HAL open science

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. 2009. hal-00389343v1

HAL Id: hal-00389343

https://hal.science/hal-00389343v1

Preprint submitted on 28 May 2009 (v1), last revised 1 Jul 2010 (v2)

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.

Submitted to the Annals of Statistics

ADAPTIVE BAYESIAN DENSITY ESTIMATION WITH LOCATION-SCALE MIXTURES

By Willem Kruijer*, Judith Rousseau and Aad van der Vaart

Université Paris Dauphine and VU University Amsterdam

We study convergence rates of Bayesian density estimators based on finite location-scale mixtures of a kernel $C_p \exp\{-|x|^p\}$. 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 supremum-norm and 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 $\log f_0$. Examples of priors which satisfy these conditions include Dirichlet and Polya-tree priors for the weights, and Poisson processes for the locations.

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 [12] and Geman and Hwang [4]. Although originally introduced in a maximum likelihood context, there has been a large number of Bayesian papers in the following years. See, among many others, [17], [2], and [3]. Whereas much progress has been made regarding the computational problems in nonparametric Bayesian inference (see for example the review by Marin et al.[16]), results on convergence rates were found only recently, especially for the case when the underlying distribution is not a mixture itself.

For the estimation of a C^2-density using continuous normal mixtures with a Dirichlet prior on the mixing distribution, [10] found optimal rates under certain conditions on the prior. Convergence rates of normal mixtures have also been studied by Scricciolo [19] and Genovese and Wasserman [5]. For beta-mixtures with rational parameters (i.e. corresponding to Bernstein polynomials), rates can be found in [7] and [14]. In these papers optimal

*Much of this work was carried out when the first author was a PhD-student at the VU University Amsterdam.

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

rates were only obtained when the smoothness β of the underlying density is at most 2 (normal mixtures) or 1 (beta-mixtures). For continuous beta-mixtures Rousseau [18] recently obtained a (near)-optimal [1] rate for any β, the posterior being rate-adaptive with respect to β. In the present work a similar result is derived for location-scale mixtures. This is achieved using the approximations studied in section 2. These are established following the approach of [18], which consists of finding a continuous mixing density f_k for which the mixture is close to f_0 and next discretizing f_k. If f_k is allowed to be different from f_0, the approximation can also be optimal when the smoothness is larger than 2. A difference with [18] is that we can exploit properties of the convolution operator (see Lemma 1), which is not possible for beta-mixtures. A remarkable difference with the frequentist approach is that the underlying density f_0 can be approximated by convoluting a function different from f_0, instead of, for example, convoluting f_0 itself with a kernel that is not strictly nonnegative.

Whereas much of the (asymptotical) literature on nonparametric mixtures is restricted to normal mixtures with Dirichlet process priors, we give general conditions for the priors on μ and w, under which the posterior is rate-optimal and adaptive. Conditions for consistency were recently given by Wu and Ghosal [20]. In section 4 we give examples of priors that satisfy these conditions. It appears that optimal rates are achieved in many other models than the Dirichlet mixtures of normals. For example, one can equally well use Laplace mixtures with a Poisson process prior on the locations and a Polya-tree prior on the weights.

We obtain posterior convergence rates for 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

$$\psi(x) = C_p e^{-|x|^p}, \tag{2}$$

for a normalizing constant C_p. The inverse $\psi^{-1}(y) = \left(\log \frac{C_p}{y}\right)^{1/p}$ is defined on $(0, C_p]$.

Notation For any nonnegative α, let

$$\nu_\alpha = \int x^\alpha \psi(x) dx. \tag{3}$$

[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 $(K_\sigma h) - h$. 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 $\sup_{x \neq y} |h^{(r)}(x) - h^{(r)}(y)|/|y - x|^{\beta - r} \leq L$. Let H_β be the Hölder-space $\cup_{L>0}\mathcal{H}(\beta, L)$, and given some function $h \in H_\beta$, let $L_{h,\beta-r} = \sup_{x \neq y} |h^{(r)}(x) - h^{(r)}(y)|/|y - x|^{\beta - r}$. When $\beta - r = 1$, this equals $\|h^{(r+1)}\|_\infty$. Let $\Delta_k = \{x \in \mathbb{R}^k : x_i \geq 0, \sum_{i=1}^k x_i = 1\}$ denote the $(k-1)$-dimensional unit-simplex and $S_k = \{x \in \mathbb{R}^k : x_i \geq 0, \sum_{i=1}^k x_i \leq 1\}$. For $b, d \in \mathbb{R}^k$, $H_k[b, d]$ denotes the hypercube $\{x \in \mathbb{R}^k \mid x_i \in [b_i, d_i]\}$. 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, $T_{c,\epsilon}$ denotes the interval $[-c|\log \epsilon|^{1/p}, c|\log \epsilon|^{1/p}]$. Given $\epsilon > 0$ and fixed points $x \in \mathbb{R}^k$ and $y \in \Delta_k$, define the l_1-balls $B_k(x, \epsilon) = \{z \in \mathbb{R}^k \mid \sum_{i=1}^k \mid z_i - x_i \mid \leq \epsilon\}$ and $\Delta_k(y, \epsilon) = \{z \in \Delta_k \mid \sum_{i=1}^k \mid z_i - y_i \mid \leq \epsilon\}$. 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 $N(I)$. Finally, the convergence rate is a sequence tending to zero such that $n\epsilon_n^2 \to \infty$ and $\Pi(d(f_0, f) > M\epsilon_n \mid X_1 \ldots, X_n) \to 0$ in F_0^n-probability, for some sufficiently large constant M, d being the Hellinger- or L_1-metric.

Conditions on f_0. The observations $X_1, \ldots, X_n$ are assumed to be an i.i.d. sample from a density f_0, satisfying the following conditions.

1. Smoothness. $\log f_0$ is assumed to be locally β-Hölder, with derivatives $l_j(x) = \frac{d^j}{dx^j} \log f(x)$. In addition we assume the existence of a polynomial L and a constant $\gamma > 0$ such that

$$\log f_0(y) \leq \log f_0(x) + \sum_{j=1}^r \frac{l_j(x)}{j!}(y - x)^j + L(x)|y - x|^\beta \tag{4}$$

$$\log f_0(y) \geq \log f_0(x) + \sum_{j=1}^r \frac{l_j(x)}{j!}(y - x)^j - L(x)|y - x|^\beta \tag{5}$$

   for all x, y with $|y - x| \leq \gamma$. Hence $\log f_0$ is locally Hölder with an additional uniformity condition. As L is polynomial, the l_j's are bounded by polynomials. In the remainder, (4) and (5) are always used when $|x - y|$ is smaller than a multiple of $\sigma|\log \sigma|^{1/p}$, and throughout the paper it is assumed that σ is small enough to have $|x - y| < \gamma$.
2. Moments. The functions L and the l_j satisfy

$$F_0|l_j|^{2\beta/j} < \infty, j = 1, \ldots, r, \qquad F_0 L^2 < \infty. \tag{6}$$

3. Monotonicity. f_0 is strictly positive, and there are numbers $x_m < x_M$ 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 [6].

4. Tails. f_0 has smaller tails than the kernel, i.e. there are constants T and M_f such that $f_0(x) \leq M_f \psi(x)$ when $|x| \geq T$. Combined with the monotonicity condition, this implies that there exists a finite constant c_f such that for all sufficiently small ϵ,

$$\{x : f_0(x) \geq \epsilon\} \subset [-c_f |\log \epsilon|^{1/p}, c_f |\log \epsilon|^{1/p}]. \tag{7}$$

The constant c_f depends on f by the constant M_f in the tail condition. This property is used in the proof of Lemma 6.

Prior (Π_n) 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 priors on the number of components, locations and weights satisfy the conditions (8)-(11) below, where d_1, d_2, d_3 and d_4 may be arbitrary positive constants.

First, the marginal distribution ρ of the number of components K is such that for all integers m

$$\sum_{k=m}^{\infty} \rho(k) \lesssim e^{-d_1 m (\log m)^r}, \tag{8}$$

where the constant r affects the logarithmic factor in the convergence rate in Theorem 1 if it is smaller than one.

Second, the joint distribution of (K, μ) satisfies

$$\Pi(N([-y, y]^c) > 0) \lesssim e^{-|y|^{d_2}}, \tag{9}$$

and for all $c > 0$ there exists a constant $d_3 > 0$ such that

$$\Pi(K = k, \mu \in B_k(\mu_0, \epsilon)) \gtrsim \exp\Big\{-d_3 k \log \frac{1}{\epsilon}\Big\} \tag{10}$$

when $y > 0$, $k \in \mathbb{N}$, $\epsilon < \frac{1}{k}$, and the elements of μ_0 are contained in $T_{c,\epsilon} = [-c|\log \epsilon|^{1/p}, c|\log \epsilon|^{1/p}]$. Condition (9) requires that the number of points (N) outside $[-y, y]$ is exponentially small.

Finally, the prior distribution of the weight vector w is independent of μ, such that for all k, $\epsilon < \frac{1}{k}$, and $w_0 \in \Delta_k$,

$$\Pi(w \in \Delta_k(w_0, \epsilon) \mid K = k) \gtrsim \exp\Big\{-d_4 k (\log k)^b \log \frac{1}{\epsilon}\Big\}, \tag{11}$$

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

We can now state our main result.

THEOREM 1. *If the prior satisfies conditions* (8)-(11) *and* f_0 *satisfies Conditions 1-4, then* $\Pi_n(\cdot \mid X_1, \ldots, X_n)$ *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* (8) *and* (11)*, and* $t > (2 + b + p^{-1})/(2 + p^{-1}) + \max(0, (1 - r)/2)$.

The proof is based on Theorem 5 of Ghosal and van der Vaart [10], which is included here in appendix A. Informally speaking, the conditions (10) and (11) translate the general prior mass condition (51) in Theorem 2 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; therefore (10) and (11) could be seen uniformity conditions. Considering only proper priors, this requires a restriction of the form $\mu_0 \in T^k_{c,\epsilon}$. Due to conditions (8) and (10), the marginal distribution of the number of components has to be exponential in k, as in Ghosal [7], who studied convergence rate of Bernstein-polynomials.

2. Approximation of smooth densities. In the proof of Theorem 1 we need a set of finite mixtures whose Kullback-Leibler- (KL) divergence with respect to f_0 is $O(\sigma^{2\beta})$. The usual approach is to bound the supremum-norm between f_0 and $K_\sigma f_0$, and then show that under certain conditions on f_0, this also gives a small KL-divergence. But as $\|f_0 - K_\sigma f_0\|_\infty$ remains of order σ^2 when $\beta > 2$, this approach becomes rather difficult. In this section we propose an alternative mixing distribution, following the construction of Rousseau [18] for beta-mixtures. In Lemma 1 we construct, by substracting the convoluted approximation error, a function f_k such that $\|f_0 - K_\sigma f_k\|_\infty = O(\sigma^\beta)$. Under extra assumptions on f_0, the error is even relative with respect to f_0 (Lemma 2). This f_k is not necessarily a density however, and after some modification and normalization we obtain a density h_k. In Lemma 3 it is shown that h_k still has the desired approximative properties. It then follows that h_k is also close to f_0 in the KL-sense (Lemma 4). After a result quantifying the L_∞- and L_1-distances between mixtures whose parameters are close, given in Lemma 5, h_k is discretized in Lemma 6. In the remainder of this section, we often write f for f_0 for notational convenience.

It is well known that for any $f \in H_\beta$,

$$\|\Delta_\sigma f\|_\infty \leq \nu_\beta L_{f,\beta}\sigma^\beta, \quad \beta \in (0,1] \tag{12}$$

$$\|\Delta_\sigma f\|_\infty \leq \nu_{\beta\wedge 2} L_{f',(\beta-1)\wedge 1}\sigma^{\beta\wedge 2}, \quad \beta > 1. \tag{13}$$

When $\beta \geq 2$, the latter inequality can be written $\|\Delta_\sigma f\|_\infty \leq \nu_2\sigma^2\|f^{(2)}\|_\infty$. The approximation error is therefore of order σ^2, even when β is larger than two. The following calculation illustrates how this can be improved if we take $f_1 = f - \Delta_\sigma f = 2f - K_\sigma f$ as mixing density instead of f itself. If $f \in H_\beta$ with $\beta \in (2,4]$, the approximation error $|(K_\sigma f_1)(x) - f(x)|$ equals

$$\left|\int \psi_\sigma(x-\mu)\left\{(f(\mu)-f(x)) - \int \psi_\sigma(\epsilon-\mu)(f(\epsilon)-f(\mu))d\epsilon\right\}d\mu\right|$$
$$= \left|\frac{\sigma^2}{2}f''(x) + O(\sigma^\beta) - \frac{\sigma^2}{2}\int \psi_\sigma(x-\mu)f''(\mu)d\mu - O(\sigma^\beta)\right| = O(\sigma^\beta).$$

If $4 < \beta \leq 6$, it can be shown that an error of $O(\sigma^\beta)$ is achieved using $K_\sigma f_2$, where $f_2 = f - \Delta_\sigma f_1$. For larger β, this procedure of substracting the convoluted error of the previous approximation can be continued; we find a sequence $f_{i+1} = f - \Delta_\sigma f_i$. It can be shown by induction that

$$f_k = \sum_{i=0}^{k}(-1)^i\binom{k+1}{i+1}K_\sigma^i f, \tag{14}$$

$$\Delta_\sigma^k f = \sum_{i=0}^{k}(-1)^{k-i}\binom{k}{i}K_\sigma^i f. \tag{15}$$

The following elementary result is included for convenience. Its proof 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 is not in H_β (e.g. when $p = 1$).

Second, it follows from the symmetry of ψ that $K_\sigma f^{(k)} = \frac{d^k}{dx^k}K_\sigma f$, i.e. the kth derivative of the convolution of f equals the convolution of $f^{(k)}$.

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

PROOF. For $k = 1$ it can be seen that $K_\sigma(f_1) - f = \Delta_\sigma(f - \Delta_\sigma(f)) - \Delta_\sigma(f) = -\Delta_\sigma\Delta_\sigma f$. From (13) it follows that $\|\Delta_\sigma\Delta_\sigma f\|_\infty \leq \nu_2\sigma^2\|(\Delta_\sigma f)''\|_\infty$. 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 (14) and (15), $\|K_\sigma f_k - f\|_\infty = \|\Delta_\sigma^{k+1} f\|_\infty$. In fact, by induction 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 $f^{(2m)} \in H_{\beta-2m}$. □

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 $f \in H_\beta$ (or an f such that $\log f \in H_\beta$), 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 one half of the largest even number strictly smaller than β.

The following approximation result, whose proof can be found in Appendix B, will be essential for controlling the KL-divergence between f and $K_\sigma f_k$. A similar result for beta-mixtures is contained in Theorem 3.1 in [18].

Lemma 2. *Let f be a density satisfying conditions 1-4, and let f_k be the first function in the sequence $\{f_i\}_{i\in\mathbb{N}}$ for which an error of order $\sigma^{2\beta}$ is achieved. Then for all sufficiently small σ and for all x contained in the set*

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

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

$$R(x) = r_{r+1}|L(x)| + \sum_{i=1}^{r} r_i |l_1(x)|^{\beta/i}, \tag{18}$$

for nonnegative constants r_i.

Hence the result of Lemma 1 can be strengthened considerably: on a set on which the l_j's are sufficiently controlled, the approximation error $(K_\sigma f_k)(x) - f(x)$ is now relative to $f(x)$, apart from a term σ^H where H can be arbitrarily large. Note that the powers of the $|l_j|$'s appearing in R are half of those appearing in the moment condition (6); this is because of the application

of (17) (after a small modification in Lemma 3) in Lemma 4, where it is required that $\int R^2 dF < \infty$.

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 are however not necessarily nonnegative, which can be easily seen by considering a compactly supported f. To obtain a probability density, we define

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

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

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

The constant $\frac{1}{2}$ in (19) and (20) is arbitrary and could be replaced by any other number between zero and one. From our conditions on f it will follow that these h_j still have the desired approximation error, i.e. we will show in Lemma 3 that for any $f \in H_\beta$ the normalizing constant $\int g_k$ is $1 + O(\sigma^\beta)$.

We define

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

where (at this point) the constant H_1 has to be at least 4β, but will have to be chosen larger in the remainder. By property (7) we have that for $\epsilon = \sigma^{H_1}$ and a certain constant c, $E_\sigma \subset T_{c,\epsilon}$.

Lemma 3. *Let f be a density satisfying conditions 1-4. Then*

$$\int_{A^c_\sigma} (K^m_\sigma f)(x) dx = O(\sigma^{2\beta}), \qquad \int_{E^c_\sigma} (K^m_\sigma f)(x) dx = O(\sigma^{2\beta}) \tag{23}$$

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

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

Consequently,

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

for all $x \in A_\alpha \cap E_\sigma$, i.e. if we make the further restriction to E_σ, (17) holds with f_k replaced by h_k.

PROOF. To show that the first integral is of order $\sigma^{2\beta}$, consider the sets

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

indexed by $\delta \le 1$. If $m = 0$, the moment condition (6) and Markov's inequality imply $F(A_\sigma^c) = O(\sigma^{2\beta})$. 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

(26)
$$\begin{aligned} P(X+\sigma U \in A_\sigma^c) &\le P(X+\sigma U \in A_\sigma^c, \mid U \mid \le k'|\log\sigma|^{1/p}) + P(|U| \ge k'|\log\sigma|^{1/p}) \\ &= O(\sigma^{2\beta}) + P(X+\sigma U \in A_\sigma^c, X \in A_{\sigma,\delta}, \mid U \mid \le k'|\log\sigma|^{1/p}) \\ &\quad + P(X+\sigma U \in A_\sigma^c, X \in A_{\sigma,\delta}^c, \mid U \mid \le k'|\log\sigma|^{1/p}) \end{aligned}$$

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 second term on the last line is zero for sufficiently small δ. We show that the conditions on f and the fact that $X \in A_{\sigma,\delta}$ and $X + \sigma U \in A_{\sigma,1}^c$ imply that $|U|$ is large, contradicting $\mid U \mid \le k'|\log\sigma|^{1/p}$.

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 $j = 1,\dots,r$, whereas $X + \sigma U \in A_{\sigma,1}^c$ implies that $|L(X+\sigma U)| \ge B\sigma^{-\beta}|\log\sigma|^{-\beta/p}$ or that $|l_i(X+\sigma U)| \ge \delta B\sigma^{-i}|\log\sigma|^{-i/p}$ for some $i \in \{1,\dots,r\}$. By (4) and (5), it follows that for all $i = 1,\dots,r$

$$\begin{aligned} |l_i(X+\sigma U)| &= \left| \sum_{j=i}^r \frac{l_j(X)}{j!}(\sigma U)^{j-i} + O(L(X)\sigma^{\beta-i}(\sigma U)^{\beta-i}) \right| \\ &\le \sum_{j=i}^r \frac{\delta B\sigma^{-j}}{j!}|\log\sigma|^{-j/p}|\sigma U|^{j-i} + \delta B\sigma^{-\beta}|\log\sigma|^{-\beta/p}|\sigma U|^{\beta-i} \\ &\le B\sigma^{-i}|\log\sigma|^{-i/p} \end{aligned}$$

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_σ^c. Hence it suffices to show that $|L(X)| \le \delta B\sigma^{-\beta}|\log\sigma|^{-\beta/p}$ and $\mid U \mid \le k'|\log\sigma|^{1/p}$ is in contradiction with $|L(X+\sigma U)| \ge B\sigma^{-\beta}|\log\sigma|^{-\beta/p}$. This follows from the assumption that L is polynomial.

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 (26) is then bounded by $P(X \in A_{\sigma,\delta}^c)$, 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), first note that for $m = 0$

$$\int_{E_\sigma^c} f(x)dx \leq \sigma^{H_1/2} \int_{E_\sigma^c} \sqrt{f(x)}dx = O(\sigma^{2\beta}). \tag{27}$$

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 $O(\sigma^{2\beta})$ by the preceding paragraph. To bound the second integral, consider the sets

$$E_{\sigma,\delta} = \{x : \log f(x) \geq \delta H_1 \log \sigma\}, \tag{28}$$

indexed by $\delta \leq 1$. We can use the inequality (26) 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 (27), 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 $\mid l(X+\sigma U)-l(X) \mid = O(1)$, whereas $\mid l(X+\sigma U)-l(X) \mid \geq (1-\delta)H_1 \mid \log\sigma \mid$. 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 that $A_\sigma \cap E_\sigma \subset R_{\sigma,k}$ we use that $f(x) > \sigma^{H_1} > \sigma^H$ and $f(x)(1+O(R(x)\sigma^\beta)) + O\left((1+R(x))\sigma^H\right)$ if $x \in A_\sigma \cap E_\sigma$. Because R is bounded by a polynomial and $E_\sigma \subset T_{c,\epsilon}$ for some c and $\epsilon = \sigma^{H_1}$, it follows that for any $\rho_1 < 1$,

$$f_1(x) = 2f(x) - K_\sigma f(x) = 2f(x) - (1+O(R(x))\sigma^\beta)f(x) - O(1+R(x))\sigma^H > \rho_1 f(x)$$

for small enough σ. Similarly,

$$f_2(x) = 2f_1(x) - K_\sigma f_1(x) = 2f_1(x) - (1+O(R(x))\sigma^\beta)f(x) - O(1+R(x))\sigma^H > \rho_2 f(x),$$

where ρ_2 can be arbitrarily close to one if we choose ρ_1 close enough to one. 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 $\rho_1 > \rho_2 > \ldots > \rho_{k-1}$ are chosen large enough, ρ_k can be at least one half; hence $A_\sigma \cap E_\sigma \subset R_{\sigma,k}$. To see that (23) now implies (24), note that the integrand $\frac{1}{2}f - f_k$ is a linear combination of $K_\sigma^m f$, $m = 0, \ldots, k$. □

Remark 1. *As a further consequence we have $h_k \geq f/(2(1+O(\sigma^\beta)))$, and the fact that $K_\sigma f$ is lower bounded by a multiple of f implies that the same is true for $K_\sigma h_k$.*

REMARK 2. *Clearly, the $O(\sigma^{2\beta})$ in (27) can be improved by choosing a larger H_1. This will be used in the proof of Lemma 6.*

LEMMA 4. *Given $\beta > 0$, let $f \in H_\beta$ be a density that satisfies the Conditions 1-4 above, and let h_k be defined by (21). Then for all small enough σ,*

$$\int f \log \frac{f}{K_\sigma h_k} = O(\sigma^{2\beta}), \quad \int f \left(\log \frac{f}{K_\sigma h_k}\right)^2 = O(\sigma^{2\beta}). \tag{29}$$

PROOF. Since

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

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

$$\begin{aligned} \int f(x) \log \frac{f(x)}{K_\sigma h_k(x)} dx &\le \int_{A_\sigma \cap E_\sigma} \frac{(f(x) - K_\sigma h_k(x))^2}{K_\sigma h_k(x)} dx \\ &+ \int_{A_\sigma^c \cup E_\sigma^c} f(x) \log \frac{f(x)}{K_\sigma h_k(x)} dx + \int_{A_\sigma^c \cup E_\sigma^c} (K_\sigma h_k(x) - f(x)) dx. \end{aligned} \tag{30}$$

The first integral on the right can be bounded by application of (25) and Remark 1 following Lemma 3. On $A_\sigma \cap E_\sigma$ the integrand is bounded by $f(x)O(\sigma^\beta R(x)) - 2O(\sigma^{\beta+H} R(x)) + O((1+R(x))^2)\sigma^{2H}/f(x)$. If we choose $H \ge H_1 + \beta$, it follows from the definition of $R(x)$ and the moment condition (6) 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 $f(x) \ge \sigma^{H_1}$ on E_σ and the Lebesgue measure of this interval is at most σ^{-H_1}. To bound the second integral in (30) 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 (29) 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 result of the preceding lemma is discretized in Lemma 6 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 [10]. The inequality for the L_1-norm will be used in the entropy calculation in the proof of Theorem 1.

LEMMA 5. *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,

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

for any norm. We have the following inequalities:

$$\begin{aligned}
\|\psi_\sigma(z - \mu_i) - \psi_\sigma(z - \tilde{\mu}_i)\|_1 &= 2\left|\Psi\left(\frac{\mu_i - \tilde{\mu}_i}{2\sigma}\right) - \Psi\left(\frac{\tilde{\mu}_i - \mu_i}{2\sigma}\right)\right| \\
&\leq 2\|\psi\|_\infty \frac{\mid \tilde{\mu}_i - \mu_i \mid}{\sigma} \leq \frac{2\|\psi\|_\infty}{\sigma \wedge \tilde{\sigma}} \mid \tilde{\mu}_i - \mu_i \mid, \\
\|\psi_\sigma - \psi_{\tilde{\sigma}}\|_1 &\leq \frac{1}{\sigma \wedge \tilde{\sigma}} \int \mid \psi(\frac{x}{\sigma}) - \psi(\frac{x}{\tilde{\sigma}}) \mid dx \leq \frac{1}{\sigma \wedge \tilde{\sigma}} \mid \sigma - \tilde{\sigma} \mid, \\
\|\psi_\sigma - \psi_{\tilde{\sigma}}\|_\infty &\leq \frac{1}{(\sigma \wedge \tilde{\sigma})^2} \|\frac{d}{dz} g_x\|_\infty \mid \sigma - \tilde{\sigma} \mid, \qquad (31) \\
\|\psi_\sigma(z - \mu_i) - \psi_\sigma(z - \tilde{\mu}_i)\|_\infty &\lesssim \frac{1}{(\sigma \wedge \tilde{\sigma})^2} \mid \tilde{\mu}_i - \mu_i \mid .
\end{aligned}$$

To prove (31), 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 4 still holds. The discretization relies on Lemma 3.13 in [13], which is included in Appendix C.

LEMMA 6. *Let the constant H_1 in the definition of E_σ be at least $4(\beta+p)$. Given $\beta > 0$, let $f \in H_\beta$ be a density that satisfies the Conditions 1-4 above. 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*

$$\int f \log \frac{f}{m} = O(\sigma^{2\beta}), \quad \int f \left(\log \frac{f}{m}\right)^2 = O(\sigma^{2\beta}). \tag{32}$$

Furthermore, (32) 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 $\delta' \geq 1 + \beta/H_1$.

PROOF. We bound the second integral in (32); 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

$$\begin{aligned}\int f \left(\log \frac{f}{m}\right)^2 &= \int_{E_\sigma} f \left(\log \frac{f}{K_\sigma h_k} + \int f \log \frac{K_\sigma h_k}{K_\sigma \tilde{h}_k} + \int f \log \frac{K_\sigma \tilde{h}_k}{m}\right)^2 \\ &\quad + \int_{E_\sigma^c} f \left(\log \frac{f}{K_\sigma h_k} + \log \frac{K_\sigma h_k}{m}\right)^2.\end{aligned} \tag{33}$$

The integral of $f(\log f/K_\sigma h_k)^2$ over E_σ is $O(\sigma^{2\beta})$ by Lemma 4. By Lemma 3 and Remark 2 following the proof of this lemma, 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. 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 Lemma 14, with $\epsilon = \sigma^{\delta' H_1+1}$ and $\delta' \geq 1 + 2\beta/H_1$. The requirement that $a \lesssim \psi^{-1}(\epsilon)$ is satisfied by the monotonicity and tail conditions on f (see (7)). The number of components k_σ 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 5 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

(34)
$$\int_{E_\sigma^c} f(x)\left(\log \frac{K_\sigma h_k}{m}(x)\right)^2 dx \leq \int_{E_\sigma^c} f(x)\left(\log \frac{C}{\sigma^{2\beta}\psi_\sigma(x)}\right)^2 dx$$
$$= \int_{E_\sigma^c} f(x)\left(\log(C_p^{-1}C) + 2\beta|\log\sigma| + \frac{|x|^p}{\sigma^p}\right)^2 dx.$$

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 4, 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 5, 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 (34), where the $|x - \sigma^{2\beta}|^{2p}$ that comes in the place of $|x|^{2p}$ can be handled with Jensen's inequality.

□

3. The proof of Theorem 1. For the entropy calculations we need the following lemma.

LEMMA 7. *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*

$$D(\epsilon, \Delta_k, l_1) \leq \left(\frac{5}{\epsilon}\right)^{k-1}, \tag{35}$$

$$D(\epsilon, H_k[b, d], l_1) \leq \frac{k!\prod_{i=1}^k (d_i - b_i + 2\epsilon)}{(2\epsilon)^k}. \tag{36}$$

PROOF. A proof of (35) can be found in [8]; 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, $\mid z_i - y_i \mid \leq \frac{\epsilon}{2}$ for all i. Because for each coordinate we have the bounds $\mid z_i \mid \leq \mid y_i \mid + \mid z_i - y_i \mid \leq d_i + \frac{\epsilon}{2}$ and $\mid z_i \mid \geq b_i - \frac{\epsilon}{2}$, 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 1. The proof is an application of Theorem 2, with rate $\widetilde{\epsilon}_n = n^{-\beta/(1+2\beta)}(\log n)^{t_1}$ 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 $k_n = k_0 n^{1/(1+2\beta)}(\log n)^{1+(1-t_1)/p}$ be the number of components in Lemma 6 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

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

their locations being contained in the set E_σ 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

(37)
$$\begin{aligned}\Pi(\sigma \in [\sigma_n, \sigma_n + {\sigma_n}^b]) &= \int_{\sigma_n}^{\sigma_n + {\sigma_n}^b} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\lambda/x} dx \\ &\geq \int_{\sigma_n}^{\sigma_n + {\sigma_n}^b} \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-2\lambda/x} dx \geq 4 \frac{\lambda^{\alpha+1}}{\Gamma(\alpha)} {\sigma_n}^{b-2} e^{-\lambda {\sigma_n}^{-1}},\end{aligned}$$

which is larger than $\exp\{-n\widetilde{\epsilon}_n^2\}$ for any choice of $t_1 \geq 0$. From the conditions (10) and (11) on the priors for μ and w it then follows that

$$\Pi(KL(f_0, \epsilon_n)) \gtrsim \exp\left\{-d_3 k_n \log {\sigma_n}^{-(\delta' H_1 + 2)} - d_4 k_n (\log k_n)^b \log {\sigma_n}^{-(\delta' H_1 + 1)}\right\}.$$

The exponent is $O(n^{1/(1+2\beta)}(\log n)^{2+b+(1-t_1)/p})$; therefore $\Pi(KL(f_0, \widetilde{\epsilon}_n)) \geq \exp\{-n\widetilde{\epsilon}_n^2\} = \exp\{-n^{1/(1+2\beta)}(\log n)^{2t_1}\}$ if $t_1 > (2 + b + p^{-1})/(2 + p^{-1})$.

We then have to find sets $\mathcal{F}_n$ such that (50) and (52) 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 $b_n^{d_2} > n^{1/(1+2\beta)}$, with d_2 as in (9), we define

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

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 and the prior mass on mixtures with at least one support point outside $[-b_n, b_n]$ is controlled by conditions (8) and (9). Combining these bounds, we find

$$\Pi(\mathcal{F}_n^c) \leq \Pi(S_n^c) + \sum_{k=r_n}^\infty \rho(k) + \Pi(N([-b_n, b_n]^c > 0)) \lesssim e^{-d_1 r_n (\log n)^r}.$$

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

To control the sum in (50), we partition $\mathcal{F}_n$ using

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

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

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

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

(38)

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

As the L_1-distance is bounded by the Hellinger-distance, condition (50) only needs to be verified for the L_1-distance. We further decompose the $\mathcal{F}_{n,j}$'s and write

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

It will be convenient to replace the covering numbers N in (50) 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 $D(\epsilon, A \times B, d_1 + d_2) \leq D(\frac{\epsilon}{2}, A, d_1) D(\frac{\epsilon}{2}, B, d_2)$, Lemma 5 implies that for all $k \geq 1$, $D(\bar{\epsilon}_n, \mathcal{F}_{n,j,k}, \| \cdot \|_1)$ 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 7 provides the following bounds:

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

For some constant C, we find that

$$\begin{aligned}
D(\bar{\epsilon}_n, \mathcal{F}_{n,j}, \| \cdot \|_1) &\leq r_n D(\bar{\epsilon}_n, \mathcal{F}_{n,j,r_n}, \| \cdot \|_1) \\
&\lesssim r_n C^{r_n} r_n! (\bar{\epsilon}_n)^{-2r_n} s_{n,j} s_{n,j-1}^{-r_n+1} (\max(b_n, \bar{\epsilon}_n s_{n,j-1}))^{r_n}.
\end{aligned} \tag{39}$$

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 (38) is bounded by $-\lambda b_n^{-1} \bar{\epsilon}_n/(1+\widetilde{\epsilon}_n)$. A combination of (38), (39) and Stirling's bound on $r_n^{r_n}$ then imply that $\sqrt{\Pi_n(\mathcal{F}_{n,j})}\sqrt{N(\bar{\epsilon}_n, \mathcal{F}_{n,j}, d)}$ 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 $\sqrt{\Pi_n(\mathcal{F}_{n,j})}\sqrt{N(\bar{\epsilon}_n, \mathcal{F}_{n,j}, d)}$ 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 $t > (2 + b + p^{-1})/(2 + p^{-1}) + \max(0, (1-r)/2)$. □

4. Examples of priors.

4.1. *Priors for the locations.* We show that conditions (8)-(10) hold for two important types of priors for (k, μ).

First we consider hierarchical priors, where K is sampled from a prior $\rho(\cdot)$ on $\mathbb{N}$, such that (8) holds by assumption. In addition it is assumed that for some constant D

$$\rho(k) \gtrsim e^{-Dk\log k}\,, \tag{40}$$

which we need to obtain the lower bound in (10). Given $K = k$, the locations $\mu_1, \ldots, \mu_k$ are drawn independently from a prior p_μ on $\mathbb{R}$ satisfying

$$p_\mu(x) \gtrsim \psi(x), \tag{41}$$

$$p_\mu(x) \lesssim e^{-a_1|x|^{a_2}} \quad \text{for constants } a_1 > 0 \text{ and } a_2 \le p\,. \tag{42}$$

The latter assumption implies that for any $y > 0$,

$$\Pi(\mid \mu_i \mid > y) = \int_{[-y,y]^c} p_\mu(x)dx \lesssim y^{\max\{0,1-a_2\}} e^{-y^{a_2}} \lesssim e^{-|y|^{d_2}} \tag{43}$$

for some constant $d_2 > 0$. Because $E_\rho K < \infty$ by condition (8), (9) follows from (43):

$$\begin{aligned} \Pi(N([-y,y]^c) > 0) &= \sum_{k=1}^{\infty} \rho(k)\, \Pi(\max_{i=1,\ldots,k} \mid \mu_i \mid > y \mid K = k) \\ &\le \sum_{k=1}^{\infty} \rho(k) k\, \Pi(\mid \mu_i \mid > y) \lesssim (E_\rho K)\, e^{-|y|^{d_2}}. \end{aligned} \tag{44}$$

To verify (10), let $c > 0$, $k > 0$, $\epsilon < \frac{1}{k}$, and $\mu_0 \in T^k_{c,\epsilon}$. Because $p_\mu \gtrsim \psi$ and for all $i = 1, \ldots, k$, $\mid \mu_i^0 \mid$ is at most $c|\log\epsilon|^{1/p}$,

$$\Pi(\mid \mu_i - \mu_i^0 \mid \le \frac{\epsilon}{k}) \gtrsim \int_{t_\epsilon - \frac{\epsilon}{k}}^{t_\epsilon + \frac{\epsilon}{k}} p_\mu(x)dx \gtrsim \frac{\epsilon}{k}\psi(c|\log\epsilon|^{1/p}) = \frac{C_p}{k}\epsilon^{c^p - 1}.$$

As the l_1-ball $B_k(\mu^0, \epsilon)$ contains the l_∞-ball $\{\mu \in \mathbb{R}^k : \mid \mu_i - \mu_i^0 \mid \leq \frac{\epsilon}{k},\ 1 \leq i \leq k\}$, we conclude that there is a constant $d_3 > 0$ such that

$$\Pi(K = k, \mu \in B_k(\mu^0, \epsilon)) \gtrsim e^{-Dk \log k} \left(\frac{C_p}{k} \epsilon^{c^p - 1} \right)^k \gtrsim \exp\{-d_3 k \log \frac{1}{\epsilon}\}, \tag{45}$$

where we used (40) and the fact that $k \leq \epsilon^{-1}$.

Conditions (8) and (40) imply that ρ needs to be of exponential form. If for example for some positive constants B_1, B_2 and $A_1 \geq A_2$

$$B_1 e^{-A_1 k} \leq \rho(k) \leq B_2 e^{-A_2 k}, \tag{46}$$

condition (8) holds for $r = 0$. Such exponential bounds were used by [7] for density estimation with mixtures of beta-densities. If ρ is Poisson with intensity ν, we have $\sum_{k=m+1}^{\infty} \rho(k) \leq \frac{\nu^{m+1}}{(m+1)!}$, and using Stirling's bound for $(m+1)!$ it can be seen that (8) holds with $r = 1$. For a geometric prior, $r = 0$.

Poisson processes are another popular choice for the location prior. We consider a Poisson point process with base measure P_μ on $\mathbb{R}$ and intensity λ, and assume that P_μ has a density p_μ for which (41) and (42) hold. Again a lower bound on $\Pi(K = k, \mu \in B_k(\mu^0, \epsilon))$ can be obtained by bounding $\Pi(\{K = k, \mu \in \mathbb{R}^k : \mid \mu_i - \mu_i^0 \mid \leq \frac{\epsilon}{k},\ 1 \leq i \leq k\})$. For some integer $l \leq k$, we can find disjoint intervals $I_1, \ldots, I_l \subset T_{c,\epsilon}$ of length $\frac{\epsilon}{k}$, containing $\mu_1^0, \ldots, \mu_k^0$. Let k_i be the number of points in I_i, and I^c the complement of $I_1 \cup \ldots \cup I_l$ in $T_{c,\epsilon}$. Since all I_i are contained in $T_{c,\epsilon}$ and $p_\mu \gtrsim \psi$,

$$P_\mu(I_i) = \int_{I_i} p_\mu(x) dx \gtrsim \int_{I_i} \psi(x) dx \gtrsim \lambda(I_i) \psi(c|\log \epsilon|^{1/p}) = \frac{\epsilon}{k} C_p \epsilon^{c^p}. \tag{47}$$

Again the tail assumptions on p_μ can be used to verify (9):

$$\begin{aligned} \Pi(N([-y, y]^c) > 0) &= 1 - \Pi(N([-y, y]^c) = 0) \\ &= 1 - \exp(-\lambda P_\mu([-y, y]^c)) \leq \lambda P_\mu([-y, y]^c) \lesssim e^{-|y|^{d_2}}. \end{aligned}$$

Finally, (10) follows from (47), as we find a constant d_3 such that

$$\begin{aligned} \Pi(K = k, \mu \in B_k(\mu^0, \epsilon)) &\geq P\Big(N(I_1) = k_1, \ldots, N(I_l) = k_l, N(I^c) = 0, N(T_{c,\epsilon}^c) = 0\Big) \\ &= \frac{\exp\{-\lambda\}}{k_1! \cdot \ldots \cdot k_l!} \prod_{i=1}^{l} (P_\mu(I_i)\lambda)^{k_i} \\ &\geq \frac{\exp\{-\lambda\}}{k!} (C_p \epsilon^{c^p} \frac{\epsilon}{k} \lambda)^k \gtrsim \exp\{-d_3 k \log \frac{1}{\epsilon}\}. \end{aligned}$$

4.2. *Priors on the weights.* In this section two classes of priors on the simplex are discussed. In both cases the Dirichlet distribution, the most popular choice in the literature, appears as a special case. The proof of Theorem 1 requires lower bounds for the prior mass on l_1-balls around some fixed point in the simplex. These bounds are given in Lemmas 8 and 10 below.

As it is a well known fact that 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 $(x_1, \ldots, x_{k-1}) \in S_{k-1}$,

(48)
$$P(X_1 \le x_1, \ldots, X_{k-1} \le x_{k-1}) = \int_0^\infty P(Y_1 \le x_1 y, \ldots Y_{k-1} \le x_{k-1} y)\, dP^{Y_1+\ldots+Y_k}(y)$$
$$= \int_0^\infty \int_0^{x_1 y} \int_0^{x_2 y} \cdots \int_0^{x_{k-1} y} f_k(y - \sum_{i=1}^{k-1} s_i) \prod_{i=1}^{k-1} f_i(s_i) ds_1 \cdots ds_{k-1} dy.$$

The corresponding density is

$$\begin{aligned} f^{X_1,\ldots,X_{k-1}}(x_1, \ldots, x_{k-1}) &= \int_0^\infty y^{k-1} f_k(y - \sum_{i=1}^{k-1} x_i y) \prod_{i=1}^{k-1} f_i(x_i y) dy \\ &= \int_0^\infty y^{k-1} \prod_{i=1}^{k} f_i(x_i y) dy, \end{aligned} \tag{49}$$

where $x_k = 1 - \sum_{i=1}^{k-1} x_i$. We obtain a result similar to lemma 8 in [10].

LEMMA 8. *Let $X_1, \ldots, X_k$ have a joint distribution with a density of the form* (49). *Assume there are positive constants $c_1(k)$, $c_2(k)$ and c_3 such that for $i = 1, \ldots, k$, $f_i(z) \ge c_1(k) z^{c_3}$ if $z \in [0, c_2(k)]$. Then there are constants c and C such that for all $y \in \Delta_k$ and all $\epsilon \le (\frac{1}{k} \wedge c_1(k) c_2(k)^{c_3+1})$*

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

PROOF. As in [10] it is assumed that $y_k \ge k^{-1}$. 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 $i = 1, \ldots, k-1$, then $\sum_{i=1}^k \mid x_i - y_i \mid \le 2 \sum_{i=1}^{k-1} \mid x_i - y_i \mid \le 2(k-1)\epsilon^2 \le \epsilon$. Note that $(x_1, \ldots, x_{k-1}) \in S_k$, as $\sum_{j=1}^{k-1} x_j \le \frac{k-1}{k} + (k-1)\epsilon^2 < 1$. Since all x_i in (49) are at most one,

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

Because

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

$$P\Big(X \in B_k(y,\epsilon)\Big) \geq \frac{1}{k^{2c_3}} \frac{\big(c_2(k)^{c_3+1}c_1(k)\big)^k}{(c_3+1)k} \prod_{j=1}^{k-1} \int_{\underline{\delta_j}}^{\bar{\delta}_j} x_j^{c_3} dx_j \geq \frac{\big(c_2(k)^{c_3+1}c_1(k)\big)^k}{(c_3+1)^2 k} \epsilon^{2k(c_3+1)-2}$$

$$\geq \exp\{k\log(c_2(k)^{c_3+1}c_1(k)) - \log(c_3+1) - \log(k) - 2k\log(\frac{\sqrt{2}}{\epsilon})\}.$$

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 $Ce^{-ck\log(\frac{1}{\epsilon})}$. □

Alternatively, the Dirichlet distribution can be seen as a Polya tree. Following Lavine [15] 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

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

where $(U_\delta, \delta \in E_*^{m-1})$ is a family of beta random variables with parameters $((\alpha_{\delta_1}, \alpha_{\delta_2}), \delta \in E_*^{m-1})$. 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 $X_{\delta 1} = (1 - U_\delta)X_\delta$, and $X_\emptyset = 1$ by construction. If $\alpha_\delta = \frac{1}{2}\alpha_{\delta_1\cdots\delta_{i-1}}$ for all $1 \leq i \leq m$ and $\delta \in E^i$, X is Dirichlet distributed.

Lemma 9. *Let X have a Polya distribution with parameters α_δ, $\delta \in E_*^{m-1}$. Then for all $y \in \Delta_{2^m}$ and $\eta > 0$,*

$$p_m(y,\eta) = P\Big(X \in \Delta_k(y,\eta)\Big) = P(\sum_{\epsilon \in E^m} \mid X_\epsilon^m - y_\epsilon^m \mid \leq \eta)$$

$$\geq \prod_{i=1}^{m} P(\max_{\partial \in E^{i-1}} \mid U_\delta - \frac{y_{\delta 0}}{y_\delta} \mid \leq \frac{\eta}{2^{m-i+2}}).$$

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

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

Consequently,

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

Hence,

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

as

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

□

With $\delta \in E^{i-1}$ fixed, we can lower-bound $P(\mid U_\delta - \frac{y_{\delta 0}}{y_\delta} \mid \leq \frac{\eta}{2^{m-i+2}})$ 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 $i = 1, \ldots, m$. For increasing $\alpha_i \geq 1$, U_δ has a unimodal beta-density, and we can restrict to the "worst case" where $\frac{y_{\delta 0}}{y_\delta} = 0$. If the α_i are decreasing, and smaller than one, the worst case is $\frac{y_{\delta 0}}{y_\delta} = \frac{1}{2}$. In both cases Lemma 11 in appendix A is used to lower bound the normalizing constant of the beta-density.

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

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

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

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

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

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

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

We have the following application of these results.

LEMMA 10. *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 $i = 1, \ldots, m$,

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

in accordance with the result in [8].

5. Conclusion. We obtained posteriors that adapt to the smoothness of the estimated 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 [13], 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 [13] 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 (37), we get an extra factor k_n in the exponent, and the argument only holds 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 [21]. 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 [9].

APPENDIX A

The following theorem is taken from [10] (Theorem 5), and slightly adapted to facilitate the entropy calculations in the proof of Theorem 1. Their condition $\Pi_n(\mathcal{F}_n|X_1,\ldots,X_n) \to 0$ in F_0^n-probability is a consequence of (51) and (52) below. This follows from a simplification of the proof of Theorem 2.1 in [8], 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 [10], with $\bar{\epsilon}_n \geq \tilde{\epsilon}_n$ and $\bar{\epsilon}_n \to 0$, the result of Theorem 5 in this paper still holds.

THEOREM 2 (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 $\tilde{\epsilon}_n$ and $\bar{\epsilon}_n \geq \tilde{\epsilon}_n$ with $\bar{\epsilon}_n \to 0$ and $n\tilde{\epsilon}_n^2 \to \infty$,*

$$\sum_{j=-\infty}^{\infty} \sqrt{N(\bar{\epsilon}_n, \mathcal{F}_{n,j}, d)}\sqrt{\Pi_n(\mathcal{F}_{n,j})}e^{-n\bar{\epsilon}_n^2} \quad \to \quad 0, \tag{50}$$

$$\Pi_n(KL(f_0, \tilde{\epsilon}_n)) \quad \geq \quad e^{-n\tilde{\epsilon}_n^2}, \tag{51}$$

$$\Pi_n(\mathcal{F}_n^c) \leq e^{-4n\tilde{\epsilon}_n^2}, \tag{52}$$

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

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

Then $\Pi_n(f \in \mathcal{F} : d(f, f_0) > 8\bar{\epsilon}_n \mid X_1, \ldots, X_n) \to 0$ in F_0^n-probability.

The advantage of the above version is that (52) is easier to verify for a faster sequence $\tilde{\epsilon}_n$. The use of the same sequence ϵ_n in (50) and (52) 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 11. *For any $\alpha > 0$,*

$$\Gamma(\alpha) = \sqrt{2\pi} e^{-\alpha} \alpha^{\alpha - \frac{1}{2}} e^{\theta(\alpha)}, \tag{53}$$

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 $\Gamma(\alpha) \le \frac{1}{\alpha}$ and $\Gamma(\alpha) \ge \frac{c}{\alpha}$, 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 (53) it follows that for all $\alpha > 0$ and all integers $j \ge 1$,

$$\frac{\sqrt{\Gamma(\frac{2j+1}{1+\alpha})}}{j!} \le \frac{1}{\sqrt{2\pi}} e^{\frac{\alpha}{1+\alpha}(j+1)} \left(\frac{2}{1+\alpha}\right)^{\frac{j}{1+\alpha}} (j+1)^{-\frac{\alpha j}{1+\alpha}}, \tag{54}$$

$$\frac{\Gamma(\frac{j+1}{1+\alpha})}{j!} \le e^{\frac{\alpha}{1+\alpha}(j+1) + \frac{1}{12}} \left(\frac{1}{1+\alpha}\right)^{\frac{j}{1+\alpha}} (j+1)^{-\frac{\alpha j}{1+\alpha}}. \tag{55}$$

The following lemma will be required for the proof of Lemma 2 in the next section.

LEMMA 12. *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 $\alpha \ge 0$ there is a number $k' = k'(p, \alpha, m)$ such that for all sufficiently small $\sigma > 0$,*

$$\int_{|x| > k' |\log \sigma|^{1/p}} \varphi(x) |x|^\alpha dx = \sigma^H. \tag{56}$$

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

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

For $m = 1$, we have

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

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 $\alpha \ge 1$ 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 (56) 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: PROOF OF LEMMA 2

Under our assumptions on f, the fact that $\log f$ is β-Hölder implies that also f itself is β-Hölder. Consequently, Lemma 1 implies that for any fixed x, $K_\sigma f_k - f$ is $O(\sigma^\beta)$. To prove Lemma 2, it suffices to show that, apart from a term σ^H, the multiplicative constant is a multiple of $R(x)$. For this we need the restriction $x \in A_\sigma$ and the additional smoothness conditions (4) and (5). 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 consider the case $f \in H_\beta$ with $\beta \in (1,2]$ and $r = 1$; the case $\beta \in (0,1]$ is easier and can be handled similarly. Using (4) we establish

$$K_\sigma f(x) \le (1+O((|L(x)|+|l_1^\beta(x)|)\sigma^\beta))f(x)+O(1+|L(x)|+|l_1^\beta(x)|)\sigma^H. \tag{57}$$

The proof of the inverse inequality using (5) is completely analogous. For any $x \in \mathbb{R}$, let

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

for a large enough constant k' to be chosen below.

Assuming that $k'\sigma|\log\sigma|^{1/p} \le \gamma$, for γ as in Condition 1 on page 3, we can rewrite (4) as $f(y) \le f(x)\exp\{B_{f,1}(x,y)\}$, and

$$K_\sigma f(x) \le f(x)\int_{D_x} e^{B_{f,r}(x,y)}\psi_\sigma(y-x)dy + \int_{D_x^c} f(y)\psi_\sigma(y-x)dy. \tag{58}$$

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

(59)
$$\begin{aligned} e^{B_{f,r}(x,y)} &= \sum_{m=0}^{r}\frac{1}{m!}B_{f,r}^m(x,y) + \frac{e^\xi}{(r+1)!}B_{f,r}^{r+1}(x,y) \le 1 + B_{f,r}(x,y) + MB_{f,r}^2(x,y) \\ &= 1 + l_1(x)(y-x) + L(x)|y-x|^\beta \\ &\quad + M\left(l_1^2(x)(y-x)^2 + 2l_1(x)L(x)(y-x)|y-x|^\beta + L^2(x)|y-x|^{2\beta}\right). \end{aligned}$$

Integrating over D_x, the terms with a factor $(y-x)$ disappear, so that the first integral on the right in (58) is bounded by

(60)
$$\int_{D_x}\psi_\sigma(y-x)\left\{1 + L(x)|y-x|^\beta + M(k'B)^{2-\beta}|l_1(x)(y-x)|^\beta + Mk'^\beta B|L(x)(y-x)|^\beta\right\}dy,$$

since $|l_1(x)(y-x)| \le k'B$ and $|L(x)(y-x)| \le k'^\beta B$ 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 $\alpha \ge 0$, when k' in the definition of D_x is sufficiently large (see Lemma 12 in Appendix A), (58), (59) and (60) imply that for constants $k_1 = M(k'B)^{2-\beta}$ and $k_2 = 1+Mk'^\beta B$,

(61)
$$\begin{aligned} (K_\sigma f)(x) &\le \int_{\mathbb{R}}\psi_\sigma(y-x)\{1 + k_1|l_1(x)|^\beta|y-x|^\beta + k_2|L(x)||y-x|^\beta\}dy \\ &\quad + (\|f\|_\infty + 1 + k_1|l_1(x)|^\beta + k_2|L(x)|)O(\sigma^H), \end{aligned}$$

which completes the proof of (57) for $\beta \in (1,2]$. Using the same arguments the inverse inequality can be obtained when we define $B_{f,1}(x,y) = l_1(x)(y-x) - L(x)|y-x|^\beta$. Consequently, we have shown that

(62)
$$(K_\sigma f)(x) = f(x)\left(1 + O(|l_1(x)|^\beta + |L(x)|)\sigma^\beta\right) + O(1 + |l_1(x)|^\beta + |L(x)|)\sigma^H.$$

For the inequality in this result resulting from (61), the term $O(|l_1(x)|^\beta + |L(x)|)\sigma^\beta f(x)$ could be replaced by $(k_1\nu_\beta|l_1(x)|^\beta + k_2\nu_\beta|L(x)|)\sigma^\beta f(x)$, with ν_β as in (3). However to combine this result with the corresponding inverse inequality we use the O-notation.

Now let f be a function for which (4) and (5) hold with $\beta \in (2,4]$. Treating $K_\sigma f$ and $K_\sigma K_\sigma f$ separately, we show that for $f_1 = 2f - K_\sigma f$

$$(K_\sigma f_1)(x) = f(x)\left(1 + O(R(x)\sigma^\beta)\right) + O\left((1 + R(x))\sigma^H\right), \tag{63}$$

where R is defined as in (18). First we give an expression for $K_\sigma f$ when $\beta \in (2,4]$, again by showing inequality in both directions. For the upper

bound on $K_\sigma f$ when $r = 3$ for example, note that when $x \in A_\sigma$ and $y \in D_x$, $e^B \leq 1 + B + \frac{1}{2}B^2 + \frac{1}{6}B^3 + MB^4$ for some constant M, with $B(x,y) = l_1(x)(y-x) + \frac{1}{2}l_1(x)(y-x) + \frac{1}{6}l_3(x)(y-x)^3 + L(x)|y-x|^\beta$. For this bound on e^B we redo the calculations given in (58), (59), (60) and (61), and we find that

(64)
$$K_\sigma f(x) = f(x)\left(1 + \frac{\nu_2}{2}(l_1^2(x) + l_2(x))\sigma^2 + O(R(x)\sigma^\beta)\right) + O\left((1 + R(x))\sigma^H\right).$$

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 $k > 2$ in similar way as in (60). 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 $|L|$. For instance, the term $\frac{1}{8}l_2^2(x)(x-y)^4$ resulting from $\frac{1}{2}B^2$ can be written as $\frac{1}{8}|l_2(x)|^{1-\beta/2}|x-y|^{2-\beta}|x-y|^\beta|l_2(x)|^{\beta/2} \lesssim |x-y|^\beta|l_2(x)|^{\beta/2}$. For terms that are products of different l_j's we first completely 'compensate' the lower order l_j's by the appropriate powers of $|x-y|$. The term $\frac{1}{6}l_1(x)l_3(x)(y-x)^4$ coming from $\frac{1}{2}B^2$ for example, is bounded by

$$\frac{k'B}{6}|l_3(x)||y-x|^3 = |l_3(x)|^{1-\beta/3}|y-x|^{3-\beta}|y-x|^\beta|l_3(x)|^{\beta/3} \lesssim |y-x|^\beta|l_3(x)|^{\beta/3}.$$

Similarly, the terms $\frac{1}{3}|L(x)|l_1^2(x)|y-x|^{2+\beta}$ and $\frac{1}{6}|L(x)|^3|y-x|^{3\beta}$ coming from $\frac{1}{6}B^3$ are bounded by multiples of $|L(x)||y-x|^\beta$. Hence all the terms resulting from e^B with a factor $|y-x|^k$ with $k > 2$ can be bounded by a multiple of $|y-x|^\beta|l_1|^\beta$, $|y-x|^\beta|l_1|^{\beta/2}$, $|y-x|^\beta|l_1|^{\beta/3}$ or $|y-x|^\beta|L|$, leading to the sum defined by (18). Note that also the terms with $|y-x|^k$ with $k > \beta$ are reduced to $|x-y|^\beta$, as terms with σ^k would not improve the result (at least not for our purpose) and would require stronger moment conditions on l_j's.

To find a similar expression for $K_\sigma K_\sigma f$, note that (64) does not depend on the explicit form of the kernel ψ. We only used the properties that ψ is symmetric with $\int \psi(x)|x|^\alpha dx = \nu_\alpha < \infty$ for certain numbers ν_α, and that $\int_{|x|>k'|\log\sigma|^{1/p}} \psi(x)|x|^\alpha dx = \sigma^H$ when k' is sufficiently large. For the kernel $\varphi = \psi * \psi$ these properties follow from Lemma 12 in Appendix A. Consequently, (64) also holds with K_σ denoting convolution over φ and ν_2 replaced by $\nu_{\varphi,2} = \int \varphi(x)|x|^\alpha dx = 2\nu_2$. Plugging in these results in $K_\sigma f_1 = 2K_\sigma - K_\sigma K_\sigma f$, we find that the σ^2-terms cancel out, completing the proof of (63).

When $k > 1$, $\beta \in (2k, 2k+2]$ and $\log f$ is satisfying (4) and (5), it can be shown that all terms with $\sigma^2, \sigma^4, \ldots, \sigma^{2k}$ cancel out. This follows

directly from Lemma 1, but can also 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 (14). The terms $O(R(x))\sigma^\beta f(x)$ and $O(1+R(x))\sigma^H$ are obtained by the same arguments used above for the case $\beta \leq 4$.

APPENDIX C: DISCRETIZATION

The following lemmas can be found in [13], p.59-60. They are straightforward extensions of the corresponding results for normal mixtures, contained in lemma 3.1 of [11] and lemma 2 of [10]. Lemma 14 is used in the proof of Lemma 6 in the present work.

LEMMA 13. *Given $p > 0$, 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 14. *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] 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.

[3] Michael D. Escobar and Mike West. Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.*, 90(430):577–588, 1995.

[4] Stuart Geman and Chii-Ruey Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.*, 10(2):401–414, 1982.

[5] Christopher R. Genovese and Larry Wasserman. Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.*, 28(4):1105–1127, 2000.

[6] S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.*, 27(1):143–158, 1999.

[7] Subhashis Ghosal. Convergence rates for density estimation with Bernstein polynomials. *Ann. Statist.*, 29(5):1264–1280, 2001.

[8] Subhashis Ghosal, Jayanta K. Ghosh, and Aad W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 2000.

[9] 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.

[10] Subhashis Ghosal and Aad van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.*, 35(2):697–723, 2007.

[11] 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.

[12] Ulf Grenander. *Abstract inference*. John Wiley & Sons Inc., New York, 1981. Wiley Series in Probability and Mathematical Statistics.

[13] 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.

[14] 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.

[15] Michael Lavine. Some aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.*, 20(3):1222–1235, 1992.

[16] J.M. Marin, K. Mengersen, and C.P. Robert. *Bayesian modelling and inference on mixtures of distributions*. Elsevier-Sciences, 2005.

[17] 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.

[18] Judith Rousseau. Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *Ann. Statist.*, 2009.

[19] C. Scricciolo. Convergence rates of posterior distributions for dirichlet mixtures of normal densities. working paper 2001-21. Technical report, 2001.

[20] Yuefeng Wu and Subhashis Ghosal. Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electron. J. Stat.*, 2:298–331, 2008.

[21] Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Trans. Inform. Theory*, 52(4):1307–1321, 2006.

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