Electronic Journal of Statistics
Vol. 16 (2022) 2774–2795
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1972

Estimation of the global mode of a density: Minimaxity, adaptation, and computational complexity

Ery Arias-Castro

Department of Mathematics, University of California, San Diego
e-mail: eariascastro@ucsd.edu

Wanli Qiao

Department of Statistics, George Mason University
e-mail: wqiao@gmu.edu

and

Lin Zheng

Department of Mathematics, University of California, San Diego
e-mail: liz176@ucsd.edu

Abstract: We consider the estimation of the global mode of a density under some decay rate condition around the global mode. We show that the maximum of a histogram, with proper choice of bandwidth, achieves the minimax rate that we establish for the setting that we consider. This is based on knowledge of the decay rate. To address the situation where the decay rate is unknown, we propose a multiscale variant consisting in the recursive refinement of a histogram. We show that this variant is minimax adaptive. These methods run in linear time, and we prove in an appendix that this is best possible: There is no estimation procedure running in sublinear time that achieves the minimax rate.

Keywords and phrases: Mode estimation, minimax adaptive, multiscale estimation, histogram-based estimation.

Received June 2021.

1. Introduction

The global mode of a bounded density f on $\mathbb{R}^d$ is simply $\arg\max_{x\in\mathbb{R}^d} f(x)$, which we assume here to be a singleton. It is a particularly important parameter when the density is assumed to be (strongly) unimodal — in which case it is simply referred to as the mode. In what follows, we use 'mode' to refer to the global mode, even when the density may have multiple local maxima. The problem of estimating the mode of a density dates back to Parzen (1962), who considered a plug-in estimator consisting first in estimating the entire density by kernel density estimation — a method which had been proposed by Rosenblatt (1956) only a few years earlier — and then in locating the mode of that

density estimate. The problem has received a considerable amount of attention since then, partly because it is a prototypical example of a nonparametric point estimation problem — indeed, one does not need to work through a density estimator to estimate other location parameters such as the mean or median. We refer the reader interested in this long and rich history to a recent survey paper by Chacón (2020), where the estimation of multiple local modes is also discussed in light of its intimate connection to the problem of clustering (Hartigan, 1975).

Although a number of methods have been proposed in the literature, Parzen's approach and its close variants appear to be the most popular and have been thoroughly studied over the years. Parzen (1962) proved that the estimator was asymptotically normal under some conditions. These conditions were refined over the years, including in a paper by Chernoff (1964) who looked at using the uniform kernel (he calls the resulting kernel density estimator "naive"), which does not satisfy the conditions imposed in (Parzen, 1962). The asymptotic normality of the kernel density plug-in estimator was extended to the multivariate setting by Konakov (1973) and Samanta (1973), and, much later, by Mokkadem and Pelletier (2003), who also established various laws of the iterated logarithm.

In this paper we study a closely related method which is based on histograms rather than kernel density estimates. Although the method cannot be said to be really new, it enjoys a number of desirable properties. In the process of establishing and discussing these properties, we also consider questions of convergence rates, computational complexity, and parameter tuning (the bandwidth in Parzen's method).

1.1. Working assumptions

Our basic assumption is that the underlying density behaves like the power function with exponent β near its mode and that it is bounded away from its maximum elsewhere. Specifically, we assume that *f has a unique mode at x_0*, and that, for some $0 < c_0 < C_0$, $h_0 > 0$ and $\beta > 0$,

$$f(x_0) - C_0\|x - x_0\|^\beta \le f(x) \le f(x_0) - c_0\|x - x_0\|^\beta, \quad \text{when } \|x - x_0\| \le h_0, \tag{1}$$

$$f(x) \le f(x_0) - c_0 h_0^\beta, \quad \text{when } \|x - x_0\| \ge h_0. \tag{2}$$

For convenience, we will also assume that

$$f \text{ has compact support.} \tag{3}$$

1.2. Convergence rates

Rates of convergence are already implied in (Parzen, 1962), and were subsequently studied under various assumptions on the underlying density in (Abraham, Biau and Cadre, 2004; Donoho and Liu, 1991; Eddy, 1980; Romano, 1988a; Vieu, 1996), as well as the other publications on the topic mentioned so far. For

example, when the density is twice differentiable with bounded second derivative, Parzen's estimator with optimal choice of bandwidth achieves the rate $O(n^{-1/5})$. This was shown to be minimax optimal by Donoho and Liu (1991) under the same conditions as ours displayed in Section 1.1. Essentially, the density is assumed to have a unique global mode and to behave quadratically in a neighborhood of that mode. See also the lower bound derived by Tsybakov (1990), although the setting is a little bit different. Romano (1988a) derives local minimax rates with respect to a neighborhood around the density defined by densities that are close up to a certain order: if the underlying density is C^p then the neighborhood consists of densities which are themselves and their derivatives up to order $p-1$ pointwise close in a neighborhood of the mode. Actual minimax rates are derived by Klemelä (2005) under similar smoothness assumptions. As it turns out, assuming that the density is log-concave in addition to twice differentiable does not change the minimax rate of convergence (Balabdaoui, Rufibach and Wellner, 2009).

Contribution 1. *We extend the minimax result of Donoho and Liu (1991) to the general situation where the density behaves like a power function with arbitrary exponent $\beta > 0$ in a neighborhood of its mode. We complement this by showing that the methods we propose achieve the minimax rate.*

Confidence intervals or regions for the mode are discussed in a number of publications (Balabdaoui, Rufibach and Wellner, 2009; Doss and Wellner, 2019; Dümbgen and Walther, 2008; Eckle et al., 2018; Genovese et al., 2016; Romano, 1988b; Rufibach and Walther, 2010) under various settings, and they are at least implicit in the papers mentioned earlier discussing the asymptotic normality of the mode, since such an asymptotic limit implies an asymptotically valid confidence region (most often an ellipsoid) when the scale parameters are estimated by plug-in.

1.3. Computational complexity

The main reason we work with a histogram rather than a kernel density estimator is computational ease: the maximizer of a histogram can be computed in (average) linear time, both in the dimension and the sample size.

The question of computational complexity has received some attention and has led to variants such as that of Abraham, Biau and Cadre (2003) who suggest maximizing a kernel density estimator among the sample points, thus avoiding a possibly costly grid search. This might also be a motivation behind some proposals based on nearest neighbors (or spacings in dimension one) as presented in (Dalenius, 1965; Dasgupta and Kpotufe, 2014; Sager, 1978; Venter, 1967). Gradient-based estimates such as the mean-shift algorithm of Fukunaga and Hostetler (1975) and the closely related procedure proposed by Tsybakov (1990) may also have a computational advantage over a grid search approach depending on the refinement of the grid and the number of iterates.

Contribution 2. *The methods we propose achieve the minimax estimation rate*

while having linear computational complexity. We show that this is best possible in the sense that no method with sublinear computational complexity can achieve the minimax rate.

Remark 1. As a reviewer pointed out, unless the number of points is gigantic, in practice the computational burden of locating a mode is typically not a concern. This is true unless the operation needs to be repeated many times, perhaps in the context of a resampling approach used to provide a confidence interval for the mode.

1.4. Parameter tuning

Parzen's method requires a choice of bandwidth. Most of the effort in this direction has been to optimize the accuracy of estimating the density itself and not so much the mode. For example, one can use cross-validation to choose the bandwidth with the intention of minimizing some measure of estimation error for the density — see (Arlot and Celisse, 2010) and references therein — and then proceed with Parzen's approach, meaning compute the kernel density estimate with this choice of bandwidth and locate the mode of that estimate. However, in our setting where we impose a condition on the behavior of the density only in a neighborhood of its mode, it is not at all clear that such an estimator would achieve the minimax rate. It turns out that it does in a setting where the density is assumed twice differentiable everywhere and with strictly negative second derivative at its mode. Balabdaoui, Rufibach and Wellner (2009) operate under a different global assumption, that the density is log-concave. A maximum likelihood estimator exists under this so-called 'shape constraint' alone, and its mode is shown to be minimax optimal under the additional assumption that the density is twice differentiable at its mode. Klemelä (2005) approaches the problem using (and extending) Lepski's method to select the kernel density estimator bandwidth but tailored to the estimation of the mode. The performance rate of the corresponding procedure is established and shown to match the minimax lower bound also derived in the paper for this adaptive setting. This is done assuming that the density is smooth near its mode.

The problem of selecting a tuning parameter for the estimation of a mode is otherwise addressed via testing for the significance of modes. This is done in a number of papers (Chacón and Duong, 2013; Duong et al., 2008; Genovese et al., 2016; Godtliebsen, Marron and Chaudhuri, 2002; Rufibach and Walther, 2010; Silverman, 1981). This is closely related to the problem of testing for unimodality. We refer the reader to additional references in (Eckle et al., 2018) where that connection is made.

Contribution 3. *We propose a parameter-free method that operates in linear time and achieves the non-adaptive minimax rate in our setting. The method is multiscale in nature and performs some sort of bisection search.*

Our approach has antecedents. Indeed, Robertson and Cryer (1974) describe a method that, in dimension one, iteratively focuses on the shortest interval with

a certain number of data points, with that number decreasing at a certain rate. The method is shown to be consistent under some mild conditions. Sager (1979) considers a multivariate version based on convex sets. The method is shown to be consistent, and a (suboptimal) rate of convergence is derived in the one-dimensional setting. Devroye (1979) discusses a method based on kernel density estimates at various bandwidth sizes. The method is shown to be consistent but no rate of convergence is provided.

1.5. Content

In Section 2 we consider the situation where the behavior of the density near its mode is known, meaning that the constants appearing in Section 1.1 are known. We propose a method based on computing a histogram and locating the bin with maximum count, whose performance we establish. We also state a minimax lower bound for this setting, which matches the performance of our method up to a multiplicative constant. In Section 3 we consider the situation where the behavior of the density near its mode is as described in Section 1.1, but the constants introduced there are unknown. We propose a form of recursive partitioning, which we show achieves the minimax rate established in Section 2, meaning that the method does as well (up to a multiplicative constant) as an optimal method with oracle knowledge of the behavior of the density in the vicinity of its mode.

1.6. Notation

Here and elsewhere in the paper, we work with the supnorm, $\|x\| := \max_i |x_i|$ when $x = (x_1, \dots, x_d)$. This is really without loss of generality as we assume the dimension d to be fixed throughout. (For the problem of estimating a density mode in the nonparametric setting of (1)-(2) there is a standard curse of dimensionality.)

2. Known behavior near the mode: monoscale approach

In this section we assume that we know the parameters describing (in fact, constraining) the behavior of the density, specifically, the constants c_0, C_0, h_0, β in (1) and (2). (The density f and its mode x_0 remain, of course, unknown.) This assumption is rather unrealistic in practice, but it is a good place to start, with the question: *What would we do, and how well would we do, if we knew these constants?*

With knowledge of these constants, we propose a very simple method, perhaps the simplest one can think of, which effectively amounts to building a histogram and returning the bin with the largest count. The method is obviously very close to a Parzen's method. The histogram construction is apparently cruder than its smoother kernel density estimate analog, but both methods achieve the same performance rate and the histogram has the advantage of being faster to compute.

2.1. Method

The method we have in mind is very simple: It amounts to partitioning the space into bins of equal size and simply returning the location of the bin with the largest count. We represent a bin by its leftmost point, although other choices (e.g., midpoint) are possible. The bin size needs to be chosen appropriately, based on the behavior of the density near its mode, in order to achieve the minimax rate. The method is compactly described in Algorithm 1. Despite what is hinted at in the literature, the algorithm clearly runs in $O(dn)$ time if we loop over the sample rather than loop over the bins.

Algorithm 1 Mono-scale Mode Hunting

Input: point set $x_1, \ldots, x_n$ in $\mathbb{R}^d$ (assumed drawn iid from a density), bin size h
Output: a point $\hat{x}$ (meant to estimate the mode of the underlying density)

Create a sparse array of bin counts, where $\mathsf{BinCount}(k)$ for $k \in \mathbb{Z}^d$ stores the number of points in the hypercube $[kh, (k+1)h)$, with all the counts initialized to 0
For $i = 1, \ldots, n$, store $k_i = \mathsf{floor}(x_i/h)$ and update $\mathsf{BinCount}(k_i) \leftarrow \mathsf{BinCount}(k_i) + 1$
Identify $\hat{k} := \arg\max_{i=1,\ldots,n} \mathsf{BinCount}(k_i)$
Return $\hat{x} := \hat{k}h$

We quantify the performance of this method by means of the following probabilistic result.

Theorem 1. *There is a constant $A > 0$ depending on the constants in (1) and (2) such that the mode estimator returned by Algorithm 1 is within distance Ah of the true location of the mode with probability at least $1 - A\exp(-nh^{d+2\beta}/A)$.*

Proof. Since f is assumed to be compactly supported, it is enough to establish the result when the bin size h is small, and in particular we may take it substantially smaller than h_0. Also, since $1 - A\exp(-1/A) < 0$ for A large enough, it suffices to establish the result when $nh^{d+2\beta} \geq 1$, which we assume henceforth. Below $A_1, A_2, \ldots$ are constants that do not depend on n or h.

Define

$$p_k := \int_{[kh,(k+1)h)} f(x)\mathrm{d}x, \quad k \in \mathbb{Z}^d,$$

which is the probability of one draw from f falling in the bin $[kh, (k+1)h)$. Also, for a set $\mathcal{S} \subset \mathbb{R}^d$, let $N(\mathcal{S})$ denote the number of data points in $\mathcal{S}$, namely, $N(\mathcal{S}) := \#\{i : X_i \in \mathcal{S}\}$. For $\mathcal{S}$ measurable, we have that $N(\mathcal{S})$ is binomial with parameters n and $\int_{\mathcal{S}} f(x)\mathrm{d}x$. We also let N_k be short for $N([kh, (k+1)h))$, which is the count for bin k.

At the mode. First, let's consider what happens at the mode. Let $k_0 = \mathsf{floor}(x_0/h)$ so that $[k_0h, (k_0+1)h)$ is the bin that contains the mode. Based on (1) and the fact that $\|x - x_0\| \leq h \leq h_0$ for all x in that bin, we have

$$p_{k_0} = \int_{[k_0h,(k_0+1)h)} f(x)\mathrm{d}x$$

$$
\begin{aligned}
&\geq \int_{[k_0 h,(k_0+1)h)} (f(x_0) - C_0 \|x - x_0\|^\beta) \mathrm{d}x \\
&= f(x_0) h^d - C_0 \int_{[k_0 h,(k_0+1)h)} \|x - x_0\|^\beta \mathrm{d}x \\
&\geq f(x_0) h^d - C_0 \int_{[k_0 h,(k_0+1)h)} \|x - k_0 h\|^\beta \mathrm{d}x \\
&= f(x_0) h^d - C_1 h^{d+\beta}, \quad \text{where } C_1 := C_0 \int_{[0,1)^d} \|x\|^\beta \mathrm{d}x.
\end{aligned}
$$

Hence, N_{k_0} is stochastically larger than the binomial distribution with parameters n and $p'_{k_0} := f(x_0)h^d - C_1 h^{d+\beta}$, assuming (as we do) that h is small enough that $p'_{k_0} > 0$. Applying Bernstein's inequality, we thus establish

$$
\mathbb{P}\left(N_{k_0} > n p'_{k_0} - s\sqrt{n f(x_0) h^d}\right) \geq 1 - \exp\left(-\tfrac{1}{4}(s^2 \wedge s\sqrt{n f(x_0) h^d})\right),
$$

when h is small enough that $f(x_0)h^d \leq 1/2$. Here by choosing s such that $s\sqrt{n f(x_0) h^d} = C_1 h^{d+\beta} n$, and given that we assume that $nh^{d+2\beta} \geq 1$, we find that

$$
N_{k_0} > \tau := n f(x_0) h^d - 2 n C_1 h^{d+\beta}, \tag{4}
$$

with probability at least $1 - \exp(-nh^{d+2\beta}/A_0)$.

Away from the mode. We now turn to a bin away from the bin containing the mode. Based on (1)-(2), we have

$$
\begin{aligned}
p_k &= \int_{[kh,(k+1)h)} f(x) \mathrm{d}x \\
&\leq \int_{[kh,(k+1)h)} \big(f(x_0) - c_0(\|x - x_0\| \wedge h_0)^\beta\big) \mathrm{d}x \\
&= f(x_0) h^d - c_0 \int_{[kh,(k+1)h)} (\|x - x_0\| \wedge h_0)^\beta \mathrm{d}x \\
&\leq f(x_0) h^d - c_0((\|k - k_0\| - 2) \wedge (h_0/h))^\beta h^{d+\beta},
\end{aligned}
$$

by the triangle inequality. For $q \geq 1$ integer, define $\mathcal{K}_q := \{q \in \mathbb{Z}^d : \|k - k_0\| = q + 2\}$, and note that

$$
p_k \leq f(x_0) h^d - c_0(q^\beta \wedge (h_0/h)^\beta) h^{d+\beta}, \quad \forall k \in \mathcal{K}_q,
$$

and that $\mathcal{K}_q$ has cardinality $|\mathcal{K}_q| \leq A_1 q^{d-1}$. We assume henceforth that h is small enough that $c_0(h_0/h)^\beta \geq 4C_1$ and restrict ourselves to $q \geq q_0$ where $q_0 \geq 2$ is an integer large enough that $c_0 q_0^\beta \geq 4C_1$. We now bound the probability that $\hat{k}$ belongs to some $\mathcal{K}_q, q \geq q_0$. In view of (4), we only need to look at the event

$$
\max_{q \geq q_0} \max_{k \in \mathcal{K}_q} N_k > \tau.
$$

Note that we may restrict our attention to $q \leq A_2/h$, since $p_k = 0$ when kh is sufficiently large by the fact that f has compact support. We may assume that $A_2 \geq h_0$. Therefore, take $q \leq A_2/h$ so that $p_k \leq f(x_0)h^d - c_0 q^\beta h^{d+\beta}$. Then, for $k \in \mathcal{K}_q$, since $N_k \sim \mathrm{Bin}(n, p_k)$, using Bernstein's inequality, we derive

$$\begin{aligned}
\mathbb{P}(N_k > \tau) &= \mathbb{P}\left(N_k > np_k + \tau - np_k\right) \\
&\leq \mathbb{P}\left(N_k - np_k > n(c_0 q^\beta - 2C_1)h^{d+\beta}\right) \\
&\leq \mathbb{P}\left(N_k - np_k > n\tfrac{1}{2}c_0 q^\beta h^{d+\beta}\right) \\
&\leq \exp\left(-\frac{\frac{1}{2}(n\frac{1}{2}c_0 q^\beta h^{d+\beta})^2}{np_k(1-p_k) + \frac{1}{3}(n\frac{1}{2}c_0 q^\beta h^{d+\beta})}\right) \\
&\leq \exp\left(-\frac{\frac{1}{2}(n\frac{1}{2}c_0 q^\beta h^{d+\beta})^2}{nf(x_0)h^d + \frac{1}{3}(n\frac{1}{2}c_0 (h_0/h)^\beta h^{d+\beta})}\right) \\
&\leq \exp\left(-q^{2\beta} nh^{d+2\beta}/A_3\right).
\end{aligned}$$

Using the union bound, we thus obtain

$$\begin{aligned}
\mathbb{P}\left(\max_{q\geq q_0}\max_{k\in\mathcal{K}_q} N_k > \tau\right) &\leq \sum_{q=q_0}^{\mathsf{floor}(A_2/h)} A_1 q^{d-1} \exp\left(-q^{2\beta} nh^{d+2\beta}/A_3\right) \\
&\leq \sum_{q=q_0}^{\infty} A_1 q^{d-1} \exp\left(-q^{2\beta} nh^{d+2\beta}/A_3\right) \\
&\leq \sum_{q=q_0}^{\infty} \exp\left(-q^{2\beta} nh^{d+2\beta}/A_4\right) \\
&\leq \int_{q_0-1}^{\infty} \exp\left(-u^{2\beta} nh^{d+2\beta}/A_4\right)\mathrm{d}u \\
&\leq \frac{\exp\left(-(q_0-1)^{2\beta} nh^{d+2\beta}/A_4\right)}{2\beta(q_0-1)^{2\beta-1} nh^{d+2\beta}/A_4} \\
&\leq A_5 \exp\left(-q_0^{2\beta} nh^{d+2\beta}/A_5\right),
\end{aligned}$$

using the fact that $q_0 \geq 2$ and $nh^{d+2\beta} \geq 1$ multiple times. The integral was bounded using integration by parts.

We thus have that

$$N_{k_0} > \tau \geq \max_{q\geq q_0}\max_{k\in\mathcal{K}_q} N_k$$

with probability at least

$$1 - \exp(-nh^{d+2\beta}/A_0) - A_5 \exp\left(-q_0^{2\beta} nh^{d+2\beta}/A_5\right),$$

and from this we conclude. □

2.2. Information bound

Based on the performance bound established in Theorem 1, we can say that Algorithm 1 achieves the rate $O(n^{-1/(d+2\beta)})$, that is,

$$\sup_{f \in \mathcal{F}_\beta} \mathbb{E}_f \, \|\hat{x} - x_f\| = O(n^{-1/(d+2\beta)}), \tag{5}$$

where $\hat{x}$ is the output of Algorithm 1, $x_f = x_0$ is the mode of f, and $\mathcal{F}_\beta \equiv \mathcal{F}_\beta(c_0, C_0, h_0)$ is the class of density functions satisfying the properties (1)-(2)-(3). It turns out that this rate is best possible in a minimax sense. This was already known for the exponent $\beta = 2$ (Donoho and Liu, 1991) — see also (Tsybakov, 1990), where the assumed conditions are a little different. We complete the picture by establishing this as the minimax rate for any value of $\beta > 0$.

Theorem 2. *There is a constant A and two densities satisfying the basic properties (1)-(2)-(3) with modes separated by $n^{-1/(d+2\beta)}/A$ that cannot be distinguished with more accuracy than a probability of error of $1/5$, based on a sample of size n.*

Proof. The proof is based on Le Cam's two-point prior argument for which a standard reference is (Tsybakov, 2009, Sec 2.2-2.4). The idea is to craft two densities, both satisfying the basic properties (1)-(2)-(3), that are impossible to distinguish with a high degree of certainty based on a sample of size n and whose modes are on the order of $\asymp n^{-1/(d+2\beta)}$ apart. These densities are denoted by f_1 and f_2 below.

Let f_1 be a density on $\mathbb{R}^d$, compactly supported, symmetric about the origin (i.e., even as a function), strictly unimodal (and therefore with a unique mode at the origin), and such that $f_1(t) = 1 - \|t\|^\beta$ in a neighborhood of the origin. This implies that there exists $h_0 \in (0,1)$ such that $f_1(t) = 1 - \|t\|^\beta$ if $\|t\| \le h_0$, and $f_1(t) \le 1 - h_0^\beta$ if $\|t\| \ge h_0$. Clearly, f_1 satisfies the basic properties (1), (2), and (3).

Denote the origin by $\underline{0}$ and let $\underline{h} = (h, \cdots, h) \in \mathbb{R}^d$ for any $h > 0$. We consider $h \le h_0$ below. Define f_2 on $\mathbb{R}^d$ as follows

$$f_2(t) = \begin{cases} f_1(t), & \text{if } t \in \mathbb{R}^d \setminus (-\underline{h}, \underline{h}), \\ 1 - h^\beta, & \text{if } t \in (-\underline{h}, \underline{h}) \setminus (\underline{0}, \underline{h}), \\ 1 + (2^d - 1)h^\beta - 2^{d+\beta}\|t - \underline{h}/2\|^\beta, & \text{if } t \in (\underline{0}, \underline{h}). \end{cases}$$

See Figure 1 for an illustration.

Notice that we can also write $f_2(t) = f_1(t) + g(t)$, where

$$\begin{aligned} g(t) =& (\|t\|^\beta - h^\beta)\, 1\{t \in (-\underline{h}, \underline{h}) \setminus (\underline{0}, \underline{h})\} \\ &+ [\|t\|^\beta + (2^d - 1)h^\beta - 2^{d+\beta}\|t - \underline{h}/2\|^\beta]\, 1\{t \in (\underline{0}, \underline{h})\}. \end{aligned}$$

Here f_2 is indeed a density function because $f_2 \ge 0$ and $\int g(x)dx = 0$ using the fact that

$$\int_{(-\underline{h}, \underline{h})} \|x\|^\beta \mathrm{d}x = \int_0^h s^\beta (2d)(2s)^{d-1} ds = \frac{d}{d+\beta} 2^d h^{d+\beta}.$$

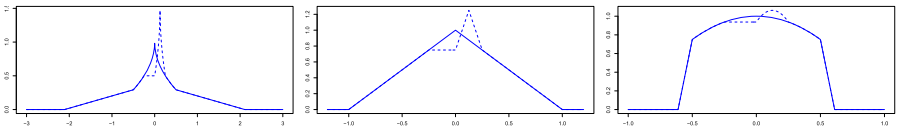

FIG 1. *Examples of pairs of functions f_1 (solid) and f_2 (dashed) for $\beta = 1/2$, $\beta = 1$, and $\beta = 2$.*

Observing that f_2 has a unique mode at $\underline{h}/2$, below we show that, for all t such that $\|t - \underline{h}/2\| \le h_0 - h/2$,

$$f_2(\underline{h}/2) - 2^{d+\beta}\|t - \underline{h}/2\|^\beta \le f_2(t) \le f_2(\underline{h}/2) - 2^{-\beta}\|t - \underline{h}/2\|^\beta. \tag{6}$$

It is easy to see that (6) holds for $t \in (-\underline{h}, \underline{h})$. For $t \in [\underline{h} - \underline{h_0}, \underline{h_0}] \setminus (-\underline{h}, \underline{h})$, we have

$$\begin{aligned}
&f_2(\underline{h}/2) - 2^{d+\beta}\|t - \underline{h}/2\|^\beta \\
&= (1 - \|2t - \underline{h}\|^\beta) + (2^d - 1)(h^\beta - \|2t - \underline{h}\|^\beta) \\
&\le 1 - \|t\|^\beta = f_2(t),
\end{aligned}$$

and

$$\begin{aligned}
&f_2(\underline{h}/2) - 2^{-\beta}\|t - \underline{h}/2\|^\beta \\
&= 1 + (2^d - 1)h^\beta - 2^{-\beta}\|t - \underline{h}/2\|^\beta \\
&\ge 1 + (2^d - 1)h^\beta - [(h/2)^\beta + \|t\|^\beta] \\
&\ge 1 - \|t\|^\beta = f_2(t).
\end{aligned}$$

The last calculation can also be used to show that $f_2(t) \le f_2(\underline{h}/2) - 2^{-\beta}(h_0 - h/2)^\beta$ for all t such that $\|t - \underline{h}/2\| \ge h_0 - h/2$. Hence f_2 also satisfies the properties (1), (2), and (3).

Define $\chi^2(f_2, f_1) := \int f_2^2/f_1 - 1$, which is sometimes called the chi-squared divergence of f_2 with respect to f_1. According to (Tsybakov, 2009, Sec 2.2-2.4), to conclude it suffices to prove that $n\chi^2(f_2, f_1)$ becomes arbitrarily small when $nh^{d+2\beta}$ is small enough. We prove this by showing below that $\chi^2(f_2, f_1) = O(h^{d+2\beta})$. Indeed, elementary calculations yield

$$\begin{aligned}
&\int \frac{f_2(x)^2}{f_1(x)}\mathrm{d}x - 1 \\
&= \int \frac{g(x)^2}{f_1(x)}\mathrm{d}x \\
&\le \int_{(-\underline{h},\underline{h})} \frac{2(\|t\|^\beta - h^\beta)^2}{f_1(x)}\mathrm{d}x + \int_{(\underline{0},\underline{h})} \frac{2[(2^d - 1)h^\beta - 2^{d+\beta}\|t - \underline{h}/2\|^\beta]^2}{f_1(x)}\mathrm{d}x \\
&\le \frac{2^{d+1}h^{d+2\beta}}{1 - h_0^\beta} + \int_{(\underline{0},\underline{h})} \frac{2^{2(d+\beta)+1}[h^\beta + \|t - \underline{h}/2\|^\beta]^2}{1 - h_0^\beta}\mathrm{d}x
\end{aligned}$$

$$\leq \frac{2^{2(d+\beta+3)}}{1-h_0^\beta} h^{d+2\beta},$$

and from this we conclude. □

Using (Tsybakov, 2009, Sec 2.2-2.4), it is straightforward to get

$$\liminf_{n\to\infty} \inf_{\hat{x}_n} \sup_{f\in\mathcal{F}_\beta} n^{1/(d+2\beta)}\, \mathbb{E}_f \|\hat{x}_n - x_f\| > 0,$$

where $\hat{x}_n$ is an arbitrary mode estimator, and $\mathcal{F}_\beta$ is defined right below (5). Theorem 1 and Theorem 2, together, establish $n^{-1/(d+2\beta)}$ as the minimax rate for estimating the mode under the conditions (1)-(2)-(3) — where the emphasis should be on the first one. This extends the result of Donoho and Liu (1991), who proved this for $\beta = 2$ in dimension $d = 1$. (The method they studied and showed to be minimax was none other than Parzen's method with a proper choice of bandwidth.) It turns out that this is the same rate as under the more restrictive assumption that the density is twice differentiable with bounded second derivative in the vicinity of the mode and with negative definite Hessian at the mode. This was established by Tsybakov (1990), who went further: If the density is Hölder-α with $\alpha \geq 2$ near its mode, and the Hessian there is negative definite, then the minimax rate is $n^{-(\alpha-1)/(d+2\alpha)}$, and is achieved by a gradient ascent method proposed in the same paper. This rate is faster than what it is in our setting which, under the same assumptions[1], still corresponds to $\beta = 2$. This is simply due to the ability to estimate the underlying density to higher precision — which Tsybakov does by using a kernel of appropriate order. (Tsybakov shows that the rate achieved by the estimator he proposes is minimax rate-optimal in the context that he considers.)

Computational complexity We thus have established that Algorithm 1, which runs in linear time, is minimax rate optimal when its tuning parameter (the bin size h) is properly chosen. Is it possible, however, to do even better in the sense of designing an algorithm that runs in sublinear time that also achieves the minimax estimation rate? The answer is 'No', and this is general: In a very broad sense, it is not possible to achieve a minimax rate in sublinear time in an estimation problem where, as is the case here, that rate is a negative power of the sample size (perhaps with some poly-logarithmic multiplicative factor). See the Appendix for details.

3. Unknown behavior near the mode: multiscale approach

Choosing the bin size correctly in Algorithm 1 is very important. It is completely analogous to choosing the bandwidth to build a histogram or to perform kernel density estimation. All the methods we are aware of necessitate the tuning of

[1] The settings — ours and Tsybakov's — only intersect at $\beta = 2$, since Tsybakov assumes that the Hessian is non-degenerate at the mode.

parameters whose optimal value, as in our case, depends on the behavior of the density near its mode.

In the special situation where the density is twice differentiable everywhere and has a negative definite Hessian at the mode — which necessarily forces $\beta = 2$ — Parzen's estimate with bandwidth chosen by cross-validation appears to achieve the minimax rate because *(i)* the optimal choice of bandwidth is the same, in order of magnitude, for the problem of density estimation and the problem of mode estimation; and *(ii)* a choice of bandwidth based on cross-validation achieves the optimal rate for the problem of density estimation as established by Hall (1983) and Stone (1984).

We propose a multiscale method that is able, under some conditions, to zoom in on the mode without assuming much of the underlying density and achieve the minimax error rate. Moreover, the method still operates in (expected) linear time. The method, in principle, still depends on a couple of parameters, but these can be chosen with much less knowledge of the underlying density. *And by letting these parameters diverge to infinity arbitrarily slowly, the method is, in effect, parameter-free.*

We are not aware of any method that is able to choose these tuning parameters automatically while achieving the minimax performance rate, except for that of Klemelä (2005). In that paper, the general approach advocated by Lepski for selecting tuning parameters is implemented and shown to yield a choice of bandwidth which leads to an *adaptive* minimax rate. Indeed, the paper also derives minimax rates for when the density smoothness at the mode is unknown, and these rates are different from those when the smoothness at the mode is known: there is a price to pay. Under the looser conditions that we operate under, it turns out that there is no price to pay.

3.1. Method

When the parameters in (1)-(2), in particular the exponent β, are unknown, we adopt a recursive partitioning approach. The resulting method is described in Algorithm 2 where the bin counts are implicitly computed as in Algorithm 1.

Algorithm 2 Multi-scale Mode Hunting

Input: point set $x_1, \ldots, x_n$ in $\mathbb{R}^d$, scale multiplier $b \geq 2$, margin $\kappa \geq 0$
Output: a point $\hat{x}$ (meant to estimate the mode of the underlying density)

Define the finest scale $s_{\max} := \mathsf{floor}(\log(n)/d \log b)$
Initialize the active set to be $I_{\text{active}} \leftarrow \{1, \ldots, n\}$
For $s = 1, \ldots, s_{\max}$
For $k \in \mathbb{Z}^d$ identify $I(k, s) = \{i \in I_{\text{active}} : x_i \in [kb^{-s}, (k+1)b^{-s})\}$
EndFor
Identify $\hat{k}(s) := \arg\max_k \#I(k, s)$
Update $I_{\text{active}} \leftarrow \bigcup\{I(k, s) : \|k - \hat{k}(s)\| \leq \kappa\}$
EndFor
Return $\hat{x} := \hat{k}(s_{\max}) b^{-s_{\max}}$

Proposition 1. *Algorithm 2 runs in linear expected time.*

Proof. At each scale, the main computational task is to identify the bins where the sample points that remain active reside. We saw when discussing the computational complexity of Algorithm 1 that doing this can be done in time $O(dm)$ if there are m active points. At scale $s = 1$, all points are active, and so the resulting time complexity is $O(dn)$. At scale $s > 1$, we expect at most $n\big(f(x_0)[(2\kappa+1)b^{-s+1}]^d \wedge 1\big) \asymp n\big((\bar\kappa b^{-s})^d \wedge 1\big)$ active points to process, where $\bar\kappa := \kappa \vee 1$, resulting in a complexity of $O(dn(\bar\kappa^d b^{-ds} \wedge 1))$ at that scale. Summing these expected computational costs over $s = 1, \dots, s_{\max}$ yields an overall computational cost bounded by a constant multiple of

$$dn + \sum_{s \geq 1} dn(\bar\kappa^d b^{-ds} \wedge 1) \asymp dn\big(\tfrac{\log \bar\kappa}{\log b} + 1\big).$$

(We have assumed that d is constant, as there is a real curse of dimensionality in the context that interests us here, but we carried it throughout these computations to display its influence, which can be seen to be rather benign.) □

Theorem 3. *There is a constant $A > 0$ depending on the constants in (1) and (2) such that the mode estimator returned by Algorithm 2 is within distance $tn^{-1/(d+2\beta)}$ of the true location of the mode with probability at least $1 - (A/\log b)(\kappa b^2/t)^{d+2\beta} \exp(-(t/\kappa b^2)^{d+2\beta}/A)$ whenever $t \geq 1$, $b \geq A$, and $\kappa \geq A$, as well as $(\kappa+1)b^2/t \leq n^{2\beta/d(d+2\beta)}$.*

The statement is a bit complicated but the core message is simple: if t, b, and κ are understood as remaining constant while the sample size becomes large, the estimator is within distance $tn^{-1/(d+2\beta)}$ with probability at least $1 - A' \exp(-t^{d+2\beta}/A')$ when $t \geq 1$, where this time A' also depends on b and κ. Thus, the same result as in (5) still holds if $\hat x$ is the output of Algorithm 2.

Proof. First, by a simple modification of the arguments underlying Theorem 1, there is a constant A_0 which depends on the constants (1) and (2) such that at scale s, whenever $n(b^{-s})^{d+2\beta} \geq 1$ and $b \geq A_0$ as well as $\kappa \geq A_0$,

$$\|\hat k(s) - k_0(s)\| \leq A_0, \quad k_0(s) := \mathsf{floor}(x_0 b^s),$$

with probability at least $1 - A_0 \exp(-n(b^{-s})^{d+2\beta}/A_0)$. This comes from considering b^{-s} as playing the role of h and κ as playing the role of q in the proof of Theorem 1, and realizing that restricting the density to $[(\hat k(s-1)-\kappa)h, (\hat k(s-1)+\kappa)h)$ does not have any substantial effect. In what follows, we assume that b and κ are indeed sufficiently large that $b \geq A_0$ and $\kappa \geq A_0$.

Second, for $\hat x$ to be within distance δ of x_0 it suffices that, $\|\hat k(s) - k_0(s)\| \leq \kappa$ for some s satisfying $(\kappa+1)b^{-s} \leq \delta$. This is simply because, by design,

$$\|\hat x - \hat k(s) b^{-s}\| \leq \kappa b^{-s}, \quad \forall s = 1, \dots, s_{\max},$$

and by definition $\|x_0 - k_0(s)b^{-s}\| \leq b^{-s}$. Therefore, $\|\hat x - x_0\| \leq \delta$ when $\|\hat k(s) - k_0(s)\| \leq \kappa$ for $s = 1, \dots, \bar s(\delta) := \mathsf{ceiling}(\log_b((\kappa+1)/\delta))$, where $\log_b(x) := \log(x)/\log(b)$ is the logarithm in base b.

With these preliminaries, we now proceed by bounding the probability that $\|\hat{k}(s)-k_0(s)\| \leq \kappa$ for $s = 1, \ldots, \bar{s}(\delta)$ with δ chosen as $\delta := tn^{-1/(d+2\beta)}$ where $t > 0$. Note that $\bar{s}(\delta) \leq s_{\max}$ as our assumptions include $(\kappa+1)b^2/t \leq n^{2\beta/d(d+2\beta)}$. Using the union bound, this probability is

$$\begin{aligned}
&\geq 1 - \sum_{s=1}^{\bar{s}(\delta)} A_0 \exp(-n(b^{-s})^{d+2\beta}/A_0) \\
&\geq 1 - \int_1^{\bar{s}(\delta)+1} A_0 \exp(-n(b^{-s})^{d+2\beta}/A_0)\mathrm{d}s \\
&= 1 - \int_{nb^{-(d+2\beta)(\bar{s}(\delta)+1)}/A_0}^{nb^{-(d+2\beta)}/A_0} A_0 \exp(-u)\frac{1}{u(d+2\beta)\log b}\mathrm{d}u \\
&\geq 1 - \frac{A_0}{(d+2\beta)\log b}\frac{\exp(-nb^{-(d+2\beta)(\bar{s}(\delta)+1)}/A_0)}{nb^{-(d+2\beta)(\bar{s}(\delta)+1)}/A_0}.
\end{aligned}$$

Some elementary calculations give that

$$nb^{-(d+2\beta)(\bar{s}(\delta)+1)} \geq (t/(\kappa+1)b^2)^{d+2\beta},$$

and from this we conclude. □

We have thus proved that Algorithm 2 achieves the minimax rate without knowledge of the exponent β driving the behavior of the density in the vicinity of its mode as prescribed in (1). The algorithm can thus be said to be 'adaptive' in that sense. This is in contrast with the more structured situation that Tsybakov (1990) considered. In that situation, Klemelä (2005) showed that there is a cost to adaptation, although a small one: a poly-logarithmic factor; and Klemelä proposed a Lepski-type method that he shows to be adaptive minimax rate-optimal. The fact that there is a cost to adaptation in this other context and not in ours may be due to the fact that, to be competitive there, a method may be forced to rely on an accurate estimate of the density (at least around the mode) to take advantage of the assumed smoothness, and there is a limit to how well the density can be estimated when its smoothness is unknown.

4. Numerical experiments

We performed some basic numerical experiments to probe our theory. We present the result of these experiments below, subdivided into $d = 1$ and $d = 2$ settings. Three cases are studied: $\beta = 1/2$, $\beta = 1$ and $\beta = 2$. We compare our main method, Algorithm 2, with Algorithm 1 with bin size chosen by cross-validation (Rudemo, 1982). In fact, to make the comparison as fair and meaningful as we can, Algorithm 1 goes through the exact *same* histograms as Algorithm 2.

A sensitivity analysis shows that the dependence on the parameters b and κ is mild. Nevertheless, in an effort to make Algorithm 2 fully automatic, we implement a stability approach not unlike that advocated for the choice of tuning

parameters in clustering algorithms (von Luxburg, 2009), and also akin to the Lepski method proposed in the context of mode estimation by Klemelä (2005). In detail, we draw multiple subsamples independently and apply Algorithm 2 to each of these subsamples to get mode estimates. We do so for multiple choices of b and κ, and each time compute the sum of pairwise distances between the estimated modes, and we choose the values of these tuning parameters that minimize that quantity.

Below, in particular in some figures, we let x denote the mode and $\hat{x}$ its estimate computed based on Algorithm 1 (with CV) or on Algorithm 2. Note that x will be at the origin in all our experiments.

4.1. One-dimension setting

We start with the setting where $d = 1$, and consider the following simple but emblematic examples of densities:[2]

$$\text{Case } \beta = \frac{1}{2}: \ f(t) = \begin{cases} 1 - \sqrt{|t|} & |t| \le 0.5, \\ \frac{5\sqrt{2}-4}{8} - \frac{9\sqrt{2}-12}{4}|t| & 0.5 \le |t| < \frac{4\sqrt{2}+7}{6}, \\ 0 & \text{otherwise;} \end{cases} \tag{7}$$

$$\text{Case } \beta = 1: \ f(t) = \begin{cases} 1 - |t| & |t| \le 1, \\ 0 & \text{otherwise;} \end{cases} \tag{8}$$

$$\text{Case } \beta = 2: \ f(t) = \begin{cases} 1 - t^2 & |t| \le 0.5, \\ \frac{99}{24} - \frac{27}{4}|t| & 0.5 \le |t| < \frac{11}{18}, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Sensitivity analysis We first perform a sensitivity analysis to the tuning parameters b and κ in Algorithm 2. We limit ourselves to the density (8). We use a sample size of $n = 1000$ and 500 repeats. The results are shown in Figure 2, where we report the mean of $|\hat{x} - x|/n^{-1/(d+2\beta)}$ (over the repeats) as a function of b and κ. In addition to tracking how much the input of Algorithm 2 depends on these parameters, we do the same for Algorithm 1. We can see that the algorithms are not overly sensitive to the choice of these tuning parameters. That said, the performance deteriorates as b increases, but this is to be expected since the larger b is, the coarser the histograms — but the faster the procedures. We also note an upward trend as κ increases, meaning that Algorithm 2 will lose some performance if the choice of κ is too conservative.

Comparison We then compare Algorithm 1 and Algorithm 2. We do so under each of the densities above — (7), (8), and (9) — and different values of the

[2] As a reviewer pointed out, the mode is at the origin, which may be seen as a little artificial as this is exactly a bin corner in the histogram constructions that we consider. We agree, although it is clear that this should not change how the experiments presented here should be interpreted.

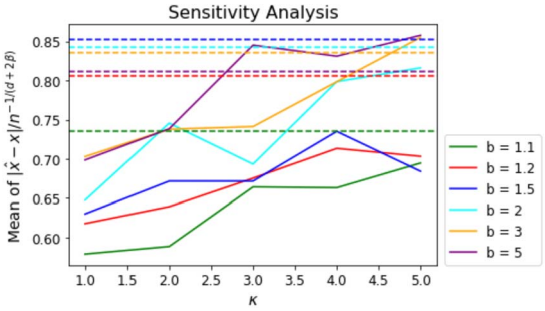

FIG 2. *The means of* $|\hat{x} - x|/n^{-1/(d+2\beta)}$ *over 500 repeats for Algorithm 1 (dashed) and Algorithm 2 (solid) for different values of* b *and* κ*. (Recall that Algorithm 1 does not depend on* κ*.) The density is given in (8) (so that* $\beta = 1$*) and the sample size is* $n = 1000$*.*

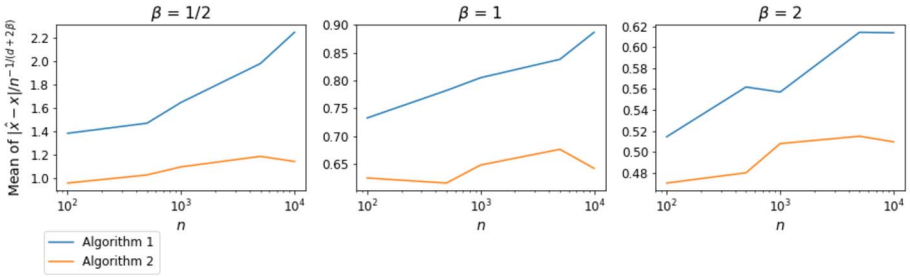

FIG 3. *The means of* $|\hat{x} - x|/n^{-1/(d+2\beta)}$ *over 500 repeats for Algorithm 1 (blue) and Algorithm 2 (orange) for different values of* b *and* κ*. (Recall that Algorithm 1 does not depend on* κ*.) The density is given in (8) (so that* $\beta = 1$*) and the sample size is* $n = 1000$*.*

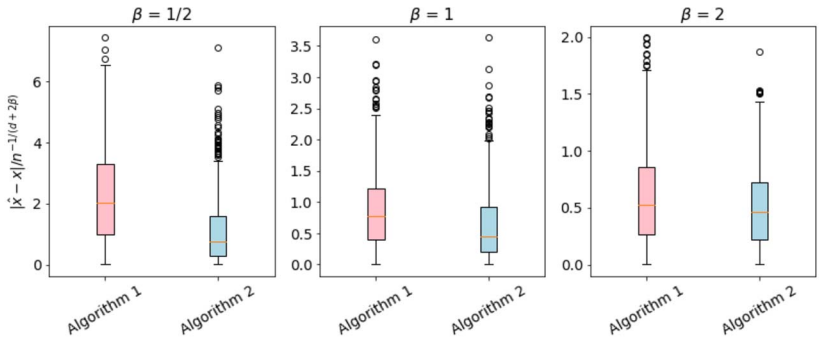

FIG 4. *Boxplots of* $|\hat{x} - x|/n^{-1/(d+2\beta)}$ *based on* 500 *repeats. Here the sample size is* $n = 10000$*.*

sample size $n \in \{100, 500, 1000, 5000, 10000\}$. We again use 500 repeats. The results are reported in Figure 3. Algorithm 2 performs uniformly better than Algorithm 1. We also show the boxplots for the situation where $n = 10000$ in Figure 4.

4.2. Two-dimension setting

We now move to the setting where $d = 2$, and consider the following three examples of densities:[3]

$$\text{Case } \beta = \frac{1}{2}: \quad f(t) = \begin{cases} 1 - \sqrt{\|t\|} & \|t\| \le 0.5, \\ 1 - \frac{\sqrt{2}}{2} + \frac{5(2-\sqrt{2})}{\sqrt{465-240\sqrt{2}}-15}(1 - 2\|t\|) & 0.5 \le |t| < \frac{\sqrt{465+240\sqrt{2}}-5}{20}, \\ 0 & \text{otherwise;} \end{cases} \tag{10}$$

$$\text{Case } \beta = 1: \quad f(t) = \begin{cases} \frac{3}{4}(1 - \|t\|) & \|t\| \le 1, \\ 0 & \text{otherwise;} \end{cases} \tag{11}$$

$$\text{Case } \beta = 2: \quad f(t) = \begin{cases} 1 - \|t\|^2 & \|t\| \le 0.5, \\ \frac{3\sqrt{11}+12}{4} - \frac{3\sqrt{11}+9}{2}\|t\| & 0.5 \le \|t\| < \frac{\sqrt{11}-1}{4}, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

Sensitivity analysis As in the setting where $d = 1$, we perform a similar sensitivity analysis focusing on the density (11) (so that $\beta = 1$) from which we draw $n = 1000$ observations. The number of repeats is 500 and we report the mean of $|\hat{x} - x|/n^{-1/(d+2\beta)}$ with different values of b and κ in Figure 5.

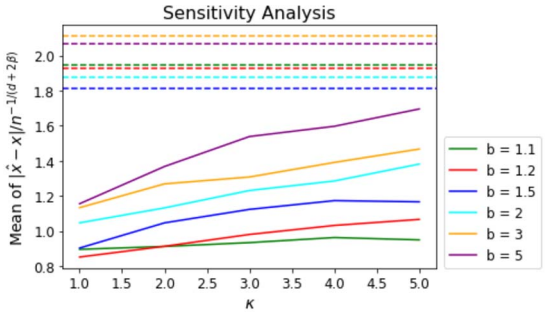

FIG 5. *The means of $|\hat{x} - x|/n^{-1/(d+2\beta)}$ over 500 repeats for Algorithm 1 (dashed) and Algorithm 2 (solid) for different values of b and κ. (Recall that Algorithm 1 does not depend on κ.) The density is given in (11) (so that $\beta = 1$) and the sample size is $n = 1000$.*

Comparison We then compare on Algorithm 1 and Algorithm 2 on samples of different sizes $n \in \{100, 500, 1000, 5000, 10000\}$ from the three densities displayed above — (10), (11), and (12). Each setting is repeated 500 times. The average errors are reported in Figure 6 and some boxplots are given in Figure 7 limited to the case where $n = 10000$. Again, at least in these simulations, Algorithm 2 is clearly superior to Algorithm 1.

[3] As in the one-dimensional case, the mode is at the origin, which is again a little artificial.

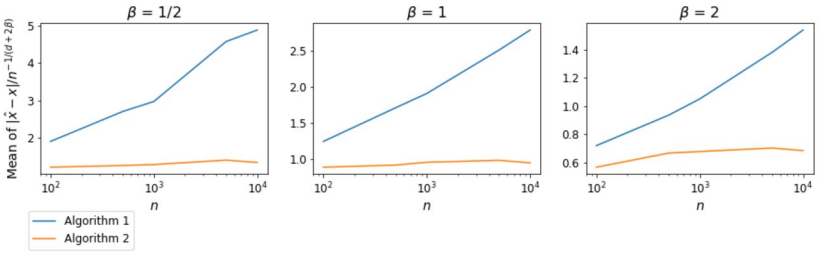

FIG 6. *The means of* $|\hat{x} - x|/n^{-1/(d+2\beta)}$ *based on* 500 *repeats. Here the sample size is* $n = 10000$.

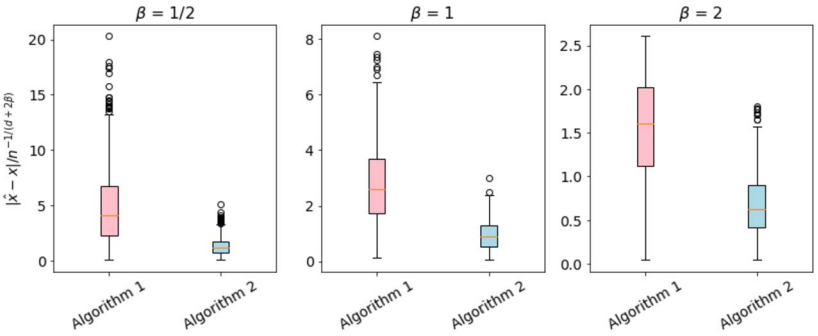

FIG 7. *Boxplots of* $|\hat{x}-x|/n^{-1/(d+2\beta)}$ *based on* 500 *repeats. Here the sample size is* $n = 10000$.

Appendix: Minimax rates in sublinear time

We establish here that, under rather general conditions, *achieving a minimax estimation rate in sublinear time is impossible when the minimax rate converges to zero faster than some negative power of the sample size* — a situation that is quite general indeed, although there are exceptions such as deconvolution problems (Fan, 1991). The fundamental idea is quite straightforward and is based on the fact that a sublinear-time algorithm is not even able to 'look' at all observations and thus effectively operates as if on a sample of size sublinear in the available sample size, so that its performance is under the purview of the minimax rate corresponding to that smaller sample size. The remainder of this section is simply devoted to formalizing this discussion.

Remark 2. We focus here on a minimax rate based on the sample size and not other parameters of the problem such as the dimension. Since there is a real curse of dimensionality for the problem of estimating of a density mode, we have assumed the dimension to be fixed throughout, but we do believe that no algorithm which is sublinear in the dimension can achieve the minimax rate.

Consider a general statistical problem where we have a family of distributions $\{P_\theta : \theta \in \Theta\}$ on some measurable space $\mathbb{X}$. The dataset consists in a sample,

$X_1, \dots, X_n$, drawn iid from a distribution in that family, say P_{θ_0}. Given some dissimilarity measure on the space, namely $\mathcal{L} : \mathbb{X} \times \mathbb{X} \to \mathbb{R}_+$ measurable, which plays the role of loss function, the risk of an estimator $\hat{\theta} = \varphi(X_1, \dots, X_n)$ at $\theta \in \Theta$ is defined as

$$\mathsf{risk}_n(\varphi, \theta) := \mathbb{E}_\theta \left[\mathcal{L}(\varphi(X_1, \dots, X_n), \theta) \right],$$

and its worst-case risk is the supremum of that over the entire parameter space,

$$\mathsf{risk}_n(\varphi) := \sup_{\theta \in \Theta} \mathsf{risk}_n(\varphi, \theta).$$

Note that φ is a (measurable) function on finite sequences of elements of $\mathbb{X}$ and $\mathbb{E}_\theta$ above is the expectation with respect to $X_1, \dots, X_n$ iid from P_θ. The minimax risk for this estimation problem is simply the infimum of this quantity over all estimators,

$$R(n) := \inf_{\varphi} \mathsf{risk}_n(\varphi).$$

We assume throughout that $R(n) < \infty$, at least for n large enough, for otherwise the setting is trivial. In that case, R is non-increasing as a real-valued function on the positive integers.

Theorem 4. *Consider a setting as described above where*

$$\limsup_{a \to \infty} \limsup_{n \to \infty} R(an)/R(n) = 0.$$

Then $\mathsf{risk}_n(\varphi) \gg R(n)$ *for any estimator* φ *which can be computed in* $o(n)$ *time when applied to a sample of size* n.

We assume below that it takes a unit of time to simply register an observation for further processing.

Proof. Let φ be such an estimator and let $b(n)$ denote the time it takes to compute φ on a sample of size n so that $b(n) = o(n)$ by assumption. (We assume that φ is not randomized in what follows, but similar arguments apply when in the situation where it relies on an exogenous source of randomness.) Assuming, without loss of generality, that φ registers the first observation first, φ is computed as follows: $\varphi(x_1, \dots, x_n) = \psi_k(x_{i_1}, \dots, x_{i_k})$ where i_1 is constant equal to 1, i_2 is a function of x_1, etc, and i_k is a function of $x_{i_1}, x_{i_2}, \dots, x_{i_{k-1}}$, and ψ_k is a function of k variables. The number of entries taken in, k, need not be constant, but ignoring some variables as needed, we may take k to be constant, and we then let ψ denote ψ_k. And given our assumption on the computational complexity of φ, necessarily, $k \le b(n)$. When applied to an iid sample, by independence, $I_2 := i_2(X_1)$ is independent of $X_2, \dots, X_n$ and so I_2 is effectively uniformly distributed on $\{2, \dots, n\}$; given $I_2 = i_2$, $I_3 := i_3(X_{i_1}, X_{i_2})$ (remember $i_1 = 1$) is independent of $\{X_i : i \notin \{i_1, i_2\}\}$ and therefore uniform in $[n] \setminus \{i_1, i_2\}$; etc; and given $I_2 = i_2, \dots, I_{k-1} = i_{k-1}$, $I_k := i_k(X_{i_1} \dots, X_{i_{k-1}})$ is independent of $\{X_i : i \notin \{i_1, \dots, i_{k-1}\}\}$ and therefore uniform in $[n] \setminus \{i_1, \dots, i_{k-1}\}$. We may therefore conclude that $\varphi(X_1, \dots, X_n)$ has the same law of $\psi(X_1, \dots, X_k)$.

Having established this, we then have

$$\begin{aligned}\mathsf{risk}_n(\varphi,\theta) &= \mathbb{E}_\theta\left[\mathcal{L}(\varphi(X_1,\ldots,X_n),\theta)\right]\\ &= \mathbb{E}_\theta\left[\mathcal{L}(\psi(X_1,\ldots,X_k),\theta)\right]\\ &= \mathsf{risk}_k(\psi,\theta), \quad \forall\theta,\end{aligned}$$

implying that

$$\mathsf{risk}_n(\varphi) \geq \mathsf{risk}_k(\psi) \geq R(k) \geq R(b(n)).$$

We then conclude with the fact that $R(b(n))/R(n) \to \infty$ due to the fact that $b(n)/n \to 0$ and our assumption on R. □

Acknowledgments

We are grateful to two anonymous referees for their comments that helped improve the paper. This work was partially supported by an NSF grant (DMS 1821154).

References

Abraham, C., Biau, G. and Cadre, B. (2003). Simple estimation of the mode of a multivariate density. *Canadian Journal of Statistics* **31** 23–34. MR1985502

Abraham, C., Biau, G. and Cadre, B. (2004). On the asymptotic properties of a simple estimate of the mode. *ESAIM: Probability and Statistics* **8** 1–11. MR2085601

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys* **4** 40–79. MR2602303

Balabdaoui, F., Rufibach, K. and Wellner, J. A. (2009). Limit distribution theory for maximum likelihood estimation of a log-concave density. *Annals of statistics* **37** 1299. MR2509075

Chacón, J. E. (2020). The modal age of statistics. *International Statistical Review* **88** 122–141. MR4088013

Chacón, J. E. and Duong, T. (2013). Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. *Electronic Journal of Statistics* **7** 499–532. MR3035264

Chernoff, H. (1964). Estimation of the mode. *Annals of the Institute of Statistical Mathematics* **16** 31–41. MR0172382

Dalenius, T. (1965). The Mode — A Neglected Statistical Parameter. *Journal of the Royal Statistical Society: Series A* **128** 110–117. MR0185720

Dasgupta, S. and Kpotufe, S. (2014). Optimal rates for k-NN density and mode estimation. *Advances in Neural Information Processing Systems* **3** 2555–2563.

Devroye, L. (1979). Recursive estimation of the mode of a multivariate density. *Canadian Journal of Statistics* **7** 159–167. MR0570537

Donoho, D. L. and Liu, R. C. (1991). Geometrizing rates of convergence, II. *Annals of Statistics* **19** 633–667. MR1105839

Doss, C. R. and Wellner, J. A. (2019). Inference for the mode of a log-concave density. *Annals of Statistics* **47** 2950–2976. MR3988778

Dümbgen, L. and Walther, G. (2008). Multiscale inference about a density. *Annals of Statistics* **36** 1758–1785. MR2435455

Duong, T., Cowling, A., Koch, I. and Wand, M. P. (2008). Feature significance for multivariate kernel density estimation. *Computational Statistics & Data Analysis* **52** 4225–4242. MR2432459

Eckle, K., Bissantz, N., Dette, H., Proksch, K. and Einecke, S. (2018). Multiscale inference for a multivariate density with applications to x-ray astronomy. *Annals of the Institute of Statistical Mathematics* **70** 647–689. MR3785711

Eddy, W. F. (1980). Optimum kernel estimators of the mode. *Annals of Statistics* **8** 870–882. MR0572631

Fan, J. (1991). On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems. *Annals of Statistics* **19** 1257–1272. MR1126324

Fukunaga, K. and Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory* **21** 32–40. MR0388638

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I. and Wasserman, L. (2016). Non-parametric inference for density modes. *Journal of the Royal Statistical Society: Series B* **78** 99–126. MR3453648

Godtliebsen, F., Marron, J. and Chaudhuri, P. (2002). Significance in scale space for bivariate density estimation. *Journal of Computational and Graphical Statistics* **11** 1–21. MR1937281

Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Annals of Statistics* **11** 1156–1174. MR0720261

Hartigan, J. A. (1975). *Clustering Algorithms*. John Wiley & Sons, Inc. MR0405726

Klemelä, J. (2005). Adaptive estimation of the mode of a multivariate density. *Journal of Nonparametric Statistics* **17** 83–105. MR2112688

Konakov, V. D. (1973). On asymptotic normality of the sample mode of multivariate distributions. *Theory of Probability and its Applications* **18** 836–842. MR0336874

Mokkadem, A. and Pelletier, M. (2003). The law of the iterated logarithm for the multivariate kernel mode estimator. *ESAIM: Probability and Statistics* **7** 1–21. MR1956072

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *Annals of Mathematical Statistics* **33** 1065–1076. MR0143282

Robertson, T. and Cryer, J. D. (1974). An iterative procedure for estimating the mode. *Journal of the American Statistical Association* **69** 1012–1016. MR0431499

Romano, J. P. (1988a). On Weak Convergence and Optimality of Kernel Density Estimates of the Mode. *Annals of Statistics* **16** 629–647. MR0947566

ROMANO, J. P. (1988b). Bootstrapping the mode. *Annals of the Institute of Statistical Mathematics* **40** 565–586. MR0964293

ROSENBLATT, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *Annals of Mathematical Statistics* **27** 832 – 837. MR0079873

RUDEMO, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics* **9** 65–78. MR0668683

RUFIBACH, K. and WALTHER, G. (2010). The block criterion for multiscale inference about a density, with applications to other multiscale problems. *Journal of Computational and Graphical Statistics* **19** 175–190. MR2654403

SAGER, T. W. (1978). Estimation of a Multivariate Mode. *Annals of Statistics* **6** 802–812. MR0491553

SAGER, T. W. (1979). An iterative method for estimating a multivariate mode and isopleth. *Journal of the American Statistical Association* **74** 329–339. MR0548023

SAMANTA, M. (1973). Nonparametric estimation of the mode of a multivariate density. *South African Statistical Journal* **7** 109–117. MR0331618

SILVERMAN, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B* **43** 97–99. MR0610384

STONE, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Annals of Statistics* **12** 1285–1297. MR0760688

TSYBAKOV, A. B. (1990). Recursive estimation of the mode of a multivariate distribution. *Problemy Peredachi Informatsii* **26** 38–45. MR1051586

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer Science+Business Media. MR2724359

VENTER, J. (1967). On Estimation of the Mode. *Annals of Mathematical Statistics* **38** 1446–1455. MR0216698

VIEU, P. (1996). A note on density mode estimation. *Statistics & Probability Letters* **26** 297–307. MR1393913

VON LUXBURG, U. (2009). Clustering Stability: An Overview. *Machine Learning* **2** 235–274.