Uniform Convergence Rates for Kernel Density Estimation

Heinrich Jiang [1]

Abstract

Kernel density estimation (KDE) is a popular nonparametric density estimation method. We (1) derive finite-sample high-probability density estimation bounds for multivariate KDE under mild density assumptions which hold uniformly in $x \in \mathbb{R}^d$ and bandwidth matrices. We apply these results to (2) mode, (3) density level set, and (4) class probability estimation and attain optimal rates up to logarithmic factors. We then (5) provide an extension of our results under the manifold hypothesis. Finally, we (6) give uniform convergence results for local intrinsic dimension estimation.

1. Introduction

KDE (Rosenblatt, 1956; Parzen, 1962) is a foundational aspect of nonparametric statistics. It is a powerful method to estimate the probability density function of a random variable. Moreover, it is simple to compute and has played a significant role in a very wide range of practical applications. Its convergence properties have been studied for a long time with most of the work dedicated to its asymptotic behavior or mean-squared risk (Tsybakov, 2008). However, there is still a surprising amount not yet fully understood about its convergence behavior. In this paper, we focus on the uniform finite-sample facet of KDE convergence theory. We handle the multivariate KDE setting in $\mathbb{R}^d$ which allows a $d \times d$ bandwidth matrix $\mathbf{H}$. This generalizes the scalar bandwidth $h > 0$ i.e. $\mathbf{H} = h^2\mathbf{I}$. Such a generalization is significant to multivariate statistics e.g. Silverman (1986); Simonoff (1996).

Our work begins by using VC-based Bernstein-type uniform convergence bounds to attain finite-sample rates for a fixed unknown density f over $\mathbb{R}^d$ (Theorem 1). These bounds hold with high-probability under general assumptions on f and the kernel i.e. we only require f to be bounded as well as decay assumptions on the kernel functions. Moreover, these bounds hold uniformly over $\mathbb{R}^d$ and bandwidth matrices $\mathbf{H}$.

We then show the versatility of our results by applying it to the related areas of KDE rates under ℓ_∞, mode estimation, density level-set estimation, and class probability estimation. We then extend our analysis to the manifold setting. Finally, we provide uniform finite-sample results for local intrinsic dimension estimation. Each of these contributions are significant on their own.

2. Contributions and Related Works

ℓ_∞ bounds for KDE: It must first be noted that bounding $|\widehat{f}_h - f|_\infty$ where $\widehat{f}_h$ is the KDE of f for scalar $h > 0$ is a more difficult problem than for example bounding the mean-squared error $\mathbb{E}_f[(\widehat{f}_h - f)^2]$. Gine & Guillon (2002) and Einmahl & Mason (2005) give asymptotic convergence results on KDE for $|\hat{f}_h - \mathbb{E}_f \hat{f}_h|_\infty$. In their work about density clustering, Rinaldo & Wasserman (2010) extends the results of the former to obtain high-probability finite-sample bounds. This is to our knowledge the strongest and most general uniform finite-sample result about KDE thus far.

We show a general bound of form $|\widehat{f}_h(x) - f(x)| \lesssim \epsilon_x + \sqrt{\log n / nh^d}$ where ϵ_x is a function of the kernel and the smoothness of f at x which holds with probability $1 - 1/n$ uniformly over $x \in \mathbb{R}^d$ and h (Theorem 1). An almost direct consequence is that if we take f to be α-Hölder continuous then under the optimal choice for $h \approx n^{-1/(2\alpha+d)}$, we have $|\hat{f}_h - f|_\infty \lesssim n^{-\alpha/(2\alpha+d)}$ with probability $1 - 1/n$ (Theorem 2). This matches the known lower bound (Tsybakov, 2008).

When comparing our finite-sample results to that of Rinaldo & Wasserman (2010), there are a few notable differences. Our results hold uniformly across bandwidths and the probability that the bounds hold are independent of the bandwidth (in fact, holds with probability $1 - 1/n$). Our results also extends to general bandwidth matrix $\mathbf{H}$.

This can be significant to analyze KDE-based procedures with *adaptive* bandwidths– i.e. when the bandwidths change depending on the region. Then the need for bounds which hold simultaneously over bandwidth choices be-

[1]Google. Correspondence to: Heinrich Jiang <heinrich.jiang@gmail.com>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

comes clear. Such an example includes adaptive or variable KDE (Terrell & Scott, 1992; Botev et al., 2010) which extends KDE to bandwidths that vary over the data space.

Thus our result for uniform finite-sample KDE bounds can be seen as a refinement to existing results.

Mode estimation Estimating the modes of a distribution has a long history e.g. Parzen (1962); Chernoff (1964); Eddy (1980); Silverman (1981); Cheng (1995); Abraham et al. (2004); Li et al. (2007); Dasgupta & Kpotufe (2014); Genovese et al. (2015); Jiang & Kpotufe (2017). The modes can be viewed as the central tendancies of a distribution and this line of work has played a significant role in areas such as clustering, image segmentation, and anomaly detection.

Much of the early work focused on the estimator $\text{argmax}_{x\in\mathbb{R}^d}\widehat{f}_h(x)$. While many useful insights have come from studying this, it is difficult to algorithmically compute. Abraham et al. (2004) turned to the simple estimator $\text{argmax}_{x\in X}\widehat{f}_h(x)$ and showed that it behaves asymptotically as $\text{argmax}_{x\in\mathbb{R}^d}\widehat{f}_h(x)$ where X is the data. In this paper, we show that this estimator is actually a rate-optimal estimator of the mode under finite samples with appropriate bandwidth choice. This would not have been possible without the appropriate bounds on KDE. This approach is similar to that of Dasgupta & Kpotufe (2014), who apply their k-NN density estimation bounds to show that the k-NN analogue of the estimator is rate-optimal.

Another approach to mode estimation that must be noted is mean-shift (Fukunaga & Hostetler, 1975; Cheng, 1995; Comaniciu & Meer, 2002; Arias-Castro et al., 2016), which is a popular clustering algorithm amongst practitioners based on performing a gradient-ascent of the KDE. Its theoretical analysis however is still far from complete; the difficulty comes from analyzing KDE's ability to estimate gradients. Here we are focused on density estimation rather than density derivative estimation so our results do not appear immediately applicable to mean-shift.

Density level-set estimation The problem of density-level set estimation has been extensively studied e.g. Carmichael et al. (1968); Hartigan (1975); Cuevas & Fraiman (1997); Tsybakov (1997); Cadre (2006); Rigollet & Vert (2009); Singh et al. (2009); Rinaldo & Wasserman (2010); Steinwart (2011); Jiang (2017). It involves estimating $\{x : f(x) \geq \lambda\}$ for some $\lambda > 0$ and density f based on samples drawn from f. This turns out to be one of the earliest and still currently most popular means of modeling clusters in the context of density-based clustering. The level-sets also influenced much of the work on hierarchical clustering (Chaudhuri & Dasgupta, 2010).

Naturally, we must use some density estimator to get a handle on λ. It turns out that in order to obtain the most general uniform recovery bounds (e.g. finite-sample Hausdorff rates (Singh et al., 2009)), one also needs similar uniform density estimation bounds. The strongest known results thus far use density estimators that are often impractical (e.g. histogram density estimator) in order to obtain these theoretical rates over a practical one such as KDE. Much of the work, especially ones using more practical density estimators have focused on bounding metrics such as symmetric set difference, which are computed as an expectation over f. This is considerably weaker than the Hausdorff metric, which imposes a uniform guarantee over each estimated point and each point in the level-set.

We show that a simple KDE-based estimator is consistent under the Hausdorff metric; moreoever when the bandwidth is appropriately chosen, it attains the minimax optimal rate established by Tsybakov (1997).

Class probability estimation Class probability estimation involves estimating the probability distribution over a set of classes for a given input. In other words, it is an approach to classification which involves first estimating the marginal density $f(Y|X)$ (where X is the observation and Y is its category) and then choosing the category with highest probability. This density-based approach to classification has been studied in many places under nonparametric assumptions. e.g. Rigollet (2007); Chaudhuri et al. (2009). However, there are still aspects about its convergence properties that haven't been fully understood. In the current work, we give uniform rates on the approximation of $f(Y|X)$. Much of the related work assume the binary classification case and derive a hard classifier based on the marginal density and compare the risk between that and the Bayes-optimal classifier. Our work differs in that we give uniform bounds on the recovery of the marginal density, which is a considerably stronger notion of consistency. This is important in situations where a worst-case bound on classifier performance is required.

Density Estimation on Manifolds Density estimation on manifolds has received much less attention than the full-dimensional counterpart. However, understanding density estimation in situations where the intrinsic dimension can be much lower than the ambient dimension is becoming ever more important: modern systems are able to capture data at an increasing resolution while the number of degrees of freedom stays relatively constant. One of the limiting aspects of density-based approaches is their performance in high dimensions. It takes an exponential in dimension number of samples to estimate the density – this is the so-called curse of dimensionality. Here we give results whose rates of convergence depend on the dimension of the manifold d_M compared to a much higher ambient dimension d; thus the convergence properties become much more attractive under the manifold hypothesis.

Local Intrinsic Dimension Estimation Many learning algorithms require the intrinsic dimension as an input in order to take advantage of the lower dimensional structure that arises. There has been much work on estimating the intrinsic dimension of the data given finite samples e.g. (Kégl, 2003). However, the more interesting problem of estimating the *local* intrinsic dimension has received much less attention. The bulk of the work in this area e.g. (Costa et al., 2005; Houle, 2013; Amsaleg et al., 2015) provide interesting estimators, but are unable to establish strong finite-sample guarantees under nonparametric assumptions. In this paper, we consider a simple notion of local intrinsic dimension based on the doubling dimension and utilize a simple estimator. We then give a uniform finite-sample convergence result for the estimator under nonparametric assumptions. To the best of our knowledge, this is perhaps the strongest finite-sample result obtained this far for this problem.

3. Background and Setup

Definition 1. *Let f be a probability density over $\mathbb{R}^d$ with corresponding distribution $\mathcal{F}$. Let $X = \{X_1, ..., X_n\}$ be n i.i.d. samples drawn from it and let $\mathcal{F}_n$ denote the empirical distribution w.r.t. X. i.e. $\mathcal{F}_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \in A\}$.*

We only require that f is bounded.

Assumption 1. $||f||_\infty < \infty$.

Definition 2. *Define kernel function $K : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ where $\mathbb{R}_{\geq 0}$ denotes the non-negative real numbers such that*

$$\int_{\mathbb{R}^d} K(u)du = 1.$$

We make the following mild regularity assumptions on K.

Assumption 2. *(Spherically Symmetric and non-increasing) There exists non-increasing function $k : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that $K(u) = k(|u|)$ for $u \in \mathbb{R}^d$.*

Assumption 3. *(Exponential Decays) There exists $\rho, C_\rho, t_0 > 0$ such that for $t > t_0$,*

$$k(t) \leq C_\rho \cdot \exp(-t^\rho).$$

Remark 1. *These assumptions allow the popular kernels such as Gaussian, Exponential, Silverman, uniform, triangular, tricube, Cosine, and Epanechnikov.*

Assumption 3 implies the next result which will be useful later on. The proof is elementary and is omitted.

Lemma 1. *For all $m > 0$, we have*

$$\int_{\mathbb{R}^d} K(u)|u|^m du < \infty.$$

Definition 3 (Bandwidth matrix)**.** *$\mathbf{H}$ is a valid bandwidth matrix if it is a positive definite and symmetric $d \times d$ matrix. $\mathbf{H_0}$ is a unit bandwidth matrix if it is a valid bandwidth matrix and $|\mathbf{H_0}| = 1$.*

Let $\sigma_1(\mathbf{H_0}) \geq \cdots \geq \sigma_d(\mathbf{H_0}) > 0$ be the eigenvalues of $\mathbf{H_0}$.

Remark 2. *In the scalar bandwidth case, $\mathbf{H_0} = \mathbf{I}$.*

Remark 3. *It will be useful later on that if $\mathbf{H} = h^2\mathbf{H_0}$ where $\mathbf{H_0}$ is a unit bandwidth, then for $u \in \mathbb{R}^d$,*

$$\sqrt{\sigma_d(\mathbf{H_0})} \cdot h \cdot |u| \leq |\mathbf{H}^{1/2}u| \leq \sqrt{\sigma_1(\mathbf{H_0})} \cdot h \cdot |u|.$$

Definition 4 (Kernel Density Estimation)**.** *Given a kernel K and $h > 0$ and $\mathbf{H_0}$, the KDE for $\mathbf{H} := h^2\mathbf{H_0}$ is given by*

$$\widehat{f}_{\mathbf{H}}(x) := \frac{1}{n} \cdot |\mathbf{H}|^{-d/2} \sum_{i=1}^{n} K\left(\mathbf{H}^{-1/2}(x - X_i)\right)$$
$$= \frac{1}{n \cdot h^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{H_0}^{-1/2}(x - X_i)}{h}\right).$$

4. Uniform Convergence Bounds

The following is a paraphrase of Bousquet et al. (2004), which was given in Chaudhuri & Dasgupta (2010).

Lemma 2. *Let $\mathcal{G}$ be a class of functions from X to $\{0, 1\}$ with VC dimension $d < \infty$, and $\mathcal{F}$ a probability distribution on X. Let $\mathbb{E}$ denote expectation with respect to $\mathcal{F}$. Suppose n points are drawn independently at random from $\mathcal{F}$; let $\mathbb{E}_n$ denote expectation with respect to this sample. Then with probability at least $1 - 1/n$, the following holds for all $g \in \mathcal{G}$:*

$$-\min\{\beta_n\sqrt{\mathbb{E}_n g}, \beta_n^2 + \beta_n\sqrt{\mathbb{E}g}\}$$
$$\leq \mathbb{E}g - \mathbb{E}_n g \leq \min\{\beta_n^2 + \beta_n\sqrt{\mathbb{E}_n g}, \beta_n\sqrt{\mathbb{E}g}\},$$

where $\beta_n \geq \sqrt{4(d+3)\log 2n/n}$.

Chaudhuri & Dasgupta (2010) takes $\mathcal{G}$ to be the indicator functions over balls. Dasgupta & Kpotufe (2014) uses this to provide similar bounds for the k-NN density estimator as in this paper. Here, we extend this idea to ellipsoids by taking $\mathcal{G} = \mathcal{B}$ (the indicator functions over ellipsoids), which has VC dimension $(d^2 + 3d)/2$ as determined by Akama & Irie (2011).

Lemma 3. *Define ellipsoid $B_{\mathbf{H_0}}(x, r) := \{x' \in \mathbb{R}^d : |\mathbf{H_0}^{-1/2}(x - x')| \leq r\}$, and $\mathcal{B} := \{B_{\mathbf{H_0}}(x, r) : x \in \mathbb{R}^d, r > 0, \mathbf{H_0}$ is a unit bandwidth$\}$. With probability at least $1 - 1/n$, the following holds uniformly for every $B \in \mathcal{B}$ and $\gamma \geq 0$:*

$$\mathcal{F}(B) \geq \gamma \Rightarrow \mathcal{F}_n(B) \geq \gamma - \beta_n\sqrt{\gamma} - \beta_n^2,$$
$$\mathcal{F}(B) \leq \gamma \Rightarrow \mathcal{F}_n(B) \leq \gamma + \beta_n\sqrt{\gamma} + \beta_n^2,$$

where $\beta_n = 8d\sqrt{\log n/n}$.

Remark 4. *We could have alternatively used a fixed confidence* δ *so that our results would hold with probability at least* $1-\delta$. *This would only require a modification of* β_n *(e.g. by taking* $\beta_n = 4d\sqrt{2(\log n + \log(1/\delta))/n}$*). In this paper, we have simply taken* $\delta = 1/n$.

5. KDE Bound

Define the following which characterizes how much the density can respectively decrease and increase from x in $B(x,r)$.

Definition 5.

$$\breve{u}_x(r) := f(x) - \inf_{x' \in B(x,r)} f(x').$$
$$\hat{u}_x(r) := \sup_{x' \in B(x,r)} f(x') - f(x).$$

The first are general upper and lower bounds for $\widehat{f}_{\mathbf{H}}$.

Theorem 1. *[Uniform Upper and Lower Bounds for* $\widehat{f}_{\mathbf{H}}$*] Let* v_d *be the volume of the unit ball in* $\mathbb{R}^d$. *Then the following holds uniformly in* $x \in \mathbb{R}^d$, $\epsilon > 0$, *unit bandwidths* $\mathbf{H_0}$, *and* $h > (\log n/n)^{1/d}$ *with probability at least* $1 - 1/n$. *Let* $\mathbf{H} := h^2\mathbf{H_0}$.

$$\widehat{f}_{\mathbf{H}}(x) > f(x) - \epsilon - C\sqrt{\frac{\log n}{n \cdot h^d}},$$

if $\int_{\mathbb{R}^d} K(u) \cdot \breve{u}_x(h|u|/\sqrt{\sigma_d(\mathbf{H_0})})du < \epsilon$, *and*

$$\widehat{f}_{\mathbf{H}}(x) < f(x) + \epsilon + C\sqrt{\frac{\log n}{n \cdot h^d}},$$

if $\int_{\mathbb{R}^d} K(u) \cdot \hat{u}_x(h|u|/\sqrt{\sigma_d(\mathbf{H_0})})du < \epsilon$, *where* $C = 8d\sqrt{\cdot v_d \cdot ||f||_\infty}\left(\int_0^\infty k(t) \cdot t^{d/2}dt + 1\right) + 64d^2 \cdot k(0)$.

Remark 5. *The conditions on* $\breve{u}_x(h|u|/\sqrt{\sigma_d})$ *and* $\hat{u}_x(h|u|/\sqrt{\sigma_d})$ *can be interpreted as a bound on their expectations over the probability measure* K *(i.e.* $\int_{\mathbb{R}^d} K(u)du = 1$*). These conditions can be satisfied by taking* h *sufficiently small.*

Remark 6. *The parameter* ϵ *allows us the amount of slack in the estimation errors. This is useful in a few aspects. Oftentimes, we don't require tight bounds, especially when reasoning about low density regions thus having a large* ϵ *allows us to satisfy the conditions more easily. In the case that we want tight bounds, the additive error controlled by the pointwise smoothness of the density can be encoded in* ϵ*, so to not require global smoothness assumptions.*

Remark 7. *Besides the* $||f||_\infty$ *factor, the value of* C *at the end of the theorem statement is a quantity which can be known without any a priori knowledge of* f. *We can bound* $||f||_\infty$ *in terms of known quantities given smoothness assumptions near* $\text{argmax}_x f(x)$. *This is used in later results where knowing a value of* C *is important.*

In order to prove Theorem 1, we first define the following two functions which serve to approximate K as a step-wise linear combination of uniform kernels.

Definition 6. *Let* $\Delta > 0$.

$$\underline{K_\Delta}(u) := \sum_{j=0}^{\infty}(k(j\Delta) - k((j+1)\Delta)) \cdot \mathbb{1}\{|u| < j\Delta\},$$
$$\overline{K_\Delta}(u) := \sum_{j=0}^{\infty}(k(j\Delta) - k((j+1)\Delta)) \cdot \mathbb{1}\{|u| < (j+1)\Delta\}.$$

Then it is clear that the following holds for all $\Delta > 0$.

$$\underline{K_\Delta}(u) \le K(u) \le \overline{K_\Delta}(u).$$

The next Lemma is useful in computing the expectations of functions over the kernel measure.

Lemma 4. *Suppose* g *is an integrable function over* $\mathbb{R}_{\ge 0}$ *and let* v_d *denote the volume of a unit ball in* $\mathbb{R}^d$. *Then*

$$\int_{\mathbb{R}^d} K(u)g(|u|)du = v_d \cdot \int_0^\infty k(t) \cdot t^d g(|u|)du.$$

Proof of Lemma 4. Let $S_d = 2\pi^{d/2}/\Gamma(d/2)$ denote the surface area of the unit ball in $\mathbb{R}^d$.

$$\int_{\mathbb{R}^d} K(u)g(|u|)du = S_d \int_0^\infty k(t) \cdot t^{d-1} \cdot g(|t|)dt$$
$$= \frac{S_d}{d}\int_0^\infty (k(t) \cdot g(|t|))t^d dt = v_d \int_0^\infty (k(t) \cdot g(|t|))t^d dt,$$

where the second last equality follows from integration by parts and the last follows from the fact that $v_d = S_d/d$. □

The following follows immediately from Lemma 4.

Corollary 1.

$$\int_{\mathbb{R}^d} K(u)\breve{u}_x(h|u|)du = v_d \cdot \int_0^\infty k(t) \cdot t^d \cdot \breve{u}_x(ht)dt,$$
$$\int_{\mathbb{R}^d} K(u)\hat{u}_x(h|u|)du = v_d \cdot \int_0^\infty k(t) \cdot t^d \cdot \hat{u}_x(ht)dt.$$

Proof of Theorem 1. Assume that the event that Lemma 3 holds, which occurs with probability at least $1 - 1/n$. We first show the lower bound for $\widehat{f}_{\mathbf{H}}(x)$. Define

$$\widehat{f}_{\Delta,\mathbf{H}}(x) := \frac{1}{n \cdot h^d}\sum_{i=1}^n \underline{K_\Delta}\left(\frac{\mathbf{H_0}^{-1/2}(x - X_i)}{h}\right).$$

It is clear that $\widehat{f}_{\mathbf{H}}(x) \ge \widehat{f}_{\Delta,\mathbf{H}}(x)$ for all $x \in \mathbb{R}^d$. Let us use the following shorthand $\Delta_{k,j} := k(j\Delta)$. We have

$$\widehat{f}_{\Delta,\mathbf{H}}(x) = \frac{1}{h^d}\sum_{j=0}^\infty (\Delta_{k,j} - \Delta_{k,j+1}) \cdot \mathcal{F}_n\left(B_{\mathbf{H_0}}(x, jh\Delta)\right).$$

We next get a handle on each $\mathcal{F}_n\left(B_{\mathbf{H_0}}(x, jh\Delta)\right)$. We have

$$\mathcal{F}\left(B_{\mathbf{H_0}}(x, jh\Delta)\right) \geq v_d \cdot (jh\Delta)^d \cdot F_j,$$

where $F_j := \max\{0, f(x) - \check{u}_x(jh\Delta/\sqrt{\sigma_d(\mathbf{H_0})})\}$. Thus, by Lemma 3, we have

$$\begin{aligned}
&\mathcal{F}_n\left(B_{\mathbf{H_0}}(x, jh\Delta)\right) \\
&\geq v_d \cdot (jh\Delta)^d \cdot F_j - \beta_n\sqrt{v_d} \cdot (jh\Delta)^{d/2} \cdot \sqrt{F_j} - \beta_n^2 \\
&\geq v_d \cdot (jh\Delta)^d \cdot F_j - \beta_n\sqrt{v_d \cdot ||f||_\infty} \cdot (jh\Delta)^{d/2} - \beta_n^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\widehat{f}_{\Delta,h}(x) \\
&\geq v_d \sum_{j=0}^{\infty}(\Delta_{k,j} - \Delta_{k,j+1})(j\Delta)^d \cdot f(x) \\
&- v_d \sum_{j=0}^{\infty}(\Delta_{k,j} - \Delta_{k,j+1})(j\Delta)^d \cdot \check{u}_x\left(\frac{jh\Delta}{\sqrt{\sigma_d(\mathbf{H_0})}}\right) \\
&- \frac{\beta_n\sqrt{v_d \cdot ||f||_\infty}}{h^{d/2}} \cdot \sum_{j=0}^{\infty}(\Delta_{k,j} - \Delta_{k,j+1})(j\Delta)^{d/2} \\
&- \beta_n^2\frac{k(0)}{h^d}.
\end{aligned}$$

We handle each term separately. For the first term, we have

$$\begin{aligned}
&\lim_{\Delta\to 0} v_d \sum_{j=0}^{\infty}(\Delta_{k,j} - \Delta_{k,j+1})(j\Delta)^d \\
&= v_d \int_0^\infty k(t)t^d dt = 1.
\end{aligned}$$

where the last equality follows from Lemma 4. Next, we have

$$\begin{aligned}
&\lim_{\Delta\to 0} v_d \sum_{j=0}^{\infty}(\Delta_{k,j} - \Delta_{k,j+1})(j\Delta)^d \cdot \check{u}_x\left(\frac{jh\Delta}{\sqrt{\sigma_d(\mathbf{H_0})}}\right) \\
&= v_d \int_0^\infty k(t) \cdot t^d \cdot \check{u}_x(th/\sqrt{\sigma_d(\mathbf{H_0})})dt < \epsilon.
\end{aligned}$$

Finally, we have

$$\begin{aligned}
&\lim_{\Delta\to 0} \sum_{j=0}^{\infty}(\Delta_{k,j} - \Delta_{k,j+1})(j\Delta)^{d/2} \\
&= \int_0^\infty k(t) \cdot t^{d/2} dt < \infty.
\end{aligned}$$

Thus, taking $\Delta \to 0$ we get

$$\begin{aligned}
\widehat{f}_{\mathbf{H}}(x) \geq f(x) - \epsilon - \frac{\beta_n\sqrt{v_d \cdot ||f||_\infty}}{h^{d/2}} \cdot \int_0^\infty k(t) \cdot t^{d/2} dt \\
- \beta_n^2\frac{k(0)}{h^d}.
\end{aligned}$$

This gives us the lower bound. Next we derive an upper bound. Let us redefine

$$\widehat{f}_{\Delta,\mathbf{H}}(x) := \frac{1}{n \cdot h^d}\sum_{i=1}^{n}\overline{K_m}\left(\frac{x - X_i}{h}\right).$$

It is clear that $\widehat{f}_{\mathbf{H}}(x) \leq \widehat{f}_{\Delta,\mathbf{H}}(x)$ for all $x \in \mathbb{R}^d$ and

$$\begin{aligned}
&\widehat{f}_{\Delta,\mathbf{H}}(x) \\
&= \frac{1}{h^d}\sum_{j=0}^{\infty}(\Delta_{k,j} - \Delta_{k,j+1}) \cdot \mathcal{F}_n\left(B_{\mathbf{H_0}}(x, (j+1)h\Delta)\right).
\end{aligned}$$

We next get a handle on each $\mathcal{F}_n\left(B_{\mathbf{H_0}}(x, jh\Delta)\right)$. We have

$$\mathcal{F}(B_{\mathbf{H_0}}(x, jh\Delta)) \leq v_d \cdot (jh\Delta)^d \cdot F_j$$

where $F_j = \min\{||f||_\infty, f(x) + \hat{u}(jh\Delta/\sqrt{\sigma_d(\mathbf{H_0})})\}$. Thus by Lemma 3 we have

$$\begin{aligned}
&\mathcal{F}_n\left(B_{\mathbf{H_0}}(x, jh/m)\right) \\
&\leq v_d(jh\Delta)^d F_j + \beta_n(jh\Delta)^{d/2}\sqrt{v_d \cdot F_j} + \beta_n^2.
\end{aligned}$$

Using this, we now have

$$\begin{aligned}
&\widehat{f}_{\Delta,\mathbf{H}}(x) \\
&\leq v_d \sum_{j=0}^{\infty}(\Delta_{k,j} - \Delta_{k,j+1})((j+1)\epsilon)^d \cdot f(x) \\
&+ v_d \sum_{j=0}^{\infty}(\Delta_{k,j} - \Delta_{k,j+1})((j+1)\Delta)^d \cdot \hat{u}_x\left(\frac{(j+1)h\Delta}{\sqrt{\sigma_d(\mathbf{H_0})}}\right) \\
&+ \frac{\beta_n\sqrt{v_d \cdot ||f||_\infty}}{h^{d/2}} \cdot \sum_{j=0}^{\infty}(\Delta_{k,j} - \Delta_{k,j+1})((j+1)\Delta)^{d/2} \\
&+ \beta_n^2\frac{k(0)}{h^d}.
\end{aligned}$$

We proceed the same way as the other direction. Thus taking $\Delta \to 0$ we get

$$\begin{aligned}
\widehat{f}_{\Delta,\mathbf{H}}(x) \leq f(x) + \epsilon + \frac{\beta_n\sqrt{v_d \cdot ||f||_\infty}}{h^{d/2}} \cdot \int_0^\infty k(t) \cdot t^{d/2} dt \\
+ \beta_n^2\frac{k(0)}{h^d}.
\end{aligned}$$

The result follows. □

6. Sup-norm Bounds for KDE

Theorem 2. *[ℓ_∞ bound for α-Hölder continuous functions] If f is Hölder-continuous (i.e. $|f(x) - f(x')| \leq C_\alpha|x - x|^\alpha$ for $x, x' \in \mathbb{R}^d$ and $0 < \alpha \leq 1$), then there exists positive constant $C' \equiv C'(C, C_\alpha, \alpha, K)$ such that the following holds with probability at least $1 - 1/n$ uniformly in $h > (\log n/n)^{1/d}$ and unit bandwidths $\mathbf{H_0}$. Let $\mathbf{H} := h^2\mathbf{H_0}$.*

$$\sup_{x\in\mathbb{R}^d}|\widehat{f}_{\mathbf{H}}(x) - f(x)| < C' \cdot \left(\frac{h^\alpha}{\sigma_d(\mathbf{H_0})^{\alpha/2}} + \sqrt{\frac{\log n}{n \cdot h^d}}\right).$$

Remark 8. *Taking $h = n^{-1/(2\alpha+d)}$ in the above r.h.s. optimizes the rates to $n^{-\alpha/(2\alpha+d)}$ (ignoring log factors).*

Remark 9. *We can attain similar results (although not uniform in bandwidth) by a straightforward application of Theorem 3.1 of Sriperumbudur & Steinwart (2012) or Proposition 9 of Rinaldo & Wasserman (2010).*

7. Mode Estimation Results

The goal of this section is to utilize the KDE to estimate the mode of a uni-modal distribution from its samples. We borrow the estimator from Abraham et al. (2004)

$$\hat{x} := \text{argmax}_{x \in X} \widehat{f}_{\mathbf{H}}(x),$$

where $\mathbf{H} := h^2\mathbf{I}$.

We adopt the mode estimation framework assumptions from Dasgupta & Kpotufe (2014) which are summarized below.

Definition 7. *x_0 is a mode of f if $f(x) < f(x_0)$ for all $x \in B(x_0, r)\backslash\{x_0\}$ for some $r > 0$.*

Assumption 4.

- *f has a single mode x_0.*
- *f is twice differentiable in a neighborhood around x_0.*
- *f has a negative-definite Hessian at x_0.*

These assumptions lead to the following.

Lemma 5 ((Dasgupta & Kpotufe, 2014))**.** *Let f satisfy Assumption 4. Then there exists $\hat{C}, \check{C}, r_0, \lambda > 0$ such that the following holds.*

$$\check{C} \cdot |x_0 - x|^2 \leq f(x_0) - f(x) \leq \hat{C} \cdot |x_0 - x|^2$$

for all $x \in A_x$ where A_0 is a connected component of $\{x : f(x) \geq \lambda\}$ and A_0 contains $B(x_0, r_0)$.

We obtain the following result for the estimation error of $\hat{x}$.

Theorem 3. *Suppose that Assumptions 1, 2, 3, 4 hold. Choose h such that $(\log n)^{2/\rho} \cdot h \to 0$ and $\log n/(nh^d) \to 0$ as $n \to \infty$. Then, for n sufficiently large depending on $d, ||f||_\infty, K, \hat{C}, \check{C}, r_0$ the following holds with probability least $1 - 1/n$.*

$$|\hat{x} - x_0|^2 \leq \max\left\{\frac{32\hat{C}}{\check{C}}(\log n)^{4/\rho} \cdot h^2, 17 \cdot C\sqrt{\frac{\log n}{n \cdot h^d}}\right\}.$$

Remark 10. *Taking $h = n^{-1/(4+d)}$ optimizes the above expression so that $|\hat{x} - x_0| \lesssim n^{-1/(4+d)}$ (ignoring log factors) which matches the lower bound rate for mode estimation as established in Tsybakov (1990).*

Remark 11. *This result can be extended to multi-modal distributions as done by Dasgupta & Kpotufe (2014) by using the connected components of nearest neighbor graphs at appropriate empirical density levels to isolate the modes away from each other.*

8. Density Level Set Estimation Results

In this section, we estimate the density level set $L_f(\lambda) := \{x : f(x) \geq \lambda\}$ where $\lambda > 0$ is given. We make the following standard regularity assumptions e.g. (Singh et al., 2009). To simplify the analysis, let us take $\mathbf{H} = h^2\mathbf{I}$. It is clear that the results that follow can be extended to arbitrary $\mathbf{H_0}$.

Assumption 5 (β-regularity)**.** *Let $0 < \beta < \infty$. There exists $0 < \lambda_0 < \lambda$ and $\check{C}_\beta, \hat{C}_\beta, \bar{r} > 0$ such that the following holds for $x \in L_f(\lambda_0)\backslash L_f(\lambda)$.*

$$\check{C}_\beta \cdot d(x, L_f(\lambda))^\beta \leq \lambda - f(x) \leq \hat{C}_\beta \cdot d(x, L_f(\lambda))^\beta,$$

where $d(x, A) := \inf_{x' \in A}\{|x - x'|\}$. and $B(L_f(\lambda), \bar{r}) \subseteq L_f(\lambda_0)$ where $B(A, r) := \{x : d(x, A) \leq r\}$.

Then we consider following estimator.

$$\widehat{L}_f := \left\{x \in X : \widehat{f}_{\mathbf{H}}(x) > \lambda - \tilde{C}\sqrt{\frac{\log n}{n \cdot h^d}}\right\}.$$

where $\tilde{C}$ is obtained by taking C and replacing the $||f||_\infty$ factor by $1+5\max_{x \in X_{n_0}} \widehat{f}_{\mathbf{H}}(x)$ where X_{n_0} is a fixed sample of size n_0. Then, $\tilde{C}$ can be viewed as a constant w.r.t. n and can be known without any a priori knowledge of f while ensuring that $\tilde{C} \geq \max\{1, 2C\}$.

We use the following Hausdorff metric.

Definition 8 (Hausdorff Distance)**.**

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

Theorem 4. *Suppose that Assumptions 1, 2, 3, 5 hold and that f is α-Hölder continuous for some $0 < \alpha \leq 1$. Choose h such that $(\log n)^{2/\rho} \cdot h \to 0$ and $\log n/(nh^d) \to 0$ as $n \to \infty$. Then, for n sufficiently large depending on $d, C, \tilde{C}, K, \hat{C}_\beta, \check{C}_\beta, \beta, \bar{r}$ the following holds with probability least $1 - 1/n$.*

$$d_H(\widehat{L}_f, L_f(\lambda)) \leq C''\left((\log n)^{2/\rho} \cdot h + \left(\frac{\log n}{n \cdot h^d}\right)^{1/(2\beta)}\right),$$

where $C'' \equiv C''(C, \tilde{C}, \hat{C}_\beta, \check{C}_\beta, \tilde{C}, \beta)$.

Remark 12. *Choosing $h = n^{-\beta/(2\beta+d)}$ gives us a density-level set estimation rate of $O(n^{-1/(2\beta+d)})$. This matches the lower bound (ignoring log factors) determined by Tsybakov (1997).*

Remark 13. *This result can be extended so that we can recover each component separately (i.e. identify which points correspond to which connected components of $L_f(\lambda)$). Similar to the mode estimation result, this can be done using nearest neighbor graphs at the appropriate level to isolate the connected components of $L_f(\lambda)$ away from each other. This has been done extensively in the related area of cluster tree estimation e.g. (Chaudhuri & Dasgupta, 2010).*

Remark 14. *The global α-Hölder continuous assumption is not required and is only here for simplicity. Smoothness in a neighborhood around a maximizer of f is sufficient so that for n_0 large enough, $\tilde{C} \geq 2C$.*

9. Class Probability Estimation Results

We consider the setting where we have observations from compact subset $\mathcal{X} \subset \mathbb{R}^d$ and labels $y \in \{1, ..., L\}$. Given a label y, an instance $x \in \mathbb{R}^d$ has density $f_y(x)$ where f_y is w.r.t. the uniform measure on $\mathbb{R}^d$. Instance-label pairs (X, Y) are thus drawn according to a mixture distribution where Y is chosen from $\{1, ..., L\}$ with corresponding probabilities $\pi_1, ..., \pi_L$ (i.e. $\sum_{j=1}^{L} \pi_j = 1$) and then X is chosen according to f_Y.

Then given $x \in \mathcal{X}$, we can define the marginal distribution as follows.

$$g(x) := (g_1(x), ..., g_L(x)),$$
$$g_y(x) := f(Y = y | X = x) = \frac{\pi_y f_y(x)}{\sum_j \pi_j f_j(x)}.$$

The goal of class probability estimation is to learn g based on samples $(x_1, y_1), ..., (x_n, y_n)$. We define our estimator naturally as follows. Let $\widehat{f}_{h,y}$ be the KDE of f_y w.r.t. to bandwidth matrix $\mathbf{H} = h^2\mathbf{I}$.

$$\hat{g}_h(x) := (\hat{g}_{h,1}(x), ..., \hat{g}_{h,L}(x)),$$
$$\hat{g}_{h,y}(x) := \frac{\hat{\pi}_y \widehat{f}_{h,y}(x)}{\sum_j \hat{\pi}_j \widehat{f}_{h,j}(x)} \text{ and } \hat{\pi}_y := \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}[y = y_i].$$

We make the following regularity assumption on f_y.

Assumption 6. *(α-Hölder densities) For each $y \in \{1, ..., L\}$ and $x \in \mathbb{R}^d$ we have*

$$|f_y(x) - f_y(x')| \leq C_\alpha |x - x'|^\alpha,$$

where $0 < \alpha \leq 1$.

We state the result below:

Theorem 5. *Suppose that Assumptions 1, 2, 3, 6 hold. Then for n sufficiently large depending on $\min_y \pi_y$, there exists positive constants $C'' \equiv C''(L, C, C_\alpha, \alpha, K)$ and $\tilde{C} \equiv \tilde{C}(\min_y \pi_y, L)$ such that the following holds with probability at least $1 - \tilde{C}/n$ uniformly in $h > (\log n/n)^{1/d}$.*

$$\sup_{x \in \mathbb{R}^d} ||\hat{g}_h(x) - g(x)||_\infty \leq C'' \cdot \left(h^\alpha + \sqrt{\frac{\log n}{n \cdot h^d}} \right).$$

Remark 15. *This corresponds to an optimized rate of $\widetilde{O}(n^{-\alpha/(2\alpha+d)})$. This matches the lower bounds up to log factors for misclassification as established in related works e.g. Audibert et al. (2007); Chaudhuri & Dasgupta (2014). Note that misclassification rate for a hard classifier is a slightly different but very related to what is done here, which is directly estimating the marginal density.*

10. Extension to Manifolds

We make the following regularity assumptions which are standard among works in manifold learning e.g. (Baraniuk & Wakin, 2009; Genovese et al., 2012; Balakrishnan et al., 2013).

Assumption 7. *$\mathcal{F}$ is supported on M where:*

- *M is a d_M-dimensional smooth compact Riemannian manifold without boundary embedded in compact subset $\mathcal{X} \subseteq \mathbb{R}^D$.*
- *The volume of M is bounded above by a constant.*
- *M has condition number $1/\tau$, which controls the curvature and prevents self-intersection.*

Let f be the density of $\mathcal{F}$ with respect to the uniform measure on M.

In this section, we assume that our density estimator is w.r.t. to d_M instead of the ambient dimension d.

$$\widehat{f}_{\mathbf{H}}(x) := \frac{1}{n \cdot h^{d_M}} \sum_{i=1}^{n} K\left(\frac{\mathbf{H_0}^{-1/2}(x - X_i)}{h} \right).$$

Remark 16. *It is then the case that we must know the intrinsic dimension d_M. There are numerous known techniques for doing so e.g. (Kegl, 2002; Levina & Bickel, 2004; Hein & Audibert, 2005; Farahmand et al., 2007).*

Next, we need the following guarantee on the volume of the intersection of a Euclidean ball and M; this is required to get a handle on the true mass of the ball under $\mathcal{F}$ in later arguments. The upper and lower bounds follow from Chazal (2013) and Lemma 5.3 of Niyogi et al. (2008). The proof can be found in (Jiang, 2017).

Lemma 6 (Ball Volume)**.** *If $0 < r < \min\{\tau/4d_M, 1/\tau\}$, and $x \in M$ then*

$$1 - \tau^2 r^2 \leq \frac{vol_{d_M}(B(x, r) \cap M)}{v_{d_M} r^{d_M}} \leq 1 + 4d_M r/\tau,$$

where vol_{d_M} is the volume w.r.t. the uniform measure on M.

We then give analogues to Theorem 1 and Theorem 2.

Theorem 6. *[Manifold Case Uniform Upper and Lower Bounds for $\widehat{f}_{\mathbf{H}}$] There exists $C_M \equiv C_M(d_M, d, K, ||f||_\infty, \tau)$ such that the following holds*

uniformly in $x \in M$, $\epsilon > 0$, *unit bandwidths* $\mathbf{H_0}$, *and* $h > (\log n/n)^{1/d_M}$ *with probability at least* $1 - 1/n$. *Let* $\mathbf{H} := h^2\mathbf{H_0}$.

$$\widehat{f}_{\mathbf{H}}(x) > f(x) - \epsilon - C_M\left(h^2 + \sqrt{\frac{\log n}{n \cdot h^{d_M}}}\right),$$

if $\int_{\mathbb{R}^d} K(u) \cdot \breve{u}_x(h|u|/\sqrt{\sigma_d(\mathbf{H_0})})du < \epsilon$, *and*

$$\widehat{f}_{\mathbf{H}}(x) < f(x) + \epsilon + C_M\left(h + \sqrt{\frac{\log n}{n \cdot h^{d_M}}}\right),$$

if $\int_{\mathbb{R}^d} K(u) \cdot \hat{u}_x(h|u|/\sqrt{\sigma_d(\mathbf{H_0})})du < \epsilon$.

Remark 17. *The extra* h^2 *and* h *term in the lower and upper bounds respectively come from the approximation of the volume of the full-dimensional balls w.r.t. the uniform measure on* M *in Lemma 6.*

Proof Sketch of Theorem 6. The proof mirrors that of the full dimensional case so we only highlight the differences. For the lower bound, instead of

$$\mathcal{F}\left(B_{\mathbf{H_0}}(x, jh\delta)\right) \geq v_d \cdot (jh\delta)^d \cdot F_j,$$

we have

$$\begin{aligned}\mathcal{F}\left(B_{\mathbf{H_0}}(x, jh\delta)\right) &\geq v_{d_M}(jh\delta)^{d_M} F_j(1 - \tau^2(jh\delta)^2) \\ &= v_{d_M}(jh\delta)^{d_M} F_j - h^{d_M+2} v_{d_M}\tau^2||f||_\infty (j\delta)^{d_M+2}.\end{aligned}$$

The first term can be treated in the same way as before, while the second term contributes in an extra term with an h^2 factor after taking the total summation.

For the upper bound, instead of

$$\mathcal{F}(B_{\mathbf{H_0}}(x, jh\delta)) \leq v_d \cdot (jh\delta)^d \cdot F_j,$$

we have

$$\mathcal{F}(B_{\mathbf{H_0}}(x, jh\delta)) \leq v_{d_M} \cdot (jh\delta)^{d_M} \cdot F_j(1 + 4d_M(jh\delta)/\tau).$$

Similary, this contributes an extra term with an h factor after taking the total summation. □

Theorem 7. *[Manifold Case* ℓ_∞ *bound for* α*-Hölder continuous functions] If* f *is Hölder-continuous (i.e.* $|f(x) - f(x')| \leq C_\alpha|x - x|^\alpha$ *for* $x, x' \in \mathbb{R}^d$ *with* $0 < \alpha \leq 1$*), then there exists positive constant* $C'_M \equiv C'_M(||f||_\infty, C_\alpha, \alpha, K, \tau, d_M, d, \sigma_d(\mathbf{H_0}))$ *such that the following holds with probability at least* $1 - 1/n$ *uniformly in* h *satisfying* $(\log n/n)^{1/d_M} < h < 1$.

$$\sup_{x \in M} |\widehat{f}_{\mathbf{H}}(x) - f(x)| < C'_M \cdot \left(h^\alpha + \sqrt{\frac{\log n}{n \cdot h^{d_M}}}\right).$$

11. Local Intrinsic Dimension Estimation

In this section, we only assume a distribution $\mathcal{F}$ on $\mathbb{R}^d$ whose support is defined as $\mathcal{X} := \{x \in \mathbb{R}^d : \mathcal{F}(B(x, h)) > 0\ \forall h > 0\}$ and $\mathcal{X}$ is assumed to be compact. We use the following notion of intrinsic dimension, which is based on the doubling dimension and adapted from previous works such as Houle (2013).

Definition 9. *For* $x \in \mathcal{X}$, *define the following local intrinsic dimension wherever the quantity exists*

$$ID(x) := \lim_{h \to 0} \log_2\left(\frac{\mathcal{F}(B(x, 2h))}{\mathcal{F}(B(x, h))}\right).$$

We can then define our estimator of local intrinsic dimension at $x \in \mathcal{X}$ as follows:

$$\widehat{\mathrm{ID}}_{n,h}(x) := \log_2\left(\frac{\mathcal{F}_n(B(x, 2h))}{\mathcal{F}_n(B(x, h))}\right).$$

The following is a uniform convergence result for $\widehat{\mathrm{ID}}_{n,h}(x)$.

Theorem 8. *Define the following*

$$ID_h(x) := \log_2\left(\frac{\mathcal{F}(B(x, 2h))}{\mathcal{F}(B(x, h))}\right).$$

Suppose that $h > 0$ *and* n *satisfy* $\beta_n < \frac{1}{10}\inf_{x' \in \mathcal{X}}\sqrt{\mathcal{F}(B(x', h))}$. *Then the following holds with probability at least* $1 - 1/n$ *uniformly in* $x \in \mathcal{X}$.

$$|\widehat{ID}_{n,h}(x) - ID_h(x)| \leq \frac{6\beta_n}{\inf_{x' \in \mathcal{X}}\sqrt{\mathcal{F}(x', 2h)}}.$$

Remark 18. *The r.h.s. goes to* 0 *as* $n \to \infty$. *Moreover, if* $ID_h(x)$ *converges to* $ID(x)$ *uniformly in* $x \in \mathcal{X}$, *then simultaneously taking* $h \to 0$ *and* $n \to \infty$ *such that* $\beta_n \cdot \left(\inf_{x' \in \mathcal{X}}\sqrt{\mathcal{F}(x', 2h)}\right)^{-1} \to 0$ *gives us a finite-sample uniform convergence rate for local intrinsic dimension estimation.*

Remark 19. *If we assume a global intrinsic dimension* d_0 *and a density, the condition* $\beta_n < \frac{1}{10}\inf_{x' \in \mathcal{X}}\sqrt{\mathcal{F}(B(x', h))}$ *can be interpreted as* $\frac{\log n}{nh^{d_0}} \to 0$ *and the r.h.s. of the bound is on the order of* $\sqrt{\frac{\log n}{nh^{d_0}}}$.

In fact, this result is similar to the uniform convergence results for the KDE for estimating the smoothed density. e.g. $|\widehat{f}_h - f_h|_\infty = O\left(\sqrt{\frac{\log n}{nh^d}}\right)$ when (ignoring some log factors) $nh^d \to \infty$ where f_h is the density convolved with the uniform kernel with bandwidth h. It is interesting that an analogous result comes up when estimating the intrinsic dimension with our notion of smoothed ID.

References

Abraham, C., Biau, G., and Cadre., B. On the asymptotic properties of a simple estimate of the mode. *ESAIM: Probability and Statistics*, 2004.

Akama, Yohji and Irie, Kei. Vc dimension of ellipsoids. *arxiv*, 2011.

Amsaleg, Laurent, Chelly, Oussama, Furon, Teddy, Girard, Stéphane, Houle, Michael E, Kawarabayashi, Ken-ichi, and Nett, Michael. Estimating local intrinsic dimensionality. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 29–38. ACM, 2015.

Arias-Castro, Ery, Mason, David, and Pelletier, Bruno. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *Journal of Machine Learning Research*, 2016.

Audibert, Jean-Yves, Tsybakov, Alexandre B, et al. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633, 2007.

Balakrishnan, S., Narayanan, S., Rinaldo, A., Singh, A., and Wasserman, L. Cluster trees on manifolds. In *Advances in Neural Information Processing Systems*, pp. 2679–2687, 2013.

Baraniuk, Richard G and Wakin, Michael B. Random projections of smooth manifolds. *Foundations of computational mathematics*, 9(1):51–77, 2009.

Botev, Zdravko I, Grotowski, Joseph F, Kroese, Dirk P, et al. Kernel density estimation via diffusion. *The Annals of Statistics*, 38(5):2916–2957, 2010.

Bousquet, O., Boucheron, S., and Lugosi, G. Introduction to statistical learning theory. *Lecture Notes in Artificial Intelligence*, 2004.

Cadre, Benoit. Kernel estimation of density level sets. *Journal of Multivariate Analysis*, 2006.

Carmichael, J., George, G., and Julius, R. Finding natural clusters. *Systematic Zoology*, 1968.

Chaudhuri, K. and Dasgupta, S. Rates for convergence for the cluster tree. *Advances in Neural Information Processing Systems*, 2010.

Chaudhuri, Kamalika and Dasgupta, Sanjoy. Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pp. 3437–3445, 2014.

Chaudhuri, Probal, Ghosh, Anil K, and Oja, Hannu. Classification based on hybridization of parametric and nonparametric classifiers. *IEEE transactions on pattern analysis and machine intelligence*, 31(7):1153–1164, 2009.

Chazal, F. An upper bound for the volume of geodesic balls in submanifolds of euclidean spaces. 2013.

Cheng, Y. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1995.

Chernoff, Herman. Estimation of the mode. *Annals of the Institute of Statistical Mathematics*, 1964.

Comaniciu, D and Meer, P. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002.

Costa, Jose A, Girotra, Abhishek, and Hero, AO. Estimating local intrinsic dimension with k-nearest neighbor graphs. In *Statistical Signal Processing, 2005 IEEE/SP 13th Workshop on*, pp. 417–422. IEEE, 2005.

Cuevas, A. and Fraiman, R. A plug-in approach to support estimation. *Annals of Statistics*, 1997.

Dasgupta, S. and Kpotufe, S. Optimal rates for k-nn density and mode estimation. *Advances in Neural Information Processing Systems*, 2014.

Eddy, William F. Optimum kernel estimators of the mode. *Annals of Statistics*, 1980.

Einmahl, Uwe and Mason, David M. Uniform in bandwidth consistency of kernel-type function estimators. *Annals of Statistics*, 2005.

Farahmand, A., Szepesvari, C., and Audibert, J. Manifold-adaptive dimension estimation. *ICML*, 2007.

Fukunaga and Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 1975.

Genovese, Christopher, Perone-Pacifico, Marco, Verdinelli, Isabella, and Wasserman, Larry. Minimax manifold estimation. *Journal of machine learning research*, 13(May):1263–1291, 2012.

Genovese, Christopher R., Verdinelli, Marco Perone-Pacificoand Isabella, and Wasserman, Larry. Nonparametric inference for density modes. *Series B Statistical Methodology*, 2015.

Gine and Guillon. Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. H. Poincare Probab. Statist.*, 2002.

Hartigan, J. Clustering algorithms. *Wiley*, 1975.

Hein, Matthias and Audibert, Jean-Yves. Intrinsic dimensionality estimation of submanifolds in rd. *ICML*, 2005.

Houle, Michael E. Dimensionality, discriminability, density and distance distributions. In *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*, pp. 468–473. IEEE, 2013.

Jiang, Heinrich. Density level set estimation on manifolds with dbscan. *International Conference on Machine Learning (ICML)*, 2017.

Jiang, Heinrich and Kpotufe, Samory. Modal-set estimation with an application to clustering. *AISTATS*, 2017.

Kegl, Balazs. Intrinsic dimension estimation using packing numbers. *NIPS*, 2002.

Kégl, Balázs. Intrinsic dimension estimation using packing numbers. In *Advances in neural information processing systems*, pp. 697–704, 2003.

Levina, Elizaveta and Bickel, Peter J. Maximum likelihood estimation of intrinsic dimension. *NIPS*, 2004.

Li, J., Ray, S., and Lindsay, B. A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, 2007.

Niyogi, P., Smale, S., and Weinberger, S. Finding the homology of submanifolds with high confidence from random samples. *Discrete and Computational Geometry*, 2008.

Parzen, Emanual. On estimation of a probability density function and mode. *Annals of mathematical statistics*, 1962.

Rigollet, P. and Vert, R. Fast rates for plug-in estimators of density level sets. *Bernoulli*, 15(4):1154–1178, 2009.

Rigollet, Philippe. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 8(Jul):1369–1392, 2007.

Rinaldo, A. and Wasserman, L. Generalized density clustering. *Annals of Statistics*, 2010.

Rosenblatt, M. Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, 1956.

Silverman, B. *Density Estimation for Statistics and Data Analysis.* CRC Press, 1986.

Silverman, B. W. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1981.

Simonoff, Jeffrey S. *Smoothing Methods in Statistics*. Springer, 1996.

Singh, Aarti, Scott, Clayton, and Nowak, Robert. Adaptive hausdorff estimation of density level sets. *Annals of Statistics*, 2009.

Sriperumbudur, Bharath K. and Steinwart, Ingo. Consistency and rates for clustering with dbscan. *AISTATS*, 2012.

Steinwart, I. Adaptive density level set clustering. *24th Annual Conference on Learning Theory*, 2011.

Terrell, George R and Scott, David W. Variable kernel density estimation. *The Annals of Statistics*, pp. 1236–1265, 1992.

Tsybakov, A. Recursive estimation of the mode of a multivariate distribution. *Problemy Peredachi Informatsii*, 1990.

Tsybakov, A. Introduction to nonparametric estimation. *Springer*, 2008.

Tsybakov, A. B. On non-parametric estimation of density level sets. *Annals of Statistics*, 1997.