On the Asymptotic Properties of a Simple Estimate of the Mode

Christophe ABRAHAM [a], Gérard BIAU [b,*] and Benoît CADRE [c]

[a] ENSAM-INRA, UMR Biométrie et Analyse des Systèmes,
2, Place Pierre Viala, 34060 Montpellier Cedex 1, France;
[b] Laboratoire de Statistique Théorique et Appliquée,
Université Pierre et Marie Curie – Paris VI,
Boîte 158, 175 rue du Chevaleret, 75013 Paris, France;
[c] Laboratoire de Probabilités et Statistique, Université Montpellier II,
Cc 051, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France

Abstract

We consider an estimate of the mode θ of a multivariate probability density f with support in $\mathbb{R}^d$ using a kernel estimate f_n drawn from a sample $X_1, \ldots, X_n$. The estimate θ_n is defined as any x in $\{X_1, \ldots, X_n\}$ such that $f_n(x) = \max_{i=1,\ldots,n} f_n(X_i)$. It is shown that θ_n behaves asymptotically as any maximizer $\hat{\theta}_n$ of f_n. More precisely, we prove that for any sequence $(r_n)_{n\geq 1}$ of positive real numbers such that $r_n \to \infty$ and $r_n^d \log n/n \to 0$, one has $r_n \|\theta_n - \hat{\theta}_n\| \to 0$ in probability. The asymptotic normality of θ_n follows without further work.

Index Terms **— Multivariate probability density, mode, kernel estimate, central limit theorem.**

AMS 2000 Classification**: 62G05.**

1 Introduction

The problem of estimating the mode of a probability density has received considerable attention in the literature. For a historical and mathematical survey, we refer the reader to Sager [12]. One of the most recent application of mode estimation is in unsupervised *cluster analysis*, where one tries to break a complex data set into a series of piecewise similar groups or structures. The

*Corresponding author. Email: biau@ccr.jussieu.fr .

nonparametric approach is based on the premise that groups correspond to modes of a density. The goal then is to estimate the modes and assign each observation to the "domain of attraction" of a mode. But there are many other fields where the knowledge of the mode is of great interest. For example, the estimation of contours, or isopleths, is a natural extension of the estimation of modal points.

In this paper, we consider the problem of estimating the mode θ of a multivariate unimodal probability density f with support in $\mathbb{R}^d$ from independent random variables $X_1, \ldots, X_n$ with density f. This problem has been studied by many authors, see for example Parzen [8], Konakov [5], Samanta [13], Devroye [2], Romano [10], Vieu [15], Leclerc and Pierre-Loti-Viaud [6], Mokkadem and Pelletier [7] and the references therein. Mostly, the estimate $\hat{\theta}_n$ of θ is defined as any maximizer of f_n, *i.e.*,

$$\hat{\theta}_n \in \mathrm{argmax}_{\mathbb{R}^d} f_n\,, \tag{1.1}$$

where f_n is a *kernel density estimate* (Rosenblatt [11], Parzen [8], Devroye [3]). Recall that f_n is defined for all $x \in \mathbb{R}^d$ by

$$f_n(x) = \frac{1}{nh_n^d} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h_n}\Big),$$

where $(h_n)_{n\geq 1}$ is a sequence of positive real numbers such that $h_n \to 0$ and $K : \mathbb{R}^d \to \mathbb{R}$ is an integrable function with $\displaystyle\int_{\mathbb{R}^d} K(x)\,\mathrm{d}x = 1$.

The estimate (1.1) is widely used, though it is hard to compute. Indeed, in addition to the calculation of f_n, it involves a numerical step for the computation of the argmax. As noticed by Devroye [2], classical search methods of the argmax perform satisfactorily only when f_n is sufficiently regular (continuous, unimodal, etc.) Thus, in practice, the argmax is usually computed over a finite grid. This failing is seldom discussed by authors, although it may affect the asymptotic properties of the estimate. Moreover, when the dimension of the sample space is large, or when accurate estimation is needed, the grid size (which exponentially increases with the dimension) leads to time-consuming computations. Finally, the search grid should be located around high density areas. In high dimension, this is a difficult task and the search grid usually includes low density areas.

As an attempt to remedy these problems, we proposed in a previous paper (Abraham, Biau and Cadre [1]) a concurrent estimate. Denoting by S_n the set $\{X_1, \ldots, X_n\}$, we let the estimate θ_n be defined as

$$\theta_n \in \operatorname{argmax}_{S_n} f_n \,,$$

i.e.,

$$\theta_n \in \big\{x \in S_n : f_n(x) = \max_{i=1,\ldots,n} f_n(X_i)\big\}.$$

We emphasize that the main advantage of using θ_n instead of the argmax estimate (1.1) is that the former is easily computed in a finite number of operations. Moreover, since the sample points are naturally concentrated in high density areas, the set S_n can be regarded as the most natural (random) grid for approximating the mode. As pointed out by the referees, θ_n may also be an appropriate choice for a start value of any optimization algorithm to approximate $\hat{\theta}_n$. In [1], we established, under the condition $nh_n^d/\log n \to \infty$, the strong consistency of θ_n towards θ and provided an almost sure rate of convergence without any differentiability condition on f around the mode. This rate relies on the sharpness of the density near θ, which is measured by a *peak index*. For discussion, examples and numerical illustration, we refer the reader to [1].

One question still unanswered is whether the maximization over a finite sample alters the rate of convergence of the estimate θ_n compared to that of $\hat{\theta}_n$. In the present paper, we prove that the estimates θ_n and $\hat{\theta}_n$ have the same asymptotic behavior. In Section 2, we set up notation and assumptions and provide the main results. Proofs are gathered in Section 3.

2 Notation, hypotheses and main results

2.1 Asymptotic proximity of θ_n and $\hat{\theta}_n$

Throughout the paper, we will denote by $\|.\|$ the usual Euclidean norm for matrices or vectors and by $\mathrm{H}g(x)$ the Hessian matrix at the point x of any function $g : \mathbb{R}^d \to \mathbb{R}$ twice continuously differentiable in a neighborhood of x. The notation diam S will stand for the diameter of any set $S \subset \mathbb{R}^d$, *i.e.*,

$$\operatorname{diam} S = \sup_{x,y \in S} \|x - y\|.$$

For all $\varepsilon > 0$, the *level set* $A(\varepsilon)$ defined by

$$A(\varepsilon) = \{x \in \mathbb{R}^d \,:\, f(x) \geq f(\theta) - \varepsilon\}$$

will play a crucial role. If $\stackrel{P}{\to}$ stands for the convergence in probability, we finally introduce the following hypotheses.

H1 The application f is twice continuously differentiable on a neighborhood $\mathcal{V}$ of θ and the matrix $\mathrm{H}f(\theta)$ is negative definite;

H2 The convergence diam $A(\varepsilon) \to 0$ as $\varepsilon \to 0$ holds;

H3 The sequence $\big(v_n(\hat{\theta}_n - \theta)\big)_{n\geq 1}$ is tight for some sequence $(v_n)_{n\geq 1}$ of positive real numbers with $v_n \to \infty$;

H4 The kernel K is twice continuously differentiable on $\mathbb{R}^d$ and moreover $\sup_{x\in\mathcal{V}} \|Hf_n(x) - Hf(x)\| \xrightarrow{P} 0$;

H5 One has $\sup_{x\in\mathbb{R}^d} |f_n(x) - \mathbf{E}f_n(x)| \xrightarrow{P} 0$.

Let us comment on these hypotheses. **H1** is a mild regularity assumption which is usually required to obtain rates of convergence in mode kernel estimation (see for example Parzen [8] and Romano [10]). Assumption **H2** has been introduced to avoid high density areas arbitrarily far from the mode. It can be shown that **H2** is equivalent to the classical condition

$$\sup_{x\notin\mathcal{U}} f(x) < f(\theta)$$

for any open vicinity $\mathcal{U}$ of θ. For further discussion on this condition, we refer to Abraham, Biau and Cadre [1]. **H3** is a weak assumption which is in particular true when the sequence $\big(v_n(\hat{\theta}_n - \theta)\big)_{n\geq 1}$ converges in distribution. In this respect, sufficient conditions are to be found in Romano [10] (for $d = 1$) and in Mokkadem and Pelletier [7] (for $d \geq 1$). Regarding **H4**, we refer to Silverman [14] for the univariate case and to Mokkadem and Pelletier [7] for the multivariate case. Assumption **H5** holds for example if K is of the form $K(x) = \Psi\big(\|x\|\big)$ where Ψ is a real valued function with bounded variation and $nh_n^d/\log n \to \infty$ (see Pollard [9], Theorem 37, page 34). More generally, it can be shown that **H5** holds whenever K satisfies a *covering number condition*, see Mokkadem and Pelletier [7] for a detailed discussion.

We are now ready to state the main result of the paper.

Theorem 2.1 (ASYMPTOTIC PROXIMITY OF θ_n AND $\hat{\theta}_n$) *Assume that* **H1**-**H5** *hold. For any sequence $(r_n)_{n\geq 1}$ of positive real numbers such that $r_n \to \infty$ and $r_n^d \log n/n \to 0$, we have*

$$r_n \|\theta_n - \hat{\theta}_n\| \xrightarrow{P} 0.$$

This theorem gains in interest if we realize that the very weak condition imposed on r_n allows to derive asymptotic properties of θ_n from analogous asymptotic properties of $\hat{\theta}_n$. Examples are presented in the next paragraph.

2.2 Application

The following corollary follows from Theorem 2.1 without further work. We let $\stackrel{\mathcal{D}}{\to}$ denote the convergence in distribution.

Corollary 2.1 (LIMIT LAW) *Assume that the assumptions* **H1**, **H2**, **H4** *and* **H5** *hold,* $n^{d/2-1}h_n^{d(d+2)/2}\log n \to 0$ *and*

$$\sqrt{nh_n^{d+2}}\,(\hat{\theta}_n - \theta) \stackrel{\mathcal{D}}{\to} Z,$$

where Z *is some* $\mathbb{R}^d$*-valued random variable. Then*

$$\sqrt{nh_n^{d+2}}\,(\theta_n - \theta) \stackrel{\mathcal{D}}{\to} Z.$$

The weak convergence of $\hat{\theta}_n$ to θ was first studied in the univariate framework by Parzen [8] who proved that if h_n is chosen such that $nh_n^6 \to \infty$ and $nh_n^7 \to 0$, then

$$\sqrt{nh_n^3}\,(\hat{\theta}_n - \theta) \stackrel{\mathcal{D}}{\to} \mathcal{N}\left(0, \frac{f(\theta)}{\left[f''(\theta)\right]^2}\int_{\mathbb{R}} K'^2(x)\,\mathrm{d}x\right),$$

where $\mathcal{N}$ is the Gaussian distribution. Eddy [4] and Romano [10] then proved that this central limit theorem still holds when the condition $nh_n^6 \to \infty$ is weakened to $nh_n^5/\log n \to \infty$. Recently, Mokkadem and Pelletier [7] extended these results to the multivariate framework. Precisely, these authors show, under suitable assumptions, that

$$\sqrt{nh_n^{d+2}}\,(\hat{\theta}_n - \theta) \stackrel{\mathcal{D}}{\to} \mathcal{N}\left(0, f(\theta)\left[\mathrm{H}f(\theta)\right]^{-1}\mathrm{G}\left[\mathrm{H}f(\theta)\right]^{-1}\right), \tag{2.1}$$

where G is the $d \times d$ matrix defined by

$$\mathrm{G}_{i,j} = \int_{\mathbb{R}^d} \frac{\partial K}{\partial x_i}(x)\frac{\partial K}{\partial x_j}(x)\,\mathrm{d}x.$$

Therefore, under the assumptions of these authors, which imply **H1**, **H2**, **H4** and **H5**, the results above transfer to θ_n.

Following the remark of a referee, we would like to shed light on the fact that there are some problems associated with the use of results of this type. As an example, if one is interested in constructing confidence sets, it will be necessary to estimate the limiting variance matrix, which involves not only $f(\theta)$ but also the local sharpness around the peak, that is, the Hessian matrix $\mathrm{H}f(\theta)$. A possible answer is to use the weakly consistent estimates $f_n(\theta_n)$

and $\mathrm{H}f_n(\theta_n)$ of $f(\theta)$ and $\mathrm{H}f(\theta)$ as well as (2.1) in order to obtain, under suitable assumptions,

$$\sqrt{nh_n^{d+2}}\Big[f_n(\theta_n)\big[\mathrm{H}f_n(\theta_n)\big]^{-1}\mathrm{G}\big[\mathrm{H}f_n(\theta_n)\big]^{-1}\Big]^{-1/2}(\theta_n-\theta)\xrightarrow{\mathcal{D}}\mathcal{N}(0,\mathrm{I}),$$

where I denotes the $d\times d$ identity matrix.

3 Proofs

3.1 Proof of Theorem 2.1

In the sequel, $B(a,\varepsilon)$ stands for the closed ball in $\big(\mathbb{R}^d,\|.\|\big)$ with center at a and radius $\varepsilon>0$. For all $n\geq 1$, $A_n(\varepsilon)$ will denote the random set

$$A_n(\varepsilon)=\{x\in\mathbb{R}^d\,:\,f_n(x)\geq f_n(\hat{\theta}_n)-\varepsilon\}.$$

First of all, we state two fundamental results. For the sake of clarity, their proofs are delayed to the end of the section.

Lemma 3.1 *Assume that* **H1**-**H3** *and* **H5** *hold. Then, for any sequence $(\alpha_n)_{n\geq 1}$ of positive random variables vanishing in probability, we have*

$$\operatorname{diam} A_n(\alpha_n)\xrightarrow{P}0.$$

Observe the correspondence between Lemma 3.1 and assumption **H2**, where A and ε have been replaced by the random quantities A_n and α_n.

Proposition 3.1 *Assume that* **H1** *and* **H3** *hold, and let $(u_n)_{n\geq 1}$ be a sequence of positive real numbers such that $u_n\to 0$ and $nu_n^d/\log n\to\infty$. Then*

$$\mathbf{P}\big(\forall i\leq n\,:\,\|X_i-\hat{\theta}_n\|\geq u_n\big)\to 0.$$

The interest of Proposition 3.1 is in the assertion that there exists with high probability an observation within a distance u_n of $\hat{\theta}_n$. The main idea of the proof of Theorem 2.1 is to show that such an observation is also close to θ_n.

Proof of Theorem 2.1 Let $\eta>0$ and, for all $n\geq 1$, $B_n=B(\hat{\theta}_n,c\eta/r_n)$, where $c\geq 1$ denotes a constant to be specified later. We have

$$\begin{aligned}\mathbf{P}\Big(\theta_n\notin B_n\Big)\leq{}&\mathbf{P}\Big(\theta_n\notin B_n,\,\exists k\leq n\,:\,\|X_k-\hat{\theta}_n\|<\eta/r_n\Big)\\&+\mathbf{P}\Big(\forall k\leq n\,:\,\|X_k-\hat{\theta}_n\|\geq\eta/r_n\Big).\end{aligned}$$

By Proposition 3.1, the second term of the right member vanishes as n grows to infinity. Thus, one only needs to prove that the first term tends to 0. Note first that, since $c \geq 1$, the event

$$\left[\theta_n \notin B_n,\, \exists k \leq n \,:\, \|X_k - \hat{\theta}_n\| < \eta/r_n\right]$$

is contained in the event

$$\left[\max_{X_i \notin B_n,\, i \leq n} f_n(X_i) \geq \max_{X_i \in B_n,\, i \leq n} f_n(X_i) \geq f_n(X_k)\right],$$

where the data X_k satisfies $\|X_k - \hat{\theta}_n\| < \eta/r_n$. Now, denoting by $\nabla f_n(x)$ the gradient of f_n at the point x, we have, according to the previous condition on X_k,

$$\begin{aligned}
|f_n(X_k) - f_n(\hat{\theta}_n)| &\leq \sup_{x \in B(\hat{\theta}_n, \eta/r_n)} \|\nabla f_n(x)\| \frac{\eta}{r_n} \\
&= \sup_{x \in B(\hat{\theta}_n, \eta/r_n)} \|\nabla f_n(x) - \nabla f_n(\hat{\theta}_n)\| \frac{\eta}{r_n} \\
&\qquad \text{(since } \nabla f_n(\hat{\theta}_n) = 0) \\
&\leq a_n \left(\frac{\eta}{r_n}\right)^2,
\end{aligned}$$

where $a_n = \sup_{x \in B(\hat{\theta}_n, \eta/r_n)} \|\mathrm{H} f_n(x)\|$. Consequently,

$$\begin{aligned}
&\mathbf{P}\Big(\theta_n \notin B_n,\, \exists k \leq n \,:\, \|X_k - \hat{\theta}_n\| < \eta/r_n\Big) \\
&\quad \leq \mathbf{P}\Big(\max_{X_i \notin B_n,\, i \leq n} f_n(X_i) \geq f_n(\hat{\theta}_n) - a_n(\eta/r_n)^2\Big) \\
&\quad \leq 1 - \mathbf{P}\Big(\sup_{B_n^c} f_n < f_n(\hat{\theta}_n) - a_n(\eta/r_n)^2\Big).
\end{aligned}$$

Hence, one only needs now to show that

$$\mathbf{P}\Big(\sup_{B_n^c} f_n < f_n(\hat{\theta}_n) - a_n(\eta/r_n)^2\Big) \to 1 \quad \text{as } n \text{ tends to infinity.}$$

To this aim, observe that if $\alpha_n = 2a_n(\eta/r_n)^2$, the event

$$\left[\forall x \in B_n^c \,:\, f_n(x) < f_n(\hat{\theta}_n) - \alpha_n\right]$$

is contained in the event

$$\left[\sup_{B_n^c} f_n < f_n(\hat{\theta}_n) - a_n(\eta/r_n)^2\right].$$

The first of the two events equals the event $[A_n(\alpha_n) \subset B_n]$. Consequently, the problem is reduced to showing that

$$\mathbf{P}\Big(A_n(\alpha_n) \subset B_n\Big) \to 1.$$

Using Taylor's formula and **H4** we have, for all $n \geq 1$ and $x \in A_n(\alpha_n)$,

$$f_n(x) - f_n(\hat{\theta}_n) = \frac{1}{2}(x - \hat{\theta}_n)^t \mathrm{R}_n(x)(x - \hat{\theta}_n), \tag{3.1}$$

where $\mathrm{R}_n(x) = \mathrm{H}f_n\big(x + \xi_x^n(x - \hat{\theta}_n)\big)$ and $\xi_x^n \in (0, 1)$. Let us introduce the event $\mathcal{E}_n$ defined by

$$\mathcal{E}_n = \Big[\forall x \in A_n(\alpha_n) \, : \, (x - \hat{\theta}_n)^t \mathrm{R}_n(x)(x - \hat{\theta}_n) \leq \frac{1}{2}(x - \hat{\theta}_n)^t \mathrm{H}f(\theta)(x - \hat{\theta}_n)\Big].$$

From (3.1), it is deduced that on the event $\mathcal{E}_n$,

$$\forall x \in A_n(\alpha_n) \, : \, -(x - \hat{\theta}_n)^t \mathrm{H}f(\theta)(x - \hat{\theta}_n) \leq 4\alpha_n,$$

and consequently, that

$$A_n(\alpha_n) \subset B\big(\hat{\theta}_n, 2\sqrt{\alpha_n/\gamma}\big), \tag{3.2}$$

where $\gamma = \inf_{\|x\|=1} \|\big(- \mathrm{H}f(\theta)\big)^{1/2} x\|^2$. Obviously $\gamma > 0$ according to **H1**. Therefore, according to (3.2), on the event $\mathcal{E}_n$, $A_n(\alpha_n) \subset B_n$ as soon as $c \geq 2\sqrt{2a_n/\gamma}$. Thus, recalling that $c \geq 1$, we choose

$$c = 2 \max\left(1, 2\sqrt{\frac{2\|\mathrm{H}f(\theta)\|}{\gamma}}\right).$$

We obtain as sort

$$\begin{aligned} \mathbf{P}\Big(A_n(\alpha_n) \subset B_n\Big) &\geq \mathbf{P}\Big(A_n(\alpha_n) \subset B_n, \mathcal{E}_n, c \geq 2\sqrt{2a_n/\gamma}\Big) \\ &= \mathbf{P}\Big(\mathcal{E}_n, c \geq 2\sqrt{2a_n/\gamma}\Big). \end{aligned}$$

Since $\hat{\theta}_n \xrightarrow{P} \theta$ by **H3**, $\sup_{x \in \mathcal{V}} \|\mathrm{H}f_n(x) - \mathrm{H}f(x)\| \xrightarrow{P} 0$ by **H4** and $\mathrm{H}f$ is continuous on $\mathcal{V}$ by **H1**, we get

$$\mathbf{P}\Big(c \geq 2\sqrt{2a_n/\gamma}\Big) \to 1.$$

Thus, it remains to prove that $\mathbf{P}(\mathcal{E}_n) \to 1$. First note that

$$\sup_{x\in A_n(\alpha_n)} \|\mathrm{R}_n(x) - \mathrm{H}f(\theta)\| \le \sup_{x\in A_n(\alpha_n)} \|\mathrm{R}_n(x) - \mathrm{H}f\big(x + \xi_x^n(x-\hat{\theta}_n)\big)\| + \sup_{x\in A_n(\alpha_n)} \|\mathrm{H}f\big(x + \xi_x^n(x-\hat{\theta}_n)\big) - \mathrm{H}f(\theta)\|.$$

Observe now that

$$\sup_{x\in A_n(\alpha_n)} \|x + \xi_x^n(x-\hat{\theta}_n) - \hat{\theta}_n\| \le 2\,\mathrm{diam}\ A_n(\alpha_n)$$

and that the bound vanishes in probability according to Lemma 3.1. Therefore, since $\hat{\theta}_n \xrightarrow{P} \theta$, we deduce from **H1** and **H4** that

$$\sup_{x\in A_n(\alpha_n)} \|\mathrm{R}_n(x) - \mathrm{H}f(\theta)\| \xrightarrow{P} 0. \tag{3.3}$$

Consequently,

$$\begin{aligned}
\mathbf{P}(\mathcal{E}_n) &\ge \mathbf{P}\Big(\forall x \in A_n(\alpha_n),\ \forall\|y\| = 1\ :\ -y^t\mathrm{R}_n(x)y \ge -1/2\, y^t\mathrm{H}f(\theta)y\Big)\\
&\ge \mathbf{P}\Big(\forall\|y\| = 1\ :\ \sup_{x\in A_n(\alpha_n)} |y^t\big(\mathrm{R}_n(x) - \mathrm{H}f(\theta)\big)y| \le -1/2\, y^t\mathrm{H}f(\theta)y\Big)\\
&\qquad \text{(using the triangle inequality)}\\
&\ge \mathbf{P}\Big(\sup_{x\in A_n(\alpha_n)} \|\mathrm{R}_n(x) - \mathrm{H}f(\theta)\| \le -1/2 \inf_{\|y\|=1} y^t\mathrm{H}f(\theta)y\Big) \qquad (3.4)
\end{aligned}$$

where, for the last inequality, we used the fact that

$$\sup_{\|y\|=1}\ \sup_{x\in A_n(\alpha_n)} |y^t\big(\mathrm{R}_n(x) - \mathrm{H}f(\theta)\big)y| \le \sup_{x\in A_n(\alpha_n)} \|\mathrm{R}_n(x) - \mathrm{H}f(\theta)\|.$$

The probability (3.4) tends to 1 according to (3.3), since $\mathrm{H}f(\theta)$ is negative definite. Consequently, $\mathbf{P}(\mathcal{E}_n) \to 1$, hence the theorem is proved. ■

3.2 Proof of Lemma 3.1

For all $n \ge 1$ and $\mu > 0$, we set

$$\beta_n = \alpha_n + \sup_{x\in\mathbb{R}^d} |f_n(x) - \mathbf{E}f_n(x)| + |f(\theta) - f_n(\hat{\theta}_n)|$$

and

$$D_n(\mu) = \left\{x \in \mathbb{R}^d\ :\ \mathbf{E}f_n(x) \ge f(\theta) - \mu\right\}.$$

We note that the first two terms in β_n go to zero in virtue of **H5**. With respect to the third term, it is bounded by

$$|f(\theta) - f(\hat{\theta}_n)| + |f(\hat{\theta}_n) - f_n(\hat{\theta}_n)|.$$

The first of the two terms above tends to 0 in probability by the continuity of f around θ and the fact that $\hat{\theta}_n \xrightarrow{P} \theta$. Finally, the second term vanishes under **H5** and the uniform continuity of f around its mode.

Moreover, attention shows that

$$A_n(\alpha_n) \subset D_n(\beta_n).$$

Let $\varepsilon > 0$. Then, for all $\gamma > 0$, the following chain of inequalities is valid.

$$\begin{aligned}
\mathbf{P}\Big(\text{diam } A_n(\alpha_n) > 4\varepsilon\Big) &\leq \mathbf{P}\Big(\text{diam } \ D_n(\beta_n) > 4\varepsilon\Big) \\
&\leq \mathbf{P}\Big(\text{diam } \ D_n(\beta_n) > 4\varepsilon, \beta_n \leq \gamma\Big) + \mathbf{P}\Big(\beta_n \geq \gamma\Big) \\
&\leq \mathbf{P}\Big(\text{diam } \ D_n(\gamma) > 4\varepsilon\Big) + \mathbf{P}\Big(\beta_n \geq \gamma\Big).
\end{aligned}$$

As mentioned above, we have $\beta_n \xrightarrow{P} 0$. Thus, one only needs to prove the existence of $\gamma > 0$ such that, for all n large enough,

$$D_n(\gamma) \subset B(\theta, 2\varepsilon).$$

It is easy to deduce from **H2** (see also Abraham, Biau and Cadre [1], Lemma 1) the existence of $\gamma > 0$ such that

$$\sup_{B(\theta,\varepsilon)^c} f < f(\theta) - 2\gamma.$$

Now, since K is integrable, there exists a compact set $T \subset \mathbb{R}^d$ with

$$f(\theta) \int_{T^c} |K(y)| \, \mathrm{d}y < \gamma.$$

Consequently, if n is large enough, we have, for all $x \in B(\theta, 2\varepsilon)^c$,

$$\begin{aligned}
\mathbf{E} f_n(x) &= \int_T K(y) f(x - h_n y) \, \mathrm{d}y + \int_{T^c} K(y) f(x - h_n y) \, \mathrm{d}y \\
&\leq \sup_{B(\theta,\varepsilon)^c} f + f(\theta) \int_{T^c} |K(y)| \, \mathrm{d}y \\
&< f(\theta) - \gamma.
\end{aligned}$$

Hence, for all n large enough, $D_n(\gamma) \subset B(\theta, 2\varepsilon)$, and the proof of the lemma is complete. ■

3.3 Proof of Proposition 3.1

We have divided the proof of Proposition 3.1 into a sequence of two lemmas. Before stating these two lemmas, we need to introduce some additional notations. From now on, ρ denotes a fixed positive real number such that $\inf_{B(\theta,\rho)} f > 0$. Note that such a ρ does exist under assumption **H1**. For any $\mathbb{R}^d$-valued random variable Z, we let Z^* be the random variable defined as follows:

$$Z^* = \begin{cases} \theta + \rho(Z-\theta)/\|Z-\theta\| & \text{if } Z \notin B(\theta,\rho) \\ Z & \text{if } Z \in B(\theta,\rho). \end{cases}$$

It is worth pointing out that Z^* is a *truncated version* of Z. Indeed, for $Z \notin B(\theta,\rho)$, Z^* is defined as the intersection of the line θZ with the sphere $\{x : \|x-\theta\| = \rho\}$.

We also introduce a *ghost sample* $Y_1, \ldots, Y_n$ of independent and identically distributed random variables with density f. This sample is assumed to be *independent* of the sample $X_1, \ldots, X_n$. Finally, for any subsets $E, F \subset \mathbb{R}^d$, we define $\delta(E,F)$ as

$$\delta(E,F) = \sup_{x\in E} \inf_{y \in F} \|x-y\|.$$

The quantity $\delta(E,F)$ may be regarded as a distance between the sets E and F reminiscent of the usual Hausdorff metric.

Lemma 3.2 *Assume that* **H1** *holds and let $(u_n)_{n\geq 1}$ be a sequence of positive real numbers such that $u_n \to 0$ and $nu_n^d/\log n \to \infty$. Then*

$$\mathbf{P}\Big(\delta\big(\{Y_1^*, \ldots, Y_n^*\}, \{X_1^*, \ldots, X_n^*\}\big) \geq u_n\Big) \to 0.$$

Proof of Lemma 3.2 Set $Y = Y_1$ and $\delta_n^* = \delta\big(\{Y_1^*, \ldots, Y_n^*\}, \{X_1^*, \ldots, X_n^*\}\big)$.

We have, for all $n \geq 1$,

$$\begin{aligned}
\mathbf{P}\Big(\delta_n^* \geq u_n\Big) &= \mathbf{P}\Big(\exists i \leq n,\, \forall j \leq n \,:\, \|Y_i^* - X_j^*\| \geq u_n\Big) \\
&= 1 - \mathbf{P}\Big(\forall i \leq n,\, \exists j \leq n \,:\, \|Y_i^* - X_j^*\| < u_n\Big) \\
&= 1 - \mathbf{E}\Big[\mathbf{P}\Big(\exists j \leq n \,:\, \|Y^* - X_j^*\| < u_n | X_1^*, \ldots, X_n^*\Big)^n\Big] \\
&\leq 1 - \mathbf{P}\Big(\exists j \leq n \,:\, \|Y^* - X_j^*\| < u_n\Big)^n \\
&\qquad \text{(using Jensen's inequality)} \\
&= 1 - \Big(1 - \mathbf{P}\Big(\forall j \leq n \,:\, \|Y^* - X_j^*\| \geq u_n\Big)\Big)^n \\
&= 1 - \Big(1 - \mathbf{E}\Big[\mathbf{P}\Big(\|Y^* - X^*\| \geq u_n | Y^*\Big)^n\Big]\Big)^n,
\end{aligned}$$

where $X = X_1$. Hence, one only needs to prove that

$$n\,\mathbf{E}\Big[\mathbf{P}\Big(\|Y^* - X^*\| \geq u_n | Y^*\Big)^n\Big] \to 0 \quad \text{as } n \to \infty.$$

By the very definition of X^*, we have

$$\begin{aligned}
\mathbf{P}\Big(\|Y^* - X^*\| \geq u_n | Y^*\Big) &\leq 1 - \mathbf{P}\Big(X^* \in B(Y^*, u_n/2) | Y^*\Big) \\
&\leq 1 - \mathbf{P}\Big(X^* \in B(Y^*, u_n/2), X \in B(\theta, \rho) | Y^*\Big) \\
&= 1 - \mathbf{P}\Big(X \in B(Y^*, u_n/2) \cap B(\theta, \rho) | Y^*\Big) \\
&= 1 - \int_{B(Y^*, u_n/2) \cap B(\theta,\rho)} f(x)\,\mathrm{d}x \\
&\leq 1 - \lambda\big(B(Y^*, u_n/2) \cap B(\theta, \rho)\big) \inf_{B(\theta,\rho)} f\,,
\end{aligned}$$

where λ denotes the Lebesgue measure on $\mathbb{R}^d$. Since $Y^* \in B(\theta, \rho)$, we deduce that for some constant $c > 0$ depending only on ρ, f and d, we have, with probability one,

$$\mathbf{P}\Big(\|Y^* - X^*\| \geq u_n | Y^*\Big) \leq 1 - c\,u_n^d\,.$$

Therefore

$$n\,\mathbf{E}\Big[\mathbf{P}\Big(\|Y^* - X^*\| \geq u_n | Y^*\Big)^n\Big] \leq n\big(1 - c\,u_n^d\big)^n,$$

and the bound vanishes under the conditions $u_n \to 0$ and $nu_n^d / \log n \to \infty$.■

Lemma 3.3 *Assume that* **H1** *and* **H3** *hold, and let* $(u_n)_{n\geq 1}$ *be a sequence of positive real numbers such that* $u_n \to 0$ *and* $nu_n^d \to \infty$. *Then*

$$\mathbf{P}\Big(\forall i \leq n \,:\, \|Y_i - \hat{\theta}_n\| \geq u_n\Big) \to 0.$$

Proof of Lemma 3.3 Let $\varepsilon > 0$. By **H3**, there exists $a > 0$ such that $\sup_{n\geq 1} \mathbf{P}(v_n\|\hat{\theta}_n - \theta\| > a) \leq \varepsilon$. If, for all $n \geq 1$, T_n denotes the ball $B(\theta, a/v_n)$, ν_n denotes the law of $\hat{\theta}_n$ and $Y = Y_1$, we deduce from the independence of $\hat{\theta}_n$ and $(Y_1, \ldots, Y_n)$ that

$$\begin{aligned}
\mathbf{P}\Big(\forall i \leq n \,:\, \|Y_i - \hat{\theta}_n\| \geq u_n\Big) &= \int \mathbf{P}\Big(\forall i \leq n \,:\, \|Y_i - t\| \geq u_n\Big)\nu_n(\mathrm{d}t) \\
&\leq \varepsilon + \int_{T_n} \mathbf{P}\Big(\|Y - t\| \geq u_n\Big)^n \nu_n(\mathrm{d}t) \\
&\leq \varepsilon + \sup_{t\in T_n} \mathbf{P}\Big(\|Y - t\| \geq u_n\Big)^n.
\end{aligned}$$

Now, for all n large enough,

$$\begin{aligned}
\sup_{t\in T_n} \mathbf{P}\Big(\|Y - t\| \geq u_n\Big)^n &= \sup_{t\in T_n} \Big(1 - \int_{B(t,u_n)} f(x)\,\mathrm{d}x\Big)^n \\
&\quad \text{(using the compactness of } T_n\text{)} \\
&\leq \big(1 - c\, u_n^d \inf_{B(\theta,\rho)} f\big)^n,
\end{aligned}$$

where $c > 0$ is a constant which only depends on the dimension d. Since $u_n \to 0$, $nu_n^d \to \infty$ and $\inf_{B(\theta,\rho)} f > 0$, the lemma is proved. ■

We are now in a position to prove Proposition 3.1.

Proof of Proposition 3.1 As in the proof of Lemma 3.2, we use the notation $\delta_n^* = \delta\big(\{Y_1^*, \ldots, Y_n^*\}, \{X_1^*, \ldots, X_n^*\}\big)$. For all $n \geq 1$, we have

$$\begin{aligned}
\mathbf{P}\Big(\forall i \leq n \,:\, \|X_i - \hat{\theta}_n\| \geq u_n\Big) &\leq \mathbf{P}\Big(\forall i \leq n \,:\, \|X_i - \hat{\theta}_n\| \geq u_n, \hat{\theta}_n \in B(\theta, \rho/2)\Big) \\
&\quad + \mathbf{P}\Big(\hat{\theta}_n \notin B(\theta, \rho/2)\Big).
\end{aligned}$$

By **H3**, the last term vanishes. Moreover, assuming that n is large enough to ensure that $u_n \le \rho/2$, we can write

$$\begin{aligned}
&\mathbf{P}\Big(\forall i \le n \,:\, \|X_i - \hat\theta_n\| \ge u_n, \hat\theta_n \in B(\theta, \rho/2)\Big)\\
&\quad = \mathbf{P}\Big(\forall i \le n \,:\, \|X_i^* - \hat\theta_n\| \ge u_n, \hat\theta_n \in B(\theta, \rho/2)\Big)\\
&\quad \le \mathbf{P}\Big(\forall i \le n \,:\, \|X_i^* - \hat\theta_n\| \ge u_n, \delta_n^* \le u_n/2, \hat\theta_n \in B(\theta, \rho/2)\Big)\\
&\qquad + \mathbf{P}\Big(\delta_n^* > u_n/2\Big).
\end{aligned}$$

By Lemma 3.2 the latter term tends to 0. Finally, by the very definition of the Y_j^*'s, we obtain

$$\begin{aligned}
&\mathbf{P}\Big(\forall i \le n \,:\, \|X_i^* - \hat\theta_n\| \ge u_n, \delta_n^* \le u_n/2, \hat\theta_n \in B(\theta, \rho/2)\Big)\\
&\quad = \mathbf{P}\Big(\forall i \le n \,:\, \|X_i^* - \hat\theta_n\| \ge u_n, \forall j \le n, \exists k \le n \,:\, \|Y_j^* - X_k^*\| \le u_n/2,\\
&\qquad\qquad \hat\theta_n \in B(\theta, \rho/2)\Big)\\
&\quad \le \mathbf{P}\Big(\forall j \le n, \exists k \le n \,:\, \|X_k^* - \hat\theta_n\| \ge u_n, \|Y_j^* - X_k^*\| \le u_n/2, \hat\theta_n \in B(\theta, \rho/2)\Big)\\
&\quad \le \mathbf{P}\Big(\forall j \le n \,:\, \|Y_j^* - \hat\theta_n\| \ge u_n/2, \hat\theta_n \in B(\theta, \rho/2)\Big)\\
&\quad = \mathbf{P}\Big(\forall j \le n \,:\, \|Y_j - \hat\theta_n\| \ge u_n/2, \hat\theta_n \in B(\theta, \rho/2)\Big),
\end{aligned}$$

where the last equality holds if n is large enough, to ensure that $u_n \le \rho$. The lemma is then a straightforward consequence of Lemma 3.3. ■

Acknowledgments. The authors greatly thank the anonymous referees and an Associate Editor for a careful reading of the paper and many helpful comments.

References

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

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

[3] Devroye, L. (1987). *A Course in Density Estimation.* Birkhäuser, Boston.

[4] Eddy, W.F. (1980). Optimum kernel estimates of the mode. *The Annals of Statistics.* **8** 870–882.

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

[6] Leclerc, J. and Pierre-Loti-Viaud, D. (2000). Vitesse de convergence presque sûre de l'estimateur à noyau du mode. *Comptes Rendus de l'Académie des Sciences de Paris.* **331** 637–640.

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

[8] Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics.* **33** 1065–1076.

[9] Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer–Verlag, New York.

[10] Romano, J.P. (1988). On weak convergence and optimality of kernel density estimates of the mode. *The Annals of Statistics.* **16** 629–647.

[11] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics.* **27** 832–837.

[12] Sager, T.W. (1983). Estimating modes and isopleths. *Communications in Statistics – Theory and Methods.* **12(5)** 529–557.

[13] Samanta, M. (1973). Nonparametric estimation of the mode of a multivariate density. *South African Statistical Journal.* **7** 109–117.

[14] Silverman, B. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *The Annals of Statistics.* **6** 177–184.

[15] Vieu, P. (1996). A note on density mode estimation. *Statistics and Probability Letters.* **26** 297–307.