Electronic Journal of Statistics
Vol. 12 (2018) 4313–4376
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1501

Bandwidth selection for kernel density estimators of multivariate level sets and highest density regions

Charles R. Doss* **and Guangwei Weng†**

School of Statistics
University of Minnesota
Minneapolis, MN 55455
e-mail: cdoss@stat.umn.edu; wengx076@umn.edu

Abstract: We consider bandwidth matrix selection for kernel density estimators of density level sets in $\mathbb{R}^d$, $d \geq 2$. We also consider estimation of highest density regions, which differs from estimating level sets in that one specifies the probability content of the set rather than specifying the level directly. This complicates the problem. Bandwidth selection for KDEs is well studied, but the goal of most methods is to minimize a global loss function for the density or its derivatives. The loss we consider here is instead the measure of the symmetric difference of the true set and estimated set. We derive an asymptotic approximation to the corresponding risk. The approximation depends on unknown quantities which can be estimated, and the approximation can then be minimized to yield a choice of bandwidth, which we show in simulations performs well. We provide an R package **lsbs** for implementing our procedure.

MSC 2010 subject classifications: 62G07.
Keywords and phrases: Level set estimation, highest density region estimation, kernel density estimator, bandwidth selection.

Received June 2018.

Contents

*Supported in part by NSF Grant DMS-1712664
†Supported in part by a University of Minnesota Grant-in-Aid grant

1. Introduction

As computing power has become greater and as data sets have become simultaneously larger and more complicated, demand for statistical methods that are increasingly flexible and data driven has increased. Two related methods for capturing the complex structure of a data set from a true density f_0 are to estimate either the density's *level sets* (LS's) or the density's *highest-density regions* (HDR's). (We will explain the difference between estimating LS's and estimating HDR's shortly.) For a density function f_0 defined on $\mathbb{R}^d$ and a given constant $c > 0$, the c-level set (sometimes known as a *density contour*) of f_0 is $\beta(c) := \{\boldsymbol{x} \in \mathbb{R}^d : f_0(\boldsymbol{x}) = c\}$, and the corresponding super-level set is

$$\mathcal{L}(c) := \{\boldsymbol{x} \in \mathbb{R}^d : f_0(\boldsymbol{x}) \geq c\}. \tag{1}$$

Under some basic regularity conditions, the density super-level set is a set of minimum volume having f_0-probability at least $\int_{\mathcal{L}(c)} f_0(\boldsymbol{x})\, d\boldsymbol{x}$ (Garcia et al., 2003). For this reason, perhaps the most common use for HDR estimation occurs in Bayesian statistics. An HDR of a posterior density is a so-called (minimum volume) credible region, which is one of the most fundamental tools in Bayesian statistics. There are quite a wide range of other applications for estimation of density LS's or density HDR's and these estimation problems have received increasing attention in the statistics and machine learning literatures in recent years. (We consider estimation of density level sets and estimation of density super-level sets to be equivalent tasks.) The applications of LS or HDR estimation include outlier/novelty detection (Lichman and Smyth, 2014; Park, Huang and Ding, 2010), discriminant analysis (Mammen and Tsybakov, 1999) and clustering analysis (Hartigan, 1975; Rinaldo and Wasserman, 2010; Cuevas, Febrero and Fraiman, 2001). LS estimation is one of the fundamental tools in estimation of cluster trees and persistence diagrams, used in topological data analysis (Chen (2017), Wasserman (2016)).

A common way to estimate the density super-level set $\mathcal{L}(c)$ based on independent and identically distributed (i.i.d.) $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ is to replace the density function in (1) with a kernel density estimator (KDE)

$$\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) := \frac{1}{n} \sum_{i=1}^{n} K(\boldsymbol{H}^{-1/2}(\boldsymbol{x} - \boldsymbol{X}_i))|\boldsymbol{H}|^{-1/2}, \tag{2}$$

where $\boldsymbol{H} \in \mathbb{R}^{d\times d}$ is a symmetric positive definite bandwidth matrix and K is a kernel function. This gives us the so-called plug-in estimator

$$\widehat{\mathcal{L}}_{n,\boldsymbol{H}}(c) := \{\boldsymbol{x} \in \mathbb{R}^d : \widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) \geq c\}. \tag{3}$$

We now explain the difference between "LS estimation" and "HDR estimation." Often the level of interest is only specified indirectly through a given probability $\tau \in (0,1)$ which yields a level $f_{\tau,0} := \inf\{y > 0 : \int_{\mathbb{R}^d} f_0(\boldsymbol{x})\mathbb{1}_{\{f(\boldsymbol{x})\geq y\}}\, d\boldsymbol{x} \leq 1-\tau\}$. Then the corresponding super-level set is

$$\mathcal{L}(f_{\tau,0}) := \{\boldsymbol{x} \in \mathbb{R}^d : f_0(\boldsymbol{x}) \geq f_{\tau,0}\}, \tag{4}$$

and the corresponding plug-in estimators are

$$\hat{f}_{\tau,n} := \inf\left\{y \in (0,\infty) : \int_{\mathbb{R}^d} \widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x})\mathbb{1}_{\{\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x})\geq y\}}\, d\boldsymbol{x} \leq 1-\tau\right\}$$

and

$$\widehat{\mathcal{L}}_{n,\boldsymbol{H}}(\hat{f}_{\tau,n}) := \{\boldsymbol{x} \in \mathbb{R}^d : \widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) \geq \hat{f}_{\tau,n}\}. \tag{5}$$

Estimating (4) based on specifying τ is known as the *HDR estimation* problem; this has extra complication over the LS estimation problem because $f_{\tau,0}$ has to be estimated rather than being fixed in advance. Thus we use the phrase *LS estimation* to mean estimation of (1) with c fixed in advance (equivalently, estimation of (4) with $f_{\tau,0}$ fixed). When we use the phrase *HDR estimation* we mean estimation of (4) with τ (but not $f_{\tau,0}$) fixed in advance. Thus, LS's and HDR's are mathematically equivalent, but estimating LS's and estimating HDR's are statistically different tasks.

Early work on LS or HDR estimation includes Hartigan (1987), Müller and Sawitzki (1991), Polonik (1995), Tsybakov (1997), and Walther (1997). Some recent work has focused on asymptotic properties of KDE plug-in estimators, including results about consistency, limit distribution theory, and statistical inference. Baíllo, Cuesta-Albertos and Cuevas (2001) show that the probability content of the plug-in estimator converges to the probability of the true super-level set as the sample size tends to infinity. Baíllo (2003) proves the strong consistency of the plug-in estimator under an integrated symmetric difference metric. Cadre (2006) further obtains the rate of convergence of the plug-in estimator when the loss is given by the generalized symmetric difference of sets. Mason and Polonik (2009) give the asymptotic normality of estimated super-level sets under the same metric as Cadre (2006). Chen, Genovese and Wasserman (2017) find a more practically usable limiting distribution of the plug-in estimator for LS's by using Hausdorff distance as the metric for set difference and provide methods for constructing confidence regions for LS's based on this limiting distribution. Jankowski and Stanberry (2012) and Mammen and Polonik (2013) also investigate the formation of confidence regions for LS's.

It is well known that KDE's are sensitive to the choice of the bandwidth (matrix). The optimal bandwidth (matrix) depends on the objective of estimation.

There are many tools that have been developed for selecting the bandwidth when $d = 1$ or the bandwidth matrix when $d > 1$; these include minimizing an asymptotic approximation to an appropriate risk function, as well as computational methods such as the bootstrap or cross-validation, and are largely focused on globally estimating the density or its derivatives well. A good summary of those methods can be found in Wand and Jones (1995), Sain, Baggerly and Scott (1994a), or Jones, Marron and Sheather (1996).

However, Duong, Koch and Wand (2009, page 505) state that, "a number of practical issues in highest density region estimation, such as good data-driven rules for choosing smoothing parameters, are yet to be resolved." Samworth and Wand (2010) is the only published work we know of that investigates the problem of selecting bandwidths for HDR estimation (and we know of no published works that directly investigate bandwidth selection for LS estimation). Samworth and Wand (2010) study the KDE plug-in estimator when $d = 1$, and show by simulation that the kernel density estimator aiming for HDR estimation can be very different from the one aiming for global density estimation. They also propose an asymptotic approximation to a risk function that is suitable for HDR estimation and a corresponding bandwidth selection procedure based on the approximation, all when $d = 1$.

In this paper, we consider the multivariate setting, where $d \geq 2$. In this case, we are estimating a level set manifold, which involves some added technical difficulties over the case $d = 1$ (in which case the level set is a finite point set), but we believe that LS or HDR estimation when $d \geq 2$ is of great practical interest because of the large variety of complicated structures that multivariate level sets can reveal. We derive asymptotic approximations to a risk function for LS estimation and to a risk function for HDR estimation. We believe that our approximations and derivations will be very valuable for any future procedures that do (either) LS or HDR bandwidth selection. Our calculations shed light on the important quantities relating to LS or HDR estimation. Furthermore, we develop a "plug-in" bandwidth selector method based on minimizing an estimate of the LS or the HDR risk approximation. This approach can be used to optimize over all positive definite bandwidth matrices or over restricted classes of matrices (e.g., diagonal ones). Our theory applies for all $d \geq 2$. We have developed code to implement our bandwidth selector when $d = 2$. It is straightforward to implement a numeric approximation to Hausdorff integrals that appear in our approximations (see Subsection 2.1 for discussion of the Hausdorff measure) when $d = 2$. It is less immediately obvious how to implement such approximations when $d \geq 3$, although we indeed believe that implementation is feasible for such approximations. In fact, we believe that computational feasibility is an important benefit of using a closed-form approximation to the risk, particularly in the multivariate setting that we consider in this paper. As will be discussed later in the paper, many simple problems in the univariate setting are more complicated in the multivariate setting and must be solved by Monte Carlo. Thus performing bootstrap or cross-validation, which involves nested Monte Carlo computations, quickly becomes infeasible.

During the development of the present paper we became aware of the recent

related work, Qiao (2018). Qiao (2018) also considers problems about bandwidth selection for KDE's in settings related to level set estimation. However, the main focus of Qiao (2018) is somewhat different than the one here. In fact, Qiao (2018) states that bandwidth selection for multivariate HDR estimation is "far from trivial" and does not consider this problem. We will discuss the approach taken by Qiao (2018) again in the Discussion section.

The structure of the paper is as follows. We present our two asymptotic risk approximation theorems, as well as corollaries about the risk approximation minimizers, in Section 2. We present methodology to select bandwidth matrices in Section 3. In Section 4 we study the performance of our bandwidth selector in simulation experiments as well as in analysis of two real data sets, the Wisconsin Breast Cancer Diagnostic data and the Banknote Authentication data. We give concluding discussion in Section 5. Proofs of the main results are given in Appendix A, and further details, technical results, and intermediate lemmas are given in Appendix B and Appendix C. Some notation and assumptions are presented in Subsections 2.1 and 2.2.

2. Asymptotic risk results

2.1. Notation

We use the following notation throughout. For a density function f_0 on $\mathbb{R}^d$ and a Borel measurable set $A \subset \mathbb{R}^d$, define the measure $\mu_{f_0}(A) = \int_A f_0(\boldsymbol{x})\, d\boldsymbol{x}$. For a function f on $\mathbb{R}^d$, a measure P, and $1 \leq p < \infty$, we let $\|f\|_{p,P}^p = \int_{\mathbb{R}^d} |f(\boldsymbol{z})|^p dP(\boldsymbol{z})$ if this quantity is finite. If P is Lebesgue measure we abbreviate $\|f\|_{p,P} \equiv \|f\|_p$, $1 \leq p < \infty$. Let $\|f\|_\infty = \sup_{\boldsymbol{z}\in\mathbb{R}^d} |f(\boldsymbol{z})|$, and for a function g with vector or matrix values, that is, $g : \mathbb{R}^d \to \mathbb{R}^{p\times q}$, let $\|g\|_\infty = \max_{1\leq i\leq p, 1\leq j\leq q} \|g_{ij}\|_\infty$. We let $\|\boldsymbol{x}\| = (\sum_{i=1}^d x_i^2)^{1/2}$ for $\boldsymbol{x} \in \mathbb{R}^d$. Let ∇f be the gradient (column) vector of f and let $\nabla^2 f$ be the Hessian matrix $\left(\frac{\partial^2}{\partial x_i \partial x_j}[f]\right)_{i,j}$. Let $\mathcal{H}$ be $d-1$ dimensional Hausdorff measure (Evans and Gariepy, 2015). The Hausdorff measure is useful for measuring the volume of lower dimensional sets, like manifolds, embedded in a higher dimensional ambient space. Let λ denote Lebesgue measure. Recall that $\beta(c) := \{\boldsymbol{x} \in \mathbb{R}^d : f_0(\boldsymbol{x}) = c\}$ and $\mathcal{L}(c) := \{\boldsymbol{x} \in \mathbb{R}^d : f_0(\boldsymbol{x}) \geq c\}$, we let $\mathcal{L}_\tau \equiv \mathcal{L}(f_{\tau,0})$ and $\widehat{\mathcal{L}}_{\tau,\boldsymbol{H}} \equiv \widehat{\mathcal{L}}_{\boldsymbol{H}}(\widehat{f}_{\tau,n})$. We generally use bold to denote vectors. We use "$\equiv$" to denote notational equivalences and "$:=$" or "$=:$" for definitions. Any integral whose domain is not specified explicitly is taken over all of $\mathbb{R}^d$. We will occasionally omit the integrating variable when there's no confusion in doing so. We use $\mathcal{S}$ to denote the set of all $d \times d$ symmetric positive definite matrices. For a symmetric matrix $\boldsymbol{A}$, we use $\lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A})$ to denote the largest and the smallest eigenvalues of $\boldsymbol{A}$ respectively. In this paper, we will use the f_0-probability volume of the symmetric difference as the distance between the true set and its estimator. We use Δ to denote the symmetric difference operation between two sets: for two sets A and B, $A\Delta B := (A\cup B)\backslash(A\cap B)$ where "$\backslash$" is set difference. Figure 1 shows the symmetric difference between the 0.02 super-level set of standard bivariate normal

distribution and an "estimated" super-level set. We let A^c be the complement of a set A. For $\delta > 0$ and $\boldsymbol{x} \in \mathbb{R}^d$, let $B(\boldsymbol{x}, \delta) := \{\boldsymbol{y} \in \mathbb{R}^d : \|\boldsymbol{y} - \boldsymbol{x}\| \leq \delta\}$, and for a set A, let $A^\delta := \cup_{\boldsymbol{x} \in A} B(\boldsymbol{x}, \delta)$.

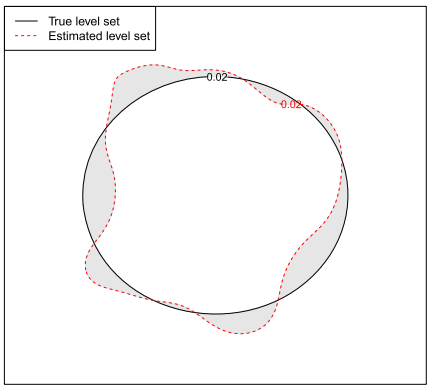

FIG 1. *Symmetric difference between the true level set and an estimated level set. The solid black line is the boundary of the true level set and the dashed red line is the boundary of the estimated level set. The shaded area is the symmetric difference of the two sets.*

2.2. Assumptions

To derive our asymptotic expansion, we make the following basic assumptions on the underlying density, kernel function and bandwidth matrix.

Assumption D1a.

1. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be i.i.d. from a bounded density f_0 on $\mathbb{R}^d$, $d \geq 2$.
2. Fix $\inf_{x \in \mathbb{R}^d} f_0(x) < c < \|f_0\|_\infty$. There exists a constant $a > 0$ such that (a) f_0 has two bounded continuous partial derivatives over the set $U_a := \{\boldsymbol{x} : c - a \leq f_0(\boldsymbol{x}) \leq c + a\}$, (b) $\inf_{U_a} \|\nabla f_0\| > 0$, and (c) U_a is contained in $\beta(c)^\delta$ for some $\delta > 0$.

Assumption D1b.

1. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be i.i.d. from a bounded density f_0 on $\mathbb{R}^d$, $d \geq 2$.
2. The density f_0 has two bounded continuous partial derivatives for all $\boldsymbol{x} \in \mathbb{R}^d$.
3. There exists a constant $a > 0$ such that $U_a := \{\boldsymbol{x} : f_{\tau,0} - a \leq f_0(\boldsymbol{x}) \leq f_{\tau,0} + a\}$ satisfies (a) $\inf_{U_a} \|\nabla f_0\| > 0$, and (b) U_a is contained in β_τ^δ for some $\delta > 0$.

Assumption D1a will be used for LS estimation and Assumption D1b for HDR estimation. We need the stronger global twice differentiability assumption in HDR estimation because of the need to estimate $f_{\tau,0}$ (which involves estimating the f_0-probability content of $\mathcal{L}_\tau$). The global twice differentiability assumption

in Assumption D1b could be weakened to an assumption of twice differentiability either on $\mathcal{L}_\tau^\delta$ or on $(\mathcal{L}_\tau^c)^\delta$.

Assumptions D1a and D1b entail that the gradient of f_0 is nonzero on (a neighborhood of) the level set of interest. This implies by the preimage theorem that the level set β, taken to be either $\beta(c)$ or β_τ, is a $(d-1)$-dimensional (boundaryless) manifold (Guillemin and Pollack, 1974). The only additional assumption we need is one of compactness, which rules out only very pathological cases, where f_0 has "spikes" of increasingly small width going out towards infinity.

Assumption D2. Let $\inf_{x\in\mathbb{R}^d} f_0(x) < c < \|f_0\|_\infty$ or $0 < \tau < 1$ be as in Assumptions D1a and D1b. Assume that $\beta(c)$ or β_τ is compact.

Our assumption on the kernel will come in the form of a so-called *Vapnik-Chervonenkis (VC)* (Dudley, 1999) type of assumption. For a metric space (T, d) and $\tau > 0$, the covering number $N(T, d, \tau)$ is the smallest number of balls of radius τ (and centers which may or may not be in T) needed to cover T. If a class of functions $\mathcal{F}$ is a VC class, we have that

$$\sup_P N(\mathcal{F}, \|\cdot\|_{2,P}, \tau\|F\|_{2,P}) \leq \left(\frac{A}{\tau}\right)^v \tag{6}$$

for some positive A, v, where the sup is over all probability measures P, and where F is the envelope of $\mathcal{F}$ meaning $\sup_{f\in\mathcal{F}}|f| \leq F$ (Chapter 2.6, van der Vaart and Wellner (1996)). We will simply directly assume that the needed classes satisfy (6). Thus our assumptions are as follows.

Assumption K.

1. The kernel K is an everywhere continuously differentiable bounded density on $\mathbb{R}^d$ with bounded partial derivatives. Both $\int K^2\,d\lambda$ and $\int(\nabla K)(\nabla K)'\,d\lambda$ are finite or have finite entries, respectively. Assume that $\int K(\boldsymbol{x})\boldsymbol{x}\,d\boldsymbol{x} = \boldsymbol{0}$, $\int \boldsymbol{x}\boldsymbol{x}'K(\boldsymbol{x})\,d\boldsymbol{x} = \mu_2(K)\boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix and $\mu_2(K) = \int x_i^2 K(\boldsymbol{x})\,d\boldsymbol{x}$ is independent of i.
2. Assume that (6) is satisfied with $\mathcal{F}$ taken to be

$$\left\{K\left(\boldsymbol{H}^{-1/2}(t-\cdot)\right) : t\in\mathbb{R}^d, \boldsymbol{H}\in\mathcal{S}\right\} \qquad \text{and} \tag{7}$$

$$\left\{\|\nabla K\big(\boldsymbol{H}^{-1/2}(t-\cdot)\big)\| : t\in\mathbb{R}^d, \boldsymbol{H}\in\mathcal{S}\right\}. \tag{8}$$

Let $R(K) := \int K^2 d\lambda$ and let $R(\nabla K)$ be the largest eigenvalue of $\int(\nabla K)(\nabla K)'\,d\lambda$.

Assumption H.

1. Let $\boldsymbol{H} \equiv \boldsymbol{H}_n \in \mathcal{S}$, such that for some $c > 0$, $|\boldsymbol{H}| \searrow 0$, $n|\boldsymbol{H}|^{1/2}/\log|\boldsymbol{H}|^{-1/2} \to \infty$, $\log\log n/\log|\boldsymbol{H}|^{-1/2} \to 0$, as $n\to\infty$, and $|\boldsymbol{H}_n|^{1/2} \leq c|\boldsymbol{H}_{2n}|^{1/2}$.
2. Assume that $\lambda_{\max}(\boldsymbol{H}) = O\{\lambda_{\min}(\boldsymbol{H})\}$ and $n|\boldsymbol{H}|^{1/2}\lambda_{\min}(\boldsymbol{H})/\log|\boldsymbol{H}|^{-1/2} \to \infty$ and $\lambda_{\max} = O(n^{-2/(4+d)})$ as $n\to\infty$.

Here, $a_n \searrow 0$ means that a_n decreases monotonically to 0. Assumptios D1a and D1b are standard in the KDE literature (see, e.g., page 95 of Wand and Jones (1995)). Note that Assumption 3 of Assumption D1b implies that there exists a constant $L > 0$ such that for $\delta > 0$ small enough that $\lambda(f_0^{-1}([f_{\tau,0} - \delta, f_{\tau,0}+\delta])) \leq L\delta$; this is a standard type of assumption that appears in the level set estimation literature (Polonik, 1995). Assumption D2 is not very limiting and only rules out pathological cases.

Our Assumption K on the kernel function is not restrictive and all of the conditions imposed are fairly standard. For Assumption 1 see, e.g., page 95 of Wand and Jones (1995) where similar conditions are imposed. Assumption 2 is also fairly standard in the KDE literature (e.g., Chen, Genovese and Wasserman (2017) uses similar conditions in the context of inference for level sets). This assumption is needed to apply the results of Giné and Guillou (2002) to get almost sure convergence rates of $\widehat{f}_{n,\boldsymbol{H}}$ and $\nabla\widehat{f}_{n,\boldsymbol{H}}$. Assumption K_1 of Giné and Guillou (2002) (or Assumption K, page 2572, of Giné, Koltchinskii and Zinn (2004)) is an easy-to-verify condition that implies Assumption 2 holds, and shows that Assumption 2 holds for Gaussian kernels and for many compactly supported kernels.

The expansions given in our Theorem 2.1 and 2.2 hold for the range of bandwidths given in Assumption H. This is sufficient to develop a practical bandwidth selector, since larger or smaller bandwidths can be easily ruled out. See Corollaries 2.1 and 2.2.

2.3. Asymptotic risk expansions

Our main results are stated in the following two theorems. The first gives the asymptotic risk expansion for level set estimation. Let $\Phi(\cdot)$ and $\phi(\cdot)$ denote the standard normal distribution function and density function, respectively.

Theorem 2.1. *For given constant c with $\inf_{\boldsymbol{x}\in\mathbb{R}^d} f_0(\boldsymbol{x}) < c < \|f_0\|_\infty$, let Assumptions K, H, D1a and D2 hold. Moreover, the kernel function K has bounded support. Then*

$$\mathbb{E}\left[\mu_{f_0}\{\mathcal{L}(c)\Delta\widehat{\mathcal{L}}_{\boldsymbol{H}}(c)\}\right] = \mathrm{LS}(\boldsymbol{H}) + o\left\{(n|\boldsymbol{H}|^{1/2})^{-1/2} + \mathrm{tr}(\boldsymbol{H})\right\}$$

as $n \to \infty$, where

$$\mathrm{LS}(\boldsymbol{H}) := \frac{c}{\sqrt{n|\boldsymbol{H}|^{1/2}}} \int_{\beta(c)} \frac{2\phi(B_{\boldsymbol{x}}(\boldsymbol{H})) + 2\Phi(B_{\boldsymbol{x}}(\boldsymbol{H}))B_{\boldsymbol{x}}(\boldsymbol{H}) - B_{\boldsymbol{x}}(\boldsymbol{H})}{-A_{\boldsymbol{x}}}\, d\mathcal{H}(\boldsymbol{x}),$$

$$A_{\boldsymbol{x}} := -\frac{\|\nabla f_0(\boldsymbol{x})\|}{\sqrt{R(K)c}}, \quad \text{and} \quad B_{\boldsymbol{x}}(\boldsymbol{H}) := -\frac{\sqrt{n|\boldsymbol{H}|^{1/2}}D_1(\boldsymbol{x},\boldsymbol{H})}{\sqrt{R(K)c}}, \tag{9}$$

with $D_1(\boldsymbol{x}, H) := \frac{1}{2}\mu(K)\,\mathrm{tr}(\boldsymbol{H}\nabla^2 f_0(\boldsymbol{x}))$.

Note that the first summand (including the factor $c/\sqrt{n|\boldsymbol{H}|^{1/2}}$) in the integral defining $\mathrm{LS}(\boldsymbol{H})$ is of the order of magnitude of a variance term in a

mean-squared error decomposition, and the second two summands are of the same order of magnitude of a squared bias term. The next theorem gives the HDR asymptotic risk expansion.

Theorem 2.2. *Let Assumptions D1b,D2,K and H hold. Then*

$$\mathbb{E}\left[\mu_{f_0}\{\mathcal{L}_\tau \Delta \widehat{\mathcal{L}}_{\tau,\boldsymbol{H}}\}\right] = \mathrm{HDR}(\boldsymbol{H}) + o\left\{(n|\boldsymbol{H}|^{1/2})^{-1/2} + \mathrm{tr}(\boldsymbol{H})\right\}$$

as $n \to \infty$*, where*

$$\mathrm{HDR}(\boldsymbol{H}) := \frac{f_{\tau,0}}{\sqrt{n|\boldsymbol{H}|^{1/2}}} \int_{\beta_\tau} \frac{2\phi(C_{\boldsymbol{x}}(\boldsymbol{H})) + 2\Phi(C_{\boldsymbol{x}}(\boldsymbol{H}))C_{\boldsymbol{x}}(\boldsymbol{H}) - C_{\boldsymbol{x}}(\boldsymbol{H})}{-A_{\boldsymbol{x}}}\, d\mathcal{H}(\boldsymbol{x}),$$

$$C_{\boldsymbol{x}}(\boldsymbol{H}) := B_{\boldsymbol{x}}(\boldsymbol{H}) + \sqrt{\frac{n|\boldsymbol{H}|^{1/2}}{R(K) f_{\tau,0}}} D_2(\boldsymbol{H}).$$

$A_{\boldsymbol{x}}$ *and* $B_{\boldsymbol{x}}(\boldsymbol{H})$ *are defined in the same way as in Theorem 2.1 with* c *replaced by* $f_{\tau,0}$*. And*

$$D_2(\boldsymbol{H}) := w_0 \left\{V_1(\boldsymbol{H}) + V_2(\boldsymbol{H})\right\},$$

with $w_0 := (\int_{\beta_\tau} 1/\nabla f_0 \, d\mathcal{H})^{-1}$ *and*

$$V_1(\boldsymbol{H}) := \int_{\beta_\tau} \frac{D_1(\boldsymbol{x}, \boldsymbol{H})}{\|\nabla f_0(\boldsymbol{x})\|}\, d\mathcal{H}(\boldsymbol{x}) \qquad V_2(\boldsymbol{H}) := \frac{1}{f_{\tau,0}} \int_{\mathcal{L}_\tau} D_1(\boldsymbol{x}, \boldsymbol{H})\, d\boldsymbol{x}.$$

We defer the proofs to the appendix. Next, we would like to study the theoretical behavior of the minimizers of LS$(\cdot)$ and HDR$(\cdot)$. Note that the minimizers of LS$(\cdot)$ or of HDR$(\cdot)$ are not practically usable bandwidth matrices, since LS$(\cdot)$ and HDR$(\cdot)$ depend on the true, unknown density f_0. We will discuss estimation of HDR$(\cdot)$ and of LS$(\cdot)$ and practical bandwidth selectors in the next section. Presently, we consider the minimizers of LS$(\cdot)$ and HDR$(\cdot)$, which serve as *oracle* bandwidth selectors.

Unfortunately, LS$(\cdot)$ and HDR$(\cdot)$ are quite complicated functions so studying their minimizers in general is not at all straightforward. Thus we will make some simplifying assumptions. We will consider f_0 that is unimodal and spherically symmetric about some point (taken to be the origin in Corollary 2.1 and 2.2). We will consider optimizing over the subclass $\mathcal{S}_1 := \left\{h^2\boldsymbol{I} : h > 0\right\}$ of bandwidth matrices, where $\boldsymbol{I}$ is the $d \times d$ identity matrix. These assumptions are made largely for simplicity and ease of presentation of the following two corollaries, and are far from necessary for the conclusions to hold. We discuss these assumptions again after the corollaries. By a slight abuse of notation, we let $\mathrm{LS}(h) \equiv \mathrm{LS}(h^2\boldsymbol{I})$ and $\mathrm{HDR}(h) \equiv \mathrm{HDR}(h^2\boldsymbol{I})$.

Corollary 2.1. *Let the assumptions of Theorem 2.1 hold. Assume further that* $f_0(x) = g(\|x\|)$ *and that the function* $g(r)$ *defined for* $r > 0$ *is strictly decreasing on* $[0, \infty)$*. Then there exists a constant* s_{opt} *depending on* f_0 *and* K *(but not*

on n) such that there is a unique positive number $h_{opt} = \operatorname{argmin}_{h\in[0,\infty)} \mathrm{LS}(h)$ satisfying

$$h_{opt} = s_{opt} n^{-1/(d+4)} \quad \text{and} \quad h_0 = h_{opt}(1+o(1)) \qquad \text{as } n \to \infty,$$

where h_0 is any minimizer of $\mathbb{E}[\mu_{f_0}\{\mathcal{L}(c)\Delta\widehat{\mathcal{L}}_{\boldsymbol{H}}(c)\}]$.

Corollary 2.2. *Let the assumptions of Theorem 2.2 hold. Assume further that $f_0(x) = g(\|x\|)$ and that the function $g(r)$ defined for $r > 0$ is strictly decreasing on $[0,\infty)$. Then there exists a constant s_{opt} depending on f_0 and K (but not on n) such that there is a unique positive number $h_{opt} = \operatorname{argmin}_{h\in[0,\infty)} \mathrm{HDR}(h)$ satisfying*

$$h_{opt} = s_{opt} n^{-1/(d+4)} \quad \text{and} \quad h_0 = h_{opt}(1+o(1)) \qquad \text{as } n \to \infty,$$

where h_0 is any minimizer of $\mathbb{E}[\mu_{f_0}\{\mathcal{L}_\tau\Delta\widehat{\mathcal{L}}_{\tau,\boldsymbol{H}}\}]$.

The proof of the two corollaries follows exactly the same way, so we provide the proof for HDR estimation and omit that for LS estimation. The corollaries tell us the order of magnitude of the true optimal bandwidths and of the oracle bandwidths. We used the assumptions of unimodality and spherical symmetry because these assumptions imply that f_0, ∇f_0, and $\nabla^2 f_0$ are constant on β_τ and $\beta(c)$. We believe that (an analogous form of) the conclusions of Corollary 2.1 and 2.2 hold for $\boldsymbol{H}_{\mathrm{opt}} \in \operatorname{argmin}_{\boldsymbol{H}\in\mathcal{S}} \mathrm{HDR}(\boldsymbol{H})$ and for $\boldsymbol{H}_{\mathrm{opt}} \in \operatorname{argmin}_{\boldsymbol{H}\in\mathcal{S}} \mathrm{LS}(\boldsymbol{H})$, and for much more general densities f_0. Our simulations show that our practical bandwidth selector (studied in the next section) does not require such extreme assumptions.

3. Bandwidth selection methodology

In the previous section, we provided asymptotic expansions of symmetric risks for HDR estimation and LS estimation, which could be used as guidance for bandwidth selection in those two scenarios. Minimizers of $\mathrm{LS}(\boldsymbol{H})$ and $\mathrm{HDR}(\boldsymbol{H})$ are natural bandwidth selectors for HDR estimation and LS estimation, respectively. The theoretical performance of the bandwidth selector using "oracle" knowledge of the functionals of the true density is studied in Corollary 2.1 and 2.2. Of course, in practice, one does not have this oracle knowledge. In the present section, we develop an effective practical bandwidth selection procedure for HDR estimation (a procedure for level set estimation is simpler and can be derived in a similar way). We will also study the theoretical performance of our bandwidth selector restricted to a simplified class $\mathcal{S}_1 = \{h^2\boldsymbol{I}, h > 0\}$.

Since there are unknown quantities that $\mathrm{HDR}(\boldsymbol{H})$ depends on, a natural "plug-in" approach is to estimate those quantities using different kernel density estimators and plug the estimates in. Moreover, the unknown functionals depend on the truth through $f_0, \nabla f_0, \nabla^2 f_0$, so we will use three pilot kernel density estimators. To be specific, we use $\widehat{f}_{n,\boldsymbol{H}_0}$ to estimate $f_{\tau,0}$ and $\mathcal{L}_\tau$; we use $\nabla\widehat{f}_{n,\boldsymbol{H}_1}$ to estimate ∇f_0, and β_τ combined with the pilot estimator of $f_{\tau,0}$; we use

$\nabla^2 \widehat{f}_{n,\boldsymbol{H}_2}$ to estimate $\nabla^2 f_0$, where $\boldsymbol{H}_0$, $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ are corresponding pilot bandwidth matrices for the three kernel density estimators. (One could also use three different kernels for $\widehat{f}_{n,\boldsymbol{H}_i}$, $i = 0, 1, 2$, but we will use the same kernel for all three.) For our theoretical results to hold, we require just the bandwidth matrix $\boldsymbol{H}_r$ to be of the optimal order for estimating the rth derivatives of f_0 (see Corollary 3.2 and Assumption H2, below). We use two-stage direct plug-in estimators for the pilot bandwidths in our algorithm below, which converge at the correct rate. A detailed description about plug-in estimators could be found in Wand and Jones (1995, Chapter 3) and Chacón and Duong (2010).

Once we have those estimated functionals, we can plug them into $\mathrm{HDR}(\boldsymbol{H})$ to obtain an estimated loss function $\widehat{\mathrm{HDR}}(\boldsymbol{H})$. Note $\boldsymbol{H}$ appears in the integrand of a Hausdorff integral and cannot be factored out of the integral; thus minimizing $\widehat{\mathrm{HDR}}(\boldsymbol{H})$ directly is infeasible. Instead, we minimize a discretized approximation to $\widehat{\mathrm{HDR}}(\boldsymbol{H})$. To illustrate this idea, we use the minimization of $\mathrm{HDR}(\boldsymbol{H})$ as an example. Let $\mathcal{A} = \{A_i\}_{i=1}^m$ be a partition of β_τ such that $\mathcal{H}(A_i)$ is sufficiently small for $i = 1, 2, \ldots, m$. Then $w_0 = (\int_{\beta_\tau} \frac{1}{\|\nabla f_0\|}\, d\mathcal{H})^{-1}$ can be approximated by $\tilde{w}_0 = \sum_{i=1}^m \frac{1}{\|\nabla f_0(\tilde{\boldsymbol{x}}_i)\|}\mathcal{H}(A_i)$, where $\tilde{\boldsymbol{x}}_i$ is an arbitrary point belonging to A_i. Note for $d = 2$, $\mathcal{H}(A_i)$ is well approximated by the length of the line segment connecting the boundary points of A_i. $V_1(\boldsymbol{H})$ and $V_2(\boldsymbol{H})$ can be computed approximately in similar ways. Replacing w_0, $V_1(\boldsymbol{H})$, $V_2(\boldsymbol{H})$ with corresponding discretized approximations in $C_{\boldsymbol{x}}(\boldsymbol{H})$ gives us an approximation $\tilde{C}_{\boldsymbol{x}}(\boldsymbol{H})$ for each $\boldsymbol{x}$. Then

$$\begin{aligned}\mathrm{HDR}(\boldsymbol{H}) &\approx \frac{f_{\tau,0}}{\sqrt{n|\boldsymbol{H}|^{1/2}}} \int_{\beta_\tau} \frac{2\phi(\tilde{C}_{\boldsymbol{x}}(\boldsymbol{H})) + 2\Phi(\tilde{C}_{\boldsymbol{x}}(\boldsymbol{H}))\tilde{C}_{\boldsymbol{x}}(\boldsymbol{H}) - \tilde{C}_{\boldsymbol{x}}(\boldsymbol{H})}{-A_{\boldsymbol{x}}}\, d\mathcal{H}(\boldsymbol{x}) \\ &\approx \frac{f_{\tau,0}}{\sqrt{n|\boldsymbol{H}|^{1/2}}} \sum_{i=1}^m \frac{2\phi(\tilde{C}_{\tilde{\boldsymbol{x}}_i}(\boldsymbol{H})) + 2\Phi(\tilde{C}_{\tilde{\boldsymbol{x}}_i}(\boldsymbol{H}))\tilde{C}_{\tilde{\boldsymbol{x}}_i}(\boldsymbol{H}) - \tilde{C}_{\tilde{\boldsymbol{x}}_i}(\boldsymbol{H})}{-A_{\tilde{\boldsymbol{x}}_i}} \\ &\quad \times \mathcal{H}(A_i). \end{aligned} \tag{10}$$

The last line above provides a computable, optimizable and close approximation to $\mathrm{HDR}(\boldsymbol{H})$ as long as $\mathcal{H}(A_i)$ is small enough for each i. We use $K = \phi$ throughout the algorithm.

The full algorithm for the HDR bandwidth selector is as follows:

1. With given i.i.d random sample $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$, estimate $\boldsymbol{H}_0$, $\boldsymbol{H}_1$, $\boldsymbol{H}_2$ using two-stage direct plug-in strategies.
2. Obtain the pilot estimator of f_0, ∇f_0, $\nabla^2 f_0$ based on the kernel density estimators $\widehat{f}_{n,\boldsymbol{H}_0}$, $\widehat{f}_{n,\boldsymbol{H}_1}$, $\widehat{f}_{n,\boldsymbol{H}_2}$.
3. Let $\widehat{f}_{\tau,n,\boldsymbol{H}_0} := \inf\{y \in (0,\infty) : \int_{\mathbb{R}^d} \widehat{f}_{n,\boldsymbol{H}_0}(\boldsymbol{x}) \mathbb{1}_{\{\widehat{f}_{n,\boldsymbol{H}_0}(\boldsymbol{x}) \geq y\}}\, d\boldsymbol{x} \leq 1 - \tau\}$ be the pilot estimator of $f_{\tau,0}$, $\widehat{\mathcal{L}}_{\tau,\boldsymbol{H}_0} := \{\boldsymbol{x} \in \mathbb{R}^d : \widehat{f}_{n,\boldsymbol{H}_0}(\boldsymbol{x}) \geq \widehat{f}_{\tau,n,\boldsymbol{H}_0}\}$ be the pilot estimator of $\mathcal{L}_\tau$ and $\widehat{\beta}_{\tau,\boldsymbol{H}_1} := \{\boldsymbol{x} \in \mathbb{R}^d : \widehat{f}_{n,\boldsymbol{H}_1}(\boldsymbol{x}) = \widehat{f}_{\tau,n,\boldsymbol{H}_0}\}$ be the pilot estimator of β_τ.
4. Substitute the estimators from Step 2 and 3 into the expressions for $C_{\boldsymbol{x}}$

and $A_{\boldsymbol{x}}$ to obtain $\widehat{C}_{\boldsymbol{x}}$ and $\widehat{A}_{\boldsymbol{x}}$. Then $\widehat{\mathrm{HDR}}(\boldsymbol{H})$ is equal to

$$\frac{\widehat{f}_{\tau,n,\boldsymbol{H}_0}}{\sqrt{n|\boldsymbol{H}|^{1/2}}} \int_{\widehat{\beta}_{\tau,\boldsymbol{H}_1}} \frac{2\phi(\widehat{C}_{\boldsymbol{x}}(\boldsymbol{H})) + 2\Phi(\widehat{C}_{\boldsymbol{x}}(\boldsymbol{H}))\widehat{C}_{\boldsymbol{x}}(\boldsymbol{H}) - \widehat{C}_{\boldsymbol{x}}(\boldsymbol{H})}{-\widehat{A}_{\boldsymbol{x}}} \, d\mathcal{H}(\boldsymbol{x}).$$

5. Minimize the discretized approximation of $\widehat{\mathrm{HDR}}(\boldsymbol{H})$ described in the previous paragraph with Newton's method to obtain the estimated optimal HDR bandwidth.

Note for the above procedure, in step 3, unlike the pilot estimator for $\mathcal{L}_\tau$, the pilot estimator for β_τ is obtained using $\widehat{f}_{n,\boldsymbol{H}_1}$ with $\widehat{f}_{\tau,n,\boldsymbol{H}_0}$ as the level. The reason we use $\widehat{f}_{n,\boldsymbol{H}_1}$ instead of $\widehat{f}_{n,\boldsymbol{H}_0}$ is because the error bound for estimating β_τ depends on the difference between the gradient of true density and that of the kernel density estimator and using $\widehat{f}_{n,\boldsymbol{H}_1}$ yields a better error bound (See Lemma B.5 and proof of Corollary 3.1, 3.2 for details).

Newton's method does not guarantee the optimum will be a positive definite bandwidth matrix. Luckily, in practice the global minimum appears to always be positive definite. The objective function $\widehat{\mathrm{HDR}}$ appear to be locally convex although not globally convex (see Figures 2 and 3 for some plots of $\mathrm{LS}(\cdot)$ and $\mathrm{HDR}(\cdot)$), so one has to be slightly careful about starting values for Newton's algorithm.

Notice also that in Step 3 of the above algorithm we need to calculate the level $\widehat{f}_{\tau,n,\boldsymbol{H}_0}$ having $\widehat{f}_{n,\boldsymbol{H}_0}$-probability $1-\tau$. Hyndman (1996) suggests two similar methods for calculating $\widehat{f}_{\tau,n,\boldsymbol{H}_0}$. One is to use an appropriate empirical quantile of the values $\widehat{f}_{n,\boldsymbol{H}_0}(\boldsymbol{X}_i)$, $i=1,\ldots,n$ ("Approach H1"). An approach of this type is studied by Cadre, Pelletier and Pudlo (2013) (and by Chen (2016) in calculating his $\hat{\alpha}_n(x)$). However, this estimator is not equal to $\widehat{f}_{\tau,n,\boldsymbol{H}_0}$, and we have not yet quantified the difference, so we choose not to use this approach. Alternatively, Hyndman (1996) suggests resampling $\tilde{\boldsymbol{X}}_1,\ldots,\tilde{\boldsymbol{X}}_M \overset{\text{iid}}{\sim} \widehat{f}_{n,\boldsymbol{H}}$, and then using the appropriate empirical quantile of $\widehat{f}_{n,\boldsymbol{H}_0}(\tilde{\boldsymbol{X}}_i)$, $i=1,\ldots,M$ ("Approach H2"). Any desired accuracy can be attained by taking M large enough. Another method is to simply use numeric integration: one can do a binary search over $(0, \|\widehat{f}_{n,\boldsymbol{H}_0}\|_\infty)$, computing the integral (numerically) at each level until one arrives at $\widehat{f}_{\tau,n,\boldsymbol{H}_0}$ within desired accuracy. When $d=2$, we found the numeric integration and binary search to be the fastest method for calculating $\widehat{f}_{\tau,n,\boldsymbol{H}_0}$. We suspect for higher dimensions, Approach H2 will be faster than numeric integration. Of course, Approach H1 is faster than the other two, and so it would be helpful to study how the Approach H1 estimator compares to $\widehat{f}_{\tau,n,\boldsymbol{H}_0}$.

In our pilot estimation process when $d=2$, we use numerical interpolation to generate points on $\widehat{\beta}_{\tau,\boldsymbol{H}_1}$ and to calculate $\mathcal{A}$. In more detail: we generate dense grid points along both the x-axis and the y-axis, and we estimate the density values at those grid points. Then we perform interpolation between grid points to get points such that the estimated density values at those points are (approximately) $\widehat{f}_{\tau,n,\boldsymbol{H}_0}$, and those points induce a partition of $\widehat{\beta}_{\tau,\boldsymbol{H}_1}$. Then for any A_i in

the partition, A_i is defined by two end points, and $\mathcal{H}(A_i)$ can be approximated by the length of the line segment connecting those two end points. By generating enough dense and equally spaced grid points, we expect those line segments will approximate the true partition $\mathcal{A}$ well and thus the Hausdorff integral will also be well approximated. However, this method is hard to implement in dimension larger than 2 because there is no simple approximation for the volumes of corresponding partition sets of $\widehat{\beta}_{\tau,\boldsymbol{H}_1}$. One approach that may be fruitful for solving this problem is to use Quasi-Monte Carlo integration to calculate the Hausdorff integral (see De Marchi and Elefante, 2018). The idea is to generate a set of points $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$ on the manifold β such that those points are approximately uniformly distributed and then we can approximate $\int_\beta \gamma(\boldsymbol{x})\, d\mathcal{H}$ by $\frac{1}{m}\sum_{i=1}^m \gamma(\boldsymbol{b}_i)$. Analysis and numerical simulation for the method has been done for special Hausdorff integrals over special manifolds (cone, cylinder, sphere and torus). There is further work needed to extend the method to the more general manifolds that arise in our problem, which we believe is non-trivial and beyond the scope of this paper.

Note that the method just described for computing the approximation (10) can be implemented as a so-called midpoint method of numerical integration, for which classical analysis shows an error rate of $O(m^{-2})$ (m is the number of equi-sized partitioning sets of the interval), provided that the function being integrated has bounded second derivative and the domain being integrated is a compact interval in $\mathbb{R}$ (Hämmerlin and Hoffmann, 1991). The same error applies for using the midpoint method to numerically compute Hausdorff integrals over one dimensional compact manifolds embedded in $\mathbb{R}^2$, by the change of variables Theorem 2 (page 99) of Evans and Gariepy (2015). Thus the errors for our selected bandwidths in the corollaries below will also have an error dependent on m, but in our experience m can be chosen large enough that this is negligible (when $d = 2$), so we do not include it in the analysis.

To give the asymptotic performance of our bandwidth selector, we need the following additional assumptions.

Assumption D3. The true density function f_0 has four continuous bounded and square integrable derivatives.

Assumption K2. K is symmetric along each coordinate, i.e., for $i = 1, \ldots, d$, we have $K(x_1, \ldots, x_i, \ldots, x_d) = K(x_1, \ldots, -x_i, \ldots, x_d)$. And all the first and second partial derivatives of K are square integrable.

Assumption H2. For $r = 0, 1, 2$, the bandwidth matrix $\boldsymbol{H}_r$ is symmetric, positive definite, such that $\boldsymbol{H}_r \to 0$ elementwise, and $n^{-1}|\boldsymbol{H}_r|^{-1/2}(\boldsymbol{H}_r^{-1})^{\otimes r} \to 0$ as $n \to \infty$, where $\otimes$ stands for Kronecker product.

This assumption and notation is as in Chacón, Duong and Wand (2011). Here for a matrix $\boldsymbol{A}$, $\boldsymbol{A}^{\otimes 0} = 1 \in \mathbb{R}$ and $\boldsymbol{A}^{\otimes 1} = \boldsymbol{A}$. Now, recall that $\mathrm{LS}(h) \equiv \mathrm{LS}(h^2\boldsymbol{I})$ and

$$\mathrm{LS}(h^2\boldsymbol{I}) = \frac{c}{(nh^d)^{1/2}} \int_{\beta(c)} \frac{\phi(B_{\boldsymbol{x}}(h)) + 2\Phi(B_{\boldsymbol{x}}(h))B_{\boldsymbol{x}}(h) - B_{\boldsymbol{x}}(h)}{-A_{\boldsymbol{x}}}\, d\mathcal{H}(\boldsymbol{x}),$$

where $B_{\boldsymbol{x}}(h) = (bh^{d+4})^{1/2}F_{\boldsymbol{x}}$ with $F_{\boldsymbol{x}} = -\frac{1}{2}\mu(K)\operatorname{tr}(\nabla^2 f_0(\boldsymbol{x}))/\sqrt{R(K)c}$. And $\mathrm{HDR}(h) \equiv \mathrm{HDR}(h^2\boldsymbol{I})$ with

$$\mathrm{HDR}(h^2\boldsymbol{I}) = \frac{f_{\tau,0}}{(nh^d)^{1/2}} \int_{\beta_\tau} \frac{\phi(C_{\boldsymbol{x}}(h)) + 2\Phi(C_{\boldsymbol{x}}(h))C_{\boldsymbol{x}}(h) - C_{\boldsymbol{x}}(h)}{-A_{\boldsymbol{x}}}\, d\mathcal{H}(\boldsymbol{x}),$$

where $C_{\boldsymbol{x}}(h) = (nh^{d+4})^{1/2}G_{\boldsymbol{x}}$, and

$$G_{\boldsymbol{x}} = -\frac{\mu(K)\operatorname{tr}(\nabla^2 f_0(\boldsymbol{x}))}{\sqrt{R(K)f_{\tau,0}}} + \frac{w_0 \int_{\beta_\tau} \frac{\mu(K)\operatorname{tr}(\nabla^2 f_0)}{2\|\nabla f_0\|}\, d\mathcal{H} + \frac{w_0}{f_{\tau,0}} \int_{\mathcal{L}_\tau} \frac{\mu(K)\operatorname{tr}(\nabla^2 f_0)}{2}\, d\lambda}{\sqrt{R(K)f_{\tau,0}}}.$$

By letting $s = (nh^{d+4})^{1/2}$, we see that minimizing $\mathrm{LS}(h)$ is equivalent to minimizing

$$\mathrm{AR}_{\mathrm{LS}}(s) := s^{-d/(d+4)} \int_{\beta(c)} \frac{\phi(sF_{\boldsymbol{x}}) + 2\Phi(sF_{\boldsymbol{x}})sF_{\boldsymbol{x}} - sF_{\boldsymbol{x}}}{-A_{\boldsymbol{x}}}\, d\mathcal{H}(\boldsymbol{x}),$$

and minimizing $\mathrm{HDR}(h)$ is equivalent to minimizing

$$\mathrm{AR}_{\mathrm{HDR}}(s) := s^{-d/(d+4)} \int_{\beta_\tau} \frac{\phi(sG_{\boldsymbol{x}}) + 2\Phi(sG_{\boldsymbol{x}})sG_{\boldsymbol{x}} - sG_{\boldsymbol{x}}}{-A_{\boldsymbol{x}}}\, d\mathcal{H}(\boldsymbol{x}).$$

he following corollaries show the convergence rate of the estimated optimal bandwidth for $\boldsymbol{H} \in \mathcal{S}_1$.

Corollary 3.1. *Let Assumptions D1a, D2, D3, K, K2 and H2 hold. Assume further that s_{opt} is a unique minimizer of $AR_{\mathrm{LS}}(s)$ for $s > 0$ and $AR''_{\mathrm{LS}}(s_{opt}) > 0$. Then*

$$\frac{\hat{h}_{opt}}{h_{opt}} = 1 + O_p\left(n^{-2/(d+8)}\right) \quad \text{and} \quad \frac{\hat{h}_{opt}}{h_0} = 1 + O_p\left(n^{-2/(d+8)}\right),$$

as $n \to \infty$, where $\hat{h}_{opt}$ is the minimizer of $\widehat{LS}(h)$, h_{opt} is the minimizer of $LS(h)$ and h_0 is any minimizer of $\mathbb{E}[\mu_{f_0}\{\mathcal{L}(c)\Delta\widehat{\mathcal{L}}_{\boldsymbol{H}}(c)\}]$ over the class $\mathcal{S}_1 = \{h^2\boldsymbol{I}, h > 0\}$.

Corollary 3.2. *Let Assumptions D1b, D3, K, K2 and H2 hold. Assume further that s_{opt} is a unique minimizer of $AR_{\mathrm{HDR}}(s)$ for $s > 0$ and $AR''_{\mathrm{HDR}}(s_{opt}) > 0$. Then*

$$\frac{\hat{h}_{opt}}{h_{opt}} = 1 + O_p\left(n^{-2/(d+8)}\right),$$

as $n \to \infty$, where $\hat{h}_{opt}$ is the minimizer of $\widehat{HDR}(h)$ and h_{opt} is the minimizer of $HDR(h)$.

Corollaries 3.1 and 3.2 both assume existence of a point s_{opt}. Corollary 2.1 and 2.2 show the existence of s_{opt} under one set of assumptions, although (as discussed after those corollaries) this conclusion holds in many other scenarios.

Remark 3.1. In Corollary 3.1, we provide the rates of convergence for both the estimated optimal bandwidth to the oracle bandwidth selector and the estimated optimal bandwidth to the true minimizer of $\mathbb{E}[\mu_{f_0}\{\mathcal{L}(c)\Delta\widehat{\mathcal{L}}_{\boldsymbol{H}}(c)\}]$, while in Corollary 3.2, we only provide the rate of convergence for the estimated optimal bandwidth to the oracle bandwidth selector. The main difficulty for proving the convergence rate of the estimated optimal bandwidth to the true minimizer of $\mathbb{E}[\mu_{f_0}\{\mathcal{L}_\tau\Delta\widehat{\mathcal{L}}_{\tau,\boldsymbol{H}}\}]$, as we can see from the proof of Theorem 2.2, is understanding the Var $\widehat{f}_{\tau,n}$ term. At present, we can only show that Var $\widehat{f}_{\tau,n}$ is $o(\frac{1}{n|\boldsymbol{H}|^{1/2}})$, but do not have a more explicit expression. Thus (even with higher order derivative assumptions) we cannot say anything stronger about Var $\widehat{f}_{\tau,n}$, which is different than when β_τ is a discrete point set, in the $d = 1$ case.

Remark 3.2. The rates of convergence given in Corollaries 3.1 and 3.2 are known as *relative rates of convergence* since they are of the form $(\hat{h}_{\text{opt}} - \tilde{h})/\tilde{h}$ for some $\tilde{h}$ (which is itself converging to 0) (Wand and Jones, 1995). One can compare the relative rates from Corollaries 3.1 and 3.2 to the relative rates of other KDE bandwidth selectors. If we plug $d = 1$ into the rate $n^{-2/(d+8)}$ we recover the rate that arose in Theorem 3 of Samworth and Wand (2010). We can also make comparisons to bandwidth selector relative rates based on global loss functions. Duong and Hazelton (2005) study relative rates of convergence for various bandwidth selectors to the bandwidth matrix that minimizes mean integrated squared error, $E\int_{\mathbb{R}^d}(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) - f_0(\boldsymbol{x}))^2\,d\boldsymbol{x}$. (An alternative benchmark is the bandwidth that minimizes *integrated squared error*, $\int_{\mathbb{R}^d}(\widehat{f}_{n,h}(\boldsymbol{x}) - f_0(\boldsymbol{x}))^2\,d\boldsymbol{x}$, for which e.g., LSCV performs well (Hall and Marron, 1987), but the relative rates for that problem behave quite differently than the ones we study in Corollaries 3.1 and 3.2, so we do not mention them here.) Table 1 of Duong and Hazelton (2005) presents the convergence rates for plug-in, unbiased cross validation, biased cross validation, and smoothed cross validation bandwidth matrix estimators. (See also Sain, Baggerly and Scott (1994b); Wand and Jones (1994); Duong and Hazelton (2003); Scott and Terrell (1987); Sheather and Jones (1991); Hall, Marron and Park (1992).) Consider $d \geq 2$. The unbiased and biased cross validation methods have relative convergence rates of $n^{-\min(d,4)/(2d+8)}$. The smoothed cross validation method and the plug-in method of Duong and Hazelton (2003) both have rates of $n^{-2/(d+6)}$. The plug-in method of Wand and Jones (1994) has a rate of $n^{-4/(d+12)}$ which is the fastest rate for all d. The rate presented in our corollaries is faster than $n^{-\min(d,4)/(2d+8)}$ but slower than $n^{-2/(d+6)}$. This suggests that more careful development of our plug-in procedure, perhaps involving more careful pilot bandwidth selection procedures, could potentially improve the asymptotic rate. However the analysis (in particular understanding how $\text{Var}(\widehat{f}_{\tau,n})$ behaves) may not be trivial. Also, procedures with better asymptotics may be inferior until the sample size is unrealistically large (this

is somewhat common in bandwidth selection settings (Wand and Jones, 1995, Section 3.8)).

4. Simulations and data analysis

In Section 3, we used LS($\boldsymbol{H}$) and HDR($\boldsymbol{H}$) to develop a bandwidth selection procedure for level set and HDR estimation. We have implemented our procedure in an R (R Core Team, 2018) package **lsbs**. In this section, we assess the accuracy of LS($\boldsymbol{H}$) and HDR($\boldsymbol{H}$) at approximating the true risks. We also use simulation to compare our procedure with the least square cross validation procedure (LSCV), An established ISE-based bandwidth selector (See Rudemo, 1982; Bowman, 1984). We simulate from the 12 bivariate normal mixture densities constructed by Wand and Jones (1993). These densities have a variety of shapes and have between 1 and 4 modes. In addition to those 12 density functions, we also simulate from

$$\frac{2}{3}N\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1/4 & 0\\0 & 1\end{pmatrix}\right)+\frac{1}{3}N\left(\begin{pmatrix}0\\0\end{pmatrix},\frac{1}{50}\begin{pmatrix}1/4 & 0\\0 & 1\end{pmatrix}\right), \tag{11}$$

which is constructed to play a bivariate analogy to the sharp mode density 4 in Marron and Wand (1992) (see also Figure 1 of Samworth and Wand (2010)). The specific form in (11) is chosen to match that used by Qiao (2018).

We will close this section with a real data analysis in which we apply HDR estimation to novelty detection for the Wisconsin Diagnostic Breast Cancer dataset and Banknote Authentication dataset, which are available on the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/).

4.1. Assessment of approximation and estimation comparison

Since it is infeasible to exactly evaluate the true symmetric risk $\mathbb{E}[\mu_{f_0}\{\mathcal{L}_\tau\Delta\widehat{\mathcal{L}}_{\tau,\boldsymbol{H}}\}]$, we approximate the true risk through Monte Carlo. For given $n,\tau,\boldsymbol{H}$, for a large Monte Carlo sample size M, $\mathbb{E}[\mu_{f_0}\{\mathcal{L}_\tau\Delta\widehat{\mathcal{L}}_{\tau,\boldsymbol{H}}\}]\approx\frac{1}{M}\sum_{i=1}^{M}\mu_{f_0}\{\mathcal{L}_\tau\Delta\widehat{\mathcal{L}}^{[i]}_{\tau,\boldsymbol{H}}\}$, where $\widehat{\mathcal{L}}^{[1]}_{\tau,\boldsymbol{H}},\widehat{\mathcal{L}}^{[2]}_{\tau,\boldsymbol{H}},\ldots,\widehat{\mathcal{L}}^{[M]}_{\tau,\boldsymbol{H}}$ are M independent realizations of $\widehat{\mathcal{L}}_{\tau,\boldsymbol{H}}$. In a multivariate KDE the bandwidth matrix contains $d(d+1)/2$ parameters. For the purpose of visualization, we restrict $\boldsymbol{H}\in\mathcal{S}_1=\{h^2\boldsymbol{I}\}$ so that it can be parametrized by a single parameter h.

Figures 2 and 3 compare the asymptotic risk approximation with the simulated true risk for HDR estimation and LS estimation, respectively, for densities corresponding to Densities C, D, E and K of Wand and Jones (1993). Contour plots of the densities are given in the top row of the figures. In Figure 3, we choose τ to be 0.2, 0.5 and 0.8 while in Figure 2, we use the same levels but with true level values computed from the underlying true density functions. For both scenarios, the sample size is chosen to be 2000 and the kernel is set to be the Gaussian kernel throughout the simulation (Theorem 2.1 requires K to be

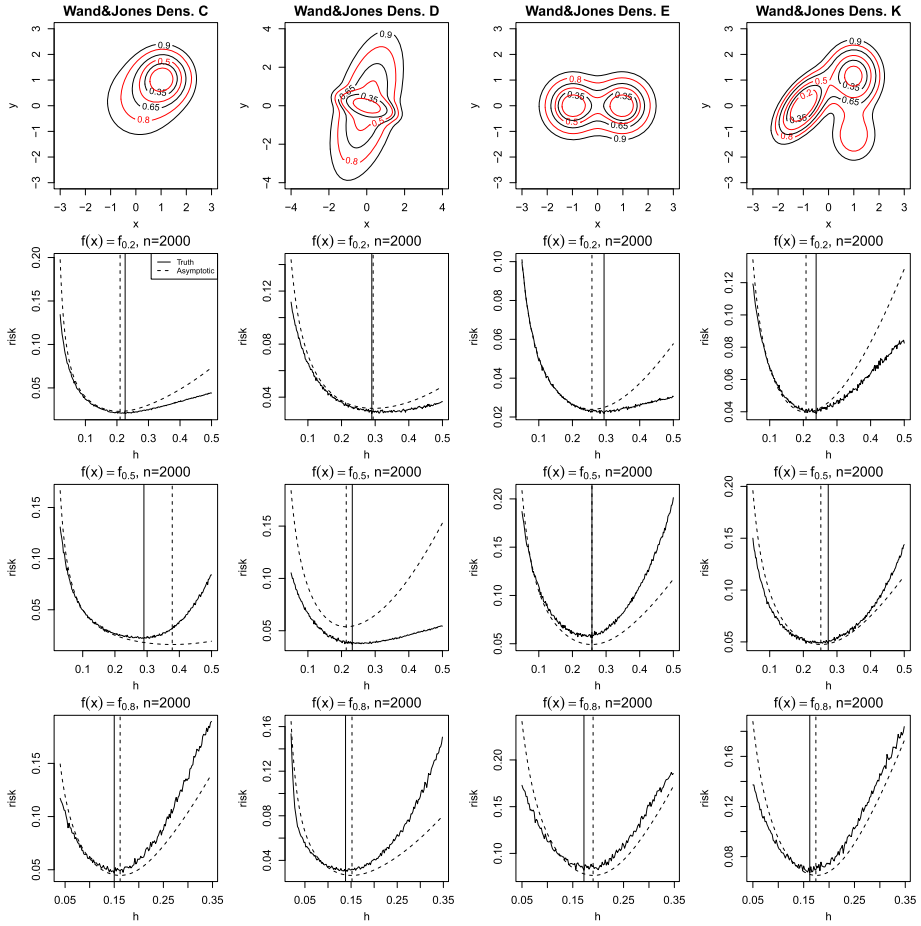

FIG 2. *Comparison of the simulated true risk function* $\mathbb{E}[\mu_{f_0}\{\mathcal{L}(c)\Delta\widehat{\mathcal{L}}_{\boldsymbol{H}}(c)\}]$ *with LS*($\boldsymbol{H}$) *for four densities in Wand and Jones (1993). The panels in the first row are the contour plots for four densities with the contours of interest plotted in red color. The panels in the rest of the rows are the comparison plots for the simulated true risk (solid line) and LS*($\boldsymbol{H}$) *(dashed line) corresponding to the density at the top of the column for* $\tau = 0.2, 0.5, 0.8$. *The positions of the solid vertical line and the dashed line stand for the optimal bandwidths obtained from the simulated true risk and the asymptotic approximation respectively over the restricted class* $\mathcal{S}_1$. *The sample size for all the cases is 2000.*

compactly supported, but nonetheless, the simulation results are not sensitive to the choice of Guassian kernel). We can see from Figures 2 and 3, in both scenarios, our asymptotic expansions provide a good approximation to the truth. The approximation works fairly well for the small values of bandwidth but the discrepancy becomes obvious when h is larger, which is unlike what was observed from the simulation in univariate cases (see Samworth and Wand, 2010). This is consistent with our Assumption H which imposes an upper bound on

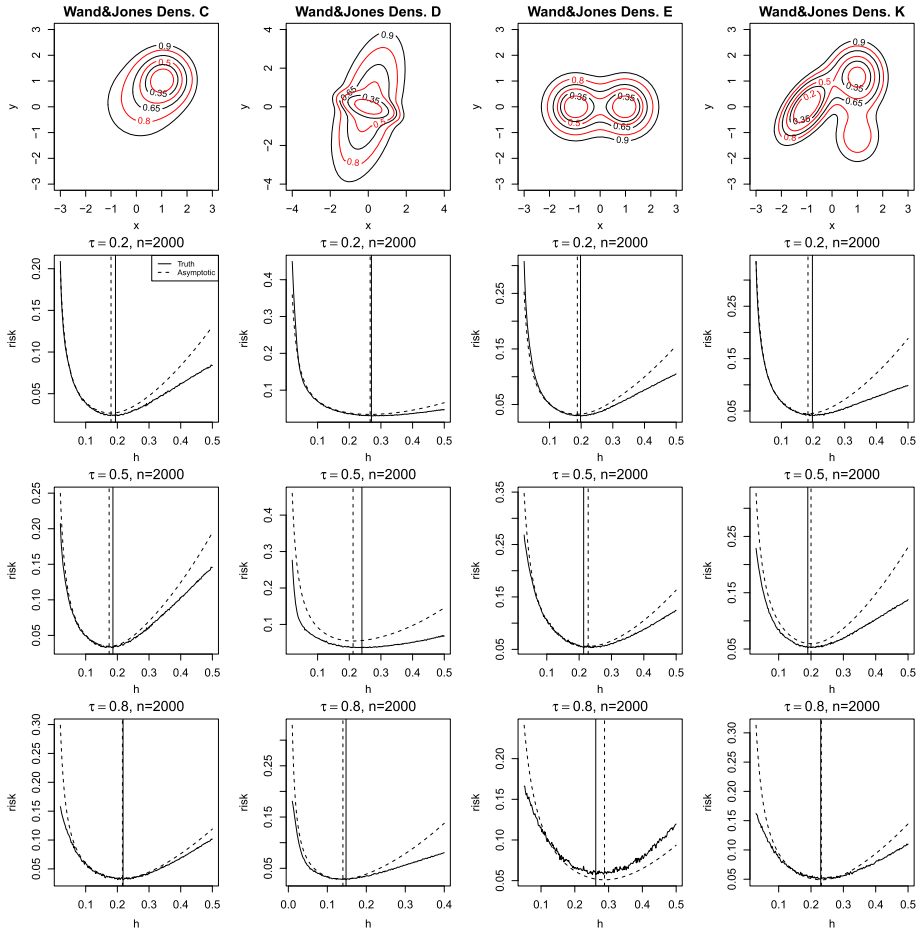

FIG 3. *Comparison of the simulated true risk function* $\mathbb{E}[\mu_{f_0}\{\mathcal{L}_\tau \Delta \widehat{\mathcal{L}}_{\tau,\boldsymbol{H}}\}]$ *with HDR*($\boldsymbol{H}$) *for four densities in Wand and Jones (1993). The panels in the first row are the contour plots for four densities with the contours of interest plotted in red color. The panels in the rest of the rows are the comparison plots for the simulated true risk (solid line) and the HDR*($\boldsymbol{H}$) *(dashed line) corresponding to the density at the top of column for* $\tau = 0.2, 0.5, 0.8$. *The positions of the solid vertical line and the dashed line stand for the optimal bandwidths obtained from the simulated true risk and the asymptotic approximation respectively over the restricted class* $\mathcal{S}_1$. *The sample size for all the cases is 2000.*

the largest eigenvalue of the bandwidth matrix, restricting it not to converge too slowly. One more thing to notice from these two figures is that the optimal bandwidth chosen from the asymptotic expansion serves as a good approximation to the true optimal bandwidth, as we can see they are quite close in most cases in simulation.

We ran a simulation study to compare the performance of our bandwidth selection method with LSCV for all the 12 densities in Wand and Jones (1993)

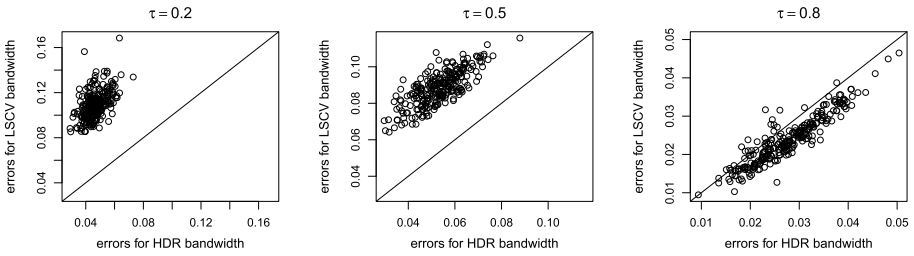

FIG 4. *Plot of simulated errors generated by HDR-tailored bandwidth and LSCV for the sharp mode density* (11). *The horizontal axis stands for errors of HDR bandwidth and vertical axis stands for errors of LSCV bandwidth.*

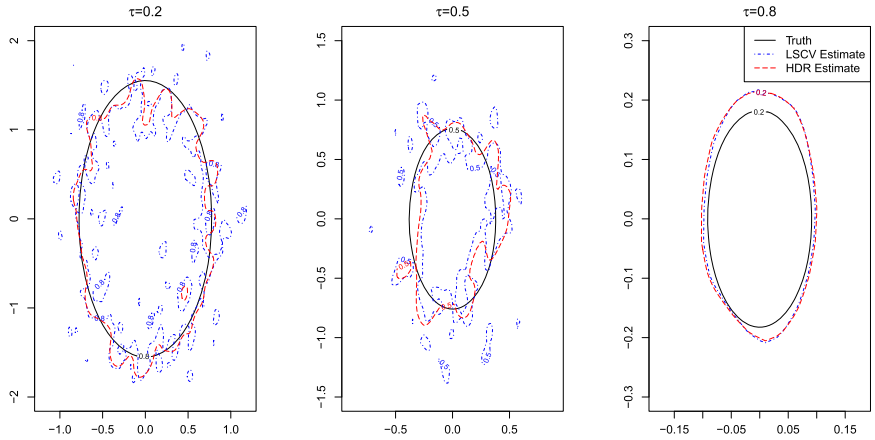

FIG 5. *Plot of boundaries of true HDR, HDR estimated by HDR bandwidth and HDR estimated by LSCV bandwidth from one simulated sample with 2000 observations. The three panels correspond to* $\tau = 0.2, 0.5, 0.8$ *respectively.*

and for density (11). For each density function, 250 Monte Carlo samples with 2000 observations were generated. For each sample, we estimated the 0.2, 0.5, 0.8 HDR with bandwidth matrices chosen by our method and LSCV respectively. The HDR error $\mu_{f_0}\{\mathcal{L}_\tau \Delta \widehat{\mathcal{L}}_{\tau,\boldsymbol{H}}\}$ was calculated for each method in each replication. Figure 4 shows the plot of the estimation errors generated by the two methods for density (11). Figure 5 shows the boundaries of the estimated HDR by HDR bandwidth and by the LSCV bandwidth selector from one of the simulated samples. We can see for $\tau = 0.2, 0.5$, the performance of HDR bandwidth selector outperformed LSCV bandwidth selector greatly for each simulated instance. For $\tau = 0.8$, the HDR bandwidth performed slightly less well than the LSCV bandwidth on average. One hypothesis for why our method suffers when $\tau = .8$ is that Assumption D1b requires that $\|\nabla f_0\| > 0$ in a neighborhood of the HDR. However, when $\tau = .8$, f_0 is close to having gradient zero on the true HDR which is close to the density mode.

It is worth noticing in Figure 5 that the HDR estimated by our method discovers the true underlying topological structure of the density, while the HDR estimated by LSCV does a very poor job of revealing the topological structure when $\tau = .2$ or .5 (the LSCV estimates have many spurious separate connected components rather than a single one).

Applying the Wilcoxon signed rank test to the simulated paired errors generated by our HDR bandwidth and LSCV bandwidth showed that for $\tau = 0.2$, our method outperformed LSCV for 12 out of 13 density functions; for $\tau = 0.5$, our method did better for 8 out of 13 density functions; for $\tau = 0.8$, our method did better in 8 out of 13 density functions.

Note that for any given fixed density, it is likely to be the case for some HDR that the MISE-optimal bandwidth and the HDR-optimal bandwidth will approximately coincide. Thus we may not expect our method to be better than LSCV for all densities and levels simultaneously. Of course, in practice one does not know whether LSCV will work well for the τ value one is interested in. Our HDR method appears to work well for lower τ values, which are the useful values in many applications of HDR estimation. For example in novelty detection, the value of τ equals the probability of type-I error which is often set to be 0.05 or 0.1; in clustering analysis, τ corresponds to fraction of the data that will be discarded during analysis and is also set to be a value close to 0. As mentioned in the previous paragraph, this may be related to the assumption that $\|\nabla f_0\| > 0$ on the HDR boundary. Relaxing this assumption is an important direction for future work, but seems likely to involve somewhat different approximations than the ones used in this paper.

4.2. Real data analysis

We now discuss two real datasets. The Wisconsin Diagnostic Breast Cancer data contains 699 instances of breast cancer cases with 458 of them being benign instances and 241 being malignant instances. Nine cancer-related features were measured for each instance. For the Banknote Authentication data, images were taken of 1372 banknotes, some fake and some genuine. Wavelet transformation tools were used to extract four descriptive features of the images. For both datasets, we reduced the original features to the first two principal components. We apply our method to perform novelty detection for the two data sets. Novelty detection is like a classification problem where only the "normal" class is observed in the training data. Then, for a new data point $\boldsymbol{x}_{\text{new}}$, we want to test the null hypothesis $H_0 : \boldsymbol{x}_{\text{new}}$ is a normal point (or, alternatively, to classify $\boldsymbol{x}_{\text{new}}$ as "normal" or "anomalous"). For level set (HDR) based novelty detection, we can consider an oracle decision rule, or acceptance region, $A := \{\boldsymbol{x} : f_0(\boldsymbol{x}) \geq c\}$ (based on knowing f_0); if $f_0(\boldsymbol{x}_{\text{new}}) \in A$, we accept the null hypothesis, and we reject otherwise. For the breast cancer data, "normal" means healthy, and for the banknote data, "normal" means genuine. If we take $c = f_\tau$, then the oracle decision rule will have type-I error, or False Positive Rate (FPR), of τ (under a regularity condition). Additionally, under regularity conditions, A has the minimum volume of any acceptance rule with FPR of τ, since HDR's are minimum

volume sets (Garcia et al., 2003). This property is beneficial for controlling the type-II error rate, or False Negative Rate (although the actual False Negative Rate depends on the unknown "anomaly" distribution).

In this section, for each of the two data sets we use a KDE with our bandwidth selection procedure to estimate an HDR based on the "normal" class data and use the estimated HDR to perform classification. We delete the observations with missing values for any covariates and randomly split the data set into two parts, training data and testing data. For the Wisconsin Breast Cancer data, 345 benign instances are contained in the training data and 200 (with half being benign and another half being malignant) are contained in the testing data. For the Banknote Authentication data, 400 genuine instances are contained in the training data and again, 200 (with half being genuine and another half being fake) are contained in the testing data. We estimate the 90% HDR using our method based on the training data. The first row of Figure 6 shows the plot of the data and the boundaries of the 90% HDR which are the decision boundaries for the two classification problems. The asymptotic FPR in these two classification problems is $\tau = 0.1$. For the Wisconsin Breast Cancer data, on the test data, the observed FPR is 0.09 and the True Positive Rate (TPR) is 0.99. For the Banknote Authentication data, the observed FPR is 0.04, and the observed TPR is 0.61. We also generated full ROC curves for the two datasets which are shown in the second row of Figure 6. The ROC curves are based on 30 different splits of the data into training and test sets (with the reported FPR and TPR given by the averages over the 30 test sets). The ROC curve clearly shows that the Wisconsin Breast Cancer data is an example where HDR-based anomaly detection is highly effective. The Banknote data is not as easy for our method; it may be the case that using an HDR based on all four variables improves the classification performance. We leave the very interesting question of how best to combine HDR-based classification with dimension reduction for future work.

5. Discussion

In this paper, we derive asymptotic expansions of the symmetric risk for LS estimation and HDR estimation based on kernel density estimators. We provide an efficient bandwidth selection procedure using a plug-in strategy. We also study by theory and by simulation the performance of our bandwidth selector. Simulation studies show that both our asymptotic expansion and our bandwidth selector are effective tools. The two asymptotic risk approximations we provide may also be useful in the analysis of other procedures, developed in future work, for doing LS or HDR bandwidth selection.

As discussed in the Introduction, the interesting paper Qiao (2018) also considers problems of bandwidth selection for KDE's via minimizing asymptotic expansions of risk functions that are based on loss functions related to level sets. Qiao (2018) does not consider HDR estimation. Qiao (2018) does consider the LS estimation problem. Our Theorem 2.1 is similar to Qiao (2018)'s

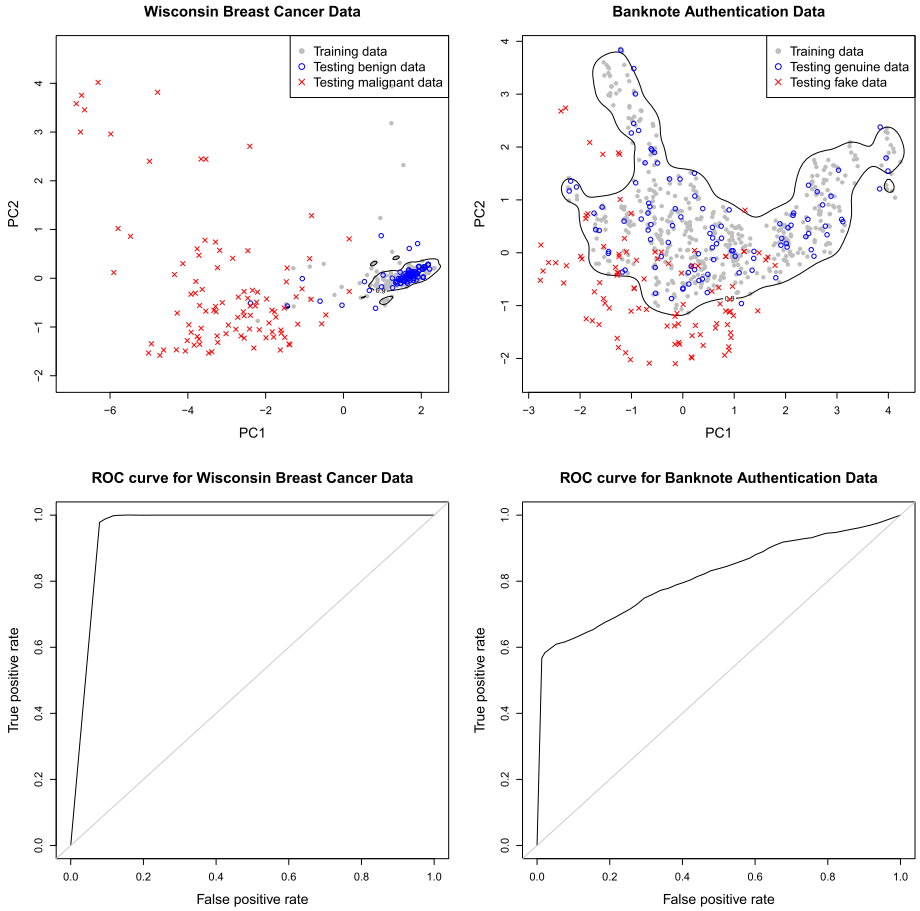

FIG 6. *Plot of data and boundary of estimated* 90% *HDR for the Wisconsin Diagnostic Breast Cancer Data and Banknote Authentication Data. Solid dots correspond to training data, circles are testing data of normal instances and crosses are testing data of anomaly instances. The two panels in the second row are the corresponding ROC curves for the two classification problems.*

Corollary 3.1; both results consider the LS estimation setting, and give risk expansions based on loss functions that are given by integrating the symmetric set differences against f_0 (or against something similar). Our theorem requires only that f_0 have two continuous derivatives in a neighborhood of $\beta(c)$ (which we believe to be approximately the weakest possible conditions), whereas Qiao (2018) requires four continuous derivatives. On the other hand, Qiao (2018) allows for using higher order kernels if one has higher order smoothness of f_0. While Qiao (2018)'s Corollary 3.1 studies the same risk function approximation, LS$(\cdot)$, that we study in our Theorem 2.1, Qiao (2018) does not present any algorithm for minimizing LS$(\cdot)$ and thus presents no simulations related to

LS(·). Rather, Qiao (2018) focuses more attention on a different risk function (the "excess risk") approximation that allows for an analytic solution, at least when $d = 2$.

There are many interesting avenues for extending the work done in the present paper. We describe a few here.

(A). (Regression and classification) In the present paper we have considered only the density estimation context, but estimation of level sets of regression functions estimated by kernel-based methods is also interesting, as is consideration of classification problems.
Regression level set estimation has received less attention than density level set estimation, although it has been studied in some settings; Cavalier (1997) studies multivariate nonparametric regression level set minimax rates of convergence.
One method for classification is to estimate densities for different classes and then classify a point by the class density having highest value at the point. In that case, rather than estimating a level set of one density, one is estimating the 0 level set of a difference of two densities. Mason and Polonik (2009, page 1110) discuss this approach to classification. In the context of an application in flow cytometry, Duong, Koch and Wand (2009) also study estimation of HDR's of density differences (without specifically focusing on classification). We believe the methods of this paper can be extended to those contexts.

(B). (Topological data analysis and critical points) Another important avenue of research is to consider modifications of the assumptions under which our approximations hold. Level set estimation is one of the main tools in topological data analysis (TDA). Estimation of LS's which have zero gradient (at some points) on the boundary (which is ruled out by our assumptions) is of great interest in TDA, because the topology of level sets can change as the level crosses critical points (points having zero gradient). In fact, in the context of using tools based on level set estimates, Wasserman (2016, Section 5) states that "the problem of choosing tuning parameters is one of the biggest open challenges in TDA". Thus, developing tools for bandwidth selection when the gradient is zero would be very useful for TDA. Unfortunately, at points where the gradient is zero we cannot apply the inverse function theorem which is used in Lemma A.1 (implicitly) and by several results in Appendix B, so a very different analysis than the one we completed here may be necessary in such cases. In general, there are very few theoretical works on level set estimation at levels that contain critical values (points where ∇f_0 is 0). In fact, the only one we know of is Chen (2016), in which a rate of convergence of $\lambda\left\{\mathcal{L}(c)\Delta\widehat{\mathcal{L}}_{\boldsymbol{H}}(c)\right\}$ (where λ is Lebesgue measure) is derived.

(C). (MCMC level sets) The work in this paper is restricted to the case where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent. An important extension is to allow the $\boldsymbol{X}_i$ to be samples from a Markov chain. It is well known that KDE's often work similarly when the data exhibit weak dependence as when they are

independent (Wand and Jones, 1995). This would allow our tools for HDR estimation to be used to form credible regions based on Markov chain Monte Carlo output in Bayesian statistical analyses. At present, ad-hoc methods are often used for forming credible regions based on Markov chain Monte Carlo output.

Appendix A: Proof of main results

A.1. Proof of Theorem 2.2

First, we observe that

$$\begin{aligned}\mu_{f_0}(\mathcal{L}_\tau \Delta \hat{\mathcal{L}}_{\tau,\boldsymbol{H}}) &= \int_{\mathbb{R}^d} f_0(\boldsymbol{x}) \left| \mathbb{1}_{\{\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) \geq \widehat{f}_{\tau,n}\}} - \mathbb{1}_{\{f_0(\boldsymbol{x}) \geq f_{\tau,0}\}} \right| d\boldsymbol{x} \\ &= \int_{\mathcal{L}_\tau^c} f_0(\boldsymbol{x}) \mathbb{1}_{\{\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) \geq \widehat{f}_{\tau,n}\}} d\boldsymbol{x} + \int_{\mathcal{L}_\tau} f_0(\boldsymbol{x}) \mathbb{1}_{\{\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) < \widehat{f}_{\tau,n}\}} d\boldsymbol{x}.\end{aligned}$$

Then by Tonelli's Theorem (Folland, 1999, Theorem 2.37), we have

$$\begin{aligned}\mathbb{E}\left[\mu_{f_0}\{\mathcal{L}_\tau \Delta \widehat{\mathcal{L}}_{\tau,\boldsymbol{H}}\}\right] &= \int_{\mathcal{L}_\tau^c} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) \geq \widehat{f}_{\tau,n}\right) d\boldsymbol{x} \\ &\quad + \int_{\mathcal{L}_\tau} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) < \widehat{f}_{\tau,n}\right) d\boldsymbol{x}.\end{aligned} \tag{12}$$

For a density function f on $\mathbb{R}^d$, let $f_\tau(f) := \inf\{\boldsymbol{y} \geq 0 : \int_{\mathbb{R}^d} f(\boldsymbol{x}) \mathbb{1}_{\{f(\boldsymbol{x}) \geq \boldsymbol{y}\}} d\boldsymbol{x} \leq 1 - \tau\}$. By this definition, $f_{\tau,0} \equiv f_\tau(f_0)$. The following lemma bounds the modulus of continuity of f_τ when the difference between two density functions is sufficiently small.

Lemma A.1. *Let the assumptions of Theorem 2.2 hold. Let $\tilde{f}$ be another uniformly continuous density function on $\mathbb{R}^d$ and $\tilde{f}_\tau \equiv f_\tau(\tilde{f})$. Then there exists a constant $C_1 \geq 1$ such that for all $\varepsilon > 0$ sufficiently small, $|\tilde{f}_\tau - f_{\tau,0}| \leq C_1 \varepsilon$ whenever $\|\tilde{f} - f_0\|_\infty \leq \varepsilon$.*

It is intuitively believable that when the sample size n is sufficiently large, the values of the two integrals on the right of (12) are mostly governed by the integrals over a small neighborhood of β_τ. To shrink the region of integration, for $\delta > 0$, and for a given level $t > 0$, we let $\beta^\delta(t) := \bigcup_{\boldsymbol{x} \in \beta(t)} B(\boldsymbol{x}, \delta)$, and $\beta_\tau^\delta \equiv \beta^\delta(f_{\tau,0})$. We also let

$$\mathcal{L}_\delta(f_{\tau,0}) := \bigcup_{\boldsymbol{x} \in \mathcal{L}_\tau} B(\boldsymbol{x}, \delta) \quad \text{and} \quad \mathcal{L}_{-\delta}(f_{\tau,0}) := \mathcal{L}(f_{\tau,0}) \backslash \beta_\tau^\delta.$$

Then we can shrink the integral region using the following lemma.

Lemma A.2. *Let the assumptions of Theorem 2.2 hold. Then for a sequence $\delta_n > 0$ converging to 0 such that $\lambda_{\max}(\boldsymbol{H}) = o(\delta_n)$, we will have*

$$\begin{aligned}\int_{\mathcal{L}_{\delta_n}(f_{\tau,0})^c} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) \geq \widehat{f}_{\tau,n}\right) d\boldsymbol{x} \\ + \int_{\mathcal{L}_{-\delta_n}(f_{\tau,0})} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) < \widehat{f}_{\tau,n}\right) d\boldsymbol{x}\end{aligned} \tag{13}$$

is $o(n^{-1})$ as $n \to \infty$.

The definition of $\widehat{f}_{\tau,n}$ is simple and straightforward, however there is no explicit form for this quantity. So we want to seek an asymptotic expansion for $\widehat{f}_{\tau,n}$. For a uniformly continuous density f on $\mathbb{R}^d$ and $y \geq 0$, we define

$$\psi(f, y) := \int_{\mathbb{R}^d} f(\boldsymbol{x}) \mathbb{1}_{\{f(\boldsymbol{x}) \geq y\}} \, d\boldsymbol{x}.$$

First, we observe for $\varepsilon > 0$ sufficiently small,

$$\begin{aligned}&\left|\psi(f_0, f_{\tau,0} + \varepsilon) - \psi(f_0, f_{\tau,0}) - \varepsilon \int_{\beta_\tau} \frac{f_0(\boldsymbol{x})}{\|\nabla f_0(\boldsymbol{x})\|} \, d\mathcal{H}(\boldsymbol{x})\right| \\ &\quad = \left|\int_{\mathbb{R}^d} f_0(\boldsymbol{x}) \mathbb{1}_{\{f_{\tau,0} \leq f_0(\boldsymbol{x}) \leq f_{\tau,0} + \varepsilon\}} \, d\boldsymbol{x} - \varepsilon \int_{\beta_\tau} \frac{f_0(\boldsymbol{x})}{\|\nabla f_0(\boldsymbol{x})\|} \, d\mathcal{H}(\boldsymbol{x})\right| = O(\varepsilon^2),\end{aligned} \tag{14}$$

as $\varepsilon \searrow 0$, where the last line comes from a similar argument of (67) and (68). A similar argument shows the same result when $\varepsilon \nearrow 0$. Next, we look at

$$\begin{aligned}&\left|\psi(\tilde{f}, \tilde{f}_\tau) - \psi(f_0, \tilde{f}_\tau) - f_{\tau,0} \int_{\beta_\tau} \frac{g}{\|\nabla f_0\|} \, d\mathcal{H} - \int_{\mathcal{L}_\tau} g \, d\lambda\right| \\ &\quad = \left|\int \tilde{f} \mathbb{1}_{\{\tilde{f} \geq \tilde{f}_\tau\}} \, d\lambda - \int f_0 \mathbb{1}_{\{f_0 \geq \tilde{f}_\tau\}} \, d\lambda - f_{\tau,0} \int_{\beta_\tau} \frac{g}{\|\nabla f_0\|} \, d\mathcal{H} - \int_{\mathcal{L}_\tau} g \, d\lambda\right| \\ &\quad = \left|\int f_0 (\mathbb{1}_{\{\tilde{f} \geq \tilde{f}_\tau\}} - \mathbb{1}_{\{f_0 \geq \tilde{f}_\tau\}}) \, d\lambda - f_{\tau,0} \int_{\beta_\tau} \frac{g}{\|\nabla f_0\|} \, d\mathcal{H}\right. \\ &\qquad \left. + \int g (\mathbb{1}_{\{\tilde{f} \geq \tilde{f}_\tau\}} - \mathbb{1}_{\{f_0 \geq f_{\tau,0}\}}) \, d\lambda\right|,\end{aligned} \tag{15}$$

where $g(\boldsymbol{x}) = \tilde{f}(\boldsymbol{x}) - f_0(\boldsymbol{x})$. For the first integral on the last line, since $\mathbb{1}_{\{\tilde{f} \geq \tilde{f}_\tau\}} - \mathbb{1}_{\{f_0 \geq \tilde{f}_\tau\}} \neq 0$ indicates that $\tilde{f}(\boldsymbol{x}) \geq \tilde{f}_\tau, f_0(\boldsymbol{x}) < \tilde{f}_\tau$ or $\tilde{f}(\boldsymbol{x}) < \tilde{f}_\tau, f_0(\boldsymbol{x}) \geq \tilde{f}_\tau$, we have $f_0(\boldsymbol{x}) \in [\tilde{f}_\tau - |g(\boldsymbol{x})|, \tilde{f}_\tau + |g(\boldsymbol{x})|]$. Combining (16) with our result in Lemma A.1 yields

$$f_0(\boldsymbol{x}) = f_\tau + O(\|g\|_\infty), \tag{16}$$

for $\boldsymbol{x} \in \{\boldsymbol{y} : \tilde{f}(\boldsymbol{y}) \geq \tilde{f}_\tau\} \Delta \{\boldsymbol{y} : f_0(\boldsymbol{y}) \geq \tilde{f}_\tau\}$. Next we need the following lemmas.

Lemma A.3. *Let the assumptions of Theorem 2.2 hold and the notation be as defined above. As* $\|g\|_\infty^2 + \|g\|_\infty \|\nabla g\|_\infty \to 0$, *we have*

$$\int \mathbb{1}_{\{\tilde{f} \geq \tilde{f}_\tau\}} - \mathbb{1}_{\{f_0 \geq \tilde{f}_\tau\}} \, d\lambda = \int_{\beta_\tau} \frac{g}{\|\nabla f_0\|} \, d\mathcal{H} + O(\|g\|_\infty^2 + \|g\|_\infty \|\nabla g\|_\infty). \quad (17)$$

Lemma A.4. *Let the assumptions of Theorem 2.2 hold and the notation be as defined above. As* $\|g\|_\infty^2 + \|g\|_\infty \|\nabla g\|_\infty \to 0$, *we have*

$$\int_{\mathbb{R}^d} g(\boldsymbol{x}) \left(\mathbb{1}_{\{\tilde{f}(\boldsymbol{x}) \geq \tilde{f}_\tau\}} - \mathbb{1}_{\{f_0(\boldsymbol{x}) \geq f_\tau\}} \right) d\boldsymbol{x} = O(\|g\|_\infty^2).$$

Now with Lemma A.3, A.4 and (16), we can see that (15) equals $O(\|g\|_\infty^2 + \|g\|_\infty \|\nabla g\|_\infty)$. Note that if $\|\nabla g\|_\infty \to 0$, then $\psi(\tilde{f}, \tilde{f}_\tau) = 1 - \tau$. Combining this with (14) and the order of (15), we have

$$\begin{aligned} 0 &= \psi(\tilde{f}, \tilde{f}_\tau) - \psi(f, f_{\tau,0}) \\ &= \psi(\tilde{f}, \tilde{f}_\tau) - \psi(f, \tilde{f}_\tau) + \psi(f, \tilde{f}_\tau) - \psi(f, f_{\tau,0}) \\ &= -(\tilde{f}_\tau - f_{\tau,0}) f_{\tau,0} \int_{\beta_\tau} \frac{1}{\|\nabla f_0\|} \, d\mathcal{H} + f_{\tau,0} \int_{\beta_\tau} \frac{g}{\|\nabla f_0\|} \, d\mathcal{H} \\ &\quad + \int_{\mathcal{L}_\tau} g \, d\boldsymbol{x} + O(\|g\|_\infty^2 + \|g\|_\infty \|\nabla g\|_\infty) \end{aligned} \quad (18)$$

as $\|g\|_\infty^2 + \|g\|_\infty \|\nabla g\|_\infty \to 0$. We want to apply (18) with $\tilde{f} = \widehat{f}_{n,\boldsymbol{H}}$, so that $g = \widehat{f}_{n,\boldsymbol{H}} - f_0$. To do this, note by Theorem B.1 that $\|\widehat{f}_{n,\boldsymbol{H}} - \mathbb{E}\widehat{f}_{n,\boldsymbol{H}}\|_\infty = O_{\text{a.s.}}\left(\sqrt{\frac{\log|\boldsymbol{H}|^{-1/2}}{n|\boldsymbol{H}|^{1/2}}}\right)$, $\|\nabla \widehat{f}_{n,\boldsymbol{H}} - \mathbb{E}\nabla \widehat{f}_{n,\boldsymbol{H}}\|_\infty = O_{a.s.}\left(\sqrt{\frac{\log|\boldsymbol{H}|^{-1/2}}{n|\boldsymbol{H}|^{1/2}\lambda_{\min}(\boldsymbol{H})}}\right)$, by (71), $\|\mathbb{E}(\widehat{f}_{n,\boldsymbol{H}}) - f_0\|_\infty = O\{\lambda_{\max}(\boldsymbol{H})\}$. We also have that $\|\mathbb{E}\nabla \widehat{f}_{n,\boldsymbol{H}} - \nabla f_0\|_\infty = O\{\lambda_{\max}^{1/2}(\boldsymbol{H})\}$. Then applying the above results, we have

$$\begin{aligned} \widehat{f}_{\tau,n} - f_{\tau,0} = w_0 \left\{ \int_{\beta_\tau} \frac{\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) - f_0(\boldsymbol{x})}{\|\nabla f_0(\boldsymbol{x})\|} \, d\mathcal{H}(\boldsymbol{x}) + \frac{1}{f_{\tau,0}} \int_{\mathcal{L}_\tau} \widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) - f_0(\boldsymbol{x}) \, d\boldsymbol{x} \right\} \\ + O_p\left(\frac{\log|\boldsymbol{H}|^{-1/2}}{n|\boldsymbol{H}|^{1/2}\sqrt{\lambda_{\max}(\boldsymbol{H})}} + \lambda_{\max}^{3/2}(\boldsymbol{H}) \right). \end{aligned} \quad (19)$$

Note from (2) and (19), for fixed $\boldsymbol{x}$, $\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) - \widehat{f}_{\tau,n}$ can be expressed as the average of i.i.d. random variables with a negligible stochastic error term. This motivates us to use the Berry-Essen Theorem (Ferguson, 1996) to approximate the two probabilities appearing on the right of (13). In order to do so, we will need to approximate the mean and variance of $\widehat{f}_{\tau,n}$, which we do in the next lemmas.

Lemma A.5. *Let the assumptions of Theorem 2.2 hold and the notation be as defined above. Then we have*

$$\mathbb{E}\widehat{f}_{\tau,n} - f_{\tau,0} = w_0 \{V_1(\boldsymbol{H}) + V_2(\boldsymbol{H})\} + o\{\text{tr}(\boldsymbol{H})\}, \quad (20)$$

as $n \to \infty$.

Recall V_1 and V_2 are defined in Theorem 2.2. The next lemma shows $\operatorname{Var}\widehat{f}_{\tau,n}$ is negligible compared with other terms in the expansion.

Lemma A.6. *Let the assumptions of Theorem 2.2 hold and the notation be as defined above. Then* $\operatorname{Var}\widehat{f}_{\tau,n} = o(n^{-1}|\boldsymbol{H}|^{-1/2})$.

Now according to Lemma A.2 and (12), we have

$$\begin{aligned}
&\mathbb{E}\mu_{f_0}(\mathcal{L}_\tau \Delta \hat{\mathcal{L}}_{\tau,\boldsymbol{H}}) \\
&\quad = \int_{\mathcal{L}_\tau^c \setminus \mathcal{L}_{\delta_n}(f_\tau)^c} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) \geq \widehat{f}_{\tau,n}\right) d\boldsymbol{x} \\
&\qquad + \int_{\mathcal{L}_\tau \setminus \mathcal{L}_{-\delta_n}(f_{\tau,0})} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) < \widehat{f}_{\tau,n}\right) d\boldsymbol{x} + o\left(n^{-1}\right) \\
&\quad = \int_{\beta_\tau^{\delta_n}} f_0(\boldsymbol{x}) \left| P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) < \widehat{f}_{\tau,n}\right) d\boldsymbol{x} - \mathbb{1}_{\{f_0(\boldsymbol{x}) < f_{\tau,0}\}} \right| d\boldsymbol{x} + o\left(n^{-1}\right).
\end{aligned}$$

Then by Lemma B.4 when δ_n is small enough, the dominating term on the last line above is equal to

$$\int_{\beta_\tau} \int_{-\delta_n}^{\delta_n} f_0(\boldsymbol{x} + t u_{\boldsymbol{x}}) \left| P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x} + t u_{\boldsymbol{x}}) < \widehat{f}_{\tau,n}\right) - \mathbb{1}_{\{f_0(\boldsymbol{x} + t u_{\boldsymbol{x}}) < f_{\tau,0}\}} \right| dt d\mathcal{H}(\boldsymbol{x}) + O(\delta_n^2), \tag{21}$$

where $u_{\boldsymbol{x}} := -\nabla f_0(\boldsymbol{x}) / \|\nabla f_0(\boldsymbol{x})\|$ is the unit outer normal vector of β_τ at $\boldsymbol{x}$. Now for a fixed $\boldsymbol{x} \in \beta_\tau$, let $\boldsymbol{x}^t = \boldsymbol{x} + \frac{t}{\sqrt{n|\boldsymbol{H}|^{1/2}}} u_{\boldsymbol{x}}$ for $t \in [-\sqrt{n|\boldsymbol{H}|^{1/2}}\delta_n, \sqrt{n|\boldsymbol{H}|^{1/2}}\delta_n]$, we see (21) equals

$$\frac{1}{\sqrt{n|\boldsymbol{H}|^{1/2}}} \int_{\beta_\tau} \int_{-\sqrt{n|\boldsymbol{H}|^{1/2}}\delta_n}^{\sqrt{n|\boldsymbol{H}|^{1/2}}\delta_n} f_0\left(\boldsymbol{x}^t\right) \left| P\left(\widehat{f}_{n,\boldsymbol{H}}\left(\boldsymbol{x}^t\right) < \widehat{f}_{\tau,n}\right) - \mathbb{1}_{\{t>0\}} \right| dt d\mathcal{H}(\boldsymbol{x}) \tag{22}$$

$$+ O(\delta_n^2). \tag{23}$$

By Taylor Expansion, we have

$$f_0\left(\boldsymbol{x} + \frac{t}{\sqrt{n|\boldsymbol{H}|^{1/2}}} u_{\boldsymbol{x}}\right) = f_0(\boldsymbol{x}) + \nabla f_0\left(\boldsymbol{x} + \frac{st}{\sqrt{n|\boldsymbol{H}|^{1/2}}} u_{\boldsymbol{x}}\right)' \frac{t}{\sqrt{n|\boldsymbol{H}|^{1/2}}} u_{\boldsymbol{x}},$$

for some $s \in [0,1]$. Since by Assumption D1b, f_0 has bounded first derivatives, we see the dominating term in (22) equals

$$\frac{f_{\tau,0}}{\sqrt{n|\boldsymbol{H}|^{1/2}}} \int_{\beta_\tau} \int_{-\sqrt{n|\boldsymbol{H}|^{1/2}}\delta_n}^{\sqrt{n|\boldsymbol{H}|^{1/2}}\delta_n} \left| P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}^t) < \widehat{f}_{\tau,n}\right) - \mathbb{1}_{\{t>0\}} \right| dt d\mathcal{H}(\boldsymbol{x}) + O(\delta_n^2), \tag{24}$$

as $n \to \infty$.

We can further shrink the region of interest by the following lemma.

Lemma A.7. *Let the assumptions of Theorem 2.2 hold and the notation be as defined above. Then for n sufficiently large, $\mathbb{E}\{\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}^t) - \widehat{f}_{\tau,n}\}$ is a strictly monotone function of $t \in [-\sqrt{n|\boldsymbol{H}|^{1/2}}\delta_n, \sqrt{n|\boldsymbol{H}|^{1/2}}\delta_n]$, with a unique zero $t^*_{\boldsymbol{x}}$. For a sequence t_n diverging to infinity and $t_n = O(\sqrt{n|\boldsymbol{H}|^{1/2}}\delta_n)$, let*

$$I^n_{\boldsymbol{x}} = [-\sqrt{n|\boldsymbol{H}|^{1/2}}\delta_n, \sqrt{n|\boldsymbol{H}|^{1/2}}\delta_n] \backslash [t^*_{\boldsymbol{x}} - t_n, t^*_{\boldsymbol{x}} + t_n].$$

We have

$$\int_{\beta_\tau} \int_{I^n_{\boldsymbol{x}}} |P(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}^t) < \widehat{f}_{\tau,n}) - \mathbb{1}_{\{t>0\}}| \, dt d\mathcal{H}(\boldsymbol{x}) \to 0 \tag{25}$$

as $n \to \infty$.

To complete the proof of Theorem 2.2, by (24) and Lemma A.7 it suffices to show that there exists a sequence t_n diverging to infinity slowly such that

$$\begin{aligned} &\frac{f_{\tau,0}}{\sqrt{n|\boldsymbol{H}|^{1/2}}} \int_{\beta_\tau} \int_{t^*_{\boldsymbol{x}} - t_n}^{t^*_{\boldsymbol{x}} + t_n} |P(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}^t) < \widehat{f}_{\tau,n}) - \mathbb{1}_{\{t<0\}}| \, dt d\mathcal{H}(\boldsymbol{x}) \\ &\quad = \mathrm{HDR}(\boldsymbol{H}) + o\left\{(n|\boldsymbol{H}|^{1/2})^{-1/2} + \mathrm{tr}(\boldsymbol{H})\right\}. \end{aligned}$$

For $i = 1, 2, \ldots, n$, let $Z_{ni}(\boldsymbol{x}) = K_{\boldsymbol{H}}(\boldsymbol{x} - \boldsymbol{X}_i)$ and $\bar{Y}_n = n^{-1}\sum_{i=1}^n Y_{ni}$, where

$$\begin{aligned} Y_{ni} = Z_{ni}(\boldsymbol{x}^t) - f_{\tau,0} - \left\{\int_{\beta_\tau} \frac{1}{\|\nabla f_0\|} \, d\mathcal{H}\right\}^{-1} \left\{\int_{\beta_\tau} \frac{Z_{ni}(\boldsymbol{x}) - f_0(\boldsymbol{x})}{\|\nabla f_0(\boldsymbol{x})\|} \, d\mathcal{H}(\boldsymbol{x}) \right. \\ \left. + \frac{1}{f_{\tau,0}} \int_{\mathcal{L}_\tau} Z_{ni}(\boldsymbol{x}) - f_0(\boldsymbol{x}) \, d\boldsymbol{x}\right\}. \end{aligned}$$

Then by (18) and (19), we can write $\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}^t) - \widehat{f}_{\tau,n} = \bar{Y}_n + R_n$, where $R_n - \mathbb{E}(R_n) = o_p\left(\frac{1}{\sqrt{n|\boldsymbol{H}|^{1/2}}}\right)$. By Lemma A.6, we know $\mathrm{Var}(\bar{Y}_n)$ is $O(n^{-1}|\boldsymbol{H}|^{-1/2})$ uniformly in t and $\boldsymbol{x}$. Let t_n diverge slowly such that for fixed $\boldsymbol{x} \in \beta_\tau$,

- $P\left(\frac{|R_n - \mathbb{E}(R_n)|}{\mathrm{Var}^{1/2}(\bar{Y}_n)} > \frac{1}{t_n^2}\right) \le \frac{1}{t_n^2}$ uniformly for $t \in [t^*_{\boldsymbol{x}} - t_n, t^*_{\boldsymbol{x}} + t_n]$.
- $\mathbb{E}(\bar{Y}_n + R_n) = \{\frac{t}{\sqrt{n|\boldsymbol{H}|^{1/2}}}\|\nabla f_0(\boldsymbol{x})\| + D_1(\boldsymbol{x}, \boldsymbol{H}) - D_2(\boldsymbol{x}, \boldsymbol{H})\}\left\{1 + o(t_n^{-2})\right\}$, uniformly for $t \in [t^*_{\boldsymbol{x}} - t_n, t^*_{\boldsymbol{x}} + t_n]$ and $\boldsymbol{x} \in \beta_\tau$, by Assumption D1b part 3.
- $n|\boldsymbol{H}|^{1/2}\,\mathrm{Var}\,\bar{Y}_n = R(K) f_{\tau,0} + o(t_n^{-2})$ uniformly for $t \in [t^*_{\boldsymbol{x}} - t_n, t^*_{\boldsymbol{x}} + t_n]$ and $\boldsymbol{x} \in \beta_\tau$.

Then

$$\begin{aligned} &P(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}^t) < \widehat{f}_{\tau,n}) - \Phi\left(A_{\boldsymbol{x}} t + C_{\boldsymbol{x}}(\boldsymbol{H})\right) \\ &\quad = P(\bar{Y}_n + R_n - \mathbb{E}(\bar{Y}_n + R_n) < -\mathbb{E}(\bar{Y}_n + R_n)) - \Phi\left(A_{\boldsymbol{x}} t + C_{\boldsymbol{x}}(\boldsymbol{H})\right) \\ &\quad \le P\left(\frac{|R_n - \mathbb{E}(R_n)|}{\mathrm{Var}^{1/2}(\bar{Y}_n)} > \frac{1}{t_n^2}\right) + P\left(\frac{\bar{Y}_n - \mathbb{E}(\bar{Y}_n)}{\mathrm{Var}^{1/2}(\bar{Y}_n)} \le \frac{-\mathbb{E}(\bar{Y}_n + R_n)}{\mathrm{Var}^{1/2}(\bar{Y}_n)} + \frac{1}{t_n^2}\right) \end{aligned}$$

$$
\begin{aligned}
&\quad - \Phi\left(A_{\boldsymbol{x}} t + C_{\boldsymbol{x}}(\boldsymbol{H})\right) \\
&= O\left(\frac{1}{t_n^2}\right) + P\left(\frac{\bar{Y}_n - \mathbb{E}(\bar{Y}_n)}{\mathrm{Var}^{1/2}(\bar{Y}_n)} \leq \frac{-\mathbb{E}(\bar{Y}_n + R_n)}{\mathrm{Var}^{1/2}(\bar{Y}_n)} + \frac{1}{t_n^2}\right) \\
&\quad - \Phi\left(A_{\boldsymbol{x}} t + C_{\boldsymbol{x}}(\boldsymbol{H})\right).
\end{aligned}
$$

Applying the Berry-Esseen theorem (Ferguson, 1996) to the last two terms on the last line yields

$$
\begin{aligned}
&\left| P\left(\frac{\bar{Y}_n - \mathbb{E}(\bar{Y}_n)}{\mathrm{Var}^{1/2}(\bar{Y}_n)} \leq \frac{-\mathbb{E}(\bar{Y}_n + R_n)}{\mathrm{Var}^{1/2}(\bar{Y}_n)} + \frac{1}{t_n^2}\right) - \Phi\left(\frac{-\mathbb{E}(\bar{Y}_n + R_n)}{\mathrm{Var}^{1/2}(\bar{Y}_n)} + \frac{1}{t_n^2}\right)\right| \\
&\qquad \leq \frac{C\mathbb{E}|Y_{ni}|^3}{\mathrm{Var}^{3/2}(Y_{ni})\sqrt{n}}.
\end{aligned}
$$

Now since we know $\mathrm{Var}(\bar{Y}_n) = R(K) f_{\tau,0}/(n|\boldsymbol{H}|^{1/2}) + o(n^{-1}|\boldsymbol{H}|^{-1/2})$ uniformly, $\mathrm{Var}(Y_{ni}) = R(K) f_{\tau,0}/(|\boldsymbol{H}|^{1/2}) + o(|\boldsymbol{H}|^{-1/2})$ and it can be shown that $\mathbb{E}|Y_{ni}|^3 = O(|\boldsymbol{H}|^{-1})$, we further have

$$
\begin{aligned}
&\left| P\left(\frac{\bar{Y}_n - \mathbb{E}(\bar{Y}_n)}{\mathrm{Var}^{1/2}(\bar{Y}_n)} \leq \frac{-\mathbb{E}(\bar{Y}_n + R_n)}{\mathrm{Var}^{1/2}(\bar{Y}_n)} + \frac{1}{t_n^2}\right) - \Phi\left(\frac{-\mathbb{E}(\bar{Y}_n + R_n)}{\mathrm{Var}^{1/2}(\bar{Y}_n)} + \frac{1}{t_n^2}\right)\right| \\
&\qquad = O\left(\frac{1}{\sqrt{n|\boldsymbol{H}|^{1/2}}}\right),
\end{aligned}
$$

and then

$$
\begin{aligned}
&P(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}^t) < \widehat{f}_{\tau,n}) - \Phi\left(A_{\boldsymbol{x}} t + C_{\boldsymbol{x}}(\boldsymbol{H})\right) \\
&\qquad \leq O\left(\frac{1}{t_n^2} + \frac{1}{\sqrt{n|\boldsymbol{H}|^{1/2}}}\right) + \Phi\left(\frac{-\mathbb{E}(\bar{Y}_n + R_n)}{\mathrm{Var}^{1/2}(\bar{Y}_n)}\right) - \Phi\left(A_{\boldsymbol{x}} t + C_{\boldsymbol{x}}(\boldsymbol{H})\right),
\end{aligned}
$$

uniformly in t and $\boldsymbol{x}$. A similar argument shows a lower bound of the same order. Now we look at the integrated error

$$
\frac{1}{\sqrt{n|\boldsymbol{H}|^{1/2}}} \int_{\beta_\tau} \int_{t_{\boldsymbol{x}}^* - t_n}^{t_{\boldsymbol{x}}^* + t_n} \left| \Phi\left(\frac{-\mathbb{E}(\bar{Y}_n + R_n)}{\mathrm{Var}^{1/2}(\bar{Y}_n)}\right) - \Phi\left(A_{\boldsymbol{x}} t + C_{\boldsymbol{x}}(\boldsymbol{H})\right) \right| dt\, d\mathcal{H}(\boldsymbol{x}).
$$

We can see that $\left|\Phi\left(\frac{-\mathbb{E}(\bar{Y}_n + R_n)}{\mathrm{Var}^{1/2}(\bar{Y}_n)}\right) - \Phi\left(A_{\boldsymbol{x}} t + C_{\boldsymbol{x}}(\boldsymbol{H})\right)\right|$ is bounded by

$$
\left\{ (t_n + |t_{\boldsymbol{x}}^*|)\|\nabla f_0\|_\infty + \sqrt{n|\boldsymbol{H}|^{1/2}}|D_1(\boldsymbol{x}, \boldsymbol{H})| + \sqrt{n|\boldsymbol{H}|^{1/2}}|D_2(\boldsymbol{x}, \boldsymbol{H})| \right\} o(t_n^{-2}).
$$

uniformly in $\boldsymbol{x}$. From (86) we know $|t_{\boldsymbol{x}}^*|$ is uniformly $O(\sqrt{n|\boldsymbol{H}|^{1/2}}\,\mathrm{tr}(\boldsymbol{H}))$, then

$$
\frac{1}{\sqrt{n|\boldsymbol{H}|^{1/2}}} \int_{\beta_\tau} \int_{t_{\boldsymbol{x}}^* - t_n}^{t_{\boldsymbol{x}}^* + t_n} (t_n + |t_{\boldsymbol{x}}^*|)\|\nabla f_0\|_\infty o(t_n^{-2})\, dt\, d\boldsymbol{x}
$$

$$= o\left(\frac{1}{\sqrt{n|\boldsymbol{H}|^{1/2}}} + \operatorname{tr}(\boldsymbol{H})\right),$$

and similarly

$$\frac{1}{\sqrt{n|\boldsymbol{H}|^{1/2}}} \int_{\beta_\tau} \int_{t^*_{\boldsymbol{x}}-t_n}^{t^*_{\boldsymbol{x}}+t_n} \left\{\sqrt{n|\boldsymbol{H}|^{1/2}}\left(|D_1(\boldsymbol{x},\boldsymbol{H})| + |D_2(\boldsymbol{x},\boldsymbol{H})|\right)\right\} o(t_n^{-2})\, dt\, d\boldsymbol{x}$$
$$= o\left(\frac{1}{\sqrt{n|\boldsymbol{H}|^{1/2}}} + \operatorname{tr}(\boldsymbol{H})\right).$$

So we have

$$\frac{f_{\tau,0}}{\sqrt{n|\boldsymbol{H}|^{1/2}}} \int_{\beta_\tau} \int_{t^*_{\boldsymbol{x}}-t_n}^{t^*_{\boldsymbol{x}}+t_n} \left|P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}^t) < \widehat{f}_{\tau,n}\right) - \mathbb{1}_{\{t<0\}}\right| dt\, d\mathcal{H}(\boldsymbol{x})$$
$$= \frac{f_{\tau,0}}{\sqrt{n|\boldsymbol{H}|^{1/2}}} \int_{\beta_\tau} \int_{t^*_{\boldsymbol{x}}-t_n}^{t^*_{\boldsymbol{x}}+t_n} \left|\Phi\left(A_{\boldsymbol{x}} t + C_{\boldsymbol{x}}(\boldsymbol{H})\right) - \mathbb{1}_{\{t<0\}}\right| dt\, d\mathcal{H}(\boldsymbol{x})$$
$$+ o\left(\frac{1}{\sqrt{n|\boldsymbol{H}|^{1/2}}} + \operatorname{tr}(\boldsymbol{H})\right).$$

It remains to see from Lemma B.3 that

$$\frac{f_{\tau,0}}{\sqrt{n|\boldsymbol{H}|^{1/2}}} \int_{\beta_\tau} \int_{-\infty}^{\infty} \left|\Phi\left(A_{\boldsymbol{x}} t + C_{\boldsymbol{x}}(\boldsymbol{H})\right) - \mathbb{1}_{\{t<0\}}\right| dt\, d\mathcal{H}(\boldsymbol{x})$$
$$= \frac{f_{\tau,0}}{\sqrt{n|\boldsymbol{H}|^{1/2}}} \int_{\beta_\tau} \frac{2\phi(C_{\boldsymbol{x}}(\boldsymbol{H})) + 2\Phi(C_{\boldsymbol{x}}(\boldsymbol{H}))C_{\boldsymbol{x}}(\boldsymbol{H}) - C_{\boldsymbol{x}}(\boldsymbol{H})}{A_{\boldsymbol{x}}}\, d\mathcal{H}(\boldsymbol{x}).$$

A.2. Proof of Theorem 2.1

We also provide a brief proof for Theorem 2.1, which is a simpler and shares the same idea as that of Theorem 2.2. First, we have

$$\mathbb{E}\left[\mu_{f_0}\{\mathcal{L}(c)\Delta\widehat{\mathcal{L}}_{\boldsymbol{H}}(c)\}\right]$$
$$= \mathbb{E}\int_{\mathbb{R}^d} f_0(\boldsymbol{x}) \left|\mathbb{1}_{\{\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x})\geq c\}} - \mathbb{1}_{\{f_0(\boldsymbol{x})\geq c\}}\right| d\boldsymbol{x}$$
$$= \int_{\mathcal{L}(c)^c} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) \geq c\right) d\boldsymbol{x} + \int_{\mathcal{L}(c)} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) < c\right) d\boldsymbol{x}. \tag{26}$$

Like Lemma A.2, we can shrink the region of interest. We show that for each $\delta > 0$ sufficiently small, we have

$$\int_{\mathcal{L}_\delta(c)^c} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) \geq c\right) d\boldsymbol{x} + \int_{\mathcal{L}(c)} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) < c\right) d\boldsymbol{x} = o(n^{-1}),$$

as $n \to \infty$.

Observe that under Assumption D1a if $\delta > 0$ is sufficiently small, then there exists $\epsilon > 0$ s.t $f_0(\boldsymbol{x}) \leq c - \epsilon$ for $\boldsymbol{x} \in \mathcal{L}_\delta(c)^c$ and $f_0(\boldsymbol{x}) \geq c + \epsilon$ for $\boldsymbol{x} \in \mathcal{L}_{-\delta}(c)$. By reducing $\delta > 0$ if necessary, for $\boldsymbol{x} \in \mathcal{L}_\delta(c)^c$,

$$\begin{aligned} P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) \geq c\right) &= P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) - c \geq 0\right) \\ &\leq P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) - c + c - f_0(\boldsymbol{x}) \geq \epsilon\right) \\ &\leq P\left(\|\widehat{f}_{n,\boldsymbol{H}} - f_0\|_\infty \geq \epsilon\right). \end{aligned}$$

Similarly we can show the same bound for $\boldsymbol{x} \in \mathcal{L}_{-\delta}(c)$. Then

$$\begin{aligned} &\int_{\mathcal{L}_\delta(c)^c} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) \geq c\right) d\boldsymbol{x} + \int_{\mathcal{L}_{-\delta}(c)} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) < c\right) d\boldsymbol{x} \\ &\quad \leq P\left(\|\widehat{f}_{n,\boldsymbol{H}} - f_0\|_\infty \geq \epsilon\right) \\ &\quad \leq P\left(\|\widehat{f}_{n,\boldsymbol{H}} - \mathbb{E}\widehat{f}_{n,\boldsymbol{H}}\|_\infty \geq \frac{\epsilon}{2}\right) + P\left(\|\mathbb{E}\widehat{f}_{n,\boldsymbol{H}} - f_0\|_\infty \geq \frac{\epsilon}{2}\right), \end{aligned}$$

where $P(\|\mathbb{E}\widehat{f}_{n,\boldsymbol{H}} - f_0\|_\infty \geq \frac{\epsilon}{2}) = 0$ for n large enough. So with the same argument in proof Lemma A.2, the above quantity is $o(n^{-1})$. Further, we have that for a sequence δ_n converging to 0 such that $\lambda_{\max}(\boldsymbol{H}) = o(\delta_n)$,

$$\begin{aligned} &\int_{\mathcal{L}_{\delta_n}(c)^c} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) \geq c\right) d\boldsymbol{x} + \int_{\mathcal{L}_{-\delta_n}(c)} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) < c\right) d\boldsymbol{x} \\ &\quad = o(n^{-1}). \end{aligned} \tag{27}$$

and we also prove this by showing that $E(\delta, \delta_n)$ which is defined as

$$\begin{aligned} &\int_{\mathcal{L}_{\delta_n}(c)^c} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) \geq c\right) d\boldsymbol{x} + \int_{\mathcal{L}_{-\delta_n}(c)} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) < c\right) d\boldsymbol{x} \\ &- \left\{ \int_{\mathcal{L}_\delta(c)^c} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) \geq c\right) d\boldsymbol{x} + \int_{\mathcal{L}_{-\delta}(c)} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) < c\right) d\boldsymbol{x} \right\} \\ &= \int_{\mathcal{L}_{\delta_n}(c)^c \backslash \mathcal{L}_{\delta_n}(c)^c} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) \geq c\right) d\boldsymbol{x} \\ &\quad + \int_{\mathcal{L}_{-\delta_n}(c) \backslash \mathcal{L}_{-\delta}(c)} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) < c\right) d\boldsymbol{x} \end{aligned}$$

is $o(n^{-1})$ as $n \to \infty$. Note that there exits a constant c_2 small s.t if we take $\epsilon_n = c_2\delta_n$, then we have $|f_0(\boldsymbol{x}) - c| \geq \epsilon_n$ when $\boldsymbol{x} \in \mathcal{L}_{\delta_n}(c)^c \backslash \mathcal{L}_{\delta_n}(c)^c \cup \mathcal{L}_{-\delta_n}(c) \backslash \mathcal{L}_{-\delta}(c)$.Then for $\boldsymbol{x} \in \mathcal{L}_{\delta_n}(c)^c \backslash \mathcal{L}_{\delta_n}(c)^c$,

$$\begin{aligned} P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) \geq c\right) &\leq P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) - c + c - f_0(\boldsymbol{x}) \geq \epsilon_n\right) \\ &\leq P\left(\|\widehat{f}_{n,\boldsymbol{H}} - f_0\|_\infty \geq \epsilon_n\right). \end{aligned}$$

We can derive the same bound for $\boldsymbol{x} \in \mathcal{L}_{-\delta_n}(c) \backslash \mathcal{L}_{-\delta}(c)$. Then

$$\begin{aligned} E(\delta, \delta_n) &\leq P\left(\|\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) - f_0(\boldsymbol{x})\|_\infty \geq \epsilon_n\right) \\ &\leq P\left(\|\widehat{f}_{n,\boldsymbol{H}} - \mathbb{E}\widehat{f}_{n,\boldsymbol{H}}\|_\infty \geq \frac{\epsilon_n}{2}\right) + P\left(\|\mathbb{E}\widehat{f}_{n,\boldsymbol{H}} - f_0\|_\infty \geq \frac{\epsilon_n}{2}\right) \end{aligned}$$

is $o(n^{-1})$ when n is large enough.

Now the risk function can be expressed as

$$\begin{aligned} \mathbb{E}\mu_{f_0}\left\{\mathcal{L}(c)\Delta\widehat{\mathcal{L}}(c)\right\} &= \int_{\mathcal{L}(c)^c \backslash \mathcal{L}_{\delta_n}(c)^c} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) \geq c\right) d\boldsymbol{x} \\ &\quad + \int_{\mathcal{L}(c) \backslash \mathcal{L}_{-\delta_n}(c)} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) < c\right) d\boldsymbol{x} + o(n^{-1}) \\ &= \int_{\beta(c)^{\delta_n}} f_0(\boldsymbol{x}) \left| P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) < c\right) - \mathbb{1}_{\{f_0(\boldsymbol{x}) < c\}} \right| d\boldsymbol{x} \\ &\quad + o(n^{-1}). \end{aligned}$$

Then according to Lemma B.4, when δ_n is small enough

$$\begin{aligned} &\int_{\beta(c)^{\delta_n}} f_0(\boldsymbol{x}) \left| P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) < c\right) - \mathbb{1}_{\{f_0(\boldsymbol{x}) < c\}} \right| d\boldsymbol{x} \\ &= \int_{\beta(c)} \int_{-\delta_n}^{\delta_n} f_0(\boldsymbol{x} + tu_x) \left| P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x} + tu_x) < c\right) - \mathbb{1}_{\{f_0(\boldsymbol{x}+tu_x) < c\}} \right| dt d\mathcal{H}(\boldsymbol{x}) \\ &\quad + O(\delta_n^2), \end{aligned}$$

where $u_{\boldsymbol{x}}$ is the unit normal outer vector at $\boldsymbol{x} \in \beta(c)$. And by simple transformation,

$$\begin{aligned} &\int_{\beta(c)} \int_{-\delta_n}^{\delta_n} f_0(\boldsymbol{x} + tu_x) \left| P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x} + tu_x) < c\right) - \mathbb{1}_{\{f_0(\boldsymbol{x}+tu_x) < c\}} \right| dt d\mathcal{H}(\boldsymbol{x}) \\ &= \frac{1}{\sqrt{n|\boldsymbol{H}|^{1/2}}} \int_{\beta(c)} \int_{-\sqrt{n|\boldsymbol{H}|^{1/2}}\delta_n}^{\sqrt{n|\boldsymbol{H}|^{1/2}}\delta_n} f_0\left(\boldsymbol{x}^t\right) \left| P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}^t) < c\right) - \mathbb{1}_{\{t<0\}} \right| dt d\mathcal{H}(\boldsymbol{x}) \\ &= \frac{c}{\sqrt{n|\boldsymbol{H}|^{1/2}}} \int_{\beta(c)} \int_{-\sqrt{n|\boldsymbol{H}|^{1/2}}\delta_n}^{\sqrt{n|\boldsymbol{H}|^{1/2}}\delta_n} \left| P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}^t) < c\right) - \mathbb{1}_{\{t<0\}} \right| dt d\mathcal{H}(\boldsymbol{x}) \\ &\quad + O(\delta_n^2). \end{aligned}$$

To further shrink the intervals of interest, we also argue that when n is large enough, $\mathbb{E}\{\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}^t\}$ is a strictly monotone function of t when t falls within the interval $[-\sqrt{n|\boldsymbol{H}|^{1/2}}\delta_n, \sqrt{n|\boldsymbol{H}|^{1/2}}\delta_n]$ with a unique zero t_x^*. Now we claim for a sequence t_n diverging to infinity,

$$\int_{\beta(c)} \int_{I_{\boldsymbol{x}}^n} \left| P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}^t) < c\right) - \mathbb{1}_{\{t<0\}} \right| dt d\mathcal{H}(\boldsymbol{x}) \to 0$$

as $n \to \infty$, where $I_x^n = [-\sqrt{n|\boldsymbol{H}|^{1/2}}\delta_n, \sqrt{n|\boldsymbol{H}|^{1/2}}\delta_n] \backslash [t_x^* - t_n, t_x^* + t_n]$. For detail of proof, please refer to the proof of Theorem 2.2.

Now by previous steps, we know

$$\begin{aligned}
&\mathbb{E}\mu_{f_0}\{\mathcal{L}(c)\Delta\hat{\mathcal{L}}(c)\} \\
&\quad = \frac{c}{\sqrt{n|\boldsymbol{H}|^{1/2}}} \int_{\beta(c)} \int_{t_x^*-t_n}^{t_x^*+t_n} \left| P\left(\widehat{f}_{n,\boldsymbol{H}}\left(\boldsymbol{x}^t\right) < c\right) - \mathbb{1}_{\{t<0\}} \right| dt d\mathcal{H}(\boldsymbol{x}) \\
&\qquad + o\left(\frac{1}{\sqrt{n|\boldsymbol{H}|^{1/2}}}\right),
\end{aligned}$$

To complete the proof, it suffices to show the dominating term above is equal to $\mathrm{LS}(\boldsymbol{H}) + o(1/\sqrt{n|\boldsymbol{H}|^{1/2}})$. Let $Z_{ni}(\boldsymbol{x}) = K_{\boldsymbol{H}}(\boldsymbol{x} - \boldsymbol{X}_i)$ and $Y_{ni} = Z_{ni}(\boldsymbol{x}^t) - c$. Then $\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}^t) - c = \bar{Y}_n$. Now let t_n diverge slowly such that

$$\mathbb{E}(\bar{Y}_n) = \left\{ \frac{t}{\sqrt{n|\boldsymbol{H}|^{1/2}}} \|\nabla f_0(\boldsymbol{x})\| + D_1(\boldsymbol{x}, \boldsymbol{H}) \right\} \{1 + o(t_n^{-2})\}, \tag{28}$$

by Assumption D1b part 3, and

$$n|\boldsymbol{H}|^{1/2} \operatorname{Var} \bar{Y}_n = R(K)c + o(t_n^{-2}), \tag{29}$$

uniformly for $\in [t_{\boldsymbol{x}}^* - t_n, t_{\boldsymbol{x}}^* + t_n]$ and $\boldsymbol{x} \in \beta(c)$. Then

$$\begin{aligned}
&P\left(\widehat{f}_{n,\boldsymbol{H}}\left(\boldsymbol{x}^t\right) < c\right) - \Phi(A_{\boldsymbol{x}} t + B_{\boldsymbol{x}}(\boldsymbol{H})) \\
&\quad = P\left(\frac{\bar{Y}_n - \mathbb{E}\bar{Y}_n}{\operatorname{Var}^{1/2} \bar{Y}_n} \le \frac{-\mathbb{E}\bar{Y}_n}{\operatorname{Var}^{1/2} \bar{Y}_n}\right) - \Phi(A_{\boldsymbol{x}} t + B_{\boldsymbol{x}}(\boldsymbol{H})),
\end{aligned}$$

applying the Berry-Esseen theorem (Ferguson, 1996, Page 31) to the first term above yields

$$\begin{aligned}
&\left| P\left(\frac{\bar{Y}_n - \mathbb{E}\bar{Y}_n}{\operatorname{Var}^{1/2} \bar{Y}_n} \le \frac{-\mathbb{E}\bar{Y}_n}{\operatorname{Var}^{1/2} \bar{Y}_n}\right) - \Phi\left(\frac{-\mathbb{E}\bar{Y}_n}{\operatorname{Var}^{1/2} \bar{Y}_n}\right) \right| \\
&\qquad \le \frac{C\mathbb{E}|Y_{ni}|^3}{\operatorname{Var}^{3/2}(Y_{ni})\sqrt{n}} = O\left(\frac{1}{\sqrt{n|\boldsymbol{H}|^{1/2}}}\right),
\end{aligned}$$

and

$$\begin{aligned}
&P\left(\frac{\bar{Y}_n - \mathbb{E}\bar{Y}_n}{\operatorname{Var}^{1/2} \bar{Y}_n} \le \frac{-\mathbb{E}\bar{Y}_n}{\operatorname{Var}^{1/2} \bar{Y}_n}\right) - \Phi(A_{\boldsymbol{x}} t + B_{\boldsymbol{x}}(\boldsymbol{H})) \\
&\quad \le \Phi\left(\frac{-\mathbb{E}\bar{Y}_n}{\operatorname{Var}^{1/2} \bar{Y}_n}\right) - \Phi(A_{\boldsymbol{x}} t + B_{\boldsymbol{x}}(\boldsymbol{H})) + O\left(\frac{1}{\sqrt{n|\boldsymbol{H}|^{1/2}}}\right),
\end{aligned}$$

uniformly for $\in [t_{\boldsymbol{x}}^* - t_n, t_{\boldsymbol{x}}^* + t_n]$ and $\boldsymbol{x} \in \beta(c)$. A similar argument shows a lower bound of the same order. Next, with a similar argument as we had in the last

step of proof for Theorem 2.2, we can show the integrated error

$$\frac{1}{\sqrt{n|\boldsymbol{H}|^{1/2}}}\int_{\beta(c)}\int_{t_{\boldsymbol{x}}^*-t_n}^{t_{\boldsymbol{x}}^*+t_n}\left|\Phi\left(\frac{-\mathbb{E}\bar{Y}_n}{\mathrm{Var}^{1/2}\,\bar{Y}_n}\right)-\Phi(A_{\boldsymbol{x}}+B_{\boldsymbol{x}}(\boldsymbol{H}))\right|\,dtd\mathcal{H}(\boldsymbol{x})$$
$$=o\left(\frac{1}{\sqrt{n|\boldsymbol{H}|^{1/2}}}+\mathrm{tr}(\boldsymbol{H})\right).$$

So we have

$$\frac{c}{\sqrt{n|\boldsymbol{H}|^{1/2}}}\int_{\beta(c)}\int_{t_x^*-t_n}^{t_x^*+t_n}\left|P\left(\widehat{f}_{n,\boldsymbol{H}}\left(\boldsymbol{x}^t\right)<c\right)-\mathbb{1}_{\{t<0\}}\right|\,dtd\mathcal{H}(\boldsymbol{x})$$
$$=\frac{c}{\sqrt{n|\boldsymbol{H}|^{1/2}}}\int_{\beta(c)}\int_{t_x^*-t_n}^{t_x^*+t_n}\left|\Phi(A_{\boldsymbol{x}}+B_{\boldsymbol{x}}(\boldsymbol{H}))-\mathbb{1}_{\{t<0\}}\right|\,dtd\mathcal{H}(\boldsymbol{x})$$
$$+o\left(\frac{1}{\sqrt{n|\boldsymbol{H}|^{1/2}}}+\mathrm{tr}(\boldsymbol{H})\right).$$

By Lemma B.3,

$$\frac{c}{\sqrt{n|\boldsymbol{H}|^{1/2}}}\int_{\beta(c)}\int_{-\infty}^{\infty}\left|\Phi(A_{\boldsymbol{x}}+B_{\boldsymbol{x}}(\boldsymbol{H}))-\mathbb{1}_{\{t<0\}}\right|\,dtd\mathcal{H}(\boldsymbol{x})=\mathrm{LS}(\boldsymbol{H}).$$

This completes the proof.

A.3. Proof of Corollary 3.1

Let $\boldsymbol{H}=h^2\boldsymbol{I}$. If h^2 is of order $n^{-2/(d+4)}$ then by Assumption D3,

$$\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x})=f_0(\boldsymbol{x})+\frac{1}{2}\,\mathrm{tr}\{\boldsymbol{H}\nabla^2 f_0(\boldsymbol{x})\}+O\left(n^{-4/(d+4)}\right),$$
$$\mathrm{Var}\,\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x})=n^{-1}|\boldsymbol{H}|^{-1/2}R(K)f_0(\boldsymbol{x})+O\left(n^{-6/(d+4)}\right),$$

uniformly in $\boldsymbol{x}$. From the proof of Theorem 2.1, if we specifically pick $t_n=\sqrt{\log n}$, and $\delta_n=\sqrt{\log n/(n|\boldsymbol{H}|^{1/2})}$, we can further quantify the error in equations (28) and (29) as

$$\mathbb{E}(\bar{Y}_n)=\left\{\frac{t}{\sqrt{n|\boldsymbol{H}|^{1/2}}}\|\nabla f_0(\boldsymbol{x})\|+D_1(\boldsymbol{x},\boldsymbol{H})\right\}\{1+O(n^{-2/(d+4)}\sqrt{\log n})\},$$
$$n|\boldsymbol{H}|^{1/2}\,\mathrm{Var}\,\bar{Y}_n=R(K)c+O(n^{-2/(d+4)}\sqrt{\log n}),$$

and further

$$\mathbb{E}[\mu_{f_0}\{\mathcal{L}(c)\Delta\widehat{\mathcal{L}}_{\boldsymbol{H}}(c)\}]=\mathrm{LS}(h)+O\left(n^{-4/(d+4)}(\log n)^{3/2}\right).$$

With a similar argument as in Corollary 2.2, we can see that $h_{\mathrm{opt}}/h_0=1+O(n^{-2/(d+4)}(\log n)^{3/2})$.

Now we study $\hat{h}_{\text{opt}}/h_{\text{opt}}$. Let $g_{n,\boldsymbol{H}_0} = \widehat{f}_{n,\boldsymbol{H}_0} - f_0$, $g_{n,\boldsymbol{H}_1} = \widehat{f}_{n,\boldsymbol{H}_1} - f_0$, $g_{n,\boldsymbol{H}_2} = \widehat{f}_{n,\boldsymbol{H}_0} - f_0$. Let

$$m(\boldsymbol{x}) := \frac{\phi(sF_{\boldsymbol{x}}) + 2\Phi(sF_{\boldsymbol{x}})sF_{\boldsymbol{x}} - sF_{\boldsymbol{x}}}{-A_{\boldsymbol{x}}},$$

and with slight abuse of notation, we let $\widehat{m}(\boldsymbol{x})$ be defined similarly where we substitute f_0 with $\widehat{f}_{n,\boldsymbol{H}_0}$, ∇f_0 with $\nabla\widehat{f}_{n,\boldsymbol{H}_1}$ and $\nabla^2 f_0$ with $\nabla^2\widehat{f}_{n,\boldsymbol{H}_0}$. We look at the difference

$$\begin{aligned}
&\left|\int_{\beta(c)} m(\boldsymbol{x})\,d\mathcal{H}(\boldsymbol{x}) - \int_{\widehat{\beta}_{n,\boldsymbol{H}_1}(c)} \widehat{m}(\boldsymbol{x})\,d\mathcal{H}(\boldsymbol{x})\right| \\
&\quad\le \left|\int_{\beta(c)} m(\boldsymbol{x})\,d\mathcal{H}(\boldsymbol{x}) - \int_{\beta(c)} \widehat{m}(\boldsymbol{x})\,d\mathcal{H}(\boldsymbol{x})\right| \\
&\quad + \left|\int_{\beta(c)} \widehat{m}(\boldsymbol{x})\,d\mathcal{H}(\boldsymbol{x}) - \int_{\widehat{\beta}_{n,\boldsymbol{H}_1}(c)} \widehat{m}(\boldsymbol{x})\,d\mathcal{H}(\boldsymbol{x})\right|,
\end{aligned} \tag{30}$$

where $\widehat{\beta}_{n,\boldsymbol{H}_1}(c) := \widehat{f}^{-1}_{n,\boldsymbol{H}_1}(c)$. By Lemma B.5, when we have $\boldsymbol{H}_1 \to 0$ and $n^{-1}|\boldsymbol{H}_1|^{-1/2}(\boldsymbol{H}_1^{-1})^{\otimes 2} = O(1)$ as $n \to \infty$, the term on the last line above is $O_p(\sup_{\boldsymbol{x}\in\beta(c)} E[|g_{n,\boldsymbol{H}_1}(\boldsymbol{x})| + \|\nabla^2 g_{n,\boldsymbol{H}_1}(\boldsymbol{x})\||g_{n,\boldsymbol{H}_1}(\boldsymbol{x})| + \|\nabla g_{n,\boldsymbol{H}1}(\boldsymbol{x})\|])$. For term on the second line above, by Jensen's inequality we know $(\int_{\beta(c)} m(\boldsymbol{x})\,d\mathcal{H}(\boldsymbol{x}) - \int_{\beta(c)} \widehat{m}(\boldsymbol{x})\,d\mathcal{H}(\boldsymbol{x}))^2 \le \int_{\beta(c)}(m(\boldsymbol{x}) - \widehat{m}(\boldsymbol{x}))^2\,d\mathcal{H}(\boldsymbol{x})$. So for any (large) $M > 0$ we have

$$\begin{aligned}
&P\left(\left|\int_{\beta(c)} m(\boldsymbol{x})\,d\mathcal{H}(\boldsymbol{x}) - \int_{\beta(c)} \widehat{m}(\boldsymbol{x})\,d\mathcal{H}(\boldsymbol{x})\right| > M\right) \\
&\quad = P\left(\left|\int_{\beta(c)} m(\boldsymbol{x})\,d\mathcal{H}(\boldsymbol{x}) - \int_{\beta(c)} \widehat{m}(\boldsymbol{x})\,d\mathcal{H}(\boldsymbol{x})\right|^2 > M^2\right)
\end{aligned}$$

which is bounded above by

$$P\left(\int_{\beta(c)} \{m(\boldsymbol{x}) - \widehat{m}(\boldsymbol{x})\}^2\,d\mathcal{H}(\boldsymbol{x}) > M^2\right) \le \frac{\mathbb{E}\int_{\beta(c)} \{m(\boldsymbol{x}) - \widehat{m}(\boldsymbol{x})\}^2\,d\mathcal{H}(\boldsymbol{x})}{M^2},$$

by Markov's inequality. By Tonelli's Theorem, we can change the order of integrals so that $\mathbb{E}\int_{\beta(c)} \{m(\boldsymbol{x}) - \widehat{m}(\boldsymbol{x})\}^2\,d\mathcal{H}(\boldsymbol{x}) = \int_{\beta(c)} \mathbb{E}\{m(\boldsymbol{x}) - \widehat{m}(\boldsymbol{x})\}^2\,d\mathcal{H}(\boldsymbol{x})$.

Since we assume the true density function has 4 continuous bounded derivatives, by Theorem 4 in Chacón, Duong and Wand (2011) with slight modification, it can be easily seen $|\widehat{f}_{n,\boldsymbol{H}_1}(\boldsymbol{x}) - f_0(\boldsymbol{x})| = O_p(n^{-2/(d+6)})$, $\|\nabla\widehat{f}_{n,\boldsymbol{H}_1}(\boldsymbol{x}) - \nabla f_0(\boldsymbol{x})\| = O_p(n^{-2/(d+6)})$, $\|\nabla^2\widehat{f}_{n,\boldsymbol{H}_2}(\boldsymbol{x}) - \nabla^2 f_0(\boldsymbol{x})\|$ is $O_p(n^{-2/(d+8)})$. Thus $\widehat{F}_{\boldsymbol{x}} = F_{\boldsymbol{x}} + O_p(n^{-2/(d+8)})$, and $\widehat{A}_{\boldsymbol{x}} = A_{\boldsymbol{x}} + O_p(n^{-2/(d+8)})$. And we can also see

$$\sup_{\boldsymbol{x}\in\beta(c)} E[|g_{n,\boldsymbol{H}_1}(\boldsymbol{x})| + \|\nabla^2 g_{n,\boldsymbol{H}_1}(\boldsymbol{x})\||g_{n,\boldsymbol{H}_1}(\boldsymbol{x})| + \|\nabla g_{n,\boldsymbol{H}1}(\boldsymbol{x})\|] = O(n^{-2/(d+6)}),$$

Thus, we can check that $\int_{\beta(c)} \mathbb{E}\{m(\boldsymbol{x}) - \widehat{m}(\boldsymbol{x})\}^2 \, d\mathcal{H}(\boldsymbol{x})$ is $O(n^{-4/(d+8)})$, and the first term on the last line of (30) is $O_p(n^{-2/(d+8)})$. We can conclude that for any $0 < s_1 < s_2 < \infty$, we have $\widehat{\mathrm{AR}}_{\mathrm{LS}}(s) = \mathrm{AR}_{\mathrm{LS}}(s)\{1 + O_p(n^{-2/(d+8)})\}$ uniformly for $s \in [s_1, s_2]$. Then we have $\widehat{\mathrm{AR}}'_{\mathrm{LS}}(\hat{s}_{\mathrm{opt}}) = \mathrm{AR}'_{\mathrm{LS}}(\hat{s}_{\mathrm{opt}})\{1 + O_p(n^{-2/(d+8)})\} = \mathrm{AR}''_{\mathrm{LS}}(\tilde{s})(\hat{s}_{\mathrm{opt}} - s_{\mathrm{opt}})\{1 + O_p(n^{-2/(d+8)})\}$, where $\mathrm{AR}''_{\mathrm{LS}}(\tilde{s}) > 0$ and is bounded from 0 as $n \to \infty$. This gives us $\hat{s}_{\mathrm{opt}}/s_{\mathrm{opt}} = 1 + O_p(n^{-2/(d+8)})$, and recall that $\widehat{h}_{\mathrm{opt}} = \hat{s}_{\mathrm{opt}}^{2/(d+4)} n^{-1/(d+4)}$, $h_{\mathrm{opt}} = s_{\mathrm{opt}}^{2/(d+4)} n^{-1/(d+4)}$, we conclude

$$\frac{\hat{h}_{\mathrm{opt}}}{h_{\mathrm{opt}}} = 1 + O_p\left(n^{-2/(d+8)}\right).$$

Combining this result with $h_{\mathrm{opt}}/h_0 = 1 + O(n^{-2/(d+4)}(\log n^{3/2}))$, we can see $\hat{h}_{\mathrm{opt}}/h_0 = O_p(n^{-2/(d+8)})$.

A.4. Proof of Corollary 3.2

Similar to the proof of Corollary 3.1, let $g_{n,\boldsymbol{H}_0} = \widehat{f}_{n,\boldsymbol{H}_0} - f_0$, $g_{n,\boldsymbol{H}_1} = \widehat{f}_{n,\boldsymbol{H}_1} - f_0$, $g_{n,\boldsymbol{H}_2} = \widehat{f}_{n,\boldsymbol{H}_0} - f_0$, and let $\epsilon = \widehat{f}_{\tau,n,\boldsymbol{H}_0} - f_{\tau,0}$. Since we assume the true density function has 4 continuous bounded derivatives, again by Theorem 4 in Chacón, Duong and Wand (2011), it can be easily seen $|\widehat{f}_{n,\boldsymbol{H}_0}(\boldsymbol{x}) - f_0(\boldsymbol{x})| = O_p(n^{-2/(d+4)})$, $|\widehat{f}_{n,\boldsymbol{H}_1}(\boldsymbol{x}) - f_0(\boldsymbol{x})| = O_p(n^{-2/(d+6)})$, $\|\nabla\widehat{f}_{n,\boldsymbol{H}_1}(\boldsymbol{x}) - \nabla f_0(\boldsymbol{x})\| = O_p(n^{-2/(d+6)})$, $\|\nabla^2\widehat{f}_{n,\boldsymbol{H}_2}(\boldsymbol{x}) - \nabla^2 f_0(\boldsymbol{x})\| = O_p(n^{-2/(d+8)})$. And by Lemma A.1, $|\widehat{f}_{\tau,n,\boldsymbol{H}_0} - f_{\tau,0}| = O_p(n^{-2/(d+8)})$. We first look at the difference

$$\begin{aligned}
&\left|\int_{\beta_\tau} \frac{1}{\|\nabla f_0\|}\, d\mathcal{H} - \int_{\widehat{\beta}_{\tau,\boldsymbol{H}_1}} \frac{1}{\|\nabla\widehat{f}_{n,\boldsymbol{H}_1}\|}\, d\mathcal{H}\right| \\
&\quad\le \left|\int_{\beta_\tau} \frac{1}{\|\nabla f_0\|}\, d\mathcal{H} - \int_{\beta_\tau} \frac{1}{\|\nabla\widehat{f}_{n,\boldsymbol{H}_1}\|}\, d\mathcal{H}\right| \\
&\qquad + \left|\int_{\beta_\tau} \frac{1}{\|\nabla\widehat{f}_{n,\boldsymbol{H}_1}\|}\, d\mathcal{H} - \int_{\widehat{\beta}_{\tau,\boldsymbol{H}_1}} \frac{1}{\|\nabla\widehat{f}_{n,\boldsymbol{H}_1}\|}\, d\mathcal{H}\right|.
\end{aligned} \tag{31}$$

Since $\|\nabla f_0 - \nabla\widehat{f}_{n,\boldsymbol{H}_1}\| = O_p(n^{-2/(d+6)})$ by Chacón, Duong and Wand (2011), it is easy to see the term on the second line above is $O_p(n^{-2/(d+6)})$. Recalling $\epsilon = \widehat{f}_{\tau,n,\boldsymbol{H}_0} - f_{\tau,0}$, we can bound the last term as

$$\begin{aligned}
&\left|\int_{\beta_\tau} \frac{1}{\|\nabla\widehat{f}_{n,\boldsymbol{H}_1}\|}\, d\mathcal{H} - \int_{\widehat{\beta}_{\tau,\boldsymbol{H}_1}} \frac{1}{\|\nabla\widehat{f}_{n,\boldsymbol{H}_1}\|}\, d\mathcal{H}\right| \\
&\quad\le \left|\int_{\{f_0 = f_{\tau,0}\}} \frac{1}{\|\nabla\widehat{f}_{n,\boldsymbol{H}_1}\|}\, d\mathcal{H} - \int_{\{\widehat{f}_{n,\boldsymbol{H}_1} = f_{\tau,0}\}} \frac{1}{\|\nabla\widehat{f}_{n,\boldsymbol{H}_1}\|}\, d\mathcal{H}\right| \\
&\qquad + \left|\int_{\{f_0 = f_{\tau,0}\}} \frac{1}{\|\nabla\widehat{f}_{n,\boldsymbol{H}_1}\|}\, d\mathcal{H} - \int_{\{\widehat{f}_{n,\boldsymbol{H}_1} = f_{\tau,0}+\epsilon\}} \frac{1}{\|\nabla\widehat{f}_{n,\boldsymbol{H}_1}\|}\, d\mathcal{H}\right|,
\end{aligned}$$

By Lemma B.5, the difference on the second line above $|\int_{\{f_0=f_{\tau,0}\}} \frac{1}{\|\nabla \widehat{f}_{n,\boldsymbol{H}_1}\|}\, d\mathcal{H} - \int_{\{\widehat{f}_{n,\boldsymbol{H}_1}=f_{\tau,0}\}} \frac{1}{\|\nabla \widehat{f}_{n,\boldsymbol{H}_1}\|}\, d\mathcal{H}|$ can been seen of order $O_p(\sup_{\boldsymbol{x}\in\beta(c)} E[|g_{n,\boldsymbol{H}_1}(\boldsymbol{x})| + \|\nabla^2 g_{n,\boldsymbol{H}_1}(\boldsymbol{x})\||g_{n,\boldsymbol{H}_1}(\boldsymbol{x})| + \|\nabla g_{n,\boldsymbol{H}1}(\boldsymbol{x})\|])$. Next, by Taylor expansion, we have

$$\begin{aligned}
&\int_{\{\widehat{f}_{n,\boldsymbol{H}_1}=f_{\tau,0}+\epsilon\}} \frac{1}{\|\nabla \widehat{f}_{n,\boldsymbol{H}_1}\|}\, d\mathcal{H} \\
&\quad = \int_{\{\widehat{f}_{n,\boldsymbol{H}_1}=f_{\tau,0}\}} \frac{1}{\|\nabla \widehat{f}_{n,\boldsymbol{H}_1}\|}\, d\mathcal{H} - \left(\frac{d}{de} \int_{\{\widehat{f}_{n,\boldsymbol{H}_1}=f_{\tau,0}+e\}} \frac{1}{\|\nabla \widehat{f}_{n,\boldsymbol{H}_1}\|}\, d\mathcal{H}\bigg|_{e=s\epsilon} \right) \epsilon,
\end{aligned}$$

where $s \in [0,1]$. Then we have

$$\begin{aligned}
&\left| \int_{\{\widehat{f}_{n,\boldsymbol{H}_1}=f_{\tau,0}\}} \frac{1}{\|\nabla \widehat{f}_{n,\boldsymbol{H}_1}\|}\, d\mathcal{H} - \int_{\{\widehat{f}_{n,\boldsymbol{H}_1}=f_{\tau,0}+\epsilon\}} \frac{1}{\|\nabla \widehat{f}_{n,\boldsymbol{H}_1}\|}\, d\mathcal{H} \right| \\
&\qquad = \left| \left(\frac{d}{de} \int_{\{\widehat{f}_{n,\boldsymbol{H}_1}=f_{\tau,0}+e\}} \frac{1}{\|\nabla \widehat{f}_{n\boldsymbol{H}_1}\|}\, d\mathcal{H}\bigg|_{e=s\epsilon} \right) \epsilon \right|.
\end{aligned} \tag{32}$$

From the proof of Lemma B.5, we can see when n is sufficiently large, the derivative on the last line of (32) is uniformly bounded. Moreover, by Lemma A.1, $\epsilon = O(\|g_{n,\boldsymbol{H}_0}\|_\infty)$, so we have

$$\begin{aligned}
&\left| \int_{\beta_\tau} \frac{1}{\|\nabla \widehat{f}_{n,\boldsymbol{H}_1}\|}\, d\mathcal{H} - \int_{\widehat{\beta}_{\tau,\boldsymbol{H}_1}} \frac{1}{\|\nabla \widehat{f}_{n,\boldsymbol{H}_1}\|}\, d\mathcal{H} \right| \\
&\quad = O(\|g_{n,\boldsymbol{H}_0}\|_\infty) \\
&\qquad + O_p\Big(\sup_{\boldsymbol{x}\in\beta(c)} E[|g_{n,\boldsymbol{H}_1}(\boldsymbol{x})| + \|\nabla^2 g_{n,\boldsymbol{H}_1}(\boldsymbol{x})\||g_{n,\boldsymbol{H}_1}(\boldsymbol{x})| + \|\nabla g_{n,\boldsymbol{H}1}(\boldsymbol{x})\|] \Big) \\
&\quad = O_p(n^{-2/(d+6)}).
\end{aligned}$$

Then

$$\left| \int_{\beta_\tau} \frac{1}{\|\nabla f_0\|}\, d\mathcal{H} - \int_{\widehat{\beta}_{\tau,\boldsymbol{H}_1}} \frac{1}{\|\nabla \widehat{f}_{n,\boldsymbol{H}_1}\|}\, d\mathcal{H} \right| = O_p(n^{-2/(d+6)}),$$

and thus $|w_0 - \hat{w}_0| = O_p(n^{-2/(d+6)})$. Using exactly the same trick, we can show

$$\left| \int_{\beta_\tau} \frac{\mu(K)\,\mathrm{tr}(\nabla^2 f_0)}{2\|\nabla f_0\|}\, d\mathcal{H} - \int_{\widehat{\beta}_{\tau,\boldsymbol{H}_1}} \frac{\mu(K)\,\mathrm{tr}(\nabla^2 \widehat{f}_{n,\boldsymbol{H}_2})}{2\|\nabla \widehat{f}_{n,\boldsymbol{H}_1}\|}\, d\mathcal{H} \right| = O_p(n^{-2/(d+8)}).$$

Next, we provide the bound for $|\int_{\mathcal{L}_\tau} \frac{\mu(K)\,\mathrm{tr}(\nabla^2 f_0)}{2}\,d\lambda - \int_{\widehat{\mathcal{L}}_{\tau,\boldsymbol{H}_0}} \frac{\mu(K)\,\mathrm{tr}(\nabla^2 \widehat{f}_{n,\boldsymbol{H}_2})}{2}\,d\lambda|$. Similarly, we have

$$\begin{aligned}
&\left|\int_{\mathcal{L}_\tau} \frac{\mu(K)\,\mathrm{tr}(\nabla^2 f_0)}{2}\,d\lambda - \int_{\widehat{\mathcal{L}}_{\tau,\boldsymbol{H}_0}} \frac{\mu(K)\,\mathrm{tr}(\nabla^2 \widehat{f}_{n,\boldsymbol{H}_2})}{2}\,d\lambda\right| \\
&\quad\le \left|\int_{\mathcal{L}_\tau} \frac{\mu(K)\,\mathrm{tr}(\nabla^2 f_0)}{2}\,d\lambda - \int_{\widehat{\mathcal{L}}_{\tau,\boldsymbol{H}_0}} \frac{\mu(K)\,\mathrm{tr}(\nabla^2 f_0)}{2}\,d\lambda\right| \\
&\qquad + \left|\int_{\widehat{\mathcal{L}}_{\tau,\boldsymbol{H}_0}} \frac{\mu(K)\,\mathrm{tr}(\nabla^2 f_0)}{2}\,d\lambda - \int_{\widehat{\mathcal{L}}_{\tau,\boldsymbol{H}_0}} \frac{\mu(K)\,\mathrm{tr}(\nabla^2 \widehat{f}_{n,\boldsymbol{H}_2})}{2}\,d\lambda\right|,
\end{aligned} \tag{33}$$

and since we assume f_0 has bounded second derivatives, the difference on the second line above $|\int_{\mathcal{L}_\tau} \frac{\mu(K)\,\mathrm{tr}(\nabla^2 f_0)}{2}\,d\lambda - \int_{\widehat{\mathcal{L}}_{\tau,\boldsymbol{H}_0}} \frac{\mu(K)\,\mathrm{tr}(\nabla^2 f_0)}{2}\,d\lambda| = O\{|\lambda(\mathcal{L}_\tau) - \lambda(\hat{\mathcal{L}}_{\tau,\boldsymbol{H}_0})|\}$. Now we show $|\lambda(\mathcal{L}_\tau) - \lambda(\hat{\mathcal{L}}_{\tau,\boldsymbol{H}})| = O(\|g_{n,\boldsymbol{H}_0}\|_\infty)$. It can be seen that

$$\begin{aligned}
&\{\boldsymbol{x} : f_0(\boldsymbol{x}) \ge f_{\tau,0} + \epsilon + \|g_{n,\boldsymbol{H}_0}\|_\infty\} \subset \widehat{\mathcal{L}}_{\tau,\boldsymbol{H}_0} \qquad \text{and} \\
&\widehat{\mathcal{L}}_{\tau,\boldsymbol{H}_0} \subset \{\boldsymbol{x} : f_0(\boldsymbol{x}) \ge f_{\tau,0} + \epsilon - \|g_{n,\boldsymbol{H}_0}\|_\infty\},
\end{aligned}$$

and then

$$\begin{aligned}
&|\lambda(\mathcal{L}_\tau) - \lambda(\widehat{\mathcal{L}}_{\tau,\boldsymbol{H}_0})| \\
&\quad\le |\lambda(\mathcal{L}_\tau) - \lambda\{\boldsymbol{x} : f_0(\boldsymbol{x}) \ge f_{\tau,0} + \epsilon + \|g_{n,\boldsymbol{H}_0}\|_\infty\}| \\
&\qquad + |\lambda(\mathcal{L}_\tau) - \lambda\{\boldsymbol{x} : f_0(\boldsymbol{x}) \ge f_{\tau,0} + \epsilon - \|g_{n,\boldsymbol{H}_0}\|_\infty\}|.
\end{aligned}$$

Further by Proposition A.1 of Cadre (2006),

$$\begin{aligned}
&|\lambda(\mathcal{L}_\tau) - \lambda\{\boldsymbol{x} : f_0(\boldsymbol{x}) \ge f_{\tau,0} + \epsilon + \|g_{n,\boldsymbol{H}_0}\|_\infty\}| \\
&\quad= \left|(\epsilon + \|g_{n,\boldsymbol{H}_2}\|_\infty) \int_{\beta_\tau} \frac{1}{\|\nabla f_0\|}\,d\mathcal{H}\right| + o(\epsilon + \|g_{n,\boldsymbol{H}_0}\|_\infty),
\end{aligned}$$

and

$$\begin{aligned}
&|\lambda(\mathcal{L}_\tau) - \lambda\{\boldsymbol{x} : f_0(\boldsymbol{x}) \ge f_{\tau,0} + \epsilon - \|g_{n,\boldsymbol{H}_0}\|_\infty\}| \\
&\quad= \left|(\epsilon - \|g_{n,\boldsymbol{H}_0}\|_\infty) \int_{\beta_\tau} \frac{1}{\|\nabla f_0\|}\,d\mathcal{H}\right| + o(\epsilon - \|g_{n,\boldsymbol{H}_0}\|_\infty),
\end{aligned}$$

and thus

$$|\lambda(\mathcal{L}_\tau) - \lambda(\widehat{\mathcal{L}}_{\tau,\boldsymbol{H}_0})| = O(\|g_{n,\boldsymbol{H}_0}\|_\infty),$$

when $\|g_{n,\boldsymbol{H}_0}\|_\infty \to 0$.

Now for the last term of (33), by Jensen's inequality we have

$$\left(\int_{\widehat{\mathcal{L}}_{\tau,\boldsymbol{H}_0}} \frac{\mu(K)\,\mathrm{tr}(\nabla^2 g_{n,\boldsymbol{H}_2})}{2}\,d\lambda\right)^2 \le \int_{\widehat{\mathcal{L}}_{\tau,\boldsymbol{H}_0}} \left(\frac{\mu(K)\,\mathrm{tr}(\nabla^2 g_{n,\boldsymbol{H}_2})}{2}\right)^2 d\lambda$$

$$\leq \int \left(\frac{\mu(K)\,\mathrm{tr}(\nabla^2 g_{n,\boldsymbol{H}_2})}{2} \right)^2 d\lambda,$$

and then for any (large) $M > 0$,

$$P\left(\left| \int_{\widehat{\mathcal{L}}_{\tau,\boldsymbol{H}_0}} \frac{\mu(K)\,\mathrm{tr}(\nabla^2 g_{n,\boldsymbol{H}_2})}{2}\, d\lambda \right| > M \right) \leq \frac{\mathbb{E}\int \left(\frac{\mu(K)\,\mathrm{tr}(\nabla^2 g_{n,\boldsymbol{H}_2})}{2} \right)^2 d\lambda}{M^2},$$

where we applied Markov's inequality to obtain the upper bound. Applying Tonelli's Theorem yields

$$\begin{aligned} &\mathbb{E}\int \left(\frac{\mu(K)\,\mathrm{tr}(\nabla^2 g_{n,\boldsymbol{H}_2})}{2} \right)^2 d\lambda \\ &\quad = \int \mathbb{E}\left(\frac{\mu(K)\,\mathrm{tr}(\nabla^2 g_{n,\boldsymbol{H}_2})}{2} \right)^2 d\lambda \\ &\quad = O_p(n^{-4/(d+8)}). \end{aligned}$$

So $|\int_{\mathcal{L}_\tau} \frac{\mu(K)\,\mathrm{tr}(\nabla^2 f_0)}{2}\, d\lambda - \int_{\widehat{\mathcal{L}}_{\tau,\boldsymbol{H}_0}} \frac{\mu(K)\,\mathrm{tr}(\nabla^2 \widehat{f}_{n,\boldsymbol{H}_2})}{2}\, d\lambda| = O_p(n^{-2/(d+8)})$.

Now from Chacón, Duong and Wand (2011), we know that $\widehat{G}_{\boldsymbol{x}} = G_{\boldsymbol{x}} + O_p(n^{-2/(d+8)})$, $\widehat{A}_{\boldsymbol{x}} = A_{\boldsymbol{x}} + O_p(n^{-2/(d+8)})$. And using a similar trick as for w_0, we have

$$|\widehat{\mathrm{AR}}_{\mathrm{HDR}}(s) - \mathrm{AR}_{\mathrm{HDR}}(s)| = O_p(n^{-2/(d+6)}).$$

And we can conclude that for any $0 < s_1 < s_2 < \infty$, we have $\widehat{\mathrm{AR}}(s) = \mathrm{AR}(s)\{1 + O_p(n^{-2/(d+8)})\}$ uniformly for $s \in [s_1, s_2]$. And $\widehat{\mathrm{AR}}'_{\mathrm{HDR}}(\hat{s}_{\mathrm{opt}}) = \mathrm{AR}'_{\mathrm{HDR}}(\hat{s}_{\mathrm{opt}})\{1 + O_p(n^{-2/(d+8)})\} = \mathrm{AR}''_{\mathrm{HDR}}(\tilde{s})(\hat{s}_{\mathrm{opt}} - s_{\mathrm{opt}})\{1 + O_p(n^{-2/(d+8)})\}$, where $\mathrm{AR}''_{\mathrm{HDR}}(\tilde{s}) > 0$ and is bounded from 0 as $n \to \infty$. This gives us $\hat{s}_{\mathrm{opt}}/s_{\mathrm{opt}} = 1 + O_p(n^{-2/(d+8)})$. Finally, recall that $\widehat{h}_{\mathrm{opt}} = \hat{s}_{\mathrm{opt}}^{2/(d+4)} n^{-1/(d+4)}$ and $h_{\mathrm{opt}} = s_{\mathrm{opt}}^{2/(d+4)} n^{-1/(d+4)}$, we conclude

$$\frac{\hat{h}_{\mathrm{opt}}}{h_{\mathrm{opt}}} = 1 + O_p\left(n^{-2/(d+8)}\right).$$

Appendix B: Additional theorems and proofs

The following theorem is a slight extension of Theorem 2.3 of Giné and Guillou (2002) to allow general bandwidth matrices and to apply to gradient estimation. Its proof is essentially the same as that of their Theorem 2.3, so is omitted.

Theorem B.1. *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be i.i.d. from a bounded density on $\mathbb{R}^d$, and let Assumptions K, and H hold. We have*

$$\limsup_{n\to\infty} \sqrt{\frac{n|\boldsymbol{H}_n|^{1/2}}{\log|\boldsymbol{H}_n|^{-1/2}}} \|\widehat{f}_{n,\boldsymbol{H}_n} - \mathbb{E}\widehat{f}_{n,\boldsymbol{H}_n}\|_\infty = C_{0,1} \qquad a.s., \tag{34}$$

and

$$\limsup_{n\to\infty} \sqrt{\frac{n|\boldsymbol{H}_n|^{1/2}\lambda_{\min}(\boldsymbol{H})}{\log|\boldsymbol{H}_n|^{-1/2}}} \|\nabla \widehat{f}_{n,\boldsymbol{H}_n} - \mathbb{E}\nabla \widehat{f}_{n,\boldsymbol{H}_n}\|_\infty \le C_{0,2} \qquad a.s., \tag{35}$$

Here $C_{0,1}$ and $C_{0,2}$ depend on d, K, and $\|f_0\|_\infty$.

The proof of Theorem B.1 also yields the following probability bound which we need in particular.

Corollary B.1. *Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be i.i.d. from a bounded density on $\mathbb{R}^d$, and let Assumptions K, and H hold. Then for some constant $C > 0$ and for $0 < \epsilon \le C\|K\|_2^2\|f_0\|_\infty/\|K\|_\infty$, we have*

$$P\left\{\left\|\widehat{f}_{n,\boldsymbol{H}} - \mathbb{E}\widehat{f}_{n,\boldsymbol{H}}\right\|_\infty > \epsilon\right\} \le L\exp\left\{-C_{0,1}\epsilon^2 n|\boldsymbol{H}_n|^{1/2}\right\}, \tag{36}$$

where $C_{0,1}$ depends on K, d, and $\|f_0\|_\infty$. Similarly, for $0 < \epsilon$ small enough (with bound depending on ∇K and $\|f_0\|_\infty$),

$$P\left\{\left\|\nabla\widehat{f}_{n,\boldsymbol{H}} - \mathbb{E}\nabla\widehat{f}_{n,\boldsymbol{H}}\right\|_\infty > \epsilon\right\} \le L\exp\left\{-C_{0,2}\epsilon^2 n|\boldsymbol{H}_n|^{1/2}\lambda_{\boldsymbol{H}}\right\}, \tag{37}$$

where $C_{0,2} > 0$ depends on ∇K, d, and $\|f_0\|_\infty$, and where $\lambda_{\boldsymbol{H}}$ is the smallest eigenvalue of $\boldsymbol{H}$.

Proof. We let

$$\mathcal{F}_{K,\boldsymbol{H}_n} := \left\{K(\boldsymbol{H}_n^{-1/2}(\boldsymbol{t} - \cdot)) : \boldsymbol{t} \in \mathbb{R}^d\right\},$$

(which is a VC class by Assumption K). We have that for $\epsilon > 0$

$$P\left\{\left\|\widehat{f}_{n,\boldsymbol{H}} - \mathbb{E}\widehat{f}_{n,\boldsymbol{H}}\right\|_\infty > \epsilon\right\} = P\left\{\frac{1}{n|\boldsymbol{H}_n|^{1/2}}\left\|\sum_{i=1}^n f(\boldsymbol{X}_i) - \mathbb{E}f(\boldsymbol{X}_i)\right\|_{\mathcal{F}_{K,\boldsymbol{H}}} > \epsilon\right\}. \tag{38}$$

Thus we set

$$\sigma_n^2 := |\boldsymbol{H}_n|^{1/2}\|K\|_2^2\|f_0\|_\infty \quad \text{and} \quad U := \|K\|_\infty$$

which satisfy the conditions of Corollary 2.2 of Giné and Guillou (2002) so we have L and C (depending on K and d) from the corollary, so we set $t = \epsilon n|\boldsymbol{H}_n|^{1/2}$, and $\lambda = C$ so that (7) in Giné and Guillou (2002) is satisfied (using that $n|\boldsymbol{H}_n|^{1/2} \to \infty$ for the lower bound). We conclude that (38) is bounded above by

$$L\exp\left\{-\frac{D\epsilon^2 n|\boldsymbol{H}_n|^{1/2}}{\|K\|_2^2\|f_0\|_\infty}\right\}$$

where $D := (\log(1 + C/4L))/LC$, completing the proof.

A similar proof shows that (37) holds. Let $K_{\boldsymbol{H}} := |\boldsymbol{H}|^{-1/2} K(\boldsymbol{H}^{-1/2}\cdot)$. Then $\nabla K_{\boldsymbol{H}}(\boldsymbol{y}) = |\boldsymbol{H}|^{-1/2}\boldsymbol{H}^{-1/2}\nabla K(\boldsymbol{H}^{-1/2}\boldsymbol{y})$, so $P\left\{\left\|\nabla \widehat{f}_{n,\boldsymbol{H}} - \mathbb{E}\nabla \widehat{f}_{n,\boldsymbol{H}}\right\|_\infty > \epsilon\right\}$ is bounded above by

$$dP\left\{\frac{1}{n|\boldsymbol{H}_n|^{1/2}}\|a_{\boldsymbol{H}}\|\left\|\sum_{i=1}^n \|\nabla K(\boldsymbol{H}^{-1/2}(\cdot - \boldsymbol{X}_i)) - \mathbb{E}\nabla K(\boldsymbol{H}^{-1/2}(\cdot - \boldsymbol{X}_i))\|\right\|_\infty > \epsilon\right\}$$

by the Cauchy-Schwarz inequality where $a'_{\boldsymbol{H}}$ is a row of $\boldsymbol{H}^{-1/2}$ (and, recall, $\|\cdot\|$ is just Euclidean norm). Since $\|a_{\boldsymbol{H}}\| \le \lambda_{\boldsymbol{H}}^{-1/2}$ where $\lambda_{\boldsymbol{H}}^{-1/2}$ is the largest eigenvalue of $\boldsymbol{H}^{-1/2}$, the previous display is bounded above by

$$P\left\{\frac{1}{n|\boldsymbol{H}_n|^{1/2}\lambda_{\boldsymbol{H}}^{1/2}}\left\|\sum_{i=1}^n f(\boldsymbol{X}_i) - \mathbb{E}f(\boldsymbol{X}_i)\right\|_{\mathcal{F}_{K,\boldsymbol{H}}} > \epsilon\right\}$$

where

$$\mathcal{F}_{K,\boldsymbol{H}_n} := \left\{\|\nabla K(\boldsymbol{H}_n^{-1/2}(\boldsymbol{t} - \cdot))\| : \boldsymbol{t} \in \mathbb{R}^d\right\},$$

is a VC class by Assumption K. We thus take $\sigma_n^2 := |\boldsymbol{H}_n|^{1/2} R(\nabla K)\|f_0\|_\infty$ and $U := \sqrt{d}\|\nabla K\|_\infty$ and apply Corollary 2.2 of Giné and Guillou (2002). Here $R(\nabla K)$ is the largest eigenvalue of $\int (\nabla K)(\nabla K)' d\lambda$. We take $t = \epsilon n|\boldsymbol{H}|^{1/2}\lambda_{\boldsymbol{H}}^{1/2}$ and $\lambda = C$. Then (7) of Giné and Guillou (2002) is satisfied since we have $n^{1/2}|\boldsymbol{H}|^{1/4}\lambda_{\boldsymbol{H}}^{1/2}/\sqrt{\log|\boldsymbol{H}|^{-1/2}} \to \infty$. This yields (37). □

The following is referred to as the ϵ-Neighborhood Theorem by Guillemin and Pollack (1974). It states that for certain manifolds, so-called Tubular Neighborhoods exist.

Theorem B.2 (page 69, Guillemin and Pollack (1974))**.** *For a compact boundaryless manifold Y in $\mathbb{R}^d$ and $\epsilon > 0$, let Y^ϵ be the open set of points in $\mathbb{R}^d$ with distance less than ϵ from Y. If ϵ is small enough, then each point $w \in Y^\epsilon$ possesses a unique closest point in Y, denoted $\pi(w)$. Moreover, the map $\pi : Y^\epsilon \to Y$ is a submersion.*

A map between manifolds is a submersion if, at all points, the Jacobian map between corresponding tangent spaces is of full rank; see page 20 of Guillemin and Pollack (1974).

Theorem B.3 (Taylor's Theorem in Several Variables)**.** *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is of class C^{k+1} on an open convex set S. If $\boldsymbol{a} \in S$ and $\boldsymbol{a} + \boldsymbol{h} \in S$, then*

$$f(\boldsymbol{a} + \boldsymbol{h}) = \sum_{|\alpha| \le k} \frac{\partial^\alpha f(\boldsymbol{a})}{\alpha!}\boldsymbol{h}^\alpha + R_{\boldsymbol{a},k}(\boldsymbol{h}), \tag{39}$$

where the remainder is given in Lagrange's form by

$$R_{\boldsymbol{a},k}(\boldsymbol{h}) = \sum_{|\alpha|=k+1} \partial^\alpha f(\boldsymbol{a}+c\boldsymbol{h})\frac{\boldsymbol{h}^\alpha}{\alpha!} \tag{40}$$

for some $c \in (0,1)$.

Lemma B.1. *Let* $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)'$ *be a* d*-dimensional vector and* $\boldsymbol{A} = \{a_{ij}\}$ *be a* $d \times d$ *matrix. Then* $|\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}| \le d\|\boldsymbol{A}\|_\infty \|\boldsymbol{x}\|^2$, *where* $\|\boldsymbol{A}\|_\infty = \max_{i,j}|a_{ij}|$.

Proof. We have

$$|\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}| \le \sum_{i,j}|a_{ij}x_ix_j| \le \sum_{i,j}|a_{ij}|\frac{x_i^2+x_j^2}{2} \le \|\boldsymbol{A}\|_\infty \sum_{i,j}\frac{x_i^2+x_j^2}{2} = d\|\boldsymbol{A}\|_\infty \|x\|^2.$$

□

Lemma B.2. *Let Assumption D1b and D2 hold, the for* $\delta_n > 0$ *small enough, there exists constant* $c_2 > 0$ *and another sequence* $\varepsilon_n > 0$ *such that* $\varepsilon_n = c_2\delta_n$ *and* $|f_0(\boldsymbol{x}) - f_{\tau,0}| \ge \varepsilon_n$ *when* $\boldsymbol{x} \in (\mathcal{L}_{\delta_n}(f_{\tau,0})^c \backslash \mathcal{L}_\delta(f_{\tau,0})^c) \cup (\mathcal{L}_{-\delta_n}(f_{\tau,0}) \backslash \mathcal{L}_{-\delta}(f_{\tau,0}))$.

Proof. The existence of such c_2 can be proved by Theorem B.2, which says for all $\delta > 0$ sufficiently small, then for each $\boldsymbol{x} \in \bigcup_{\boldsymbol{y}\in\beta} B(\boldsymbol{y},\delta)$ there exist a unique $\boldsymbol{\theta} \in I_d$ and $|s| \le \delta$ such that $\boldsymbol{x} = \boldsymbol{y}(\boldsymbol{\theta}) + s\boldsymbol{u}(\boldsymbol{\theta})$, where

$$\boldsymbol{u}(\boldsymbol{\theta}) = -\frac{\nabla f_0(\boldsymbol{y})}{\|\nabla f_0(\boldsymbol{y})\|},$$

is outer unit normal vector of β_τ at $\boldsymbol{y} \equiv \boldsymbol{y}(\boldsymbol{\theta})$. And here we pick $\delta > 0$ sufficiently small such that not only the Tubular Neighborhood Theorem (Theorem B.2) but also the following hold: When $\|\boldsymbol{y}_1 - \boldsymbol{y}_2\| \le \delta$, $|\frac{\partial f_0(\boldsymbol{y}_1)}{x_i} - \frac{\partial f_0(\boldsymbol{y}_2)}{x_i}| \le \gamma, i = 1, 2, \ldots, d$, for some $\gamma > 0$.

Note these two conditions are both feasible because under Assumption D1b, f_0 has two continuous bounded derivatives, which indicates both f_0 and ∇f_0 are Lipschitz. Then for $\boldsymbol{x} \in (\mathcal{L}_{\delta_n}(f_{\tau,0})^c \backslash \mathcal{L}_\delta(f_{\tau,0})^c) \cup (\mathcal{L}_{-\delta_n}(f_{\tau,0}) \backslash \mathcal{L}_{-\delta}(f_{\tau,0}))$,

$$|f_0(\boldsymbol{x}) - f_{\tau,0}| = |f_0(\boldsymbol{y} + s\boldsymbol{u}) - f_0(\boldsymbol{y})| = |\nabla f_0(\boldsymbol{\xi})'\boldsymbol{u}s|,$$

where $\boldsymbol{\xi} = \boldsymbol{y} + ls\boldsymbol{u}$ for some $0 \le l \le 1$, $\boldsymbol{y} \in \beta_\tau$. So

$$|f_0(\boldsymbol{x}) - f_{\tau,0}| = \left|\nabla f_0(\boldsymbol{\xi})'\frac{\nabla f_0(\boldsymbol{y})}{\|\nabla f_0(\boldsymbol{y})\|}s\right|.$$

Note that

$$|\nabla f_0(\boldsymbol{\xi})'\nabla f_0(\boldsymbol{y})| = |\|\nabla f_0(\boldsymbol{y})\| + (\nabla f_0(\boldsymbol{\xi}) - \nabla f_0(\boldsymbol{y}))'\nabla f_0(\boldsymbol{y})|$$

Let $b := \inf_{\boldsymbol{y}\in\beta_\tau}\|\nabla f_0(\boldsymbol{y})\|$, so by Assumption D1b, $b > 0$. Then by Cauchy-Schwarz inequality

$$|(\nabla f_0(\boldsymbol{\xi}) - \nabla f_0(\boldsymbol{y}))'\nabla f_0(\boldsymbol{y})| \le \|\nabla f_0(\boldsymbol{\xi}) - \nabla f_0(\boldsymbol{y})\|\|f_0(\boldsymbol{y})\| \le \sqrt{d}\gamma b$$

We can choose $\gamma > 0$ sufficiently small such that $|\nabla f_0(\boldsymbol{\xi})'\nabla f_0(\boldsymbol{y})| \geq \frac{1}{2}b$. Then since $\|\nabla f_0(\boldsymbol{y})\|$ is bounded, $|f_0(\boldsymbol{x}) - f_{\tau,0}| \geq \frac{1}{2\sup_{\boldsymbol{y}\in\beta}\|\nabla f_0(\boldsymbol{y})\|}|s|$. Now for $\boldsymbol{x} \in (\mathcal{L}_{\delta_n}(f_{\tau,0})^c \backslash \mathcal{L}_{\delta}(f_{\tau,0})^c) \cup (\mathcal{L}_{-\delta_n}(f_{\tau,0}) \backslash \mathcal{L}_{-\delta}(f_{\tau,0}))$, $|s| \geq \delta_n$, so $|f_0(\boldsymbol{x}) - f_{\tau,0}| \geq \varepsilon_n = \frac{1}{2\sup_{\boldsymbol{y}\in\beta}\|\nabla f_0(\boldsymbol{y})\|}\delta_n$. □

Lemma B.3. *Let $a < 0$ and $b \in \mathbb{R}$ be two constants, then*

$$\int_{\mathbb{R}} |\Phi(ax+b) - \mathbb{1}_{\{x<0\}}|\, dx = \frac{2\phi(b) + 2\Phi(b)b - b}{-a}.$$

Proof. Note

$$\int_{\mathbb{R}} |\Phi(ax+b) - \mathbb{1}_{\{x<0\}}|\, dx = \int_{-\infty}^{0} (1 - \Phi(ax+b))\, dx + \int_{0}^{\infty} \Phi(ax+b)\, dx.$$

And

$$\int_{-\infty}^{0} (1 - \Phi(ax+b)\, dx = x(1 - \Phi(ax+b))|_{-\infty}^{0} + \int_{-\infty}^{0} x\phi(ax+b)a\, dx$$

which equals

$$\begin{aligned}
\int_{-\infty}^{0} x\phi(ax+b)a\, dx &= \frac{1}{a}\int_{\infty}^{b} (y-b)\phi(y)\, dy \\
&= -\frac{1-\Phi(b)}{a}\int_{b}^{\infty} (y-b)\frac{\phi(y)}{1-\Phi(b)}\, dy \\
&= -\frac{1-\Phi(b)}{a}\left(\frac{\phi(b)}{1-\Phi(b)} - b\right) \\
&= -\frac{\phi(b) - (1-\Phi(b))b}{a}.
\end{aligned}$$

Also,

$$\int_{0}^{\infty} \Phi(ax+b)\, dx = x\Phi(ax+b)|_{0}^{\infty} - \int_{0}^{\infty} ax\phi(ax+b)\, dx$$

which equals

$$\begin{aligned}
-\int_{0}^{\infty} ax\phi(ax+b)\, dx = \frac{1}{a}\int_{-\infty}^{b} (y-b)\phi(y)\, dy &= \frac{\Phi(b)}{a}\int_{-\infty}^{b} (y-b)\frac{\phi(y)}{\Phi(b)}\, dy \\
&= \frac{\Phi(b)}{a}\left(\frac{-\phi(b)}{\Phi(b)} - b\right) \\
&= \frac{-\phi(b) - \Phi(b)b}{a}.
\end{aligned}$$

Thus

$$\int_{\mathbb{R}} |\Phi(ax+b) - \mathbb{1}_{\{x<0\}}|\, dx = \frac{2\phi(b) + 2\Phi(b)b - b}{-a}.$$ □

Recall that $\beta^\delta := \cup_{\boldsymbol{x}\in\beta} B(\boldsymbol{x},\delta)$ and that we let $u_{\boldsymbol{x}}$ be the unit outer normal vector to the manifold β at $\boldsymbol{x}$. The following lemma gives a very useful approximate change of variables type of theorem.

Lemma B.4. *Let either Assumption D1a or Assumption D1b hold, and let D2 hold for the density f_0. Let either $\beta := f_0^{-1}(c)$ in the LS setting or let $\beta := f_0^{-1}(f_{\tau,0})$ in the HDR setting. Let $\delta > 0$ be such that the conclusion of Theorem B.2 holds for β^δ. Let h be a bounded Lebesgue measurable function on β^δ and let $H(\boldsymbol{x}) := \int_{-\delta}^{\delta} h(\boldsymbol{x}+tu_x)dt$. Then*

$$\left|\int_{\beta^\delta} h(\boldsymbol{x})d\boldsymbol{x} - \int_{\beta} H(\boldsymbol{z})d\mathcal{H}^{d-1}(\boldsymbol{z})\right| \le C\sup_{\boldsymbol{x}\in\beta}\int_{-\delta}^{\delta} th(\boldsymbol{x}+tu_{\boldsymbol{x}})dt \tag{41}$$

where C is a constant depending on f_0.

Proof. Since β is compact (by Assumption D1b), it admits a finite "atlas", $\{(U^\alpha,\varphi_\alpha)\}_\alpha$, meaning $\{U^\alpha\}_\alpha$ is an open cover of β, that $\varphi_\alpha : V^\alpha \to U^\alpha$ is a diffeomorphism, and that V^α is open in $\mathbb{R}^{d-1}$. Let $V_\delta^\alpha := V^\alpha \times (-\delta,\delta)$. Let $\Phi_\alpha : V_\delta^\alpha \to \beta^\delta$ be defined by

$$\Phi_\alpha(\boldsymbol{\theta},t) := \varphi_\alpha(\boldsymbol{\theta}) + tu_{\varphi_\alpha(\boldsymbol{\theta})} \quad \text{where} \quad u_{\boldsymbol{x}} := -\frac{\nabla f_0}{\|\nabla f_0\|}(\boldsymbol{x}).$$

Thus $u_{\boldsymbol{x}}$ is the unit outer normal to β at $\boldsymbol{x}\in\beta$. By the change of variables Theorem 2 (page 99) of Evans and Gariepy (2015) (see also the example on page 101),

$$\int_{V^\alpha} h(\varphi_\alpha(\boldsymbol{\theta})))J\varphi_\alpha(\boldsymbol{\theta})d\boldsymbol{\theta} = \int_{U^\alpha} h(\boldsymbol{y})d\mathcal{H}^{d-1}(\boldsymbol{y}). \tag{42}$$

Here,

$$J\varphi_\alpha(\boldsymbol{\theta}) := \det\left[(\nabla\varphi_\alpha(\boldsymbol{\theta}))'\nabla\varphi_\alpha(\boldsymbol{\theta})\right]^{1/2} \tag{43}$$

by Theorem 3 (page 88) of Evans and Gariepy (2015). Similarly,

$$\int_{V_\delta^\alpha} h(\Phi_\alpha((\boldsymbol{\theta},t))J\Phi_\alpha(\boldsymbol{\theta},t)d(\boldsymbol{\theta},t) = \int_{U_\delta^\alpha} h(\boldsymbol{y})d\boldsymbol{y} \tag{44}$$

where $J\Phi_\alpha = |\det\nabla\Phi_\alpha|$ and $U_\delta^\alpha := \Phi_\alpha(V_\delta^\alpha)$. We can see that

$$\nabla\Phi_\alpha(\boldsymbol{\theta},t) = \left(\nabla\varphi_\alpha(\boldsymbol{\theta}) + t\nabla u_{\varphi_\alpha(\boldsymbol{\theta})} | u_{\varphi_\alpha(\boldsymbol{\theta})}\right). \tag{45}$$

Thus, because $u_{\boldsymbol{x}}$ is perpendicular to the tangent space of β at $\boldsymbol{x}$, and this tangent space is equal to the span of the columns of $\nabla\varphi_\alpha(\boldsymbol{\theta})$ for $t\in[-\delta,\delta]$, letting $\boldsymbol{x} = \varphi_\alpha(\boldsymbol{\theta})$, we have

$$\nabla\Phi_\alpha(\boldsymbol{\theta},t)'\nabla\Phi_\alpha(\boldsymbol{\theta},t) = \begin{pmatrix} A_t & t\nabla u_{\boldsymbol{x}}' u_{\boldsymbol{x}} \\ tu_{\boldsymbol{x}}'\nabla u_{\boldsymbol{x}} & 1 \end{pmatrix} \tag{46}$$

where

$$A_t := \nabla\varphi_\alpha(\boldsymbol{\theta})'\nabla\varphi_\alpha(\boldsymbol{\theta}) + t\nabla\varphi_\alpha(\boldsymbol{\theta})'\nabla u_{\boldsymbol{x}} + t\nabla u_{\boldsymbol{x}}'\nabla\varphi_\alpha(\boldsymbol{\theta}) + t^2\nabla u_{\boldsymbol{x}}'\nabla u_{\boldsymbol{x}}. \tag{47}$$

Note that from (46) we have

$$J\Phi_\alpha(\boldsymbol{\theta},0) = J\varphi_\alpha(\boldsymbol{\theta}). \tag{48}$$

Now

$$\det(A+\epsilon AX) = \det A + \epsilon \det A \operatorname{tr} X + O(\epsilon^2) \tag{49}$$

as $\epsilon \to 0$ (Magnus and Neudecker, 1999) for any square matrices A and X of the same dimension. Thus

$$\begin{aligned} J\Phi_\alpha(\boldsymbol{\theta},t) &= \left(\det \nabla\Phi_\alpha(\boldsymbol{\theta},t)'\nabla\Phi_\alpha(\boldsymbol{\theta},t)\right)^{1/2} \\ &= \left(\det \nabla\Phi_\alpha(\boldsymbol{\theta},0)'\nabla\Phi_\alpha(\boldsymbol{\theta},0) + O(t)\right)^{1/2} \qquad \text{by (49), (46), and (47),} \end{aligned}$$

which equals

$$J\Phi_\alpha(\boldsymbol{\theta},0) + O(t) \quad \text{as} \quad t \to 0, \tag{50}$$

by differentiability of $z \mapsto z^{1/2}$ away from 0, since $J\Phi_\alpha(\boldsymbol{\theta},0)$ is uniformly bounded away from 0. The $O(t)$ term is uniform in $\boldsymbol{\theta}$. Thus, by (48), (42), and (44),

$$\int_{U_\delta^\alpha} h(\boldsymbol{y})\, d\boldsymbol{y} = \int_{U^\alpha} H(\boldsymbol{y})\, d\mathcal{H}^{d-1}(\boldsymbol{y}) + E \tag{51}$$

where $|E| \le C \int_{I^d} \int_{-\delta}^{\delta} th(\Phi_\alpha(\boldsymbol{\theta},t))dtd\boldsymbol{\theta}$ where C is the constant from the $O(t)$ term in (50). This proves the lemma if β is parameterizable by a single open set; for the general case, we use a partition of unity. Let $\{\rho_i\}$ be a finite (smooth) partition of unity subordinate to $\{U^\alpha\}$ (Spivak, 1965, page 63). Define $\rho_i^\delta(\boldsymbol{x} + tu_{\boldsymbol{x}}) := \rho_i(\boldsymbol{x})$ for $t \in (-\delta,\delta)$ (which thus forms a partition of unity of β^δ subordinate to $\{U_\delta^\alpha\}_\alpha$). Then replacing h in (51) by $\rho_i^\delta \cdot h$, since each ρ_i is bounded, smooth, and zero outside one of the U_α,

$$\begin{aligned} \int_{\beta^\delta} h(\boldsymbol{y})d\boldsymbol{y} &= \sum_i \int_{\beta^\delta} \rho_i^\delta(\boldsymbol{y})h(\boldsymbol{y})d\boldsymbol{y} \\ &= \sum_i \int_\beta \rho_i H d\mathcal{H}^{d-1} + E_2 = \int_\beta H d\mathcal{H}^{d-1} + E_2 \end{aligned}$$

since $\rho_i^\delta(\boldsymbol{x}+tu_{\boldsymbol{x}}) = \rho_i(\boldsymbol{x})$, and where $|E_2| \le C_2 \sup_{\boldsymbol{x}\in\beta} \int_{-\delta}^{\delta} th(\boldsymbol{x}+tu_{\boldsymbol{x}})dt$. □

Lemma B.5. *Let Assumption D1a hold.*

1. *Assume that γ is a continuously differentiable function on an open neighborhood of β_τ in $\mathbb{R}^d$. For ϵ near 0, let $\beta_\epsilon := f_0^{-1}(f_{\tau,0}+\epsilon)$ and assume β_ϵ is compact for all ϵ in a neighborhood of 0. Then $\epsilon \mapsto \int_{\beta_\epsilon} \gamma d\mathcal{H}$ is continuously differentiable in a neighborhood of $\epsilon = 0$.*
2. *Let $\widehat{f}_{n,\boldsymbol{H}}$ be the KDE (defined in (2)), where K satisfies Assumptions K and K2, and $\boldsymbol{H}$ satisfies $\boldsymbol{H} \to 0$ and $n^{-1}|\boldsymbol{H}|^{-1/2}(\boldsymbol{H}^{-1})^{\otimes 2} = O(1)$ as $n \to \infty$. Let $g_n := \widehat{f}_{n,\boldsymbol{H}} - f_0$. Let $\check{\beta}_{\tau,n} := \widehat{f}_{n,\boldsymbol{H}}^{-1}(f_{\tau,0})$. Assume $\gamma_n \equiv \gamma$ is potentially random but satisfies $\sup_{\boldsymbol{x}\in\beta_\tau^\delta} |\gamma(\boldsymbol{x})| = O_p(1)$ and $\sup_{\boldsymbol{x}\in\beta_\tau^\delta} \|\nabla\gamma(\boldsymbol{x})\| =$*

$O_p(1)$, *for some* $\delta > 0$. *Then*

$$\left| \int_{\beta_\tau} \gamma d\mathcal{H}^{d-1} - \int_{\check{\beta}_{\tau,n}} \gamma d\mathcal{H}^{d-1} \right| = O_p\Big(\sup_{\boldsymbol{x} \in \beta_\tau} \mathbb{E}\left[|g_n(\boldsymbol{x})| + \|\nabla^2 g_n(\boldsymbol{x})\| |g_n(\boldsymbol{x})| + \|\nabla g_n(\boldsymbol{x})\| \right] \Big)$$

as $n \to \infty$.

Proof. **Proof of Part 1**: Fix $\boldsymbol{x}_0 \in \beta_0$. By Assumption D1a, we may assume without loss of generality that $\frac{\partial}{\partial x_d} f(\boldsymbol{x}_0) \neq 0$. Define

$$F(x_1, \ldots, x_d) := (x_1, \ldots, x_{d-1}, f(x_1, \ldots, x_d))$$

and note that $\det \nabla F(\boldsymbol{x}_0) = \frac{\partial}{\partial x_d} f(\boldsymbol{x}_0) \neq 0$. Since f is twice continuously differentiable at x_0 (Assumption D1a), F is twice continuously differentiable at x_0. By the inverse function theorem (pages 67–68, Bredon (1993)), F^{-1} exists and is twice continuously differentiable in a neighborhood of $F(\boldsymbol{x}_0)$. Clearly $F^{-1}(y_1, \ldots, y_d)$ equals $(y_1, \ldots, y_{d-1}, k(y_1, \ldots, y_d))$ for some k that is twice continuously differentiable and satisfies

$$f(y_1, \ldots, y_{d-1}, k(y_1, \ldots, y_d)) = y_d.$$

Thus

$$\varphi_\epsilon(y_1, \ldots, y_{d-1}) := (y_1, \ldots, y_{d-1}, k(y_1, \ldots, y_{d-1}, f_{\tau,0} + \epsilon))$$

is a twice-continuously differentiable invertible parameterization (is a "C^2 diffeomorphism") from an open set $U \subset \mathbb{R}^{d-1}$ to $V_\epsilon \subset \beta_\epsilon$ where $V_\epsilon \ni \boldsymbol{x}_0$ is open in β_ϵ. Each $\boldsymbol{x}_0 \in \beta_\epsilon$ has such a C^2 diffeomorphism onto an open neighborhood $V_\epsilon \subset \beta_\epsilon$; since β_ϵ is compact, we can pick a finite number of them that cover β_ϵ and construct a partition of unity (Spivak, 1965, page 63) on the cover. We will continue considering our fixed $\boldsymbol{x}_0 \in \beta_\epsilon$ and the above-constructed parameterization on a neighborhood of $\boldsymbol{x}_0$. At the end of the proof, our local result can be made global by using the partition of unity.

Now, $\int_{\beta_\epsilon} \gamma d\mathcal{H} = \int_U (\gamma \circ \varphi_\epsilon) J\varphi_\epsilon d\lambda^{d-1}$ where λ^{d-1} is Lebesgue measure (Evans and Gariepy, 2015). Here $J\varphi_\epsilon = \det(\nabla \varphi_\epsilon' \nabla \varphi_\epsilon)^{1/2}$ is continuously differentiable in ϵ (in a neighborhood of 0) since k is twice continuously differentiable and since $\det(\nabla \varphi_\epsilon' \nabla \varphi_\epsilon) \neq 0$. We also know that $\gamma \circ \varphi_\epsilon$ is continuously differentiable in ϵ since γ is assumed continuously differentiable. Since $\frac{\partial}{\partial \epsilon}((\gamma \circ \varphi_\epsilon) J\varphi_\epsilon)$ is continuous so is bounded on $U \times [-\tilde{\epsilon}, \tilde{\epsilon}]$, some $\tilde{\epsilon} > 0$, we can apply the Leibniz rule (Billingsley, 2012) to see that

$$\frac{\partial}{\partial \epsilon} \int_{V_\epsilon} \gamma \, d\mathcal{H}^{d-1} = \frac{\partial}{\partial \epsilon} \int_U (\gamma \circ \varphi_\epsilon) J\varphi_\epsilon d\lambda^{d-1} = \int_U \frac{\partial}{\partial \epsilon} ((\gamma \circ \varphi_\epsilon) J\varphi_\epsilon) d\lambda^{d-1}$$

Thus, the derivative on the left side of the previous display exists, meaning that $\int_{\beta_\epsilon} \gamma \, d\mathcal{H}$ is indeed differentiable for ϵ near 0, as desired. This is true on the neighborhood V_ϵ; it extends to the case where V_ϵ is replaced by β_ϵ by using the partition of unity we constructed above.

Proof of Part 2: We write $g \equiv g_n$, suppressing dependence on n. For $\boldsymbol{x} \in \mathbb{R}^d$, let $h(\boldsymbol{x}, \delta) := f_0(\boldsymbol{x}) + \delta g(\boldsymbol{x})$, and let $\beta_\delta := h_\delta^{-1}(f_{\tau,0})$. We will explicitly construct $\phi_\delta : U \to V_\delta$, for some open $U \subset \mathbb{R}^{d-1}$ and $V_\delta \subset \beta_\delta$, by the inverse function theorem, and then check that $\frac{\partial}{\partial \delta}\phi_\delta(\boldsymbol{z})$ is $O_p(|g(\phi_\delta(\boldsymbol{z}))|)$ and that $\frac{\partial}{\partial \delta} J\phi_\delta(\boldsymbol{z})$ is $O_p(\|\nabla^2 g(\phi_\delta(\boldsymbol{z}))\||g(\phi_\delta(\boldsymbol{z}))| + |g(\phi_\delta(\boldsymbol{z}))| + \|\nabla g(\phi_\delta(\boldsymbol{z}))\|)$. Then the proof can be finished as the proof of the previous part was finished.

Fix $\boldsymbol{x}_0 \in \beta_\tau \equiv \beta_0$. Define $F(x_1, \ldots, x_d, \delta) := (x_1, \ldots, x_{d-1}, h(\boldsymbol{x}, \delta), \delta)$. As in the proof of the previous part, note that $\det \nabla F(\boldsymbol{x}_0) \neq 0$ (when $\|\nabla g(\boldsymbol{x}_0\|$ is small), so by the inverse function theorem F^{-1} exists, is twice continuously differentiable in a neighborhood of $F(\boldsymbol{x}_0)$, and clearly satisfies $F^{-1}(y_1, \ldots, y_d, \delta) = (y_1, \ldots, y_{d-1}, k(y_1, \ldots, y_d, \delta), \delta)$. Let $\boldsymbol{z} := (\boldsymbol{x}, \delta)$ and note by definition

$$k(F(\boldsymbol{z})) = k(x_1, \ldots, x_{d-1}, h(\boldsymbol{z}), \delta) = x_d. \tag{52}$$

From this we will derive formulas for the first and second derivatives of k. In this proof, for a function $f\colon \mathbb{R}^p \to \mathbb{R}$ we use the notation $f_i(\boldsymbol{x})$ for $\frac{\partial}{\partial x_i} f(x_1, \ldots, x_d)$ and $f_{ij}(\boldsymbol{x})$ for $\frac{\partial^2}{\partial x_i \partial x_j} f(x_1, \ldots, x_d)$. Taking $\frac{\partial}{\partial x_i}$ of (52) for $1 \leq i \leq d-1$, we see that

$$k_i(F(\boldsymbol{z})) = -k_d(F(\boldsymbol{z}))h_i(\boldsymbol{x}). \tag{53}$$

Applying $\frac{\partial}{\partial x_d}$ to (52), we get that

$$k_d(F(\boldsymbol{z}))h_d(\boldsymbol{z}) = 1 \quad \text{or} \quad k_d(F(\boldsymbol{z})) = 1/h_d(\boldsymbol{z}), \tag{54}$$

and applying $\frac{\partial}{\partial \delta}$ to (52), we get

$$k_d(F(\boldsymbol{z}))h_{d+1}(\boldsymbol{z}) + k_{d+1}(F(\boldsymbol{z})) = 0, \quad \text{or} \quad k_{d+1}(F(\boldsymbol{z})) = -\frac{h_{d+1}(\boldsymbol{z})}{h_d(\boldsymbol{z})} = -\frac{g(\boldsymbol{x})}{h_d(\boldsymbol{z})}. \tag{55}$$

Applying $\frac{\partial}{\partial \delta}$ to (53) yields

$$\begin{aligned} &k_{i,d}(F(\boldsymbol{z}))h_{d+1}(\boldsymbol{z}) + k_{i,d+1}(F(\boldsymbol{z})) \\ &\quad = -\big(k_{d,d}(F(\boldsymbol{z}))h_{d+1}(\boldsymbol{z}) + k_{d,d+1}(F(\boldsymbol{z}))\big)h_i(\boldsymbol{z}) - k_d(F(\boldsymbol{z}))h_{i,d+1}(\boldsymbol{x}) \end{aligned} \tag{56}$$

and, letting $\boldsymbol{y} := F(\boldsymbol{z})$, since $h_{d+1}(\boldsymbol{z}) = g(\boldsymbol{x})$ and $h_{i,d+1}(\boldsymbol{z}) = g_i(\boldsymbol{x})$, this implies that

$$k_{i,d+1}(\boldsymbol{y}) = -k_{i,d}(\boldsymbol{y})g(\boldsymbol{x}) - \Big(k_{d,d}(\boldsymbol{y})g(\boldsymbol{x}) + k_{d,d+1}(\boldsymbol{y})\Big)h_i(\boldsymbol{z}) - k_d(\boldsymbol{y})g_i(\boldsymbol{x}). \tag{57}$$

To understand the expression in (57) we need to control $k_{i,d}$, $k_{d,d}$, and $k_{d,d+1}$. Applying $\frac{\partial}{\partial \delta}$ to (54) we see that

$$k_{d,d}(F(\boldsymbol{z}))h_{d+1}(\boldsymbol{z}) + k_{d,d+1}(F(\boldsymbol{z})) = -\frac{h_{d,d+1}(\boldsymbol{z})}{h_d^2(\boldsymbol{z})},$$

so

$$k_{d,d+1}(F(\boldsymbol{z})) = k_{d,d}(F(\boldsymbol{z}))g(\boldsymbol{x}) - \frac{g_d(\boldsymbol{x})}{h_d^2(\boldsymbol{z})}. \tag{58}$$

We will next verify that $k_{i,d}$ and $k_{d,d}$ are $O_p(1+\|\nabla^2 g\|)$ (which is $O_p(1)$ under our assumption on $\boldsymbol{H}$ (Chacón, Duong and Wand, 2011)). Then by (57) and (58), we will see, uniformly for $\delta \in [-1,1]$, that

$$k_{i,d+1}(F(\boldsymbol{z})) = O_p(|g(\boldsymbol{x})| + \|\nabla g(\boldsymbol{x})\| + \|\nabla^2 g(\boldsymbol{x})\||g(\boldsymbol{x})|) \quad \text{as} \quad n \to \infty. \tag{59}$$

Note that by (54), $k_d(F(\boldsymbol{z})) = O_p(1)$ and $1/h_d(\boldsymbol{z}) = O_p(1)$ (since by assumption $\frac{\partial}{\partial x_d} f(\boldsymbol{x}_0) \neq 0$ and $\|\nabla g(\boldsymbol{x})\| \to_p 0$).

Now applying $\partial/\partial x_d$ to (54), we see

$$k_{d,d}(F(\boldsymbol{z})) = -\frac{k_d^2(F(\boldsymbol{z}))h_{d,d}(\boldsymbol{z})}{h_d(\boldsymbol{z})} = -\frac{h_{d,d}(\boldsymbol{z})}{h_d^3(\boldsymbol{z})}, \tag{60}$$

so $k_{d,d}(F(\boldsymbol{z})) = O_p(1+\|\nabla^2 g(\boldsymbol{x})\|)$. Applying $\partial/\partial x_i$ to (the left expression in) (54) yields

$$k_{i,d}(F(\boldsymbol{z})) + k_{d,d}(F(\boldsymbol{z}))h_i(\boldsymbol{z}) = -\frac{h_{d,i}(\boldsymbol{z})}{h_d^2(\boldsymbol{z})}. \tag{61}$$

Thus by (60) we see $k_{i,d}(F(\boldsymbol{z})) = O_p(1+\|\nabla^2 g(\boldsymbol{x})\|)$, so (59) holds.

Now we let

$$\phi_\delta(y_1,\dots,y_{d-1}) := (y_1,\dots,y_{d-1},k(y_1,\dots,y_{d-1},f_{\tau,0},\delta)),$$

which we have shown is a C^2 parameterization from an open set $U \subset \mathbb{R}^{d-1}$ to $V_\delta \subset \beta_\delta$ where $V_\delta \ni \boldsymbol{x}_0$ is open in β_δ. We can check that $J\phi_\delta = \det(\nabla\phi_\delta'\nabla\phi_\delta)^{1/2}$ is continuously differentiable in δ for $\delta \in [-1,1]$ by (59), and, by three Taylor expansions,

$$\int_{V_1} \gamma d\mathcal{H}^{d-1} = \int_U (\gamma\circ\phi_1)J\phi_1 d\lambda^{d-1} = \int_U ((\gamma\circ\phi_0(\boldsymbol{y}))J\phi_0(\boldsymbol{y}) + \epsilon(\boldsymbol{y})\,d\boldsymbol{y} \tag{62}$$

where $\epsilon(\boldsymbol{y}) = O_p(|g(\boldsymbol{x})|+\|\nabla g(\boldsymbol{x})\|+\|\nabla^2 g(\boldsymbol{x})\||g(\boldsymbol{x})|)$, since we have $\frac{\partial}{\partial\delta}J\phi_\delta(\boldsymbol{y})$ is $O_p(|g(\boldsymbol{x})|+\|\nabla g(\boldsymbol{x})\|+\|\nabla^2 g(\boldsymbol{x})\||g(\boldsymbol{x})|)$ uniformly for $\delta\in[-1,1]$, since $\frac{\partial}{\partial\delta}\phi_\delta(\boldsymbol{y}) = O_p(|g(\boldsymbol{x})|)$ uniformly for $\delta\in[-1,1]$ (by (55)), and since γ is continuously differentiable in a neighborhood of β_τ. In fact, we can see that $\mathbb{E}|\epsilon(\boldsymbol{y})|$ is be bounded by $C\mathbb{E}\left[|g(\boldsymbol{x})| + \|\nabla g(\boldsymbol{x})\| + \|\nabla^2 g(\boldsymbol{x})\||g(\boldsymbol{x})|\right]$ for a constant $C>0$. By the Fubini-Tonelli theorem, $\mathbb{E}\int_U |\epsilon(\boldsymbol{y})|\,d\boldsymbol{y} = \int \mathbb{E}|\epsilon(\boldsymbol{y})|\,d\boldsymbol{y}$, so we can see

$$\int_U \epsilon(\boldsymbol{y})\,d\boldsymbol{y} = O_p \sup_{\boldsymbol{x}\in\beta_\tau} \mathbb{E}\left[|g(\boldsymbol{x})| + \|\nabla g(\boldsymbol{x})\| + \|\nabla^2 g(\boldsymbol{x})\||g(\boldsymbol{x})|\right] \tag{63}$$

by Markov's inequality. Combining (62), (63), and $\int_U((\gamma\circ\phi_0(\boldsymbol{y}))J\phi_0(\boldsymbol{y}) = \int_{V_0}\gamma d\mathcal{H}^{d-1}$ we get

$$\int_{V_1}\gamma d\mathcal{H}^{d-1} = \int_{V_0}\gamma d\mathcal{H}^{d-1} + O_p \sup_{\boldsymbol{x}\in\beta_\tau} \mathbb{E}\left[|g(\boldsymbol{x})| + \|\nabla g(\boldsymbol{x})\| + \|\nabla^2 g(\boldsymbol{x})\||g(\boldsymbol{x})|\right].$$

Then the proof can be finished as in the proof of Part 1, including using a partition of unity to extend V_1 to $\check{\beta}_\tau$ and V_0 to β_τ to conclude from the previous display

$$\int_{\check{\beta}_\tau} \gamma d\mathcal{H}^{d-1} = \int_{\beta_\tau} \gamma d\mathcal{H}^{d-1} + O_p \sup_{\boldsymbol{x}\in\beta_\tau} \mathbb{E}\left[|g(\boldsymbol{x})| + \|\nabla g(\boldsymbol{x})\| + \|\nabla^2 g(\boldsymbol{x})\||g(\boldsymbol{x})|\right].$$

□

B.1. Proof of Corollary 2.2

By our assumptions of unimodality and spherical symmetry of f_0, we have that ∇f_0 and $\nabla^2 f_0$ are constant on β_τ, and we denote these two quantities as $\nabla_\tau f_0$ and $\nabla^2_\tau f_0$. Then for $h > 0$ we can write

$$B(h) = -(nh^{d+4})^{1/2} F_1 \quad \text{where} \quad F_1 := \frac{\mu_2(K)\,\mathrm{tr}(\nabla^2_\tau f_0)}{2\sqrt{R(K) f_{\tau,0}}},$$

and $C(h) = B(h) + (nh^{d+4})^{1/2} F_2$ where

$$F_2 := \|\nabla_\tau f_0\| \left(\int_{\beta_\tau} d\mathcal{H}\right)^{-1} \left\{\frac{\mu_2(K)\,\mathrm{tr}(\nabla^2_\tau f_0)}{2\|\nabla_\tau f_0\|} \int_{\beta_\tau} d\mathcal{H} + \frac{\mu_2(K)}{2 f_{\tau,0}}\,\mathrm{tr}(\nabla^2_\tau f_0) \int_{\mathcal{L}_\tau} d\boldsymbol{x}\right\}.$$

Then

$$\mathrm{HDR}(h) = \frac{f_{\tau,0}}{A}\left(\int_{\beta_\tau} d\mathcal{H}\right)(nh^d)^{-1/2}\left(2\phi(C(h)) + (2\Phi(C(h)) - 1)C(h)\right)$$

where $A = \|\nabla_\tau f_0\|/\sqrt{R(K) f_{\tau,0}}$. Note that $2\phi(C(h)) + (2\Phi(C(h)) - 1)C(h) = 2\phi(|C(h)|) + (2\Phi(|C(h)|) - 1)|C(h)|$. Let $G := |C(h)|/(nh^{d+4})^{1/2} = |F_2 - F_1|$, We will thus minimize

$$\begin{aligned} & n^{2/(d+4)} \left(\frac{A}{f_{\tau,0}} \int_{\beta_\tau} d\mathcal{H}\right)^{-1} \mathrm{HDR}(h) \\ &\quad = (n^{1/2} h^{(d+4)/2})^{-d/(d+4)} \\ &\qquad \times \left(2\phi(G(nh^{d+4})^{1/2}) + G(nh^{d+4})^{1/2}(2\Phi(G(nh^{d+4})^{1/2}) - 1)\right) \end{aligned} \tag{64}$$

over $h \geq 0$. By the change of variables

$$s = (nh^{d+4})^{1/2}, \tag{65}$$

minimizing (64) is equivalent to minimizing

$$\mathrm{HDR}^*(s) := 2s^{-d/d+4}\phi(Gs) + Gs^{4/d+4}(2\Phi(Gs) - 1).$$

Note that $\mathrm{HDR}^*(s) \to \infty$ as $s \to \infty$ and as $s \searrow 0$, so $\mathrm{HDR}^*(s)$ attains its minimum on $(0, \infty)$. Now, HDR^* has a unique minimum if $(\mathrm{HDR}^*)'(s)$ has a unique 0, and by calculation,

$$(\mathrm{HDR}^*)'(s) = 2\frac{-d}{d+4} s^{\frac{-2d-4}{d+4}} \phi(Gs) + G\frac{4}{d+4} s^{\frac{-d}{d+4}}(2\Phi(Gs) - 1),$$

$(\mathrm{HDR}^*)'(s)$ has a unique 0 if and only if

$$(\mathrm{HDR}^*)'(s)(2(d/d+4)s^{-(2d+4)/d+4}\phi(Gs))^{-1} = -1 + \frac{2}{d}\frac{Gs(2\Phi(Gs)-1)}{\phi(Gs)} \quad (66)$$

has a unique 0. We can compute the derivative of (66) to be

$$G\frac{2}{d}\left(2Gs + \frac{(1+G^2s^2)(2\Phi(Gs)-1)}{\phi(Gs)}\right) > 0$$

for $s \in (0,\infty)$. Thus (66) is strictly increasing on $(0,\infty)$, is negative at 0, and approaches ∞ as $c \to \infty$, and so (66) has a unique zero. Let $s_{\mathrm{opt}} > 0$ be the unique minimum of $\mathrm{HDR}^*(s)$, and let $h_{\mathrm{opt}} := s_{\mathrm{opt}}^{2/d+4} n^{-1/d+4}$. By (65), h_{opt} minimizes (64), and so minimizes $\mathrm{HDR}(h)$. By Theorem 2.2, we conclude that for any h_0 that minimizes $\mathbb{E}[\mu_{f_0}\{\mathcal{L}_\tau \Delta\hat{\mathcal{L}}_{\tau,\boldsymbol{H}}\}]$, $h_0 = h_{\mathrm{opt}}(1+o(1))$.

Appendix C: Proof of intermediate results

Proof of Lemma A.1. Let $C_1 > 1 + 2\lambda(\{f_0(\boldsymbol{x}) \geq f_{\tau,0}\}) / \int_{\beta_\tau} \frac{f_0}{\|\nabla f_0\|}\, d\mathcal{H}$. Then when $\varepsilon > 0$ is sufficiently small,

$$\begin{aligned}
\int \tilde{f}(\boldsymbol{x}) \mathbb{1}_{\{\tilde{f}(\boldsymbol{x}) \geq f_{\tau,0} - C_1\varepsilon\}}\, d\boldsymbol{x} &\geq \int (f_0(\boldsymbol{x}) - \varepsilon)\mathbb{1}_{\{f(\boldsymbol{x}) \geq f_{\tau,0} - (C_1-1)\varepsilon\}}\, d\boldsymbol{x} \\
&= 1 - \tau + \int f_0(\boldsymbol{x}) \mathbb{1}_{\{f_{\tau,0} - (C_1-1)\varepsilon \leq f_0(\boldsymbol{x}) < f_{\tau,0}\}}\, d\boldsymbol{x} \\
&\quad - \varepsilon\lambda(\{f_0(\boldsymbol{x}) \geq f_{\tau,0} - (C_1-1)\varepsilon\}) \\
&\geq 1 - \tau + \int f_0(\boldsymbol{x}) \mathbb{1}_{\{f_{\tau,0} - (C_1-1)\varepsilon \leq f_0(\boldsymbol{x}) < f_{\tau,0}\}}\, d\boldsymbol{x} \\
&\quad - 2\varepsilon\lambda(\{f_0(\boldsymbol{x}) \geq f_{\tau,0}\}).
\end{aligned}$$

By Proposition A.1 of Cadre (2006),

$$\int f_0(\boldsymbol{x}) \mathbb{1}_{\{f_{\tau,0} - (C_1-1)\varepsilon \leq f_0(\boldsymbol{x}) < f_{\tau,0}\}}\, d\boldsymbol{x} = \int_{f_{\tau,0} - (C_1-1)\varepsilon}^{f_{\tau,0}} \int_{\beta(s)} \frac{f_0(\boldsymbol{x})}{\|\nabla f_0(\boldsymbol{x})\|}\, d\mathcal{H}(\boldsymbol{x})\, ds. \quad (67)$$

So we can express $\int f_0(\boldsymbol{x}) \mathbb{1}_{\{f_{\tau,0} - (C_1-1)\varepsilon \leq f_0(\boldsymbol{x}) < f_{\tau,0}\}}\, d\boldsymbol{x}$ as

$$(C_1 - 1)\varepsilon \int_{\beta_\tau} \frac{f_0(\boldsymbol{x})}{\|\nabla f_0(\boldsymbol{x})\|}\, d\mathcal{H}(\boldsymbol{x}) + O(\varepsilon^2), \quad (68)$$

by Lemma B.5, and thus see that

$$\begin{aligned}
&\int \tilde{f}(\boldsymbol{x}) \mathbb{1}_{\{\tilde{f}(\boldsymbol{x}) \geq f_{\tau,0} - C_1\varepsilon\}}\, d\boldsymbol{x} \\
&\quad \geq 1 - \tau + (C_1 - 1)\varepsilon \int_{\beta_\tau} \frac{f_0(\boldsymbol{x})}{\|\nabla f_0(\boldsymbol{x})\|}\, d\mathcal{H}(\boldsymbol{x}) + o(\varepsilon) - 2\varepsilon\lambda(\{\boldsymbol{x} : f_0(\boldsymbol{x}) \geq f_{\tau,0}\})
\end{aligned}$$

$$> 1 - \tau,$$

when $\varepsilon > 0$ is sufficiently small. So $\tilde{f}_\tau > f_{\tau,0} - C_1\varepsilon$. For the upper bound, with a similar argument, we get $\tilde{f}_\tau < f_{\tau,0} + C_1\varepsilon$. So we proved $|\tilde{f}_\tau - f_{\tau,0}| \leq C_1\varepsilon$ for $\varepsilon > 0$ sufficiently small. □

Proof of Lemma A.2. We first prove an intermediate result that

$$\int_{\mathcal{L}_\delta(f_{\tau,0})^c} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) \geq \widehat{f}_{\tau,n}\right) d\boldsymbol{x} + \int_{\mathcal{L}_{-\delta}(f_{\tau,0})} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) < \widehat{f}_{\tau,n}\right) d\boldsymbol{x} \tag{69}$$

is $o(n^{-1})$ as $n \to \infty$ for fixed $\delta > 0$ sufficiently small. Observe that under Assumption D1b if $\delta > 0$ is sufficiently small, then there exists $\varepsilon > 0$ such that $f_0(\boldsymbol{x}) \leq f_{\tau,0} - \varepsilon$ for $\boldsymbol{x} \in \mathcal{L}_\delta(f_{\tau,0})^c$ and $f_0(\boldsymbol{x}) \geq f_{\tau,0} + \varepsilon$ for $\boldsymbol{x} \in \mathcal{L}_{-\delta}(f_{\tau,0})$. By reducing $\delta > 0$ if necessary, for $\boldsymbol{x} \in \mathcal{L}_\delta(f_{\tau,0})^c$,

$$\begin{aligned} P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) \geq \widehat{f}_{\tau,n}\right) &\leq P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) - f_0(\boldsymbol{x}) - (\widehat{f}_{\tau,n} - f_{\tau,0}) \geq \varepsilon\right) \\ &\leq P\left(\|\widehat{f}_{n,\boldsymbol{H}} - f_0\|_\infty \geq \varepsilon/2\right) + P\left(|\widehat{f}_{\tau,n} - f_{\tau,0}| \geq \varepsilon/2\right) \\ &\leq P\left(\|\widehat{f}_{n,\boldsymbol{H}} - f_0\|_\infty \geq \frac{\varepsilon}{2C_1}\right) + P\left(|\widehat{f}_{\tau,n} - f_{\tau,0}| \geq \frac{\varepsilon}{2}\right), \end{aligned}$$

where $C_1 \geq 1$ is the constant we defined in Lemma A.1 ; by that lemma, we have

$$P\left(|\widehat{f}_{\tau,n} - f_{\tau,0}| \geq \frac{\varepsilon}{2}\right) \leq P\left(\|\widehat{f}_{n,\boldsymbol{H}} - f_0\|_\infty \geq \frac{\varepsilon}{2C_1}\right),$$

so

$$P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) \geq \widehat{f}_{\tau,n}\right) \leq 2P\left(\|\widehat{f}_{n,\boldsymbol{H}} - f_0\|_\infty \geq \frac{\varepsilon}{2C_1}\right). \tag{70}$$

A similar argument yields the same upper bound for $P(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) < \widehat{f}_{\tau,n})$ when $\boldsymbol{x} \in \mathcal{L}_{-\delta}(f_{\tau,0})$. Now by Assumption D1b,

$$\|\mathbb{E}(\widehat{f}_{n,\boldsymbol{H}}) - f_0\|_\infty \to 0,$$

as $n \to \infty$. Together with the inequality (70) together, this yields that for n sufficiently large,

$$\begin{aligned} &\int_{\mathcal{L}_\delta(f_{\tau,0})^c} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) \geq \widehat{f}_{\tau,n}\right) d\boldsymbol{x} + \int_{\mathcal{L}_{-\delta}(f_{\tau,0})} f_0(\boldsymbol{x}) P\left(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) < \widehat{f}_{\tau,n}\right) d\boldsymbol{x} \\ &\quad \leq 2P\left(\|\widehat{f}_{n,\boldsymbol{H}} - f_0\|_\infty \geq \frac{\varepsilon}{2C_1}\right) \end{aligned}$$

which is bounded above by

$$2P\left(\|\widehat{f}_{n,\boldsymbol{H}} - \mathbb{E}(\widehat{f}_{n,\boldsymbol{H}})\|_\infty \geq \frac{\varepsilon}{4C_1}\right) + 2P\left(\|\mathbb{E}(\widehat{f}_{n,\boldsymbol{H}}) - f_0\|_\infty \geq \frac{\varepsilon}{4C_1}\right)$$

$$
\begin{aligned}
&= 2P\left(\|\widehat{f}_{n,\boldsymbol{H}} - \mathbb{E}(\widehat{f}_{n,\boldsymbol{H}})\|_\infty \geq \frac{\varepsilon}{4C_1}\right) \\
&\leq L\exp\left\{-\frac{C_{0,1}\varepsilon^2 n|\boldsymbol{H}|^{1/2}}{16C_1^2}\right\} \\
&= o(n^{-1}),
\end{aligned}
$$

where the last inequality comes from Corollary B.1.

Now it suffices to show that $E(\delta,\delta_n) = o(n^{-1})$, where $E(\delta,\delta_n)$ is defined as

$$
\begin{aligned}
&\int_{\mathcal{L}_{\delta_n}(f_{\tau,0})^c\backslash\mathcal{L}_{\delta}(f_{\tau,0})^c} f_0(\boldsymbol{x})P(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) \geq \widehat{f}_{\tau,n})\,d\boldsymbol{x} \\
&\quad + \int_{\mathcal{L}_{-\delta_n}(f_{\tau,0})\backslash\mathcal{L}_{-\delta}(f_{\tau,0})} f_0(\boldsymbol{x})P(\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) < \widehat{f}_{\tau,n})\,d\boldsymbol{x}.
\end{aligned}
$$

Using Taylor expansion, we have

$$
\|\mathbb{E}(\widehat{f}_{n,\boldsymbol{H}}) - f_0\|_\infty = \sup_{\boldsymbol{x}\in\mathbb{R}^d}\left|\int K(\boldsymbol{z})\left\{\frac{1}{2}(\boldsymbol{H}^{1/2}\boldsymbol{z})'\nabla^2 f_0(\boldsymbol{x}_z)\boldsymbol{H}^{1/2}\boldsymbol{z}\right\}d\boldsymbol{z}\right|,
$$

where $\boldsymbol{x}_z = \boldsymbol{x} - c\boldsymbol{H}^{1/2}\boldsymbol{z}$ for some $c\in(0,1)$. Under Assumption D1b, f_0 has bounded second derivatives and let $A>0$ be such that $\|\nabla^2 f\|_\infty \leq A$. Then

$$
\|\mathbb{E}(\widehat{f}_{n,\boldsymbol{H}}) - f_0\|_\infty \leq \frac{1}{2}dA\int K(\boldsymbol{z})\boldsymbol{z}'\boldsymbol{H}\boldsymbol{z}\,d\boldsymbol{z} = \frac{1}{2}dA\mu_2(K)\operatorname{tr}(\boldsymbol{H}) = O\left\{\lambda_{\max}(\boldsymbol{H})\right\}, \tag{71}
$$

as $|\boldsymbol{H}|\to 0$. Now by Lemma B.2, there exists a constant c_2 small enough that if we take $\varepsilon_n = c_2\delta_n$, then we have $|f_0(\boldsymbol{x}) - f_{\tau,0}| \geq \varepsilon_n$ for all $\boldsymbol{x}\in(\mathcal{L}_{\delta_n}(f_{\tau,0})^c\backslash\mathcal{L}_{\delta}(f_{\tau,0})^c)\cup(\mathcal{L}_{-\delta_n}(f_{\tau,0})\backslash\mathcal{L}_{-\delta}(f_{\tau,0}))$. Moreover, $\lambda_{\max}(\boldsymbol{H}) = o(\epsilon_n)$ by our assumption, so for n sufficiently large, by (71), $P(\|\mathbb{E}(\widehat{f}_{n,\boldsymbol{H}}) - f_0\|_\infty \geq \frac{\varepsilon_n}{4C}) = 0$. Then for n large enough,

$$
\begin{aligned}
E(\delta,\delta_n) &\leq 2P\left(\|\widehat{f}_{n,\boldsymbol{H}} - \mathbb{E}(\widehat{f}_{n,\boldsymbol{H}})\|_\infty \geq \frac{\varepsilon_n}{4C}\right) \\
&\leq L\exp\left\{-\frac{C_{0,1}\varepsilon_n^2 n|\boldsymbol{H}|^{1/2}}{16C_1^2}\right\} = o(n^{-1}).
\end{aligned} \tag{72}
$$

as $n\to\infty$. □

Proof of Lemma A.3. Let $\boldsymbol{z}\in\{\boldsymbol{x}\in\mathbb{R}^d : f_0(\boldsymbol{x}) = \tilde{f}_\tau\}$ and let $\boldsymbol{y}\in\{\boldsymbol{x}\in\mathbb{R}^d : \tilde{f}(\boldsymbol{x}) = \tilde{f}_\tau\}$ be such that $\boldsymbol{z} = \boldsymbol{x} + \eta_1 u_{\boldsymbol{x}}$ and $\boldsymbol{y} = \boldsymbol{x} + \eta_2 u_{\boldsymbol{x}}$ for some $\boldsymbol{x}\in\beta_\tau$ and $\eta_i \equiv \eta_i(\boldsymbol{x})\in\mathbb{R}$, $i = 1,2$. By Taylor expansion, we have

$$
\begin{aligned}
f_0(\boldsymbol{z}) &= f_0(\boldsymbol{x} + \eta_1 u_{\boldsymbol{x}}) \\
&= f_0(\boldsymbol{x}) + \eta_1 u_{\boldsymbol{x}}'\nabla f_0(\boldsymbol{x}) + \frac{1}{2}u_{\boldsymbol{x}}'\nabla^2 f_0(\boldsymbol{x} + s_1\eta_1 u_{\boldsymbol{x}})u_{\boldsymbol{x}}\eta_1^2,
\end{aligned}
$$

where $s_1 \in [0,1]$, and

$$\begin{aligned}\tilde{f}(\boldsymbol{y}) &= \tilde{f}(x+\eta_2 u_{\boldsymbol{x}}) \\ &= \tilde{f}(\boldsymbol{x}) + \eta_2 u_{\boldsymbol{x}}' \nabla \tilde{f}(\boldsymbol{x}+s_2\eta_2 u_{\boldsymbol{x}}).\end{aligned}$$

We then see

$$\begin{aligned}0 &= f_0(\boldsymbol{z}) - \tilde{f}(\boldsymbol{y}) \\ &= f_0(\boldsymbol{x}) + \eta_1 u_{\boldsymbol{x}}' \nabla f_0(\boldsymbol{x}) + \frac{1}{2} u_{\boldsymbol{x}}' \nabla^2 f_0(\boldsymbol{x}+s_1\eta_1 u_{\boldsymbol{x}}) u_{\boldsymbol{x}} \eta_1^2 \\ &\quad - \tilde{f}(\boldsymbol{x}) - \eta_2 u_{\boldsymbol{x}}' \nabla \tilde{f}(\boldsymbol{x}+s_2\eta_2 u_{\boldsymbol{x}}),\end{aligned} \tag{73}$$

where $s_2 \in [0,1]$. We thus have

$$\begin{aligned}\eta_1 - \eta_2 = \frac{f_0(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x})}{\|\nabla f_0(\boldsymbol{x})\|} + \frac{u_{\boldsymbol{x}}' \nabla^2 f_0(\boldsymbol{x}+s_1\eta_1 u_{\boldsymbol{x}}) u_{\boldsymbol{x}} \eta_1^2}{2\|\nabla f_0(\boldsymbol{x})\|} \\ + \eta_2 \frac{\left\langle \nabla \tilde{f}(\boldsymbol{x}+s_2\eta_2 u_{\boldsymbol{x}}) - \nabla f(\boldsymbol{x}), \nabla f_0(\boldsymbol{x}) \right\rangle}{\|\nabla f_0(\boldsymbol{x})\|^2}.\end{aligned} \tag{74}$$

A similar analysis as in (73), beginning from the identity $\tilde{f}_\tau - f_{\tau,0} = \tilde{f}(\boldsymbol{y}) - f_0(\boldsymbol{x})$ shows that $\eta_2 = O(\|g\|_\infty)$ since by Lemma A.1, $\tilde{f}_\tau - f_{\tau,0} = O(\|g\|_\infty)$. (Similarly, $\eta_1 = O(\|g\|_\infty)$.) Since by Assumption D1b, f_0 has bounded second derivatives,the second term on the right in (74) is $O(\|g\|_\infty^2)$. For the second term, note

$$\begin{aligned}&\frac{\left\langle \nabla \tilde{f}(\boldsymbol{x}+s_2\eta_2 u_{\boldsymbol{x}}) - \nabla f(\boldsymbol{x}), \nabla f_0(\boldsymbol{x}) \right\rangle}{\|\nabla f_0(\boldsymbol{x})\|^2} \\ &\quad = \frac{\left\langle \nabla \tilde{f}(\boldsymbol{x}+s_2\eta_2 u_{\boldsymbol{x}}) - \nabla f_0(\boldsymbol{x}+s_2\eta_2 u_{\boldsymbol{x}}), \nabla f_0(\boldsymbol{x}) \right\rangle}{\|\nabla f_0(\boldsymbol{x})\|^2} \\ &\qquad + \frac{\langle \nabla f_0(\boldsymbol{x}+s_2\eta_2 u_{\boldsymbol{x}}) - \nabla f_0(\boldsymbol{x}), \nabla f_0(\boldsymbol{x}) \rangle}{\|\nabla f_0(\boldsymbol{x})\|^2},\end{aligned}$$

and by Assumption D1b, $\nabla f_0(\boldsymbol{x})$ is Lipschitz, we have the third term on the right of (74) is $O(\|g\|_\infty \|\nabla g\|_\infty + \|g\|_\infty^2)$.

We will apply Lemma B.4 to $h(\boldsymbol{y}) = \mathbb{1}_{\{\tilde{f}(\boldsymbol{y}) \geq \tilde{f}_\tau\}} - \mathbb{1}_{\{f_0(\boldsymbol{y}) \geq \tilde{f}_\tau\}}$. For $\|g\|_\infty$ small enough, $\{\tilde{f} \geq \tilde{f}_\tau\} \Delta \{f_0 \geq \tilde{f}_\tau\} \subset \beta_\tau^\delta$ for some $\delta > 0$, by Lemma A.1, and by Assumption D1b (a) and (b). Thus the left side of (17) equals $\int_{\beta_\tau^\delta} h(\boldsymbol{y}) d\boldsymbol{y}$. We may shrink δ so that the conclusion of Theorem B.2 holds, so that for each $\boldsymbol{y} \in \beta_\tau^\delta$ there is a unique closest $\boldsymbol{x}_{\boldsymbol{y}} \in \beta_\tau$. Now, for δ small enough, considering $\mathbb{1}_{\{\tilde{f}(\boldsymbol{x}+tu_{\boldsymbol{x}}) \geq \tilde{f}_\tau\}}$ as a function of $t \in [-\delta, \delta]$, we can see that $\mathbb{1}_{\{\tilde{f}(\boldsymbol{x}+tu_{\boldsymbol{x}}) \geq \tilde{f}_\tau\}} = \mathbb{1}_{\{-\delta \leq t \leq \eta_2(\boldsymbol{x})\}}$, because $\nabla \tilde{f}(\boldsymbol{x})' \nabla f_0(\boldsymbol{x}) > 0$, so $\tilde{f}$ is locally strictly decreasing in the direction of $u_{\boldsymbol{x}} = -\nabla f_0(\boldsymbol{x}) / \|\nabla f_0(\boldsymbol{x})\|$. Similarly $\mathbb{1}_{\{f_0(\boldsymbol{x}+tu_{\boldsymbol{x}}) \geq \tilde{f}_\tau\}} = \mathbb{1}_{\{-\delta \leq t \leq \eta_1(\boldsymbol{x})\}}$. Thus for $\boldsymbol{y} \in \beta_\tau^\delta$,

$$h(\boldsymbol{y}) = \mathbb{1}_{\{\eta_1(\boldsymbol{x}_{\boldsymbol{y}}) \leq t \leq \eta_2(\boldsymbol{x}_{\boldsymbol{y}})\}} - \mathbb{1}_{\{\eta_2(\boldsymbol{x}_{\boldsymbol{y}}) \leq t \leq \eta_1(\boldsymbol{x}_{\boldsymbol{y}})\}}, \tag{75}$$

(where $\mathbb{1}_{\{a\le t\le b\}}$ is just identically 0 if $b<a$) and so for $\boldsymbol{x}\in\beta_\tau$,

$$H(\boldsymbol{x}) := \int_{-\delta}^{\delta} h(\boldsymbol{x}+tu_{\boldsymbol{x}})\,dt = \eta_2(\boldsymbol{x})-\eta_1(\boldsymbol{x}). \tag{76}$$

We can now apply Lemma B.4 to see

$$\int_{\beta_\tau^\delta} h(\boldsymbol{x})d\boldsymbol{x} = \int_{\beta_\tau} H(\boldsymbol{x})\,d\mathcal{H}(\boldsymbol{x}) + O(\sup_{\boldsymbol{x}\in\beta_\tau}\eta_2(\boldsymbol{x})-\eta_1(\boldsymbol{x}))^2$$

as $\sup_{\boldsymbol{x}\in\beta_\tau}\eta_2(\boldsymbol{x})-\eta_1(\boldsymbol{x})\to 0$ and further we have

$$\int_{\beta_\tau^\delta} h(\boldsymbol{x})d\boldsymbol{x} = \int_{\beta_\tau} \frac{\tilde{f}(\boldsymbol{x})-f_0(\boldsymbol{x})}{\|\nabla f_0(\boldsymbol{x})\|}\,d\mathcal{H}(\boldsymbol{x}) + O(\|g\|_\infty^2) + O(\|g\|_\infty\|\nabla g\|_\infty)$$

as $\|g\|_\infty^2+\|g\|_\infty\|\nabla g\|_\infty\to 0$, by (76) and (74) (and because $\sup_{\boldsymbol{x}\in\beta_\tau}\eta_2(\boldsymbol{x})-\eta_1(\boldsymbol{x})=O(\|g\|_\infty)$ and the term on the right of (74) is $O(\|g\|_\infty\|\nabla g\|_\infty)$). □

Proof of Lemma A.4. Let $\boldsymbol{y}\in\{\boldsymbol{x}\in\mathbb{R}^d:\tilde{f}(\boldsymbol{x})=\tilde{f}_\tau\}$ be such that $\boldsymbol{y}=\boldsymbol{x}+\eta u_{\boldsymbol{x}}$ for some $\boldsymbol{x}\in\beta_\tau$. Then

$$\tilde{f}(\boldsymbol{y}) = \tilde{f}(\boldsymbol{x}) + \eta\nabla\tilde{f}(\boldsymbol{x}+s\eta u_{\boldsymbol{x}})'u_{\boldsymbol{x}},$$

where $s\in[0,1]$ depends on $\boldsymbol{x}$. Then subtracting $f_0(\boldsymbol{x})$ on both sides yields

$$\tilde{f}_\tau - f_{\tau,0} = \tilde{f}(\boldsymbol{x}) - f_0(\boldsymbol{x}) + \eta\nabla\tilde{f}(\boldsymbol{x}+s\eta u_{\boldsymbol{x}})'u_{\boldsymbol{x}},$$

so

$$\eta = \frac{\tilde{f}_\tau - f_{\tau,0} - g(\boldsymbol{x})}{\nabla f(\boldsymbol{x}+s\eta u_{\boldsymbol{x}})'u_{\boldsymbol{x}}},$$

and by Lemma A.1, $\tilde{f}_\tau - f_{\tau,0} - g(\boldsymbol{x}) = O(\|g\|_\infty)$. We also know $\nabla f(\boldsymbol{x}+s\eta u_{\boldsymbol{x}})'u_{\boldsymbol{x}}$ is bounded away from zero as $\|g\|_\infty^2+\|g\|_\infty\|\nabla g\|_\infty\to 0$. Then

$$\begin{aligned}
&\int_{\mathbb{R}^d} g(\boldsymbol{x})\left(\mathbb{1}_{\{\tilde{f}(\boldsymbol{x})\ge\tilde{f}_\tau\}} - \mathbb{1}_{\{f(\boldsymbol{x})\ge f_\tau\}}\right)d\boldsymbol{x}\\
&\quad\le \|g\|_\infty\int_{\mathbb{R}^d}\left|\mathbb{1}_{\{\tilde{f}(\boldsymbol{x})\ge\tilde{f}_\tau\}} - \mathbb{1}_{\{f(\boldsymbol{x})\ge f_\tau\}}\right|d\boldsymbol{x}\\
&\quad= O(\|g\|_\infty^2).
\end{aligned}$$

□

Proof of Lemma A.5. It is well known (e.g., Wand and Jones (1995)) that

$$\mathbb{E}\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) = f_0(\boldsymbol{x}) + \frac{1}{2}\mu_2(K)\,\mathrm{tr}\{\boldsymbol{H}\nabla^2 f_0(\boldsymbol{x})\} + o\{\mathrm{tr}(\boldsymbol{H})\}.$$

This statement and all asymptotic statements in this proof are as $n\to\infty$ (implying $\boldsymbol{H}\to 0$). Now we show

$$\int_{\beta_\tau}\frac{\mathbb{E}\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x})-f_0(\boldsymbol{x})}{\|\nabla f_0(\boldsymbol{x})\|}\,d\mathcal{H}(\boldsymbol{x}) + \frac{1}{f_{\tau,0}}\int_{\mathcal{L}_\tau}\mathbb{E}\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x})-f_0(\boldsymbol{x})\,d\boldsymbol{x}$$

$$= V_1(\boldsymbol{H}) + V_2(\boldsymbol{H}) + o\{\mathrm{tr}(\boldsymbol{H})\}.$$

For fixed $\boldsymbol{x} \in \beta_\tau$, by change of variable and a Taylor expansion, we have

$$\begin{aligned} &\mathbb{E}\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) - f_0(\boldsymbol{x}) - \frac{1}{2}\mu_2(K)\,\mathrm{tr}\{\boldsymbol{H}\nabla^2 f_0(\boldsymbol{x})\} \qquad (77)\\ &\quad \le \frac{1}{2}\int_{\mathbb{R}^d} K(\boldsymbol{z})(\boldsymbol{H}^{1/2}\boldsymbol{z})^T \left|\nabla^2 f_0(\boldsymbol{x}_{\boldsymbol{z}}) - \nabla^2 f_0(\boldsymbol{x})\right| (\boldsymbol{H}^{1/2}\boldsymbol{z})\, d\boldsymbol{z}, \end{aligned}$$

where $\boldsymbol{x}_{\boldsymbol{z}} = \boldsymbol{x} - s_{\boldsymbol{z}}\boldsymbol{H}^{1/2}\boldsymbol{z}$ for some $s_{\boldsymbol{z}} \in (0,1)$ depending on $\boldsymbol{z}$. Now let $M(\boldsymbol{x},\boldsymbol{z}) = \max\left\{\left|\nabla^2 f_0(\boldsymbol{x}_{\boldsymbol{z}}) - \nabla^2 f_0(\boldsymbol{x})\right|\right\}_{i,j}$ which also implicitly depends on $\boldsymbol{H}$ and is uniformly bounded since $\nabla^2 f_0$ is uniformly bounded. Then (77) is bounded above by $\frac{1}{2}\,\mathrm{tr}\left(\boldsymbol{H}\int_{\mathbb{R}^d} M(\boldsymbol{x},\boldsymbol{z})K(\boldsymbol{Z})\boldsymbol{z}\boldsymbol{z}^T\,d\boldsymbol{z}\right)$. Then

$$\begin{aligned} &\int_{\beta_\tau} \frac{\mathbb{E}\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) - f_0(\boldsymbol{x}) - \frac{1}{2}\mu_2(K)f_0(\boldsymbol{x})}{\|\nabla f_0(\boldsymbol{x})\|}\, d\mathcal{H}(\boldsymbol{x}) \qquad (78)\\ &\quad \le \int_{\beta_\tau} \frac{1}{\|\nabla f_0(\boldsymbol{x})\|}\frac{1}{2}\,\mathrm{tr}\left(\boldsymbol{H}\int_{\mathbb{R}^d} M(\boldsymbol{x},\boldsymbol{z})K(\boldsymbol{z})\boldsymbol{z}\boldsymbol{z}^T\,d\boldsymbol{z}\right) d\mathcal{H}(\boldsymbol{x}) \qquad (79) \end{aligned}$$

which equals

$$\frac{1}{2}\,\mathrm{tr}\left(\boldsymbol{H}\int_{\beta_\tau}\frac{1}{\|\nabla f_0(\boldsymbol{x})\|}\int_{\mathbb{R}^d} M(\boldsymbol{x},\boldsymbol{z})K(\boldsymbol{z})\boldsymbol{z}\boldsymbol{z}^T\,d\boldsymbol{z}\,d\mathcal{H}(\boldsymbol{x})\right).$$

Applying the Dominated Convergence theorem to both the outer integral and the inner integral yields

$$\int_{\beta_\tau}\frac{1}{\|\nabla f_0(\boldsymbol{x})\|}\int_{\mathbb{R}^d} M(\boldsymbol{x},\boldsymbol{z})K(\boldsymbol{z})\boldsymbol{z}\boldsymbol{z}^T\,d\boldsymbol{z}\,d\mathcal{H}(\boldsymbol{x}) \to 0,$$

and thus (78) equals

$$\begin{aligned} \int_{\beta_\tau} \frac{\mathbb{E}\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) - f_0(\boldsymbol{x}) - \frac{1}{2}\mu_2(K)f_0(\boldsymbol{x})}{\|\nabla f_0(\boldsymbol{x})\|}\, d\mathcal{H}(\boldsymbol{x}) &= \int_{\beta_\tau}\frac{\mathbb{E}\widehat{f}_{n,\boldsymbol{H}} - f_0}{\|\nabla f_0\|}\, d\mathcal{H} - V1(\boldsymbol{H})\\ &= o\left\{\mathrm{tr}(\boldsymbol{H})\right\}. \end{aligned}$$

With the same argument, we can show

$$\frac{1}{f_{\tau,0}}\int_{\mathcal{L}_\tau}\mathbb{E}\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) - f_0(\boldsymbol{x})\,d\boldsymbol{x} - V_2(\boldsymbol{H}) = o\left\{\mathrm{tr}(\boldsymbol{H})\right\}.$$

In order to finish the proof it is sufficient to show that for any $\eta > 0$,

$$\mathbb{E}\left|\widehat{f}_{\tau,n} - f_{\tau,0} - w_0\left\{V_1(\boldsymbol{H}) + V_2(\boldsymbol{H})\right\}\right| \mathbb{1}_{\{\|\widehat{f}_{n,\boldsymbol{H}} - f_0\|_\infty + \|\nabla\widehat{f}_{n,\boldsymbol{H}} - \nabla f_0\|_\infty > \eta\}} \qquad (80)$$

is $o\{\mathrm{tr}(\boldsymbol{H})\}$. It can be show that $\widehat{f}_{\tau,n} = O(1)$. And we have

$$\begin{aligned} &P(\|\widehat{f}_{n,\boldsymbol{H}} - f_0\|_\infty + \|\nabla\widehat{f}_{n,\boldsymbol{H}} - \nabla f_0\|_\infty > \eta)\\ &\quad \le P(\|\widehat{f}_{n,\boldsymbol{H}} - f_0\|_\infty > \eta/2) + P(\|\nabla\widehat{f}_{n,\boldsymbol{H}} - \nabla f_0\|_\infty > \eta/2) = o(n^{-1}). \end{aligned}$$

Then by the Cauchy-Schwarz inequality (80) is $o\{\mathrm{tr}(\boldsymbol{H})\}$. □

Proof of Lemma A.6. First, we show

$$\operatorname{Var}\left\{\int_{\mathcal{L}_\tau} \widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) - f_0(\boldsymbol{x})\,d\boldsymbol{x}\right\} = O(n^{-1}). \tag{81}$$

We write the left side of (81) as

$$\begin{aligned} n^{-1}\operatorname{Var}\left\{\int_{\mathcal{L}_\tau} K_{\boldsymbol{H}}(\boldsymbol{x}-\boldsymbol{X}_i)\,d\boldsymbol{x}\right\} &= n^{-1}\mathbb{E}\left\{\int_{\mathcal{L}_\tau} K_{\boldsymbol{H}}(\boldsymbol{x}-\boldsymbol{X}_i)\,d\boldsymbol{x}\right\}^2 \\ &\quad - n^{-1}\left[\mathbb{E}\left\{\int_{\mathcal{L}_\tau} K_{\boldsymbol{H}}(\boldsymbol{x}-\boldsymbol{X}_i)\,d\boldsymbol{x}\right\}\right]^2. \end{aligned} \tag{82}$$

We first consider the first term on the right side of (82). If $\boldsymbol{y}$ is an interior point of $\mathcal{L}_\tau$, there exists $r>0$ such that $B(\boldsymbol{y},r)\subset\mathcal{L}_\tau$. Then we have

$$\begin{aligned} \int_{\mathcal{L}_\tau} |\boldsymbol{H}|^{-1/2}K(\boldsymbol{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{y}))\,d\boldsymbol{x} &\geq \int_{B(\boldsymbol{y},r)} |\boldsymbol{H}|^{-1/2}K(\boldsymbol{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{y}))\,d\boldsymbol{x} \\ &= \int \mathbb{1}_{\{\|\boldsymbol{H}^{1/2}\boldsymbol{z}\|<r\}}K(\boldsymbol{z})\,d\boldsymbol{z}, \end{aligned}$$

and $\mathbb{1}_{\{\|\boldsymbol{H}^{1/2}\boldsymbol{z}\|<r\}}\to 1$ as $\boldsymbol{H}\to 0$ for every $\boldsymbol{z}$; thus by the Dominated Convergence Theorem, $\int_{\mathcal{L}_\tau}|\boldsymbol{H}|^{-1/2}K(\boldsymbol{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{y}))\,d\boldsymbol{x}\to 1$ as $\boldsymbol{H}\to 0$. Similarly, if $\boldsymbol{y}$ is an exterior point of $\{\boldsymbol{x}|f_0(\boldsymbol{x})\geq f_{\tau,0}\}$, that is, there exists $r>0$ such that $B(\boldsymbol{y},r)\cap\mathcal{L}_\tau=\emptyset$. Then

$$\begin{aligned} \int_{\mathcal{L}_\tau} K_{\boldsymbol{H}}(\boldsymbol{x}-\boldsymbol{y})\,d\boldsymbol{x} &\leq 1 - \int_{B(\boldsymbol{y},r)} |\boldsymbol{H}|^{-1/2}K(\boldsymbol{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{y}))\,d\boldsymbol{x} \\ &= 1 - \int \mathbb{1}_{\{\|\boldsymbol{H}^{1/2}\boldsymbol{z}\|<r\}}K(\boldsymbol{z})\,d\boldsymbol{z} \to 0 \end{aligned}$$

as $\boldsymbol{H}\to 0$. And by Assumption D1b, $P(f_0(\boldsymbol{x})=f_{\tau,0})=0$. So we have that almost surely $\left(\int_{\mathcal{L}_\tau} K_{\boldsymbol{H}}(\boldsymbol{x}-\boldsymbol{X}_i)\,d\boldsymbol{x}\right)^p\to\mathbb{1}_{\mathcal{L}_\tau}$, as $n\to\infty$, for $p=1,2$. Applying the Dominated Convergence Theorem to the two expectations on the right of (82) yields

$$n\operatorname{Var}\left\{\int_{\mathcal{L}_\tau} \widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) - f_0(\boldsymbol{x})\,d\boldsymbol{x}\right\} \to \mathbb{P}(f_0(\boldsymbol{X}_i)\geq f_{\tau,0})(1-\mathbb{P}(f_0(\boldsymbol{X}_i)\geq f_{\tau,0})),$$

as $n\to\infty$, which shows $\operatorname{Var}\left\{\int_{\mathcal{L}_\tau} \widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) - f_0(\boldsymbol{x})\,d\boldsymbol{x}\right\} = O(n^{-1})$.

Next, we show

$$\operatorname{Var}\left\{\int_{\beta_\tau} \frac{\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) - f_0(\boldsymbol{x})}{\|\nabla f_0(\boldsymbol{x})\|}\,d\mathcal{H}(\boldsymbol{x})\right\} = o\left(\frac{1}{n|\boldsymbol{H}|^{1/2}}\right).$$

The left side of the previous display equals

$$n^{-1}\operatorname{Var}\int_{\beta_\tau}\frac{K_{\boldsymbol{H}}(\boldsymbol{x}-\boldsymbol{X}_i)}{\|\nabla f_0(\boldsymbol{x})\|}\,d\mathcal{H}(\boldsymbol{x})$$
$$=\frac{1}{n}\mathbb{E}\left[\left\{\int_{\beta_\tau}\frac{K_{\boldsymbol{H}}(\boldsymbol{x}-\boldsymbol{X}_i)}{\|\nabla f_0(\boldsymbol{x})\|}\,d\mathcal{H}(\boldsymbol{x})\right\}^2\right]$$
$$-\frac{1}{n}\left[\mathbb{E}\left\{\int_{\beta_\tau}\frac{K_{\boldsymbol{H}}(\boldsymbol{x}-\boldsymbol{X}_i)}{\|\nabla f_0(\boldsymbol{x})\|}\,d\mathcal{H}(\boldsymbol{x})\right\}\right]^2,$$

and $\left\{\int_{\beta_\tau}\frac{|\boldsymbol{H}|^{-1/2}K(\boldsymbol{H}^{1/2}(\boldsymbol{x}-\boldsymbol{X}_i))}{\|\nabla f_0(\boldsymbol{x})\|}\,d\mathcal{H}(\boldsymbol{x})\right\}^2$ can be written as

$$\int_{\beta_\tau}\frac{|\boldsymbol{H}|^{-1/2}K(\boldsymbol{H}^{1/2}(\boldsymbol{x}-\boldsymbol{X}_i))}{\|\nabla f_0(\boldsymbol{x})\|}\,d\mathcal{H}(\boldsymbol{x})\int_{\beta_\tau}\frac{|\boldsymbol{H}|^{-1/2}K(\boldsymbol{H}^{1/2}(\boldsymbol{y}-\boldsymbol{X}_i))}{\|\nabla f_0(\boldsymbol{y})\|}\,d\mathcal{H}(\boldsymbol{y}).$$

By taking the expectation over $\boldsymbol{X}_i$ and reordering the integrals by Tonelli's theorem, we can then see that $n^{-1}\mathbb{E}\left\{\int_{\beta_\tau}\frac{K_{\boldsymbol{H}}(\boldsymbol{x}-\boldsymbol{X}_i)}{\|\nabla f_0(\boldsymbol{x})\|}\,d\mathcal{H}(\boldsymbol{x})\right\}^2$ equals

$$\frac{1}{n|\boldsymbol{H}|}\int_{\beta_\tau}\int_{\beta_\tau}\frac{1}{\|\nabla f_0(\boldsymbol{x})\|}\frac{1}{\|\nabla f_0(\boldsymbol{y})\|}\times$$
$$\int_{\mathbb{R}^d}K(\boldsymbol{H}^{-1/2}(\boldsymbol{x}-a))K(\boldsymbol{H}^{-1/2}(\boldsymbol{y}-a))f_0(\boldsymbol{a})\,d\boldsymbol{a}\,d\mathcal{H}(\boldsymbol{x})d\mathcal{H}(\boldsymbol{y}). \tag{83}$$

And

$$\int_{\mathbb{R}^d}K(\boldsymbol{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{a}))K(\boldsymbol{H}^{-1/2}(\boldsymbol{y}-\boldsymbol{a}))f_0(\boldsymbol{a})d\boldsymbol{a}$$
$$=\int_{\mathbb{R}^d}K(\boldsymbol{z})K(\boldsymbol{z}+\boldsymbol{H}^{-1/2}(\boldsymbol{y}-\boldsymbol{x}))f_0(\boldsymbol{x}-\boldsymbol{H}^{1/2}\boldsymbol{z})|\boldsymbol{H}|^{1/2}d\boldsymbol{z}$$

by the change of variables $\boldsymbol{z}=\boldsymbol{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{a})$. And by first-order Taylor expansion, the previous display equals

$$|\boldsymbol{H}|^{1/2}\int_{\mathbb{R}^d}K(\boldsymbol{z})K(\boldsymbol{z}+\boldsymbol{H}^{-1/2}(\boldsymbol{y}-\boldsymbol{x}))\left\{f_0(\boldsymbol{x})-\boldsymbol{H}^{1/2}\boldsymbol{z}\nabla f_0(\boldsymbol{x}-s\boldsymbol{H}^{1/2}\boldsymbol{z})\right\}d\boldsymbol{z}$$

where $s\in[0,1]$ depends on $\boldsymbol{z}$. Since by Assumption D1b, $\nabla f_0(\boldsymbol{x})$ is bounded, we can express (83) as

$$\frac{f_{\tau,0}}{n|\boldsymbol{H}|^{1/2}}\int_{\beta_\tau}\int_{\beta_\tau}\frac{1}{\|\nabla f_0(\boldsymbol{x})\|\|\nabla f_0(\boldsymbol{y})\|}\times$$
$$\int_{\mathbb{R}^d}K(\boldsymbol{z})K(\boldsymbol{z}+\boldsymbol{H}^{-1/2}(\boldsymbol{y}-\boldsymbol{x}))d\boldsymbol{z}d\mathcal{H}(\boldsymbol{y})d\mathcal{H}(\boldsymbol{x}) \tag{84}$$
$$+o\left(n^{-1}|\boldsymbol{H}|^{-1/2}\right).$$

Note that if $\boldsymbol{x} \neq \boldsymbol{y}$, then

$$\int_{\mathbb{R}^d} K(\boldsymbol{z})K(\boldsymbol{z} + \boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{x}))d\boldsymbol{z} \to 0,$$

as $\boldsymbol{H} \to 0$ by the Dominated Convergence Theorem. For fixed $\boldsymbol{x} = \boldsymbol{y}$, we have $\int_{\mathbb{R}^d} K(\boldsymbol{z})K(\boldsymbol{z} + \boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{x}))d\boldsymbol{z} = R(K)$, so

$$\int_{\mathbb{R}^d} K(\boldsymbol{z})K(\boldsymbol{z} + \boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{x}))d\boldsymbol{z} \to R(K)\mathbb{1}_{\{\boldsymbol{x}=\boldsymbol{y}\}},$$

as $\boldsymbol{H} \to 0$. Then applying the Dominated Convergence Theorem shows that the first summand in (84) converges to

$$\frac{1}{n|\boldsymbol{H}|^{1/2}} \int_{\beta_\tau} f_0(\boldsymbol{x}) \int_{\beta_\tau} \frac{1}{\|\nabla f_0(\boldsymbol{x})\|\|\nabla f_0(\boldsymbol{y})\|} R(K)\mathbb{1}_{\{\boldsymbol{x}=\boldsymbol{y}\}} d\mathcal{H}(\boldsymbol{y})d\mathcal{H}(\boldsymbol{x}) = 0.$$

So we proved

$$\operatorname{Var}\left\{ \int_{\beta_\tau} \frac{\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) - f_0(\boldsymbol{x})}{\|\nabla f_0(\boldsymbol{x})\|} d\mathcal{H}(\boldsymbol{x}) \right\} = o\left(\frac{1}{n|\boldsymbol{H}|^{1/2}} \right). \tag{85}$$

To complete the proof, it remains to show that for any $\eta > 0$,

$$\mathbb{E}\left\{ \widehat{f}_{\tau,n} - \mathbb{E}(\widehat{f}_{\tau,n}) \right\}^2 \mathbb{1}_{\{\|\widehat{f}_{n,\boldsymbol{H}} - f_0\|_\infty + \|\nabla \widehat{f}_{n,\boldsymbol{H}} - \nabla f_0\|_\infty > \eta\}} = o\left(\frac{1}{n|\boldsymbol{H}|^{1/2}} \right),$$

which follows the same steps we used at the end of the proof of Lemma A.5. The proof is then complete by (19). □

The reader may be surprised by the conclusion of (85), since $\operatorname{Var} \widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) = O(n^{-1}|\boldsymbol{H}|^{-1/2})$; for intuition, it may help to recall that $\int_{\mathbb{R}^d} \widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x})d\boldsymbol{x} = 1$, so has variance 0.

Proof of Lemma A.7. Note by Theorem B.1, we have that $\mathbb{E}\nabla \widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x})$ converges to $\nabla \widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x})$ uniformly in $\boldsymbol{x} \in \mathbb{R}^d$. By Durrett (2010, Theorem A.5.1), we have $\nabla \mathbb{E}\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) = \mathbb{E}\nabla \widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x})$, thus we also have $\nabla \mathbb{E}\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x})$ also converges to $\nabla f_0(\boldsymbol{x})$ uniformly in $\boldsymbol{x}$.

Now, we show $\mathbb{E}\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x} + tu_{\boldsymbol{x}})$ is strictly monotone for $t \in [-\delta_n, \delta_n]$ when n is sufficiently. From our assumption, ∇f_0 is Lipschitz. So when n large enough and δ_n small enough, for each $t \in [-\delta_n, \delta_n]$ there exists ϵ_t such that $\nabla \mathbb{E}\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x} + tu_{\boldsymbol{x}}) = \nabla f_0(\boldsymbol{x}) + \epsilon_t$ and $\|\epsilon_t\| < \frac{l}{2}$, where $l = \inf_{\boldsymbol{x} \in \beta_\tau} \|\nabla f_0(\boldsymbol{x})\|$ and we know $l > 0$ from Assumption D1b. Then

$$\begin{aligned} \frac{d\mathbb{E}\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x} + tu_{\boldsymbol{x}})}{dt} &= \nabla \mathbb{E}\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x} + tu_{\boldsymbol{x}})u_{\boldsymbol{x}} \\ &= (\nabla f_0(\boldsymbol{x}) + \epsilon_t)\frac{-\nabla f_0(\boldsymbol{x})}{\|\nabla f_0(\boldsymbol{x})\|} \end{aligned}$$

$$= -\|\nabla f_0(\boldsymbol{x})\| - \frac{\nabla f_0(\boldsymbol{x})'\epsilon_t}{\|\nabla f_0(\boldsymbol{x})\|}$$
$$< -\frac{l}{2},$$

for all $t \in [-\delta_n, \delta_n]$ by the Cauchy-Schwarz inequality. Moreover, from Lemma A.5 we have

$$\mathbb{E}\widehat{f}_{\tau,n} = f_{\tau,0} + \left\{\int_{\beta_\tau} \frac{1}{\|\nabla f_0\|}\, d\mathcal{H}\right\}^{-1} \left\{\int_{\beta_\tau} \frac{\mu_2(K)\,\mathrm{tr}\left(\boldsymbol{H}\nabla^2 f_0\right)}{2\|\nabla f_0\|}\, d\mathcal{H} + \int_{\mathcal{L}_\tau} \frac{1}{2}\mu_2(K)\,\mathrm{tr}\left(\boldsymbol{H}\nabla^2 f_0\right)\, d\lambda\right\} + o\{\mathrm{tr}(\boldsymbol{H})\},$$

and we also know

$$\begin{aligned}
&\mathbb{E}\widehat{f}_{n,\boldsymbol{H}}\left(\boldsymbol{x} + \frac{t}{\sqrt{n|\boldsymbol{H}|^{1/2}}}u_{\boldsymbol{x}}\right) \\
&\quad = f_0\left(\boldsymbol{x} + \frac{t}{\sqrt{n|\boldsymbol{H}|^{1/2}}}u_{\boldsymbol{x}}\right) \\
&\qquad + \frac{1}{2}\int \boldsymbol{z}'\boldsymbol{H}^{1/2}\nabla^2 f_0\left(\boldsymbol{x} + \frac{t}{\sqrt{n|\boldsymbol{H}|^{1/2}}}u_{\boldsymbol{x}} - s_{\boldsymbol{z}}\boldsymbol{H}^{1/2}\boldsymbol{z}\right)\boldsymbol{H}^{1/2}\boldsymbol{z}K(\boldsymbol{z})\, d\boldsymbol{z} \\
&\quad = f_0(\boldsymbol{x}) + \nabla f_0\left(\boldsymbol{x} + \frac{w_{\boldsymbol{x}}t}{\sqrt{n|\boldsymbol{H}|^{1/2}}}u_{\boldsymbol{x}}\right)'\frac{t}{\sqrt{n|\boldsymbol{H}|^{1/2}}}u_{\boldsymbol{x}} \\
&\qquad + \frac{1}{2}\int \boldsymbol{z}'\boldsymbol{H}^{1/2}\nabla^2 f_0\left(\boldsymbol{x} + \frac{t}{\sqrt{n|\boldsymbol{H}|^{1/2}}}u_{\boldsymbol{x}} - s_{\boldsymbol{z}}\boldsymbol{H}^{1/2}\boldsymbol{z}\right)\boldsymbol{H}^{1/2}\boldsymbol{z}K(\boldsymbol{z})\, d\boldsymbol{z},
\end{aligned}$$

and $\frac{1}{2}\int \boldsymbol{z}^T\boldsymbol{H}^{1/2}\nabla^2 f_0\left(\boldsymbol{x} + \frac{t}{\sqrt{n|\boldsymbol{H}|^{1/2}}}u_{\boldsymbol{x}} - s_{\boldsymbol{z}}\boldsymbol{H}^{1/2}\boldsymbol{z}\right)\boldsymbol{H}^{1/2}\boldsymbol{z}K(\boldsymbol{z})\, d\boldsymbol{z}$ is $O(\mathrm{tr}(\boldsymbol{H}))$ uniformly in $\boldsymbol{x}$. Then

$$\begin{aligned}
&\frac{t^*_{\boldsymbol{x}}}{\sqrt{n|\boldsymbol{H}|^{1/2}}}\nabla f_0\left(\boldsymbol{x} + \frac{w_{\boldsymbol{x}}t}{\sqrt{n|\boldsymbol{H}|^{1/2}}}u_{\boldsymbol{x}}\right)' u_{\boldsymbol{x}} \\
&\quad = \left[w_0\left\{\int_{\beta_\tau}\frac{D_1(\boldsymbol{x},\boldsymbol{H})}{\|\nabla f_0\|}\, d\mathcal{H} + \int_{\mathcal{L}_\tau} D_1(\boldsymbol{x},\boldsymbol{H})\, d\lambda\right\}\right. \\
&\qquad \left. -\frac{1}{2}\int \boldsymbol{z}^T\boldsymbol{H}^{1/2}\nabla^2 f_0\left(\boldsymbol{x} + \frac{t}{\sqrt{n|\boldsymbol{H}|^{1/2}}}u_{\boldsymbol{x}} - s_{\boldsymbol{z}}\boldsymbol{H}^{1/2}\boldsymbol{z}\right)\boldsymbol{H}^{1/2}\boldsymbol{z}K(\boldsymbol{z})\, d\boldsymbol{z}\right] \\
&\qquad (1 + o(1)). \qquad (86)
\end{aligned}$$

Since ∇f_0 is Lipschitz, when n is large enough $\nabla f_0\left(\boldsymbol{x} + \frac{t}{\sqrt{n|\boldsymbol{H}|^{1/2}}}u_{\boldsymbol{x}}\right)' u_{\boldsymbol{x}} < -\frac{l}{2}$ for all $\boldsymbol{x} \in \beta_\tau$ and all $t \in \left[-\sqrt{n|\boldsymbol{H}|^{1/2}}\delta_n, \sqrt{n|\boldsymbol{H}|^{1/2}}\delta_n\right]$. To prove the last line

of the lemma, since

$$\frac{d\mathbb{E}\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}+tu_{\boldsymbol{x}})}{dt} < -\frac{l}{2},$$

for all $t \in [-\delta_n, \delta_n]$ and all $\boldsymbol{x} \in \beta_\tau$,

$$\frac{d\mathbb{E}\widehat{f}_{n,\boldsymbol{H}}\left(\boldsymbol{x}+\frac{t}{\sqrt{n|\boldsymbol{H}|^{1/2}}}u_{\boldsymbol{x}}\right)}{dt} \leq -\frac{l}{2\sqrt{n|\boldsymbol{H}|^{1/2}}},$$

for all $\boldsymbol{x} \in \beta_\tau$ and all $t \in [-\sqrt{n|\boldsymbol{H}|^{1/2}}\delta_n, \sqrt{n|\boldsymbol{H}|^{1/2}}\delta_n]$. Then by first order Taylor expansion, it is easy to get when $t \in I^n_{\boldsymbol{x}}$,

$$\left|\mathbb{E}\left\{\widehat{f}_{n,\boldsymbol{H}}\left(\boldsymbol{x}+\frac{t}{\sqrt{n|\boldsymbol{H}|^{1/2}}}u_{\boldsymbol{x}}\right)-\widehat{f}_{\tau,n}\right\}\right| \geq \frac{l}{2\sqrt{n|\boldsymbol{H}|^{1/2}}}\left|t-t^*_{\boldsymbol{x}}\right|,$$

when n is large enough.

And then when $t \leq 0$,

$$\begin{aligned}
&\left|P\left\{\widehat{f}_{n,\boldsymbol{H}}\left(\boldsymbol{x}^t\right) < \widehat{f}_{\tau,n}\right\} - \mathbb{1}_{\{t>0\}}\right| \\
&\quad = P\left\{\widehat{f}_{n,\boldsymbol{H}}\left(\boldsymbol{x}^t\right) < \widehat{f}_{\tau,n}\right\} \\
&\quad \leq P\left\{\widehat{f}_{\tau,n} - \widehat{f}_{n,\boldsymbol{H}}\left(\boldsymbol{x}^t\right) + \mathbb{E}\left(\widehat{f}_{n,\boldsymbol{H}}\left(\boldsymbol{x}^t\right) - \widehat{f}_{\tau,n}\right) \geq \frac{l}{2\sqrt{n|\boldsymbol{H}|^{1/2}}}|t-t^*_{\boldsymbol{x}}|\right\} \\
&\quad \leq P\left\{\left|\widehat{f}_{n,\boldsymbol{H}}\left(\boldsymbol{x}^t\right) - \mathbb{E}\widehat{f}_{n,\boldsymbol{H}}\left(\boldsymbol{x}^t\right)\right| \geq \frac{l}{4\sqrt{n|\boldsymbol{H}|^{1/2}}}|t-t^*_{\boldsymbol{x}}|\right\} \\
&\qquad + P\left\{\left|\widehat{f}_{\tau,n} - \mathbb{E}\widehat{f}_{\tau,n}\right| \geq \frac{l}{4\sqrt{n|\boldsymbol{H}|^{1/2}}}|t-t^*_{\boldsymbol{x}}|\right\}
\end{aligned} \tag{87}$$

and we can show the same bound for $t > 0$. Since

$$\operatorname{Var}\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}) = n^{-1}\left[|\boldsymbol{H}|^{-1/2}\int K(\boldsymbol{z})f\left(\boldsymbol{x}-\boldsymbol{H}^{1/2}\boldsymbol{z}\right)d\boldsymbol{z} - \left\{\int K(\boldsymbol{z})f\left(\boldsymbol{x}-\boldsymbol{H}^{1/2}\boldsymbol{z}\right)d\boldsymbol{z}\right\}^2\right],$$

$\operatorname{Var}\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x})$ is uniformly $O(n^{-1}|\boldsymbol{H}|^{-1/2})$. And we know $\operatorname{Var}\widehat{f}_{\tau,n}$ is also of order $o(n^{-1}|\boldsymbol{H}|^{-1/2})$ from Lemma A.6. Then there exists $C_2 > 0$ such that (87) can be further bounded as

$$\begin{aligned}
&\left|P\left\{\widehat{f}_{n,\boldsymbol{H}}\left(\boldsymbol{x}^t\right) < \widehat{f}_{\tau,n}\right\} - \mathbb{1}_{\{t>0\}}\right| \\
&\qquad \leq P\left\{\left|\frac{\widehat{f}_{n,\boldsymbol{H}}\left(\boldsymbol{x}^t\right) - \mathbb{E}\widehat{f}_{n,\boldsymbol{H}}\left(\boldsymbol{x}^t\right)}{\operatorname{Var}\widehat{f}_{n,\boldsymbol{H}}\left(\boldsymbol{x}^t\right)}\right| \geq C_2|t-t^*_{\boldsymbol{x}}|\right\}
\end{aligned}$$

$$+ P\left\{\left|\frac{\widehat{f}_{\tau,n} - \mathbb{E}\widehat{f}_{\tau,n}}{\operatorname{Var}\widehat{f}_{\tau,n}}\right| \geq C_2|t - t^*_{\boldsymbol{x}}|\right\}$$
$$\leq \frac{2C_2}{(t - t^*_{\boldsymbol{x}})^2}$$

for all $\boldsymbol{x} \in \beta_\tau, t \in \bigcup_{\boldsymbol{x}\in\beta_\tau} I^n_{\boldsymbol{x}}$ by Chebyshev inequality. And note that $(t-t^*_{\boldsymbol{x}})^2 \geq t_n^2$ for all $t \in \bigcup_{\boldsymbol{x}\in\beta_\tau} I^n_{\boldsymbol{x}}$. So $\mathbb{1}_{I^n_{\boldsymbol{x}}} \cdot |P\{\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}^t) < \widehat{f}_{\tau,n}\} - \mathbb{1}_{\{t>0\}}|$ converges to 0 uniformly in t and is dominated by $\max\{1/(t-t^*_{\boldsymbol{x}})^2, 1\}$ which is a integrable function over $\mathbb{R}$. Then by Dominate Convergence Theorem, we have

$$\int_{I^n_{\boldsymbol{x}}} \left|P\left\{\widehat{f}_{n,\boldsymbol{H}}\left(\boldsymbol{x}^t\right) < \widehat{f}_{\tau,n}\right\} - \mathbb{1}_{\{t>0\}}\right| dt \to 0,$$

as $n \to \infty$. Also note that $\int_{I^n_{\boldsymbol{x}}} |P\{\widehat{f}_{n,\boldsymbol{H}}(\boldsymbol{x}^t) < \widehat{f}_{\tau,n}\} - \mathbb{1}_{\{t>0\}}|\, dt \leq \int \max\{\frac{1}{t^2}, 1\}\, dt$ for all $\boldsymbol{x} \in \beta_\tau$. So we have

$$\int_{\beta_\tau}\int_{I^n_{\boldsymbol{x}}} \left|P\left\{\widehat{f}_{n,\boldsymbol{H}}\left(\boldsymbol{x}^t\right) < \widehat{f}_{\tau,n}\right\} - \mathbb{1}_{\{t>0\}}\right| dt d\mathcal{H}(\boldsymbol{x}) \to 0,$$

as $n \to \infty$. □

References

Baíllo, A. (2003). Total error in a plug-in estimator of level sets. *Statist. Probab. Lett.* **65** 411–417. MR2039885

Baíllo, A., Cuesta-Albertos, J. A. and Cuevas, A. (2001). Convergence rates in nonparametric estimation of level sets. *Statist. Probab. Lett.* **53** 27–35. MR1843338

Billingsley, P. (2012). *Probability and Measure. Wiley Series in Probability and Statistics*. John Wiley & Sons, Inc., Hoboken, NJ. MR2893652

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

Bredon, G. E. (1993). *Topology and geometry. Graduate Texts in Mathematics* **139**. Springer-Verlag, New York. MR1224675

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

Cadre, B., Pelletier, B. and Pudlo, P. (2013). Estimation of density level sets with a given probability content. *Journal of Nonparametric Statistics* **25** 261–272. MR3039981

Cavalier, L. (1997). Nonparametric estimation of regression level sets. *Statistics* **29** 131–160. MR1484386

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

Chen, Y.-C. (2016). Generalized cluster trees and singular measures. *arXiv*.

Chen, Y.-C. (2017). A tutorial on kernel density estimation and recent advances. *arXiv:1704.03924v1*.

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., Febrero, M. and Fraiman, R. (2001). Cluster analysis: a further approach based on density estimation. *Comput. Statist. Data Anal.* **36** 441–459. MR1855727

De Marchi, S. and Elefante, G. (2018). Quasi-monte carlo integration on manifolds with mapped low-discrepancy points and greedy minimal riesz s-energy points. *Applied Numerical Mathematics* **127** 110–124. MR3760224

Dudley, R. M. (1999). *Uniform Central Limit Theorems* **63**. Cambridge University Press, Cambridge. MR1720712

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

Duong, T. and Hazelton, M. L. (2005). Cross-Validation Bandwidth Matrices for Multivariate Kernel Density Estimation. *Scandinavian Journal of Statistics* **32** 485–506. MR2204631

Duong, T., Koch, I. and Wand, M. P. (2009). Highest Density Difference Region Estimation with Application to Flow Cytometric Data. *Biometrical Journal* **51** 504–521. MR2750050

Durrett, R. (2010). *Probability: theory and examples*. Cambridge university press. MR2722836

Evans, L. C. and Gariepy, R. F. (2015). *Measure Theory and Fine Properties of Functions*, revised ed. *Textbooks in Mathematics*. CRC Press, Boca Raton, FL. MR3409135

Ferguson, T. S. (1996). *A course in large sample theory. Texts in Statistical Science Series*. Chapman & Hall, London. MR1699953

Folland, G. B. (1999). *Real analysis*, second ed. *Pure and Applied Mathematics*. John Wiley & Sons, Inc., New York. MR1681462

Garcia, J. N., Kutalik, Z., Cho, K.-H. and Wolkenhauer, O. (2003). Level sets and minimum volume sets of probability density functions. *International journal of approximate reasoning* **34** 25–47. MR2017778

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. En l'honneur de J. Bretagnolle, D. Dacunha-Castelle, I. Ibragimov. MR1955344

Giné, E., Koltchinskii, V. and Zinn, J. (2004). Weighted uniform consistency of kernel density estimators. *Ann. Probab.* **32** 2570–2605. MR2078551

Guillemin, V. and Pollack, A. (1974). *Differential Topology*. Prentice-Hall, Inc., Englewood Cliffs, N.J. MR0348781

Hall, P. and Marron, J. S. (1987). Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probab. Theory Related Fields* **74** 567–581. MR0876256

Hall, P., Marron, J. S. and Park, B. U. (1992). Smoothed cross-validation. *Probab. Theory Related Fields* **92** 1–20. MR1156447

Hämmerlin, G. and Hoffmann, K.-H. (1991). *Numerical mathematics. Undergraduate Texts in Mathematics*. Springer-Verlag, New York. Translated from the German by Larry Schumaker. MR1088482

Hartigan, J. A. (1975). *Clustering algorithms* **209**. Wiley New York. MR0405726

Hartigan, J. A. (1987). Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.* **82** 267–270. MR0883354

Hyndman, R. J. (1996). Computing and graphing highest density regions. *Amer. Statist.* **50** 120–126.

Jankowski, H. and Stanberry, L. (2012). Confidence regions in level set estimation. *Preprint*.

Jones, M. C., Marron, J. S. and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *J. Amer. Statist. Assoc.* **91** 401–407. MR1394097

Lichman, M. and Smyth, P. (2014). Modeling human location data with mixtures of kernel densities. In *the 20th ACM SIGKDD international conference* 35–44. ACM Press, New York, New York, USA.

Magnus, J. R. and Neudecker, H. (1999). *Matrix differential calculus with applications in statistics and econometrics. Wiley Series in Probability and Statistics*. John Wiley & Sons, Ltd., Chichester. Revised reprint of the 1988 original. MR1698873

Mammen, E. and Polonik, W. (2013). Confidence regions for level sets. *J. Multivariate Anal.* **122** 202–214. MR3189318

Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808–1829. MR1765618

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. *Ann. Appl. Probab.* **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

Park, C., Huang, J. Z. and Ding, Y. (2010). A computable plug-in estimator of minimum volume sets for novelty detection. *Oper. Res.* **58** 1469–1480. MR2560548

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. *arXiv:1707.09697*.

R Core Team, (2018). R: A language and environment for statistical computing. r foundation for statistical computing, Vienna, Austria.

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

Sain, S. R., Baggerly, K. A. and Scott, D. W. (1994a). Cross-validation of multivariate densities. *J. Amer. Statist. Assoc.* **89** 807–817. MR1294726

Sain, S. R., Baggerly, K. A. and Scott, D. W. (1994b). Cross-validation of multivariate densities. *Journal of the American Statistical Association* **89** 807–817. MR1294726

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, D. W. and Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association* **82** 1131–1146. MR0922178

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* **53** 683–690. MR1125725

Spivak, M. (1965). *Calculus on Manifolds*. W. A. Benjamin, Inc., New York-Amsterdam.

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

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes. Springer Series in Statistics*. Springer-Verlag, New York. MR1385671

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

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** 97–116. MR1280754

Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing. Monographs on Statistics and Applied Probability* **60**. Chapman and Hall, Ltd., London. MR1319818

Wasserman, L. (2016). Topological data analysis. *arXiv:1609.08227v1*.