}(Y_n\geqslant\delta_k)+\delta_k.$$

Taking limit, we have

$$B_2=\liminf_{n\to\infty}\mathbb{E}(Y_n)\leqslant B_1\liminf_{n\to\infty}\mathbb{P}(Y_n\geqslant\delta_k)+\delta_k\leqslant(B_1+1)\delta_k.$$

Since $\delta_k\to 0$, we lead to a contradiction. □

### *7.1. Approximation GCV*

$$\widehat{f}_\lambda = \underset{f \in \mathcal{H}_S}{\operatorname{argmin}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{X}_i) + \int_{\mathcal{X}} e^{f(\boldsymbol{x})} d\boldsymbol{x} + \lambda \|f\|_{\mathcal{H}}^2 \right\},$$

$$\mathcal{H}_S = \left\{ f \in \mathcal{H} : f(\boldsymbol{x}) = \boldsymbol{c}^\top S \boldsymbol{\Psi}(\boldsymbol{x}) \text{ for } \boldsymbol{c} \in \mathbb{R}^p \right\}.$$

$$\widehat{f}_{\lambda,-i} = \underset{f \in \mathcal{H}_S}{\operatorname{argmin}} \left\{ -\frac{1}{n-1} \sum_{j=\neq i} f(\boldsymbol{X}_j) + \int_{\mathcal{X}} e^{f(\boldsymbol{x})} d\boldsymbol{x} + \lambda \|f\|_{\mathcal{H}}^2 \right\},$$

$S \in \mathbb{R}^{m\times n}$, $S^\top \boldsymbol{c}$

$$\begin{aligned} Q_\lambda(\boldsymbol{c}) &= -\frac{1}{n} \mathbf{1}^\top R S^\top \boldsymbol{c} + \int_{\mathcal{X}} e^{\boldsymbol{c}^\top S \boldsymbol{\Psi}(\boldsymbol{x})} d\boldsymbol{x} + \lambda \boldsymbol{c}^\top S R S^\top \boldsymbol{c}, \\ Q_{\lambda,-i}(\boldsymbol{c}) &= -\frac{1}{n-1} (\mathbf{1} - e_i)^\top R S^\top \boldsymbol{c} + \int_{\mathcal{X}} e^{\boldsymbol{c}^\top S \boldsymbol{\Psi}(\boldsymbol{x})} d\boldsymbol{x} + \lambda \boldsymbol{c}^\top S R S^\top \boldsymbol{c}, \end{aligned}$$

Let $\widehat{\boldsymbol{c}}_\lambda = \operatorname{argmin}_{\boldsymbol{c} \in \mathbb{R}^p} Q_\lambda(\boldsymbol{c})$ and $\widehat{\boldsymbol{c}}_{\lambda,-i} = \operatorname{argmin}_{\boldsymbol{c} \in \mathbb{R}^p} Q_{\lambda,-i}(\boldsymbol{c})$. Hence, it follows that xxx.

$$\begin{aligned} \dot{Q}_\lambda(\boldsymbol{c}) &= -\frac{1}{n} S R \mathbf{1} + \int_{\mathcal{X}} e^{\boldsymbol{c}^\top S \boldsymbol{\Psi}(\boldsymbol{x})} S \boldsymbol{\Psi}(\boldsymbol{x}) d\boldsymbol{x} + 2\lambda S R S^\top \boldsymbol{c}, \\ \dot{Q}_{\lambda,-i}(\boldsymbol{c}) &= -\frac{1}{n-1} S R (\mathbf{1} - e_i) + \int_{\mathcal{X}} e^{\boldsymbol{c}^\top S \boldsymbol{\Psi}(\boldsymbol{x})} S \boldsymbol{\Psi}(\boldsymbol{x}) d\boldsymbol{x} + 2\lambda S R S^\top \boldsymbol{c}, \\ &= -\frac{1}{n-1} S R (\mathbf{1} - e_i) + \dot{Q}_\lambda(\boldsymbol{c}) + \frac{1}{n} S R \mathbf{1} \\ &= -\frac{1}{n(n-1)} S R \mathbf{1} + \frac{1}{n-1} S R e_i + \dot{Q}_\lambda(\boldsymbol{c}), \\ \ddot{Q}_\lambda(\boldsymbol{c}) = \ddot{Q}_{\lambda,-i}(\boldsymbol{c}) &= S \left\{ \int_{\mathcal{X}} e^{\boldsymbol{c}^\top S \boldsymbol{\Psi}(\boldsymbol{x})} \boldsymbol{\Psi}(\boldsymbol{x}) \boldsymbol{\Psi}^\top(\boldsymbol{x}) d\boldsymbol{x} + 2\lambda R \right\} S^\top. \end{aligned}$$

$$\begin{aligned} \ddot{Q}(\widetilde{\boldsymbol{c}})(\boldsymbol{c} - \widetilde{\boldsymbol{c}}) &= -\dot{Q}(\widetilde{\boldsymbol{c}}), \\ \ddot{Q}(\widetilde{\boldsymbol{c}})\boldsymbol{c} &= -\dot{Q}(\widetilde{\boldsymbol{c}}) + \ddot{Q}(\widetilde{\boldsymbol{c}})\widetilde{\boldsymbol{c}}. \end{aligned}$$

$$
\begin{aligned}
L_{f,g}(t) &= \int_{\mathcal{X}} e^{f(\boldsymbol{x})+tg(\boldsymbol{x})} d\boldsymbol{x}, \\
\dot{L}_{f,g}(t) &= \int_{\mathcal{X}} g(\boldsymbol{x}) e^{f(\boldsymbol{x})+tg(\boldsymbol{x})} d\boldsymbol{x}, \\
\dot{L}_{f,g}(0) &= \int_{\mathcal{X}} g(\boldsymbol{x}) e^{f(\boldsymbol{x})} d\boldsymbol{x} = \mu_f(g), \\
\ddot{L}_{f,g}(t) &= \int_{\mathcal{X}} g^2(\boldsymbol{x}) e^{f(\boldsymbol{x})+tg(\boldsymbol{x})} d\boldsymbol{x}, \\
\ddot{L}_{f,g}(0) &= \int_{\mathcal{X}} g^2(\boldsymbol{x}) e^{f(\boldsymbol{x})} d\boldsymbol{x} = V_f(g),
\end{aligned}
$$

$$
\begin{aligned}
\int_{\mathcal{X}} e^{f(\boldsymbol{x})} d\boldsymbol{x} = L_{\widetilde{f}, f-\widetilde{f}}(1) &\approx L_{\widetilde{f}, f-\widetilde{f}}(0) + \mu_{\widetilde{f}}(f-\widetilde{f}) + \frac{1}{2} V_{\widetilde{f}}(f-\widetilde{f}) \\
&= \mu_{\widetilde{f}}(f) - V_{\widetilde{f}}(f, \widetilde{f}) + \frac{1}{2} V_{\widetilde{f}}(f) + \text{const.}
\end{aligned}
$$

$$
-\frac{1}{n-1} \sum_{j \neq i} f(\boldsymbol{X}_j) + \mu_{\widetilde{f}}(f) - V_{\widetilde{f}}(f, \widetilde{f}) + \frac{1}{2} V_{\widetilde{f}}(f) + \lambda \|f\|_{\mathcal{H}}^2.
$$

$$
-\frac{1}{n} \sum_{j=1}^{n} f(\boldsymbol{X}_j) + \mu_{\widetilde{f}}(f) - V_{\widetilde{f}}(f, \widetilde{f}) + \frac{1}{2} V_{\widetilde{f}}(f) + \lambda \|f\|_{\mathcal{H}}^2.
$$

Since $\dot{Q}_\lambda(\widehat{\boldsymbol{c}}_\lambda) = 0$, it follows that

$$
\begin{aligned}
\widehat{\boldsymbol{c}}_{\lambda,-i} &\approx \widehat{\boldsymbol{c}}_\lambda - \ddot{Q}_{\lambda,-i}^{-1}(\widehat{\boldsymbol{c}}_\lambda) \dot{Q}_{\lambda,-i}(\widehat{\boldsymbol{c}}_\lambda) \\
&= \widehat{\boldsymbol{c}}_\lambda - \ddot{Q}_{\lambda}^{-1}(\widehat{\boldsymbol{c}}_\lambda) \left( -\frac{1}{n(n-1)} SR\mathbf{1} + \frac{1}{n-1} SRe_i + \dot{Q}_\lambda(\widehat{\boldsymbol{c}}_\lambda) \right) \\
&= \widehat{\boldsymbol{c}}_\lambda + \frac{1}{n(n-1)} \ddot{Q}_{\lambda}^{-1}(\widehat{\boldsymbol{c}}_\lambda) SR\mathbf{1} - \frac{1}{n-1} \ddot{Q}_{\lambda}^{-1}(\widehat{\boldsymbol{c}}_\lambda) SRe_i.
\end{aligned}
$$

$$
\begin{aligned}
\widehat{f}_{\lambda,-i}(\boldsymbol{X}_i) &\approx \boldsymbol{\Psi}^\top(\boldsymbol{X}_i)S^\top\widehat{c}_{\lambda,-i} \\
&= \widehat{f}_\lambda(\boldsymbol{X}_i) + \frac{1}{n(n-1)}\boldsymbol{\Psi}_\lambda^\top(\boldsymbol{X}_i)S^\top\ddot{Q}_\lambda^{-1}(\widehat{\boldsymbol{c}}_\lambda)SR\mathbf{1} - \frac{1}{n-1}\boldsymbol{\Psi}^\top(\boldsymbol{X}_i)S^\top\ddot{Q}_\lambda^{-1}(\widehat{\boldsymbol{c}}_\lambda)SRe_i \\
&= \widehat{f}_\lambda(\boldsymbol{X}_i) - \frac{1}{n-1}\boldsymbol{\Psi}^\top(\boldsymbol{X}_i)S^\top\ddot{Q}_\lambda^{-1}(\widehat{\boldsymbol{c}}_\lambda)SR\left(e_i - \mathbf{1}/n\right) \\
&= \widehat{f}_\lambda(\boldsymbol{X}_i) - \frac{1}{n-1}\left(\boldsymbol{\Psi}(\boldsymbol{X}_i) - R\mathbf{1}/n\right)^\top S^\top\ddot{Q}_\lambda^{-1}(\widehat{\boldsymbol{c}}_\lambda)SR\left(e_i - \mathbf{1}/n\right) \\
&\quad - \frac{1}{n(n-1)}\mathbf{1}^\top R^\top S^\top\ddot{Q}_\lambda^{-1}(\widehat{\boldsymbol{c}}_\lambda)SR\left(e_i - \mathbf{1}/n\right) \\
&= \widehat{f}_\lambda(\boldsymbol{X}_i) - \frac{1}{n-1}\left(Re_i - R\mathbf{1}/n\right)^\top S^\top\ddot{Q}_\lambda^{-1}(\widehat{\boldsymbol{c}}_\lambda)SR\left(e_i - \mathbf{1}/n\right) \\
&\quad - \frac{1}{n(n-1)}\mathbf{1}^\top R^\top S^\top\ddot{Q}_\lambda^{-1}(\widehat{\boldsymbol{c}}_\lambda)SR\left(e_i - \mathbf{1}/n\right) \\
&= \widehat{f}_\lambda(\boldsymbol{X}_i) - \frac{1}{n-1}\left(e_i - \mathbf{1}/n\right)^\top R^\top S^\top\ddot{Q}_\lambda^{-1}(\widehat{\boldsymbol{c}}_\lambda)SR\left(e_i - \mathbf{1}/n\right) \\
&\quad - \frac{1}{n(n-1)}\mathbf{1}^\top R^\top S^\top\ddot{Q}_\lambda^{-1}(\widehat{\boldsymbol{c}}_\lambda)SR\left(e_i - \mathbf{1}/n\right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^n\widehat{f}_{\lambda,-i}(\boldsymbol{X}_i) &\approx \frac{1}{n}\sum_{i=1}^n\widehat{f}_\lambda(\boldsymbol{X}_i) - \frac{1}{n(n-1)}\sum_{i=1}^n\left(e_i - \mathbf{1}/n\right)^\top R^\top S^\top\ddot{Q}_\lambda^{-1}(\widehat{\boldsymbol{c}}_\lambda)SR\left(e_i - \mathbf{1}/n\right) \\
&= \frac{1}{n}\sum_{i=1}^n\widehat{f}_\lambda(\boldsymbol{X}_i) - \frac{1}{n(n-1)}\sum_{i=1}^n\left(e_i - \mathbf{1}/n\right)^\top R^\top S^\top\ddot{Q}_\lambda^{-1}(\widehat{\boldsymbol{c}}_\lambda)SR\left(e_i - \mathbf{1}/n\right) \\
&= \frac{1}{n}\sum_{i=1}^n\widehat{f}_\lambda(\boldsymbol{X}_i) - \frac{1}{n(n-1)}Tr\left\{(I-P_1)R^\top S^\top\ddot{Q}_\lambda^{-1}(\widehat{\boldsymbol{c}}_\lambda)SR(I-P_1)\right\}.
\end{aligned}
$$

Here $P_1 = I - \mathbf{1}\mathbf{1}^\top/n$.

Minimize

$$
AGCV(\lambda) = -\frac{1}{n}\sum_{i=1}^n\widehat{f}_\lambda(\boldsymbol{X}_i) + \int_{\mathcal{X}} e^{\widehat{f}_\lambda(\boldsymbol{x})}d\boldsymbol{x} + \frac{1}{n(n-1)}Tr\left\{(I-P_1)R^\top S^\top\ddot{Q}_\lambda^{-1}(\widehat{\boldsymbol{c}}_\lambda)SR(I-P_1)\right\}
$$

$$
GCV(\lambda) = -\frac{1}{n}\sum_{i=1}^n\widehat{f}_{\lambda,-i}(\boldsymbol{X}_i) + \int_{\mathcal{X}} e^{\widehat{f}_\lambda(\boldsymbol{x})}d\boldsymbol{x}
$$

## References


Lin, Y. (2000). Tensor product space anova models. *Annals of Statistics*, 28(3):734–755.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer Science & Business Media.