# Adaptive Bayesian Density Estimation with Location-Scale Mixtures 

Willem Kruijer, Judith Rousseau, Aad A.W. van Der Vaart

## To cite this version:

Willem Kruijer, Judith Rousseau, Aad A.W. van Der Vaart. Adaptive Bayesian Density Estimation with Location-Scale Mixtures. Electronic Journal of Statistics , 2010, 4, pp.1225-1257. hal00389343v2

HAL Id: hal-00389343
https://hal.science/hal-00389343v2
Submitted on 1 Jul 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Adaptive Bayesian Density Estimation with Location-Scale Mixtures 

Willem Kruijer, Judith Rousseau and Aad van der Vaart<br>Address of the First and Second authors: CEREMADE<br>Université Paris Dauphine<br>Place du Marchal De Lattre De Tassigny<br>75775 PARIS CEDEX 16<br>France<br>e-mail: kruijer@ceremade.dauphine.fr; rousseau@ceremade.dauphine.fr

Address of the third author:
Department of Mathematics
Vrije Universiteit Amsterdam
Boelelaan 1081a
1081 HV Amsterdam
The Netherlands
e-mail: aad@math.vu.nl


#### Abstract

We study convergence rates of Bayesian density estimators based on finite location-scale mixtures of a kernel proportional to $\exp \left\{-|x|^{p}\right\}$. We construct a finite mixture approximation of densities whose logarithm is locally $\beta$-Hölder, with squared integrable Hölder constant. Under additional tail and moment conditions, the approximation is minimax for both the Kullback-Leibler divergence. We use this approximation to establish convergence rates for a Bayesian mixture model with priors on the weights, locations, and the number of components. Regarding these priors, we provide general conditions under which the posterior converges at a near optimal rate, and is rate-adaptive with respect to the smoothness of the logarithm of the true density.


AMS 2000 subject classifications: Primary 62G07, 62G20.
Keywords and phrases: Rate-adaptive density estimation, Bayesian density estimation, Nonparametric density estimation, Convergence Rates, LocationScale Mixtures.

## 1. Introduction

When the number of components in a mixture model can increase with the sample size, it can be used for nonparametric density estimation. Such models were called mixture sieves by Grenander [15] and Geman and Hwang [7]. Although originally introduced in a maximum likelihood context, there has been a large number of Bayesian papers in recent years; among many others, see [25], [5], and [6]. Whereas much progress has been made regarding the computational problems in nonparametric Bayesian inference (see for example the review by Marin et al.[22]), results on convergence rates were found only recently, especially for the case when the underlying distribution is not a mixture itself. Also
the approximative properties of mixtures needed in the latter case are not well understood.

In this paper we find conditions under which a probability density of any Hölder-smoothness can be efficiently approximated by a location-scale mixture. Using these results we then considerably generalize existing results on posterior convergence of location-scale mixtures. In particular our results are adaptive to any degree of smoothness, and allow for more general kernels and priors on the mixing distribution. Moreover, the bandwidth prior can be any inverse-gamma distribution, whose support neither has to be bounded away from zero, nor to depend on the sample size.

We consider location-scale mixtures of the type

$$
\begin{equation*}
m(x ; k, \mu, w, \sigma)=\sum_{j=1}^{k} w_{j} \psi_{\sigma}\left(x-\mu_{j}\right) \tag{1}
\end{equation*}
$$

where $\sigma>0, w_{j} \geq 0, \sum_{j=1}^{k} w_{j}=1, \mu_{j} \in \mathbb{R}$ and, for $p \in \mathbb{N}$,

$$
\begin{equation*}
\psi_{\sigma}(x)=\frac{1}{2 \sigma \Gamma\left(1+\frac{1}{p}\right)} e^{-(|x| / \sigma)^{p}} \tag{2}
\end{equation*}
$$

Approximation theory (see for example [3]) tells us that for a compactly supported kernel and a compactly supported $\beta$-Hölder function, being not necessarily nonnegative, the approximation error will be of order $k^{-\beta}$, provided $\sigma \sim k^{-1}$ and the weights are carefully chosen. This remains the case if both the kernel and the function to be approximated have exponential tails, as we consider in this work. If the function is a probability density however, this raises the question whether the approximation error $k^{-\beta}$ can also be achieved using nonnegative weights only. To our knowledge, this question has been little studied in the approximation theory literature.

Ghosal and Van der Vaart [13] approximate twice continuously differentiable densities with mixtures of Gaussians, but it is unclear if their construction can be extended to other kernels, or densities of different smoothness. In particular, for functions with more than two derivatives, the use of negative weights seems at first sight to be inevitable. A recent result by Rousseau [26] however does allow for nonnegative approximation of smooth but compactly supported densities by beta-mixtures. We will derive a similar result for location-scale mixtures of a kernel $\psi$ as in (2), for any $p \in \mathbb{N}$. Although the same differencing technique is used to construct the desired apprimations, there are various differences. First, we are dealing with a noncompact support, which required investigation of the tail conditions under which approximations can be established. Second, we are directly dealing with location-scale mixtures, hence there is no need for a 'location-scale mixture' approximation as in [26].

The parameters $k, \sigma, w$ and $\mu$ in (1) can be given a prior distribution $\Pi$; when there are observations $X_{1}, \ldots, X_{n}$ from an unknown density $f_{0}$, Bayes'
formula gives the posterior

$$
\Pi\left(A \mid X_{1}, \ldots, X_{n}\right)=\frac{\int_{A} \prod_{i=1}^{n} m\left(X_{i} ; k, \mu, w, \sigma\right) d \Pi(k, \mu, w, \sigma)}{\int \prod_{i=1}^{n} m\left(X_{i} ; k, \mu, w, \sigma\right) d \Pi(k, \mu, w, \sigma)} .
$$

The posterior (or its mean) can be used as a Bayesian density estimator of $f_{0}$. Provided this estimator is consistent, it is then of interest to see how fast it converges to the Dirac-mass at $f_{0}$. More precisely, let the convergence rate be a sequence $\epsilon_{n}$ tending to zero such that $n \epsilon_{n}^{2} \rightarrow \infty$ and

$$
\begin{equation*}
\Pi\left(d\left(f_{0}, f\right)>M \epsilon_{n} \mid X_{1} \ldots, X_{n}\right) \rightarrow 0 \tag{3}
\end{equation*}
$$

in $F_{0}^{n}$-probability, for some sufficiently large constant $M, d$ being the Hellingeror $L_{1}$-metric. The problem of finding general conditions for statistical models under which (3) holds has been studied in among others [11], [13], [32], [17], [8] and [29]. In all these papers, the complexity of the model needs to be controlled, typically by verifying entropy conditions, and at the same time the prior mass on Kullback-Leibler balls around $f_{0}$ needs to be lower bounded. It is for the latter condition that the need for good approximations arises. Our approximation result allows to prove (3) with $\epsilon_{n}=n^{-\frac{\beta}{2 \beta+1}}(\log n)^{t}$ for location-scale mixtures of the kernel $\psi$, provided $p$ is even and $f_{0}$ is locally Hölder and has tails bounded by $\psi$. The constant $t$ in the rate depends on the choice of the prior. We only consider priors independent of $\beta$, hence the posterior adapts to the unknown smoothness of $f_{0}$, which can be any $\beta>0$. The adaptivity relies on the approximation result that allows to approximate $f_{0}$ with $f_{1} * \psi$, for a density $f_{1}$ that may be different from $f_{0}$. In previous work on density estimation with finite location-scale mixtures (see e.g. [27], [8] and [13]) $f_{0}$ is approximated with $f_{0} * \psi$, which only gives minimax-rates for $\beta \leq 2$. For regression-models based on location-scale mixtures, fully adaptive posteriors have recently been obtained by De Jonge and Van Zanten [2]; their work was written at the same time and independently of the present work. For continuous beta-mixtures (near)-optimal ${ }^{1}$ rates have been derived by Rousseau [26]. Another related work is [28], where also kernels of type (2) are studied; however it is assumed that the true density is a mixture itself. In a clustering and variable selection framework using multivariate Gaussian mixtures, Maugis and Michel [23] give non-asymptotic bounds on the risk of a penalized maximum likelihood estimator. Finally, for a general result on consistency of location scale mixtures, see [31].

Notation Let $C_{p}$ denote the normalizing constant $\left(2 \Gamma\left(1+\frac{1}{p}\right)\right)^{-1}$. The inverse $\psi_{\sigma}^{-1}(y)=\sigma\left(\log \frac{C_{p}}{y}\right)^{1 / p}$ is defined on $\left(0, C_{p}\right]$. When $\sigma=1$ we also write $\psi(x)=\psi_{1}(x)=C_{p} \exp \left\{-|x|^{p}\right\}$ and $\psi^{-1}(y)=\psi_{1}^{-1}(y)$. For any nonnegative $\alpha$, let

$$
\begin{equation*}
\nu_{\alpha}=\int x^{\alpha} \psi(x) d x \tag{4}
\end{equation*}
$$

[^0]For any function $h$, let $K_{\sigma} h$ denote the convolution $h * \psi_{\sigma}$, and let $\Delta_{\sigma} h$ denote the error $\left(K_{\sigma} h\right)-h$.

The ( $k-1$ )-dimensional unit-simplex and the $k$-dimensional bounded quadrant are denoted

$$
\Delta_{k}=\left\{x \in \mathbb{R}^{k}: x_{i} \geq 0, \sum_{i=1}^{k} x_{i}=1\right\}, \quad S_{k}=\left\{x \in \mathbb{R}^{k}: x_{i} \geq 0, \sum_{i=1}^{k} x_{i} \leq 1\right\}
$$

and $H_{k}[b, d]=\left\{x \in \mathbb{R}^{k} \mid x_{i} \in\left[b_{i}, d_{i}\right]\right\}$, where $b, d \in \mathbb{R}^{k}$. When no confusion can result we write $H_{k}[b, d]:=H_{k}[(b, \ldots, b),(d, \ldots, d)]$ for real numbers $b$ and $d$. For positive numbers $c$ and $\epsilon$,

$$
\begin{equation*}
T_{c, \epsilon}=\left[-c|\log \epsilon|^{1 / p}, c|\log \epsilon|^{1 / p}\right] . \tag{5}
\end{equation*}
$$

Given $\epsilon>0$ and fixed points $x \in \mathbb{R}^{k}$ and $y \in \Delta_{k}$, define the $l_{1}$-balls

$$
\begin{aligned}
& B_{k}(x, \epsilon)=\left\{z \in \mathbb{R}^{k} ; \sum_{i=1}^{k}\left|z_{i}-x_{i}\right| \leq \epsilon\right\}, \\
& \Delta_{k}(y, \epsilon)=\left\{z \in \Delta_{k} ; \sum_{i=1}^{k}\left|z_{i}-y_{i}\right| \leq \epsilon\right\} .
\end{aligned}
$$

Inequality up to a multiplicative constant is denoted with $\lesssim$ and $\gtrsim$ (for $\lesssim$ we also use $O$ ). The number of integer points in an interval $I \in \mathbb{R}$ is denoted $N(I)$. Integrals of the form $\int g d F_{0}$ are also denoted $F_{0} g$.

## 2. Main results

We now state our conditions on $f_{0}$ and the prior. Note that some of them will not be used in some of our results. For instance in Theorem 1 below, (C3) is not required.
Conditions on $f_{0}$. The observations $X_{1}, \ldots, X_{n}$ are an i.i.d. sample from a density $f_{0}$ satisfying the following conditions.
(C1) Smoothness. $\log f_{0}$ is assumed to be locally $\beta$-Hölder, with derivatives $l_{j}(x)=\frac{d^{j}}{d x^{j}} \log f(x)$. We assume the existence of a polynomial $L$ and a constant $\gamma>0$ such that

$$
\begin{equation*}
\left|l_{r}(x)-l_{r}(y)\right| \leq r!L(x)|x-y|^{\beta-r} \tag{6}
\end{equation*}
$$

for all $x, y$ with $|y-x| \leq \gamma$.
(C2) Tails. There exists $\epsilon>0$ such that the functions $l_{j}$ and $L$ satisfy

$$
\begin{equation*}
F_{0}\left|l_{j}\right|^{\frac{2 \beta+\epsilon}{j}}<\infty, j=1, \ldots, r, \quad F_{0} L^{2+\frac{\epsilon}{\beta}}<\infty \tag{7}
\end{equation*}
$$

and there exist constants $\alpha>2, T>0$ and $c>0$ such that when $|x|>T$,

$$
\begin{equation*}
f_{0}(x) \leq c x^{-\alpha} \tag{8}
\end{equation*}
$$

(C3) A stronger tail condition: $f_{0}$ has smaller tails than the kernel, i.e. there exist constants $T$ and $M_{f_{0}}$ such that

$$
\begin{equation*}
f_{0}(x) \leq M_{f_{0}} \psi(x), \quad|x| \geq T . \tag{9}
\end{equation*}
$$

(C4) Monotonicity. $f_{0}$ is strictly positive, and there exist $x_{m}<x_{M}$ such that $f_{0}$ is nondecreasing on $\left(-\infty, x_{m}\right)$ and nonincreasing on $\left(x_{M}, \infty\right)$. Without loss of generality we assume that $f_{0}\left(x_{m}\right)=f_{0}\left(x_{M}\right)=c$ and that $f_{0}(x) \geq c$ for all $x_{m}<x<x_{M}$. The monotonicity in the tails implies that $K_{\sigma} f_{0} \gtrsim$ $f_{0}$; see the remark on p. 149-150 in [9].

Assumption (C3) is only needed in the proofs of Lemma 4 and Theorem 2.
We can now state the approximation result which will be the main ingredient in the proof of Theorem 2, but which is also interesting on its own right.
Theorem 1. Let $f$ be a density satisfying conditions (C1),(C2) and (C4), and let $K_{\sigma}$ denote convolution over the kernel $\psi$ defined in (2), for any $p \in \mathbb{N}$. Then there exists a density $h_{k}$ such that for all small enough $\sigma$,

$$
\begin{equation*}
\int f \log \frac{f}{K_{\sigma} h_{k}}=O\left(\sigma^{2 \beta}\right), \quad \int f\left(\log \frac{f}{K_{\sigma} h_{k}}\right)^{2}=O\left(\sigma^{2 \beta}\right) \tag{10}
\end{equation*}
$$

The construction of the approximation $h_{k}$ is detailed in section 3. As our smoothness condition is only local, the class of densities satisfying (C1),(C2) and (C4) is quite large. In particular, all (log)-spline densities are permitted, provided they are sufficiently differentiable at the knots. Condition (7) rules out superexponential densities like $\exp \left\{-\exp \left\{x^{2}\right\}\right\}$. In fact the smallest possible $\tilde{L}(x)$ such that (6) holds, does not have to be of polynomial form, but in that case it should be bounded by some polynomial $L$ for which (7) holds. Note that when $\beta=2, L$ is an upper bound for $\frac{d^{2}}{d x^{2}} \log f_{0}(x)=f_{0}^{\prime \prime}(x) / f_{0}(x)-\left(f^{\prime}(x) / f(x)\right)^{2}$, and apart from the additional $\epsilon$ in (7), this assumption is equivalent to the assumption in [13] that $F_{0}\left(f_{0}^{\prime \prime} / f_{0}\right)^{2}$ and $F_{0}\left(f_{0}^{\prime} / f_{0}\right)^{4}$ be finite. Also the monotonicity condition can be weakened, as in fact it suffices to have an upper and lower bound on $f_{0}$ for which ( C 4 ) hold. For the clarity of presentation however we assume monotonicity of $f_{0}$ itself.

We now describe the family of priors we consider to construct our estimate.
Prior (П) The prior on $\sigma$ is the inverse Gamma distribution with scale parameter $\lambda>0$ and shape parameter $\alpha>0$, i.e. $\sigma$ has prior density $\frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\lambda / x}$ and $\sigma^{-1}$ has the Gamma-density $\frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$.

The other parameters have a hierarchical prior, where the number of components $k$ is drawn, and given $k$ the locations $\mu$ and weights $w$ are independent. The priors on $k, \mu$ and $w$ satisfy the conditions (11)-(14) below.

The prior on $k$ is such that for all integers $k>0$

$$
\begin{equation*}
B_{0} e^{-b_{0} k(\log k)^{r}} \leq \Pi(k) \leq B_{1} e^{-b_{1} k(\log k)^{r}} \tag{11}
\end{equation*}
$$

for some constants $0<B_{0} \leq B_{1}, 0<b_{1} \leq b_{0}$ and $r \geq 0$. The constant $r$ affects the logarithmic factor in the convergence rate in Theorem 2 if it is smaller than one.

Given $k$, the locations $\mu_{1}, \ldots, \mu_{k}$ are drawn independently from a prior density $p_{\mu}$ on $\mathbb{R}$ satisfying

$$
\begin{align*}
& p_{\mu}(x) \gtrsim \psi(x)  \tag{12}\\
& p_{\mu}(x) \lesssim e^{-a_{1}|x|^{a_{2}}} \quad \text { for constants } a_{1}>0 \text { and } a_{2} \leq p \tag{13}
\end{align*}
$$

Given $k$, the prior distribution of the weight vector $w=\left(w_{1}, \ldots, w_{k}\right)$ is independent of $\mu$, and there is a constant $d_{1}$ such that for $\epsilon<\frac{1}{k}$, and $w_{0} \in \Delta_{k}$,

$$
\begin{equation*}
\Pi\left(w \in \Delta_{k}\left(w_{0}, \epsilon\right) \mid K=k\right) \gtrsim \exp \left\{-d_{1} k(\log k)^{b} \log \frac{1}{\epsilon}\right\} \tag{14}
\end{equation*}
$$

for some nonnegative constant $b$, which affects the logarithmic factor in the convergence rate.

Theorem 2. Let the bandwidth $\sigma$ be given an inverse-gamma prior, and assume that the prior on the weights and locations satisfies conditions (11)-(14). Given a positive even integer $p$, let $\psi$ be the kernel defined in (2), and consider the family of location-scale mixtures defined in (1), equipped with the prior described above. If $f_{0}$ satisfies conditions (C1)-(C4), then $\Pi\left(\cdot \mid X_{1}, \ldots, X_{n}\right)$ converges to $f_{0}$ in $F_{0}^{n}$-probability, with respect to the Hellinger or $L_{1}$-metric, with rate $\epsilon_{n}=n^{-\beta /(1+2 \beta)}(\log n)^{t}$, where $r$ and $b$ are as in (11) and (14), and $t>(2+$ $\left.b+p^{-1}\right) /\left(2+\beta^{-1}\right)+\max (0,(1-r) / 2)$.
The proof is based on Theorem 5 of Ghosal and van der Vaart [13], which is included here in appendix A.

Condition (11) is usual in finite mixture models, see for instance [10], [20] and [26] for beta-mixtures. It controls both the approximating properties of the support of the prior and its entropy. For a Poisson prior, we have $r=1$ and for a geometric prior $r=0$.

Conditions (12) and (14) translate the general prior mass condition (41) in Theorem 3 to conditions on the priors for $\mu$ and $w$. The prior is to put enough mass near $\mu_{0}$ and $w_{0}$, which are the locations and weights of a mixture approximating $f_{0}$. Since $\mu_{0}$ and $w_{0}$ are unknown, the conditions in fact require that there is a minimal amount of prior mass around all their possible values. The restriction to kernels with even $p$ in Theorem 2 is assumed to discretize the approximation $h_{k}$ obtained from Theorem 1. Results on minimax-rates for Laplace-mixtures $(p=1)$ (see [18]) suggest that this assumption is in fact necessary. Note that also [2] and [28] require analytic kernels.

## 3. Approximation of smooth densities

In many statistical problems it is of interest to bound the Kullback-Leibler divergence $D_{K L}\left(f_{0}, m\right)=\int f_{0} \log \frac{f_{0}}{m}$ between $f_{0}$ and densities contained in the model under consideration, in our case finite location-scale mixtures $m$. When $\beta \leq 2$, the usual approach to find an $m$ such that $D_{K L}\left(f_{0}, m\right)=O\left(\sigma^{2 \beta}\right)$, is to discretize the continuous mixture $K_{\sigma} f_{0}$, and show that $\left\|K_{\sigma} f_{0}-m\right\|_{\infty}$ and
$\left\|f_{0}-K_{\sigma} f_{0}\right\|_{\infty}$ are both $O\left(\sigma^{\beta}\right)$. Under additional assumptions on $f_{0}$, this then gives a KL-divergence of $O\left(\sigma^{2 \beta}\right)$. But as $\left\|f_{0}-K_{\sigma} f_{0}\right\|_{\infty}$ remains of order $\sigma^{2}$ when $\beta>2$, this approach appears to be inefficient for smooth $f_{0}$. In this section we propose an alternative mixing distribution $\tilde{f}_{0}$ such that $D_{K L}\left(f_{0}, K_{\sigma} \tilde{f}_{0}\right)=$ $O\left(\sigma^{2 \beta}\right)$. To do so, we first construct a not necessarily positive function $f_{k}$ such that under a global Hölder condition, $\left\|f_{0}-K_{\sigma} f_{k}\right\|_{\infty}=O\left(\sigma^{\beta}\right)$. However, as we only assume the local Hölder condition (C1), the approximation error of $O\left(\sigma^{\beta}\right)$ will in fact include the local Hölder constant, which is made explicit in Lemma 1. Modifying $f_{k}$ we obtain a density which still has the desired approximative properties (Lemma 2). Using this result we then prove Theorem 1. Finally we prove that the continuous mixture can be approximated by a discrete mixture (Lemmas 3 and 4). In the remainder of this section, we write $f$ instead of $f_{0}$ for notational convenience, unless stated otherwise.

To illustrate the problem that arises when approximating a smooth density $f$ with its convolution $K_{\sigma} f$, let us consider a three times continuously differentiable density $f$ such that $\left\|f^{\prime \prime}\right\|_{\infty}=L .{ }^{2}$ Then $\left\|f-K_{\sigma} f\right\|_{\infty} \leq \frac{1}{2} \nu_{2} L \sigma^{2}$, where $\nu_{2}$ is defined as in (4). Although the regularity of $f$ is larger than two, the approximation error remains order $\sigma^{2}$. The following calculation illustrates how this can be improved if we take $f_{1}=f-\Delta_{\sigma} f=2 f-K_{\sigma} f$ as the mixing density instead of $f$. The approximation error is

$$
\begin{aligned}
& \left|\left(K_{\sigma} f_{1}\right)(x)-f(x)\right|=\left|\int \psi_{\sigma}(x-\mu)\left\{\left(f-\Delta_{\sigma} f\right)(\mu)-f(x)\right\} d \mu\right| \\
& \quad=\left|\int \psi_{\sigma}(x-\mu)\left\{(f(\mu)-f(x))-\int \psi_{\sigma}(\epsilon-\mu)(f(\epsilon)-f(\mu)) d \epsilon\right\} d \mu\right| \\
& \quad=\left|\frac{\sigma^{2} \nu_{2}}{2} f^{\prime \prime}(x)+O\left(\sigma^{3}\right)-\frac{\sigma^{2}}{2} \int \psi_{\sigma}(x-\mu) f^{\prime \prime}(\mu) d \mu-O\left(\sigma^{3}\right)\right|=O\left(\sigma^{3}\right) .
\end{aligned}
$$

Likewise, the error is $O\left(\sigma^{\beta}\right)$ when $f$ is of Hölder regularity $\beta \in(2,4]$. When $\beta>4$, this procedure can be repeated, yielding a sequence

$$
\begin{equation*}
f_{j+1}=f-\Delta_{\sigma} f_{j}, \quad j \geq 0, \quad f_{0}:=f \tag{15}
\end{equation*}
$$

Once the approximation error $O\left(\sigma^{\beta}\right)$ is achieved with a certain $f_{k}$, the approximation clearly doesn't improve any more for $f_{j}$ with $j>k$. In the context of a fixed $\beta>0$ and a density $f$ of Hölder regularity $\beta, f_{k}$ will be understood as the first function in the sequence $\left\{f_{i}\right\}_{i \in \mathbb{N}}$ for which an error of order $\sigma^{\beta}$ is achieved, i.e. $k$ is such that $\beta \in(2 k, 2 k+2]$. The construction of the sequence $\left\{f_{i}\right\}_{i \in \mathbb{N}}$ is related to the use of superkernels in kernel density estimation (see e.g. [30] and [4]), or to the twicing kernels used in econometrics (see [24]). However, instead of finding a kernel $\psi_{k}$ such that $\left\|f-\psi_{k} * f\right\|_{\infty}=O\left(\sigma^{\beta}\right)$, we construct a function $f_{k}$ fow which $\left\|f-\psi * f_{k}\right\|_{\infty}=O\left(\sigma^{\beta}\right)$.

In Lemma 11 in appendix B we show that for any $\beta>0,\left\|f-K_{\sigma} f_{k}\right\|_{\infty}=$ $O\left(\sigma^{\beta}\right)$ when $f$ is (globally) $\beta$-Hölder. In Theorems 1 and 2 however we have

[^1]instead the local Hölder condition ( C 1$)$ on $\log f$, along with the tail and monotonicity conditions (C2) and (C4). With only a local Hölder condition, the approximation error will depend in some way on the local Hölder constant $L(x)$ as well as the derivatives $l_{j}(x)$ of $\log f$. This is made explicit in the following approximation result, whose proof can be found in Appendix C. A similar result for beta-mixtures is contained in Theorem 3.1 in [26].
Lemma 1. Given $\beta>0$, let $f$ be a density satisfying condition (C1), for any possible function $L$, not necessarily polynomial. Let $k$ be such that $\beta \in$ $(2 k, 2 k+2]$, and let $f_{k}$ be defined as in (15). Then for all sufficiently small $\sigma$ and for all $x$ contained in the set
\[

$$
\begin{equation*}
A_{\sigma}=\left\{x:\left|l_{j}(x)\right| \leq B \sigma^{-j}|\log \sigma|^{-\frac{j}{p}}, j=1, \ldots, r,|L(x)| \leq B \sigma^{-\beta}|\log \sigma|^{-\frac{\beta}{p}}\right\} \tag{16}
\end{equation*}
$$

\]

we have

$$
\begin{equation*}
\left(K_{\sigma} f_{k}\right)(x)=f(x)\left(1+O\left(R(x) \sigma^{\beta}\right)\right)+O\left((1+R(x)) \sigma^{H}\right) \tag{17}
\end{equation*}
$$

where $H>0$ can be chosen arbitrarily large and

$$
\begin{equation*}
R(x)=r_{r+1}|L(x)|+\sum_{i=1}^{r} r_{i}\left|l_{i}(x)\right|^{\beta / i} \tag{18}
\end{equation*}
$$

for nonnegative constants $r_{i}$.
Compared to the uniform result that can be obtained under a global Hölder condition (Lemma 11 in appendix B) the approximation error $\left(K_{\sigma} f_{k}\right)(x)-f(x)$ depends on $R(x)$. The good news however, is that on a set on which the $l_{j}$ 's are sufficiently controlled, it is also relative to $f(x)$, apart from a term $\sigma^{H}$ where $H$ can be arbitrarily large. Note that no assumptions were made regarding $L$, but obviously the result is only of interest when $L$ is known to be bounded in some way. In the remainder we require $L$ to be polynomial.

Since $K_{\sigma} f_{j}$ is a density when $f_{j}$ is a density, we have that for any nonnegative integer $j$ ( $f_{0}$ denoting the density $f$ itself) $f_{j}$ integrates to one. For $j>0$ the $f_{j}$ 's are however not necessarily nonnegative. To obtain a probability density, we define

$$
\begin{align*}
J_{\sigma, j} & =\left\{x: f_{j}(x)>\frac{1}{2} f(x)\right\}  \tag{19}\\
g_{j}(x) & =f_{j}(x) 1_{J_{\sigma, j}}+\frac{1}{2} f(x) 1_{J_{\sigma, j}^{c}}  \tag{20}\\
h_{j}(x) & =g_{j}(x) / \int g_{j}(x) d x \tag{21}
\end{align*}
$$

The constant $\frac{1}{2}$ in (19) and (20) is arbitrary and could be replaced by any other number between zero and one. In the following lemma, whose proof can be found in Appendix D, we show that the normalizing constant $\int g_{k}$ is $1+O\left(\sigma^{\beta}\right)$. For this purpose, we first control integrals over the sets $A_{\sigma}$ defined in (16) and

$$
\begin{equation*}
E_{\sigma}=\left\{x: f(x) \geq \sigma^{H_{1}}\right\} \tag{22}
\end{equation*}
$$

for a sufficiently large constant $H_{1}$.
Lemma 2. Let $f$ be a density satisfying conditions (C1),(C2) and (C4). Then for all small enough $\sigma$ and all nonnegative integers $m$ and all $K>0$,

$$
\begin{equation*}
\int_{A_{\sigma}^{c}}\left(K_{\sigma}^{m} f\right)(x) d x=O\left(\sigma^{2 \beta}\right), \quad \int_{E_{\sigma}^{c}}\left(K_{\sigma}^{m} f\right)(x) d x=O\left(\sigma^{K}\right) \tag{23}
\end{equation*}
$$

provided that $H_{1}$ in (22) is sufficiently large. Furthermore, $A_{\sigma} \cap E_{\sigma} \subset J_{\sigma, k}$ for small enough $\sigma$. Consequently,

$$
\begin{equation*}
\int g_{k}(x) d x=1+\int_{J_{\sigma, k}^{c}}\left(\frac{1}{2} f-f_{k}\right) d x=1+O\left(\sigma^{2 \beta}\right) \tag{24}
\end{equation*}
$$

Finally, when $\beta>2$, and $f_{k}$ is defined as in Lemma 1 and $h_{k}$ as in (21),

$$
\begin{equation*}
K_{\sigma} h_{k}(x)=f(x)\left(1+O\left(R(x) \sigma^{\beta}\right)\right)+O\left((1+R(x)) \sigma^{H}\right) \tag{25}
\end{equation*}
$$

for all $x \in A_{\sigma} \cap E_{\sigma}$, i.e. in (17) we can replace $f_{k}$ by $h_{k}$, provided we assume that $x$ is also contained in $E_{\sigma}$.

Remark 1. From (20), (21) and (24) it follows that $h_{k} \geq f /\left(2\left(1+O\left(\sigma^{\beta}\right)\right)\right)$. The fact that $K_{\sigma} f$ is lower bounded by a multiple of $f$ then implies that the same is true for $K_{\sigma} h_{k}$.

Remark 2. The integrals over $A_{\sigma}^{c}$ in (23) can be shown to be $O\left(\sigma^{2 \beta}\right)$ only using conditions (C1) and (C2), whereas for the integrals over $E_{\sigma}^{c}$ also condition (C4) is required.

Using this result we can now prove Theorem 1:
Proof. Since

$$
\int_{S} p \log \frac{p}{q} \leq \int_{S} p \frac{p-q}{q}=\int_{S} \frac{(p-q)^{2}}{q}+\int_{S}(p-q)=\int_{S} \frac{(p-q)^{2}}{q}+\int_{S^{c}}(q-p)
$$

for any densities $p$ and $q$ and any set $S$, we have the bound

$$
\begin{align*}
& \int f(x) \log \frac{f(x)}{K_{\sigma} h_{k}(x)} d x \leq \int_{A_{\sigma} \cap E_{\sigma}} \frac{\left(f(x)-K_{\sigma} h_{k}(x)\right)^{2}}{K_{\sigma} h_{k}(x)} d x  \tag{26}\\
& \quad+\int_{A_{\sigma}^{c} \cup E_{\sigma}^{c}} f(x) \log \frac{f(x)}{K_{\sigma} h_{k}(x)} d x+\int_{A_{\sigma}^{c} \cup E_{\sigma}^{c}}\left(K_{\sigma} h_{k}(x)-f(x)\right) d x .
\end{align*}
$$

The first integral on the right can be bounded by application of (25) and Remark 1 following Lemma 2. On $A_{\sigma} \cap E_{\sigma}$ the integrand is bounded by $f(x) O\left(\sigma^{\beta} R(x)\right)-$ $2 O\left(\sigma^{\beta+H} R(x)\right)+O\left((1+R(x))^{2}\right) \sigma^{2 H} / f(x)$. Let $H_{1}$ be such that the second integral in (23) is $O\left(\sigma^{2 \beta}\right)$ (i.e. $K=2 \beta$ ), and choose $H \geq H_{1}+\beta$. It follows from the definition of $R(x)$ and (7) that the integral over $A_{\sigma} \cap E_{\sigma}$ is $O\left(\sigma^{2 \beta}\right)$ for each of these terms. For example, $\int(1+R(x))^{2} \sigma^{2 H} / f(x) d x=\int f(x)(1+$
$R(x))^{2} \sigma^{2 H} / f^{2}(x) d x \lesssim \sigma^{2\left(H-H_{1}\right)}$, as $f(x) \geq \sigma^{H_{1}}$ on $E_{\sigma}$ and the Lebesgue measure of this interval is at most $\sigma^{-H_{1}}$. To bound the second integral in (26) we use once more that $K_{\sigma} h_{k} \gtrsim f$, and then apply (23) with $m=0$. For the last integral we use (23) with $m=0, \ldots, k+1$; recall that $h_{k}$ is a linear combination of $K_{\sigma}^{m} f, m=0, \ldots, k$.

The second integral in (10) is bounded by

$$
\int_{A_{\sigma}^{c} \cup E_{\sigma}^{c}} f(x)\left(\log \frac{f(x)}{K_{\sigma} h_{k}(x)}\right)^{2} d x+\int_{A_{\sigma} \cap E_{\sigma}} \frac{\left(f(x)-K_{\sigma} h_{k}(x)\right)^{2}}{K_{\sigma} h_{k}(x)} d x
$$

which is $O\left(\sigma^{2 \beta}\right)$ by the same arguments.

The continuous mixture approximation of Theorem 1 is discretized in Lemma 4 below. Apart from the finite mixture derived from $h_{k}$ we also need to construct a set of finite mixtures close to it, such that this entire set is contained in a KL-ball around $f$. For this purpose the following lemma is useful. A similar result can be found in Lemma 5 of [13]. The inequality for the $L_{1}$-norm will be used in the entropy calculation in the proof of Theorem 2.
Lemma 3. Let $w, \tilde{w} \in \Delta_{k}, \mu, \tilde{\mu} \in \mathbb{R}^{k}$ and $\sigma, \tilde{\sigma} \in \mathbb{R}^{+}$. Let $\psi$ be a differentiable symmetric density such that $x \psi^{\prime}(x)$ is bounded. Then for mixtures $m(x)=m(x ; k, \mu, w, \sigma)$ and $\tilde{m}(x)=m(x ; k, \tilde{\mu}, \tilde{w}, \tilde{\sigma})$ we have

$$
\begin{aligned}
\|m-\tilde{m}\|_{1} & \leq\|w-\tilde{w}\|_{1}+2\|\psi\|_{\infty} \sum_{i=1}^{k} \frac{w_{i} \wedge \tilde{w}_{i}}{\sigma \wedge \tilde{\sigma}}\left|\mu_{i}-\tilde{\mu}_{i}\right|+\frac{|\sigma-\tilde{\sigma}|}{\sigma \wedge \tilde{\sigma}} \\
\|m-\tilde{m}\|_{\infty} & \lesssim \sum_{i=1}^{k} \frac{\left|w_{i}-\tilde{w}_{i}\right|}{\sigma \wedge \tilde{\sigma}}+\sum_{i=1}^{k} \frac{w_{i} \wedge \tilde{w}_{i}}{(\sigma \wedge \tilde{\sigma})^{2}}\left|\mu_{i}-\tilde{\mu}_{i}\right|+\frac{|\sigma-\tilde{\sigma}|}{(\sigma \wedge \tilde{\sigma})^{2}}
\end{aligned}
$$

Proof. Let $1 \leq i \leq k$ and assume that $\tilde{w}_{i} \leq w_{i}$. By the triangle inequality,

$$
\begin{aligned}
& \left\|w_{i} \psi_{\sigma}\left(\cdot-\mu_{i}\right)-\tilde{w}_{i} \psi_{\tilde{\sigma}}\left(\cdot-\tilde{\mu}_{i}\right)\right\| \leq\left\|w_{i} \psi_{\sigma}\left(\cdot-\mu_{i}\right)-\tilde{w}_{i} \psi_{\sigma}\left(\cdot-\mu_{i}\right)\right\| \\
& \quad+\left\|\tilde{w}_{i} \psi_{\sigma}\left(\cdot-\mu_{i}\right)-\tilde{w}_{i} \psi_{\sigma}\left(\cdot-\tilde{\mu}_{i}\right)\right\|+\left\|\tilde{w}_{i} \psi_{\sigma}\left(\cdot-\tilde{\mu}_{i}\right)-\tilde{w}_{i} \psi_{\tilde{\sigma}}\left(\cdot-\tilde{\mu}_{i}\right)\right\|
\end{aligned}
$$

for any norm. We have the following inequalities:

$$
\begin{align*}
\left\|\psi_{\sigma}\left(z-\mu_{i}\right)-\psi_{\sigma}\left(z-\tilde{\mu}_{i}\right)\right\|_{1} & =2\left|\Psi\left(\frac{\mu_{i}-\tilde{\mu}_{i}}{2 \sigma}\right)-\Psi\left(\frac{\tilde{\mu}_{i}-\mu_{i}}{2 \sigma}\right)\right| \\
& \leq 2\|\psi\|_{\infty} \frac{\left|\tilde{\mu}_{i}-\mu_{i}\right|}{\sigma} \leq \frac{2\|\psi\|_{\infty}}{\sigma \wedge \tilde{\sigma}}\left|\tilde{\mu}_{i}-\mu_{i}\right| \\
\left\|\psi_{\sigma}-\psi_{\tilde{\sigma}}\right\|_{1} & \leq \frac{1}{\sigma \wedge \tilde{\sigma}} \int\left|\psi\left(\frac{x}{\sigma}\right)-\psi\left(\frac{x}{\tilde{\sigma}}\right)\right| d x \leq \frac{1}{\sigma \wedge \tilde{\sigma}}|\sigma-\tilde{\sigma}| \\
\left\|\psi_{\sigma}-\psi_{\tilde{\sigma}}\right\|_{\infty} & \leq \frac{1}{(\sigma \wedge \tilde{\sigma})^{2}}\left\|\frac{d}{d z} g_{x}\right\|_{\infty}|\sigma-\tilde{\sigma}|  \tag{27}\\
\left\|\psi_{\sigma}\left(z-\mu_{i}\right)-\psi_{\sigma}\left(z-\tilde{\mu}_{i}\right)\right\|_{\infty} & \lesssim \frac{1}{(\sigma \wedge \tilde{\sigma})^{2}}\left|\tilde{\mu}_{i}-\mu_{i}\right|
\end{align*}
$$

To prove (27), let $\sigma=z^{-1}$ and $\tilde{\sigma}=\tilde{z}^{-1}$, and for fixed $x$ define the function $g_{x}: z \rightarrow z \psi(z x)$. By assumption, $\frac{d}{d z} g_{x}(z)=\psi(z x)+z x \psi^{\prime}(z x)$ is bounded, and

$$
\left\|\psi_{\sigma}-\psi_{\tilde{\sigma}}\right\|_{\infty}=\sup _{x}\left|g_{x}(z)-g_{x}(\tilde{z})\right| \leq|z-\tilde{z}|\left\|\frac{d}{d z} g_{x}\right\|_{\infty} \leq \frac{1}{(\sigma \wedge \tilde{\sigma})^{2}}\left\|\frac{d}{d z} g_{x}\right\|_{\infty}|\sigma-\tilde{\sigma}| .
$$

Applying the mean value theorem to $\psi$ itself, the last inequality is obtained.
The approximation $h_{k}$ defined by (21) can be discretized such that the result of Lemma 1 still holds. The discretization relies on Lemma 3.13 in [19], which is included in Appendix F. As in [2] and [28] (XXX), we require the kernel $\psi$ to be analytic. i.e. $p$ needs to be even.
Lemma 4. Let the constant $H_{1}$ in the definition of $E_{\sigma}$ be at least $4(\beta+p)$. Given $\beta>0$, let $f$ be a density that satisfies conditions (C1)-(C4) and for $p=2,4, \ldots$ let $\psi$ be as in (2). Then there exists a finite mixture $m=m\left(\cdot ; k_{\sigma}, \mu_{\sigma}, w_{\sigma}, \sigma\right)$ with $k_{\sigma}=O\left(\sigma^{-1}|\log \sigma|^{1+p^{-1}}\right)$ support points contained in $E_{\sigma}$, such that

$$
\begin{equation*}
\int f \log \frac{f}{m}=O\left(\sigma^{2 \beta}\right), \quad \int f\left(\log \frac{f}{m}\right)^{2}=O\left(\sigma^{2 \beta}\right) \tag{28}
\end{equation*}
$$

Furthermore, (28) holds for all mixtures $m^{\prime}=m\left(\cdot ; k_{\sigma}, \mu, w, \sigma^{\prime}\right)$ such that $\sigma^{\prime} \in$ $\left[\sigma, \sigma+\sigma^{\delta^{\prime} H_{1}+2}\right], \mu \in B_{k_{\sigma}}\left(\mu_{\sigma}, \sigma^{\delta^{\prime} H_{1}+2}\right)$ and $w \in \Delta_{k_{\sigma}}\left(w_{\sigma}, \sigma^{\delta^{\prime} H_{1}+1}\right)$, where $\delta^{\prime} \geq$ $1+\beta / H_{1}$.

The proof can be found in Appendix E. A discretization assuming only (C1),(C2) and (C4) could be derived similarly, but to have sufficient control of the number of components in Theorem 2, we make the stronger assumption (C3) of exponential tails. Together with the monotonicity condition (C4) this implies the existence of a finite constant $c_{f}$ such that for all sufficiently small $\epsilon$,

$$
\begin{equation*}
\{x: f(x) \geq \epsilon\} \subset\left[-c_{f}|\log \epsilon|^{1 / p}, c_{f}|\log \epsilon|^{1 / p}\right]=T_{c_{f}, \epsilon} \tag{29}
\end{equation*}
$$

The constant $c_{f}$ depends on $f$ by the constant $M_{f}$ in (9). This property is used in the proof of Lemma 4.

## 4. The proof of Theorem 2

We first state a lemma needed for the entropy calculations.
Lemma 5. For positive vectors $b=\left(b_{1}, \ldots, b_{k}\right)$ and $d=\left(d_{1}, \ldots, d_{k}\right)$, with $b_{i}<d_{i}$ for all $i$, the packing numbers of $\Delta_{k}$ and $H_{k}[b, d]$ satisfy

$$
\begin{align*}
D\left(\epsilon, \Delta_{k}, l_{1}\right) & \leq\left(\frac{5}{\epsilon}\right)^{k-1}  \tag{30}\\
D\left(\epsilon, H_{k}[b, d], l_{1}\right) & \leq \frac{k!\prod_{i=1}^{k}\left(d_{i}-b_{i}+2 \epsilon\right)}{(2 \epsilon)^{k}} . \tag{31}
\end{align*}
$$

Proof. A proof of (30) can be found in [11]; the other result follows from a volume argument. For $\lambda_{k}$ the $k$-dimensional Lebesgue measure, $\lambda_{k}\left(S_{k}\right)=\frac{1}{k!}$ and $\lambda_{k}\left(B_{k}\left(y, \frac{\epsilon}{2}, l_{1}\right)\right)=\frac{\epsilon^{k}}{k!}$, where $B_{k}\left(y, \frac{\epsilon}{2}, l_{1}\right)$ is the $l_{1}$-ball in $\mathbb{R}^{k}$ centered at $y$, with radius $\frac{\epsilon}{2}$. Suppose $x_{1}, \ldots, x_{N}$ is a maximal $\epsilon$-separated set in $H_{k}[b, d]$. If the center $y$ of an $l_{1}$-ball of radius $\frac{\epsilon}{2}$ is contained in $H_{k}[b, d]$ then for any point $z$ in this ball, $\left|z_{i}-y_{i}\right| \leq \frac{\epsilon}{2}$ for all $i$. Because for each coordinate we have the bounds $\left|z_{i}\right| \leq\left|y_{i}\right|+\left|z_{i}-y_{i}\right| \leq d_{i}+\frac{\epsilon}{2}$ and $\left|z_{i}\right| \geq b_{i}-\frac{\epsilon}{2}, z$ is an element of $H_{k}\left[b-\frac{\epsilon}{2}, d+\frac{\epsilon}{2}\right]$. The union of the balls $B_{k}\left(x_{1}, \frac{\epsilon}{2}, l_{1}\right), \ldots, B_{k}\left(x_{N}, \frac{\epsilon}{2}, l_{1}\right)$ is therefore contained in $H_{k}\left[b-\frac{\epsilon}{2}, d+\frac{\epsilon}{2}\right]$.

Proof of Theorem 2. The proof is an application of Theorem 3 in [13] (stated below in appendix A), with sequences $\widetilde{\epsilon}_{n}=n^{-\beta /(1+2 \beta)}(\log n)^{t_{1}}$ and $\bar{\epsilon}_{n}=n^{-\beta /(1+2 \beta)}(\log n)^{t_{2}}$, where $t_{1}$ and $t_{2} \geq t_{1}$ are determined below. Let $k_{n}$ be the number of components in Lemma 4 when $\sigma=\sigma_{n}=\widetilde{\epsilon}_{n}^{1 / \beta}$. This lemma then provides a $k_{n^{-}}$ dimensional mixture $m=m\left(\cdot ; k_{n}, \mu^{(n)}, w^{(n)}, \sigma_{n}\right)$ whose KL-divergence from $f_{0}$ is $O\left(\sigma_{n}{ }^{2 \beta}\right)=O\left(\tilde{\epsilon}_{n}^{2}\right)$. The number of components is

$$
\begin{equation*}
k_{n}=O\left(\sigma_{n}^{-1}\left|\log \sigma_{n}\right|^{1+p^{-1}}\right)=O\left(n^{1 /(1+2 \beta)}(\log n)^{1+p^{-1}-t_{1} / \beta}\right) \tag{32}
\end{equation*}
$$

their locations being contained in the set $E_{\sigma_{n}}$ defined in (22). By the same lemma there are $l_{1}$-balls $B_{n}=B_{k_{n}}\left(\mu^{(n)}, \sigma_{n}{ }^{\delta^{\prime} H_{1}+2}\right)$ and $\Delta(n)=\Delta_{k_{n}}\left(w^{(n)}, \sigma_{n}{ }^{\delta^{\prime} H_{1}+1}\right)$ such that the same is true for all $k_{n}$-dimensional mixtures $m=m\left(\cdot ; k_{n}, \mu, w, \sigma\right)$ with $\sigma \in\left[\sigma_{n}, \sigma_{n}+\sigma_{n}{ }^{\delta^{\prime} H_{1}+2}\right]$ and $(\mu, w) \in B_{n} \times \Delta(n)$. It now suffices to lower bound the prior probability on having $k_{n}$ components and on $B_{n}, \Delta(n)$ and $\left[\sigma_{n}, \sigma_{n}+\sigma_{n}{ }^{\delta^{\prime} H_{1}+2}\right]$.

Let $b=\delta^{\prime} H_{1}+2$; as $\sigma^{-1}$ is inverse-gamma, it follows from the mean value theorem that

$$
\begin{align*}
\Pi\left(\sigma \in\left[\sigma_{n}, \sigma_{n}+\sigma_{n}^{b}\right]\right) & =\int_{\sigma_{n}}^{\sigma_{n}+\sigma_{n}^{b}} \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\lambda / x} d x \\
& \geq \int_{\sigma_{n}}^{\sigma_{n}+\sigma_{n}^{b}} \frac{\lambda^{\alpha}}{\Gamma(\alpha)} e^{-2 \lambda / x} d x \geq 4 \frac{\lambda^{\alpha+1}}{\Gamma(\alpha)} \sigma_{n}^{b-2} e^{-\lambda \sigma_{n}-1} \tag{33}
\end{align*}
$$

which is larger than $\exp \left\{-n \widetilde{\epsilon}_{n}^{2}\right\}$ for any choice of $t_{1} \geq 0$. Condition (11) gives a lower bound of $B_{0} \exp \left\{-b_{0} k_{n} \log ^{r} k_{n}\right\}$ on $\Pi\left(k_{n}\right)$, which is larger than $\exp \left\{-n \widetilde{\epsilon}_{n}^{2}\right\}$ when $\left(2+\beta^{-1}\right) t_{1}>1+p^{-1}+r$. Given that there are $k_{n}$ components, condition (14) gives a lower bound on $\Pi(\Delta(n))$, which is larger than $\exp \left\{-n \widetilde{\epsilon}_{n}^{2}\right\}$ when $\left(2+\beta^{-1}\right) t_{1}>2+b+p^{-1}$. The required lower-bound for $\Pi\left(B_{n}\right)$ follows from (9) and the fact that $\mu_{1}^{(n)}, \ldots, \mu_{k_{n}}^{(n)}$ are independent with prior density $p_{\mu}$ satisfying (12). The 'target' mixture given by Lemma 4 has location vector $\mu^{(n)}$, whose elements are contained in $E_{\sigma_{n}}$. By (9), $E_{\sigma_{n}}$ is contained in the interval $T_{c_{f}, \epsilon}$ defined in (29), with $\epsilon=\sigma_{n}{ }^{H_{1}}$. Since $p_{\mu} \gtrsim \psi, p_{\mu}$ is lower bounded by a multiple of $\sigma_{n} c_{f}^{p} H_{1}$ at the boundaries of this interval. Consequently, for all $i=1, \ldots, k_{n}$,

$$
\Pi\left(\left|\mu_{i}-\mu_{i}^{(n)}\right| \leq \frac{\sigma_{n}^{\delta^{\prime} H_{1}+2}}{k_{n}}\right) \gtrsim \frac{\sigma_{n}^{\delta^{\prime} H_{1}+2+c_{f}^{p} H_{1}}}{k_{n}}
$$

As the $l_{1}$-ball $B_{k_{n}}\left(\mu^{(n)}, \sigma_{n}{ }^{\delta^{\prime} H_{1}+2}\right)$ contains the $l_{\infty}$-ball $\left\{\mu \in \mathbb{R}^{k_{n}}:\left|\mu_{i}-\mu_{i}^{(n)}\right| \leq\right.$ $\left.\frac{\sigma_{n}{ }^{\delta^{\prime} H_{1}+2}}{k_{n}}, 1 \leq i \leq k_{n}\right\}$, we conclude that

$$
\Pi\left(\mu \in B_{n}\right) \gtrsim \exp \left\{-d k_{n} \log n\right\}
$$

for some constant $d>0$. Combining the above results it follows that $\Pi\left(K L\left(f_{0}, \widetilde{\epsilon}_{n}\right)\right) \geq$ $\exp \left\{-n \widetilde{\epsilon}_{n}^{2}\right\}$ when $t_{1}>\left(2+b+p^{-1}\right) /\left(2+\beta^{-1}\right)$.

We then have to find sets $\mathcal{F}_{n}$ such that (40) and (42) hold. For $r_{n}=$ $n^{\frac{1}{1+2 \beta}}(\log n)^{t_{r}}$ (rounded to the nearest integer) and a polynomially increasing sequence $b_{n}$ such that $b_{n}^{a_{2}}>n^{1 /(1+2 \beta)}$, with $a_{2}$ as in (13), we define

$$
\mathcal{F}_{n}=\left\{m(\cdot ; k, \mu, w, \sigma) \mid k \leq r_{n}, \mu \in H_{k}\left[-b_{n}, b_{n}\right], \sigma \in S_{n}\right\} .
$$

The bandwidth $\sigma$ is contained in $S_{n}=\left(\underline{\sigma}_{n}, \bar{\sigma}_{n}\right]$, where $\underline{\sigma}_{n}=n^{-A}$ and $\bar{\sigma}_{n}=$ $\exp \left\{n \tilde{\epsilon}_{n}^{2}(\log n)^{\delta}\right\}$, for arbitrary constants $A>1$ and $\delta>0$. An upper bound on $\Pi\left(S_{n}^{c}\right)$ can be found by direct calculation, for example

$$
\begin{aligned}
\int_{\bar{\sigma}_{n}}^{\infty} \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\frac{\lambda}{x}} d x & =\int_{0}^{\bar{\sigma}_{n}^{-1}} \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} d x \\
& \leq \int_{0}^{\bar{\sigma}_{n}^{-1}} \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} d x=O\left(\exp \left\{-\alpha n \tilde{\epsilon}_{n}^{2}(\log n)^{\delta}\right\}\right)
\end{aligned}
$$

Hence $\Pi\left(S_{n}^{c}\right) \leq e^{-c n \tilde{\epsilon}_{n}^{2}}$ for any constant $c$, for large enough $n$. The prior mass on mixtures with more than $r_{n}$ support points is bounded by a multiple of $\exp \left\{-b_{1} k_{n} \log ^{r_{n}} k_{n}\right\}$. The prior mass on mixtures with at least one support point outside $\left[-b_{n}, b_{n}\right]$ is controlled as follows. By conditions (11) and (13), the probability that a certain $\mu_{i}$ is outside $\left[-b_{n}, b_{n}\right]$, is

$$
\begin{equation*}
\Pi\left(\left|\mu_{i}\right|>b_{n}\right)=\int_{\left[-b_{n}, b_{n}\right] c} p_{\mu}(x) d x \lesssim b_{n}^{\max \left\{0,1-a_{2}\right\}} e^{-b_{n}^{a_{2}}} \tag{34}
\end{equation*}
$$

Since the prior on $k$ satisfies (11), $k$ clearly has finite expectation. Consequently, (34) implies that

$$
\begin{align*}
\Pi\left(N\left(\left[-b_{n}, b_{n}\right]^{c}\right)>0\right) & =\sum_{k=1}^{\infty} \Pi(K=k) \Pi\left(\max _{i=1, \ldots, k}\left|\mu_{i}\right|>b_{n} \mid K=k\right)  \tag{35}\\
& \leq \sum_{k=1}^{\infty} \Pi(k) k \Pi\left(\left|\mu_{i}\right|>b_{n}\right) \lesssim e^{-\left|b_{n}\right|^{a_{2}}}
\end{align*}
$$

Combining these bounds, we find

$$
\Pi\left(\mathcal{F}_{n}^{c}\right) \leq \Pi\left(S_{n}^{c}\right)+\sum_{k=r_{n}}^{\infty} \rho(k)+\Pi\left(N\left(\left[-b_{n}, b_{n}\right]^{c}>0\right)\right) \lesssim e^{-b_{1} r_{n}(\log n)^{r}}
$$

The right hand side decreases faster than $e^{-n \tilde{\epsilon}_{n}^{2}}$ if $t_{r}+r>2 t_{1}$.

To control the sum in (40), we partition $\mathcal{F}_{n}$ using

$$
\begin{gathered}
\mathcal{F}_{n, j}=\left\{m(\cdot ; k, \mu, w, \sigma) \mid k \leq r_{n}, \mu \in H_{k}\left[-b_{n}, b_{n}\right], \sigma \in S_{n, j}\right\}, \\
S_{n, j}=\left(s_{n, j-1}, s_{n, j}\right]=\left(\underline{\sigma}_{n}\left(1+\widetilde{\epsilon}_{n}\right)^{j-1}, \underline{\sigma}_{n}\left(1+\widetilde{\epsilon}_{n}\right)^{j}\right], \quad j=1, \ldots, J_{n}, \\
J_{n}=\left(\log \frac{\bar{\sigma}_{n}}{\underline{\sigma}_{n}}\right) / \log \left(1+\epsilon_{n}\right)=O\left(n \widetilde{\epsilon}_{n}(\log n)^{\delta}\right) .
\end{gathered}
$$

An upper bound on the prior probability on the $\mathcal{F}_{n, j}$ is again found by direct calculation:

$$
\begin{align*}
\Pi\left(\mathcal{F}_{n, j}\right) & \leq \Pi\left(S_{n, j}\right)=\Pi\left(\sigma^{-1} \in\left[\underline{\sigma}_{n}^{-1}\left(1+\widetilde{\epsilon}_{n}\right)^{-j}, \underline{\sigma}_{n}^{-1}\left(1+\widetilde{\epsilon}_{n}\right)^{1-j}\right)\right) \\
& =\int_{\underline{\sigma}_{n}^{-1}\left(1+\widetilde{\epsilon}_{n}\right)^{-j}}^{\sigma_{n}^{-1}\left(1+\widetilde{\epsilon}_{n}\right)^{1-j}} y^{\alpha-1} e^{-\lambda y} d y \\
& \leq \lambda^{-1} \max \left\{\left(\underline{\sigma}_{n}^{-1}\left(1+\widetilde{\epsilon}_{n}\right)^{-j}\right)^{\alpha-1},\left(\underline{\sigma}_{n}^{-1}\left(1+\widetilde{\epsilon}_{n}\right)^{1-j}\right)^{\alpha-1}\right\} \exp \left\{-\lambda \underline{\sigma}_{n}^{-1}\left(1+\widetilde{\epsilon}_{n}\right)^{-j}\right\} \\
& \lesssim \underline{\sigma}_{n}^{1-\alpha}\left(1+\widetilde{\epsilon}_{n}\right)^{-(\alpha-1) j} \exp \left\{-\lambda \underline{\sigma}_{n}^{-1}\left(1+\widetilde{\epsilon}_{n}\right)^{-j}\right\} . \tag{36}
\end{align*}
$$

As the $L_{1}$-distance is bounded by the Hellinger-distance, condition (40) only needs to be verified for the $L_{1}$-distance. We further decompose the $\mathcal{F}_{n, j}$ 's and write

$$
\mathcal{F}_{n, j}=\cup_{k=1}^{r_{n}} \mathcal{F}_{n, j, k}=\cup_{k=1}^{r_{n}}\left\{m(\cdot ; k, \mu, w, \sigma) \mid \mu \in H_{k}\left[-b_{n}, b_{n}\right], \sigma \in S_{n, j}\right\} .
$$

It will be convenient to replace the covering numbers $N$ in (40) by their corresponding packing numbers $D$, which are at least as big. Since for any pair of metric spaces $\left(A, d_{1}\right)$ and $\left(B, d_{2}\right)$ we have $D\left(\epsilon, A \times B, d_{1}+d_{2}\right) \leq D\left(\frac{\epsilon}{2}, A, d_{1}\right) D\left(\frac{\epsilon}{2}, B, d_{2}\right)$, Lemma 3 implies that for all $k \geq 1, D\left(\bar{\epsilon}_{n}, \mathcal{F}_{n, j, k},\|\cdot\|_{1}\right)$ is bounded by

$$
D\left(\frac{\bar{\epsilon}_{n}}{3}, \Delta_{k}, l_{1}\right) D\left(\frac{\bar{\epsilon}_{n} s_{n, j-1}}{6\|\psi\|_{\infty}}, H_{k}\left[-b_{n}, b_{n}\right], l_{1}\right) D\left(\frac{\bar{\epsilon}_{n} s_{n, j-1}}{3},\left(s_{n, j-1}, s_{n, j}\right], l_{1}\right)
$$

Lemma 5 provides the following bounds:

$$
\begin{aligned}
D\left(\frac{\bar{\epsilon}_{n}}{3}, \Delta_{k}, l_{1}\right) & \leq\left(\frac{15}{\bar{\epsilon}_{n}}\right)^{k-1} \\
D\left(\frac{\bar{\epsilon}_{n} s_{n, j-1}}{6\|\psi\|_{\infty}}, H_{k}\left[-b_{n}, b_{n}\right], l_{1}\right) & \leq k!\left(\frac{\bar{\epsilon}_{n} s_{n, j-1}}{3\|\psi\|_{\infty}}\right)^{-k} \prod_{i=1}^{k}\left(2 b_{n}+\frac{\bar{\epsilon}_{n} s_{n, j-1}}{3\|\psi\|_{\infty}}\right) \\
D\left(\frac{\bar{\epsilon}_{n} s_{n, j-1}}{3},\left(s_{n, j-1}, s_{n, j}\right], l_{1}\right) & \leq\left(s_{n, j-1} \bar{\epsilon}_{n} / 3\right)\left(\left(s_{n, j}-s_{n, j-1}\right)+\bar{\epsilon}_{n} s_{n, j-1} / 3\right)
\end{aligned}
$$

For some constant $C$, we find that

$$
\begin{align*}
D\left(\bar{\epsilon}_{n}, \mathcal{F}_{n, j},\|\cdot\|_{1}\right) & \leq r_{n} D\left(\bar{\epsilon}_{n}, \mathcal{F}_{n, j, r_{n}},\|\cdot\|_{1}\right) \\
& \lesssim r_{n} C^{r_{n}} r_{n}!\left(\bar{\epsilon}_{n}\right)^{-2 r_{n}} s_{n, j} s_{n, j-1}^{-r_{n}+1}\left(\max \left(b_{n}, \bar{\epsilon}_{n} s_{n, j-1}\right)\right)^{r_{n}} \tag{37}
\end{align*}
$$

If $b_{n} \geq \bar{\epsilon}_{n} s_{n, j-1}$, we have $\left(1+\widetilde{\epsilon}_{n}\right)^{-j} \geq \frac{\bar{\epsilon}_{n} \underline{\sigma}_{n}}{b_{n}\left(1+\tilde{\epsilon}_{n}\right)}$, and the last exponent in (36) is bounded by $-\lambda b_{n}^{-1} \bar{\epsilon}_{n} /\left(1+\widetilde{\epsilon}_{n}\right)$. A combination of (36), (37) and Stirling's bound on $r_{n}$ ! then imply that $\sqrt{\Pi\left(\mathcal{F}_{n, j}\right)} \sqrt{N\left(\bar{\epsilon}_{n}, \mathcal{F}_{n, j}, d\right)}$ is bounded by a multiple of

$$
\begin{aligned}
& \underline{\sigma}_{n}^{(1-\alpha) / 2}\left(1+\widetilde{\epsilon}_{n}\right)^{-(\alpha-1) j / 2} \sqrt{r_{n}} C^{r_{n} / 2} r_{n}^{r_{n} / 2+1 / 2}\left(\bar{\epsilon}_{n}\right)^{-r_{n}} \sqrt{s_{n, j}} \\
& \quad s_{n, j-1}^{-r_{n} / 2+1 / 2} b_{n}^{r_{n} / 2} \exp \left\{-\frac{\lambda}{2} \underline{\sigma}_{n}^{-1}\left(1+\widetilde{\epsilon}_{n}\right)^{-j}\right\} \\
& \lesssim n^{\frac{A}{2} r_{n}+\frac{\alpha-3}{2} A}\left(1+\widetilde{\epsilon}_{n}\right)^{-\frac{1}{2}(j-1)\left(r_{n}+\alpha-2\right)+\frac{1-\alpha}{2}}\left(r_{n}+1\right)^{r_{n}+1} C^{\frac{r_{n}}{2}} \bar{\epsilon}_{n}^{-r_{n}} b_{n}^{\frac{r_{n}}{2}} \exp \left\{-\lambda b_{n}^{-1} \frac{\bar{\epsilon}_{n}}{1+\widetilde{\epsilon}_{n}}\right\} \\
& \lesssim K_{0} \exp \left\{K_{1} r_{n}(\log n)\right\},
\end{aligned}
$$

for certain constants $C, K_{0}$ and $K_{1}$. If $b_{n}<\bar{\epsilon}_{n} s_{n, j-1}$ we obtain similar bound but with an additional factor $\bar{\epsilon}_{n}^{-r_{n} / 2} n^{-A r_{n} / 2}\left(1+\widetilde{\epsilon}_{n}\right)^{(j-1) r_{n} / 2}$, where the factor $\left(1+\widetilde{\epsilon}_{n}\right)^{(j-1) r_{n} / 2}$ cancels out with $\left(1+\widetilde{\epsilon}_{n}\right)^{-(j-1) r_{n} / 2}$ on the third line of the above display. There is however a remaining factor $\left(1+\widetilde{\epsilon}_{n}\right)^{\frac{1}{2}(j-1)(2-\alpha)}$. Since $J_{n}$ is defined such that $n^{-A}\left(1+\widetilde{\epsilon}_{n}\right)^{J_{n}}=\exp \left\{n \widetilde{\epsilon}_{n}^{2}(\log n)^{\delta}\right\}$, the sum of $\sqrt{\Pi\left(\mathcal{F}_{n, j}\right)} \sqrt{N\left(\bar{\epsilon}_{n}, \mathcal{F}_{n, j}, d\right)}$ over $j=1, \ldots, J_{n}$ is a multiple of $\exp \left\{K_{1} r_{n}(\log n)+\right.$ $\left.n \widetilde{\epsilon}_{n}^{2}(\log n)^{\delta}\right\}$, which increases at a slower rate than $\exp \left\{n \bar{\epsilon}_{n}^{2}\right\}$ if $2 t_{2}>\max \left(t_{r}+\right.$ $\left.1,2 t_{1}+\delta\right)$. Combined with the requirement that $t_{r}+r>2 t_{1}$ this gives $t_{2}>$ $t_{1}+\frac{1-r}{2}$. Hence the convergence rate is $\epsilon_{n}=n^{-\beta /(1+2 \beta)}(\log n)^{t}$, with $t>$ $\left(2+b+p^{-1}\right) /\left(2+p^{-1}\right)+\max (0,(1-r) / 2)$.

## 5. Examples of priors on the weights

Condition (14) on the weights-prior is known to hold for the Dirichlet distribution. We now address the question whether it also holds for other priors. Alternatives to Dirichlet-priors are increasingly popular, see for example [16]. In this section two classes of priors on the simplex are considered. In both cases the Dirichlet distribution appears as a special case. The proof of Theorem 2 requires lower bounds for the prior mass on $l_{1}$-balls around some fixed point in the simplex. These bounds are given in Lemmas 6 and 8 below.

Since a normalized vector of independent gamma distributed random variables is Dirichlet distributed, a straightforward generalization is to consider random variables with an alternative distribution on $\mathbb{R}^{+}$. Given independent random variables $Y_{1}, \ldots, Y_{k}$ with densities $f_{i}$ on $[0, \infty)$, define a vector $X$ with elements $X_{i}=Y_{i} /\left(Y_{1}+\ldots+Y_{k}\right), i=1, \ldots, k$. For $\left(x_{1}, \ldots, x_{k-1}\right) \in S_{k-1}$,

$$
\begin{align*}
P\left(X_{1}\right. & \left.\leq x_{1}, \ldots, X_{k-1} \leq x_{k-1}\right)=\int_{0}^{\infty} P\left(Y_{1} \leq x_{1} y, \ldots Y_{k-1} \leq x_{k-1} y\right) d P^{Y_{1}+\ldots+Y_{k}}(y) \\
& =\int_{0}^{\infty} \int_{0}^{x_{1} y} \int_{0}^{x_{2} y} \cdots \int_{0}^{x_{k-1} y} f_{k}\left(y-\sum_{i=1}^{k-1} s_{i}\right) \prod_{i=1}^{k-1} f_{i}\left(s_{i}\right) d s_{1} \cdots d s_{k-1} d y \tag{38}
\end{align*}
$$

The corresponding density is

$$
\begin{align*}
f^{X_{1}, \ldots, X_{k-1}}\left(x_{1}, \ldots, x_{k-1}\right) & =\int_{0}^{\infty} y^{k-1} f_{k}\left(y-\sum_{i=1}^{k-1} x_{i} y\right) \prod_{i=1}^{k-1} f_{i}\left(x_{i} y\right) d y  \tag{39}\\
& =\int_{0}^{\infty} y^{k-1} \prod_{i=1}^{k} f_{i}\left(x_{i} y\right) d y
\end{align*}
$$

where $x_{k}=1-\sum_{i=1}^{k-1} x_{i}$. We obtain a result similar to lemma 8 in [13].
Lemma 6. Let $X_{1}, \ldots, X_{k}$ have a joint distribution with a density of the form (39). Assume there are positive constants $c_{1}(k), c_{2}(k)$ and $c_{3}$ such that for $i=$ $1, \ldots, k, f_{i}(z) \geq c_{1}(k) z^{c_{3}}$ if $z \in\left[0, c_{2}(k)\right]$. Then there are constants $c$ and $C$ such that for all $y \in \Delta_{k}$ and all $\epsilon \leq\left(\frac{1}{k} \wedge c_{1}(k) c_{2}(k)^{c_{3}+1}\right)$

$$
P\left(X \in \Delta_{k}(y, 2 \epsilon)\right) \geq C e^{-c k \log \left(\frac{1}{\epsilon}\right)}
$$

Proof. As in [13] it is assumed that $y_{k} \geq k^{-1}$. Define $\underline{\delta_{i}}=\max \left(0, y_{i}-\epsilon^{2}\right)$ and $\bar{\delta}_{i}=\min \left(1, y_{i}+\epsilon^{2}\right)$. If $x_{i} \in\left(\underline{\delta_{i}}, \bar{\delta}_{i}\right)$ for $i=1, \ldots, \bar{k}-1$, then $\sum_{i=1}^{k} \mid$ $x_{i}-y_{i}\left|\leq 2 \sum_{i=1}^{k-1}\right| x_{i}-y_{i} \mid \leq 2(k-1) \epsilon^{2} \leq \epsilon$. Note that $\left(x_{1}, \ldots, x_{k-1}\right) \in S_{k}$, as $\sum_{j=1}^{k-1} x_{j} \leq \frac{k-1}{k}+(k-1) \epsilon^{2}<1$. Since all $x_{i}$ in (39) are at most one,
$f\left(x_{1}, \ldots, x_{k-1}\right) \geq \int_{0}^{c_{2}(k)} y^{k-1} \prod_{i=1}^{k}\left(c_{1}(k)\left(x_{i} y\right)^{c_{3}}\right) d y=\frac{\left(c_{2}(k)^{c_{3}+1} c_{1}(k)\right)^{k}}{\left(c_{3}+1\right) k}\left(x_{1} \cdot \ldots \cdot x_{k}\right)^{c_{3}}$.
Because

$$
\begin{gathered}
x_{k}=\left|1-\sum_{j=1}^{k-1} x_{j}\right|=\left|y_{k}+\sum_{j=1}^{k-1}\left(y_{j}-x_{j}\right)\right| \geq k^{-1}-(k-1) \epsilon^{2} \geq \epsilon^{2} \geq \frac{1}{k^{2}}, \\
P\left(X \in B_{k}(y, \epsilon)\right) \geq \frac{1}{k^{2 c_{3}}} \frac{\left(c_{2}(k)^{c_{3}+1} c_{1}(k)\right)^{k}}{\left(c_{3}+1\right) k} \prod_{j=1}^{k-1} \int_{\underline{\delta_{j}}}^{\bar{\delta}_{j}} x_{j}^{c_{3}} d x_{j} \geq \frac{\left(c_{2}(k)^{c_{3}+1} c_{1}(k)\right)^{k}}{\left(c_{3}+1\right)^{2} k} \epsilon^{2 k\left(c_{3}+1\right)-2} \\
\geq \exp \left\{k \log \left(c_{2}(k)^{c_{3}+1} c_{1}(k)\right)-\log \left(c_{3}+1\right)-\log (k)-2 k \log \left(\frac{\sqrt{2}}{\epsilon}\right)\right\} .
\end{gathered}
$$

As $\epsilon \leq\left(\frac{1}{k} \wedge c_{1}(k) c_{2}(k)^{c_{3}+1}\right)$, there are constants $c$ and $C$ for which this quantity is lower-bounded by $C e^{-c k \log \left(\frac{1}{\epsilon}\right)}$.

Alternatively, the Dirichlet distribution can be seen as a Polya tree. Following Lavine [21] we use the notation $E=\{0,1\}, E^{0}=\emptyset$ and for $m \geq 1, E^{m}=$ $\{0,1\}^{m}$. In addition, let $E_{*}^{m}=\cup_{i=0}^{m}\{0,1\}^{i}$. It is assumed that $k=2^{m}$ for some integer $m$, and the coordinates are indexed with binary vectors $\epsilon \in E^{m}$. A vector $X$ has a Polya tree distribution if

$$
X_{\epsilon}=\prod_{j=1, \epsilon_{j}=0}^{m} U_{\epsilon_{1} \cdots \epsilon_{j-1}} \prod_{j=1, \epsilon_{j}=1}^{m}\left(1-U_{\epsilon_{1} \cdots \epsilon_{j-1}}\right)
$$

where $\left(U_{\delta}, \delta \in E_{*}^{m-1}\right)$ is a family of beta random variables with parameters $\left(\left(\alpha_{\delta_{1}}, \alpha_{\delta_{2}}\right), \delta \in E_{*}^{m-1}\right)$. We only consider symmetric beta densities, for which $\alpha_{\delta}=\alpha_{\delta_{1}}=\alpha_{\delta_{2}}$. Adding pairs of coordinates, lower dimensional vectors $X_{\delta}$ can be defined for $\delta \in E_{*}^{m-1}$. For $\delta \in E_{*}^{m-1}$, let $X_{\delta 0}=U_{\delta} X_{\delta}$ and $X_{\delta 1}=\left(1-U_{\delta}\right) X_{\delta}$, and $X_{\emptyset}=1$ by construction. If $\alpha_{\delta}=\frac{1}{2} \alpha_{\delta_{1} \ldots \delta_{i-1}}$ for all $1 \leq i \leq m$ and $\delta \in E^{i}$, $X$ is Dirichlet distributed.

Lemma 7. Let $X$ have a Polya distribution with parameters $\alpha_{\delta}, \delta \in E_{*}^{m-1}$. Then for all $y \in \Delta_{2^{m}}$ and $\eta>0$,

$$
\begin{aligned}
p_{m}(y, \eta) & =P\left(X \in \Delta_{k}(y, \eta)\right)=P\left(\sum_{\epsilon \in E^{m}}\left|X_{\epsilon}^{m}-y_{\epsilon}^{m}\right| \leq \eta\right) \\
& \geq \prod_{i=1}^{m} P\left(\max _{\partial \in E^{i-1}}\left|U_{\delta}-\frac{y_{\delta 0}}{y_{\delta}}\right| \leq \frac{\eta}{2^{m-i+2}}\right) .
\end{aligned}
$$

Proof. For all $i=1, \ldots, m$ and $\delta \in E^{i-1}$,

$$
\begin{aligned}
\left|U_{\delta} X_{\delta}-y_{\delta 0}\right| & \leq U_{\delta}\left|X_{\delta}-y_{\delta}\right|+y_{\delta}\left|U_{\delta}-\frac{y_{\delta 0}}{y_{\delta}}\right| \\
\left|\left(1-U_{\delta}\right) X_{\delta}-y_{\delta 1}\right| & \leq\left(1-U_{\delta}\right)\left|X_{\delta}-y_{\delta}\right|+y_{\delta}\left|\left(1-U_{\delta}\right)-\frac{y_{\delta}-y_{\delta 0}}{y_{\delta}}\right|
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
\sum_{\delta \in E^{m}}\left|X_{\delta}-y_{\delta}\right| & =\sum_{\delta \in E^{m-1}}\left|X_{\delta 0}-y_{\delta 0}\right|+\left|X_{\delta 1}-y_{\delta 1}\right| \\
& \leq \sum_{\delta \in E^{m-1}}\left|X_{\delta}-y_{\delta}\right|+2 \sum_{\delta \in E^{m-1}} y_{\delta}\left|U_{\delta}-\frac{y_{\delta 0}}{y_{\delta}}\right| \\
& \leq \sum_{\delta \in E^{m-1}}\left|X_{\delta}-y_{\delta}\right|+2 \max _{\delta \in E^{m-1}}\left|U_{\delta}-\frac{y_{\delta 0}}{y_{\delta}}\right|
\end{aligned}
$$

Hence,

$$
\begin{aligned}
p_{m}(y, \eta) & \geq p_{m-1}\left(y, \frac{\eta}{2}\right) P\left(\max _{\partial \in E^{m-1}}\left|U_{\delta}-\frac{y_{\delta 0}}{y_{\delta}}\right| \leq \frac{\eta}{4}\right) \\
& \geq \prod_{i=2}^{m} P\left(\max _{\partial \in E^{i-1}}\left|U_{\delta}-\frac{y_{\delta 0}}{y_{\delta}}\right| \leq \frac{\eta}{2^{m-i+2}}\right) P\left(\left|U_{\emptyset}-y_{0}\right| \leq \frac{\eta}{2^{m}}\right) \\
& \geq \prod_{i=1}^{m} P\left(\max _{\partial \in E^{i-1}}\left|U_{\delta}-\frac{y_{\delta 0}}{y_{\delta}}\right| \leq \frac{\eta}{2^{m-i+2}}\right),
\end{aligned}
$$

as

$$
\begin{aligned}
p_{1}\left(\eta 2^{-m}\right) & =P\left(\left|X_{0}-y_{0}\right|+\left|X_{1}-y_{1}\right| \leq \eta 2^{-m}\right) \\
& =P\left(\left|U_{0}-y_{0}\right|+\left|\left(1-U_{0}\right)-\left(1-y_{0}\right)\right| \leq \eta 2^{-m}\right)=P\left(\left|U_{0}-y_{0}\right| \leq \eta 2^{-m-1}\right)
\end{aligned}
$$

With $\delta \in E^{i-1}$ fixed, we can lower-bound $P\left(\left|U_{\delta}-\frac{y_{\delta 0}}{y_{\delta}}\right| \leq \frac{\eta}{2^{m-i+2}}\right)$ for various values of the $\alpha_{\delta}$. In the remainder we will assume that $\alpha_{\delta}=\alpha_{i}$, for all $\delta \in E^{i-1}$, with $i=1, \ldots, m$. For increasing $\alpha_{i} \geq 1, U_{\delta}$ has a unimodal beta-density, and without loss of generality we can assume the most unfavorable case, i.e. when $\frac{y_{\delta 0}}{y_{\delta}}=0$. If the $\alpha_{i}$ are decreasing, and smaller than one, this is when $\frac{y_{\delta 0}}{y_{\delta}}=\frac{1}{2}$. In both cases Lemma 9 in appendix A is used to lower bound the normalizing constant of the beta-density.

$$
\text { If } \alpha_{i} \uparrow \infty, i=1, \ldots, m \text { when } m \rightarrow \infty \text {, then }
$$

$$
\begin{aligned}
P\left(\left|U_{\delta}\right| \leq \eta 2^{-m+i-2}\right) & =\int_{0}^{\eta 2^{-m+i-2}} \frac{\Gamma\left(2 \alpha_{i}\right)}{\Gamma^{2}\left(\alpha_{i}\right)} x^{\alpha_{i}-1}(1-x)^{\alpha_{i}-1} d x \\
& \gtrsim \int_{0}^{\eta 2^{-m+i-2}} \alpha_{i}{ }^{-\frac{1}{2}} 2^{2 \alpha_{i}-\frac{1}{2}} \frac{1}{2} x^{\alpha_{i}-1} d x=2^{-(m-i) \alpha_{i}-\frac{3}{2}} \alpha_{i}{ }^{-\frac{3}{2}} \eta^{\alpha_{i}}
\end{aligned}
$$

At the $i$ th level there are $2^{i-1}$ independent variables $U_{\delta}$ with the $\operatorname{Beta}\left(\alpha_{i}, \alpha_{i}\right)$ distribution, and therefore

$$
\begin{aligned}
\log \left(p_{m}(y, \eta)\right) & \gtrsim \log \prod_{i=1}^{m}\left(2^{-(m-i) \alpha_{i}-\frac{3}{2}} \alpha_{i}{ }^{-\frac{3}{2}} \eta^{\alpha_{i}}\right)^{2^{i-1}} \\
& =\sum_{i=1}^{m} 2^{i-1}\left\{-\alpha_{i} \log \frac{1}{\eta}-\frac{3}{2} \log \left(\alpha_{i}\right)-\alpha_{i}(m-i) \log (2)\right\} .
\end{aligned}
$$

If $\alpha_{i} \downarrow 0, i=1, \ldots, m$ when $m \rightarrow \infty$, we have

$$
\begin{aligned}
& P\left(\left|U_{\delta}-\frac{1}{2}\right| \leq \eta 2^{-m+i-2}\right)=\int_{1 / 2-\eta 2^{-m+i-2}}^{1 / 2+\eta 2^{-m+i-2}} \frac{\Gamma\left(2 \alpha_{i}\right)}{\Gamma^{2}\left(\alpha_{i}\right)} x^{\alpha_{i}-1}(1-x)^{\alpha_{i}-1} d x \\
& \gtrsim \alpha_{i} \eta 2^{-m+i-1}\left(\frac{1}{4}\right)^{\alpha_{i}-1} \\
& \log \left(p_{m}(y, \eta)\right) \gtrsim \sum_{i=1}^{m} 2^{i-1}\left\{\log \left(\alpha_{i}\right)-\left(2 \alpha_{i}+(m-i-1)\right) \log (2)-\log \frac{1}{\eta}\right\} .
\end{aligned}
$$

We have the following application of these results.
Lemma 8. Let $X_{\delta}^{m}$ be Polya distributed with parameters $\alpha_{i}$. If $\alpha_{i}=i^{b}$ for $b>0$,

$$
P\left(X \in \Delta_{k}(y, \eta)\right) \geq C \exp \left\{-c k(\log k)^{b} \log \frac{1}{\eta}\right\}
$$

for some constants c and C. By a straightforward calculation one can see that this result is also valid for $b=0$. In the Dirichlet case $\alpha_{i}=\frac{1}{2} \alpha_{i-1}$ for $i=$ $1, \ldots, m$,

$$
P\left(X \in \Delta_{k}(y, \eta)\right) \geq C \exp \left\{-c k \log \frac{1}{\eta}\right\}
$$

in accordance with the result in [11].

## 6. Conclusion

We obtained posteriors that adapt to the smoothness of the underlying density, that is assumed to be contained in a nonparametric model. It is of interest to obtain, using the same prior, a parametric rate if the underlying density is a finite mixture itself. This is the case in the location-scale-model studied in [19], and the arguments used therein could be easily applied in the present work. The result would however have less practical relevance, as the variances $\sigma_{j}^{2}$ of all components are required to be the same.

Furthermore, the prior on the $\sigma_{j}$ 's used in [19] depends on $n$, and this seems to be essential if the optimal rates and adaptivity found in the present work are to be maintained. In the lower bound for the prior mass on a $K L$-ball around $f_{0}$, given by (33), we get an extra factor $k_{n}$ in the exponent, and the argument only applies if $\lambda=\lambda_{n} \approx \sigma_{n}$. This suggests that the restriction to have the same variance for all components is necessary to have a rate-adaptive posterior based on a fixed prior, but we have not proved this. The determination of lower bounds for convergence rates deserves further investigation; some results can be found in [33]. Full adaptivity over the union of all finite mixtures and Hölder densities could perhaps be established by putting a hyperprior on the two models, as considered in [12].

## 7. Acknowledgements

We want to thank Catia Scricciolo, Bertrand Michel and Cathy Maugis for carefully reading earlier versions of this work, enabling to significantly improve our paper.

## Appendix A

The following theorem is taken from [13] (Theorem 5), and slightly adapted to facilitate the entropy calculations in the proof of Theorem 2. Their condition $\Pi\left(\mathcal{F}_{n} \mid X_{1}, \ldots, X_{n}\right) \rightarrow 0$ in $F_{0}^{n}$-probability is a consequence of (41) and (42) below. This follows from a simplification of the proof of Theorem 2.1 in [11], p. 525 , where we replace the complement of a Hellinger-ball around $f_{0}$ by $\mathcal{F}_{n}^{c}$. If we then take $\epsilon=2 \bar{\epsilon}_{n}$ in Corollary 1 in [13], with $\bar{\epsilon}_{n} \geq \widetilde{\epsilon}_{n}$ and $\bar{\epsilon}_{n} \rightarrow 0$, the result of Theorem 5 in this paper still holds.
Theorem 3 (Ghosal and van der Vaart, 2006). Given a statistical model $\mathcal{F}$, let $\left\{X_{i}\right\}_{i \geq 1}$ be an i.i.d. sequence with density $f_{0} \in \mathcal{F}$. Assume that there exists a sequence of submodels $\mathcal{F}_{n}$ that can be partitioned as $\bigcup_{j=-\infty}^{\infty} \mathcal{F}_{n, j}$ such that, for
sequences $\widetilde{\epsilon}_{n}$ and $\bar{\epsilon}_{n} \geq \widetilde{\epsilon}_{n}$ with $\bar{\epsilon}_{n} \rightarrow 0$ and $n \widetilde{\epsilon}_{n}^{2} \rightarrow \infty$,

$$
\begin{align*}
\sum_{j=-\infty}^{\infty} \sqrt{N\left(\bar{\epsilon}_{n}, \mathcal{F}_{n, j}, d\right)} \sqrt{\Pi_{n}\left(\mathcal{F}_{n, j}\right)} e^{-n \bar{\epsilon}_{n}^{2}} & \rightarrow 0  \tag{40}\\
\Pi_{n}\left(K L\left(f_{0}, \widetilde{\epsilon}_{n}\right)\right) & \geq e^{-n \widetilde{\epsilon}_{n}^{2}}  \tag{41}\\
\Pi_{n}\left(\mathcal{F}_{n}^{c}\right) \leq e^{-4 n \widetilde{\epsilon}_{n}^{2}} & \tag{42}
\end{align*}
$$

where $K L\left(f_{0}, \widetilde{\epsilon}_{n}\right)$ is the Kullback-Leibler ball

$$
\left\{f: F_{0} \log \left(f_{0} / f\right) \leq \widetilde{\epsilon}_{n}^{2}, F_{0} \log ^{2}\left(f_{0} / f\right) \leq \widetilde{\epsilon}_{n}^{2}\right\}
$$

Then $\Pi_{n}\left(f \in \mathcal{F}: d\left(f, f_{0}\right)>8 \bar{\epsilon}_{n} \mid X_{1}, \ldots, X_{n}\right) \rightarrow 0$ in $F_{0}^{n}$-probability.
The advantage of the above version is that (42) is easier to verify for a faster sequence $\widetilde{\epsilon}_{n}$. The use of the same sequence $\epsilon_{n}$ in (40) and (42) would otherwise pose restrictions for the choice of $\mathcal{F}_{n}$.

The following asymptotic formula for the Gamma function can be found in many references, see for example Abramowitz and Stegun [1].
Lemma 9. For any $\alpha>0$,

$$
\begin{equation*}
\Gamma(\alpha)=\sqrt{2 \pi} e^{-\alpha} \alpha^{\alpha-\frac{1}{2}} e^{\theta(\alpha)} \tag{43}
\end{equation*}
$$

where $0<\theta(\alpha)<\frac{1}{12 \alpha}$. If $\alpha \rightarrow \infty$, this gives the bound $\frac{\Gamma(2 \alpha)}{\Gamma(\alpha) \Gamma(\alpha)} \gtrsim \alpha^{-\frac{1}{2}} 2^{2 \alpha-\frac{1}{2}}$ for the beta function. For $\alpha \rightarrow 0$, the identity $\alpha \Gamma(\alpha)=\Gamma(\alpha+1)$ gives the bounds $\Gamma(\alpha) \leq \frac{1}{\alpha}$ and $\Gamma(\alpha) \geq \frac{c}{\alpha}$, where $c=0.8856 \ldots$ is the local minimum of the gamma function on the positive real line. Consequently, $\frac{\Gamma(2 \alpha)}{\Gamma(\alpha) \Gamma(\alpha)} \gtrsim \alpha$. From (43) it follows that for all $\alpha>0$ and all integers $j \geq 1$,

$$
\begin{align*}
\frac{\sqrt{\Gamma\left(\frac{2 j+1}{1+\alpha}\right)}}{j!} & \leq \frac{1}{\sqrt{2 \pi}} e^{\frac{\alpha}{1+\alpha}(j+1)}\left(\frac{2}{1+\alpha}\right)^{\frac{j}{1+\alpha}}(j+1)^{-\frac{\alpha j}{1+\alpha}}  \tag{44}\\
\frac{\Gamma\left(\frac{j+1}{1+\alpha}\right)}{j!} & \leq e^{\frac{\alpha}{1+\alpha}(j+1)+\frac{1}{12}}\left(\frac{1}{1+\alpha}\right)^{\frac{j}{1+\alpha}}(j+1)^{-\frac{\alpha j}{1+\alpha}} \tag{45}
\end{align*}
$$

The following lemma will be required for the proof of Lemma 1 in the next section.
Lemma 10. Given a positive integer $m$ and $\psi_{(p)}(x)=C_{p} e^{-|x|^{p}}$, let $\varphi$ be the $m$ fold convolution $\psi * \cdots * \psi$. Then for any $\alpha \geq 0$ there is a number $k^{\prime}=k^{\prime}(p, \alpha, m)$ such that for all sufficiently small $\sigma>0$,

$$
\begin{equation*}
\int_{|x|>k^{\prime}|\log \sigma|^{1 / p}} \varphi(x)|x|^{\alpha} d x=\sigma^{H} \tag{46}
\end{equation*}
$$

Proof. For any $p>0$ and a random variable $Z$ with density $\psi_{(p)}$,

$$
P(Z>y)=\int_{y}^{\infty} \psi_{(p)}(x) d x \leq p^{-1} y^{1-p} \int_{y}^{\infty} p x^{p-1} \psi_{(p)}(x) d x=p^{-1} y^{1-p} \psi_{(p)}(y)
$$

For $m=1$, we have

$$
\begin{aligned}
\int_{y}^{\infty} x^{\alpha} \psi_{(p)}(x) d x & =\int_{y^{1+\alpha}}^{\infty} \psi_{(p)}\left(z^{1 /(1+\alpha)}\right) d x=\frac{C_{p}}{C_{p /(1+\alpha)}} \int_{y^{1+\alpha}}^{\infty} \psi_{(p /(1+\alpha))}(z) d z \\
& =\frac{C_{p}}{C_{p /(1+\alpha)}} P_{Z \sim \psi_{(p / /(1+\alpha))}}\left(Z>k^{\left.k^{(1+\alpha)}|\log \sigma|^{\frac{1+\alpha}{p}}\right)}\right.
\end{aligned}
$$

for any $\alpha>0$ and $y>0$.
Now let $m>1$, and $X=\sum_{i=1}^{m} Z_{i}$ for i.i.d. random variables $Z_{i}$ with density $\psi_{(p)}$. If $\alpha \geq 1$ then, by Jensen's inequality applied to the function $x \mapsto x^{\alpha}$,

$$
\begin{aligned}
& E\left(|Z|^{\alpha} 1_{|Z|>k^{\prime}|\log \sigma|^{1 / p}}\right) \leq E\left(m^{\alpha-1}\left(\sum_{i=1}^{m}\left|Z_{i}\right|^{\alpha}\right) 1_{|Z|>k^{\prime}|\log \sigma|^{1 / p}}\right) \\
& \quad \leq m^{\alpha-1} \sum_{i=1}^{m} E\left(\left|Z_{i}\right|^{\alpha} \sum_{j=1}^{m} 1_{\left|Z_{j}\right|>\frac{k^{\prime}}{m}|\log \sigma|^{1 / p}}\right)=\sigma^{H},
\end{aligned}
$$

where we used (46) with $\alpha=0$ and the independence of the $Z_{i}$ 's to bound the terms with $i \neq j$. If $\alpha<1$, we bound $|Z|^{\alpha}$ by $|Z|$ and apply the preceding result.

## Appendix B: Approximation under a global Hölder condition

For $L>0, \beta>0$ and $r$ the largest integer smaller than $\beta$, let $\mathcal{H}(\beta, L)$ be the space of functions $h$ such that $\sup _{x \neq y}\left|h^{(r)}(x)-h^{(r)}(y)\right| /|y-x|^{\beta-r} \leq L$, where $h^{(r)}$ is the $r$ th derivative of $h$. Let $H_{\beta}$ be the Hölder-space $\cup_{L>0} \mathcal{H}(\beta, L)$, and given some function $h \in H_{\beta}$, let $L_{h, \beta-r}=\sup _{x \neq y}\left|h^{(r)}(x)-h^{(r)}(y)\right| /|y-x|^{\beta-r}$. When $\beta-r=1$, this equals $\left\|h^{(r+1)}\right\|_{\infty}$.

Lemma 11. Let $f \in H_{\beta}$, where $2 k<\beta \leq 2 k+2$ for some nonnegative integer $k$. Then $\left\|f-f_{k} * \psi_{\sigma}\right\|_{\infty}=O\left(\sigma^{\beta}\right)$, where $f_{k}$ is defined recursively by $f_{0}=f$, $f_{1}=f-\Delta_{\sigma} f=2 f-K_{\sigma} f$ and $f_{j+1}=f-\Delta_{\sigma} f_{j}, j \geq 1$.

Proof. By induction it follows that

$$
\begin{equation*}
f_{k}=\sum_{i=0}^{k}(-1)^{i}\binom{k+1}{i+1} K_{\sigma}^{i} f, \quad \Delta_{\sigma}^{k} f=\sum_{i=0}^{k}(-1)^{k-i}\binom{k}{i} K_{\sigma}^{i} f . \tag{47}
\end{equation*}
$$

The proof then depends on the following two observations. First, note that if $f \in H_{\beta}$ then $f_{1}, f_{2}, \ldots$ are also in $H_{\beta}$, even if $\psi$ itself is not in $H_{\beta}$ (e.g.
when $\psi$ is the Laplace kernel). Second, it follows from the symmetry of $\psi$ that $K_{\sigma} f^{(r)}=\frac{d^{r}}{d x^{r}} K_{\sigma} f$, i.e. the $r$ th derivative of the convolution of $f$ equals the convolution of $f^{(r)}$.

When $k=0$ and $\beta \leq 2$ the result is elementary. When $k=1$ we have $K_{\sigma}\left(f_{1}\right)-$ $f=\Delta_{\sigma}\left(f-\Delta_{\sigma}(f)\right)-\Delta_{\sigma}(f)=-\Delta_{\sigma} \Delta_{\sigma} f$, and $\left\|\Delta_{\sigma} \Delta_{\sigma} f\right\|_{\infty} \leq \nu_{2} \sigma^{2}\left\|\left(\Delta_{\sigma} f\right)^{\prime \prime}\right\|_{\infty}$. Because differentiation and the $\Delta_{\sigma}$ operator can be interchanged, we also have $\left\|\left(\Delta_{\sigma} f\right)^{\prime \prime}\right\|_{\infty}=\left\|\left(\Delta_{\sigma} f^{\prime \prime}\right)\right\|_{\infty}$. Since $f^{\prime \prime} \in H_{\beta-2}$, the latter quantity is $O\left(\sigma^{\beta-2}\right)$. Consequently, $\left\|\Delta_{\sigma} \Delta_{\sigma} f\right\|_{\infty}=O\left(\sigma^{\beta}\right)$. For $k>1$, we repeat this step and use that, as a consequence of (47), $\left\|K_{\sigma} f_{k}-f\right\|_{\infty}=\left\|\Delta_{\sigma}^{k+1} f\right\|_{\infty}$. From the following induction argument it follows that for any positive integer $k, \beta \in(2 k, 2 k+$ 2] and $f \in H_{\beta},\left\|\Delta_{\sigma}^{k+1} f\right\|_{\infty}=O\left(\sigma^{\beta}\right)$. Suppose this statement holds for $k=$ $0,1, \ldots, m-1$, and that $f \in H_{\beta}$ with $\beta \in(2 m, 2 m+2]$. Then $\left\|\Delta_{\sigma}^{m} f\right\|_{\infty}=$ $O\left(\left\|\Delta_{\sigma} f^{(2 m)}\right\|_{\infty} \sigma^{2 m}\right)$ and $\left\|\Delta_{\sigma} f^{(2 m)}\right\|_{\infty}=O\left(\sigma^{\beta-2 m}\right)$ as $f^{(2 m)} \in H_{\beta-2 m}$.

## Appendix C: Proof of Lemma 1

The smoothness condition (6) in (C1) implies that

$$
\begin{align*}
& \log f_{0}(y) \leq \log f_{0}(x)+\sum_{j=1}^{r} \frac{l_{j}(x)}{j!}(y-x)^{j}+L(x)|y-x|^{\beta}  \tag{48}\\
& \log f_{0}(y) \geq \log f_{0}(x)+\sum_{j=1}^{r} \frac{l_{j}(x)}{j!}(y-x)^{j}-L(x)|y-x|^{\beta} \tag{49}
\end{align*}
$$

again for all $x, y$ with $|y-x| \leq \gamma$.
Let $f$ be a function for which these conditions hold, $r$ being the largest integer smaller than $\beta$. We define

$$
B_{f, r}(x, y)=\sum_{j=1}^{r} \frac{l_{j}(x)}{j!}(y-x)^{j}+L(x)|y-x|^{\beta}
$$

First we assume that $\beta \in(1,2]$ and $r=1$. The case $\beta \in(0,1]$ is easier and can be handled similarly; the case $\beta>2$ is treated below. Using (48) we demonstrate below that

$$
\begin{equation*}
K_{\sigma} f(x) \leq\left(1+O\left(\left(|L(x)|+\left|l_{1}^{\beta}(x)\right|\right) \sigma^{\beta}\right)\right) f(x)+O\left(1+|L(x)|+\left|l_{1}^{\beta}(x)\right|\right) \sigma^{H} \tag{50}
\end{equation*}
$$

We omit the proof of the inequality in the other direction, which can be obtained similarly using (49). To prove (50), we define, for any $x \in \mathbb{R}$,

$$
D_{x}=\left\{y:|y-x| \leq k^{\prime} \sigma|\log \sigma|^{1 / p}\right\}
$$

for a large enough constant $k^{\prime}$ to be chosen below. Assuming that $k^{\prime} \sigma|\log \sigma|^{1 / p} \leq$ $\gamma$, for $\gamma$ as in condition (C1), we can rewrite (48) as $f(y) \leq f(x) \exp \left\{B_{f, 1}(x, y)\right\}$, and

$$
\begin{equation*}
K_{\sigma} f(x) \leq f(x) \int_{D_{x}} e^{B_{f, r}(x, y)} \psi_{\sigma}(y-x) d y+\int_{D_{x}^{c}} f(y) \psi_{\sigma}(y-x) d y \tag{51}
\end{equation*}
$$

Furthermore, if $x \in A_{\sigma}$ and $y \in D_{x}$, then for $M=\frac{1}{(r+1)!} \exp \left\{\sup _{x \in A_{\sigma}, y \in D_{x}}\left|B_{f, r}(x, y)\right|\right\}$ and some $\xi \in(0, B)$,

$$
\begin{equation*}
e^{B_{f, r}(x, y)}=\sum_{m=0}^{r} \frac{1}{m!} B_{f, r}^{m}(x, y)+\frac{e^{\xi}}{(r+1)!} B_{f, r}^{r+1}(x, y) \leq \sum_{m=0}^{r} \frac{1}{m!} B_{f, r}^{m}(x, y)+M\left|B_{f, r}\right|^{r+1}(x, y) . \tag{52}
\end{equation*}
$$

In the present case, $\beta \in(1,2]$ and $r=1$, hence

$$
\begin{align*}
e^{B_{f, r}(x, y)} \leq 1 & +B_{f, r}(x, y)+M B_{f, r}^{2}(x, y)=1+l_{1}(x)(y-x)+L(x)|y-x|^{\beta} \\
& +M\left(l_{1}^{2}(x)(y-x)^{2}+2 l_{1}(x) L(x)(y-x)|y-x|^{\beta}+L^{2}(x)|y-x|^{2 \beta}\right) . \tag{53}
\end{align*}
$$

Integrating over $D_{x}$, the terms with a factor $(y-x)$ disappear, so that the first term on the right in (51) is bounded by
$f(x) \int_{D_{x}} \psi_{\sigma}(y-x)\left\{1+L(x)|y-x|^{\beta}+M\left(k^{\prime} B\right)^{2-\beta}\left|l_{1}(x)(y-x)\right|^{\beta}+M k^{\prime \beta} B|L(x)(y-x)|^{\beta}\right\} d y$,
since $\left|l_{1}(x)(y-x)\right| \leq k^{\prime} B$ and $|L(x)||(y-x)|^{\beta} \leq k^{\prime \beta} B$ when $x \in A_{\sigma}$ and $y \in D_{x}$. Because $\int_{D_{x}^{c}} \psi_{\sigma}(y-x)|y-x|^{\alpha} d y=\sigma^{H}$ for any $\alpha \geq 0$, when $k^{\prime}$ in the definition of $D_{x}$ is sufficiently large (see Lemma 10 in Appendix A), (51), (53) and (54) imply that for constants $k_{1}=M\left(k^{\prime} B\right)^{2-\beta}$ and $k_{2}=1+M k^{\prime \beta} B$,

$$
\begin{gather*}
\left(K_{\sigma} f\right)(x) \leq f(x) \int_{\mathbb{R}} \psi_{\sigma}(y-x)\left\{1+k_{1}\left|l_{1}(x)\right|^{\beta}|y-x|^{\beta}+k_{2}|L(x)||y-x|^{\beta}\right\} d y \\
+\left(\|f\|_{\infty}+1+k_{1}\left|l_{1}(x)\right|^{\beta}+k_{2}|L(x)|\right) O\left(\sigma^{H}\right) \tag{55}
\end{gather*}
$$

which completes the proof of (50) for $\beta \in(1,2]$. Using the same arguments the inequality in the other direction (with different constants) can be obtained when we define $B_{f, 1}(x, y)=l_{1}(x)(y-x)-L(x)|y-x|^{\beta}$, and use that $e^{B_{f, r}(x, y)} \geq$ $\sum_{m=0}^{r} \frac{1}{m!} B_{f, r}^{m}(x, y)-M\left|B_{f, r}\right|^{r+1}(x, y)$ instead of (52). This finishes the proof of (17) for $k=0$.

Now let $f$ be a function for which (48) and (49) hold with $\beta \in(3,4]$ and $r=3$; the case $\beta \in(2,3]$ being similar and simpler. Before looking at $K_{\sigma} f_{1}$ we first give an expression for $K_{\sigma} f$. When $x \in A_{\sigma}$ and $y \in D_{x}, e^{B} \leq 1+B+\frac{1}{2} B^{2}+\frac{1}{6} B^{3}+M B^{4}$ and for some constant $M$, with $B(x, y)=l_{1}(x)(y-x)+\frac{1}{2} l_{1}(x)(y-x)+\frac{1}{6} l_{3}(x)(y-$ $x)^{3}+L(x)|y-x|^{\beta}$. Using this bound on $e^{B}$ we can redo the calculations given in (51), (52), (54) and (55); again by showing inequality in both directions we find that

$$
\begin{equation*}
K_{\sigma} f(x)=f(x)\left(1+\frac{\nu_{2}}{2}\left(l_{1}^{2}(x)+l_{2}(x)\right) \sigma^{2}+O\left(R(x) \sigma^{\beta}\right)\right)+O\left((1+R(x)) \sigma^{H}\right) . \tag{56}
\end{equation*}
$$

This follows from the fact that for $x \in A_{\sigma}$ and $y \in D_{x}$ we can control the terms containing a factor $|y-x|^{k}$ with $k>2$, similar to (54). All these terms can be
shown to be a multiple of $\sigma^{\beta}$ by taking out a factor $|y-x|^{\beta}$ and matching the remaining factor $|y-x|^{k-\beta}$ by a certain power of the $\left|l_{j}\right|$ 's or $|L|$.

The proof of (17) for $f_{1}$ can now be completed by the observation that (56) depends on the kernel $\psi$ only through the values of $\nu_{\alpha}$. In fact it holds for any symmetric kernel such that $\int \psi(x)|x|^{\alpha} d x=\nu_{\alpha}<\infty$ and $\int_{|x|>k^{\prime}|\log \sigma|^{1 / p}} \psi(x)|x|^{\alpha} d x=$ $\sigma^{H}$ when $k^{\prime}$ is large enough. For the kernel $\psi * \psi$ these properties follow from Lemma 10 in Appendix A. Consequently, (56) still holds when $K_{\sigma} f$ is replaced by $K_{\sigma} K_{\sigma} f$ and $\nu_{2}$ by $\nu_{\psi * \psi, 2}=\int(\psi * \psi)(x)|x|^{\alpha} d x$. As $f_{1}=2 f-K_{\sigma} f$ and $\nu_{\psi * \psi, 2}=2 \nu_{2}$, this proves (17) for $k=1$.

The same arguments can be used when $k>1$ and $\beta \in(2 k, 2 k+2]$; in that case all terms with $\sigma^{2}, \sigma^{4}, \ldots, \sigma^{2 k}$ cancel out. This can be shown by expressing the moments $\nu_{m, 2}, \ldots, \nu_{m, 2 k}$ of the kernels $K_{\sigma}^{m}, m=2, \ldots, k+1$ in terms of $\nu_{2}, \ldots, \nu_{2 k}$ and combining this with (47) in the proof of Lemma 11 in Appendix B.

## Appendix D: Proof of Lemma 2

To show that the first integral in (23) is of order $\sigma^{2 \beta}$, consider the sets
$A_{\sigma, \delta}=\left\{x:\left|l_{j}(x)\right| \leq \delta B \sigma^{-j}|\log \sigma|^{-j / p}, j=1, \ldots, r,|L(x)| \leq \delta B \sigma^{-\beta}|\log \sigma|^{-\beta / p}\right\}$,
indexed by $\delta \leq 1$. For notational convenience, let $\sum_{j=1}^{\beta}$ denote sums over $(r+1)$ terms containing respectively the functions $l_{1}, \ldots, l_{r}$ and $l_{\beta}=L$. First let $m=0$. It follows from (7) in (C2) and Markov's inequality that
$\int_{A_{\sigma}^{c}}\left(K_{\sigma}^{0} f\right)(x) d x \leq \sum_{j=1}^{\beta} P\left(\left|l_{j}(X)\right|^{\frac{2 \beta+\epsilon}{j}} \geq(\delta B)^{\frac{2 \beta+\epsilon}{j}} \sigma^{-2 \beta-\epsilon}|\log \sigma|^{-\frac{2 \beta+\epsilon}{p}}\right)=O\left(\sigma^{2 \beta}\right)$, provided that $\sigma^{-\epsilon}|\log \sigma|^{-\frac{2 \beta+\epsilon}{p}}>1$, which is the case if $\sigma$ is sufficiently small.

If $m=1$, consider independent random variables $X$ and $U$ with densities $f$ and $\psi$, respectively. Then $X+\sigma U$ has density $K_{\sigma} f$. Because $P(|U| \geq$ $\left.k^{\prime}|\log \sigma|^{1 / p}\right)=O\left(\sigma^{2 \beta}\right)$ if the constant $k^{\prime}$ is sufficiently large, we have

$$
\begin{align*}
& P\left(X+\sigma U \in A_{\sigma}^{c}\right) \leq P\left(X+\sigma U \in A_{\sigma}^{c},|U| \leq k^{\prime}|\log \sigma|^{1 / p}\right)+P\left(|U| \geq k^{\prime}|\log \sigma|^{1 / p}\right) \\
& \quad=O\left(\sigma^{2 \beta}\right)+P\left(X+\sigma U \in A_{\sigma}^{c}, X \in A_{\sigma, \delta},|U| \leq k^{\prime}|\log \sigma|^{1 / p}\right) \\
& \quad+P\left(X+\sigma U \in A_{\sigma}^{c}, X \in A_{\sigma, \delta}^{c},|U| \leq k^{\prime}|\log \sigma|^{1 / p}\right) \tag{57}
\end{align*}
$$

The last term is bounded by $P\left(X \in A_{\sigma, \delta}^{c}\right)$, which is $O\left(\sigma^{2 \beta}\right)$ for any $0<\delta \leq 1$. We show that the last term on the second line is zero for sufficiently small $\delta$. This can be shown by contradiction: together with the conditions on $f$, the fact that $X \in A_{\sigma, \delta}$ and $X+\sigma U \in A_{\sigma, 1}^{c}$ implies that $|U|$ is large, contradicting $|U| \leq k^{\prime}|\log \sigma|^{1 / p}$.

To see this, note that since $X \in A_{\sigma, \delta},|L(X)| \leq \delta B \sigma^{-\beta}|\log \sigma|^{-\beta / p}$ and $\left|l_{j}(X)\right| \leq \delta B \sigma^{-j}|\log \sigma|^{-j / p}$ for $j=1, \ldots, r$. On the other hand, $X+\sigma U \in$
$A_{\sigma, 1}^{c}$ implies that $|L(X+\sigma U)| \geq B \sigma^{-\beta}|\log \sigma|^{-\beta / p}$ or that $\left|l_{i}(X+\sigma U)\right| \geq$ $\delta B \sigma^{-i}|\log \sigma|^{-i / p}$ for some $i \in\{1, \ldots, r\}$. From (6) it follows that for all $i=$ $1, \ldots, r$
$\left|l_{i}(X+\sigma U)\right| \leq\left.\left.\left|\sum_{j=0}^{r-i} \frac{l_{i+j}(X)}{j!}(\sigma U)^{j}+\frac{r!}{(r-i)!}\right| L(X)| | \sigma U\right|^{\beta-i}\left|\leq B \sigma^{-i}\right| \log \sigma\right|^{-i / p}$
if $\delta$ is sufficiently small. Therefore it has to be a large value of $|L(X+\sigma U)|$ that forces $X+\sigma U$ to be in $A_{\sigma}^{c}$. Hence it suffices to show that $|L(X)| \leq$ $\delta B \sigma^{-\beta}|\log \sigma|^{-\beta / p}$ and $|U| \leq k^{\prime}|\log \sigma|^{1 / p}$ is in contradiction with $|L(X+\sigma U)| \geq$ $B \sigma^{-\beta}|\log \sigma|^{-\beta / p}$. We now derive the contradiction from the assumption that $L$ is polynomial. Let $q$ be its degree, and let $\eta=\max \left|z_{i}\right|, z_{i}$ being the roots of $L$. First, suppose that $|X|>\eta+1$. Then

$$
U^{j} \sigma^{j} L^{(j)}(X)=O\left(\left|U^{j} \sigma^{j} L(X)\right|\right)=O\left(\sigma^{-(\beta-j)}|\log \sigma|^{-\frac{\beta-j}{p}}\right), \quad j=1, \ldots, q
$$

This implies

$$
\begin{aligned}
|L(X+\sigma U)| & \leq|L(X)|+\left|\sum_{j=1}^{q} \frac{\sigma^{j} U^{j} L^{(j)}(X)}{j!}\right|+\frac{\sigma^{q}|U|^{q}}{q!}\left|L^{(q)}(\xi)-L^{(q)}(X)\right| \\
& \leq \delta B \sigma^{-\beta}|\log \sigma|^{-\frac{\beta}{p}}+O\left(\sigma^{-(\beta-1)}|\log \sigma|^{-\frac{\beta-1}{p}}\right)
\end{aligned}
$$

which is smaller than $B \sigma^{-\beta}|\log \sigma|^{-\frac{\beta}{p}}$ when $\sigma$ and $\delta<1$ are small enough. If $|X| \leq \eta+1$, note that this implies $|X+\sigma U| \leq \eta+2$ for sufficiently small $\sigma$, as $|U| \leq k^{\prime}|\log \sigma|^{\frac{\beta}{p}}$. Consequently,

$$
|L(X+\sigma U)| \leq \max _{|x| \leq \eta+2}|L(x)|=\bar{L} \leq B \sigma^{-\beta}|\log \sigma|^{-\frac{\beta}{p}},
$$

again for sufficiently small $\sigma$.
If $m=2$ in (23), note that the above argument remains valid if $X$ has density $K_{\sigma} f$ instead of $f$. The last term in (57) is then bounded by $P\left(X \in A_{\sigma, \delta}^{c}\right)$, which is $O\left(\sigma^{2 \beta}\right)$ by the result for $m=1$. This step can be repeated arbitrarily often, for some decreasing sequence of $\delta$ 's.

To bound the second integral in (23) for $m=0$, we need the tail condition $f(x) \leq c|x|^{-\alpha}$ in (C2). In combination with the monotonicity of $f$ required in (C4), this implies that

$$
\begin{equation*}
\int_{E_{\sigma}^{c}} f(x) d x \leq \sigma^{H_{1} / 2} \int_{E_{\sigma}^{c}} \sqrt{f(x)} d x=O\left(\sigma^{2 \beta}\right) \tag{58}
\end{equation*}
$$

which is $O\left(\sigma^{2 \beta}\right)$ when $H_{1} \geq 4 \beta$.
For $m=1$, we integrate over the sets $E_{\sigma}^{c} \cap A_{\sigma}^{c}$ and $E_{\sigma}^{c} \cap A_{\sigma}$. The integral over the first set is $O\left(\sigma^{2 \beta}\right)$ by the preceding paragraph. To bound the second integral, consider the sets

$$
\begin{equation*}
E_{\sigma, \delta}=\left\{x: \log f(x) \geq \delta H_{1} \log \sigma\right\} \tag{59}
\end{equation*}
$$

indexed by $\delta \leq 1$. We can use the inequality (57) with $A_{\sigma}^{c}, A_{\sigma, \delta}$ and $A_{\sigma, \delta}^{c}$ replaced by respectively $E_{\sigma}^{c} \cap A_{\sigma}, E_{\sigma, \delta} \cap A_{\sigma}$ and $E_{\sigma, \delta}^{c} \cap A_{\sigma}$. The probability ${ }_{P} P_{X \sim f}(X \in$ $\left.E_{\sigma, \delta}^{c}\right)$ can be shown to be $O\left(\sigma^{2 \beta}\right)$ as in (58), provided that $\delta H_{1} / 2 \geq 2 \beta$. The probability that $|U| \leq k^{\prime}|\log \sigma|^{1 / p}, X+\sigma U \in E_{\sigma}^{c} \cap A_{\sigma}$ and $X \in E_{\sigma, \delta} \cap A_{\sigma}$ is zero: due to the construction of $A_{\sigma}$ we have $|l(X+\sigma U)-l(X)|=O(1)$, whereas $|l(X+\sigma U)-l(X)| \geq(1-\delta) H_{1}|\log \sigma|$. This step can be repeated as long as the terms $P_{X \sim f}\left(X \in E_{\sigma, \delta}^{c}\right)$ remain $O\left(\sigma^{2 \beta}\right)$, which is the case if the initial $H_{1}$ is chosen large enough. This finishes the proof of (23).

To prove (25), let $\beta>2$ and $k \geq 1$ be such that $2 k<\beta \leq 2 k+2, l=\log f$ being $\beta$-Hölder. It can be seen that Lemma 1 still holds if we treat $l$ as if it was Hölder smooth of degree 2. Instead of (17), we then obtain

$$
\begin{equation*}
\left(K_{\sigma} f\right)(x)=f(x)\left(1+O\left(R^{(2)}(x) \sigma^{2}\right)\right)+O\left(\left(1+R^{(2)}(x)\right) \sigma^{H}\right) \tag{60}
\end{equation*}
$$

where $L^{(2)}=l_{2}$ and $R^{(2)}$ is a linear combination of $l_{1}^{2}$ and $\left|L^{(2)}\right|$. The key observation is that $R^{(2)}=o(1)$ uniformly on $A_{\sigma}$ when $\sigma \rightarrow 0$. Combining (60) with the lower bound for $f$ on $E_{\sigma}$, can find a constant $\rho$ close to 1 such that
$f_{1}(x)=2 f(x)-K_{\sigma} f(x)=2 f(x)-\left(1+O\left(R^{(2)}(x)\right) \sigma^{2}\right) f(x)-O\left(1+R^{(2)}(x)\right) \sigma^{H}>\rho f(x)$
for small enough $\sigma$. Similarly, when $l$ is treated as being Hölder smooth of degree 4, we find that
$f_{2}(x)=2 f_{1}(x)-K_{\sigma} f_{1}(x)=2 f_{1}(x)-\left(1+O\left(R^{(4)}(x)\right) \sigma^{4}\right) f(x)-O\left(1+R^{(4)}(x)\right) \sigma^{H}>\rho^{2} f(x)$.
Continuing in this manner, we find a constant $\rho_{k}$ such that $f_{k}(x)>\rho_{k} f(x)$ for $x \in A_{\sigma} \cap E_{\sigma}$ and $\sigma$ sufficiently small. If initially $\rho$ is chosen close enough to 1 , $\rho^{k}>\frac{1}{2}$ and hence $A_{\sigma} \cap E_{\sigma} \subset J_{\sigma, k}$. To see that (23) now implies (24), note that the integrand $\frac{1}{2} f-f_{k}$ is a linear combination of $K_{\sigma}^{m} f, m=0, \ldots, k$.

## Appendix E: Proof of Lemma 4

We bound the second integral in (28); the first integral can be bounded similarly. For $\tilde{h}_{k}$ the normalized restriction of $h_{k}$ to $E_{\sigma}$ and $m$ the finite mixture to be constructed, we write

$$
\begin{align*}
\int f\left(\log \frac{f}{m}\right)^{2}= & \int_{E_{\sigma}} f\left(\log \frac{f}{K_{\sigma} h_{k}}+f \log \frac{K_{\sigma} h_{k}}{K_{\sigma} \tilde{h}_{k}}+f \log \frac{K_{\sigma} \tilde{h}_{k}}{m}\right)^{2}  \tag{61}\\
& +\int_{E_{\sigma}^{c}} f\left(\log \frac{f}{K_{\sigma} h_{k}}+\log \frac{K_{\sigma} h_{k}}{m}\right)^{2}
\end{align*}
$$

The integral of $f\left(\log \left(f / K_{\sigma} h_{k}\right)\right)^{2}$ over $E_{\sigma}$ is $O\left(\sigma^{2 \beta}\right)$ by Theorem 1. To show that the integral of $f\left(\log \left(K_{\sigma} h_{k} / K_{\sigma} \tilde{h}_{k}\right)\right)^{2}$ over $E_{\sigma}$ is $O\left(\sigma^{2 \beta}\right)$ as well, recall the definition of $g_{k}$ and $h_{k}$ in (20) and (21). Combining (23) and (24) in Lemma 2 with the fact that $f_{k}$ is a linear combination of $K_{\sigma}^{i} f, i=0, \ldots, k$ (see (47)
in appendix B), we find that $\int_{E_{\sigma}^{c}} h_{k}=O\left(\sigma^{2 \beta}\right)$. Consequently, $\left.K_{\sigma} h_{k} / K_{\sigma} \tilde{h}_{k}\right)=$ $O\left(\sigma^{2 \beta}\right)$ and the required bound for $\int_{E_{\sigma}} f\left(\log \left(K_{\sigma} h_{k} / K_{\sigma} \tilde{h}_{k}\right)\right)^{2}$ follows. To bound the integral of $f\left(\log K_{\sigma} \tilde{h}_{k} / m\right)^{2}$ over $E_{\sigma}$, let $m=m\left(\cdot ; k_{\sigma}, \mu_{\sigma}, w_{\sigma}, \sigma\right)$ be the finite mixture obtained from Lemmas 12 and 13 , with $\epsilon=\sigma^{\delta^{\prime} H_{1}+1}$ and $\delta^{\prime} \geq 1+2 \beta / H_{1}$. The requirement that $a \lesssim \psi^{-1}(\epsilon)$ in Lemmas 12 and 13 is satisfied by the monotonicity and tail conditions on $f$ (see (29)). The number of components $k_{\sigma}$ in Lemma 13 is $O\left(\sigma^{-1}|\log \sigma|^{1+p^{-1}}\right)$. We have

$$
\int_{E_{\sigma}} f\left(\log \frac{K_{\sigma} \tilde{h}_{k}}{m}\right)^{2} \leq \int_{E_{\sigma}} f\left(\frac{m-K_{\sigma} \tilde{h}_{k}}{\sigma^{H_{1}}-\sigma^{\delta^{\prime} H_{1}}}\right)^{2} \leq \sigma^{2\left(\delta^{\prime}-1\right) H_{1}}=O\left(\sigma^{2 \beta}\right)
$$

provided that $\delta^{\prime} \geq 1+\frac{\beta}{H_{1}}$. The cross-products resulting from the square in the integral over $E_{\sigma}$ can be shown to be $O\left(\sigma^{2 \beta}\right)$ using the Cauchy-Schwartz inequality and the preceding bounds.

To bound the integral over $E_{\sigma}^{c}$, we add a component with weight $\sigma^{2 \beta}$ and mean zero to the finite mixture $m$. From Lemma 3 it can be seen that this does not affect the preceding results. Since $f$ and $h_{k}$ are uniformly bounded, so is $K_{\sigma} h_{k}$. If $C$ is an upper bound for $K_{\sigma} h_{k}$, then

$$
\begin{align*}
\int_{E_{\sigma}^{c}} f(x)\left(\log \frac{K_{\sigma} h_{k}}{m}(x)\right)^{2} d x & \leq \int_{E_{\sigma}^{c}} f(x)\left(\log \frac{C}{\sigma^{2 \beta} \psi_{\sigma}(x)}\right)^{2} d x \\
& =\int_{E_{\sigma}^{c}} f(x)\left(\log \left(C_{p}^{-1} C\right)+2 \beta|\log \sigma|+\frac{|x|^{p}}{\sigma^{p}}\right)^{2} d x \tag{62}
\end{align*}
$$

This is $O\left(\sigma^{2 \beta}\right)$ if

$$
\int_{E_{\sigma}^{c}} f(x)|x|^{2 p} d x \leq \sigma^{H_{1} / 2} \int_{E_{\sigma}^{c}} \sqrt{f(x)}|x|^{2 p} d x=O\left(\sigma^{2 \beta+2 p}\right)
$$

which is the case if $H_{1} \geq 4(\beta+p)$. The integral of $f\left(\log f / K_{\sigma} h_{k}\right)^{2}$ over $E_{\sigma}^{c}$ is $O\left(\sigma^{2 \beta}\right)$ by Lemma 1, and the integral of $f\left(\log f / K_{\sigma} h_{k}\right)\left(\log K_{\sigma} h_{k} / m\right)$ over $E_{\sigma}^{c}$ can be bounded using Cauchy-Schwartz.

If $m^{\prime}=m\left(\cdot ; k_{\sigma}, \mu, w, \sigma^{\prime}\right)$ is a different mixture with $\sigma^{\prime} \in\left[\sigma, \sigma+\sigma^{\delta^{\prime} H_{1}+2}\right]$, $\mu \in B_{k_{\sigma}}\left(\mu_{\sigma}, \sigma^{\delta^{\prime} H_{1}+2}\right)$ and $w \in \Delta_{k_{\sigma}}\left(w_{\sigma}, \sigma^{\delta^{\prime} H_{1}+1}\right)$, the $L_{\infty}$-norm between $m$ and $m^{\prime}$ is $\sigma^{\delta^{\prime} H_{1}}$ by Lemma 3, and $\int_{E_{\sigma}} f\left(\log \frac{K_{\sigma} \tilde{h}_{k}}{m^{\prime}}\right)^{2}=O\left(\sigma^{2 \beta}\right)$. The integral over $E_{\sigma}^{c}$ can be shown to be $O\left(\sigma^{2 \beta}\right)$ as in (62), where the $\left|x-\sigma^{2 \beta}\right|^{2 p}$ that comes in the place of $|x|^{2 p}$ can be handled with Jensen's inequality.

## Appendix F: Discretization

The following lemmas can be found in [19], p.59-60. They are straightforward extensions of the corresponding results for normal mixtures, contained in lemma 3.1 of [14] and lemma 2 of [13]. Lemma 13 is used in the proof of Lemma 4 in the present work.

Lemma 12. Given $p=2,4, \ldots$, let $\psi(x)=C_{p} e^{-|x|^{p}}$. Let $F$ be a probability measure on $[-a, a]$, where $a \lesssim \psi^{-1}(\epsilon)$, and assume that $\sigma \in\left[\underline{\sigma}_{n}, \bar{\sigma}_{n}\right]$ and $\epsilon<$ $\left(1 \wedge C_{p}\right)$. Then there exists a discrete distribution $F^{\prime}$ on $[-a, a]$ with at most $N=p e^{2} \log \frac{C_{p}}{\epsilon}$ support points such that $\left\|F * \psi_{\sigma}-F^{\prime} * \psi_{\sigma}\right\|_{\infty} \lesssim \epsilon$.

Lemma 13. Given $\sigma \in\left[\underline{\sigma}_{n}, \bar{\sigma}_{n}\right]$ and $F \in \mathcal{M}[-a, a]$, let $F^{\prime}$ be the discrete distribution from the previous lemma. Then $\left\|F * \psi_{\sigma}-F^{\prime} * \psi_{\sigma}\right\|_{1} \lesssim \epsilon \psi^{-1}(\epsilon)$. Moreover, for any $\sigma>0$ there exists a discrete $F^{\prime}$ with a multiple of ( $a \sigma^{-1} \vee$ 1) $\log \epsilon^{-1}$ support points, for which $\left\|F * \psi_{\sigma}-F^{\prime} * \psi_{\sigma}\right\|_{1} \lesssim \epsilon \psi^{-1}(\epsilon)$ and $\| F * \psi_{\sigma}-$ $F^{\prime} * \psi_{\sigma} \|_{\infty} \lesssim \frac{\epsilon}{\sigma}$.

## References

[1] Milton Abramowitz and Irene A. Stegun. Handbook of mathematical functions with formulas, graphs, and mathematical tables, volume 55 of National Bureau of Standards Applied Mathematics Series. For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., http://www.math.sfu.ca/~cbm/aands/, 1964.
[2] R. De Jonge and H. Van Zanten. Adaptive nonparametric Bayesian regression using location-scale mixture priors, preprint. 2009.
[3] Ronald A. DeVore and George G. Lorentz. Constructive approximation, volume 303 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, 1993.
[4] Luc Devroye. A note on the usefulness of superkernels in density estimation. Ann. Statist., 20(4):2037-2056, 1992.
[5] Jean Diebolt and Christian P. Robert. Estimation of finite mixture distributions through Bayesian sampling. J. Roy. Statist. Soc. Ser. B, 56(2):363375, 1994.
[6] Michael D. Escobar and Mike West. Bayesian density estimation and inference using mixtures. J. Amer. Statist. Assoc., 90(430):577-588, 1995.
[7] Stuart Geman and Chii-Ruey Hwang. Nonparametric maximum likelihood estimation by the method of sieves. Ann. Statist., 10(2):401-414, 1982.
[8] Christopher R. Genovese and Larry Wasserman. Rates of convergence for the Gaussian mixture sieve. Ann. Statist., 28(4):1105-1127, 2000.
[9] S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. Ann. Statist., 27(1):143-158, 1999.
[10] Subhashis Ghosal. Convergence rates for density estimation with Bernstein polynomials. Ann. Statist., 29(5):1264-1280, 2001.
[11] Subhashis Ghosal, Jayanta K. Ghosh, and Aad W. van der Vaart. Convergence rates of posterior distributions. Ann. Statist., 28(2):500-531, 2000.
[12] Subhashis Ghosal, Jüri Lember, and Aad Van Der Vaart. On Bayesian adaptation. In Proceedings of the Eighth Vilnius Conference on Probability

Theory and Mathematical Statistics, Part II (2002), volume 79, pages 165175, 2003.
[13] Subhashis Ghosal and Aad van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. Ann. Statist., 35(2):697-723, 2007.
[14] Subhashis Ghosal and Aad W. van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. Ann. Statist., 29(5):1233-1263, 2001.
[15] Ulf Grenander. Abstract inference. John Wiley \& Sons Inc., New York, 1981. Wiley Series in Probability and Mathematical Statistics.
[16] Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G. Walker (Editors). Bayesian Nonparametrics. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2010.
[17] Tzee-Ming Huang. Convergence rates for posterior distributions and adaptive estimation. Ann. Statist., 32(4):1556-1593, 2004.
[18] A. P. Korostelëv and A. B. Tsybakov. Minimax theory of image reconstruction, volume 82 of Lecture Notes in Statistics. Springer-Verlag, New York, 1993.
[19] Willem Kruijer. Convergence Rates in Nonparametric Bayesian Density Estimation. PhD-thesis. Department of Mathematics, Vrije Universiteit Amsterdam, http://www.math.vu.nl/~kruijer/PhDthesis_Kruijer.pdf, 2008.
[20] Willem Kruijer and Aad Van der Vaart. Posterior convergence rates for dirichlet mixtures of beta densities. Journal of Statistical Planning and Inference, 138(7):1981-1992, 2008.
[21] Michael Lavine. Some aspects of Pólya tree distributions for statistical modelling. Ann. Statist., 20(3):1222-1235, 1992.
[22] J.M. Marin, K. Mengersen, and C.P. Robert. Bayesian modelling and inference on mixtures of distributions. Elsevier-Sciences, 2005.
[23] C. Maugis and B. Michel. A non asymptotic penalized criterion for gaussian mixture model selection, forthcoming. ESAIM P8SS.
[24] Whitney K. Newey, Fushing Hsieh, and James M. Robins. Twicing kernels and a small bias property of semiparametric estimators. Econometrica, 72(3):947-962, 2004.
[25] Sylvia Richardson and Peter J. Green. On Bayesian analysis of mixtures with an unknown number of components. J. Roy. Statist. Soc. Ser. B, 59(4):731-792, 1997.
[26] Judith Rousseau. Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. Ann. Statist., 38:146-180, 2010.
[27] C. Scricciolo. Convergence rates of posterior distributions for dirichlet mixtures of normal densities. working paper 2001-21. Technical report, 2001.
[28] C. Scricciolo. Posterior rates of convergence for dirichlet mixtures of exponential power densities, preprint. Technical report, 2010.
[29] C. Shalizi. Dynamics of bayesian updating with dependent data and misspecified models, preprint. 2009.
[30] M. P. Wand and M. C. Jones. Kernel smoothing, volume 60 of Monographs
on Statistics and Applied Probability. Chapman and Hall Ltd., London, 1995.
[31] Yuefeng Wu and Subhashis Ghosal. Kullback Leibler property of kernel mixture priors in Bayesian density estimation. Electron. J. Stat., 2:298331, 2008.
[32] Tong Zhang. From $\epsilon$-entropy to KL-entropy: analysis of minimum information complexity density estimation. Ann. Statist., 34(5):2180-2210, 2006.
[33] Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. IEEE Trans. Inform. Theory, 52(4):1307-1321, 2006.


[^0]:    ${ }^{1}$ In the sequel, a near optimal rate is understood to be the minimax rate with an additional factor $(\log n)^{c}$.

[^1]:    ${ }^{2}$ We emphasize that this global condition is only considered here as a motivation for the construction of $f_{k}$; in the rest of the paper smoothness condition ( C 1 ) is assumed

