Optimal rates for k-NN density and mode estimation

Sanjoy Dasgupta  
University of California, San Diego, CSE  
dasgupta@eng.ucsd.edu

Samory Kpotufe *  
Princeton University, ORFE  
samory@princeton.edu

Abstract

We present two related contributions of independent interest: (1) high-probability finite sample rates for k-NN density estimation, and (2) practical mode estimators – based on k-NN – which attain minimax-optimal rates under surprisingly general distributional conditions.

1 Introduction

We prove finite sample bounds for k-nearest neighbor (k-NN) density estimation, and subsequently apply these bounds to the related problem of mode estimation. These two main results, while related, are interesting on their own.

First, k-NN density estimation [1] is one of the better known and simplest density estimation procedures. The estimate \( f_k(x) \) of an unknown density \( f \) (see Definition 1 of Section 3) is a simple functional of the distance \( r_k(x) \) from \( x \) to its \( k \)-th nearest neighbor in a sample \( X_n = \{ X_i \}_{i=1}^n \). As such it is intimately related to other functionals of \( r_k(x) \), e.g. the degree of vertices \( x \) in k-NN graphs and their variants used in modeling communities and in clustering applications (see e.g. [2]).

While this procedure has been known for a long time, its convergence properties are still not fully understood. The bulk of research in the area has concentrated on establishing its asymptotic convergence, while its finite sample properties have received little attention in comparison. Our finite sample bounds are concisely derived once the proper tools are identified. The bounds hold with high probability, under general conditions on the unknown density \( f \). This generality proves quite useful as shown in our subsequent application to the problem of mode estimation.

The basic problem of estimating the modes (local maxima) of an unknown density \( f \) has also been studied for a while (see e.g. [3] for an early take on the problem). It arises in various unsupervised problems where modes are used as a measure of typicality of a sample \( X \). In particular, in modern applications, mode estimation is often used in clustering, with the modes representing cluster centers (see e.g. [4],[5] and general applications of the popular mean-shift procedure).

While there exists a rich literature on mode estimation, the bulk of theoretical work concerns estimators of a single mode (highest maximum of \( f \)), and often concentrates on procedures that are hard to implement in practice. Given the generality of our first result on k-NN density estimation, we can prove that some simple implementable procedures yield optimal estimates of the modes of an unknown density \( f \), under surprisingly general conditions on \( f \).

Our results are overviewed in the following section, along with an overview of the rich literature on k-NN density estimation and mode estimation. This is followed by our theoretical setup in Section 3, our rates for k-NN density estimation are detailed in Section 4 while the results on mode estimation are given in Section 5.

*Much of this work was conducted when this author was at TTI-Chicago.
2 Overview of results and related Work

2.1 Rates for $k$-NN density estimates

The $k$-NN density estimator dates back perhaps to the early work of [1] where it is shown to be consistent when the unknown density $f$ is continuous on $\mathbb{R}^d$. While one of the best known and simplest procedure for density estimation, it has proved more cumbersome to analyze than its smooth counterpart, the kernel density estimator.

More general consistency results such as [6, 7] have been established since its introduction. In particular [6] shows that, for $f$ Lipschitz in a neighborhood of a point $x$, where $f(x) > 0$, and $k = k(n)$ satisfying $k \to \infty$ and $k/n^{2/(2+d)} \to 0$, the estimator is asymptotically normal, i.e. $\sqrt{k}(f_k(x) - f(x))/f(x) \overset{D}{\to} N(0,1)$. The recent work of [8], concerning generalized weighted variants of $k$-NN, shows that asymptotic normality holds under the weaker restriction $k/n^{4/(4+d)} \to 0$ if $f$ is twice differentiable at $x$.

Asymptotic normality as stated above yields some insight into the rate of convergence of $f_k$: we can expect that $|f_k(x) - f(x)| \lesssim f(x)/\sqrt{k}$ under the stated conditions on $k$. In fact, [8] shows that such a result can be obtained in expectation for $n = n(x)$ sufficiently large. In particular, their conditions on $k$ allows for a setting of $k \approx n^{4/(4+d)}$ (not allowed under the above conditions) yielding a minimax-optimal $l_2$ risk $\mathbb{E} |f_k(x) - f(x)|^2 \lesssim f(x)^2/k = O(n^{-4/(4+d)})$.

While consistency results and bounds on expected error are now well understood, we still don’t have a clear understanding of the conditions under which high probability bounds on $|f_k(x) - f(x)|$ are possible. This is particularly important given the inherent instability of nearest neighbors estimates which are based on order-statistics rather than the more stable average statistics at the core of kernel-density estimates. The recent result of [9] provides an initial answer: they obtain a high-probability bound uniformly over $x$ taking value in the sample $X_{[n]}$, however under conditions not allowing for optimal settings of $k$ (where $f$ is assumed Lipschitz).

The bounds in the present paper hold with high-probability, simultaneously for all $x$ in the support of $f$. Rather than requiring smoothness conditions on $f$, we simply give the bounds in terms of the modulus of continuity of $f$ at any $x$, i.e. how much $f$ can change in a neighborhood of $x$. This allows for a useful degree of flexibility in applying these bounds. In particular, optimal bounds under various degrees of smoothness of $f$ at $x$ easily follow. More importantly, for our application to mode estimation, the bounds allow us to handle $|f_k(x) - f(x)|$ at different $x \in \mathbb{R}^d$ with varying smoothness in $f$. As a result we can derive minimax-optimal mode estimation rates for practical procedures under surprisingly weak assumptions.

2.2 Mode estimation

There is an extensive literature on mode estimation and we unfortunately can only overview some of the relevant work. Most of the literature covers the case of a unimodal distribution, or one where there is a single maximizer $x_0$ of $f$.

Early work on estimating the (single) mode of a distribution focused primarily on understanding the consistency and rates achievable by various approaches, with much less emphasis on the ease of implementation of these approaches. The common approaches consist of estimating $x_0$ as $\hat{x} \triangleq \arg\sup_{x \in \mathbb{R}^d} f_n(x)$ where $f_n$ is an estimate of $f$, usually a kernel density estimate. Various work such as [3, 10, 11] establish consistency properties of the approach and achievable rates under various Euclidean settings and regularity assumptions on the distribution $F$. More recent work such as [12, 13] address the problem of optimal choice of bandwidth and kernel to adaptively achieve the minimax risk for mode estimation. Essentially, under smoothness $\kappa$ (e.g. $f$ is $\kappa$ times differentiable), the minimax risk ($\inf_{\hat{x}} \sup_{f} \mathbb{E}_f \|\hat{x} - x_0\|$) is of the form $n^{-1/(\kappa + 1)}$, as independently established in [14] and [15].

As noticed early in [16], the estimator $\arg\sup_{x \in \mathbb{R}^d} f_n(x)$, while yielding much insight into the problem, is hard to implement in practice. Hence, other work, apparently starting with [16, 14] have looked into so-called recursive estimators of the (single) mode which are practical and easy to update as the sample size increases. These approaches can be viewed as some form of gradient-
ascent of \( f_n \) with carefully chosen step sizes. The later versions of [14] are shown to be minimax-optimal. Another line of work is that of so-called direct mode estimators which estimate the mode from practical statistics of the data [17, 18]. In particular, [18] shows that the simple and practical estimator \( \arg \max_{x \in X_n} f_n(x) \), where \( f_n \) is a kernel-density estimator, is a consistent estimator of the mode. We show in the present paper that \( \arg \max_{x \in X_n} f_k(x) \), where \( f_k \) is a k-NN density estimator, is not only consistent, but converges at a minimax-optimal rate under surprisingly mild distributional conditions.

The more general problem of estimating all modes of distribution has received comparatively little attention. The best known practical approach for this problem is the mean-shift procedure and its variants [19, 4, 20, 21], quite related to recursive-mode-estimators, as they essentially consist of gradient ascent of \( f \) variants. The best known practical approach for this problem is the mean-shift estimator, is not only consistent, but converges at a minimax-optimal rate under surprisingly mild global distributional conditions. We show that any mode that is sufficiently salient given the inherent variability in estimating \( f \) from sample. This question was recently addressed by [23] which proposes a hypothesis test to detect false modes based on confidence intervals around Hessians estimated at the modes returned by any procedure.

Interestingly, while a k-NN density estimate \( f_k \) is far from smooth, in fact not even continuous, we show a simple practical procedure that identifies any mode of the unknown density \( f \) under mild conditions: we mainly require that \( f \) is well approximated by a quadratic in a neighborhood of any mode \( x \). While mean-shift is popular in practice, it has proved quite difficult to analyze. A recent result of [22] comes close to establishing the consistency of mean-shift, as it establishes the convergence of the procedure to the right gradient lines (essentially the ascent path to the mode) if it is seeded from fixed starting points rather than the random samples themselves. It remains unclear however whether mean-shift produces only true modes, given the inherent variability in estimating \( f \) from sample. This question was recently addressed by [23] which proposes a hypothesis test to detect false modes based on confidence intervals around Hessians estimated at the modes returned by any procedure.

As alluded to so far, our results are given in terms of local assumptions on modes rather than global distributional conditions. We show that any mode that is sufficiently salient (this is locally parametrized) w.r.t. the finite sample size \( n \), is optimally estimated, while false modes are pruned away. In particular our results allow for \( f \) having a countably infinite number of modes.

### 3 Preliminaries

Throughout the analysis, we assume access to a sample \( X_n = \{X_i\}_{i=1}^n \) drawn i.i.d. from an absolutely continuous distribution \( F \) over \( \mathbb{R}^d \), with Lebesgue-density function \( f \). We let \( \mathcal{X} \) denote the support of the density function \( f \).

The k-NN density estimate at a point \( x \) is defined as follows.

**Definition 1** (k-NN density estimate). For every \( x \in \mathbb{R}^d \), let \( r_k(x) \) denote the distance from \( x \) to its k-th nearest neighbor in \( X_n \). The density estimate is given as:

\[
    f_k(x) \triangleq \frac{k}{n \cdot v_d \cdot (r_k(x))^d},
\]

where \( v_d \) denotes the volume of the unit sphere in \( \mathbb{R}^d \).

All balls considered in the analysis are closed Euclidean balls of \( \mathbb{R}^d \).
4 \( k \)-NN density estimation rates

In this section we bound the error in estimating \( f(x) \) as \( f_k(x) \) at every \( x \in \mathcal{X} \). The main results of the section are Lemmas 5 and 4. These lemmas are easily obtained given the right tools: uniform concentration bounds on the empirical mass of balls in \( \mathbb{R}^d \), using relative Vapnik-Chervonenkis bounds, i.e. Bernstein’s type bounds rather than Chernoff type bounds (see e.g. Theorem 5.1 of [23]). We next state a form of these bounds for completion.

**Lemma 1.** Let \( \mathcal{G} \) be a class of functions from \( \mathcal{X} \) to \( \{0, 1\} \) with VC dimension \( d < \infty \), and \( \mathbb{P} \) a probability distribution on \( \mathcal{X} \). Let \( \mathbb{E} \) denote expectation with respect to \( \mathbb{P} \). Suppose \( n \) points are drawn independently at random from \( \mathbb{P} \); let \( \mathbb{E}_n \) denote expectation with respect to this sample. Then for any \( \delta > 0 \), with probability at least \( 1 - \delta \), 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 \( \beta_n = \sqrt{(4/n)(d \ln 2n + \ln(8/\delta))} \).

These sort of relative VC bounds allows for a tighter relation (than Chernoff type bounds) between empirical and true mass of sets (\( \mathbb{E}_n g \) and \( \mathbb{E} g \)) in those situations where these quantities are small, i.e. of the order of \( \beta_n^2 = \hat{O}(1/n) \) above. This is particularly useful since the balls we have to deal with are those containing approximately \( k \) points, and hence of (small) mass approximately \( k/n \).

A direct result of the above lemma is the following lemma of [26]. This next lemma essentially reworks Lemma 1 above into a form we can use more directly. We re-use \( C_{\delta,n} \) below throughout the analysis.

**Lemma 2 ([25]).** Pick \( 0 < \delta < 1 \). Let \( C_{\delta,n} \) denote \( 16 \log(2/\delta) \sqrt{d \ln n} \). Assume \( k \geq d \log n \). With probability at least \( 1 - \delta \), for every ball \( B \subset \mathbb{R}^d \) we have,

\[
\mathcal{F}(B) \geq C_{\delta,n} \sqrt{\frac{d \log n}{n}} \implies \mathcal{F}_n(B) > 0,
\]

\[
\mathcal{F}(B) \geq \frac{k}{n} + C_{\delta,n} \sqrt{\frac{k}{n}} \implies \mathcal{F}_n(B) \geq \frac{k}{n}, \text{ and}
\]

\[
\mathcal{F}(B) \leq \frac{k}{n} - C_{\delta,n} \sqrt{\frac{k}{n}} \implies \mathcal{F}_n(B) < \frac{k}{n}.
\]

The main idea in bounding \( f_k(x) \) is to bound the random term \( r_k(x) \) in terms of \( f(x) \) using Lemma 2 above. We can deduce from the lemma that if a ball \( B(x, r) \) centered has mass roughly \( k/n \), then its empirical mass is likely to be of the order \( k/n \); hence \( r_k(x) \) is likely to be close to the radius \( r \) of \( B(x, r) \). Now if \( f \) does not vary too much in \( B(x, r) \), then we can express the mass of \( B(x, r) \) in terms of \( f(x) \), and thus get our desired bound on \( r_k(x) \) and \( f_k(x) \) in terms of \( f(x) \).

Our results are given in terms of how \( f \) varies in a neighborhood of \( x \), captured as follows.

**Definition 2.** For \( x \in \mathbb{R}^d \) and \( \epsilon > 0 \), define \( \hat{r}(\epsilon, x) \) \( \hat{=} \) \( \sup \left\{ r : \sup_{\|x-x'\| \leq r} f(x') - f(x) \leq \epsilon \right\} \), and \( \hat{r}(\epsilon, x) \) \( \hat{=} \) \( \sup \left\{ r : \sup_{\|x-x'\| \leq r} f(x) - f(x') \leq \epsilon \right\} \).

The continuity parameters \( \hat{r}(\epsilon, x) \) and \( \hat{r}(\epsilon, x) \) (related to the modulus of continuity of \( f \) at \( x \)) are easily bounded under smoothness assumptions on \( f \) at \( x \). Our high-probability bounds on the estimates \( f_k(x) \) in terms of \( f(x) \) and the continuity parameters are given as follows.

**Lemma 3** (Upper-bound on \( f_k \)). Suppose \( k \geq 4C_{\delta,n}^2 \). Then, with probability at least \( 1 - \delta \), for all \( x \in \mathbb{R}^d \) and all \( \epsilon > 0 \),

\[
f_k(x) < \left(1 + 2 \frac{C_{\delta,n}}{\sqrt{k}}\right) (f(x) + \epsilon),
\]

provided \( k \) satisfies \( v_d \cdot \hat{r}(\epsilon, x)^d \cdot (f(x) + \epsilon) \geq \frac{k}{n} - C_{\delta,n} \sqrt{k} \).
Lemma 4 (Lower-bound on $f_k$). Then, with probability at least $1 - \delta$, for all $x \in \mathbb{R}^d$ and all $\epsilon > 0$,
\[
f_k(x) \geq \left(1 - \frac{C_{\delta,n}}{\sqrt{k}}\right)(f(x) - \epsilon),
\]
provided $k$ satisfies $v_d \cdot \hat{r}(\epsilon, x)^d \cdot (f(x) - \epsilon) \geq \frac{k}{n} + C_{\delta,n} \frac{\sqrt{\epsilon}}{n}$.

The proof of these results are concise applications of Lemma 2 above. They are given in the appendix (long version). The trick is in showing that, under the conditions on $k$, there exists an $r \approx (k/(n \cdot f(x)))^{1/d}$ which is at most $\hat{r}(\epsilon, x)$ or $\hat{r}(\epsilon, x)$ as appropriate; hence, $f$ does not vary much on $B(x, r)$ so we must have
\[
\mathcal{F}(B(x, r)) \approx \text{volume}(B(x, r)) \cdot f(x) = v_d \cdot r^d \cdot f(x) \approx \frac{k}{n}.
\]

Using Lemma 2 we get $r_k(x) \approx r$; plug this value into $f_k(x)$ to obtain $f_k(x) \approx (1 + 1/\sqrt{k})f(x)$.

Lemmas 3 and 4 allow a great deal of flexibility as we will soon see with their application to mode estimation. In particular we can consider various smoothness conditions simultaneously at different $x$ for different biases $\epsilon$.

Suppose for instance that $f$ is locally Hölder at $x$, i.e. $\exists r, L, \beta > 0$ s.t. for all $x' \in B(x, r)$, $|f(x) - f(x')| \leq L \|x - x'\|^\beta$. Then for small $\epsilon$, both $\hat{r}(\epsilon, x)$ and $\hat{r}(\epsilon, x')$ are at least $(\epsilon/L)^{1/\beta}$; pick $\epsilon = O(f(x)/\sqrt{k})$ for $n$ sufficiently large, then by both lemmas we have, w.h.p., $|f_k(x) - f(x)| \leq O(f(x)/\sqrt{k})$ provided $k = \Omega((\log^2 n)$ and satisfies $v_d(1/\sqrt{k})^d f(x) \geq Ck/n$ for some constant $C$. This allows for a setting of $k = \Theta(n^{2\beta/(2\beta+d)})$ for a minimax-optimal rate of $|f_k(x) - f(x)| = O(n^{-\beta/(2\beta+d)})$.

The ability to consider various biases $\epsilon$ would prove particularly helpful in the next section on mode estimation where we have to consider different approximations in different parts of space with varying smoothness in $f$. In particular, at a mode $x$, we will essentially have $\beta = 2$ ($f$ is twice differentiable) while elsewhere on $\mathcal{X}$ we might not have much smoothness in $f$.

5 Mode estimation

We start with the following definition of modes.

Definition 3. We denote the set of modes of $f$ by $\mathcal{M} \equiv \{x : \exists r > 0, \forall x' \in B(x, r), f(x') < f(x)\}$.

We need the following assumption at modes.

Assumption 1. $f$ is twice differentiable in a neighborhood of every $x \in \mathcal{M}$. We denote the gradient and Hessian of $f$ by $\nabla f$ and $\nabla^2 f$. Furthermore, $\nabla^2 f(x)$ is negative definite at all $x \in \mathcal{M}$.

Assumption 1 excludes modes at the boundary of the support of $f$ (where $f$ cannot be continuously differentiable). We note that most work on the subject consider only interior modes as we are doing here. Modes on the boundary can however be handled under additional boundary smoothness assumptions to ensure that $f$ puts sufficient mass on any ball around such modes. This however only complicates the analysis, while the main insights remain the same as for interior modes.

An implication of Assumption 1 is that for all $x \in \mathcal{M}$, $\nabla f$ is continuous in a neighborhood of $x$, with $\nabla f(x) = 0$. Together with $\nabla^2 f(x) \prec 0$ (i.e. negative definite), $f$ is well-approximated by a quadratic in a neighborhood of a mode $x \in \mathcal{M}$. This is stated in the following lemma.

Lemma 5. Let $f$ satisfy Assumption 1. Consider any $x \in \mathcal{M}$. Then there exists a neighborhood $B(x, r), r > 0$, and constants $\tilde{C}_x, \check{C}_x > 0$ such that, for all $x' \in B(x, r)$, we have
\[
\check{C}_x \|x' - x\|^2 \leq f(x) - f(x') \leq \tilde{C}_x \|x' - x\|^2. \tag{1}
\]

We can therefore parametrize a mode $x \in \mathcal{M}$ locally as follows:

Definition 4 (Critical radius $r_x$ around mode $x$). For every mode $x \in \mathcal{M}$, there exists $r_x > 0$, such that $B(x, r_x)$ is contained in a set $A_x$, satisfying the following conditions:
(i) $A_x$ is a connected component of a level set $\lambda^x \triangleq \{x' \in \mathcal{X} : f(x') > \lambda\}$ for some $\lambda > 0$.
(ii) $\exists \check{C}_x, \tilde{C}_x > 0, \forall x' \in A_x, \check{C}_x \|x' - x\|^2 \leq f(x) - f(x') \leq \tilde{C}_x \|x' - x\|^2$. (So $A_x \cap \mathcal{M} = \{x\}$.)
Theorem 1. Let $N$ outside $B$

Proof. Let $x$ be the mode of $f$.

Consider any $x$ where we analyze the practical procedure of Figure 1 where we pick the maximizer of $R$ maximizer of some density estimator (usually a kernel estimate) over the entire space $\mathbb{R}^d$. Instead we analyze the practical procedure of Figure 1 where we pick the maximizer of $f_k$ out of the finite sample $X[n]$. The rates of Theorem 1 are optimal $(O(n^{-1/(4+d)}))$ for a setting of $k = O(n^{4/(4+d)})$.

Theorem 1. Let $\delta > 0$. Assume $f$ has a single mode $x_0$ and satisfies Assumptions 1, 2. There exists $N_{x_0,\delta}$ such that the following holds for $n \geq N_{x_0,\delta}$. Let $C_{x_0}, \hat{C}_{x_0}$ be as in Definition 4. Suppose $k$ satisfies

$$
\left( \frac{24C_{\delta,n} f(x_0)}{C_{x_0}^2} \right)^2 \leq k \leq \left( \frac{1}{2} \sqrt{\frac{C_{\delta,n}}{C_{x_0}}} \right)^{4d/(4+d)} f(x_0)^{(2d+4)/(4+d)} \left( \frac{\nu_d}{4} \right)^{4/(4+d)}.
$$

Let $x$ be the mode returned in the procedure of Figure 1. With probability at least $1 - 2\delta$ we have

$$
\|x - x_0\| \leq 5 \sqrt{\frac{C_{\delta,n}}{C_{x_0}} f(x_0)} \cdot \frac{1}{k^{1/4}}.
$$

Proof. Let $r_{x_0}$ be the critical radius of Definition 4. Let $r_n(x_0) \equiv \inf \{ r : B(x_0, r) \cap X[n] \neq \emptyset \}$. Let $0 < \tau < 1$ to be later specified, and assume the event that $r_n(x_0) \leq \frac{\tau}{2} r_{x_0}$. We will bound the probability of this event once the proper setting of $\tau$ becomes clear.

Consider $\tilde{r}$ satisfying $r_{x_0} \geq \tilde{r} \geq 2r_n(x_0)/\tau$ (see Figure 3). We will first upper bound $f_k$ for any $x$ outside $B(x_0, \tilde{r})$, then lower-bound $f_k$ for $x \in B(x_0, r_n(x_0))$.

Recall $A_{x_0}$ from Definition 4. By equation (1) we have

$$
\sup_{x \in A_{x_0} \cap B(x_0, \tilde{r}/2)} f(x) \leq f(x_0) - \tilde{C}_{x_0}(\tilde{r}/2)^2 \leq \hat{F}.
$$

The above allows us to apply Lemma 3 as follows. First note that for any $x \in \mathcal{X}\setminus B(x_0, \tilde{r}/2)$, $f(x) \leq \hat{F}$ since $A_{x_0}$ is a level set of the unimodal $f$, i.e. $\sup_{x \notin A_{x_0}} f(x) \leq \inf_{x \in A_{x_0}} f(x)$. Therefore, for any $x \in \mathcal{X}\setminus B(x_0, \tilde{r})$ let $\epsilon = \hat{F} - f(x)$. By equation (3) the modulus of continuity $\tilde{r}(\epsilon, x)$ is at least
Initialize: $M_n \leftarrow \emptyset$.

For $\lambda = \max_{x \in X_{[n]}} f_n(x)$ down to 0:

- Let $\epsilon \triangleq \lambda / C_{\delta,n} / \sqrt{K}$.
- Let $\{ \tilde{A}_i \}_{i=1}^m$ be the CCs of $G(\lambda - \epsilon - \tilde{\epsilon})$ disjoint from $M_n$.
- $M_n \leftarrow M_n \cup \left\{ x_i \triangleq \arg \max_{x \in \tilde{A}_i \cap X_{[n]}} f_n(x) \right\}_{i=1}^m$.

Return the estimated modes $M_n$.

Figure 3: Estimate the modes of a multimodal $f$ from $X_{[n]}$. The parameter $\tilde{\epsilon}$ serves to prune.

Therefore, if $k$ satisfies

$$v_d \cdot \left( \frac{r}{2} \right)^d \cdot \left( f(x) - \tilde{C}_{x_0} \left( \frac{r}{2} \right)^2 \right) \geq \frac{k}{n} - C_{\delta,n} \sqrt{K},$$

we have with probability at least $1 - \delta$

$$\sup_{x \in X \setminus B(x_0,r)} f_k(x) < \left( 1 + 2 \frac{C_{\delta,n}}{\sqrt{K}} \right) \left( f(x) - \tilde{C}_{x_0} \left( \frac{r}{2} \right)^2 \right).$$

Now we turn to $x \in B(x_0,r_0(x_0))$. We have again by equation $\mathbb{1}$ that $\inf_{x \in B(x_0,\tau r)} f(x) \geq f(x_0) - \tilde{C}_{x_0} (\tau r)^2 \triangleq \tilde{F}$. Therefore, for $x \in B(x_0,r_0(x_0))$ let $\epsilon = f(x) - \tilde{F}$, we have $\tilde{r}(\epsilon, x) \geq \tau r - r_0(x_0) \geq \tau \tilde{r}/2$. It follows that, if $k$ satisfies

$$v_d \cdot \left( \frac{r}{2} \right)^d \cdot \left( f(x) - \tilde{C}_{x_0} \left( \frac{r}{2} \right)^2 \right) \geq \frac{k}{n} + C_{\delta,n} \sqrt{K},$$

we have by Lemma $\mathbb{4}$ that, with probability at least $1 - \delta$ (under the same event used in Lemma $\mathbb{5}$)

$$\inf_{x \in B(x_0,r_0(x_0))} f_k(x) \geq \left( 1 - \frac{C_{\delta,n}}{\sqrt{K}} \right) \left( f(x) - \tilde{C}_{x_0} (\tau r)^2 \right).$$

Next, with a bit of algebra, we can pick $\tau$ and $\tilde{r}$ so that the l.h.s. of $\mathbb{5}$ is less than the l.h.s. of equation $\mathbb{7}$. It suffices to pick $\tau^2 = \tilde{C}_{x_0}/8\tilde{C}_{x_0}$ and $\tilde{r}^2 \geq 24 f(x_0) C_{\delta,n} / \tilde{C}_{x_0} \sqrt{K}$. Given these settings, equations $\mathbb{4}$ and $\mathbb{6}$ are satisfied whenever $k$ satisfies equation $\mathbb{2}$ of the lemma statement.

It follows that, with probability at least $1 - \delta$, $\inf_{x \in B(x_0,r_0(x_0))} f_k(x) > \sup_{x \in X \setminus B(x_0,\tilde{r})} f_k(x)$. Therefore, the empirical mode chosen by the procedure is in $B(x_0,\tilde{r})$. We are free to choose $\tilde{r}$ as small as $\max \left\{ \frac{\sqrt{K} f(x_0)}{C_{\delta,n}}, \frac{\tilde{C}_{x_0} \sqrt{K}}{2 r_0(x_0)} \right\}$.

We’ve assumed so far the event that $r_0(x_0) \leq \frac{\tau r}{2}$. We bound the probability of this event as follows. Let $r \triangleq \sqrt{24 f(x_0) C_{\delta,n} / \tilde{C}_{x_0} \sqrt{K}}$. Under the above setting of $\tau$, the Theorem’s assumptions on $k$ imply that $r \leq r_0$, and that $v_d \cdot \left( \frac{r}{2} \right)^d \cdot \left( f(x) - \tilde{C}_{x_0} \left( \frac{r}{2} \right)^2 \right) \geq \frac{k}{n} + C_{\delta,n} \sqrt{K}$. Again, by equation $\mathbb{1}$, this implies that $\mathbb{1} F(B(x_0, \tau r)^2) \geq \frac{k}{n} + C_{\delta,n} \sqrt{K}$. By Lemma $\mathbb{2}$, with probability at least $1 - \delta$, $\mathbb{2} \mathbb{F}_n(B(x_0, \tau r)^2) \geq k/n$ and therefore $r_0(x_0) \leq (\tau r) r \leq (\tau r) r_0$.

Now becomes clear that we can just pick $\tilde{r} = r$.

### 5.2 Multiple modes

In this section we turn to the problem of estimating the modes of a more general density $f$ with an unknown number of modes.

The algorithm of Figure 3 operates on the following set of nested graphs $G(\lambda)$. These are subgraphs of a mutual $k$-NN graph on the sample $X_{[n]}$, where vertices are connected if they are in each other’s nearest neighbor sets. The connected components (CCs) of these graphs $G(\lambda)$ are known to be good estimates of the CCs of corresponding level sets of the unknown density $f$ $[9, 26, 27]$. 


Assumption 2. Suppose \( r \) satisfies Assumption 1.

Theorem 3. The proof of the main theorem of this section is based on Lemma 7.4 of [24].

5.2.2 Pruning guarantees

We will show that for a given \( n \), any sufficiently salient mode is optimally recovered; furthermore, if \( f \) is uniformly continuous on \( \mathbb{R}^d \), then the procedure returns no false mode above a level \( \lambda_n \rightarrow 0 \).

5.2.1 Optimal Recovery for Any Mode

The guarantees of this section would be given in terms of salient modes as defined below. Essentially a mode \( x_0 \) is salient if it is separated from other modes by a sufficiently wide and deep valley. We define saliency in a way similar to [9], but simpler: we only require a wide valley since the smoothness of \( f \) at the mode (as expressed in equation 1) takes care of the depth.

We start with a notion of separation between sets inspired from [26].

Definition 6 (r-separation). A, A’ ⊂ X are r-separated if there exists a (separating) set \( S \subset \mathbb{R}^d \) such that: every path from A to A’ crosses S, and \( \sup_{x \in S + B(0,r)} f(x) < \inf_{x \in A \cup A'} f(x) \).

Our notion of mode saliency follows: for a mode \( x \), we require the critical set \( A_x \) of Definition 4 to be well separated from all components at the level where it appears.

Definition 7 (r-salient Modes). A mode \( x \) of \( f \) is said to be r-salient for \( r > 0 \) if the following holds. There exist \( A_x \) as in Definition 4 (with the corresponding \( r_x \), \( C_x \) and \( \tilde{C}_x \)), which is a CC of say \( X^{\lambda_x} \triangleq \{ x \in X : f(x) \geq \lambda_x \} \). \( A_x \) is r-separated from \( X^{\lambda_x} \setminus A_x \).

The next theorem again yields the optimal rates \( O(n^{-1/(4+d)}) \) for \( k = O(n^{d/(4+d)}) \).

Theorem 2 (Recovery of salient modes). Assume \( f \) satisfies Assumptions 1, 2. Suppose \( \tilde{\epsilon} = \tilde{\epsilon}(n) \rightarrow 0 \). Let \( x_0 \) be an r-salient mode for some \( r > 0 \). Assume \( k = \Omega \left( C_{\lambda,n}^2 \right) \). Then there exist \( N = N \left( x_0, \{ \tilde{\epsilon}(n) \} \right) \) depending on \( x_0 \) and \( \tilde{\epsilon}(n) \) such that the following holds for \( n \geq N \).

Let \( x_0, \tilde{C}_x, \tilde{C}_x \) be as in Definition 4 and let \( \lambda_{x_0} \triangleq \inf_{x \in A_{x_0}} f(x) \). Let \( \delta > 0 \). Suppose \( k \) further satisfies

\[
\left( \frac{24 C_{\lambda,n} f(x_0)}{\tilde{C}_{x_0} \min \{ r_x^2/4, (r_x/\alpha)^2 \}} \right)^2 \leq k \leq \left( \frac{1}{2} \right) \left( \frac{C_{\lambda,n}}{\tilde{C}_{x_0}} \right)^{4d/4 + 4d} \lambda_x^{-2d/(4d + 4) + 4d/4 + 4d} \left( \frac{v_d}{n^4} \right)^{1/4 + 4d}.
\]

Let \( M_n \) be the modes returned by the procedure of Figure 3. With probability at least \( 1 - \delta \), there exists \( x \in M_n \), such that

\[
\| x - x_0 \| \leq \frac{\sqrt{\frac{C_{\lambda,n} f(x_0)}{\tilde{C}_{x_0}}} \cdot \frac{1}{k^{1/4}}}{\sqrt{4}}.
\]

5.2.2 Pruning guarantees

The proof of the main theorem of this section is based on Lemma 7.4 of [24].

Theorem 3. Let \( \Lambda \triangleq \sup_x f(x) \) and \( r(\epsilon) \triangleq \sup_{x \in \mathbb{R}^d} \max \{ \tilde{r}(\epsilon, x), \tilde{r}(\epsilon, x) \} \). Assume \( f \) satisfies Assumption 2. Suppose \( r(\epsilon) = \Omega \left( k/n \right)^{1/\beta} \) which is feasible whenever \( f \) is uniformly continuous on \( \mathbb{R}^d \). In particular, if \( f \) is Hölder continuous, i.e.

\[
\forall x, x' \in \mathbb{R}^d, \quad \| f(x) - f(x') \| \leq L \| x - x' \| ^\beta, \quad \text{for some} \ L > 0, 0 < \beta \leq 1,
\]

then we can just let \( \tilde{\epsilon} = \Omega \left( k/n \right)^{\beta/\beta} \) since \( r(\epsilon) \geq (\epsilon/L)^{1/\beta} \). Define

\[
\lambda_0 = \max \left\{ 2\tilde{\epsilon} \frac{\Lambda}{k} C_{\lambda,n}^2, \left( \frac{k}{n} + C_{\lambda,n} \sqrt{k} n \right) \frac{2}{v_d r(\epsilon)^d} \right\}.
\]

Assume \( k \geq 9 C_{\lambda,n}^2 \). The following holds with probability at least \( 1 - \delta \). Pick any \( \lambda \geq 2\lambda_0 \), and let \( \lambda_f = \inf_{x \in X^{\lambda}_[n]} f(x) \). All estimated modes in \( M_n \cap X^{\lambda}_[n] \) can be assigned to distinct modes in \( M \cap X^{\lambda_f} \).
References


