Adaptive Bayesian Density Estimation with Location-Scale Mixtures

Willem Kruijer, Judith Rousseau, Aad Van Der Vaart

Adaptive Bayesian Density Estimation with Location-Scale Mixtures

Willem Kruijer, Judith Rousseau and Aad van der Vaart

Address of the First and Second authors:
CEREMADE
Université Paris Dauphine
Place du Marchal De Lattre De Tassigny
75775 PARIS CEDEX 16
France
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 $\exp\{-|x|^p\}$. We construct a finite mixture approximation of densities whose logarithm is locally β-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, Location-Scale 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

$$m(x; k, \mu, w, \sigma) = \sum_{j=1}^{k} w_j \psi_\sigma(x - \mu_j), \tag{1}$$

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

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

Approximation theory (see for example [3]) tells us that for a compactly supported kernel and a compactly supported β-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 ψ 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, σ, w and μ in (1) can be given a prior distribution Π; when there are observations $X_1, \ldots, X_n$ from an unknown density f_0, Bayes' formula gives the posterior

$$\Pi(A \mid X_1, \ldots, X_n) = \frac{\int_A \prod_{i=1}^n m(X_i; k, \mu, w, \sigma) d\Pi(k, \mu, w, \sigma)}{\int \prod_{i=1}^n m(X_i; k, \mu, w, \sigma) d\Pi(k, \mu, w, \sigma)}.$$

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 ϵ_n tending to zero such that $n\epsilon_n^2 \to \infty$ and

$$\Pi(d(f_0, f) > M\epsilon_n \mid X_1 \ldots, X_n) \to 0 \tag{3}$$

in F_0^n-probability, for some sufficiently large constant M, d being the Hellinger- or 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 ψ, provided p is even and f_0 is locally Hölder and has tails bounded by ψ. The constant t in the rate depends on the choice of the prior. We only consider priors independent of β, 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 $(0, C_p]$. When $\sigma = 1$ we also write $\psi(x) = \psi_1(x) = C_p \exp\{-|x|^p\}$ and $\psi^{-1}(y) = \psi_1^{-1}(y)$. For any nonnegative α, let

$$\nu_\alpha = \int x^\alpha \psi(x) dx. \tag{4}$$

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

For any function h, let $K_\sigma h$ denote the convolution $h * \psi_\sigma$, and let $\Delta_\sigma h$ denote the error $(K_\sigma h) - h$.

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

$$\Delta_k = \{x \in \mathbb{R}^k : x_i \geq 0, \sum_{i=1}^k x_i = 1\}, \quad S_k = \{x \in \mathbb{R}^k : x_i \geq 0, \sum_{i=1}^k x_i \leq 1\}$$

and $H_k[b,d] = \{x \in \mathbb{R}^k \mid x_i \in [b_i, d_i]\}$, 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 ϵ,

$$T_{c,\epsilon} = [-c|\log\epsilon|^{1/p}, c|\log\epsilon|^{1/p}]. \tag{5}$$

Given $\epsilon > 0$ and fixed points $x \in \mathbb{R}^k$ and $y \in \Delta_k$, define the l_1-balls

$$\begin{aligned} B_k(x,\epsilon) &= \Big\{z \in \mathbb{R}^k; \sum_{i=1}^k \mid z_i - x_i \mid \leq \epsilon\Big\}, \\ \Delta_k(y,\epsilon) &= \Big\{z \in \Delta_k; \sum_{i=1}^k \mid z_i - y_i \mid \leq \epsilon\Big\}. \end{aligned}$$

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 $\int g dF_0$ 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 β-Hölder, with derivatives $l_j(x) = \frac{d^j}{dx^j}\log f(x)$. We assume the existence of a polynomial L and a constant $\gamma > 0$ such that

$$|l_r(x) - l_r(y)| \leq r! L(x)|x-y|^{\beta - r} \tag{6}$$

for all x, y with $|y - x| \leq \gamma$.

(C2) Tails. There exists $\epsilon > 0$ such that the functions l_j and L satisfy

$$F_0|l_j|^{\frac{2\beta+\epsilon}{j}} < \infty, j = 1,\ldots,r, \qquad F_0 L^{2+\frac{\epsilon}{\beta}} < \infty, \tag{7}$$

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

$$f_0(x) \leq cx^{-\alpha}. \tag{8}$$

(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

$$f_0(x) \leq M_{f_0}\psi(x), \qquad |x| \geq T. \tag{9}$$

(C4) Monotonicity. f_0 is strictly positive, and there exist $x_m < x_M$ such that f_0 is nondecreasing on $(-\infty, x_m)$ and nonincreasing on (x_M, ∞). Without loss of generality we assume that $f_0(x_m) = f_0(x_M) = c$ 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_σ denote convolution over the kernel ψ defined in* (2)*, for any $p \in \mathbb{N}$. Then there exists a density h_k such that for all small enough σ,*

$$\int f \log \frac{f}{K_\sigma h_k} = O(\sigma^{2\beta}), \qquad \int f \left(\log \frac{f}{K_\sigma h_k}\right)^2 = O(\sigma^{2\beta}). \tag{10}$$

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 super-exponential densities like $\exp\{-\exp\{x^2\}\}$. 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 $\frac{d^2}{dx^2}\log f_0(x) = f_0^{''}(x)/f_0(x) - (f'(x)/f(x))^2$, and apart from the additional ϵ in (7), this assumption is equivalent to the assumption in [13] that $F_0(f_0^{''}/f_0)^2$ and $F_0(f_0^{'}/f_0)^4$ 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 (C4) 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 (Π) The prior on σ is the inverse Gamma distribution with scale parameter $\lambda > 0$ and shape parameter $\alpha > 0$, i.e. σ has prior density $\frac{\lambda^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\lambda/x}$ and σ^{-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 μ and weights w are independent. The priors on k, μ and w satisfy the conditions (11)-(14) below.

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

$$B_0 e^{-b_0 k (\log k)^r} \leq \Pi(k) \leq B_1 e^{-b_1 k (\log k)^r}, \tag{11}$$

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_μ on $\mathbb{R}$ satisfying

$$p_\mu(x) \gtrsim \psi(x), \tag{12}$$
$$p_\mu(x) \lesssim e^{-a_1|x|^{a_2}} \quad \text{for constants } a_1 > 0 \text{ and } a_2 \leq p\,. \tag{13}$$

Given k, the prior distribution of the weight vector $w = (w_1, \ldots, w_k)$ is independent of μ, and there is a constant d_1 such that for $\epsilon < \frac{1}{k}$, and $w_0 \in \Delta_k$,

$$\Pi(w \in \Delta_k(w_0, \epsilon) \mid K = k) \gtrsim \exp\Big\{-d_1 k (\log k)^b \log \frac{1}{\epsilon}\Big\}, \tag{14}$$

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

Theorem 2. *Let the bandwidth σ 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 ψ 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 $\Pi(\cdot \mid X_1, \ldots, X_n)$ 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 $t > (2 + b + p^{-1})/(2 + \beta^{-1}) + \max(0, (1-r)/2)$.*

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 μ and w. The prior is to put enough mass near μ_0 and w_0, which are the locations and weights of a mixture approximating f_0. Since μ_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 $D_{KL}(f_0, m) = \int f_0 \log \frac{f_0}{m}$ 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 $D_{KL}(f_0, m) = O(\sigma^{2\beta})$, is to discretize the continuous mixture $K_\sigma f_0$, and show that $\|K_\sigma f_0 - m\|_\infty$ and

$\|f_0 - K_\sigma f_0\|_\infty$ are both $O(\sigma^\beta)$. Under additional assumptions on f_0, this then gives a KL-divergence of $O(\sigma^{2\beta})$. But as $\|f_0 - K_\sigma f_0\|_\infty$ remains of order σ^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 $D_{KL}(f_0, K_\sigma \tilde{f}_0) = O(\sigma^{2\beta})$. To do so, we first construct a not necessarily positive function f_k such that under a global Hölder condition, $\|f_0 - K_\sigma f_k\|_\infty = O(\sigma^\beta)$. However, as we only assume the local Hölder condition (C1), the approximation error of $O(\sigma^\beta)$ 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 $\|f''\|_\infty = L$.[2] Then $\|f - K_\sigma f\|_\infty \leq \frac{1}{2}\nu_2 L \sigma^2$, where ν_2 is defined as in (4). Although the regularity of f is larger than two, the approximation error remains order σ^2. The following calculation illustrates how this can be improved if we take $f_1 = f - \Delta_\sigma f = 2f - K_\sigma f$ as the mixing density instead of f. The approximation error is

$$\begin{aligned}
|(K_\sigma f_1)(x) - f(x)| &= \left| \int \psi_\sigma(x-\mu) \left\{ (f - \Delta_\sigma f)(\mu) - f(x) \right\} d\mu \right| \\
&= \left| \int \psi_\sigma(x-\mu) \left\{ (f(\mu) - f(x)) - \int \psi_\sigma(\epsilon - \mu)(f(\epsilon) - f(\mu)) d\epsilon \right\} d\mu \right| \\
&= \left| \frac{\sigma^2 \nu_2}{2} f''(x) + O(\sigma^3) - \frac{\sigma^2}{2} \int \psi_\sigma(x-\mu) f''(\mu) d\mu - O(\sigma^3) \right| = O(\sigma^3).
\end{aligned}$$

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

$$f_{j+1} = f - \Delta_\sigma f_j, \quad j \geq 0, \quad f_0 := f. \tag{15}$$

Once the approximation error $O(\sigma^\beta)$ 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 β, f_k will be understood as the first function in the sequence $\{f_i\}_{i\in\mathbb{N}}$ for which an error of order σ^β is achieved, i.e. k is such that $\beta \in (2k, 2k+2]$. The construction of the sequence $\{f_i\}_{i\in\mathbb{N}}$ 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 ψ_k such that $\|f - \psi_k * f\|_\infty = O(\sigma^\beta)$, we construct a function f_k fow which $\|f - \psi * f_k\|_\infty = O(\sigma^\beta)$.

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

[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 (C1) is assumed

instead the local Hölder condition (C1) 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 $\beta \in (2k, 2k+2]$, and let f_k be defined as in* (15). *Then for all sufficiently small σ and for all x contained in the set*

$$A_\sigma = \{x : |l_j(x)| \leq B\sigma^{-j}|\log\sigma|^{-\frac{j}{p}}, j = 1, \ldots, r, |L(x)| \leq B\sigma^{-\beta}|\log\sigma|^{-\frac{\beta}{p}}\} \tag{16}$$

we have

$$(K_\sigma f_k)(x) = f(x)\left(1 + O(R(x)\sigma^\beta)\right) + O\left((1 + R(x))\sigma^H\right), \tag{17}$$

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

$$R(x) = r_{r+1}|L(x)| + \sum_{i=1}^{r} r_i |l_i(x)|^{\beta/i}, \tag{18}$$

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 $(K_\sigma f_k)(x) - f(x)$ 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 σ^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

$$J_{\sigma,j} = \{x : f_j(x) > \frac{1}{2}f(x)\}, \tag{19}$$

$$g_j(x) = f_j(x)1_{J_{\sigma,j}} + \frac{1}{2}f(x)1_{J^c_{\sigma,j}}, \tag{20}$$

$$h_j(x) = g_j(x)/\int g_j(x)dx. \tag{21}$$

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 $1 + O(\sigma^\beta)$. For this purpose, we first control integrals over the sets A_σ defined in (16) and

$$E_\sigma = \{x : f(x) \geq \sigma^{H_1}\}, \tag{22}$$

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 σ and all nonnegative integers m and all $K>0$,*

$$\int_{A_\sigma^c} (K_\sigma^m f)(x)dx = O(\sigma^{2\beta}), \qquad \int_{E_\sigma^c} (K_\sigma^m f)(x)dx = O(\sigma^K), \tag{23}$$

provided that H_1 in (22) is sufficiently large. Furthermore, $A_\sigma \cap E_\sigma \subset J_{\sigma,k}$ for small enough σ. Consequently,

$$\int g_k(x)dx = 1 + \int_{J_{\sigma,k}^c} (\frac{1}{2}f - f_k)dx = 1 + O(\sigma^{2\beta}). \tag{24}$$

Finally, when $\beta > 2$, and f_k is defined as in Lemma 1 and h_k as in (21),

$$K_\sigma h_k(x) = f(x)\left(1 + O(R(x)\sigma^\beta)\right) + O\left((1+R(x))\sigma^H\right) \tag{25}$$

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_σ.

Remark 1. *From (20), (21) and (24) it follows that $h_k \geq f/(2(1+O(\sigma^\beta)))$. 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_σ^c in (23) can be shown to be $O(\sigma^{2\beta})$ only using conditions (C1) and (C2), whereas for the integrals over E_σ^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

$$\begin{aligned}\int f(x)\log\frac{f(x)}{K_\sigma h_k(x)}dx \leq & \int_{A_\sigma\cap E_\sigma} \frac{(f(x)-K_\sigma h_k(x))^2}{K_\sigma h_k(x)}dx \\ &+ \int_{A_\sigma^c\cup E_\sigma^c} f(x)\log\frac{f(x)}{K_\sigma h_k(x)}dx + \int_{A_\sigma^c\cup E_\sigma^c}(K_\sigma h_k(x) - f(x))dx.\end{aligned} \tag{26}$$

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 $f(x)O(\sigma^\beta R(x)) - 2O(\sigma^{\beta+H}R(x)) + O((1+R(x))^2)\sigma^{2H}/f(x)$. Let H_1 be such that the second integral in (23) is $O(\sigma^{2\beta})$ (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 $O(\sigma^{2\beta})$ for each of these terms. For example, $\int(1+R(x))^2\sigma^{2H}/f(x)dx = \int f(x)(1+$

$R(x))^2\sigma^{2H}/f^2(x)dx \lesssim \sigma^{2(H-H_1)}$, as $f(x) \geq \sigma^{H_1}$ on E_σ and the Lebesgue measure of this interval is at most σ^{-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

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

which is $O(\sigma^{2\beta})$ 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 ψ be a differentiable symmetric density such that $x\psi'(x)$ 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*

$$\begin{aligned}
\|m - \tilde{m}\|_1 &\leq \|w - \tilde{w}\|_1 + 2\|\psi\|_\infty \sum_{i=1}^k \frac{w_i \wedge \tilde{w}_i}{\sigma \wedge \tilde{\sigma}} \mid \mu_i - \tilde{\mu}_i \mid + \frac{\mid \sigma - \tilde{\sigma} \mid}{\sigma \wedge \tilde{\sigma}}, \\
\|m - \tilde{m}\|_\infty &\lesssim \sum_{i=1}^k \frac{\mid w_i - \tilde{w}_i \mid}{\sigma \wedge \tilde{\sigma}} + \sum_{i=1}^k \frac{w_i \wedge \tilde{w}_i}{(\sigma \wedge \tilde{\sigma})^2} \mid \mu_i - \tilde{\mu}_i \mid + \frac{\mid \sigma - \tilde{\sigma} \mid}{(\sigma \wedge \tilde{\sigma})^2}.
\end{aligned}$$

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

$$\begin{aligned}
\|w_i\psi_\sigma(\cdot - \mu_i) - \tilde{w}_i\psi_{\tilde{\sigma}}(\cdot - \tilde{\mu}_i)\| &\leq \|w_i\psi_\sigma(\cdot - \mu_i) - \tilde{w}_i\psi_\sigma(\cdot - \mu_i)\| \\
&+ \|\tilde{w}_i\psi_\sigma(\cdot - \mu_i) - \tilde{w}_i\psi_\sigma(\cdot - \tilde{\mu}_i)\| + \|\tilde{w}_i\psi_\sigma(\cdot - \tilde{\mu}_i) - \tilde{w}_i\psi_{\tilde{\sigma}}(\cdot - \tilde{\mu}_i)\|
\end{aligned}$$

for any norm. We have the following inequalities:

$$\begin{aligned}
\|\psi_\sigma(z - \mu_i) - \psi_\sigma(z - \tilde{\mu}_i)\|_1 &= 2\left| \Psi\left(\frac{\mu_i - \tilde{\mu}_i}{2\sigma} \right) - \Psi\left(\frac{\tilde{\mu}_i - \mu_i}{2\sigma} \right) \right| \\
&\leq 2\|\psi\|_\infty \frac{\mid \tilde{\mu}_i - \mu_i \mid}{\sigma} \leq \frac{2\|\psi\|_\infty}{\sigma \wedge \tilde{\sigma}} \mid \tilde{\mu}_i - \mu_i \mid, \\
\|\psi_\sigma - \psi_{\tilde{\sigma}}\|_1 &\leq \frac{1}{\sigma \wedge \tilde{\sigma}} \int \mid \psi(\frac{x}{\sigma}) - \psi(\frac{x}{\tilde{\sigma}}) \mid dx \leq \frac{1}{\sigma \wedge \tilde{\sigma}} \mid \sigma - \tilde{\sigma} \mid, \\
\|\psi_\sigma - \psi_{\tilde{\sigma}}\|_\infty &\leq \frac{1}{(\sigma \wedge \tilde{\sigma})^2} \|\frac{d}{dz} g_x\|_\infty \mid \sigma - \tilde{\sigma} \mid, \qquad (27) \\
\|\psi_\sigma(z - \mu_i) - \psi_\sigma(z - \tilde{\mu}_i)\|_\infty &\lesssim \frac{1}{(\sigma \wedge \tilde{\sigma})^2} \mid \tilde{\mu}_i - \mu_i \mid .
\end{aligned}$$

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

$$\|\psi_\sigma - \psi_{\tilde{\sigma}}\|_\infty = \sup_x \mid g_x(z) - g_x(\tilde{z}) \mid \leq \mid z - \tilde{z} \mid \|\frac{d}{dz} g_x\|_\infty \leq \frac{1}{(\sigma \wedge \tilde{\sigma})^2} \|\frac{d}{dz} g_x\|_\infty \mid \sigma - \tilde{\sigma} \mid .$$

Applying the mean value theorem to ψ 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 ψ to be analytic. i.e. p needs to be even.

Lemma 4. *Let the constant H_1 in the definition of E_σ 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 ψ be as in* (2). *Then there exists a finite mixture $m = m(\cdot; k_\sigma, \mu_\sigma, w_\sigma, \sigma)$ with $k_\sigma = O(\sigma^{-1} |\log \sigma|^{1+p^{-1}})$ support points contained in E_σ, such that*

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

Furthermore, (28) *holds for all mixtures $m' = m(\cdot; k_\sigma, \mu, w, \sigma')$ such that $\sigma' \in [\sigma, \sigma + \sigma^{\delta' H_1 + 2}]$, $\mu \in B_{k_\sigma}(\mu_\sigma, \sigma^{\delta' H_1 + 2})$ and $w \in \Delta_{k_\sigma}(w_\sigma, \sigma^{\delta' H_1 + 1})$, where $\delta' \geq 1 + \beta / H_1$.*

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 ϵ,

$$\{x : f(x) \geq \epsilon\} \subset [-c_f |\log \epsilon|^{1/p}, c_f |\log \epsilon|^{1/p}] = T_{c_f, \epsilon}. \tag{29}$$

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 $b = (b_1, \ldots, b_k)$ and $d = (d_1, \ldots, d_k)$, with $b_i < d_i$ for all i, the packing numbers of Δ_k and $H_k[b, d]$ satisfy*

$$D(\epsilon, \Delta_k, l_1) \leq \left(\frac{5}{\epsilon}\right)^{k-1}, \tag{30}$$

$$D(\epsilon, H_k[b, d], l_1) \leq \frac{k! \prod_{i=1}^k (d_i - b_i + 2\epsilon)}{(2\epsilon)^k}. \tag{31}$$

Proof. A proof of (30) can be found in [11]; the other result follows from a volume argument. For λ_k the k-dimensional Lebesgue measure, $\lambda_k(S_k) = \frac{1}{k!}$ and $\lambda_k(B_k(y, \frac{\epsilon}{2}, l_1)) = \frac{\epsilon^k}{k!}$, where $B_k(y, \frac{\epsilon}{2}, l_1)$ 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 ϵ-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, $| z_i - y_i | \leq \frac{\epsilon}{2}$ for all i. Because for each coordinate we have the bounds $| z_i | \leq | y_i | + | z_i - y_i | \leq d_i + \frac{\epsilon}{2}$ and $| z_i | \geq b_i - \frac{\epsilon}{2}$, z is an element of $H_k[b - \frac{\epsilon}{2}, d + \frac{\epsilon}{2}]$. The union of the balls $B_k(x_1, \frac{\epsilon}{2}, l_1), \ldots, B_k(x_N, \frac{\epsilon}{2}, l_1)$ is therefore contained in $H_k[b - \frac{\epsilon}{2}, d + \frac{\epsilon}{2}]$. □

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 $m = m(\cdot; k_n, \mu^{(n)}, w^{(n)}, \sigma_n)$ whose KL-divergence from f_0 is $O({\sigma_n}^{2\beta}) = O(\widetilde{\epsilon}_n^2)$. The number of components is

$$k_n = O({\sigma_n}^{-1} |\log \sigma_n|^{1+p^{-1}}) = O\left(n^{1/(1+2\beta)}(\log n)^{1+p^{-1}-t_1/\beta}\right), \qquad (32)$$

their locations being contained in the set E_{σ_n} defined in (22). By the same lemma there are l_1-balls $B_n = B_{k_n}(\mu^{(n)}, {\sigma_n}^{\delta' H_1+2})$ and $\Delta(n) = \Delta_{k_n}(w^{(n)}, {\sigma_n}^{\delta' H_1+1})$ such that the same is true for all k_n-dimensional mixtures $m = m(\cdot; k_n, \mu, w, \sigma)$ with $\sigma \in [\sigma_n, \sigma_n + {\sigma_n}^{\delta' H_1+2}]$ 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 $[\sigma_n, \sigma_n + {\sigma_n}^{\delta' H_1+2}]$.

Let $b = \delta' H_1 + 2$; as σ^{-1} is inverse-gamma, it follows from the mean value theorem that

$$\begin{aligned} \Pi(\sigma \in [\sigma_n, \sigma_n + {\sigma_n}^b]) &= \int_{\sigma_n}^{\sigma_n + {\sigma_n}^b} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\lambda/x} dx \\ &\geq \int_{\sigma_n}^{\sigma_n + {\sigma_n}^b} \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-2\lambda/x} dx \geq 4 \frac{\lambda^{\alpha+1}}{\Gamma(\alpha)} {\sigma_n}^{b-2} e^{-\lambda {\sigma_n}^{-1}}, \end{aligned} \qquad (33)$$

which is larger than $\exp\{-n\widetilde{\epsilon}_n^2\}$ for any choice of $t_1 \geq 0$. Condition (11) gives a lower bound of $B_0 \exp\{-b_0 k_n \log^r k_n\}$ on $\Pi(k_n)$, which is larger than $\exp\{-n\widetilde{\epsilon}_n^2\}$ when $(2 + \beta^{-1})t_1 > 1 + p^{-1} + r$. Given that there are k_n components, condition (14) gives a lower bound on $\Pi(\Delta(n))$, which is larger than $\exp\{-n\widetilde{\epsilon}_n^2\}$ when $(2+\beta^{-1})t_1 > 2 + b + p^{-1}$. The required lower-bound for $\Pi(B_n)$ follows from (9) and the fact that $\mu_1^{(n)}, \ldots, \mu_{k_n}^{(n)}$ are independent with prior density p_μ satisfying (12). The 'target' mixture given by Lemma 4 has location vector $\mu^{(n)}$, whose elements are contained in E_{σ_n}. By (9), E_{σ_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_μ is lower bounded by a multiple of ${\sigma_n}^{c_f^p H_1}$ at the boundaries of this interval. Consequently, for all $i = 1, \ldots, k_n$,

$$\Pi\left(| \mu_i - \mu_i^{(n)} | \leq \frac{{\sigma_n}^{\delta' H_1 + 2}}{k_n} \right) \gtrsim \frac{{\sigma_n}^{\delta' H_1 + 2 + c_f^p H_1}}{k_n}.$$

As the l_1-ball $B_{k_n}(\mu^{(n)}, {\sigma_n}^{\delta' H_1+2})$ contains the l_∞-ball $\{\mu \in \mathbb{R}^{k_n} : | \mu_i - \mu_i^{(n)} | \leq \frac{{\sigma_n}^{\delta' H_1+2}}{k_n},\ 1 \leq i \leq k_n\}$, we conclude that

$$\Pi\left(\mu \in B_n\right) \gtrsim \exp\{-dk_n \log n\}$$

for some constant $d > 0$. Combining the above results it follows that $\Pi(KL(f_0, \widetilde{\epsilon}_n)) \geq \exp\{-n\widetilde{\epsilon}_n^2\}$ when $t_1 > (2 + b + p^{-1})/(2 + \beta^{-1})$.

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

$$\mathcal{F}_n = \big\{m(\cdot; k, \mu, w, \sigma) | k \leq r_n, \mu \in H_k[-b_n, b_n], \sigma \in S_n\big\}.$$

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

$$\begin{aligned}\int_{\bar{\sigma}_n}^{\infty} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\frac{\lambda}{x}} dx &= \int_0^{\bar{\sigma}_n^{-1}} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} dx \\ &\leq \int_0^{\bar{\sigma}_n^{-1}} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} dx = O(\exp\{-\alpha n\widetilde{\epsilon}_n^2(\log n)^\delta\}).\end{aligned}$$

Hence $\Pi(S_n^c) \leq e^{-cn\widetilde{\epsilon}_n^2}$ 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 $\exp\{-b_1 k_n \log^{r_n} k_n\}$. The prior mass on mixtures with at least one support point outside $[-b_n, b_n]$ is controlled as follows. By conditions (11) and (13), the probability that a certain μ_i is outside $[-b_n, b_n]$, is

$$\Pi(| \mu_i |> b_n) = \int_{[-b_n,b_n]^c} p_\mu(x)dx \lesssim b_n^{\max\{0,1-a_2\}} e^{-b_n^{a_2}}. \tag{34}$$

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

$$\begin{aligned}\Pi(N([-b_n, b_n]^c) > 0) &= \sum_{k=1}^{\infty} \Pi(K = k)\, \Pi(\max_{i=1,\ldots,k} | \mu_i |> b_n \mid K = k) \\ &\leq \sum_{k=1}^{\infty} \Pi(k) k\, \Pi(| \mu_i |> b_n) \lesssim e^{-|b_n|^{a_2}}.\end{aligned} \tag{35}$$

Combining these bounds, we find

$$\Pi(\mathcal{F}_n^c) \leq \Pi(S_n^c) + \sum_{k=r_n}^{\infty} \rho(k) + \Pi(N([-b_n, b_n]^c > 0)) \lesssim e^{-b_1 r_n (\log n)^r}.$$

The right hand side decreases faster than $e^{-n\widetilde{\epsilon}_n^2}$ if $t_r + r > 2t_1$.

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

$$\mathcal{F}_{n,j} = \big\{m(\cdot;k,\mu,w,\sigma)|k \le r_n, \mu \in H_k[-b_n,b_n], \sigma \in S_{n,j}\big\},$$

$$S_{n,j} = (s_{n,j-1}, s_{n,j}] = (\underline{\sigma}_n(1+\widetilde{\epsilon}_n)^{j-1}, \underline{\sigma}_n(1+\widetilde{\epsilon}_n)^{j}], \quad j = 1,\dots,J_n,$$

$$J_n = \left(\log\frac{\bar{\sigma}_n}{\underline{\sigma}_n}\right) / \log(1+\epsilon_n) = O\left(n\widetilde{\epsilon}_n(\log n)^{\delta}\right).$$

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

$$\begin{aligned}
\Pi(\mathcal{F}_{n,j}) &\le \Pi(S_{n,j}) = \Pi(\sigma^{-1} \in [\underline{\sigma}_n^{-1}(1+\widetilde{\epsilon}_n)^{-j}, \underline{\sigma}_n^{-1}(1+\widetilde{\epsilon}_n)^{1-j})) \\
&= \int_{\underline{\sigma}_n^{-1}(1+\widetilde{\epsilon}_n)^{-j}}^{\underline{\sigma}_n^{-1}(1+\widetilde{\epsilon}_n)^{1-j}} y^{\alpha-1}e^{-\lambda y}dy \\
&\le \lambda^{-1}\max\{(\underline{\sigma}_n^{-1}(1+\widetilde{\epsilon}_n)^{-j})^{\alpha-1}, (\underline{\sigma}_n^{-1}(1+\widetilde{\epsilon}_n)^{1-j})^{\alpha-1}\}\exp\{-\lambda\underline{\sigma}_n^{-1}(1+\widetilde{\epsilon}_n)^{-j}\} \\
&\lesssim \underline{\sigma}_n^{1-\alpha}(1+\widetilde{\epsilon}_n)^{-(\alpha-1)j}\exp\{-\lambda\underline{\sigma}_n^{-1}(1+\widetilde{\epsilon}_n)^{-j}\}.
\end{aligned} \tag{36}$$

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

$$\mathcal{F}_{n,j} = \cup_{k=1}^{r_n}\mathcal{F}_{n,j,k} = \cup_{k=1}^{r_n}\big\{m(\cdot;k,\mu,w,\sigma)|\mu \in H_k[-b_n,b_n], \sigma \in S_{n,j}\big\}.$$

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 (A,d_1) and (B,d_2) we have $D(\epsilon, A\times B, d_1+d_2) \le D(\frac{\epsilon}{2}, A, d_1)D(\frac{\epsilon}{2}, B, d_2)$, Lemma 3 implies that for all $k \ge 1$, $D(\bar{\epsilon}_n, \mathcal{F}_{n,j,k}, \|\cdot\|_1)$ is bounded by

$$D\Big(\frac{\bar{\epsilon}_n}{3}, \Delta_k, l_1\Big)\, D\Big(\frac{\bar{\epsilon}_n s_{n,j-1}}{6\|\psi\|_\infty}, H_k[-b_n,b_n], l_1\Big)\, D\Big(\frac{\bar{\epsilon}_n s_{n,j-1}}{3}, (s_{n,j-1}, s_{n,j}], l_1\Big).$$

Lemma 5 provides the following bounds:

$$\begin{aligned}
D\Big(\frac{\bar{\epsilon}_n}{3}, \Delta_k, l_1\Big) &\le \left(\frac{15}{\bar{\epsilon}_n}\right)^{k-1}, \\
D\Big(\frac{\bar{\epsilon}_n s_{n,j-1}}{6\|\psi\|_\infty}, H_k[-b_n,b_n], l_1\Big) &\le k!\Big(\frac{\bar{\epsilon}_n s_{n,j-1}}{3\|\psi\|_\infty}\Big)^{-k}\prod_{i=1}^{k}\Big(2b_n + \frac{\bar{\epsilon}_n s_{n,j-1}}{3\|\psi\|_\infty}\Big), \\
D\Big(\frac{\bar{\epsilon}_n s_{n,j-1}}{3}, (s_{n,j-1}, s_{n,j}], l_1\Big) &\le \big(s_{n,j-1}\bar{\epsilon}_n/3\big)\Big((s_{n,j}-s_{n,j-1}) + \bar{\epsilon}_n s_{n,j-1}/3\Big).
\end{aligned}$$

For some constant C, we find that

$$\begin{aligned}
D(\bar{\epsilon}_n, \mathcal{F}_{n,j}, \|\cdot\|_1) &\le r_n D(\bar{\epsilon}_n, \mathcal{F}_{n,j,r_n}, \|\cdot\|_1) \\
&\lesssim r_n C^{r_n} r_n!(\bar{\epsilon}_n)^{-2r_n} s_{n,j} s_{n,j-1}^{-r_n+1}(\max(b_n, \bar{\epsilon}_n s_{n,j-1}))^{r_n}.
\end{aligned} \tag{37}$$

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

$$\begin{aligned}
&\underline{\sigma}_n^{(1-\alpha)/2}(1+\widetilde{\epsilon}_n)^{-(\alpha-1)j/2}\sqrt{r_n}C^{r_n/2}r_n^{r_n/2+1/2}(\bar{\epsilon}_n)^{-r_n}\sqrt{s_{n,j}}\\
&\quad s_{n,j-1}^{-r_n/2+1/2}b_n^{r_n/2}\exp\{-\frac{\lambda}{2}\underline{\sigma}_n^{-1}(1+\widetilde{\epsilon}_n)^{-j}\}\\
&\lesssim n^{\frac{A}{2}r_n+\frac{\alpha-3}{2}A}(1+\widetilde{\epsilon}_n)^{-\frac{1}{2}(j-1)(r_n+\alpha-2)+\frac{1-\alpha}{2}}(r_n+1)^{r_n+1}C^{\frac{r_n}{2}}\bar{\epsilon}_n^{-r_n}b_n^{\frac{r_n}{2}}\exp\{-\lambda b_n^{-1}\frac{\bar{\epsilon}_n}{1+\widetilde{\epsilon}_n}\}\\
&\lesssim K_0\exp\{K_1 r_n(\log n)\},
\end{aligned}$$

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 $\bar{\epsilon}_n^{-r_n/2}n^{-Ar_n/2}(1+\widetilde{\epsilon}_n)^{(j-1)r_n/2}$, where the factor $(1+\widetilde{\epsilon}_n)^{(j-1)r_n/2}$ cancels out with $(1+\widetilde{\epsilon}_n)^{-(j-1)r_n/2}$ on the third line of the above display. There is however a remaining factor $(1+\widetilde{\epsilon}_n)^{\frac{1}{2}(j-1)(2-\alpha)}$. Since J_n is defined such that $n^{-A}(1+\widetilde{\epsilon}_n)^{J_n} = \exp\{n\widetilde{\epsilon}_n^2(\log n)^\delta\}$, the sum of $\sqrt{\Pi(\mathcal{F}_{n,j})}\sqrt{N(\bar{\epsilon}_n, \mathcal{F}_{n,j}, d)}$ over $j = 1, \ldots, J_n$ is a multiple of $\exp\{K_1 r_n(\log n) + n\widetilde{\epsilon}_n^2(\log n)^\delta\}$, which increases at a slower rate than $\exp\{n\bar{\epsilon}_n^2\}$ if $2t_2 > \max(t_r+1, 2t_1+\delta)$. Combined with the requirement that $t_r + r > 2t_1$ 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 $t > (2+b+p^{-1})/(2+p^{-1}) + \max(0, (1-r)/2)$. □

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 $X_i = Y_i/(Y_1 + \ldots + Y_k)$, $i = 1, \ldots, k$. For $(x_1, \ldots, x_{k-1}) \in S_{k-1}$,

$$\begin{aligned}
P(X_1 \leq x_1, \ldots, X_{k-1} \leq x_{k-1}) &= \int_0^\infty P(Y_1 \leq x_1 y, \ldots Y_{k-1} \leq x_{k-1}y)\, dP^{Y_1+\ldots+Y_k}(y)\\
&= \int_0^\infty \int_0^{x_1 y}\int_0^{x_2 y}\cdots\int_0^{x_{k-1}y} f_k(y - \sum_{i=1}^{k-1} s_i)\prod_{i=1}^{k-1} f_i(s_i)ds_1\cdots ds_{k-1}dy.
\end{aligned} \tag{38}$$

The corresponding density is

$$\begin{aligned} f^{X_1,\dots,X_{k-1}}(x_1,\dots,x_{k-1}) &= \int_0^\infty y^{k-1} f_k(y - \sum_{i=1}^{k-1} x_i y) \prod_{i=1}^{k-1} f_i(x_i y) dy \\ &= \int_0^\infty y^{k-1} \prod_{i=1}^{k} f_i(x_i y) dy, \end{aligned} \tag{39}$$

where $x_k = 1 - \sum_{i=1}^{k-1} x_i$. We obtain a result similar to lemma 8 in [13].

Lemma 6. *Let $X_1,\dots,X_k$ 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 $i = 1,\dots,k$, $f_i(z) \geq c_1(k) z^{c_3}$ if $z \in [0, c_2(k)]$. Then there are constants c and C such that for all $y \in \Delta_k$ and all $\epsilon \leq (\frac{1}{k} \wedge c_1(k) c_2(k)^{c_3+1})$*

$$P\Big(X \in \Delta_k(y, 2\epsilon)\Big) \geq C e^{-ck\log(\frac{1}{\epsilon})}.$$

Proof. As in [13] it is assumed that $y_k \geq k^{-1}$. Define $\underline{\delta_i} = \max(0, y_i - \epsilon^2)$ and $\bar{\delta}_i = \min(1, y_i + \epsilon^2)$. If $x_i \in (\underline{\delta_i}, \bar{\delta}_i)$ for $i = 1,\dots,k-1$, then $\sum_{i=1}^k \mid x_i - y_i \mid \leq 2\sum_{i=1}^{k-1} \mid x_i - y_i \mid \leq 2(k-1)\epsilon^2 \leq \epsilon$. Note that $(x_1,\dots,x_{k-1}) \in S_k$, 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,

$$f(x_1,\dots,x_{k-1}) \geq \int_0^{c_2(k)} y^{k-1} \prod_{i=1}^k \big(c_1(k)(x_i y)^{c_3}\big) dy = \frac{\big(c_2(k)^{c_3+1} c_1(k)\big)^k}{(c_3+1)k} (x_1 \cdot \ldots \cdot x_k)^{c_3}.$$

Because

$$x_k = \left| 1 - \sum_{j=1}^{k-1} x_j \right| = \left| y_k + \sum_{j=1}^{k-1} (y_j - x_j) \right| \geq k^{-1} - (k-1)\epsilon^2 \geq \epsilon^2 \geq \frac{1}{k^2},$$

$$\begin{aligned} P\Big(X \in B_k(y,\epsilon)\Big) &\geq \frac{1}{k^{2c_3}} \frac{\big(c_2(k)^{c_3+1} c_1(k)\big)^k}{(c_3+1)k} \prod_{j=1}^{k-1} \int_{\underline{\delta_j}}^{\bar{\delta}_j} x_j^{c_3} dx_j \geq \frac{\big(c_2(k)^{c_3+1} c_1(k)\big)^k}{(c_3+1)^2 k} \epsilon^{2k(c_3+1)-2} \\ &\geq \exp\left\{ k \log(c_2(k)^{c_3+1} c_1(k)) - \log(c_3+1) - \log(k) - 2k\log(\frac{\sqrt{2}}{\epsilon}) \right\}. \end{aligned}$$

As $\epsilon \leq (\frac{1}{k} \wedge c_1(k) c_2(k)^{c_3+1})$, there are constants c and C for which this quantity is lower-bounded by $C e^{-ck\log(\frac{1}{\epsilon})}$. □

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

$$X_\epsilon = \prod_{j=1,\epsilon_j=0}^m U_{\epsilon_1\cdots\epsilon_{j-1}} \prod_{j=1,\epsilon_j=1}^m \big(1 - U_{\epsilon_1\cdots\epsilon_{j-1}}\big),$$

where $\big(U_\delta, \delta \in E_*^{m-1}\big)$ is a family of beta random variables with parameters $\big((\alpha_{\delta_1}, \alpha_{\delta_2}), \delta \in E_*^{m-1}\big)$. We only consider symmetric beta densities, for which $\alpha_\delta = \alpha_{\delta_1} = \alpha_{\delta_2}$. Adding pairs of coordinates, lower dimensional vectors X_δ can be defined for $\delta \in E_*^{m-1}$. For $\delta \in E_*^{m-1}$, let $X_{\delta 0} = U_\delta X_\delta$ and $X_{\delta 1} = \big(1-U_\delta\big)X_\delta$, and $X_\emptyset = 1$ by construction. If $\alpha_\delta = \frac{1}{2}\alpha_{\delta_1\cdots\delta_{i-1}}$ for all $1 \le i \le m$ and $\delta \in E^i$, X is Dirichlet distributed.

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

$$\begin{aligned} p_m(y,\eta) = P\Big(X \in \Delta_k(y,\eta)\Big) &= P(\sum_{\epsilon \in E^m} \mid X_\epsilon^m - y_\epsilon^m \mid \le \eta) \\ &\ge \prod_{i=1}^m P(\max_{\partial \in E^{i-1}} \mid U_\delta - \frac{y_{\delta 0}}{y_\delta} \mid \le \frac{\eta}{2^{m-i+2}}). \end{aligned}$$

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

$$\begin{aligned} \mid U_\delta X_\delta - y_{\delta 0} \mid &\le U_\delta \mid X_\delta - y_\delta \mid + y_\delta \mid U_\delta - \frac{y_{\delta 0}}{y_\delta} \mid, \\ \mid (1-U_\delta)X_\delta - y_{\delta 1} \mid &\le (1-U_\delta) \mid X_\delta - y_\delta \mid + y_\delta \mid (1-U_\delta) - \frac{y_\delta - y_{\delta 0}}{y_\delta} \mid. \end{aligned}$$

Consequently,

$$\begin{aligned} \sum_{\delta \in E^m} \mid X_\delta - y_\delta \mid &= \sum_{\delta \in E^{m-1}} \mid X_{\delta 0} - y_{\delta 0} \mid + \mid X_{\delta 1} - y_{\delta 1} \mid \\ &\le \sum_{\delta \in E^{m-1}} \mid X_\delta - y_\delta \mid + 2 \sum_{\delta \in E^{m-1}} y_\delta \mid U_\delta - \frac{y_{\delta 0}}{y_\delta} \mid \\ &\le \sum_{\delta \in E^{m-1}} \mid X_\delta - y_\delta \mid + 2 \max_{\delta \in E^{m-1}} \mid U_\delta - \frac{y_{\delta 0}}{y_\delta} \mid. \end{aligned}$$

Hence,

$$\begin{aligned} p_m(y,\eta) &\ge p_{m-1}(y, \frac{\eta}{2}) P(\max_{\partial \in E^{m-1}} \mid U_\delta - \frac{y_{\delta 0}}{y_\delta} \mid \le \frac{\eta}{4}) \\ &\ge \prod_{i=2}^m P(\max_{\partial \in E^{i-1}} \mid U_\delta - \frac{y_{\delta 0}}{y_\delta} \mid \le \frac{\eta}{2^{m-i+2}}) P(\mid U_\emptyset - y_0 \mid \le \frac{\eta}{2^m}) \\ &\ge \prod_{i=1}^m P(\max_{\partial \in E^{i-1}} \mid U_\delta - \frac{y_{\delta 0}}{y_\delta} \mid \le \frac{\eta}{2^{m-i+2}}), \end{aligned}$$

as

$$\begin{aligned} p_1(\eta 2^{-m}) &= P(\mid X_0 - y_0 \mid + \mid X_1 - y_1 \mid \le \eta 2^{-m}) \\ &= P(\mid U_0 - y_0 \mid + \mid (1-U_0) - (1-y_0) \mid \le \eta 2^{-m}) = P(\mid U_0 - y_0 \mid \le \eta 2^{-m-1}). \end{aligned}$$

□

With $\delta \in E^{i-1}$ fixed, we can lower-bound $P(| U_\delta - \frac{y_{\delta 0}}{y_\delta} |\leq \frac{\eta}{2^{m-i+2}})$ for various values of the α_δ. In the remainder we will assume that $\alpha_\delta = \alpha_i$, for all $\delta \in E^{i-1}$, with $i = 1, \dots, m$. For increasing $\alpha_i \geq 1$, U_δ 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 α_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.

If $\alpha_i \uparrow \infty$, $i = 1, \dots, m$ when $m \to \infty$, then

$$P(| U_\delta |\leq \eta 2^{-m+i-2}) = \int_0^{\eta 2^{-m+i-2}} \frac{\Gamma(2\alpha_i)}{\Gamma^2(\alpha_i)} x^{\alpha_i - 1}(1-x)^{\alpha_i - 1} dx$$
$$\gtrsim \int_0^{\eta 2^{-m+i-2}} \alpha_i{}^{-\frac{1}{2}} 2^{2\alpha_i - \frac{1}{2}} \frac{1}{2} x^{\alpha_i - 1} dx = 2^{-(m-i)\alpha_i - \frac{3}{2}} \alpha_i{}^{-\frac{3}{2}} \eta^{\alpha_i}.$$

At the ith level there are 2^{i-1} independent variables U_δ with the Beta(α_i, α_i) distribution, and therefore

$$\log\Big(p_m(y,\eta)\Big) \gtrsim \log \prod_{i=1}^m \big(2^{-(m-i)\alpha_i - \frac{3}{2}} \alpha_i{}^{-\frac{3}{2}} \eta^{\alpha_i}\big)^{2^{i-1}}$$
$$= \sum_{i=1}^m 2^{i-1} \big\{-\alpha_i \log\frac{1}{\eta} - \frac{3}{2}\log(\alpha_i) - \alpha_i(m-i)\log(2)\big\}.$$

If $\alpha_i \downarrow 0$, $i = 1, \dots, m$ when $m \to \infty$, we have

$$P(| U_\delta - \frac{1}{2} |\leq \eta 2^{-m+i-2}) = \int_{1/2 - \eta 2^{-m+i-2}}^{1/2 + \eta 2^{-m+i-2}} \frac{\Gamma(2\alpha_i)}{\Gamma^2(\alpha_i)} x^{\alpha_i - 1}(1-x)^{\alpha_i - 1} dx$$
$$\gtrsim \alpha_i \eta 2^{-m+i-1} \big(\frac{1}{4}\big)^{\alpha_i - 1},$$

$$\log\Big(p_m(y,\eta)\Big) \gtrsim \sum_{i=1}^m 2^{i-1} \big\{\log(\alpha_i) - \big(2\alpha_i + (m-i-1)\big)\log(2) - \log\frac{1}{\eta}\big\}.$$

We have the following application of these results.

Lemma 8. *Let X_δ^m be Polya distributed with parameters α_i. If $\alpha_i = i^b$ for $b > 0$,*

$$P(X \in \Delta_k(y,\eta)) \geq C \exp\{-ck(\log k)^b \log\frac{1}{\eta}\},$$

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 $i = 1, \dots, m$,

$$P(X \in \Delta_k(y,\eta)) \geq C \exp\{-ck \log\frac{1}{\eta}\},$$

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 σ_j^2 of all components are required to be the same.

Furthermore, the prior on the σ_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 KL-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 $\Pi(\mathcal{F}_n|X_1,\dots,X_n) \to 0$ 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 $\bar{\epsilon}_n \to 0$, 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 $\{X_i\}_{i\geq 1}$ 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* $\bar{\epsilon}_n \to 0$ *and* $n\widetilde{\epsilon}_n^2 \to \infty$,

$$\sum_{j=-\infty}^{\infty} \sqrt{N(\bar{\epsilon}_n, \mathcal{F}_{n,j}, d)}\sqrt{\Pi_n(\mathcal{F}_{n,j})}e^{-n\bar{\epsilon}_n^2} \quad \to \quad 0, \tag{40}$$

$$\Pi_n(KL(f_0, \widetilde{\epsilon}_n)) \quad \geq \quad e^{-n\widetilde{\epsilon}_n^2}, \tag{41}$$

$$\Pi_n(\mathcal{F}_n^c) \leq e^{-4n\widetilde{\epsilon}_n^2}, \tag{42}$$

where $KL(f_0, \widetilde{\epsilon}_n)$ *is the Kullback-Leibler ball*

$$\{f : F_0 \log(f_0/f) \leq \widetilde{\epsilon}_n^2, F_0 \log^2(f_0/f) \leq \widetilde{\epsilon}_n^2\}.$$

Then $\Pi_n(f \in \mathcal{F} : d(f, f_0) > 8\bar{\epsilon}_n \mid X_1, \dots, X_n) \to 0$ *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 ϵ_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* $\alpha > 0$,

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

where $0 < \theta(\alpha) < \frac{1}{12\alpha}$. *If* $\alpha \to \infty$, *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* $\alpha \to 0$, *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* $c = 0.8856\dots$ *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* $j \geq 1$,

$$\frac{\sqrt{\Gamma\left(\frac{2j+1}{1+\alpha}\right)}}{j!} \quad \leq \quad \frac{1}{\sqrt{2\pi}}e^{\frac{\alpha}{1+\alpha}(j+1)}\left(\frac{2}{1+\alpha}\right)^{\frac{j}{1+\alpha}}(j+1)^{-\frac{\alpha j}{1+\alpha}}, \tag{44}$$

$$\frac{\Gamma\left(\frac{j+1}{1+\alpha}\right)}{j!} \quad \leq \quad e^{\frac{\alpha}{1+\alpha}(j+1) + \frac{1}{12}}\left(\frac{1}{1+\alpha}\right)^{\frac{j}{1+\alpha}}(j+1)^{-\frac{\alpha j}{1+\alpha}}. \tag{45}$$

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* φ *be the* m*-fold convolution* $\psi * \cdots * \psi$. *Then for any* $\alpha \geq 0$ *there is a number* $k' = k'(p, \alpha, m)$ *such that for all sufficiently small* $\sigma > 0$,

$$\int_{|x| > k'|\log\sigma|^{1/p}} \varphi(x)|x|^\alpha dx = \sigma^H. \tag{46}$$

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

$$P(Z > y) = \int_y^\infty \psi_{(p)}(x)dx \le p^{-1}y^{1-p}\int_y^\infty px^{p-1}\psi_{(p)}(x)dx = p^{-1}y^{1-p}\psi_{(p)}(y).$$

For $m = 1$, we have

$$\int_y^\infty x^\alpha\psi_{(p)}(x)dx = \int_{y^{1+\alpha}}^\infty \psi_{(p)}\left(z^{1/(1+\alpha)}\right)dx = \frac{C_p}{C_{p/(1+\alpha)}}\int_{y^{1+\alpha}}^\infty \psi_{(p/(1+\alpha))}(z)dz$$
$$= \frac{C_p}{C_{p/(1+\alpha)}}P_{Z\sim\psi_{(p/(1+\alpha))}}(Z > k'^{(1+\alpha)}|\log\sigma|^{\frac{1+\alpha}{p}})$$

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 $\alpha \ge 1$ then, by Jensen's inequality applied to the function $x \mapsto x^\alpha$,

$$E\left(|Z|^\alpha 1_{|Z|>k'|\log\sigma|^{1/p}}\right) \le E\left(m^{\alpha-1}\left(\sum_{i=1}^m |Z_i|^\alpha\right)1_{|Z|>k'|\log\sigma|^{1/p}}\right)$$
$$\le m^{\alpha-1}\sum_{i=1}^m E\left(|Z_i|^\alpha\sum_{j=1}^m 1_{|Z_j|>\frac{k'}{m}|\log\sigma|^{1/p}}\right) = \sigma^H,$$

where we used (46) with $\alpha = 0$ and the independence of the Z_i's to bound the terms with $i \ne j$. 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 β, let $\mathcal{H}(\beta, L)$ be the space of functions h such that $\sup_{x\ne y}|h^{(r)}(x) - h^{(r)}(y)|/|y-x|^{\beta-r} \le L$, where $h^{(r)}$ is the rth derivative of h. Let H_β be the Hölder-space $\cup_{L>0}\mathcal{H}(\beta, L)$, and given some function $h \in H_\beta$, let $L_{h,\beta-r} = \sup_{x\ne y}|h^{(r)}(x) - h^{(r)}(y)|/|y-x|^{\beta-r}$. When $\beta - r = 1$, this equals $\|h^{(r+1)}\|_\infty$.

Lemma 11. *Let $f \in H_\beta$, where $2k < \beta \le 2k+2$ for some nonnegative integer k. Then $\|f - f_k * \psi_\sigma\|_\infty = O(\sigma^\beta)$, where f_k is defined recursively by $f_0 = f$, $f_1 = f - \Delta_\sigma f = 2f - K_\sigma f$ and $f_{j+1} = f - \Delta_\sigma f_j$, $j \ge 1$.*

Proof. By induction it follows that

$$f_k = \sum_{i=0}^k (-1)^i\binom{k+1}{i+1}K_\sigma^i f, \qquad \Delta_\sigma^k f = \sum_{i=0}^k (-1)^{k-i}\binom{k}{i}K_\sigma^i f. \tag{47}$$

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_β, even if ψ itself is not in H_β (e.g.

when ψ is the Laplace kernel). Second, it follows from the symmetry of ψ that $K_\sigma f^{(r)} = \frac{d^r}{dx^r} K_\sigma f$, i.e. the rth 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 $K_\sigma(f_1) - f = \Delta_\sigma(f - \Delta_\sigma(f)) - \Delta_\sigma(f) = -\Delta_\sigma \Delta_\sigma f$, and $\|\Delta_\sigma \Delta_\sigma f\|_\infty \leq \nu_2 \sigma^2 \|(\Delta_\sigma f)''\|_\infty$. Because differentiation and the Δ_σ operator can be interchanged, we also have $\|(\Delta_\sigma f)''\|_\infty = \|(\Delta_\sigma f'')\|_\infty$. Since $f'' \in H_{\beta-2}$, the latter quantity is $O(\sigma^{\beta-2})$. Consequently, $\|\Delta_\sigma \Delta_\sigma f\|_\infty = O(\sigma^\beta)$. For $k > 1$, we repeat this step and use that, as a consequence of (47), $\|K_\sigma f_k - f\|_\infty = \|\Delta_\sigma^{k+1} f\|_\infty$. From the following induction argument it follows that for any positive integer k, $\beta \in (2k, 2k+2]$ and $f \in H_\beta$, $\|\Delta_\sigma^{k+1} f\|_\infty = O(\sigma^\beta)$. Suppose this statement holds for $k = 0, 1, \ldots, m-1$, and that $f \in H_\beta$ with $\beta \in (2m, 2m+2]$. Then $\|\Delta_\sigma^m f\|_\infty = O(\|\Delta_\sigma f^{(2m)}\|_\infty \sigma^{2m})$ and $\|\Delta_\sigma f^{(2m)}\|_\infty = O(\sigma^{\beta-2m})$ as $f^{(2m)} \in H_{\beta-2m}$. □

Appendix C: Proof of Lemma 1

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

$$\log f_0(y) \leq \log f_0(x) + \sum_{j=1}^{r} \frac{l_j(x)}{j!}(y-x)^j + L(x)|y-x|^\beta \tag{48}$$

$$\log f_0(y) \geq \log f_0(x) + \sum_{j=1}^{r} \frac{l_j(x)}{j!}(y-x)^j - L(x)|y-x|^\beta, \tag{49}$$

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 β. 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

$$K_\sigma f(x) \leq (1 + O((|L(x)| + |l_1^\beta(x)|)\sigma^\beta))f(x) + O(1 + |L(x)| + |l_1^\beta(x)|)\sigma^H. \tag{50}$$

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 $x \in \mathbb{R}$,

$$D_x = \{y : |y-x| \leq k'\sigma|\log\sigma|^{1/p}\},$$

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

$$K_\sigma f(x) \leq f(x) \int_{D_x} e^{B_{f,r}(x,y)} \psi_\sigma(y-x)dy + \int_{D_x^c} f(y)\psi_\sigma(y-x)dy. \tag{51}$$

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

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

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

$$\begin{aligned} e^{B_{f,r}(x,y)} &\leq 1 + B_{f,r}(x,y) + MB_{f,r}^2(x,y) = 1 + l_1(x)(y-x) + L(x)|y-x|^\beta \\ &\quad + M\left(l_1^2(x)(y-x)^2 + 2l_1(x)L(x)(y-x)|y-x|^\beta + L^2(x)|y-x|^{2\beta}\right). \end{aligned} \tag{53}$$

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

$$f(x)\int_{D_x}\psi_\sigma(y-x)\left\{1 + L(x)|y-x|^\beta + M(k'B)^{2-\beta}|l_1(x)(y-x)|^\beta + Mk'^\beta B|L(x)(y-x)|^\beta\right\}dy, \tag{54}$$

since $|l_1(x)(y-x)| \leq k'B$ and $|L(x)||(y-x)|^\beta \leq k'^\beta B$ when $x \in A_\sigma$ and $y \in D_x$. Because $\int_{D_x^c}\psi_\sigma(y-x)|y-x|^\alpha dy = \sigma^H$ for any $\alpha \geq 0$, when k' in the definition of D_x is sufficiently large (see Lemma 10 in Appendix A), (51), (53) and (54) imply that for constants $k_1 = M(k'B)^{2-\beta}$ and $k_2 = 1 + Mk'^\beta B$,

$$\begin{aligned} (K_\sigma f)(x) &\leq f(x)\int_{\mathbb{R}}\psi_\sigma(y-x)\{1 + k_1|l_1(x)|^\beta|y-x|^\beta + k_2|L(x)||y-x|^\beta\}dy \\ &\quad + (\|f\|_\infty + 1 + k_1|l_1(x)|^\beta + k_2|L(x)|)O(\sigma^H), \end{aligned} \tag{55}$$

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 $e^{B_{f,r}(x,y)} \geq \sum_{m=0}^r \frac{1}{m!}B_{f,r}^m(x,y) - M|B_{f,r}|^{r+1}(x,y)$ 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 $y \in D_x$, $e^B \leq 1 + B + \frac{1}{2}B^2 + \frac{1}{6}B^3 + MB^4$ 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

$$K_\sigma f(x) = f(x)\left(1 + \frac{\nu_2}{2}(l_1^2(x) + l_2(x))\sigma^2 + O(R(x)\sigma^\beta)\right) + O\left((1 + R(x))\sigma^H\right). \tag{56}$$

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 σ^β by taking out a factor $|y-x|^\beta$ and matching the remaining factor $|y-x|^{k-\beta}$ by a certain power of the $|l_j|$'s or $|L|$.

The proof of (17) for f_1 can now be completed by the observation that (56) depends on the kernel ψ only through the values of ν_α. In fact it holds for any symmetric kernel such that $\int \psi(x)|x|^\alpha dx = \nu_\alpha < \infty$ and $\int_{|x|>k'|\log\sigma|^{1/p}} \psi(x)|x|^\alpha dx = \sigma^H$ when k' 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 ν_2 by $\nu_{\psi*\psi,2} = \int (\psi*\psi)(x)|x|^\alpha dx$. As $f_1 = 2f - K_\sigma f$ and $\nu_{\psi*\psi,2} = 2\nu_2$, this proves (17) for $k=1$.

The same arguments can be used when $k>1$ and $\beta \in (2k, 2k+2]$; in that case all terms with $\sigma^2, \sigma^4, \ldots, \sigma^{2k}$ cancel out. This can be shown by expressing the moments $\nu_{m,2}, \ldots, \nu_{m,2k}$ of the kernels K_σ^m, $m = 2, \ldots, k+1$ in terms of $\nu_2, \ldots, \nu_{2k}$ 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

$$A_{\sigma,\delta} = \{x : |l_j(x)| \le \delta B \sigma^{-j} |\log\sigma|^{-j/p}, j = 1, \ldots, r, |L(x)| \le \delta B \sigma^{-\beta} |\log\sigma|^{-\beta/p}\},$$

indexed by $\delta \le 1$. 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

$$\int_{A_\sigma^c} (K_\sigma^0 f)(x) dx \le \sum_{j=1}^{\beta} P\left(|l_j(X)|^{\frac{2\beta+\epsilon}{j}} \ge (\delta B)^{\frac{2\beta+\epsilon}{j}} \sigma^{-2\beta-\epsilon} |\log\sigma|^{-\frac{2\beta+\epsilon}{p}} \right) = O(\sigma^{2\beta}),$$

provided that $\sigma^{-\epsilon}|\log\sigma|^{-\frac{2\beta+\epsilon}{p}} > 1$, which is the case if σ is sufficiently small.

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

$$\begin{aligned} P(X+\sigma U \in A_\sigma^c) &\le P(X+\sigma U \in A_\sigma^c, |\, U \,| \le k'|\log\sigma|^{1/p}) + P(|U| \ge k'|\log\sigma|^{1/p}) \\ &= O(\sigma^{2\beta}) + P(X+\sigma U \in A_\sigma^c, X \in A_{\sigma,\delta}, |\, U \,| \le k'|\log\sigma|^{1/p}) \\ &\quad + P(X+\sigma U \in A_\sigma^c, X \in A_{\sigma,\delta}^c, |\, U \,| \le k'|\log\sigma|^{1/p}) \end{aligned} \tag{57}$$

The last term is bounded by $P(X \in A_{\sigma,\delta}^c)$, which is $O(\sigma^{2\beta})$ for any $0 < \delta \le 1$. We show that the last term on the second line is zero for sufficiently small δ. 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 $|\, U \,| \le k'|\log\sigma|^{1/p}$.

To see this, note that since $X \in A_{\sigma,\delta}$, $|L(X)| \le \delta B\sigma^{-\beta}|\log\sigma|^{-\beta/p}$ and $|l_j(X)| \le \delta B \sigma^{-j}|\log\sigma|^{-j/p}$ for $j = 1, \ldots, r$. On the other hand, $X + \sigma U \in$

$A^c_{\sigma,1}$ implies that $|L(X+\sigma U)| \geq B\sigma^{-\beta}|\log\sigma|^{-\beta/p}$ or that $|l_i(X+\sigma U)| \geq \delta B\sigma^{-i}|\log\sigma|^{-i/p}$ for some $i \in \{1,\ldots,r\}$. From (6) it follows that for all $i = 1,\ldots,r$

$$|l_i(X+\sigma U)| \leq \left|\sum_{j=0}^{r-i}\frac{l_{i+j}(X)}{j!}(\sigma U)^j + \frac{r!}{(r-i)!}|L(X)||\sigma U|^{\beta-i}\right| \leq B\sigma^{-i}|\log\sigma|^{-i/p}$$

if δ 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^c_σ. Hence it suffices to show that $|L(X)| \leq \delta B\sigma^{-\beta}|\log\sigma|^{-\beta/p}$ and $|\, U\, |\leq k'|\log\sigma|^{1/p}$ 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 $\eta = \max|z_i|$, z_i being the roots of L. First, suppose that $|X| > \eta+1$. Then

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

This implies

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

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

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 $P(X \in A^c_{\sigma,\delta})$, which is $O(\sigma^{2\beta})$ by the result for $m=1$. This step can be repeated arbitrarily often, for some decreasing sequence of δ'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

$$\int_{E^c_\sigma} f(x)dx \leq \sigma^{H_1/2}\int_{E^c_\sigma}\sqrt{f(x)}dx = O(\sigma^{2\beta}), \tag{58}$$

which is $O(\sigma^{2\beta})$ when $H_1 \geq 4\beta$.

For $m=1$, we integrate over the sets $E^c_\sigma \cap A^c_\sigma$ and $E^c_\sigma \cap A_\sigma$. The integral over the first set is $O(\sigma^{2\beta})$ by the preceding paragraph. To bound the second integral, consider the sets

$$E_{\sigma,\delta} = \{x : \log f(x) \geq \delta H_1 \log\sigma\}, \tag{59}$$

indexed by $\delta \leq 1$. We can use the inequality (57) with A_σ^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 $P_{X\sim f}(X \in E_{\sigma,\delta}^c)$ can be shown to be $O(\sigma^{2\beta})$ as in (58), provided that $\delta H_1/2 \geq 2\beta$. The probability that $|U| \leq k'|\log\sigma|^{1/p}$, $X + \sigma U \in E_\sigma^c \cap A_\sigma$ and $X \in E_{\sigma,\delta} \cap A_\sigma$ is zero: due to the construction of A_σ 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 $P_{X\sim f}(X \in E_{\sigma,\delta}^c)$ remain $O(\sigma^{2\beta})$, 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 $2k < \beta \leq 2k+2$, $l = \log f$ being β-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

$$(K_\sigma f)(x) = f(x)\left(1 + O(R^{(2)}(x)\sigma^2)\right) + O\left((1 + R^{(2)}(x))\sigma^H\right), \qquad (60)$$

where $L^{(2)} = l_2$ and $R^{(2)}$ is a linear combination of l_1^2 and $|L^{(2)}|$. The key observation is that $R^{(2)} = o(1)$ uniformly on A_σ when $\sigma \to 0$. Combining (60) with the lower bound for f on E_σ, can find a constant ρ close to 1 such that

$$f_1(x) = 2f(x) - K_\sigma f(x) = 2f(x) - (1 + O(R^{(2)}(x))\sigma^2)f(x) - O(1 + R^{(2)}(x))\sigma^H > \rho f(x)$$

for small enough σ. Similarly, when l is treated as being Hölder smooth of degree 4, we find that

$$f_2(x) = 2f_1(x) - K_\sigma f_1(x) = 2f_1(x) - (1 + O(R^{(4)}(x))\sigma^4)f(x) - O(1 + R^{(4)}(x))\sigma^H > \rho^2 f(x).$$

Continuing in this manner, we find a constant ρ_k such that $f_k(x) > \rho_k f(x)$ for $x \in A_\sigma \cap E_\sigma$ and σ sufficiently small. If initially ρ 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_σ and m the finite mixture to be constructed, we write

$$\begin{aligned}\int f\left(\log\frac{f}{m}\right)^2 = \int_{E_\sigma} f\left(\log\frac{f}{K_\sigma h_k} + f\log\frac{K_\sigma h_k}{K_\sigma \tilde{h}_k} + f\log\frac{K_\sigma \tilde{h}_k}{m}\right)^2 \\ + \int_{E_\sigma^c} f\left(\log\frac{f}{K_\sigma h_k} + \log\frac{K_\sigma h_k}{m}\right)^2.\end{aligned} \qquad (61)$$

The integral of $f(\log(f/K_\sigma h_k))^2$ over E_σ is $O(\sigma^{2\beta})$ by Theorem 1. To show that the integral of $f(\log(K_\sigma h_k/K_\sigma \tilde{h}_k))^2$ over E_σ is $O(\sigma^{2\beta})$ 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 $\int_{E_\sigma^c} h_k = O(\sigma^{2\beta})$. Consequently, $K_\sigma h_k/K_\sigma \tilde{h}_k) = O(\sigma^{2\beta})$ and the required bound for $\int_{E_\sigma} f(\log(K_\sigma h_k/K_\sigma \tilde{h}_k))^2$ follows. To bound the integral of $f(\log K_\sigma \tilde{h}_k/m)^2$ over E_σ, let $m = m(\cdot; k_\sigma, \mu_\sigma, w_\sigma, \sigma)$ be the finite mixture obtained from Lemmas 12 and 13, with $\epsilon = \sigma^{\delta' H_1+1}$ and $\delta' \geq 1+2\beta/H_1$. 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_σ in Lemma 13 is $O(\sigma^{-1}|\log\sigma|^{1+p^{-1}})$. We have

$$\int_{E_\sigma} f\left(\log\frac{K_\sigma \tilde{h}_k}{m}\right)^2 \leq \int_{E_\sigma} f\left(\frac{m - K_\sigma \tilde{h}_k}{\sigma^{H_1} - \sigma^{\delta' H_1}}\right)^2 \leq \sigma^{2(\delta'-1)H_1} = O(\sigma^{2\beta}),$$

provided that $\delta' \geq 1 + \frac{\beta}{H_1}$. The cross-products resulting from the square in the integral over E_σ can be shown to be $O(\sigma^{2\beta})$ using the Cauchy-Schwartz inequality and the preceding bounds.

To bound the integral over E_σ^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

$$\begin{aligned} \int_{E_\sigma^c} f(x)\left(\log\frac{K_\sigma h_k}{m}(x)\right)^2 dx &\leq \int_{E_\sigma^c} f(x)\left(\log\frac{C}{\sigma^{2\beta}\psi_\sigma(x)}\right)^2 dx \\ &= \int_{E_\sigma^c} f(x)\left(\log(C_p^{-1}C) + 2\beta|\log\sigma| + \frac{|x|^p}{\sigma^p}\right)^2 dx. \end{aligned} \tag{62}$$

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

$$\int_{E_\sigma^c} f(x)|x|^{2p} dx \leq \sigma^{H_1/2}\int_{E_\sigma^c} \sqrt{f(x)}|x|^{2p} dx = O(\sigma^{2\beta+2p}),$$

which is the case if $H_1 \geq 4(\beta+p)$. The integral of $f(\log f/K_\sigma h_k)^2$ over E_σ^c is $O(\sigma^{2\beta})$ by Lemma 1, and the integral of $f(\log f/K_\sigma h_k)(\log K_\sigma h_k/m)$ over E_σ^c can be bounded using Cauchy-Schwartz.

If $m' = m(\cdot; k_\sigma, \mu, w, \sigma')$ is a different mixture with $\sigma' \in [\sigma, \sigma + \sigma^{\delta' H_1+2}]$, $\mu \in B_{k_\sigma}(\mu_\sigma, \sigma^{\delta' H_1+2})$ and $w \in \Delta_{k_\sigma}(w_\sigma, \sigma^{\delta' H_1+1})$, the L_∞-norm between m and m' is $\sigma^{\delta' H_1}$ by Lemma 3, and $\int_{E_\sigma} f\left(\log\frac{K_\sigma \tilde{h}_k}{m'}\right)^2 = O(\sigma^{2\beta})$. The integral over E_σ^c can be shown to be $O(\sigma^{2\beta})$ as in (62), where the $|x-\sigma^{2\beta}|^{2p}$ that comes in the place of $|x|^{2p}$ 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 $\sigma \in [\underline{\sigma}_n, \bar{\sigma}_n]$ and $\epsilon < (1 \wedge C_p)$. Then there exists a discrete distribution F' on $[-a, a]$ with at most $N = pe^2 \log \frac{C_p}{\epsilon}$ support points such that $\|F * \psi_\sigma - F' * \psi_\sigma\|_\infty \lesssim \epsilon$.*

Lemma 13. *Given $\sigma \in [\underline{\sigma}_n, \bar{\sigma}_n]$ and $F \in \mathcal{M}[-a, a]$, let F' be the discrete distribution from the previous lemma. Then $\|F * \psi_\sigma - F' * \psi_\sigma\|_1 \lesssim \epsilon\, \psi^{-1}(\epsilon)$. Moreover, for any $\sigma > 0$ there exists a discrete F' with a multiple of $(a\sigma^{-1} \vee 1) \log \epsilon^{-1}$ support points, for which $\|F * \psi_\sigma - F' * \psi_\sigma\|_1 \lesssim \epsilon\psi^{-1}(\epsilon)$ and $\|F * \psi_\sigma - F' * \psi_\sigma\|_\infty \lesssim \frac{\epsilon}{\sigma}$.*

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):363–375, 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 165–175, 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 P&S*.

[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.*, 2:298–331, 2008.

[32] Tong Zhang. From ϵ-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.