Electronic Journal of Statistics
Vol. 14 (2020) 302–344
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1668

Asymptotics and optimal bandwidth for nonparametric estimation of density level sets

Wanli Qiao*

Department of Statistics
George Mason University
4400 University Drive, MS 4A7
Fairfax, VA 22030
e-mail: wqiao@gmu.edu

Abstract: Bandwidth selection is crucial in the kernel estimation of density level sets. A risk based on the symmetric difference between the estimated and true level sets is usually used to measure their proximity. In this paper we provide an asymptotic L^p approximation to this risk, where p is characterized by the weight function in the risk. In particular the excess risk corresponds to an L^2 type of risk, and is adopted to derive an optimal bandwidth for nonparametric level set estimation of d-dimensional density functions ($d \geq 1$). A direct plug-in bandwidth selector is developed for kernel density level set estimation and its efficacy is verified in numerical studies.

AMS 2000 subject classifications: Primary 62G20; secondary 62G05.
Keywords and phrases: Level set, optimal bandwidth, kernel density estimation, symmetric difference.

Received April 2019.

1. Introduction

For a density function f on $\mathbb{R}^d$, $d \geq 1$, its (upper) level set at a given level c is defined as

$$\mathcal{L}_c = \{x \in \mathbb{R}^d : \ f(x) \geq c\}.$$

With a given random sample from f, it is often of interest to estimate $\mathcal{L}_c$. Density level set estimation has been useful in many areas, such as clustering (Rinaldo and Wasserman, 2010), classification (Steinwart et al., 2005), tests for multimodality (Müller and Sawitzki, 1991), and topological data analysis (Fasy et al., 2014). A plug-in estimator of $\mathcal{L}_c$ using kernel density estimation is given by

$$\widehat{\mathcal{L}}_c = \{x \in \mathbb{R}^d : \ \widehat{f}(x) \geq c\},$$

where $\widehat{f}(x)$ is the kernel estimator of $f(x)$ (see (2.1)). It is well-known that the choice of bandwidth plays a crucial role in the performance of kernel-type

*Partially supported by NSF grants DMS 1821154 and FET 1900061, and a Jeffress Memorial Trust Award.

estimators. In this paper we derive an asymptotically optimal bandwidth for $\widehat{\mathcal{L}}_c$. We take the level c as a fixed value and denote $\mathcal{L} = \mathcal{L}_c$ and $\widehat{\mathcal{L}} = \widehat{\mathcal{L}}_c$ for simplicity.

The optimal bandwidth selection for $\widehat{f}$ has been studied extensively in the literature, usually based on ISE, MISE, or MIAE (see Wand and Jones, 1995). These criteria measure the proximity between $\widehat{f}$ and f over $\mathbb{R}^d$. We emphasize here that the target of our estimation $\mathcal{L}$ is a set rather than a density function; this has a critical impact on the optimal bandwidth, since the quality of density estimation should be prioritized regionally rather than over the entire domain. Figure 1 is an illustration, which shows that the overall closeness of density functions is not equivalent to the closeness of their level sets. Therefore, the criteria used for optimal bandwidth should be tailored specifically for nonparametric level set estimation.

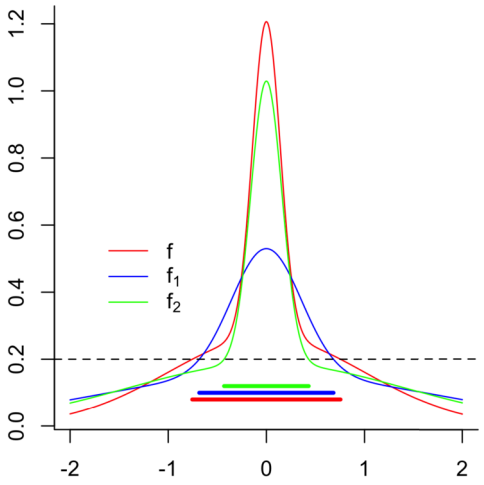

FIG 1. *Three density functions on* $\mathbb{R}$ *are shown on the graph:* f *(red curve),* f_1 *(blue curve) and* f_2 *(green curve). The level sets with* $c = 0.2$ *(dotted line) of the three density functions are considered. The thick horizontal lines underneath the density curves represent the level sets. Overall, the density function* f_2 *is closer than* f_1 *to* f*. However, the level set of* f_1 *is closer to that of* f*.*

A usual loss function used to measure the closeness between $\mathcal{L}$ and $\widehat{\mathcal{L}}$ is based on their symmetric difference. For any two sets A and B, let $A\Delta B$ be their symmetric difference, i.e., $A\Delta B = (A \cap B^{\complement}) \cup (B \cap A^{\complement})$, where we use $^{\complement}$ to denote the complement of a set. For any nonnegative integrable function g on $\mathbb{R}^d$ and any Lebesgue measurable subset A of $\mathbb{R}^d$, denote $\lambda_g(A) = \int_A g(x)dx$. In the literature $\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}})$ has been well accepted as a measure of the proximity between $\mathcal{L}$ and $\widehat{\mathcal{L}}$, due to its natural geometric interpretation. Examples of $g(x)$ include $f(x)$ and $|f(x) - c|^q$ for some $q \geq 0$. The asymptotics of $\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}})$ has been studied, e.g., in Baíllo et al. (2000), Baíllo (2003), Cadre (2006), Cuevas et al. (2006), Biau et al. (2008), and Mason and Polonik (2009).

We use the risk $\mathbb{E}\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}})$ to study the problem of optimal bandwidth for density level set estimation for $d \geq 1$. A critical step is to obtain the asymptotic expression for this risk, which is one of the main results in this paper and is first described below.

Let $\mathcal{M} = \{x \in \mathbb{R}^d : \ f(x) = c\}$, which is the boundary of $\mathcal{L}$ (i.e., $\mathcal{M} = \partial\mathcal{L}$) under a mild assumption (e.g., see our assumption (F1) below). Let Vol_{d-1} be the natural $(d-1)$-dimensional volume measure that $\mathcal{M}$ inherits as a subset of $\mathbb{R}^d$, and $d(x, \mathcal{M})$ be the distance from any $x \in \mathbb{R}^d$ to $\mathcal{M}$. Suppose $g(x)$ is approximately pth power of $d(x, \mathcal{M})$ for x in a small neighborhood of $\mathcal{M}$. Then under regularity conditions the following approximation holds asymptotically:

$$\mathbb{E}\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}}) = \mathbb{E}\int_{\mathcal{M}} |\widehat{f}(x) - f(x)|^{p+1} w_g(x) d\text{Vol}_{d-1}(x)\{1 + o(1)\}, \tag{1.1}$$

where w_g is a positive function on $\mathcal{M}$. Here we approximate a risk describing horizontal variations with the one constructed with vertical variations. A rigorous statement with appropriate assumptions for (1.1) is given in Theorem 3.1 below. Using the above expression we can interpret $\mathbb{E}\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}})$ asymptotically as a weighted L^{p+1} risk for density estimation, in the form of an integration over the boundary of $\mathcal{L}$ with respect to the $(d-1)$-dimensional volume measure. When $g \equiv 1$, λ_g is the Lebesgue measure. In this case, $p = 0$ and the above approximation corresponds to the L^1 risk called Mean Integrated Absolute Error (MIAE), which has been used as a measure of proximity for optimal bandwidth selection for kernel density estimation (See Devroye and Györfi (1985), Devroye (1987), Hall and Wand (1988), Holmström and Klemelä (1992), and Devroye and Lugosi (2001)). Alternatively, Mean Integrated Squared Error (MISE) is more tractable than MIAE. This motivates us to use the choice of g with $p = 1$ for optimal bandwidth, specifically $g(x) = |f(x) - c|$. In this case, $\mathbb{E}\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}})$ corresponds to the excess risk (or regret) in the classification literature. In fact the excess risk has been used to find the optimal tuning parameter for nonparametric classifier, where the excess risk can be asymptotically decomposed into a squared bias term and a variance term (Hall and Kang 2005, Hall et al. 2008, Samworth 2012, Cannings et al. 2017). The results in this paper provide a way of understanding the excess risk as an L^2 risk, in a more general setting. In addition to the asymptotic approximation for the risk $\mathbb{E}\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}})$, we also show the asymptotic approximation for the error

$$\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}}) = \int_{\mathcal{M}} |\widehat{f}(x) - f(x)|^{p+1} w_g(x) d\text{Vol}_{d-1}(x)\{1 + o_p(1)\}, \tag{1.2}$$

under some extra assumptions on the convergence rate of the bandwidth.

Some of the important work on level set estimation includes Hartigan (1987), Polonik (1995), Tsybakov (1997), Walther (1997), Cadre (2006), Rigollet and Vert (2009), among many others. Also see Mason and Polonik (2009) for a comprehensive review of the literature for level set estimation. Confidence regions for level sets have recently been studied in Mammen and Polonik (2013), Sommerfeld et al. (2015), Chen et al. (2017), and Qiao and Polonik (2019). In Jang

(2006), plug-in level set estimation is applied to two-dimensional astronomical sky survey data, where the selection of bandwidth is based on the classical plug-in and cross validation approaches for density estimation.

When the level c is not explicitly given but determined by a probability value $\tau \in (0,1)$ through $c = \inf\{y \in (0,\infty) : \int_{f(x)\geq y} f(x)dx \leq 1-\tau\}$, we denote $c = c(\tau)$ and $\mathcal{L}_{c(\tau)}$ is called the $100(1-\tau)\%$ highest density region (HDR) of f (see Hyndman, 1996). The corresponding plug-in estimator is $\widehat{\mathcal{L}}_{\widehat{c}(\tau)}$, where $\widehat{c}(\tau) = \inf\{y \in (0,\infty) : \int_{\widehat{f}(x)\geq y} \widehat{f}(x)dx \leq 1-\tau\}$. In the case of f being a univariate density (i.e., $d=1$), the bandwidth selection problem for estimating HDR was studied in Samworth and Wand (2010). They chose $\mathbb{E}\lambda_g(\mathcal{L}_{c(\tau)} \,\Delta\, \widehat{\mathcal{L}}_{\widehat{c}(\tau)})$ with $g = f$ as the risk function to minimize for bandwidth selection. The extension of their approach to the multivariate case is far from trivial and has been recently studied in Doss and Weng (2018).

The work in Doss and Weng (2018) also considers the bandwidth selection problem of the estimation of density level sets. The comparison of their work with an earlier arXiv version of the present paper (see Qiao, 2018) has been discussed in Doss and Weng (2018). In particular, the risk criterion they use for bandwidth selection for level set estimation is $\mathbb{E}\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}})$ with $g = f$, which is an L^1 type of risk as a special case of (1.1). See Remark 3.4 for Corollary 3.1 as well as the Discussion section for more detailed comparisons.

The rest of the paper is organized as follows. We first introduce some notation and geometric concepts in Section 2. In Section 3, after discussing the assumptions that we will use, we derive some asymptotic results for the L^p type of risks introduced above and an optimal bandwidth for density level set estimation. Specifically, Theorems 3.1 and 3.2 formulate the ideas given in (1.2) and (1.1), respectively. Corollary 3.1 gives an exact asymptotic expression for $\mathbb{E}\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}})$ when $p = 0$. The excess risk as an asymptotic L^2 type of risk, is used to find the optimal bandwidth with the result given in Theorem 3.3. Simulation results are presented in Section 4, where we show the efficacy of our bandwidth selector in finite samples. We leave all the proofs to Section 5, while some miscellaneous results are put in the appendix.

2. Notation and some geometric concepts

Let $X_1, \cdots, X_n$ be i.i.d. from the d-dimensional density function f. Denote the bandwidth vector $\mathbf{h} = (h_1, h_2, \cdots, h_d)^T$ and $\mathbf{h}^{-1} = (h_1^{-1}, h_2^{-1}, \cdots, h_d^{-1})^T$. We consider a kernel density estimator

$$\widehat{f}(x) = \frac{1}{n\Pi_{j=1}^d h_j} \sum_{i=1}^n K\left(\mathbf{h}^{-1} \odot (x - X_i)\right), \tag{2.1}$$

where K is a kernel function on $\mathbb{R}^d$ and $\odot$ is used to denote the Hadamard or element-wise product between two vectors of the same size. Here we assign a bandwidth value for each of the variables in the density estimation. This

corresponds to a diagonal bandwidth matrix, which is a compromise between flexibility (by using a full bandwidth matrix) and simplicity (by using only a scalar bandwidth). See Wand and Jones (1994) for more discussion on the impact of the form of bandwidth matrix on multivariate density estimation. We use a product kernel for K, i.e., we can write

$$K\left(\mathbf{h}^{-1}\odot(x-X_i)\right)=\prod_{j=1}^{d}\widetilde{K}\left(\frac{x_j-X_{ij}}{h_j}\right),$$

where $\widetilde{K}$ is a univariate kernel function, $X_i=(X_{i1},\cdots,X_{id})^T$, for $i=1,\cdots,n$, and $x=(x_1,\cdots,x_d)^T$. The order of a kernel is determined by its first nonzero moment. We call K a νth ($\nu\geq 2$) order kernel if $\int_{\mathbb{R}}|u^{\nu}\widetilde{K}(u)|du<\infty$ and

$$\int_{\mathbb{R}}u^l\widetilde{K}(u)du=\begin{cases}1, & \text{if } l=0,\\ 0, & \text{if } l=1,\cdots,\nu-1,\\ \kappa_\nu\neq 0, & \text{if } l=\nu.\end{cases}$$

It is obvious that ν is always even if $\widetilde{K}$ is symmetric. When $d=1$, we also denote $h=\mathbf{h}$ and write $\widehat{f}(x)=\frac{1}{nh}\sum_{i=1}^{n}K(h^{-1}(x-X_i))$.

Notation: Let $\mathcal{H}_{d-1}$ be the $(d-1)$-dimensional normalized Hausdorff measure on $\mathbb{R}^d$ (cf. Evans and Gariepy, 1992). It agrees with the $(d-1)$-dimensional volume measure Vol_{d-1} on nice sets. For $d=1$, $\mathcal{H}_0$ is the cardinality of a set, such that for $A=\{a_1,\cdots,a_m\}\subset\mathbb{R}$ and a function $g:\mathbb{R}\mapsto\mathbb{R}$, $\int_A g(x)d\mathcal{H}_0(x)=\sum_{i=1}^{m}g(a_i)$. For simplicity, we usually omit the subscript $d-1$ of $\mathcal{H}_{d-1}$ and write $\mathcal{H}=\mathcal{H}_{d-1}$ if the value of d is clear in the context. Let λ be the d-dimensional Lebesgue measure ($d\geq 1$). Recall that λ and $\mathcal{H}_d$ are equal on $\mathbb{R}^d$. Let $\mathbf{I}$ be the indicator function.

For a positive integer p, let $\mathbb{Z}_+^p$ be the set of all the p-dimensional vectors of positive integers. For any $\mathbf{i}=(i_1,\cdots,i_p)^T\in\mathbb{Z}_+^p$, let $\max(\mathbf{i})=\max(i_1,\cdots,i_p)$. Denote $\mathbb{Z}_+^{p,d}=\{\mathbf{i}\in\mathbb{Z}_+^p,\max(\mathbf{i})\leq d\}$. For any function $g:\mathbb{R}^d\mapsto\mathbb{R}$ with pth (partial) derivatives ($p\geq 1$), and for $\mathbf{i}\in\mathbb{Z}_+^{p,d}$, denote $g_{(\mathbf{i})}(x)=g_{(i_1,\cdots,i_p)}(x)=\frac{\partial^p}{\partial x_{i_1}\cdots\partial x_{i_p}}g(x)$ with the convention $g_{(\mathbf{i})}(x)=g(x)$ for $\mathbf{i}\in\mathbb{Z}_+^0$. For example, $g_{(k,l)}(x)=\frac{\partial^2}{\partial x_k\partial x_l}g(x)$ for $1\leq k,l\leq d$. If $i_1=\cdots=i_p=i$ for $1\leq i\leq d$, we denote $g_{(i*p)}(x)=g_{(i_1,\cdots,i_p)}(x)$. For $x\in\mathbb{R}^d$ and $\mathbf{i}\in\mathbb{Z}_+^{p,d}$, denote $x^{(\mathbf{i})}=x_{i_1}\times\cdots\times x_{i_p}$. For $d\geq 2$, we denote the gradient and Hessian matrix of g by ∇g and $\nabla^2 g$, respectively. With slight abuse of notation, we also use ∇g to denote the first derivative g' when $d=1$. For any Borel set $A\subset\mathbb{R}$, let $g^{-1}(A)=\{x\in\mathbb{R}^d:\ g(x)\in A\}$. Let $\|g\|_q=(\int_{\mathbb{R}^d}|g(x)|^q dx)^{1/q}$ for $q>0$ and $\|g\|_\infty=\sup_{x\in\mathbb{R}^d}|g(x)|$. For sequences $a_n,b_n\in\mathbb{R}$, we denote $a_n\asymp b_n$ if $0<\liminf_{n\to\infty}|a_n/b_n|\leq\limsup_{n\to\infty}|a_n/b_n|<\infty$. For a sequence $\mathbf{a}_n=(a_{1,n},\cdots,a_{d,n})^T\in\mathbb{R}^d$, $d\geq 2$, denote $\mathbf{a}_n\asymp b_n$ if $a_{i,n}\asymp b_n$, $i=1,\cdots,d$.

Next we introduce some geometric concepts. For any set $A\subset\mathbb{R}^d$ and $\epsilon>0$, we denote the ϵ-enlargement of A by $A\oplus\epsilon=\bigcup_{x\in A}\mathcal{B}_x(\epsilon)$, where $\mathcal{B}_x(\epsilon)=\{y\in$

$\mathbb{R}^d : \|x - y\| \leq \epsilon\}$. For any two sets $A, B \subset \mathbb{R}^d$, let $d_H(A, B)$ be the Hausdorff distance between A and B, i.e.,

$$d_H(A, B) = \max\left\{\sup_{x \in B} d(x, A),\ \sup_{x \in A} d(x, B)\right\},$$

where $d(x, A) = \inf_{y \in A} \|x - y\|$. Let $\pi_A(x)$ be the set of the closest points in A to x, i.e. $\pi_A(x) = \{y \in A : \|x - y\| = d(x, A)\}$, which is called the normal projection of x onto A.

Let A and B be two $(d-1)$-dimensional smooth submanifolds embedded in $\mathbb{R}^d$ $(d \geq 2)$. Then the normal projections $\pi_A : B \mapsto A$ and $\pi_B : A \mapsto B$ define two maps between A and B. The two manifolds A and B are called *normal compatible* if the projections π_A and π_B are homeomorphisms. See Chazal et al. (2007) and Figure 2 for a graphical illustration.

We will also use the concept of *reach* of a manifold. For a p-dimensional manifold $\mathcal{S}$ embedded in $\mathbb{R}^d$ $(p < d)$, the reach of $\mathcal{S}$, denoted by $\rho(\mathcal{S})$, is the largest δ such that the normal projection from every point in $\mathcal{S} \oplus \delta$ onto $\mathcal{S}$ is unique. See Federer (1959). A positive reach corresponds to the notion of bounded curvature of a manifold. See Niyogi et al. (2008) and Genovese et al. (2012).

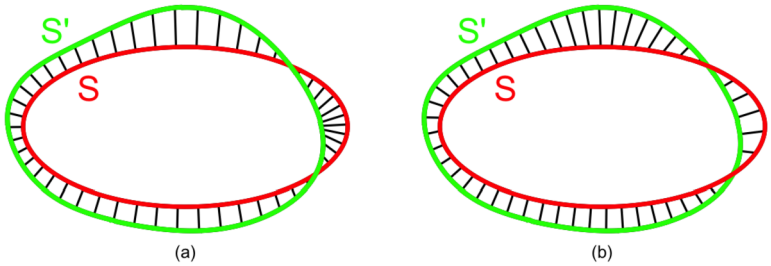

FIG 2. *Two normal compatible curves S and S'. (a) represents the normal projection from S' to S. (b) represents the normal projection from S to S'.*

3. Main results

3.1. Assumptions and their discussion

We introduce the assumptions that will be used in this paper. Let $\mathbb{R}^d_+$ be the set of vectors in $\mathbb{R}^d$ with positive coordinates. With the requirement of C^2 smoothness of the kernel function K, define the class of functions

$$\mathcal{K} = \left\{K_{(\mathbf{i})}(\mathbf{h}^{-1} \odot (x - \cdot)) : \mathbf{h} \in \mathbb{R}^d_+, x \in \mathbb{R}^d, \mathbf{i} \in \mathbb{Z}^p_+, p \in \{0, 1, 2\}, \max(\mathbf{i}) \leq d\right\}.$$

Let $\mathcal{B}$ be the Borel σ-algebra on $\mathbb{R}^d$. For any probability measure Q on $(\mathbb{R}^d, \mathcal{B})$ and $\epsilon > 0$, let $N(\mathcal{K}, L_2(Q), \epsilon)$ be the ϵ-covering number for $\mathcal{K}$ using the L^2 norm

with respect to Q, i.e., the minimal number of balls $\{g : \int_{\mathbb{R}^d} |g - \widetilde{g}|^2 dQ \leq \epsilon\}$ needed to cover $\mathcal{K}$. Denote the envelope function $F(x) = \sup_{g\in\mathcal{K}} |g(x)|$. For $\delta > 0$, denote $\mathcal{I}(\delta) = f^{-1}([c-\delta/2, c+\delta/2])$. For any $x \in \mathbb{R}^d$ such that $\nabla f(x) \neq 0$ and $s \in \mathbb{R}$, denote

$$\zeta_x(s) = x + \frac{\nabla f(x)}{\|\nabla f(x)\|} s.$$

Note that $\zeta_x(s) = x + \text{sign}(f'(x)) \times s$ when $d = 1$. For $1 \leq p \leq \infty$, denote the L^p space of Lebesgue measurable functions on $\mathbb{R}^d$ by $\mathcal{L}^p$.

Assumptions:

(F1) f is a ν times continuously differentiable pdf for some $\nu \geq 2$. The density function and all of its first to ν-th derivatives are bounded on $\mathbb{R}^d$. We also assume $C_\ell < c < C_u$, where $C_\ell = \inf_{x\in\mathbb{R}^d} f(x)$ and $C_u = \sup_{x\in\mathbb{R}^d} f(x)$, and that there exist $\delta_0 > 0$ and $\epsilon_0 > 0$ such that $\|\nabla f(x)\| > \epsilon_0$ for all $x \in \mathcal{I}(2\delta_0)$.

(G1) g is a non-negative continuous function on $\mathbb{R}^d$ and there exist $p \geq 0$ and a bounded positive function $g^{(p)}(x)$ on $\mathcal{M}$ such that as $s \to 0$,

$$\sup_{x\in\mathcal{M}} \left| \frac{g(\zeta_x(s))}{|s|^p} - g^{(p)}(x) \right| = o(1).$$

(K1) K is a symmetric product kernel function of νth order for some $\nu \geq 2$. Also $K \in \mathcal{L}^1 \cap \mathcal{L}^\infty$.

(K2) K is two times continuously differentiable. We require that $\|F\|_\infty < \infty$ and for some $C_0 > 0$ and $\eta > 0$,

$$\sup_Q N(\mathcal{K}, L_2(Q), \epsilon\|F\|_\infty) \leq C_0 \epsilon^{-\eta},$$

for $0 < \epsilon < 1$, where the supremum is taken over all the probability measures Q on $(\mathbb{R}^d, \mathcal{B})$.

Remark 3.1.

a) In assumption (F1), the global smoothness requirement for f can be weakened to only hold regionally on $\mathcal{I}(2\delta_0)$, if we choose to use a kernel function K with bounded support. Conditions similar to $\|\nabla f(x)\| > \epsilon_0$ for $x \in \mathcal{I}(2\delta_0)$ in assumption (F1) have appeared in Cadre (2006), Cuevas et al. (2006), Mammen and Polonik (2013), among others. It excludes the possibility of "flat parts" around the level set. In particular, it implies that $\mathcal{M} = \partial\mathcal{L}$, which is a compact $(d-1)$-dimensional C^1 submanifold in $\mathbb{R}^d$ (see Theorem 2 in Walther (1997)). In the case $d = 1$, $\mathcal{M}$ is a collection of separated points, i.e., there exist $x_1, \cdots, x_N$ for some positive integer N such that $\mathcal{M} = \{x_i : i = 1, 2, \cdots, N\}$.

b) An assumption similar to (G1) has appeared in Mason and Polonik (2009). Below we give the specific forms of $g^{(p)}$ for some usual functions g.

   (i) If g is a continuous function with positive values on $\mathcal{M}$, then $p = 0$ and $g^{(p)}(x) = g(x)$, $x \in \mathcal{M}$. Examples include $g(x) \equiv 1$ and $g(x) = f(x)$.

(ii) If $g(x) = f(x)^r |f(x) - c|^q$ for some $q > 0$ and any $r \geq 0$, then $p = q$ and $g^{(q)}(x) = c^r \|\nabla f(x)\|^q$, $x \in \mathcal{M}$. This is because for $x \in \mathcal{M}$, as $s \to 0$,

$$\begin{aligned} & g(\zeta_x(s)) \\ =& [f(x + s \times \nabla f(x)/\|\nabla f(x)\|)]^r \times |f(x + s \times \nabla f(x)/\|\nabla f(x)\|) - c|^q \\ =& [c + o(s)]^r \times |s\|\nabla f(x)\| + o(s)|^q \\ =& c^r |s|^q \|\nabla f(x)\|^q + o(|s|^q). \end{aligned}$$

c) For assumption (K1), it is known that using higher order kernels (i.e. $\nu > 2$), together with higher order smoothness assumptions can reduce the bias in kernel density estimation. But higher order kernels are avoided sometimes because it is possible that density estimators have negative values (see, e.g., Silverman, 1986, page 69). However, this should be of less concern for level set estimation, because the negative values of the density estimate are not (directly) involved in our level set estimator for $c > 0$.

d) Assumption (K2) is imposed to uniformly control the stochastic variation of the kernel density estimator and their derivatives around the expectations. Similar assumptions have appeared in Giné and Guillou (2002), and Einmahl and Mason (2005). Also see Chen et al. (2017). For sufficient conditions for (K2) to hold, see e.g., Nolan and Pollard (1987). In particular, the Gaussian kernel and many usual kernels with bounded support satisfy assumption (K2).

3.2. Asymptotic expressions of $\mathbb{E}\boldsymbol{\lambda}_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}})$ and $\boldsymbol{\lambda}_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}})$

For $x \in \mathcal{M}$, let

$$t_n(x) = \operatorname*{argmin}_t \left\{ |t| : \ \zeta_x(t) \in \widehat{\mathcal{M}} \right\}.$$

For $d \geq 2$, once we establish the normal compatibility between $\widehat{\mathcal{M}}$ and $\mathcal{M}$, the inverse mapping of the normal projection $\pi_{\mathcal{M}}$ will be well defined and is denoted by P_n. Namely, for any $x \in \mathcal{M}$, we have $P_n(x) \in \widehat{\mathcal{M}}$ and $P_n(x) - x$ is orthogonal to the tangent space of $\mathcal{M}$ at x. We also write $\widehat{\mathcal{M}} = P_n(\mathcal{M})$. Since $\nabla f(x)$ is a normal vector of $\mathcal{M}$ at x, we can write

$$P_n(x) = \zeta_x(t_n(x)), \tag{3.1}$$

for some unique $t_n(x) \in \mathbb{R}$. For $d = 1$, P_n is set to be equivalent to $\pi_{\widehat{\mathcal{M}}}$, i.e. it maps points in $\mathcal{M}$ to their closest points in $\widehat{\mathcal{M}}$.

Let $\{A_i : \ i = 1, 2, \cdots, N_d\}$ be a partition of $\mathcal{M}$ and a_i be a point on A_i. Since pointwisely t_n is small when n is large, the following approximation is heuristic when the partition is fine enough:

$$\lambda(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}}) \approx \sum_{i=1}^{N_d} |t_n(a_i)| \mathcal{H}(A_i) \approx \int_{\mathcal{M}} |t_n(x)| d\mathcal{H}(x).$$

The more precise form of the above idea is given in the following theorem, where we need the assumption on n and $\mathbf{h}$ as below.

(H1) The bandwidth (vector) $\mathbf{h} \in \mathbb{R}_+^d$ is dependent on n such that

$$(\log n)^{-1} n h_1 \cdots h_d \|\mathbf{h}\|^4 \to \infty \text{ and } \log(1/\|\mathbf{h}\|)/(\log\log n) \to \infty,$$

as $n \to \infty$. When $d \geq 2$, we assume $h_i \asymp h_j$, for $1 \leq i, j \leq d$.

Theorem 3.1. *Under assumptions (K1), (K2), (F1), (G1) and (H1), we have*

$$\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}}) = \frac{1}{p+1} \int_{\mathcal{M}} \frac{g^{(p)}(x)}{\|\nabla f(x)\|^{p+1}} |\widehat{f}(x) - f(x)|^{p+1} d\mathcal{H}(x)\{1 + o_p(1)\}. \quad (3.2)$$

Remark 3.2.

a) This result is related to but different from Theorem 2.1 in Cadre (2006), where assumptions are imposed to ensure that $\sqrt{nh^d}\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}}) \to \mu_g$ in probability for some $\mu_g > 0$ as $n \to \infty$. In particular the bandwidth is assumed to be small enough that the bias in the kernel density estimation can be ignored. In contrast, our focus is on revealing the asymptotic expression of $\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}})$ for the purpose of finding the optimal bandwidth, for which both the variance and bias in the kernel density estimation are involved.

b) The assumption $(\log n)^{-1} n h_1 \cdots h_d \|\mathbf{h}\|^4 \to \infty$ in this theorem is used to guarantee the normal compatibility between $\mathcal{M}$ and $\widehat{\mathcal{M}}$ for $d \geq 2$, and can in fact be relaxed and replaced with $(\log n)^{-1} n h^3 \to \infty$ for $d = 1$, which is required for the uniform consistency of the kernel estimation for the first derivative of the density. As indicated in Section 2, $\mathcal{H}_0$ is the cardinality measure. For $d = 1$, with $\mathcal{M} = \{x_i : \ i = 1, 2, \cdots, N\}$ (see the discussion after the assumptions), the result (3.2) becomes

$$\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}}) = \frac{1}{p+1} \sum_{i=1}^{N} \frac{|\widehat{f}(x_i) - f(x_i)|^{p+1}}{|f'(x_i)|^{p+1}} g^{(p)}(x_i)\{1 + o_p(1)\}.$$

The required assumption of $(\log n)^{-1} n h_1 \cdots h_d \|\mathbf{h}\|^4 \to \infty$ is critical in the above theorem. However, if we only consider the expectation of $\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}})$, it is in fact not needed. We modify (H1) into the following weaker assumption.

(H2) The bandwidth (vector) $\mathbf{h} \in \mathbb{R}_+^d$ is dependent on n such that

$$(\log n)^{-1} n h_1 \cdots h_d \to \infty \text{ and } \log(1/\|\mathbf{h}\|)/(\log\log n) \to \infty,$$

as $n \to \infty$. When $p \geq 4$ where p appears in assumption (G1), we further assume that $(\log n)^{-(p-2)} n h_1 \cdots h_d \to \infty$. When $d \geq 2$, we assume $h_i \asymp h_j$, for $1 \leq i, j \leq d$.

Let $s_n > 0$ be such that

$$s_n^2 = \frac{1}{n h_1 \cdots h_d} \|K\|_2^2 c, \quad (3.3)$$

$$\text{and } \beta_{\mathbf{h}}(x) = \frac{1}{\nu!}\kappa_\nu \sum_{k=1}^{d} h_k^\nu f_{(k*\nu)}(x). \tag{3.4}$$

Notice that $\beta_{\mathbf{h}}(x) = O(\|\mathbf{h}\|^\nu)$ if the boundedness of the ν-th derivatives of f is assumed (see F1). It is known (see, e.g., Wand and Jones, 1995; also see (6.15) and (6.16) in the proof) that under regularity conditions the bias for kernel density estimator at $x \in \mathcal{M}$ is

$$\mathbb{E}\widehat{f}(x) - f(x) = \beta_{\mathbf{h}}(x) + o(\|\mathbf{h}\|^\nu), \tag{3.5}$$

and variance is

$$\mathrm{Var}(\widehat{f}(x)) = s_n^2(1 + o(1)). \tag{3.6}$$

We have the following theorem.

Theorem 3.2. *Under assumptions (K1), (K2), (F1), (G1), (H2), we have*

$$\begin{aligned}
&\mathbb{E}\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}}) \\
=&\frac{1}{1+p}\mathbb{E}\int_{\mathcal{M}} \frac{g^{(p)}(x)}{\|\nabla f(x)\|^{p+1}} \,|s_n\, Z + \beta_{\mathbf{h}}(x)|^{p+1}\, d\mathcal{H}(x) + o(s_n^{p+1} + \|\mathbf{h}\|^{\nu(p+1)}),
\end{aligned} \tag{3.7}$$

$$=\frac{1}{1+p}\mathbb{E}\int_{\mathcal{M}} \frac{g^{(p)}(x)}{\|\nabla f(x)\|^{p+1}} |\widehat{f}(x) - f(x)|^{p+1} d\mathcal{H}(x) + o(s_n^{p+1} + \|\mathbf{h}\|^{\nu(p+1)}). \tag{3.8}$$

where Z is a standard normal random variable.

Remark 3.3. Using the symmetry of Z's distribution, we have

$$\mathbb{E}\,|s_n\, Z + \beta_{\mathbf{h}}(x)|^{p+1} = \mathbb{E}\,|s_n\, Z + |\beta_{\mathbf{h}}(x)||^{p+1} \geq \max[s_n^{p+1}\mathbb{E}(|Z|^{p+1}), |\beta_{\mathbf{h}}(x)|^{p+1}].$$

Also see (6.28) in the proof for a lower bound. So the first terms on the right-hand sides of (3.7) and (3.8) are indeed leading terms, if $|\beta_{\mathbf{h}}(x)|/\|\mathbf{h}\|^\nu$ is not zero for all $x \in \mathcal{M}$.

Notice that $g \equiv g^{(p)}$ when $p = 0$ in assumption (G1). By observing the fact for any $a \in \mathbb{R}$,

$$\mathbb{E}|Z - a| = |a|\mathbb{P}(|Z| \leq |a|) + \sqrt{\frac{2}{\pi}} e^{-a^2/2} = \gamma(|a|), \tag{3.9}$$

where

$$\gamma(u) = \sqrt{\frac{2}{\pi}} \left(u \int_0^u e^{-t^2/2} dt + e^{-u^2/2} \right),\ u \geq 0,$$

we have the following corollary which gives an exact asymptotic expression of $\mathbb{E}\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}})$ when $p = 0$, the example including $g \equiv 1$ and $g = f$. The result is comparable to Theorem 1 in Devroye and Györfi (1985, page 78), where they considered the MIAE as the risk for kernel density estimation.

Corollary 3.1. *Suppose $p = 0$ in assumption (G1). Under assumptions (K1), (K2), (F1), (G1), (H2), we have*

$$\mathbb{E}\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}}) = \int_{\mathcal{M}} \frac{s_n \gamma(|\beta_{\mathbf{h}}(x)|/s_n)}{\|\nabla f(x)\|} g(x) d\mathcal{H}(x) + o\left(\|\mathbf{h}\|^{\nu} + \frac{1}{\sqrt{nh_1 \cdots h_d}}\right). \tag{3.10}$$

Remark 3.4.

a) One can obtain asymptotic lower and upper bounds for the risk in (3.10), following similar arguments as in the proof of Theorem 2 in Devroye and Györfi (1985, page 79), or in Holmström and Klemelä (1992, page 257). For example, for the upper bound, since $\gamma(u) \le u + \sqrt{2/\pi}$, $u \ge 0$, we have

$$\begin{aligned} &\mathbb{E}\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}}) \\ &\le \left[\int_{\mathcal{M}} \frac{|\beta_{\mathbf{h}}(x)|\, g(x)}{\|\nabla f(x)\|} d\mathcal{H}(x) + \sqrt{\frac{2}{\pi}} s_n \int_{\mathcal{M}} \frac{g(x)}{\|\nabla f(x)\|} d\mathcal{H}(x)\right] + o(\|\mathbf{h}\|^{\nu} + s_n). \end{aligned} \tag{3.11}$$

The minimization of the upper bound leads to an approximation to the asymptotically optimal bandwidth, as the closed form of the minimizer for the leading term in (3.10) is difficult to obtain. See Devroye and Györfi (1985, page 107) for a similar suggestion. In the case $h_1 = \cdots = h_d = h$, the leading term in the above upper bound can be analytically minimized with respect to h. Using a numerical method following the ideas in Hall and Wand (1988), where minimizing the MIAE of kernel density estimation is considered, it is also possible to find an asymptotic optimal bandwidth selector tailored for the level set estimation by minimizing $\mathbb{E}\int_{\mathcal{M}} \frac{|\hat{f}(x) - f(x)|}{\|\nabla f(x)\|} d\mathcal{H}(x)$.

b) If we specifically choose $g = f$ and $\nu = 2$, then the result in this corollary is similar to Theorem 2.1 in Doss and Weng (2018), where they consider the selection of a bandwidth matrix for level set estimation. In fact Theorem 2.1 in Doss and Weng (2018) can be understood as a special case of this corollary, if their bandwidth matrix is restricted to be diagonal. In this corollary we approximate $\mathbb{E}\lambda_f(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}})$ as an L^1 type of risk, which is a special case of a more general result in Theorem 3.2.

In addition to the case $p = 0$ covered in Corollary 3.1, another interesting scenario is $p = 1$ in assumption (G1), which holds when $g(x) = g_r(x) := f(x)^r |f(x) - c|$ for some $r \ge 0$. Note that the choice of r only impacts up to a constant in the asymptotic form of $\mathbb{E}\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}})$ in Theorem 3.2 when $g = g_r$. This is because $g_r^{(1)}(x) = c^r \|\nabla f(x)\|$ (see the calculation in Remark 3.1 b)(ii)).

We call the quantity $\mathbb{E}\lambda_{g_r}(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}})$ the "excess risk", for $r \ge 0$. This is closely related to the concept of excess risk frequently used in the classification literature (see, e.g. Samworth, 2012). Suppose we have a random pair $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$, where Y is the class label of X. Then the Bayes optimal classifier

is $\psi(x) := \mathbf{I}(\eta(x) \geq \frac{1}{2})$, where $\eta(x) = \mathbb{E}(Y|X = x)$. The misclassification risk is $R(\psi) := \mathbb{P}(\psi(X) \neq Y)$. Given an i.i.d. sample $\mathcal{X}$ with the same distribution as (X, Y), suppose one can find an estimator $\widehat{\eta}(x)$ for $\eta(x)$ and build an empirical classifier $\widehat{\psi}(x) := \mathbf{I}(\widehat{\eta}(x) \geq \frac{1}{2})$. Then the misclassification risk for $\widehat{\psi}$ is $\mathbb{E}R(\widehat{\psi}) = \mathbb{P}(\widehat{\psi}(X) \neq Y)$. The difference $\mathbb{E}R(\widehat{\psi}) - R(\psi)$ is called the excess risk in this binary classification problem and it is well-known that one can write

$$\mathbb{E}R(\widehat{\psi}) - R(\psi) = \frac{1}{2}\mathbb{E}\int_{\mathcal{S}_{\widehat{\eta}}\Delta\mathcal{S}_{\eta}} \left|\eta(x) - \frac{1}{2}\right| dx, \tag{3.12}$$

where $\mathcal{S}_{\eta} = \{x \in \mathbb{R}^d : \ \eta(x) \geq \frac{1}{2}\}$ and $\mathcal{S}_{\widehat{\eta}} = \{x \in \mathbb{R}^d : \ \widehat{\eta}(x) \geq \frac{1}{2}\}$ are level sets of η and $\widehat{\eta}$ at the level $\frac{1}{2}$, respectively. Note that the above expression has the same form as $\mathbb{E}\lambda_{g_0}(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}})$.

3.3. Optimal bandwidth using excess risk

We use the excess risk defined above to find the asymptotic optimal bandwidth for density level set estimation, motivated by the connection to the classification literature. In fact, the excess risk is also studied in the literature of density level set estimation (see, e.g. Rinaldo and Wasserman, 2010). For any measurable set $\mathcal{A} \subset \mathbb{R}^d$, the *excess mass functional* is defined as $\mathcal{E}(\mathcal{A}) = \mathbb{P}(\mathcal{A}) - c\lambda(\mathcal{A})$. It is known that $\mathcal{E}(\mathcal{A})$ is maximized when $\mathcal{A} = \mathcal{L}$. Then it is easy to show that $\mathcal{E}(\mathcal{L}) - \mathbb{E}[\mathcal{E}(\widehat{\mathcal{L}})] = \mathbb{E}\lambda_{g_0}(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}})$. In other words, minimizing the risk $\mathbb{E}\lambda_{g_0}(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}})$ is equivalent to maximizing $\mathbb{E}[\mathcal{E}(\widehat{\mathcal{L}})]$. The excess risk $\mathbb{E}\lambda_{g_0}(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}})$ is "cost-sensitive" (Scott and Davenport, 2006) in the sense that the weight function $g_0(x) = |f(x) - c|$ penalizes more heavily at a point $x \in \mathcal{L} \,\Delta\, \widehat{\mathcal{L}}$, if its density value deviates more from the level c.

One can also understand the excess risk for density level set estimation from a binary classification perspective. Given a random vector $X \sim f$, which is independent of $X_1, \cdots, X_n$, suppose that we would like to find a set $\mathcal{A}$, such that we claim $f(X) \geq c$ when $X \in \mathcal{A}$ and $f(X) < c$ when $X \in \mathcal{A}^{\complement}$, where $\mathcal{A}^{\complement}$ is the complement of $\mathcal{A}$. Define the loss function

$$e_{\mathcal{A}}(x) = [c - f(x)][\mathbf{I}(x \in \mathcal{A}) - \mathbf{I}(x \in \mathcal{A}^{\complement})], \ x \in \mathbb{R}^d. \tag{3.13}$$

Notice that this loss function is related to the excess mass functional through $\int_{\mathbb{R}^d} e_{\mathcal{A}}(x)dx = \mathcal{E}(\mathcal{A}^{\complement}) - \mathcal{E}(\mathcal{A})$. Also it is clear that $\mathcal{A} = \mathcal{L}$ minimizes the risk function $\mathcal{R}(\mathcal{A}) := \mathbb{E}[e_{\mathcal{A}}(X)]$. Note that

$$\mathbb{E}\mathcal{R}(\widehat{\mathcal{L}}) - \mathcal{R}(\mathcal{L}) = 2\mathbb{E}\lambda_{g_1}(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}}),$$

which has a form similar to (3.12). The weight function on the right-hand side of the above equation is g_1, but as indicated below Remark 3.4, $\mathbb{E}\lambda_{g_r}(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}})$ has the same asymptotic form for all $r \geq 0$ up to a constant. The risk $\mathbb{E}\mathcal{R}(\widehat{\mathcal{L}})$ has an "empirical" form

$$\widehat{\mathcal{R}}_n(\widehat{\mathcal{L}}) = \frac{1}{n}\sum_{i=1}^{n} e_{\widehat{\mathcal{L}}}(X_i), \tag{3.14}$$

which can be used to evaluate the performance of a density level set estimator. Note that $e_{\widehat{\mathcal{L}}}(X_i)$ still depends on the unknown f, unlike its counterparts for classification or regression level set (see Willett and Nowak, 2007). Nonetheless, we still use this "empirical" risk function as one of performance metrics in our simulation study, because the density functions are known there.

Our optimal bandwidth for level set estimation is based on the excess risk $\mathbb{E}\lambda_{g_r}(\mathcal{L}\,\Delta\,\widehat{\mathcal{L}})$ for a $r \geq 0$, which is shown to resemble MISE for kernel density estimation in Theorem 3.2, where we take $p = 1$ and $g_r^{(p)}(x) = c^r\|\nabla f(x)\|$ (see Remark 3.1 b)(ii)). The following proposition provides another way of understanding this notion.

Proposition 3.1. *When $g(x) = f(x)^r|f(x)-c|$ for some $r \geq 0$, under assumptions (K1), (K2), (F1), (H2), as $\delta \searrow 0$ we have*

$$\frac{2\delta c^{-r}\mathbb{E}\lambda_g(\mathcal{L}\,\Delta\,\widehat{\mathcal{L}})}{\mathbb{E}\int_{\mathcal{I}(\delta)}|\widehat{f}(x)-f(x)|^2dx} \to 1. \tag{3.15}$$

Following this result we can interpret the excess risk as a limit of the MISE for kernel density estimation constrained in a neighborhood of $\mathcal{M}$. As discussed in the remark after Corollary 3.1, the risk $\mathbb{E}\lambda_g(\mathcal{L}\,\Delta\,\widehat{\mathcal{L}})$ with $g \equiv 1$ or $g = f$ is analogous to MIAE used for kernel density estimation. In comparison with this L^1 type of risk, using the excess risk (which is L^2 type) for bandwidth selection in level set estimation enjoys some mathematical simplicity, similar to MISE for kernel density estimation (see page 16, Wand and Jones, 1995).

In what follows we denote

$$m(\mathbf{h}) = \mathbb{E}\int_{\mathcal{M}}\frac{|\widehat{f}(x)-f(x)|^2}{\|\nabla f(x)\|}d\mathcal{H}(x), \tag{3.16}$$

$$\text{and } \widetilde{m}(\mathbf{h}) = s_n^2\int_{\mathcal{M}}\frac{1}{\|\nabla f(x)\|}d\mathcal{H}(x) + \int_{\mathcal{M}}\frac{\beta_{\mathbf{h}}(x)^2}{\|\nabla f(x)\|}d\mathcal{H}(x). \tag{3.17}$$

The assumptions in Theorem 3.2 guarantee that when $g(x) = f(x)^r|f(x)-c|$ for some $r \geq 0$,

$$\mathbb{E}\lambda_g(\mathcal{L}\,\Delta\,\widehat{\mathcal{L}}) = \frac{1}{2}c^r m(\mathbf{h}) + o(\|\mathbf{h}\|^{2\nu}+s_n^2) = \frac{1}{2}c^r\widetilde{m}(\mathbf{h}) + o(\|\mathbf{h}\|^{2\nu}+s_n^2).$$

Therefore the excess risk can be asymptotically minimized by minimizing $\widetilde{m}(\mathbf{h})$. Note that

$$\begin{aligned}\widetilde{m}(\mathbf{h}) &= \frac{1}{(\nu!)^2}\kappa_\nu^2\sum_{k=1}^d\sum_{l=1}^d h_k^\nu h_l^\nu\int_{\mathcal{M}}\frac{f_{(k*\nu)}(x)f_{(l*\nu)}(x)}{\|\nabla f(x)\|}d\mathcal{H}(x)\\ &\quad + \frac{1}{n\Pi_{j=1}^d h_j}\|K\|_2^2\int_{\mathcal{M}}\frac{c}{\|\nabla f(x)\|}d\mathcal{H}(x)\\ &= \frac{1}{(\nu!)^2}\kappa_\nu^2(\mathbf{h}^\nu)^T A(f)\mathbf{h}^\nu + \frac{cb(f)\|K\|_2^2}{n}\frac{1}{(h_1^\nu h_2^\nu\cdots h_d^\nu)^{1/\nu}},\end{aligned} \tag{3.18}$$

where we denote $\mathbf{h}^\nu = (h_1^\nu, h_2^\nu, \cdots, h_d^\nu)^T$, $b(f) = \int_{\mathcal{M}} \|\nabla f(x)\|^{-1} d\mathcal{H}(x)$ and

$$A(f) = [a_{kl}]_{1\le k,l\le d} \text{ with } a_{kl} = \int_{\mathcal{M}} \frac{f_{(k*\nu)}(x) f_{(l*\nu)}(x)}{\|\nabla f(x)\|} d\mathcal{H}(x).$$

For $\mathbf{u} = (u_1, u_2, \cdots, u_d)^T$, define the function

$$Q(\mathbf{u}; \mathbf{M}, a, \nu) = \frac{1}{(\nu!)^2} \mathbf{u}^T \mathbf{M} \mathbf{u} + \frac{a}{(u_1 u_2 \cdots u_d)^{1/\nu}}. \tag{3.19}$$

Then from (3.18) we can write

$$\widetilde{m}(\mathbf{h}) = Q\left(\mathbf{h}^\nu; \kappa_\nu^2 A(f), \frac{cb(f)\|K\|_2^2}{n}, \nu\right). \tag{3.20}$$

To ensure the uniqueness of the minimizer of $\widetilde{m}(\mathbf{h})$, we impose the following assumption.

> **(F2)** For $d = 1$, we require $A(f) > 0$; for $d \ge 2$, we require that $A(f)$ is positive semi-definite and $\inf_{\mathbf{u}\in\bar{\mathbb{R}}_+^d, \|\mathbf{u}\|\neq 0} \mathbf{u}^T A(f)\mathbf{u}/\|\mathbf{u}\|^2 > 0$, where $\bar{\mathbb{R}}_+^d$ denotes the set of vectors in $\mathbb{R}^d$ with non-negative coordinates.

An assumption similar to (F2) in the kernel regression setting appears in Yang and Tschernig (1999). A density function with linearly dependent νth partial derivatives $\{f_{(k*\nu)}(x) : k = 1, \cdots, d\}$ does not satisfy this assumption. See Yang and Tschernig (1999) for more discussions on the similar assumption.

Let $\mathbf{u}(\mathbf{M}, a, \nu)$ be the vector $\mathbf{u}$ which minimizes $Q(\mathbf{u}; \mathbf{M}, a, \nu)$. Denote the $d \times d$ identity matrix by $\mathbf{I}_d$. We have the following optimization result for $\widetilde{m}(\mathbf{h})$.

Theorem 3.3. *Under assumptions (K1), (F1), and (F2), $\widetilde{m}(\mathbf{h})$ is uniquely minimized by a bandwidth given by*

$$\widetilde{\mathbf{h}}_{opt} = \left(\frac{cb(f)\|K\|_2^2}{\kappa_\nu^2 n}\right)^{1/(d+2\nu)} \mathbf{u}^{1/\nu}(A(f), 1, \nu). \tag{3.21}$$

In addition, assume that f has bounded and continuous $(\nu+2)$ times derivatives and $\int_{\mathbb{R}} |u^{\nu+2}\widetilde{K}(u)|du < \infty$. Then as $n \to \infty$, the bandwidth $\mathbf{h}_{opt}$ which minimizes $m(\mathbf{h})$ satisfies

$$\widetilde{\mathbf{h}}_{opt} = \left\{\mathbf{I}_d + O\left(n^{-2\nu/(d+2\nu)}\right)\right\} \mathbf{h}_{opt}, \tag{3.22}$$

and

$$\widetilde{m}(\widetilde{\mathbf{h}}_{opt}) = \left\{1 + O\left(n^{-2\nu/(d+2\nu)}\right)\right\} m(\mathbf{h}_{opt}). \tag{3.23}$$

Remark 3.5.

a) The result (3.21) also contains the case $d = 1$, which we state explicitly below. For $d = 1$, we write

$$m(h) = \sum_{i=1}^{N} \frac{\mathbb{E}|\widehat{f}(x_i) - f(x_i)|^2}{|f'(x_i)|}, \tag{3.24}$$

$$\text{and } \widetilde{m}(h) = \frac{\|K\|_2^2 c}{nh} \sum_{i=1}^{N} \frac{1}{|f'(x_i)|} + \frac{h^{2\nu}}{(\nu!)^2} \kappa_\nu^2 \sum_{i=1}^{N} \frac{f^{(\nu)}(x_i)^2}{|f'(x_i)|}, \tag{3.25}$$

where $f^{(\nu)}$ is the νth derivative of f. Then $m(h) = \widetilde{m}(h) + o(\frac{1}{nh} + h^4)$. The asymptotic optimal bandwidth is given by

$$\widetilde{h}_{\text{opt}} = Cn^{-\frac{1}{1+2\nu}} \text{ with } C = \left(\frac{c(\nu!)^2 \|K\|_2^2 \sum_{i=1}^{N} |f'(x_i)|^{-1}}{2\nu\kappa_\nu^2 \sum_{i=1}^{N} [f^{(\nu)}(x_i)]^2 |f'(x_i)|^{-1}} \right)^{\frac{1}{1+2\nu}}. \tag{3.26}$$

b) If we have the restriction $h_1 = h_2 = \cdots h_d$ for $\mathbf{h}$, then $\widetilde{\mathbf{h}}_{\text{opt}} = (\widetilde{h}_{\text{opt}}, \cdots, \widetilde{h}_{\text{opt}})^T$ has a closed form with

$$\widetilde{h}_{\text{opt}} = \left(\frac{cd(\nu!)^2 b(f) \|K\|_2^2}{2n\nu\kappa_\nu^2 \sum_{k=1}^{d} \sum_{l=1}^{d} a_{kl}} \right)^{1/(d+2\nu)}.$$

c) In general, for the multivariate case, (3.21) has an analytical expression only when $d = 2$, given by $\widetilde{\mathbf{h}}_{\text{opt}} = (\widetilde{h}_{\text{opt},1}, \widetilde{h}_{\text{opt},2})^T$, where

$$\widetilde{h}_{\text{opt},1} = \left(\frac{c(\nu!)^2 b(f) \|K\|_2^2 a_{22}^{(\nu+1)/(2\nu)}}{2n\nu\kappa_\nu^2 a_{11}^{(\nu+1)/(2\nu)} (a_{11}^{1/2} a_{22}^{1/2} + a_{12})} \right)^{1/(2+2\nu)},$$

$$\text{and } \widetilde{h}_{\text{opt},2} = \left(\frac{a_{11}}{a_{22}} \right)^{1/(2\nu)} \widetilde{h}_{\text{opt},1}.$$

For $d \geq 3$, one has to use numerical methods to find the solution. See Wand and Jones (1994).

Since (3.21) contains unknown quantities, in practice we need to find estimators $\widehat{b}(f)$ and $\widehat{A}(f)$ for $b(f)$ and $A(f)$. Then the asymptotic risk function $\widetilde{m}(\mathbf{h})$ is estimated by

$$\widehat{m}(\mathbf{h}) = Q\left(\mathbf{h}^\nu; \kappa_\nu^2 \widehat{A}(f), \frac{c\widehat{b}(f)\|K\|_2^2}{n}, \nu \right). \tag{3.27}$$

Correspondingly, the plug-in optimal bandwidth becomes

$$\widehat{\mathbf{h}}_{\text{opt}} = \left(\frac{c\widehat{b}(f)\|K\|_2^2}{\kappa_\nu^2 n} \right)^{1/(d+2\nu)} \mathbf{u}^{1/\nu}(\widehat{A}(f), 1, \nu). \tag{3.28}$$

For simplicity, below we assume $\nu = 2$ in (3.20) and (3.21), but our methodology applies to general $\nu \geq 2$. Note that $b(f)$ and $A(f)$ involve the unknowns $\mathcal{M}$, ∇f and $f_{(k,k)}f_{(j,j)}$ for $1 \leq k, j \leq d$, which need to be estimated. Below we discuss our choices of estimators and the relatively rates of convergence of our plug-in bandwidth selectors for $d = 1$ and $d \geq 2$ separately, because the case $d \geq 2$ involves estimation of surface integrals on level sets, whereas the case $d = 1$ only requires point estimation.

We first consider $d = 1$. Recall that $\mathcal{M} = \{x_i : \ i = 1, 2, \cdots, N\}$ for $d = 1$ (see the discussion after the assumptions). Let $\widehat{\mathcal{M}} = \{x : \widehat{f}(x) = c\} = \{\widehat{x}_i : \ i = 1, 2, \cdots, \widehat{N}\}$, where $\widehat{N}$ is the cardinality of $\widehat{\mathcal{M}}$. Also let $\widehat{b}(f) = \sum_{i=1}^{\widehat{N}} |\widehat{f}'(\widehat{x}_i)|^{-1}$ and $\widehat{A}(f) = \sum_{i=1}^{\widehat{N}} [\widehat{f}''(\widehat{x}_i)]^2 |\widehat{f}'(\widehat{x}_i)|^{-1}$. For $d = 1$ and $\nu = 2$, the estimated risk function in (3.27) is

$$\widehat{m}(h) = \frac{\|K\|_2^2 c}{nh} \sum_{i=1}^{\widehat{N}} \frac{1}{|\widehat{f}'(x_i)|} + \frac{h^{2\nu}}{(\nu!)^2} \kappa_\nu^2 \sum_{i=1}^{\widehat{N}} \frac{[\widehat{f}''(x_i)]^2}{|\widehat{f}'(x_i)|}, \tag{3.29}$$

and the plug-in estimator in (3.28) is

$$\widehat{h}_{\text{opt}} = \widehat{C} n^{-\frac{1}{5}} \text{ with } \widehat{C} = \left(\frac{c\|K\|_2^2 \sum_{i=1}^{\widehat{N}} |\widehat{f}'(\widehat{x}_i)|^{-1}}{\kappa_2^2 \sum_{i=1}^{\widehat{N}} [\widehat{f}''(\widehat{x}_i)]^2 |\widehat{f}'(\widehat{x}_i)|^{-1}} \right)^{\frac{1}{5}}. \tag{3.30}$$

Note that in the above estimator we are essentially estimating f, f' and f' using $\widehat{f}$, $\widehat{f}'$ and $\widehat{f}''$. The kernel function K may be replaced by a different one in these estimators. However, for simplicity of notation, we keep using K in what follows. The bandwidths used in the estimators $\widehat{f}$, $\widehat{f}'$ and $\widehat{f}''$ can be chosen separately, for which we denote as $h^{(0)}$, $h^{(1)}$ and $h^{(2)}$, respectively. We propose to use the direct plug-in bandwidths for the kernel density and its first two derivatives as the pilot bandwidths $h^{(0)}$, $h^{(1)}$ and $h^{(2)}$, respectively. See Wand and Jones (1994, 1995), Duong and Hazelton (2003), and Chacón et al. (2011) for details of the direct plug-in strategies. In fact, our pilot bandwidths for $d = 1$ can be chosen following the exact procedure given in Samworth and Wand (2010, page 1777). The following theorem gives the relative rates of convergence of estimating our optimal bandwidth for $d = 1$. Recall that $\widetilde{m}$ given in (3.25) is an asymptotic approximation to the excess risk when $g(x) = |f(x) - c|$ and $\widetilde{h}_{\text{opt}}$ given in (3.26) is a minimizer of $\widetilde{m}$.

Theorem 3.4. *Suppose $d = 1$ and assumptions (F1), (F2), (K1) and (K2) hold with $\nu = 2$. In addition, assume that f has bounded continuous fourth derivatives and K has bounded continuous third derivatives of bounded variation. If $h^{(0)} \asymp n^{-1/5}$, $h^{(1)} \asymp n^{-1/7}$ and $h^{(2)} \asymp n^{-1/9}$, then for $\widehat{h}_{opt}$ in (3.30) and $\widehat{m}$ in (3.29) we have*

$$\widehat{h}_{opt} = \widetilde{h}_{opt} \left\{ 1 + O_p\left(n^{-2/9} \right) \right\}, \tag{3.31}$$

$$\text{and } \widehat{m}(\widehat{h}_{opt}) = \widetilde{m}(\widetilde{h}_{opt}) \left\{ 1 + O_p\left(n^{-2/9} \right) \right\}. \tag{3.32}$$

Remark 3.6.

It is clear from the proof of the above theorem that the relative rates of convergence in (3.31) and (3.32) are mainly determined by the choice of $h^{(2)}$ for the estimator $\widehat{f}''$. If we choose $h^{(0)} = h^{(1)} = h^{(2)} \asymp n^{-1/9}$, then the conclusion in Theorem 3.4 still holds.

Next we consider $d \geq 2$. The asymptotics for $\widehat{\mathbf{h}}_{\text{opt}}$ when $d \geq 2$ involves the estimation of integrals on level sets, which is studied in Qiao (2019). In the literature, estimating the volume of manifolds or surface integrals has been studied in, e.g., Cuevas et al. (2007) and Jiménez and Yukich (2011). We consider plug-in estimators $\widehat{b}(f) = \int_{\widehat{\mathcal{M}}} \|\nabla \widehat{f}(x)\|^{-1} d\mathcal{H}(x)$ and $\widehat{A}(f) = [\widehat{a}_{kl}]_{1\leq k,l\leq d}$, where $\widehat{\mathcal{M}} = \{x \in \mathbb{R}^d : \ \widehat{f}(x) = c\}$ and

$$\widehat{a}_{kl} = \int_{\widehat{\mathcal{M}}} \|\nabla \widehat{f}(x)\|^{-1} \widehat{f}_{(k,k)}(x) \widehat{f}_{(l,l)}(x) d\mathcal{H}(x). \tag{3.33}$$

Similar to the case $d = 1$, we can still use different bandwidths for the estimation of derivatives of different orders in $\widehat{b}(f)$ and $\widehat{A}(f)$. Here for simplicity we choose to use a common bandwidth $\mathbf{h}_{\text{pilot}}$ in $\widehat{b}(f)$ and $\widehat{A}(f)$ for the reason given in the remark after Theorem 3.4. The following theorem is a consequence of Theorem 3.1 in Qiao (2019) by noticing that $\widetilde{\mathbf{h}}_{\text{opt}}$ and $\widetilde{m}(\widetilde{\mathbf{h}}_{\text{opt}})$ are smooth functions of $b(f)$ and $A(f)$, where $\widetilde{m}$ and $\widetilde{\mathbf{h}}_{\text{opt}}$ are given in (3.17) and (3.21), respectively.

Theorem 3.5. *Suppose $d \geq 2$ and assumptions (F1), (F2), (K1) and (K2) hold with $\nu = 2$. In addition, assume that both f and K have continuous four times derivatives, and K has bounded support. Let h_n be a sequence such that $h_n \to 0$ and $(\log n)^{-1} n h_n^{d+4} \to \infty$ as $n \to \infty$. If $\mathbf{h}_{pilot} \asymp h_n$, then*

$$\widehat{\mathbf{h}}_{opt} = \widetilde{\mathbf{h}}_{opt} \left\{1 + O_p\left(\alpha_n\right)\right\}, \tag{3.34}$$

$$\text{and } \widehat{m}(\widehat{\mathbf{h}}_{opt}) = \widetilde{m}(\widetilde{\mathbf{h}}_{opt}) \left\{1 + O_p\left(\alpha_n\right)\right\}, \tag{3.35}$$

where $\alpha_n = \frac{1}{\sqrt{n h_n^5}} + \frac{1}{n h_n^{d+4}} + h_n^2$.

Remark 3.7.

a) To minimize α_n, we choose $\mathbf{h}_{\text{pilot}} \asymp n^{-1/(\max\{9,d+6\})}$, i.e., $\mathbf{h}_{\text{pilot}} \asymp n^{-1/9}$ when $d = 2$; and $\mathbf{h}_{\text{pilot}} \asymp n^{-1/(d+6)}$ when $d \geq 3$. If so, then correspondingly we have $\alpha_n \asymp n^{-2/(\max\{9,d+6\})}$. In practice, we can use $\mathbf{h}^{(1)}$ as $\mathbf{h}_{\text{pilot}}$, which is the direct plug-in optimal bandwidth for estimating the gradient of f, because $\mathbf{h}^{(1)} \asymp n^{-1/(d+6)}$. If so, then we have $\alpha_n \asymp n^{-3/16}$ when $d = 2$; and $\alpha_n \asymp n^{-2/(d+6)}$ when $d \geq 3$.

b) When $d \geq 3$, the computation of the surface integrals in $\widehat{b}(f)$ and $\widehat{A}(f)$ might be quite challenging. Alternatively, for a sequence $\epsilon_n > 0$, we can replace $\widehat{b}(f)$ and $\widehat{a}_{kl}$ in $\widehat{\mathbf{h}}_{\text{opt}}$ by the following two types of estimators using integration over small neighborhoods of $\widehat{\mathcal{M}}$, where we still use $\mathbf{h}_{\text{pilot}}$ as

the pilot bandwidth:

$$(i)\begin{cases}\widehat{b}^*(f)=\frac{1}{2\epsilon_n}\lambda(\widehat{f}^{-1}[c-\epsilon_n,c+\epsilon_n])\\ \widehat{a}^*_{kl}=\frac{1}{2\epsilon_n}\int_{\widehat{f}^{-1}[c-\epsilon_n,c+\epsilon_n]}\widehat{f}_{(k,k)}(x)\widehat{f}_{(l,l)}(x)dx\end{cases},$$

or

$$(ii)\begin{cases}\widehat{b}^\dagger(f)=\frac{1}{2\epsilon_n}\int_{\widehat{\mathcal{M}}\oplus\epsilon_n}\|\nabla\widehat{f}(x)\|^{-1}dx\\ \widehat{a}^\dagger_{kl}=\frac{1}{2\epsilon_n}\int_{\widehat{\mathcal{M}}\oplus\epsilon_n}\|\nabla\widehat{f}(x)\|^{-1}\widehat{f}_{(k,k)}(x)\widehat{f}_{(l,l)}(x)dx\end{cases}.$$

Here ϵ_n controls the width of tubes around $\widehat{\mathcal{M}}$ as domains of integration in these estimators. If we use these two types of estimators, then α_n is replaced by $\alpha_n+\epsilon_n^2$ in Theorem 3.5 under the same condition. Again this is a consequence of Theorem 3.1 in Qiao (2019). Using ϵ_n of the same order of h_n does not increase the previous relative rates of convergence $O_p(\alpha_n)$. For example, if $\mathbf{h}_{\text{pilot}}$ is chosen to be $\mathbf{h}^{(1)}$ as in a), then we can use $\min(\mathbf{h}^{(1)})$, $\max(\mathbf{h}^{(1)})$, or the average of the individual bandwidths in $\mathbf{h}^{(1)}$ as ϵ_n.

4. Simulation results

A simulation study was run to assess the performance of our bandwidth selector $\widehat{\mathbf{h}}_{\text{opt}}$ tailored for level set estimation. We compared the performance of our bandwidth selector with the least square cross validation method (see Rudemo, 1982, and Bowman, 1984), which is an ISE-based selector denoted by $\mathbf{h}_{\text{LSCV}}$, as well as the direct plug-in bandwidth selector (see e.g. Wand and Jones, 1994) denoted by $\mathbf{h}_{\text{DPI}}$. Note that both $\mathbf{h}_{\text{LSCV}}$ and $\mathbf{h}_{\text{DPI}}$ are bandwidth selectors for kernel density estimation. In order to make a fair comparison with $\widehat{\mathbf{h}}_{\text{opt}}$, $\mathbf{h}_{\text{LSCV}}$ and $\mathbf{h}_{\text{DPI}}$ are also d-dimensional vectors, which correspond to diagonal bandwidth matrices.

We first compared the performance of the three bandwidth selectors by considering a Gaussian mixture model with the distribution

$$\frac{2}{3}\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1/4&0\\0&1\end{pmatrix}\right)+\frac{1}{3}\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix},\frac{1}{50}\begin{pmatrix}1/4&0\\0&1\end{pmatrix}\right),\tag{4.1}$$

which has a sharp mode and was constructed to represent a bivariate analog to density 4 in Marron and Wand (1992). The levels of the density functions in our analysis were chosen corresponding to the 20%, 50% and 80% HDRs, respectively, that is, $c=c(\tau)$, where $\tau=0.2,0.5$, and 0.8 (see Section 1 for the definition of HDRs). 500 samples were drawn from this distribution, and for each sample we used the error $e(\mathbf{h})=\lambda_{g_0}(\mathcal{L}\Delta\widehat{\mathcal{L}})$, where $g_0(x)=|f(x)-c|$, to evaluate the performance of a density level set estimator with bandwidth $\mathbf{h}$ and the Gaussian kernel.

Figure 3 show the simulation results for the model in (4.1) with sample size $n=1000$. It can be seen that our bandwidth selector $\widehat{\mathbf{h}}_{\text{opt}}$ performed better than

$\mathbf{h}_{\mathrm{LSCV}}$ and $\mathbf{h}_{\mathrm{DPI}}$ in terms of the error $e(\mathbf{h})$ for most of the samples. The improvement of $\widehat{\mathbf{h}}_{\mathrm{opt}}$ was statistically significant at the 0.1% level for $\tau = 0.2, 0.5$, and 0.8 when the Wilcoxon tests were applied to the ratio of errors given by $e(\mathbf{h}_{\mathrm{LSCV}})/e(\widehat{\mathbf{h}}_{\mathrm{opt}})$ and $e(\mathbf{h}_{\mathrm{DPI}})/e(\widehat{\mathbf{h}}_{\mathrm{opt}})$. For each τ value, among the 500 samples, we chose the one with the ratio of errors $e(\mathbf{h}_{\mathrm{DPI}})/e(\widehat{\mathbf{h}}_{\mathrm{opt}})$ closest to the median as a representative.

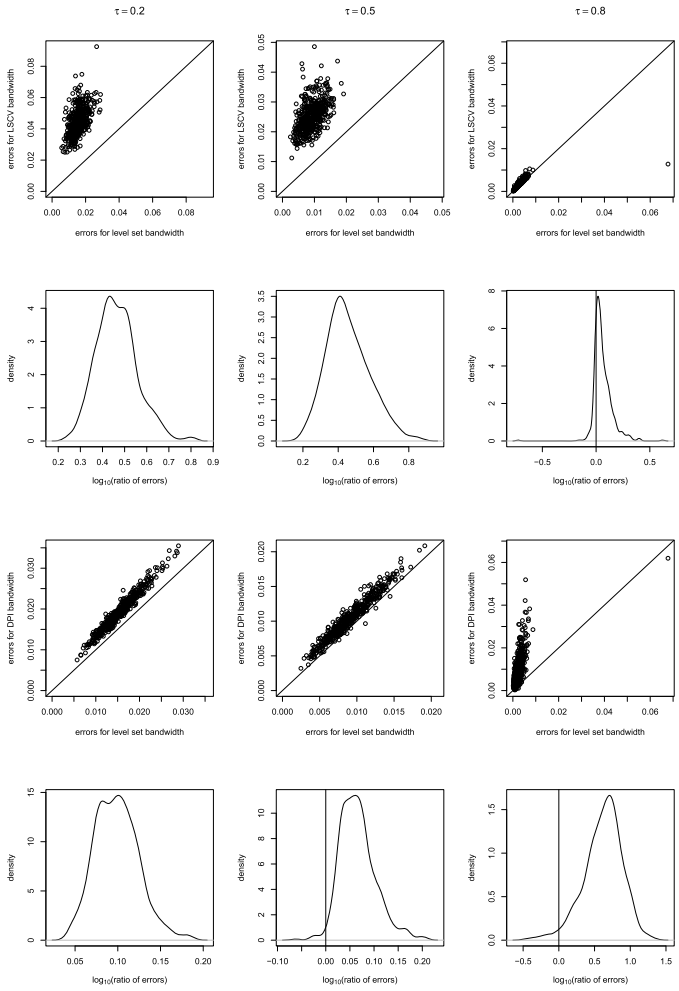

FIG 3. *Graphical comparison of the performance between* $\widehat{\mathbf{h}}_{opt}$ *and* $\mathbf{h}_{LSCV}$ *and between* $\widehat{\mathbf{h}}_{opt}$ *and* $\mathbf{h}_{DPI}$ *for* $\tau = 0.2, 0.5, 0.8$ *for the model in (4.1), with sample size of* $n = 1000$ *for 500 replications. The graphs in the first row show the scatter plots of the errors* $e(\mathbf{h})$ *for* $\widehat{\mathbf{h}}_{opt}$ *and* $\mathbf{h}_{LSCV}$*, and the graphs in the second row show the kernel density estimates of the common logarithm of ratios between the errors using* $\mathbf{h}_{LSCV}$ *and the errors using* $\widehat{\mathbf{h}}_{opt}$*. The third and fourth rows are similar comparisons between* $\widehat{\mathbf{h}}_{opt}$ *and* $\mathbf{h}_{DPI}$*.*

Figure 4 visually compares the level set estimations between the three bandwidth selectors for the representative samples for $\tau = 0.2, 0.5$, and 0.8. It can be seen that when $\widehat{\mathbf{h}}_{\text{opt}}$ or $\mathbf{h}_{\text{DPI}}$ were used, the level sets were estimated reasonably well, with $\widehat{\mathbf{h}}_{\text{opt}}$ slightly better, while using $\mathbf{h}_{\text{LSCV}}$ only captured the level sets for $\tau = 0.8$. When we decreased the sample size to $n = 500$, $\widehat{\mathbf{h}}_{\text{opt}}$ still performed better than $\mathbf{h}_{\text{LSCV}}$ for $\tau = 0.2$, 0.5 and 0.8, and better than $\mathbf{h}_{\text{DPI}}$ for $\tau = 0.2$ and 0.8 but not for $\tau = 0.5$. With this smaller sample size $\widehat{\mathcal{M}}$ using the pilot bandwidth had about 17% chance to be an empty set in the replications for $\tau = 0.8$, which corresponds to a relatively high density level, and in these cases our bandwidth selector was not computable and so we had set $e(\mathbf{h}) = \lambda_g(\mathcal{L})$. This issue arises because the kernel density estimator underestimates the density in a neighborhood of the modes on average, when a second order kernel is used (see the expansion of the bias in (3.5)). When the level c is relatively high and the sample size n is small, we suspect that using a higher order kernel or a more sophisticated pilot bandwidth in the pilot density estimate might make an improvement on this issue.

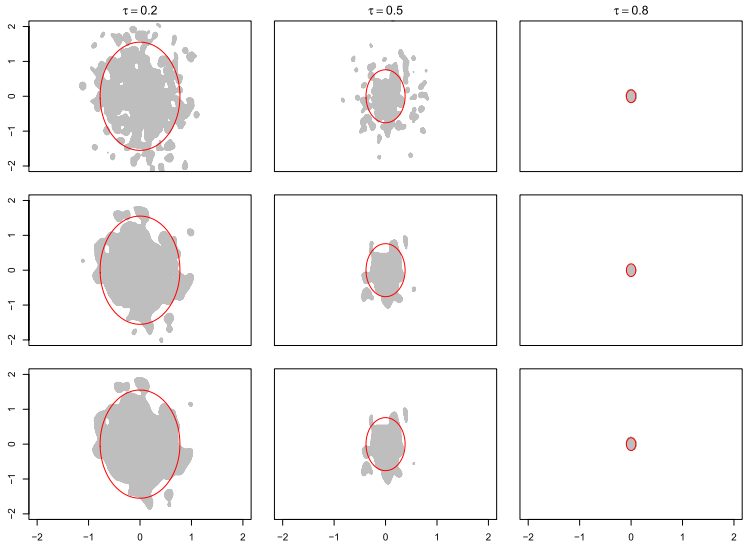

FIG 4. *Comparisons among the estimated level sets using* $\mathbf{h}_{LSCV}$ *(upper panels),* $\mathbf{h}_{DPI}$ *(middle panels) and* $\widehat{\mathbf{h}}_{opt}$ *(lower panels) for* $\tau = 0.2, 0.5, 0.8$ *for the model in (4.1). Estimated level sets are represented by the gray areas, and the true level sets are enclosed by the red curves. The samples were chosen such that the ratios between the errors using* $\mathbf{h}_{DPI}$ *and the errors using* $\widehat{\mathbf{h}}_{opt}$ *are closest to their medians in the 500 replications.*

In addition, we also considered 12 bivariate Gaussian mixture models used in Wand and Jones (1993), which cover from unimodal to quadrimodal models. We further extended these density functions to their trivariate counterparts in our simulation study as specified below. Denote a bivariate Gaussian mixture model

with $k \geq 1$ components by $\sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$, where for some $-1 < \rho_i < 1$,

$$\mu_i = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \end{pmatrix}, \text{ and } \Sigma_i = \begin{pmatrix} \sigma_{i1}^2 & \rho_i \sigma_{i1} \sigma_{i2} \\ \rho_i \sigma_{i1} \sigma_{i2} & \sigma_{i2}^2 \end{pmatrix}.$$

Its trivariate extension is $\sum_{i=1}^{k} w_i \mathcal{N}(\widetilde{\mu}_i, \widetilde{\Sigma}_i)$, where

$$\widetilde{\mu}_i = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \\ \mu_{i2} \end{pmatrix}, \text{ and } \widetilde{\Sigma}_i = \begin{pmatrix} \sigma_{i1}^2 & \rho_i \sigma_{i1} \sigma_{i2} & \rho_i \sigma_{i1} \sigma_{i2} \\ \rho_i \sigma_{i1} \sigma_{i2} & \sigma_{i2}^2 & \rho_i \sigma_{i2} \sigma_{i2} \\ \rho_i \sigma_{i1} \sigma_{i2} & \rho_i \sigma_{i2} \sigma_{i2} & \sigma_{i2}^2 \end{pmatrix}.$$

In other words, the third marginal means and variances replicate the second ones for all the components and the correlation coefficients remain the same. If $\widetilde{\Sigma}_i$ is not positive definite by this extension, we replaced ρ_i by its half, which makes the covariance matrices of all the components positive definite.

These 12 models and their extensions are used to compare the performance of the three bandwidth selectors in the following four cases, where case 1 can be viewed as a base case, and in cases 2–4 we consider the variation of the risk criteria, the dimensions d, and the orders of the kernel function ν, respectively. Recall that $e(\mathbf{h})$ denotes a error metric for density level set estimation with bandwidth $\mathbf{h}$.

Case 1: d=2, $\nu = 2$, using $e(\mathbf{h}) = \lambda_{g_0}(\mathcal{L} \Delta \widehat{\mathcal{L}})$;

Case 2: d=2, $\nu = 2$, using $e(\mathbf{h}) = \widehat{\mathcal{R}}_n(\widehat{\mathcal{L}})$ as defined in (3.14);

Case 3: d=2, $\nu = 4$, using $e(\mathbf{h}) = \lambda_{g_0}(\mathcal{L} \Delta \widehat{\mathcal{L}})$;

Case 4: d=3, $\nu = 2$, using $e(\mathbf{h}) = \lambda_{g_0}(\mathcal{L} \Delta \widehat{\mathcal{L}})$.

Before we show the simulation results, we give some details in the implementation. The Gaussian kernel was used for cases 1, 2, and 4 (i.e., $\nu = 2$). For case 3, the fourth-order kernel function is chosen to be $K(x_1, x_2) = \widetilde{K}(x_1)\widetilde{K}(x_2)$ with $\widetilde{K}(v) = \frac{1}{2}(3 - v^2)\phi(v)$, $v \in \mathbb{R}$, where ϕ is the pdf of a standard normal distribution. We used $\mathbf{h}^{(1)}$ as the pilot bandwidth $\mathbf{h}_{\text{pilot}}$ for $\nu = 2$, which has been discussed in Remark 3.7 a). For $\nu = 4$, while we still use $\mathbf{h}^{(1)}$ to estimate $\widehat{f}$ and $\nabla \widehat{f}$, we use $\mathbf{h}^{(2)}$ as the pilot bandwidth to estimate the fourth derivatives of f, where $\mathbf{h}^{(2)}$ is the direct plug-in optimal bandwidth for estimating the Hessian of f. Our bandwidth selector $\widehat{\mathbf{h}}_{\text{opt}}$ involves the calculation of line/surface integrals. The numerical approximation to line integrals on curves when $d = 2$ are straightforward (for cases 1, 2, and 3). For case 4, we generated meshes with fine triangulation and used the corresponding Riemann sums to approximate the surface integrals. As indicated in Remark 3.7 b), these surface integrals can also be approximated by integration over some small neighborhoods of the surfaces.

For each of the distributions, random sampling was replicated for 500 times. Again we used $c = c(\tau)$ with $\tau = 0.2, 0.5$, and 0.8 as the levels of the density functions. For each case, we have 36 combinations of the τ values and models. The sample sizes were chosen to be $n = 1{,}000$, $n = 2{,}000$, and $n = 10{,}000$. We applied the one-sided Wilcoxon tests to the ratios of errors given by $e(\mathbf{h}_{\text{LSCV}})/e(\widehat{\mathbf{h}}_{\text{opt}})$ and $e(\mathbf{h}_{\text{DPI}})/e(\widehat{\mathbf{h}}_{\text{opt}})$, respectively. Table 1 below summarizes the counts of scenarios when the improvement of $\widehat{\mathbf{h}}_{\text{opt}}$ was not statistically significant at the 0.1% levels for $\tau = 0.2, 0.5$, and 0.8.

TABLE 1
Simulation results

		$\widehat{\mathbf{h}}_{\text{opt}}$ vs. $\mathbf{h}_{\text{LSCV}}$				$\widehat{\mathbf{h}}_{\text{opt}}$ vs. $\mathbf{h}_{\text{DPI}}$			
		$\tau = 0.2$	$\tau = 0.5$	$\tau = 0.8$	Total	$\tau = 0.2$	$\tau = 0.5$	$\tau = 0.8$	Total
case 1	n=1000	4	3	2	9	10	8	2	20
	n=2000	3	3	2	8	9	8	0	17
	n=10000	3	1	1	5	6	2	1	9
case 2	n=1000	3	3	1	7	10	8	1	19
	n=2000	2	3	1	6	9	8	1	18
	n=10000	3	1	1	5	7	3	0	10
case 3	n=1000	6	4	3	13	7	5	1	13
	n=2000	4	3	2	9	7	3	0	10
	n=10000	3	1	0	4	3	1	0	4
case 4	n=1000	11	10	4	25	11	11	2	24
	n=2000	6	5	1	12	7	7	0	14
	n=10000	3	3	1	7	7	3	1	11

Overall we find our bandwidth selector $\widehat{\mathbf{h}}_{\text{opt}}$ performs better than $\mathbf{h}_{\text{LSCV}}$ and $\mathbf{h}_{\text{DPI}}$ for density level set estimation, especially when the sample size is moderately large. Between the two competitors $\mathbf{h}_{\text{LSCV}}$ and $\mathbf{h}_{\text{DPI}}$, in general a larger sample size is needed for $\widehat{\mathbf{h}}_{\text{opt}}$ to outperform the latter, though the needed sample size can be reduced by using higher order kernels as shown in case 3. Also it appears $\widehat{\mathbf{h}}_{\text{opt}}$ performs well for high density levels, while we need larger sample sizes for the asymptotics for $\widehat{\mathbf{h}}_{\text{opt}}$ to show effect when the levels are low. Note that data of large sample sizes are available for many application areas of density level set estimation, such as flow cytometry (Naumann and Wand, 2009) and astronomical survey (Jang, 2006).

5. Discussion

In this paper we give asymptotic L^p approximations of $\lambda_g(\mathcal{L} \Delta \widehat{\mathcal{L}})$ and $\mathbb{E}\lambda_g(\mathcal{L} \Delta \widehat{\mathcal{L}})$, where p is determined by the local behavior of g around $\mathcal{M}$. In particular, when $g(x) = f(x)^r |f(x) - c|$ for some $r \geq 0$, the excess risk $\mathbb{E}\lambda_g(\mathcal{L} \Delta \widehat{\mathcal{L}})$ has an L^2 approximation and is used to select bandwidth for density level set estimation. Numerical results verify that our bandwidth selectors tailored for level set estimation outperforms the lease square cross validation and the direct plug-in bandwidth selectors for density estimation, when the sample size is moderately large.

As indicated in the Introduction section, the work in Doss and Weng (2018) is related to some of the results in this paper, and they have given a comparison between their work with an earlier arXiv version of this paper. When focusing on the level set estimation, they only consider $g = f$ in the asymptotic approximation for $\mathbb{E}\lambda_g(\mathcal{L} \Delta \widehat{\mathcal{L}})$ (which is an L^1 type of risk as interpreted in our Corollary 3.1), and use it as a risk function for bandwidth selection for density level set estimation. We give the expressions of the asymptotic forms of both $\lambda_g(\mathcal{L} \Delta \widehat{\mathcal{L}})$ and $\mathbb{E}\lambda_g(\mathcal{L} \Delta \widehat{\mathcal{L}})$ for a general class of g, which allows us to inter-

pret them as asymptotic L^p type of loss and risk, depending on a property of g given in assumption (G1). In these approximations higher order kernel functions are also allowed if higher smoothness of the density function is assumed. Our bandwidth selection is based on an L^2 type of risk (the excess risk), which corresponds to a specific choice of g in our general result. The excess risk resembles the MISE for kernel density estimation, and is more tractable than the L^1 type of risk. Note that in order to study the theory for the minimization of the L^1 type of risk (when $g = f$), Doss and Weng (2018) assume that the density function f is unimodal and symmetric (see their Corollary 2.1). By contrast, the minimization of the excess risk does not require such assumptions on the shape of the density functions, and our optimal bandwidth can have analytical forms, depending on the structure of the bandwidth matrix (see Remark 3.5). Also as indicated in Remark 3.4, if one uses the L_1 type of risk (e.g. $g = f$ or $g \equiv 1$) for bandwidth selection, minimizing one of its upper bounds is an alternative approach, which can give closed-form solutions (again depending on the structure of the bandwidth matrix).

Note that $g(x) = f(x)^r|f(x) - c|$ for $r \geq 0$ is chosen to be used mainly because its close connection to the excess risk in the classification literature and its interpretation as local MISE given in Proposition 3.1. One can also use $g(x) = \|\nabla f(x)\| f(x)^r|f(x) - c|$, which adds the norm of the gradient as a weight into the integrand. If so then we have a simple approximation $\mathbb{E}\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}}) \approx \frac{1}{2}c^r\mathbb{E}\int_{\mathcal{M}} |\widehat{f}(x) - f(x)|^2 d\mathcal{H}(x)$ by Theorem 3.2. Using this risk, the asymptotic optimal bandwidth has a simpler form because $b(f)$ and a_{kl} in $\widetilde{m}(\mathbf{h})$ in (3.18) will be replaced by $\mathcal{H}(\mathcal{M})$ and $\int_{\mathcal{M}} f_{(k*\nu)}(x) f_{(l*\nu)}(x) d\mathcal{H}(x)$, respectively. In other words, one does not need to estimate the first derivatives in the plug-in bandwidth selector. With this form of the surface integrals, one may use U-statistic type estimators to improve the relative rates of convergence given in Theorem 3.5. See Theorem 2.1 and Corollary 3.1 in Qiao (2019).

In this paper we focus on the selection of bandwidth vectors, which correspond to diagonal bandwidth matrices. We expect that our results can be extended to full unconstrained bandwidth matrices, which might work better for level set estimation. See, e.g. Chacón and Duong (2010). A closely related question is to select bandwidths for the estimation of HDR. Here the risk criterion can be set as $\mathbb{E}\lambda_g(\mathcal{L}_{c(\tau)} \,\Delta\, \widehat{\mathcal{L}}_{\widehat{c}(\tau)})$ (see the Introduction section). Doss and Weng (2018) use $g = f$ and thus obtain an L_1 approximation of this risk criterion. It is expected that one can have an L_2 approximation to this risk using $g(x) = f(x)^r|f(x) - c|$ for $r \geq 0$ and select bandwidths using similar ideas in this paper. We leave the exploration of this idea to future work.

6. Proofs

Proof of Theorem 3.1.

We first present the proof for the case $d \geq 2$. The case $d = 1$ is briefly discussed at the end.

By Theorem 2 of Cuevas et al. (2006), we have

$$d_H(\mathcal{M}, \widehat{\mathcal{M}}) = O\left(\|\widehat{f} - f\|_\infty\right). \tag{6.1}$$

With a slight generalization of Theorem 1 of Einmahl and Mason (2005) to the case of individual bandwidth for each dimension in the kernel density estimator, it follows from assumptions (K1), (K2), (F1) and (H1) that

$$\limsup_{n\to\infty} \sqrt{\frac{nh_1\cdots h_d}{\log n}} \sup_{x\in\mathbb{R}^d} |\widehat{f}(x) - \mathbb{E}\widehat{f}(x)| \le \eta_1, \ a.s. \tag{6.2}$$

for some constant η_1. Following a standard derivation for kernel density estimation, we can show that, there exists a constant $\eta_2 > 0$ such that

$$\sup_{x\in\mathbb{R}^d} |\mathbb{E}\widehat{f}(x) - f(x)| \le \eta_2 \|\mathbf{h}\|^\nu. \tag{6.3}$$

Combining (6.1), (6.2) and (6.3), we have that there exists a positive constant η_3 such that with

$$\epsilon_n = \eta_3 \left(\sqrt{\frac{\log n}{nh_1\cdots h_d}} + \|\mathbf{h}\|^\nu\right), \tag{6.4}$$

we have $\mathcal{L} \,\Delta\, \widehat{\mathcal{L}} \subset \mathcal{M} \oplus \epsilon_n$ for n large enough with probability one, and as a result,

$$\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}}) = \int_{\mathbb{R}^d} \mathbf{I}(x \in \mathcal{L} \,\Delta\, \widehat{\mathcal{L}}) g(x) dx = \int_{\mathcal{M}\oplus\epsilon_n} \mathbf{I}(x \in \mathcal{L} \,\Delta\, \widehat{\mathcal{L}}) g(x) dx. \tag{6.5}$$

By the definition of reach for a manifold, for any ϵ with $0 < \epsilon < \rho(\mathcal{M})$, we can write $\mathcal{M} \oplus \epsilon = \{\zeta_x(s) : \ x \in \mathcal{M}, \ |s| \le \epsilon\}$. Then for large n, the map $\zeta(x, s) := \zeta_x(s)$ is a diffeomorphism from $\mathcal{M} \times [-\epsilon_n, \epsilon_n]$ to $\mathcal{M} \oplus \epsilon_n$.

The following derivation uses integration on manifolds, the theory of which can be found in e.g., Guillemin and Pollack (1974, page 168) and Gray (2004, Theorems 3.15 and 4.7). Denote $\mathcal{S}_n = \mathcal{M} \times [-\epsilon_n, \epsilon_n]$. For any $(x, s) \in \mathcal{S}_n$, let $\psi : U \mapsto \mathcal{S}_n$ be a local parameterisation of $\mathcal{S}_n$ around (x, s), where U is an open subset of $\mathbb{R}^d$. Denote function composition $\eta = \zeta \circ \psi$, which is a local parameterisation of $\mathcal{M} \oplus \epsilon_n$ around $\zeta_x(s)$. Note that both ψ and η depend on (x, s) and this dependence has been suppressed in our notation. Let $D\psi$ and $D\eta$ be the Jacobian matrices of ψ and η, respectively. Define the derivative $D\zeta_{(x,s)} : T_{(x,s)}\mathcal{S}_n \mapsto \mathbb{R}$ of ζ at (x, s) by $D\zeta_{(x,s)} = D\eta \circ (D\psi)^{-1}$, where $T_{(x,s)}\mathcal{S}_n = T_x\mathcal{M} \times \mathbb{R}$ is the tangent space of $\mathcal{S}_n$ at (x, s). Following Proposition 6 in Cannings et al. (2017), we have

$$D\zeta_{(x,s)}(v_1, v_2) = (I + sB(x)) \left(v_1 + \frac{\nabla f(x)}{\|\nabla f(x)\|} v_2\right), \tag{6.6}$$

for $v_1 \in T_x(\mathcal{M})$ and $v_2 \in \mathbb{R}$, where $B(x) = \frac{1}{\|\nabla f(x)\|}\left(I - \frac{\nabla f(x)\nabla f(x)^T}{\|\nabla f(x)\|^2}\right)\nabla^2 f(x)$ is also known as the shape operator on level sets (see Qiao, 2019). It then follows from (6.5), (6.6), and the derivation in Section 7.3 of Cannings et al. (2017) that with probability one for n large enough we have

$$\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}}) = \int_{\mathcal{M}} \int_{-\epsilon_n}^{\epsilon_n} \det(I + sB(x))\; g(\zeta_x(s))\; \mathbf{I}(\zeta_x(s) \in \mathcal{L} \,\Delta\, \widehat{\mathcal{L}}) ds d\mathcal{H}(x). \tag{6.7}$$

Since $\det(I + sB(x)) = 1 + o(1)$ as $n \to \infty$, uniformly in $s \in [-\epsilon_n, \epsilon_n]$ and $x \in \mathcal{M}$, we obtain

$$\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}}) = \int_{\mathcal{M}} \int_{-\epsilon_n}^{\epsilon_n} g(\zeta_x(s))\; \mathbf{I}(\zeta_x(s) \in \mathcal{L} \,\Delta\, \widehat{\mathcal{L}}) ds d\mathcal{H}(x)\{1 + o(1)\}. \tag{6.8}$$

For any $x \in \mathcal{M}$, recall that $P_n(x) = \zeta_x(t_n(x)) \in \widehat{\mathcal{M}}$. Using Lemma 1 in Chen et al. (2017), $\mathcal{M}$ and $\widehat{\mathcal{M}}$ are normal compatible and hence P_n is well defined. For n large enough, we have $\text{sign}(f(\zeta_x(s)) - c) = \text{sign}(s)$ and $\text{sign}(\widehat{f}(\zeta_x(s)) - c) = \text{sign}(s - t_n(x))$, for $s \in (-\epsilon_n, \epsilon_n)$. Hence for any $x \in \mathcal{M}$, the event $\zeta_x(s) \in \mathcal{L} \,\Delta\, \widehat{\mathcal{L}}$ is equivalent to $s \in [t_n(x) \wedge 0, t_n(x) \vee 0]$, where $t_n(x) \wedge 0 = \min(t_n(x), 0)$ and $t_n(x) \vee 0 = \max(t_n(x), 0)$. With probability one we have $t_n(x) < \epsilon_n$ for n large enough since $\mathcal{L} \,\Delta\, \widehat{\mathcal{L}} \subset \mathcal{M} \oplus \epsilon_n$ as indicated above. Hence from (6.8) we can write

$$\begin{aligned}\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}}) &= \int_{\mathcal{M}} \int_{-\epsilon_n}^{\epsilon_n} \mathbf{I}(s \in [t_n(x) \wedge 0, t_n(x) \vee 0])\; g\left(\zeta_x(s)\right) ds d\mathcal{H}(x)\{1 + o(1)\} \\ &= \int_{\mathcal{M}} \int_{t_n(x)\wedge 0}^{t_n(x)\vee 0} g\left(\zeta_x(s)\right) ds d\mathcal{H}(x)\{1 + o(1)\}. \end{aligned} \tag{6.9}$$

By Assumption (G1) we have

$$g\left(\zeta_x(s)\right) = g^{(p)}(x)|s|^p\{1 + o(1)\}, \tag{6.10}$$

where $o(1)$ is uniform in (x, s) for $x \in \mathcal{M}$ and $s \in [-\epsilon_n, \epsilon_n]$. Note that

$$\int_{\mathcal{M}} \int_{t_n(x)\wedge 0}^{t_n(x)\vee 0} g^{(p)}(x)|s|^p ds d\mathcal{H}(x) = \frac{1}{p+1} \int_{\mathcal{M}} g^{(p)}(x)|t_n(x)|^{p+1} d\mathcal{H}(x). \tag{6.11}$$

By using the Taylor expansion for $x \in \mathcal{M}$, we have

$$\begin{aligned} 0 = \widehat{f}(P_n(x)) - f(x) =& \widehat{f}(x) - f(x) + \frac{\nabla f(x)^T \nabla \widehat{f}(x)}{\|\nabla f(x)\|} t_n(x) \\ &+ O\left(|t_n(x)|^2 \sup_{x \in \mathcal{M} \oplus \epsilon_n} \|\nabla^2 \widehat{f}(x)\|\right). \end{aligned}$$

It follows that

$$|t_n(x)| = \frac{|\widehat{f}(x) - f(x)|}{\|\nabla f(x)\|}\{1 + o_p(1)\}, \tag{6.12}$$

where $o_p(1)$ is uniform in $x \in \mathcal{M}$. Also see Lemma 2.2 in Qiao (2019). Combining (6.11) and (6.12), we have

$$\begin{aligned}&\int_{\mathcal{M}} \int_{t_n(x)\wedge 0}^{t_n(x)\vee 0} g^{(p)}(x)|s|^p ds d\mathcal{H}(x)\\ =&\frac{1}{p+1}\int_{\mathcal{M}} \frac{g^{(p)}(x)}{\|\nabla f(x)\|^{p+1}} |\widehat{f}(x) - f(x)|^{p+1} d\mathcal{H}(x)\{1+o_p(1)\}. \end{aligned} \tag{6.13}$$

The result (3.7) immediately follows from (6.9), (6.10) and (6.13).

The proof for the case $d = 1$ can be shown by going through a similar procedure as above, but should be simplified with fewer geometric ingredients. Note that (6.5) is still valid for $d = 1$. Then using (6.10) we have

$$\begin{aligned}\lambda_g(\mathcal{L} \,\Delta\, \widehat{\mathcal{L}}) =& \sum_{x\in\mathcal{M}} \int_{t_n(x)\wedge 0}^{t_n(x)\vee 0} g\left(\zeta_x(s)\right) ds\\ =&\frac{1}{p+1}\sum_{x\in\mathcal{M}} \frac{g^{(p)}(x)}{|f'(x)|^{p+1}} |\widehat{f}(x) - f(x)|^{p+1}\{1+o_p(1)\}. \end{aligned}$$ □

Proof of Theorem 3.2.

We only show the proof for the case $d \geq 2$, as the proof is similar and simpler for the case $d = 1$, as shown in the proof of Theorem 3.1. Before we show the main steps in the proof, we give a useful property of the kernel function K under assumption (K1):

$$K \in \mathcal{L}^q, \text{ for all } 1 \leq q \leq \infty. \tag{6.14}$$

To show (6.14), notice that for $1 < q < \infty$,

$$\|K\|_q = \left(\int_{\mathbb{R}^d} |K(x)|\, |K(x)|^{q-1} dx\right)^{1/q} \leq \|K\|_\infty^{(q-1)/q} \|K\|_1^{1/q} < \infty.$$

Step 1. Let $B_n(x) = \mathbb{E}\widehat{f}(x) - f(x)$ and $\sigma_n(x) = \sqrt{\mathrm{Var}(\widehat{f}(x))}$. We will first prove the following facts, which show that s_n^2 and $\beta_{\mathbf{h}}(x)$ are the asymptotic expressions of the variance and bias of kernel density estimation uniformly in a small neighborhood of $\mathcal{M}$.

$$\sup_{x\in\mathcal{M}\oplus\epsilon_n} |B_n(x) - \beta_{\mathbf{h}}(x)| = o(\|\mathbf{h}\|^\nu), \tag{6.15}$$

$$\sup_{x\in\mathcal{M}\oplus\epsilon_n} |\sigma_n^2(x) - s_n^2| = o\left(\frac{1}{nh_1\cdots h_d}\right). \tag{6.16}$$

We first show (6.15). Note that by using the Taylor expansion for $f(x-\mathbf{h}\odot y)$ around $f(x)$, we have

$$B_n(x) = \int_{\mathbb{R}^d} [f(x - \mathbf{h}\odot y) - f(x)] K(y) dy$$

$$= \int_{\mathbb{R}^d} [f(x + \mathbf{h} \odot y) - f(x)] K(y) dy$$
$$= \sum_{\mathbf{i} \in \mathbb{Z}_+^{\nu,d}} \mathbf{h}^{(\mathbf{i})} \int_{\mathbb{R}^d} \int_0^1 \frac{(1-t)^{\nu-1}}{(\nu-1)!} f_{(\mathbf{i})}(x + t\mathbf{h} \odot y) dt \, K(y) y^{(\mathbf{i})} dy.$$

The assumption that K is a νth order symmetric kernel function implies that we can write

$$\beta_{\mathbf{h}}(x) = \int_0^1 \frac{(1-t)^{\nu-1}}{(\nu-1)!} dt \sum_{\mathbf{i} \in \mathbb{Z}_+^{\nu,d}} \left[\mathbf{h}^{(\mathbf{i})} f_{(\mathbf{i})}(x) \int_{\mathbb{R}^d} K(y) y^{(\mathbf{i})} dy \right].$$

Notice that

$$\sup_{x \in \mathcal{M} \oplus \epsilon_n} |B_n(x) - \beta_{\mathbf{h}}(x)|$$
$$\leq \sum_{\mathbf{i} \in \mathbb{Z}_+^{\nu,d}} \mathbf{h}^{(\mathbf{i})} \int_{\mathbb{R}^d} |K(y) y^{(\mathbf{i})}| \int_0^1 \frac{(1-t)^{\nu-1}}{(\nu-1)!} \sup_{x \in \mathcal{M} \oplus \epsilon_n} |f_{(\mathbf{i})}(x + t\mathbf{h} \odot y) - f_{(\mathbf{i})}(x)| dt dy.$$

We then obtain (6.15) by applying the Dominated Convergence Theorem and assumption (F1) to the right-hand side of the above inequality.

Next we show (6.16). For $s \in [-\epsilon_n, \epsilon_n]$ with the same ϵ_n given in (6.4), by using Taylor expansion, we have that

$$\sup_{x \in \mathcal{M}} |f(\zeta_x(s)) - c - s\|\nabla f(x)\|| = O(s^2) = O(\epsilon_n^2). \tag{6.17}$$

Therefore using Taylor expansion again we have

$$\sup_{x \in \mathcal{M} \oplus \epsilon_n} |\sigma_n^2(x) - s_n^2|$$
$$= \sup_{x \in \mathcal{M} \oplus \epsilon_n} \left| \frac{1}{nh_1 \cdots h_d} \int_{\mathbb{R}^d} [f(x - \mathbf{h} \odot y) - c] K^2(y) dy \right.$$
$$\left. - \frac{1}{n} \left[\int_{\mathbb{R}^d} f(x - \mathbf{h} \odot y) K(y) dy \right]^2 \right|$$
$$\leq \frac{1}{nh_1 \cdots h_d} \sup_{x \in \mathbb{R}^d} \|\nabla f(x)\| \left[\epsilon_n + \|\mathbf{h}\| \int_{\mathbb{R}^d} \|y\| K^2(y) dy \right] + \frac{1}{n} \|f\|_\infty^2 \|K\|_1^2$$
$$= o\left(\frac{1}{nh_1 \cdots h_d} \right).$$

To get the above o-term, we have used the fact that

$$\int_{\mathbb{R}^d} \|y\| K^2(y) dy = \int_{\mathcal{B}_{\mathbf{0}}(1)} \|y\| K^2(y) dy + \int_{[\mathcal{B}_{\mathbf{0}}(1)]^{\complement}} \|y\| K^2(y) dy$$

$$\begin{aligned}
&\leq \|K\|_2^2 + \|K\|_\infty \int_{\mathbb{R}^d} \|y\|^\nu |K(y)| dy \\
&\leq \|K\|_2^2 + \|K\|_\infty d^{\nu/2-1} \int_{\mathbb{R}^d} (|y_1|^\nu + \cdots + |y_d|^\nu) |K(y)| dy < \infty,
\end{aligned}$$

where $\mathbf{0}$ is the origin of $\mathbb{R}^d$, and we use (6.14) and the definition of νth order kernels.

Step 2. We prove (3.7) in this step. Note that

$$\begin{aligned}
\mathbb{E}\lambda_g(\mathcal{L}\Delta\widehat{\mathcal{L}}) &= \mathbb{E}\int \mathbf{I}(x \in \mathcal{L}\Delta\widehat{\mathcal{L}}) g(x) dx \\
&= \int \mathbb{P}(x \in \mathcal{L}\Delta\widehat{\mathcal{L}}) g(x) dx \\
&= \int \mathbb{P}(\widehat{f}(x) \geq c > f(x)) g(x) dx + \int \mathbb{P}(f(x) \geq c > \widehat{f}(x)) g(x) dx \\
&= \int_{\mathcal{L}^{\complement}} \mathbb{P}(\widehat{f}(x) \geq c) g(x) dx + \int_{\mathcal{L}} \mathbb{P}(\widehat{f}(x) < c) g(x) dx.
\end{aligned}$$

Since $\mathcal{L}\Delta\widehat{\mathcal{L}} \subset \mathcal{M} \oplus \epsilon_n$ for large n with probability one,

$$\begin{aligned}
&\mathbb{E}\lambda_g(\mathcal{L}\Delta\widehat{\mathcal{L}}) \\
&= \int_{\mathcal{L}^{\complement} \cap (\mathcal{M} \oplus \epsilon_n)} \mathbb{P}(\widehat{f}(x) \geq c) g(x) dx + \int_{\mathcal{L} \cap (\mathcal{M} \oplus \epsilon_n)} \mathbb{P}(\widehat{f}(x) < c) g(x) dx. \qquad (6.18)
\end{aligned}$$

Note that here $\mathcal{L}^{\complement} \cap (\mathcal{M} \oplus \epsilon_n) = \{\zeta_x(s) : \ x \in \mathcal{M}, -\epsilon_n < s < 0\}$ and $\mathcal{L} \cap (\mathcal{M} \oplus \epsilon_n) = \{\zeta_x(s) : \ x \in \mathcal{M}, 0 \leq s < -\epsilon_n\}$. Similar to (6.8), we have

$$\begin{aligned}
\mathbb{E}\lambda_g(\mathcal{L}\Delta\widehat{\mathcal{L}}) = \Bigg[&\int_{\mathcal{M}} \int_{-\epsilon_n}^{0} \mathbb{P}(\widehat{f}(\zeta_x(s)) \geq c) g(\zeta_x(s)) ds d\mathcal{H}(x) \\
&+ \int_{\mathcal{M}} \int_{0}^{\epsilon_n} \mathbb{P}(\widehat{f}(\zeta_x(s)) < c) g(\zeta_x(s)) ds d\mathcal{H}(x) \Bigg] (1 + o(1)). \qquad (6.19)
\end{aligned}$$

For the leading term on the right-hand side of the above expression, it follows from (6.10) that

$$\begin{aligned}
&\int_{\mathcal{M}} \int_{-\epsilon_n}^{0} \mathbb{P}(\widehat{f}(\zeta_x(s)) \geq c) g(\zeta_x(s)) ds d\mathcal{H}(x) \\
&\qquad + \int_{\mathcal{M}} \int_{0}^{\epsilon_n} \mathbb{P}(\widehat{f}(\zeta_x(s)) < c) g(\zeta_x(s)) ds d\mathcal{H}(x) \\
&= \Bigg[\int_{\mathcal{M}} g^{(p)}(x) \int_{-\epsilon_n}^{0} \mathbb{P}(\widehat{f}(\zeta_x(s)) \geq c) |s|^p ds d\mathcal{H}(x) \\
&\qquad + \int_{\mathcal{M}} g^{(p)}(x) \int_{0}^{\epsilon_n} \mathbb{P}(\widehat{f}(\zeta_x(s)) < c) |s|^p ds d\mathcal{H}(x) \Bigg] (1 + o(1)). \qquad (6.20)
\end{aligned}$$

Only focusing on the leading term again, in what follows we perform a sequence of decompositions. In general, we use L_1, L_2, and L_3 to denote dominant terms and R_1, R_2, and R_3 to denote remainder terms. We have

$$\begin{aligned}
&\int_{\mathcal{M}} g^{(p)}(x) \int_{-\epsilon_n}^{0} \mathbb{P}(\widehat{f}(\zeta_x(s)) \geq c)|s|^p ds d\mathcal{H}(x) \\
&\qquad + \int_{\mathcal{M}} g^{(p)}(x) \int_{0}^{\epsilon_n} \mathbb{P}(\widehat{f}(\zeta_x(s)) < c)|s|^p ds d\mathcal{H}(x) \\
&= \int_{\mathcal{M}} g^{(p)}(x) \int_{-\epsilon_n}^{0} \mathbb{P}\left(\frac{\widehat{f}(\zeta_x(s)) - \mathbb{E}\widehat{f}(\zeta_x(s))}{\sigma_n(\zeta_x(s))} \geq \frac{c - \mathbb{E}\widehat{f}(\zeta_x(s))}{\sigma_n(\zeta_x(s))} \right) |s|^p ds d\mathcal{H}(x) \\
&\quad + \int_{\mathcal{M}} g^{(p)}(x) \int_{0}^{\epsilon_n} \mathbb{P}\left(\frac{\widehat{f}(\zeta_x(s)) - \mathbb{E}\widehat{f}(\zeta_x(s))}{\sigma_n(\zeta_x(s))} < \frac{c - \mathbb{E}\widehat{f}(\zeta_x(s))}{\sigma_n(\zeta_x(s))} \right) |s|^p ds d\mathcal{H}(x) \\
&= L_1 + R_1, \qquad (6.21)
\end{aligned}$$

where

$$\begin{aligned}
L_1 = &\int_{\mathcal{M}} g^{(p)}(x) \int_{-\epsilon_n}^{0} \Phi\left(-\frac{c - \mathbb{E}\widehat{f}(\zeta_x(s))}{\sigma_n(\zeta_x(s))} \right) |s|^p ds d\mathcal{H}(x) \\
&+ \int_{\mathcal{M}} g^{(p)}(x) \int_{0}^{\epsilon_n} \Phi\left(\frac{c - \mathbb{E}\widehat{f}(\zeta_x(s))}{\sigma_n(\zeta_x(s))} \right) |s|^p ds d\mathcal{H}(x),
\end{aligned}$$

with Φ the standard normal distribution function. We will show in *step 4* that

$$|R_1| = o(\|\mathbf{h}\|^{\nu(p+1)} + s_n^{p+1}). \qquad (6.22)$$

Using the results given in (6.16) and (6.15), we have

$$L_1 = L_2 + R_2, \qquad (6.23)$$

where

$$\begin{aligned}
L_2 = &\int_{\mathcal{M}} g^{(p)}(x) \int_{-\epsilon_n}^{0} \Phi\left(\frac{s\|\nabla f(x)\| + \beta_{\mathbf{h}}(x)}{s_n} \right) |s|^p ds d\mathcal{H}(x) \\
&+ \int_{\mathcal{M}} g^{(p)}(x) \int_{0}^{\epsilon_n} \Phi\left(\frac{-s\|\nabla f(x)\| - \beta_{\mathbf{h}}(x)}{s_n} \right) |s|^p ds d\mathcal{H}(x).
\end{aligned}$$

We will show in *step 4* that

$$|R_2| = o(\|\mathbf{h}\|^{\nu(p+1)} + s_n^{p+1}). \qquad (6.24)$$

Let $u = s/s_n$. Then we continue to decompose L_2 as follows.

$$L_2 = L_3 + R_3, \qquad (6.25)$$

where

$$L_3 = s_n^{p+1} \int_{\mathcal{M}} g^{(p)}(x) \int_{-\infty}^{0} \Phi\left(u\|\nabla f(x)\| + \frac{\beta_{\mathbf{h}}(x)}{s_n} \right) |u|^p du d\mathcal{H}(x)$$

$$+ s_n^{p+1} \int_{\mathcal{M}} g^{(p)}(x) \int_0^{\infty} \Phi\left(-u\|\nabla f(x)\| - \frac{\beta_{\mathbf{h}}(x)}{s_n}\right) |u|^p du d\mathcal{H}(x).$$

We will show in *step 4* that

$$|R_3| = o(s_n^{p+1}). \tag{6.26}$$

Using integration by parts we can write L_3 as

$$\begin{aligned}
&(-1)^p s_n^{p+1} \int_{\mathcal{M}} g^{(p)}(x) \int_{-\infty}^{0} \Phi\left(u\|\nabla f(x)\| + \frac{\beta_{\mathbf{h}}(x)}{s_n}\right) u^p du d\mathcal{H}(x) \\
&\quad + s_n^{p+1} \int_{\mathcal{M}} g^{(p)}(x) \int_0^{\infty} \Phi\left(-u\|\nabla f(x)\| - \frac{\beta_{\mathbf{h}}(x)}{s_n}\right) u^p du d\mathcal{H}(x) \\
=&\frac{(-1)^{p+1}}{p+1} s_n^{p+1} \int_{\mathcal{M}} g^{(p)}(x)\|\nabla f(x)\| \int_{-\infty}^{0} \phi\left(u\|\nabla f(x)\| + \frac{\beta_{\mathbf{h}}(x)}{s_n}\right) u^{p+1} du d\mathcal{H}(x) \\
&+ \frac{1}{p+1} s_n^{p+1} \int_{\mathcal{M}} g^{(p)}(x)\|\nabla f(x)\| \int_{-\infty}^{0} \phi\left(u\|\nabla f(x)\| + \frac{\beta_{\mathbf{h}}(x)}{s_n}\right) u^{p+1} du d\mathcal{H}(x) \\
=&\frac{1}{p+1} s_n^{p+1} \int_{\mathcal{M}} g^{(p)}(x)\|\nabla f(x)\| \int_{-\infty}^{\infty} \phi\left(u\|\nabla f(x)\| + \frac{\beta_{\mathbf{h}}(x)}{s_n}\right) |u|^{p+1} du d\mathcal{H}(x),
\end{aligned}$$

where ϕ is the pdf of a standard normal distribution. Using the variable transformation $v = u\|\nabla f(x)\| + \beta_{\mathbf{h}}(x)/s_n$, then we have

$$\begin{aligned}
L_3 =&\frac{1}{p+1} s_n^{p+1} \int_{\mathcal{M}} \frac{g^{(p)}(x)}{\|\nabla f(x)\|^{p+1}} \int_{-\infty}^{\infty} \phi(v) \left|v - \frac{\beta_{\mathbf{h}}(x)}{s_n}\right|^{p+1} dv d\mathcal{H}(x) \\
=&\frac{1}{p+1} \int_{\mathcal{M}} \frac{g^{(p)}(x)}{\|\nabla f(x)\|^{p+1}} \int_{-\infty}^{\infty} \phi(v) \left|s_n\, v - \beta_{\mathbf{h}}(x)\right|^{p+1} dv d\mathcal{H}(x) \\
=&\frac{1}{p+1} \mathbb{E} \int_{\mathcal{M}} \frac{g^{(p)}(x)}{\|\nabla f(x)\|^{p+1}} \left|s_n\, Z + \beta_{\mathbf{h}}(x)\right|^{p+1} d\mathcal{H}(x),
\end{aligned} \tag{6.27}$$

where Z is a standard normal random variable, and in the last equality above we have used the symmetry of Z's distribution. Note that

$$\mathbb{E}\left|s_n\, Z + \beta_{\mathbf{h}}(x)\right|^{p+1} \le 2^p s_n^{p+1} \mathbb{E}|Z|^{p+1} + 2^p |\beta_{\mathbf{h}}(x)|^{p+1}. \tag{6.28}$$

Since $\mathcal{M}$ is a compact set and $\sup_{x\in\mathcal{M}} |\beta_{\mathbf{h}}(x)| \le C_0\|\mathbf{h}\|^{\nu}$ for some constant $C_0 > 0$, we obtain that

$$L_3 = O(s_n^{p+1} + \|\mathbf{h}\|^{\nu(p+1)}). \tag{6.29}$$

Then (3.7) follows from (6.18), (6.19), (6.20), (6.21), (6.22), (6.23), (6.24), (6.25), (6.26), (6.27) and (6.29).

Step 3. Now we prove (3.8), which is implied by

$$\mathbb{E} \int_{\mathcal{M}} \frac{g^{(p)}(x)}{\|\nabla f(x)\|^{p+1}} \left|\widehat{f}(x) - f(x)\right|^{p+1} d\mathcal{H}(x)$$

$$=\mathbb{E}\int_{\mathcal{M}}\frac{g^{(p)}(x)}{\|\nabla f(x)\|^{p+1}}\,|s_n\,Z+\beta_{\mathbf{h}}(x)|^{p+1}\,d\mathcal{H}(x)+o(s_n^{p+1}+\|\mathbf{h}\|^{\nu(p+1)}). \quad (6.30)$$

We will apply Lemma 1 in Horváth (1991) (see Lemma A.2 in the appendix) to show (6.30). Let $Y_i(x)=(h_1\cdots h_d)^{-1}K(\mathbf{h}^{-1}\odot(x-X_i))$. Then $\widehat{f}(x)=n^{-1}\sum_{i=1}^n Y_i(x)$ and $Var(Y_i(x))=n\sigma_n^2(x)$. Now with $w=B_n(x)/\sigma_n(x)$, we have

$$\mathbb{E}\left|\widehat{f}(x)-f(x)\right|^{p+1}=\frac{1}{n^{p+1}}\mathbb{E}\left|\sum_{i=1}^n[Y_i(x)-\mathbb{E}Y_i(x)]+n^{1/2}[n^{1/2}\sigma_n(x)]w\right|^{p+1}. \quad (6.31)$$

For $1\le k\le p+3$ and $x\in\mathcal{M}$, using the substitution $u=\mathbf{h}^{-1}\odot(x-y)$ we can write

$$\begin{aligned}\mathbb{E}|Y_1(x)|^k&=(h_1\cdots h_d)^{-k}\int_{\mathbb{R}^d}|K(\mathbf{h}^{-1}\odot(x-y))|^k f(y)dy\\&=(h_1\cdots h_d)^{-(k-1)}\int_{\mathbb{R}^d}|K(u)|^k f(x-\mathbf{h}\odot u)du\\&=(h_1\cdots h_d)^{-(k-1)}c\|K\|_k^k\{1+o(1)\},\end{aligned}$$

where the last step is a consequence of (6.14), assumption (F1), and the Dominated Convergence Theorem. Since $\mathbb{E}Y_1(x)=c+o(1)$ uniformly in $x\in\mathcal{M}$ (see (6.15)), and for $2\le k\le p+3$,

$$\begin{aligned}\sum_{j=0}^k(-1)^{k-j}\binom{k}{j}\mathbb{E}|Y_1(x)|^j|\mathbb{E}Y_1(x)|^{k-j}&\le\mathbb{E}|Y_1(x)-\mathbb{E}Y_1(x)|^k\\&\le\sum_{j=0}^k\binom{k}{j}\mathbb{E}|Y_1(x)|^j|\mathbb{E}Y_1(x)|^{k-j},\end{aligned}$$

we obtain that for $x\in\mathcal{M}$ and $2\le k\le p+3$,

$$\mathbb{E}|Y_1(x)-\mathbb{E}Y_1(x)|^k=(h_1\cdots h_d)^{-(k-1)}c\|K\|_k^k\{1+o(1)\}, \quad (6.32)$$

where $o(1)$ is uniform in $x\in\mathcal{M}\oplus\epsilon_n$.

By applying Lemma A.2 and using (6.15), (6.16), (6.31) and (6.32), there exist positive constants C_1, C_2 and C_3 such that for all $x\in\mathcal{M}$,

$$\begin{aligned}&\left|\mathbb{E}|\widehat{f}(x)-f(x)|^{p+1}-\mathbb{E}\,|\sigma_n(x)Z+B_n(x)|^{p+1}\right|\\\le&\frac{1}{n^{p+1}}C_1\left(1+\left|\frac{B_n(x)}{\sigma_n(x)}\right|^p\right)\\&\times\left\{n^{p/2}[n^{1/2}\sigma_n(x)]^{p-2}\frac{c\|K\|_3^3}{(h_1\cdots h_d)^2}+[n^{1/2}\sigma_n(x)]^{-2}\frac{c\|K\|_{p+3}^{p+3}}{(h_1\cdots h_d)^{p+2}}\right\}\end{aligned}$$

$$
\begin{aligned}
\leq & C_2\left(1+\frac{(nh_1\cdots h_d)^{p/2}}{c^{p/2}\|K\|_2^p}\,|\beta_{\mathbf{h}}(x)|^p\right)\\
& \times\left(\frac{c^{p/2}\|K\|_2^{p-2}\|K\|_3^3}{(nh_1\cdots h_d)^{p/2+1}}+\frac{\|K\|_{p+3}^{p+3}}{c\|K\|_2^4(nh_1\cdots h_d)^{p+1}}\right)\\
\leq & C_3(s_n^{p+2}+s_n^2|\beta_{\mathbf{h}}(x)|^p). \qquad (6.33)
\end{aligned}
$$

Let $\gamma_{n,\mathbf{h}}=s_n+\|\mathbf{h}\|^\nu$. Denote $A_n(x,Z)=\gamma_{n,\mathbf{h}}^{-1}[s_nZ+\beta_{\mathbf{h}}(x)]$ and $D_n(x,Z)=\gamma_{n,\mathbf{h}}^{-1}[(\sigma_n(x)-s_n)Z+B_n(x)-\beta_{\mathbf{h}}(x)]$. Then we can write

$$
\begin{aligned}
& (\gamma_{n,\mathbf{h}})^{-(p+1)}\left|\mathbb{E}\,|\sigma_n(x)Z+B_n(x)|^{p+1}-\mathbb{E}\,|s_n\,Z+\beta_{\mathbf{h}}(x)|^{p+1}\right|\\
= & \left|\mathbb{E}\,|A_n(x,Z)+D_n(x,Z)|^{p+1}-\mathbb{E}\,|A_n(x,Z)|^{p+1}\right|. \qquad (6.34)
\end{aligned}
$$

Using the expressions of s_n and $\beta_{\mathbf{h}}(x)$ in (3.3) and (3.4), we have that for $k=1,\cdots,2(p+1)$,

$$
\mathbb{E}|A_n(x,Z)|^k\leq 2^{k-1}\{|\gamma_{n,\mathbf{h}}^{-1}s_n|^k\mathbb{E}|Z|^k+|\gamma_{n,\mathbf{h}}^{-1}\beta_{\mathbf{h}}(x)|^k\}=O(1).
$$

Similarly, we have that $\mathbb{E}|D_n(x,Z)|^k=o(1)$ for $k=1,\cdots,2(p+1)$, by using (6.15) and (6.16). Therefore, on the one hand,

$$
\begin{aligned}
& \mathbb{E}\,|A_n(x,Z)+D_n(x,Z)|^{p+1}-\mathbb{E}\,|A_n(x,Z)|^{p+1}\\
\leq & \sum_{j=0}^{p}\binom{p+1}{j}\mathbb{E}[|A_n(x,Z)|^j|D_n(x,Z)|^{p+1-j}]\\
\leq & \sum_{j=0}^{p}\binom{p+1}{j}\sqrt{\mathbb{E}[|A_n(x,Z)|^{2j}]\mathbb{E}[|D_n(x,Z)|^{2(p+1-j)}]}=o(1), \qquad (6.35)
\end{aligned}
$$

where we have used the Cauchy-Schwarz inequality. On the other hand,

$$
\begin{aligned}
& |A_n(x,Z)+D_n(x,Z)|^{p+1}\\
\geq & \,||A_n(x,Z)|-|D_n(x,Z)||^{p+1}\\
= & \begin{cases}\sum_{j=0}^{p+1}\binom{p+1}{j}\|A_n(x,Z)|^j[-|D_n(x,Z)|]^{p+1-j} & \text{if } |A_n(x,Z)|\geq|D_n(x,Z)|\\ \sum_{j=0}^{p+1}\binom{p+1}{j}[-|A_n(x,Z)|]^j|D_n(x,Z)|^{p+1-j} & \text{if } |A_n(x,Z)|<|D_n(x,Z)|.\end{cases}
\end{aligned}
$$

Therefore similar to (6.35) we have

$$
\begin{aligned}
& \mathbb{E}\,|A_n(x,Z)+D_n(x,Z)|^{p+1}-\mathbb{E}\,|A_n(x,Z)|^{p+1}\\
\geq & -2\mathbb{E}|D_n(x,Z)|-\sum_{j=0}^{p}\binom{p+1}{j}\mathbb{E}\{|A_n(x,Z)|^j|D_n(x,Z)|^{p+1-j}\}=o(1).
\end{aligned} \qquad (6.36)
$$

It then follows from (6.35) and (6.36) that the right-hand side of (6.34) is of order $o(1)$, which further implies that

$$\left|\mathbb{E}\left|\sigma_n(x)Z + B_n(x)\right|^{p+1} - \mathbb{E}\left|s_n\, Z + \beta_{\mathbf{h}}(x)\right|^{p+1}\right| = o(s_n^{p+1} + \|\mathbf{h}\|^{\nu(p+1)}). \quad (6.37)$$

Then (6.30) and hence (3.8) immediately follow from (6.33), (6.37) and the fact that the $\mathcal{M}$ is compact and $g^{(p)}(x)/\|\nabla f(x)\|^{p+1}$ is bounded on $\mathcal{M}$.

Step 4. We will prove (6.22), (6.24) and (6.26), as required in *step 2*.

We first show the proof of (6.22) for R_1. By the nonuniform Berry-Esseen theorem (c.f. Theorem 14, Petrov 1975, page 125), there exists a constant $C_4 > 0$ such that for all $y \in \mathbb{R}$,

$$\begin{aligned}&\left|\mathbb{P}\left(\frac{\widehat{f}(\zeta_x(s)) - \mathbb{E}\widehat{f}(\zeta_x(s))}{\sigma_n(\zeta_x(s))} \le y\right) - \Phi(y)\right| \\ \le &\frac{C_4\mathbb{E}\left|Y_1(\zeta_x(s)) - \mathbb{E}Y_1(\zeta_x(s))\right|^3}{n^{1/2}\left(\mathbb{E}\left|Y_1(\zeta_x(s)) - \mathbb{E}Y_1(\zeta_x(s))\right|^2\right)^{3/2}(1+|y|)^3}.\end{aligned}$$

It then follows from (6.32) that there exists a constant $C_5 > 0$ such that for all $y \in \mathbb{R}$,

$$\sup_{x\in\mathcal{M}}\sup_{s\in[-\epsilon_n,\epsilon_n]}\left|\mathbb{P}\left(\frac{\widehat{f}(\zeta_x(s)) - \mathbb{E}\widehat{f}(\zeta_x(s))}{\sigma_n(\zeta_x(s))} \le y\right) - \Phi(y)\right| \le \frac{C_5}{\sqrt{nh_1\cdots h_d}(1+|y|^3)}. \quad (6.38)$$

As a result,

$$|R_1| \le \frac{C_5}{\sqrt{nh_1\cdots h_d}}\int_{\mathcal{M}} g^{(p)}(x)\int_{-\epsilon_n}^{\epsilon_n}|s|^p\left(1 + \left|\frac{c - \mathbb{E}\widehat{f}(\zeta_x(s))}{\sigma_n(\zeta_x(s))}\right|^3\right)^{-1} ds d\mathcal{H}(x). \quad (6.39)$$

Note that due to (6.15), (6.16) and (6.17), there exists a positive constant C_6 such that for all $\eta_3\|\mathbf{h}\|^\nu \le |s| \le \epsilon_n$ (where η_3 appears in (6.4)),

$$\inf_{x\in\mathcal{M}}\left|\frac{c - \mathbb{E}\widehat{f}(\zeta_x(s))}{\sigma_n(\zeta_x(s))}\right| \ge \frac{C_6|s|}{s_n}.$$

Plugging this inequality to the right-hand side of (6.39), we obtain

$$|R_1| \le R_{11} + R_{12},$$

where

$$R_{11} = \frac{C_5}{\sqrt{nh_1\cdots h_d}}\int_{\mathcal{M}} g^{(p)}(x)d\mathcal{H}(x)\int_{|s|\le\eta_3\|\mathbf{h}\|^\nu}|s|^p ds$$

$$= \frac{2C_5}{\sqrt{nh_1\cdots h_d}} \int_{\mathcal{M}} g^{(p)}(x) d\mathcal{H}(x) \frac{(\eta_3 \|\mathbf{h}\|^\nu)^{p+1}}{p+1}$$
$$= o(\|\mathbf{h}\|^{\nu(p+1)}), \tag{6.40}$$

and

$$R_{12} = \frac{C_5}{\sqrt{nh_1\cdots h_d}} \int_{\mathcal{M}} g^{(p)}(x) d\mathcal{H}(x) \int_{\eta_3\|\mathbf{h}\|^\nu \le |s| \le \epsilon_n} |s|^p \left(1 + \frac{C_6^3 |s|^3}{s_n^3}\right)^{-1} ds.$$

Using the variable transformation $t = s/s_n$ we have

$$\begin{aligned} R_{12} &= \frac{2C_5 s_n^{p+1}}{\sqrt{nh_1\cdots h_d}} \int_{\mathcal{M}} g^{(p)}(x) d\mathcal{H}(x) \int_{\eta_3\|\mathbf{h}\|^\nu/s_n \le t \le \epsilon_n/s_n} \frac{t^p}{1+C_6^3 t^3} dt \\ &\le \frac{2C_5 s_n^{p+1}}{\sqrt{nh_1\cdots h_d}} \int_{\mathcal{M}} g^{(p)}(x) d\mathcal{H}(x) \int_{0 \le t \le \epsilon_n/s_n} \frac{t^p}{1+C_6^3 t^3} dt \\ &= o(s_n^{p+1} + \|\mathbf{h}\|^{\nu(p+1)}), \end{aligned} \tag{6.41}$$

where the last rate follows from assumption (H2) and

$$\int_{0 \le t \le \epsilon_n/s_n} \frac{t^p}{1+C_6^3 t^3} dt = \begin{cases} O(1), & p = 0, 1 \\ O(\log(\epsilon_n/s_n)), & p = 2 \\ O((\epsilon_n/s_n)^{p-2}), & p \ge 3. \end{cases}$$

With (6.40) and (6.41), we thus get (6.22).

Next we show the proof of (6.24) for R_2. It follows from (6.15), (6.16) and (6.17) that for any $\epsilon > 0$ small enough, there exists $N_0 > 0$ such that for all $n > N_0$ we have that for all $|s| \le \epsilon_n$

$$\begin{aligned} &\sup_{x\in\mathcal{M}} \left| \frac{c - \mathbb{E}\widehat{f}(\zeta_x(s))}{\sigma_n(\zeta_x(s))} - \frac{-s\|\nabla f(x)\| - \beta_{\mathbf{h}}(x)}{s_n} \right| \\ \le &\sup_{x\in\mathcal{M}} \left| \frac{c - \mathbb{E}\widehat{f}(\zeta_x(s))}{\sigma_n(\zeta_x(s))} - \frac{c - \mathbb{E}\widehat{f}(\zeta_x(s))}{s_n} \right| \\ &+ \left| \frac{c - \mathbb{E}\widehat{f}(\zeta_x(s))}{s_n} - \frac{-s\|\nabla f(x)\| - \beta_{\mathbf{h}}(x)}{s_n} \right| \\ \le &\epsilon^2 \left(\frac{\|\mathbf{h}\|^\nu}{s_n} + \frac{|s|}{s_n} \right). \end{aligned}$$

Hence for large n, by possibly decreasing ϵ and increasing η_3 in (6.4) we have

$$\begin{aligned} D_n^+(s,x) &:= \left| \Phi\left(\frac{c - \mathbb{E}\widehat{f}(\zeta_x(s))}{\sigma_n(\zeta_x(s))} \right) - \Phi\left(\frac{-s\|\nabla f(x)\| - \beta_{\mathbf{h}}(x)}{s_n} \right) \right| \\ &\le \begin{cases} 1 & \text{if } 0 \le s \le \frac{\epsilon s_n - \beta_{\mathbf{h}}(x)}{\|\nabla f(x)\|} \\ \epsilon^2 \left(\frac{\|\mathbf{h}\|^\nu}{s_n} + \frac{|s|}{s_n} \right) \phi\left(\frac{s\|\nabla f(x)\| + \beta_{\mathbf{h}}(x)}{2s_n} \right) & \text{if } \frac{\epsilon s_n - \beta_{\mathbf{h}}(x)}{\|\nabla f(x)\|} < s \le \epsilon_n \end{cases}, \end{aligned}$$

and similarly

$$D_n^-(s,x) := \left|\Phi\left(-\frac{c-\mathbb{E}\widehat{f}(\zeta_x(s))}{\sigma_n(\zeta_x(s))}\right)-\Phi\left(\frac{s\|\nabla f(x)\|+\beta_{\mathbf{h}}(x)}{s_n}\right)\right|$$
$$\leq \begin{cases} 1 & \text{if } 0 \geq s \geq \frac{-\epsilon s_n-\beta_{\mathbf{h}}(x)}{\|\nabla f(x)\|} \\ \epsilon^2\left(\frac{\|\mathbf{h}\|^\nu}{s_n}+\frac{|s|}{s_n}\right)\phi\left(\frac{s\|\nabla f(x)\|+\beta_{\mathbf{h}}(x)}{2s_n}\right) & \text{if } \frac{-\epsilon s_n-\beta_{\mathbf{h}}(x)}{\|\nabla f(x)\|} > s \geq -\epsilon_n \end{cases}.$$

Therefore

$$\begin{aligned}|R_2| \leq & \int_{\mathcal{M}} g^{(p)}(x)\left[\int_{-\epsilon_n}^0 D_n^-(s,x)|s|^p ds+\int_0^{\epsilon_n} D_n^+(s,x)|s|^p ds\right]d\mathcal{H}(x)\\ \leq & R_{21}+R_{22}, \end{aligned} \tag{6.42}$$

where

$$R_{21}=\int_{\mathcal{M}} g^{(p)}(x)\int_{\frac{-\epsilon s_n-\beta_{\mathbf{h}}(x)}{\|\nabla f(x)\|}}^{\frac{\epsilon s_n-\beta_{\mathbf{h}}(x)}{\|\nabla f(x)\|}}|s|^p ds d\mathcal{H}(x),$$

and

$$R_{22}=\epsilon^2\int_{\mathcal{M}} g^{(p)}(x)\int_{-\infty}^{\infty}\left(\frac{\|\mathbf{h}\|^\nu}{s_n}+\frac{|s|}{s_n}\right)\phi\left(\frac{s\|\nabla f(x)\|+\beta_{\mathbf{h}}(x)}{2s_n}\right)|s|^p ds d\mathcal{H}(x).$$

Using the variable transformation $v=[s\|\nabla f(x)\|+\beta_{\mathbf{h}}(x)]/s_n$, we get

$$\begin{aligned}R_{21}=&s_n\int_{\mathcal{M}} g^{(p)}(x)\int_{-\epsilon}^{\epsilon}\frac{|s_n v-\beta_{\mathbf{h}}(x)|^p}{\|\nabla f(x)\|^{p+1}}dv d\mathcal{H}(x)\\ \leq&\sum_{j=0}^p\binom{p}{j}s_n^{j+1}\int_{\mathcal{M}} g^{(p)}(x)|\beta_{\mathbf{h}}(x)|^{p-j}\int_{-\epsilon}^{\epsilon}\frac{|v|^j}{\|\nabla f(x)\|^{p+1}}dv d\mathcal{H}(x)\\ =&\sum_{j=0}^p\frac{2}{j+1}\binom{p}{j}s_n^{j+1}\epsilon^{j+1}\int_{\mathcal{M}}\frac{g^{(p)}(x)|\beta_{\mathbf{h}}(x)|^{p-j}}{\|\nabla f(x)\|^{p+1}}d\mathcal{H}(x)\\ \leq&\epsilon C_7\left[(\|\mathbf{h}\|^\nu)^{p+1}+s_n^{p+1}\right], \end{aligned} \tag{6.43}$$

for some $C_7>0$. Using the variable transformation $v=[s\|\nabla f(x)\|+\beta_{\mathbf{h}}(x)]/s_n$ again, we have

$$\begin{aligned}R_{22}=&\epsilon^2 s_n\int_{\mathcal{M}} g^{(p)}(x)\int_{-\infty}^{\infty}\left(\|\mathbf{h}\|^\nu+\frac{|s_n v-\beta_{\mathbf{h}}(x)|}{\|\nabla f(x)\|}\right)\\ &\times\phi\left(\frac{v}{2}\right)\frac{|s_n v-\beta_{\mathbf{h}}(x)|^p}{\|\nabla f(x)\|^{p+1}}dv d\mathcal{H}(x)\end{aligned}$$

$$\begin{aligned}
\leq & \epsilon^2 \sum_{j=0}^{p} \binom{p}{j} s_n^{j+1} \int_{\mathcal{M}} g^{(p)}(x) \frac{|\beta_{\mathbf{h}}(x)|^{p-j}}{\|\nabla f(x)\|^{p+1}} \int_{-\infty}^{\infty} \left(\|\mathbf{h}\|^{\nu} + \frac{s_n |v| + |\beta_{\mathbf{h}}(x)|}{\|\nabla f(x)\|} \right) \\
& \times \phi\left(\frac{v}{2}\right) |v|^j dv d\mathcal{H}(x) \\
\leq & \epsilon^2 C_8 \left[(\|\mathbf{h}\|^{\nu})^{p+2} + s_n^{p+2} \right], \qquad (6.44)
\end{aligned}$$

for some $C_8 > 0$. Then (6.24) immediately follows from (6.42), (6.43) and (6.44).

Next we show the proof of (6.26) for R_3. Note that $\epsilon_n / s_n \to \infty$ as $n \to \infty$ and $\sup_{x \in \mathcal{M}} |\beta_{\mathbf{h}}(x)| \leq C_9 \epsilon_n$ for some positive constant C_9. When n is large enough,

$$\begin{aligned}
|R_3| = & s_n^{p+1} \int_{\mathcal{M}} g^{(p)}(x) \int_{-\infty}^{-\epsilon_n / s_n} \Phi\left(u \|\nabla f(x)\| + \frac{\beta_{\mathbf{h}}(x)}{s_n} \right) |u|^p du d\mathcal{H}(x) \\
& + s_n^{p+1} \int_{\mathcal{M}} g^{(p)}(x) \int_{\epsilon_n / s_n}^{\infty} \Phi\left(-u \|\nabla f(x)\| - \frac{\beta_{\mathbf{h}}(x)}{s_n} \right) |u|^p du d\mathcal{H}(x) \\
\leq & 2 s_n^{p+1} \int_{\mathcal{M}} g^{(p)}(x) \int_{-\infty}^{-\epsilon_n / s_n} \Phi\left[u(\|\nabla f(x)\| + C_9) \right] |u|^p du d\mathcal{H}(x) \\
= & o(s_n^{p+1}).
\end{aligned}$$

Hence (6.26) is proved and here we conclude the proof. □

Proof of Proposition 3.1.

An application of Proposition A.1 in Cadre (2006) (see Lemma A.1 in the appendix) leads to

$$\int_{\mathcal{I}(\delta)} |\widehat{f}(x) - f(x)|^2 dx = \int_{c-\delta/2}^{c+\delta/2} \int_{f^{-1}(\tau)} \frac{|\widehat{f}(x) - f(x)|^2}{\|\nabla f(x)\|} d\mathcal{H}(x) d\tau,$$

for small $\delta > 0$. Using the Lebesgue–Besicovitch theorem (cf. Evans and Gariepy 1992, Theorem 1, Chapter I), we obtain

$$\lim_{\delta \searrow 0} \frac{1}{\delta} \mathbb{E} \int_{\mathcal{I}(\delta)} |\widehat{f}(x) - f(x)|^2 dx = \mathbb{E} \int_{\mathcal{M}} \frac{|\widehat{f}(x) - f(x)|^2}{\|\nabla f(x)\|} d\mathcal{H}(x). \qquad (6.45)$$

Then the assertion follows from Theorem 3.2, where we take $p = 1$ and $g^{(p)}(x) = c^r \|\nabla f(x)\|$ when $g(x) = f(x)^r |f(x) - c|$. See Remark 3.1 b)(ii). □

Proof of Theorem 3.3.

Following simple algebra, we have for any $w > 0$,

$$Q(\mathbf{u}; \mathbf{M}, a, \nu) = a^{2\nu/(d+2\nu)} w^{d/(d+2\nu)} Q(a^{-\nu/(d+2\nu)} w^{\nu/(d+2\nu)} \mathbf{u}; w^{-1} \mathbf{M}, 1, \nu),$$

and correspondingly,

$$\mathbf{u}(\mathbf{M}, a, \nu) = a^{\nu/(d+2\nu)} w^{-\nu/(d+2\nu)} \mathbf{u}(w^{-1} \mathbf{M}, 1, \nu).$$

The expression of the minimizer in (3.21) then follows by noticing (3.20) with $\mathbf{u} = \mathbf{h}^{\nu}$, $w = \kappa_{\nu}^2$, $\mathbf{M} = \kappa_{\nu}^2 A(f)$ and $a = n^{-1} c b(f) \|K\|_2^2$.

We continue to use the above notation in what follows. The argument of the uniqueness of the minimizer uses similar ideas in the proof of Theorem 6 in Yang and Tschernig (1999), which we describe below. When $d = 1$, for positive $\mathbf{u}$,

$$\nabla^2 Q(\mathbf{u}; \mathbf{M}, a, \nu) = \frac{2}{(\nu!)^2}\mathbf{M} + \frac{a(\nu+1)}{\nu^2}\mathbf{u}^{-2-1/\nu} > 0. \tag{6.46}$$

When $d \geq 2$, the Hessian of $Q(\mathbf{u}; \mathbf{M}, a, \nu)$ w.r.t. $\mathbf{u}$ is given by

$$\begin{aligned}&\nabla^2 Q(\mathbf{u}; \mathbf{M}, a, \nu)\\ =&\frac{2}{(\nu!)^2}\mathbf{M} + \frac{a}{\nu^2(u_1u_2\cdots u_d)^{1/\nu}}\begin{pmatrix}(\nu+1)u_1^{-2} & (u_1u_2)^{-1} & \cdots & (u_1u_d)^{-1}\\ (u_1u_2)^{-1} & (\nu+1)u_2^{-2} & \cdots & (u_2u_d)^{-1}\\ \vdots & \vdots & \ddots & \vdots\\ (u_1u_d)^{-1} & (u_2u_d)^{-1} & \cdots & (\nu+1)u_d^{-2}\end{pmatrix},\end{aligned} \tag{6.47}$$

which is positive definite for $\mathbf{u} \in \bar{\mathbb{R}}_+^d$ under assumption (F2). Therefore by (3.20), $\widetilde{m}(\mathbf{h})$ is a strictly convex function of $\mathbf{h}^\nu$ for $\mathbf{h} \in \bar{\mathbb{R}}_+^d$, which implies that there is at most one minimizer of $\widetilde{m}(\mathbf{h})$ in $\bar{\mathbb{R}}_+^d$. Also notice that $\widetilde{m}(\mathbf{h})$ tends to infinity if either $\|\mathbf{h}\| \to \infty$ or $\|\mathbf{h}\| \to 0$. This shows that $\widetilde{m}(\mathbf{h})$ is uniquely minimized by $\widetilde{\mathbf{h}}_{\text{opt}}$.

The result (3.22) follows a standard argument as given in Hall and Marron (1987). We only sketch the proof here. By Fubini's theorem, we have from (3.16) that

$$m(\mathbf{h}) = \int_{\mathcal{M}} \frac{\mathbb{E}|\widehat{f}(x) - f(x)|^2}{\|\nabla f(x)\|^2} d\mathcal{H}(x) = \int_{\mathcal{M}} \frac{\mathrm{Var}(\widehat{f}(x)) + |\mathbb{E}\widehat{f}(x) - f(x)|^2}{\|\nabla f(x)\|^2} d\mathcal{H}(x).$$

Under the assumption that f has bounded and continuous $(\nu+2)$ times derivatives and $\int_{\mathbb{R}} |u^{\nu+2}\widetilde{K}(u)|du < \infty$, we can extend the expansions in (3.5) and (3.6) to the following:

$$\begin{aligned}\mathbb{E}\widehat{f}(x) - f(x) &= \beta_{\mathbf{h}}(x) + \frac{1}{(\nu+2)!}\kappa_{\nu+2}\sum_{k=1}^d h_k^{\nu+2} f_{(k*(\nu+2))}(x) + o(\|\mathbf{h}\|^{\nu+2}),\\ \mathrm{Var}(\widehat{f}(x)) &= s_n^2 + \frac{1}{nh_1\cdots h_d}\left(\frac{1}{2}\alpha(K)\sum_{k=1}^d h_k^2 f_{(k*2)}(x) + o(\|\mathbf{h}\|^2)\right)\\ &\quad - \frac{1}{n}\left(f(x) + \frac{1}{\nu!}k_\nu\sum_{k=1}^d h_k^\nu f_{(k*\nu)}(x) + o(\|\mathbf{h}\|^\nu)\right)^2,\end{aligned}$$

where $\kappa_{\nu+2} = \int_{\mathbb{R}} u^{\nu+2}\widetilde{K}(u)du$ and $\alpha(K) = \int_{\mathbb{R}^d} u_1^2 K(u)^2 du < \infty$. Then we can obtain the following results:

$$m(\mathbf{h}) = \widetilde{m}(\mathbf{h}) + O\left(\frac{1}{n\|\mathbf{h}\|^{d-2}} + \|\mathbf{h}\|^{2\nu+2}\right),$$

$$\nabla m(\mathbf{h}) = \nabla \widetilde{m}(\mathbf{h}) + O\left(\frac{1}{n\|\mathbf{h}\|^{d-1}} + \|\mathbf{h}\|^{2\nu+1}\right),$$
$$\nabla^2 m(\mathbf{h}) = \nabla^2 \widetilde{m}(\mathbf{h}) + O\left(\frac{1}{n\|\mathbf{h}\|^{d}} + \|\mathbf{h}\|^{2\nu}\right),$$
$$\text{and } \nabla^2 \widetilde{m}(\mathbf{h}) = O\left(\frac{1}{n\|\mathbf{h}\|^{d+2}} + \|\mathbf{h}\|^{2\nu-2}\right). \tag{6.48}$$

Using Taylor expansion, we have

$$0 = \nabla m(\mathbf{h}_{\text{opt}}) = \nabla m(\widetilde{\mathbf{h}}_{\text{opt}}) + \left[\int_0^1 \nabla^2 m(s\,\mathbf{h}_{\text{opt}} + (1-s)\,\widetilde{\mathbf{h}}_{\text{opt}})ds\right](\mathbf{h}_{\text{opt}} - \widetilde{\mathbf{h}}_{\text{opt}}),$$

which implies

$$\mathbf{h}_{\text{opt}} - \widetilde{\mathbf{h}}_{\text{opt}} = \left[\nabla^2 \widetilde{m}(\widetilde{\mathbf{h}}_{\text{opt}})\right]^{-1}\left[\nabla \widetilde{m}(\widetilde{\mathbf{h}}_{\text{opt}}) - \nabla m(\widetilde{\mathbf{h}}_{\text{opt}})\right](1 + o(1)),$$

since $\nabla \widetilde{m}(\widetilde{\mathbf{h}}_{\text{opt}}) = 0$ and $\nabla^2 \widetilde{m}(\widetilde{\mathbf{h}}_{\text{opt}})$ is a nonzero scalar when $d = 1$ by (6.46), or a nonsingular matrix when $d \geq 2$ by (6.47). Then immediately we have (3.22) and (3.23) by using (6.48). □

Proof of Theorem 3.4.

By noticing (6.2) and (6.3), we have that with probability one $\widehat{N} = N$ for n sufficiently large (also see Theorem 3.1 in Biau et al., 2007). Hence we do not distinguish between $\widehat{N}$ and N in what follows. Without loss of generality, we assume that $x_1 < \cdots < x_N$ and $\widehat{x}_1 < \cdots < \widehat{x}_N$. Denote the kernel density estimation using bandwidth $h^{(k)}$ by $\widehat{f}_k(x)$, $k = 0, 1, 2$. Also for $\ell = 0, 1, 2, \cdots$, denote the ℓth derivative of a function g on $\mathbb{R}$ by $g^{(\ell)}$ if it exists, including the convention $g^{(0)} \equiv g$.

Under assumption (F1), there exists $b_0 > 0$ such that $|f'(x)| > \epsilon_0$ for $x \in \bigcup_{i=1}^N [x_i - b_0, x_i + b_0]$. By assuming f has bounded continuous fourth derivatives and K has bounded continuous third derivatives of bounded variation, it follows from Lemmas 2 and 3 in Arias-Castro et al. (2016) that for $\widehat{f}$ using bandwidth $h > 0$ and for $\ell = 0, 1, 2$, and 3,

$$\sup_{x \in [x_i - b_0, x_i + b_0]} |\mathbb{E}\widehat{f}^{(\ell)}(x) - f^{(\ell)}(x)| = O\left(h^{\min(4-\ell, 2)}\right), \tag{6.49}$$

$$\sup_{x \in [x_i - b_0, x_i + b_0]} |\widehat{f}^{(\ell)}(x) - \mathbb{E}\widehat{f}^{(\ell)}(x)| = O\left(\sqrt{\frac{\log n}{nh^{1+2\ell}}}\right), \ a.s. \tag{6.50}$$

Due to the uniform consistency result for $\widehat{f}_0$ shown in (6.1), (6.2) and (6.3), with probability one $\widehat{x}_i \in [x_i - b_0, x_i + b_0]$, for $i = 1, \cdots, N$, for n large enough. Let $g_n(x) = \widehat{f}_0(x) - f(x)$. For $i = 1, \cdots, N$, we have

$$g_n(x_i) = O_p((nh^{(0)})^{-1/2} + (h^{(0)})^2) = O_p(n^{-2/5}). \tag{6.51}$$

Since $\widehat{f}_0(\widehat{x}_i) = f(x_i) = c$, we have

$$g_n(\widehat{x}_i) = \widehat{f}_0(\widehat{x}_i) - f(\widehat{x}_i) = f(x_i) - f(\widehat{x}_i) = f'(\widetilde{x}_i)(x_i - \widehat{x}_i),$$

where $\widetilde{x}_i$ is between x_i and $\widehat{x}_i$ by using the Taylor expansion. Note that $\widetilde{x}_i \in [x_i - b_0, x_i + b_0]$, which implies that $|f'(\widetilde{x}_i)| > \epsilon_0$ and yields

$$|x_i - \widehat{x}_i| \le \frac{1}{\epsilon_0}|g_n(\widehat{x}_i)|. \tag{6.52}$$

Another Taylor expansion for $g_n(\widehat{x}_i)$ leads to

$$g_n(\widehat{x}_i) - g_n(x_i) = g_n'(\check{x}_i)(\widehat{x}_i - x_i), \tag{6.53}$$

where $\check{x}_i$ is between $\widehat{x}_i$ and x_i and

$$|g_n'(\check{x}_i)| \le \sup_{x\in[x_i-b_0,x_i+b_0]} |g_n'(x)| = O_p\left(\sqrt{\frac{\log n}{n(h^{(0)})^3}} + (h^{(0)})^2\right) = o_p(1), \tag{6.54}$$

by using (6.49) and (6.50). Then it follows from (6.51), (6.52), (6.53) and (6.54) that

$$|\widehat{x}_i - x_i| \le \frac{1}{\epsilon_0}\left[|g_n(x_i)| + |g_n(\widehat{x}_i) - g_n(x_i)|\right] = O_p(n^{-2/5}). \tag{6.55}$$

Consequently, with $s_n(x) := \widehat{f}_2''(x) - f''(x)$, using (6.49) and (6.50) we have

$$|s_n(\widehat{x}_i) - s_n(x_i)| \le |s_n'(\check{x}_i)|\,|(\widehat{x}_i - x_i)| = o_p(n^{-2/5}), \tag{6.56}$$

where $\check{x}_i$ is between $\widehat{x}_i$ and x_i. With the choice $h^{(2)} = O(n^{-1/9})$,

$$\begin{aligned}&\widehat{f}_2''(\widehat{x}_i) - f''(x_i)\\ =&[s_n(\widehat{x}_i) - s_n(x_i)] + s_n(x_i) + [f''(\widehat{x}_i) - f''(x_i)] = O_p(n^{-2/9}).\end{aligned}$$

Similarly, we have $\widehat{f}_1'(\widehat{x}_i) - f'(x_i) = O_p(n^{-2/7})$. We then have

$$\sum_{i=1}^N [\widehat{f}_2''(\widehat{x}_i)]^2|\widehat{f}_1'(\widehat{x}_i)|^{-1} - \sum_{i=1}^N [f''(x_i)]^2|f'(x_i)|^{-1} = O_p(n^{-2/9}),$$

and

$$\sum_{i=1}^N |\widehat{f}_1'(\widehat{x}_i)|^{-1} - \sum_{i=1}^N |f'(x_i)|^{-1} = O_p(n^{-2/7}).$$

As a result, for C and $\widehat{C}$ given in (3.26) and (3.30), we have

$$\frac{\widehat{C}}{C} - 1 = O_p(h^{-2/9}).$$

and correspondingly we get (3.31) and (3.32). □

Appendix

In this appendix, we collect some known results that are used in the proofs.

Lemma A.1 (Proposition A.1. in Cadre (2006)). *Let $\phi : \mathbb{R}^d \mapsto \mathbb{R}_+$ be a continuously differentiable function such that $\phi(x) \to 0$ as $\|x\| \to \infty$, and $J \subset \mathbb{R}_+$ be an interval such that $\inf J > 0$ and $\inf_{\phi^{-1}(J)} \|\nabla\phi\| > 0$. Then, for all bounded Borel function $g : \mathbb{R}^d \mapsto \mathbb{R}$:*

$$\int_{\phi^{-1}(J)} g d\lambda = \int_J \int_{\phi^{-1}(\{s\})} \frac{g}{\|\nabla\phi\|} d\mathcal{H} ds.$$

Lemma A.2 (Lemma 1 in Horváth (1991)). *Let $Y, Y_1, \cdots, Y_n$ be i.i.d random vectors with $\mathbb{E}Y = \mu$, $var(Y) = \sigma^2$. If $\mathbb{E}(Y)^{p+2} < \infty$, then there is a constant $C = C(p)$ such that for any $w \in \mathbb{R}$,*

$$\begin{aligned}&\left|\mathbb{E}\left|\sum_{i=1}^n (Y_i - \mu) + n^{1/2}\sigma w\right|^p - n^{p/2}\sigma^p \mathbb{E}|Z + w|^p\right| \\ \leq & C(1 + |w|^{p-1})\left[n^{(p-1)/2}\sigma^{p-3}\mathbb{E}|Y - \mu|^3 + \sigma^{-2}\mathbb{E}|Y - \mu|^{p+2}\right],\end{aligned}$$

where Z is a standard normal random variable.

Acknowledgement

The author is grateful to Anand Vidyashankar, an Associate Editor and a referee for their careful reading of an earlier version of the paper and for insightful comments that lead to significant improvements. The research of Wanli Qiao was partially supported by NSF grants DMS 1821154 and FET 1900061, and a Jeffress Memorial Trust Award. The simulations in this work were run on ARGO, a research computing cluster provided by the Office of Research Computing at George Mason University, VA.

References

Arias-Castro, E., Mason, D., and Pelletier, B. (2016). On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *Journal of Machine Learning Research* **17** 1–28. MR3491137

Baíllo, A. (2003). Total error in a plug-in estimator of level sets. *Statistics & Probability Letters* **65** 411–417. MR2039885

Baíllo, A., Cuevas, A., and Justel, A. (2000). Set estimation and nonparametric detection. *The Canadian Journal of Statistics* **28** 765–782. MR1821433

Biau, G., Cadre, B., and Pelletier, B. (2007). A graph-based estimator of the number of clusters. *ESAIM Probab. Stat.* **11** 272–280. MR2320821

Biau, G., Cadre, B., and Pelletier, B. (2008). Exact rates in density support estimation. *J. Multivariate Anal.* **99** 2185–2207. MR2463383

Bowman, A.W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71** 353–360. MR0767163

Cadre, B. (2006). Kernel estimation of density level sets. *J. Multivariate Anal.* **97** 999–1023. MR2256570

Cannings, T.I., Berrett, T.B., and Samworth, R.J. (2017). Local nearest neighbour classification with applications to semi-supervised learning. arXiv:1704.00642.

Chacón, J.E. and Duong, T. (2010). Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *TEST* **19** 375–398. MR2677734

Chacón, J.E., Duong, T., and Wand, M.P. (2011). Asymptotics for general multivariate kernel density derivative estimators. *Statist. Sinica* **21** 807–840. MR2829857

Chazal, F., Lieutier, A., and Rossignac, J. (2007). Normal-map between normal-compatible manifolds. *International Journal of Computational Geometry & Applications* **17** 403–421. MR2362411

Chen, Y.-C., Genovese, C.R., and Wasserman, L. (2017). Density Level Sets: Asymptotics, Inference, and Visualization. *J. Amer. Statist. Assoc.* **112** 1684–1696. MR3750891

Cuevas, A., Fraiman, R., and Rodríguez-Casal, A. (2007). A nonparametric approach to the estimation of lengths and surface areas. *Ann. Statist.* **35** 1031–1051. MR2341697

Cuevas, A., González-Manteiga, W., and Rodríguez-Casal, A. (2006). Plug-in estimation of general level sets. *Australian & New Zealand Journal of Statistics* **48** 7–19. MR2234775

Devroye, L. (1987). *A Course in Density Estimation*. Birkhäuser, Boston. MR0891874

Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View*. Wiley, New York. MR0780746

Devroye, L. and Lugosi, G. (2001). *Combinatorial Methods in Density Estimation*. Springer-Verlag, New York. MR1843146

Doss, C.R. and Weng, G. (2018). Bandwidth selection for kernel density estimators of multivariate level sets and highest density regions. *Electronic Journal of Statistics* **12** 4313–4376. MR3892342

Duong, T. and Hazelton, M.L. (2003). Plug-in bandwidth matrices for bivariate kernel density estimation. *Journal of Nonparametric Statistics* **15** 17–30. MR1958957

Einmahl, U. and Mason, D.M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.* **33** 1380–1403. MR2195639

Evans, L.C. and Gariepy, R.F. (1992). *Measure Theory and Fine Properties of Functions*. CRC Press, Boca Raton, FL. MR1158660

Fasy, B.T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, L. (2014). Confidence sets for persistence diagrams. *Ann. Statist.* **42** 2301–2339. MR3269981

Federer, H. (1959). Curvature measures. *Trans. Amer. Math. Soc.* **93** 418–491. MR0110078

Genovese, C.R., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2012). Minimax manifold estimation. *J. Mach. Learn. Res.* **13** 1263–1291. MR2930639

Giné, E. and Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. H. Poincaré Probab. Statist.* **38**

907–921. MR1955344
Gray, A. (2004). *Tubes*, 2nd ed. *Progress in Mathematics* **221**. Birkhäuser, Basel. MR2024928
Guillemin, V. and Pollack, A. (1974). *Differential Topology*. Prentice-Hall, Englewood Cliffs. MR0348781
Hall, P. and Kang, K.-H. (2005). Bandwidth choice for nonparametric classification. *Ann. Statist.* **33** 284–306. MR2157804
Hall, P. and Marron, J.S. (1987). Extent to which least-squares cross-validation minimizes integrated square error in nonparametric density estimation. *Probab. Theory Related Fields* **74** 567–581. MR0876256
Hall, P., Park, B.U., and Samworth R.J. (2008). Choice of neighbor order in nearest-neighbor classification. *Ann. Statist.* **36** 2135–2152. MR2458182
Hall, P. and Wand, M.P. (1988). Minimizing L_1 distance in nonparametric density estimation. *J. Multivariate Anal.* **26** 59–88. MR0955204
Hartigan, J.A. (1987). Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.* **82** 267–270. MR0883354
Holmström, L. and Klemelä, J. (1992). Asymptotic bounds for the expected L_1 error of a multivariate kernel density estimator. *J. Multivariate Anal.* **42** 245–266. MR1183845
Horváth, L. (1991). On L_p-norms of multivariate density estimations. *Ann. Statist.* **19** 1933–1949. MR1135157
Hyndman, R.J. (1996). Computing and graphing highest density regions. *Amer. Statist.* **50** 120–126.
Jang, W. (2006). Nonparametric density estimation and clustering in astronomical sky surveys. *Computational Statistics & Data Analysis* **50** 760–774. MR2207006
Jiménez, R. and Yukich, J.E. (2011). Nonparametric estimation of surface integrals. *Ann. Statist.* **39** 232–260. MR2797845
Mammen, E. and Polonik, W. (2013). Confidence sets for level sets. *Journal of Multivariate Analysis* **122** 202–214. MR3189318
Marron, J.S. and Wand, M.P. (1992). Exact mean integrated squared error. *Ann. Statist.* **20** 712–736. MR1165589
Mason, D.M. and Polonik, W. (2009). Asymptotic normality of plug-in level set estimates. *The Annals of Applied Probability* **19** 1108–1142. MR2537201
Müller, D.W. and Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.* **86** 738–746. MR1147099
Naumann, U. and Wand, M.P. (2009). Automation in high-content flow cytometry screening. *Cytometry Part A* **75** 789–797.
Niyogi, P., Smale, S., and Weinberger, S. (2008). Finding the homology of submanifolds with high confidence from random samples. *Discrete and Computational Geometry* **39** 419–441. MR2383768
Nolan, D. and Pollard, D. (1987). U-processes: rates of convergences. *Ann. Statist.* **15** 780–799. MR0888439
Petrov, V.V. (1975). *Sums of Independent Random Variables*. Springer, New York. MR0388499
Polonik, W. (1995). Measuring mass concentrations and estimating density

contour clusters – an excess mass approach. *Ann. Statist.* **23** 855–881. MR1345204

Qiao, W. (2018). Asymptotics and optimal bandwidth selection for nonparametric estimation of density level sets. https://arxiv.org/abs/1707.09697v2.

Qiao, W. (2019). Nonparametric estimation of surface integrals on density level sets. arXiv:1804.03601. MR1665715

Qiao, W. and Polonik, W. (2019). Nonparametric confidence regions for level sets: statistical properties and geometry. *Electronic Journal of Statistics* **13**(1) 985–1030. MR3934621

Rigollet, P. and Vert, R. (2009). Optimal rates for plug-in estimators of density level sets. *Bernoulli* **15** 1154–1178. MR2597587

Rinaldo, A. and Wasserman, L. (2010). Generalized density clustering. *Ann. Statist.* **38** 2678–2722. MR2722453

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9** 65–78. MR0668683

Samworth, R.J. (2012). Optimal weighted nearest neighbour classifiers. *Ann. Statist.* **40** 2733–2763. MR3097618

Samworth, R.J. and Wand, M.P. (2010). Asymptotics and optimal bandwidth selection for highest density region estimation. *Ann. Statist.* **38** 1767–1792. MR2662359

Scott, C. and Davenport, M. (2006). Regression Level Set Estimation Via Cost-Sensitive Classification. *IEEE Transactions on Signal Processing* **55**(6) 2752–2757. MR1500201

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London. MR0848134

Sommerfeld, M., Sain, S., and Schwartzman, A. (2015). Confidence regions for excursion sets in asymptotically Gaussian random fields, with an application to climate. arXiv:1501.07000. MR3862360

Steinwart, I., Hush, D., and Scovel, C. (2005). A classification framework for anomaly detection. *J. Machine Learning Research* **6** 211–232. MR2249820

Tsybakov, A.B. (1997). Nonparametric estimation of density level sets. *Ann. Statist.* **25** 948–969. MR1447735

Wand, M.P. and Jones, M.C. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *J. Amer. Statist. Assoc.* **88** 520–528. MR1224377

Wand, M.P. and Jones, M.C. (1994). Multivariate plug-in bandwidth selection. *Comput. Statist.* **9** 97–117. MR1280754

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*. Chapman and Hall, London. MR1319818

Walther, G. (1997). Granulometric smoothing. *Ann. Statist.* **25** 2273–2299. MR1604445

Willett, R. and Nowak, R. (2007). Minimax optimal level-set estimation. *IEEE Trans. Image Process.* **12** 2965–2979. MR2472804

Yang, L. and Tschernig, R. (1999). Multivariate bandwidth selection for local linear regression. *J. R. Statist. Soc. B* **61** 793–815. MR1722240