

## ENTROPIES AND RATES OF CONVERGENCE FOR MAXIMUM LIKELIHOOD AND BAYES ESTIMATION FOR MIXTURES OF NORMAL DENSITIES

BY SUBHASHIS GHOSAL AND AAD W. VAN DER VAART

*University of Minnesota and Free University of Amsterdam*

We study the rates of convergence of the maximum likelihood estimator (MLE) and posterior distribution in density estimation problems, where the densities are location or location-scale mixtures of normal distributions with the scale parameter lying between two positive numbers. The true density is also assumed to lie in this class with the true mixing distribution either compactly supported or having sub-Gaussian tails. We obtain bounds for Hellinger bracketing entropies for this class, and from these bounds, we deduce the convergence rates of (sieve) MLEs in Hellinger distance. The rate turns out to be  $(\log n)^\kappa/\sqrt{n}$ , where  $\kappa \geq 1$  is a constant that depends on the type of mixtures and the choice of the sieve. Next, we consider a Dirichlet mixture of normals as a prior on the unknown density. We estimate the prior probability of a certain Kullback-Leibler type neighborhood and then invoke a general theorem that computes the posterior convergence rate in terms the growth rate of the Hellinger entropy and the concentration rate of the prior. The posterior distribution is also seen to converge at the rate  $(\log n)^\kappa/\sqrt{n}$  in, where  $\kappa$  now depends on the tail behavior of the base measure of the Dirichlet process.

**1. Introduction.** A mixture of normal densities is often used to model an unknown smooth density because of its wide range of flexibility and infinite degree of smoothness. Normal mixture models have been used for a variety of inference problems including density estimation, clustering analysis and robust estimation; see, for example, Lindsay (1995), McLachlan and Basford (1988), Banfield and Raftery (1993), Robert (1996) and Roeder and Wasserman (1997). The mixture model is a fully nonparametric class which is nevertheless appropriate for model based inference such as the maximum likelihood method or the Bayesian method.

Variants of the maximum likelihood method have been considered in the literature; see Roeder (1992) and Priebe (1994). Since the likelihood is unbounded without any restriction on the support of the mixing distribution, it is necessary to restrict the maximization over a suitable subset called a sieve, which grows with the sample size. The sieve method was introduced by Grenander (1981) and studied by many authors including Geman and Hwang (1982), van de Geer (1993), Shen and Wong (1994), Wong and Shen (1995) and Birgé and Massart (1998). While Roeder (1992) maximized a likelihood based on spacings, Priebe (1994) considered finite mixtures and described an

---

Received April 2000; revised March 2001.

AMS 2000 subject classifications. Primary 62G07, 62G20.

Key words and phrases. Bracketing, Dirichlet mixture, entropy, maximum likelihood, mixture of normals, posterior distribution, rate of convergence, sieve.

algorithm to adaptively select a finitely supported mixing distribution. These authors showed consistency of the resulting estimates. Van de Geer (1996) obtained the rate of convergence of the maximum likelihood estimate (MLE) in some mixture models, but she did not discuss the case of normal mixtures.

From a Bayesian point of view, the mixture model provides an ideal platform for density estimation where one can induce a prior distribution on the densities by attaching a prior distribution to the mixing distribution. Such an approach was taken by Ferguson (1983) and Lo (1984) who used a Dirichlet process prior on the mixing distribution and obtained expressions for Bayes estimates. Note that the Dirichlet process always selects discrete distributions and hence it cannot be directly used as a prior on densities. In a recent article, Ghosal, Ghosh and Ramamoorthi (1999a) showed that the Dirichlet mixture of normal prior gives rise to a consistent posterior under general conditions for the weak topology and the variation distance. These authors established weak posterior consistency by verifying Schwartz's (1965) condition of the positivity of the prior probabilities of Kullback-Leibler neighborhoods. In addition, by bounding the  $L_1$ -metric entropies of the class of mixtures, certain uniformly consistent tests were obtained. Existence of the uniformly consistent tests and the Schwartz condition together imply posterior consistency in the variation distance. The main purpose of the paper, among other things, is to refine this result to a rate of convergence. For a general discussion on posterior consistency for non-parametric problems, see the recent reviews Ghosal, Ghosh and Ramamoorthi (1999b) and Wasserman (1998).

Ferguson (1983) and Lo (1984) obtained analytic expressions for the Bayes estimates for the Dirichlet mixture prior. Unfortunately, these expressions are not suitable for computational purposes, because of their exponential order of complexity. We can nevertheless compute the Bayes estimates by simulation methods. Due to the substantial progress of Bayesian computing in the last decade, posterior characteristics such as the posterior mean can now be computed for many Bayesian nonparametric problems including the Dirichlet mixture model by Markov chain Monte Carlo methods. West (1992), West, Mueller and Escobar (1994) and Escobar and West (1995) among others, developed Gibbs sampling techniques to compute the Bayes estimate and other posterior quantities for the Dirichlet mixture prior. These authors also used the Dirichlet mixture prior effectively in many applications.

In this article, we obtain the rate of convergence of the MLE and sieve MLEs as well as the rate of convergence of the posterior distribution for Dirichlet mixtures. We first obtain bounds on the bracketing entropies of the class of normal mixtures. These bounds immediately give us the rates of convergence of MLE and sieve MLEs in view of the results of van de Geer (1993), Wong and Shen (1995) and Birgé and Massart (1998). Our entropy bounds are new and the method we use appears to be new too. We believe these bounds will be of independent interest as well.

To obtain the posterior rate of convergence, we further compute the concentration rate of the prior distribution on a Kullback-Leibler type neighborhood and then apply the recent results on posterior rate of convergence obtained by

Ghosal, Ghosh and van der Vaart (2000); see also Shen and Wasserman (2001) for a similar result under stronger conditions. For easy reference, the relevant result of Ghosal, Ghosh and van der Vaart (2000) is stated as Theorem 2.1 in Section 2 of this paper. The condition (2.10) in that theorem may be thought of as a quantitative analogue of the Schwartz condition. Condition (2.9) helps us to effectively reduce the size of the parameter space. Condition (2.8) implies the existence of tests with sufficiently high power for testing against the complement of a neighborhood shrinking at a certain rate and can be viewed as a quantitative refinement of the testing condition in Schwartz (1965) or the entropy conditions in Ghosal, Ghosh and Ramamoorthi [(1999a), Theorem 2] or Barron, Schervish and Wasserman [(1999), Assumption 2]. To obtain the right rate of convergence, we however need much more precise entropy and prior estimates than those used by Ghosal, Ghosh and Ramamoorthi (1999a). The relatively crude bounds obtained there sufficed for posterior consistency but are inadequate for rates.

We consider both location mixtures and location-scale mixtures where the scale is assumed to be bounded above and below and the true mixing distribution of the location is compactly supported or has sub-Gaussian tails. A near parametric rate  $(\log n)^\kappa/\sqrt{n}$  of convergence is obtained for the MLE and its variants and the posterior, where  $\kappa \geq 1$  depends on the type of mixtures and the choice of the sieve for the sieve MLE or the tail behavior of the base measure for the posterior. It should be noted here that this sharp rate in contrast with those of the popular estimators such as the kernel estimator is due to the assumption that the true density is also a mixture of normals whose scale parameters can vary only between two positive numbers, in addition to the smoothness of the normal density. When the true density lies outside that class, in order to approximate such a density, one has to let the scale take arbitrarily small positive values. It will be very interesting to study the convergence rates for this problem. In this case, one will have to consider a sieve where the scale is bounded below by a number decreasing to zero at some rate. Our basic inequalities are explicit in the lower bound of the scale, so a rate may be calculated from these bounds. However, the obtained rate does not appear to be close to the optimal rate, and hence we do not pursue it here.

Recently Genovese and Wasserman (2000) also obtained bounds for bracketing Hellinger entropies for normal mixtures where the scale parameter again lies between two positive numbers and as a result, computed the convergence rates of sieve MLEs. They considered only location mixtures and obtained the convergence rate  $n^{-1/6}(\log n)^{(1+\delta)/6}$  of the sieve MLE for some  $\delta > 0$ . In contrast, we consider the maximum likelihood as well as the Bayesian methods, treat both the location and the location-scale mixtures and at the same time obtain much faster rates.

As far as we are aware, the optimal rate of convergence relative to the Hellinger distance is unknown for our problem. However, Ibragimov (2001) and Ibragimov and Khasminskii (1982) have studied the minimax rate for the problem of estimating an entire density relative to the  $L_p$ -norms for  $p \geq 2$ . Specifically, consider estimating a density  $f$  based on an independent sample

of size  $n$  from  $f$ , where  $f$  is known to belong to the class  $\mathcal{F}$  of density functions that admit an analytic continuation  $f : \mathbb{C} \rightarrow \mathbb{C}$  such that

$$\sup_{x \in \mathbb{R}} |f(x + iy)| \leq M e^{c|y|^2}, \quad y \in \mathbb{R}.$$

Then for  $p \geq 2$  and some positive constants  $C_1$  and  $C_2$ ,

$$C_1 \frac{(\log n)^{1/4}}{\sqrt{n}} \leq \inf_{\hat{f}} \sup_{f \in \mathcal{F}} E_f \left( \int_{-\infty}^{\infty} |\hat{f}(x) - f(x)|^p dx \right)^{1/p} \leq C_2 \frac{(\log n)^{1/4}}{\sqrt{n}}.$$

The class of densities under consideration in our paper is smaller than the above class  $\mathcal{F}$  (for some  $M$  and  $c$ ). This suggests that the power of the logarithm in our results can possibly be improved a little. This is not certain as Ibragimov (2001) considered ad-hoc and not Bayesian or maximum likelihood estimators and  $L_p$ -norms for  $p \geq 2$  rather than Hellinger or  $L_1$ -norms. However, it appears plausible, even though mathematically intractable to us at this time.

The organization of the paper is as follows. In Section 2 we explain the set up and notations and discuss the necessary prerequisites. Bounds for entropies and bracketing entropies are obtained in Section 3. Using these bounds, the convergence rates for the MLE and sieve MLEs are obtained in Section 4. Convergence rates of posterior distribution for location mixtures and location-scale mixtures are respectively obtained in Section 5 and Section 6. In Section 7, we consider certain extensions and variations of the results of Section 5. Some lemmas of more general interests are presented in the Appendix.

**2. Notation and preliminaries.** Suppose we have independent observations  $X_1, X_2, \dots$  from a common density  $p(x)$  on the real line. Let  $\phi(x)$  stand for the standard normal density  $(2\pi)^{-1/2} \exp(-x^2/2)$  and let  $\phi_\sigma(x) = \sigma^{-1} \phi(x/\sigma)$  be the density of the normal distribution with mean zero and standard deviation  $\sigma$ . We model the density in three different ways in increasing order of generality. We assume that the density  $p(x)$  is either a location mixture of normals, that is,

$$(2.1) \quad p(x) = p_{F,\sigma}(x) = \int \phi_\sigma(x - z) dF(z),$$

where  $F(\cdot)$  is a probability distribution on  $\mathbb{R}$  called the mixing distribution or the latent distribution, or  $p(x)$  is a location-scale mixture of normals of the type

$$(2.2) \quad p(x) = p_{F,G}(x) = \int \int \phi_\sigma(x - z) dF(z) dG(\sigma),$$

where  $F(\cdot)$  is a probability distribution on  $\mathbb{R}$  and  $G(\cdot)$  is a probability distribution on  $(0, \infty)$ , or  $p(x)$  is a location-scale mixture of normals of the type

$$(2.3) \quad p(x) = p_H(x) = \int \phi_\sigma(x - z) dH(z, \sigma),$$

where  $H(\cdot, \cdot)$  is a probability distribution on  $\mathbb{R} \times (0, \infty)$ . In model (2.1), let  $F_0(\cdot)$  and  $\sigma_0$  be the true values of  $F(\cdot)$  and  $\sigma$  respectively. In model (2.2), let  $F_0(\cdot)$  and  $G_0(\cdot)$  be the true values of  $F(\cdot)$  and  $G(\cdot)$  respectively while the true value of  $H(\cdot, \cdot)$  in model (2.3) is denoted by  $H_0(\cdot, \cdot)$ . The possible values of  $\sigma$  in all the models are assumed to lie in a compact interval  $[\underline{\sigma}, \bar{\sigma}] \subset (0, \infty)$ . The values of  $\underline{\sigma}, \bar{\sigma}$  are kept fixed throughout. For all the three models, we write  $p_0$  for the true value of  $p$  and  $P_0$  for the probability distribution corresponding to  $p_0$ .

We determine the rates of convergence of the MLE or sieve MLEs and the posterior distribution. As a prior for  $p$ , we consider a Dirichlet mixture of normals. To be more precise, in model (2.1), we let  $F$  have the Dirichlet process distribution  $D_\alpha$ , where  $\alpha$  is a finite, positive measure on  $\mathbb{R}$  and let  $\sigma$  be distributed on  $[\underline{\sigma}, \bar{\sigma}]$  independently of  $F$ . For model (2.2), we consider independent Dirichlet process priors  $D_\alpha$  and  $D_\beta$  respectively for  $F$  and  $G$ , where  $\alpha$  is as above and  $\beta$  is a measure on  $[\underline{\sigma}, \bar{\sigma}]$ . For model (2.3),  $F$  is assumed to have the Dirichlet process prior  $D_\gamma$ , where  $\gamma$  is a finite, positive measure on  $\mathbb{R} \times [\underline{\sigma}, \bar{\sigma}]$ . Recall that the Dirichlet process on a measurable space  $\mathfrak{X}$  with a base measure  $\alpha$  is a random probability measure  $F$  on  $\mathfrak{X}$  such that for every finite partition  $(A_1, \dots, A_k)$  of  $\mathfrak{X}$ , the probability vector  $(F(A_1), \dots, F(A_k))$  has a Dirichlet distribution on the  $k$ -dimensional simplex with parameters  $(\alpha(A_1), \dots, \alpha(A_k))$ . We choose the Hellinger distance

$$d(f, g) = \left( \int (f^{1/2}(x) - g^{1/2}(x))^2 dx \right)^{1/2}$$

as the metric on the space of densities. Other possible choices are the variation or the  $L_1$ -norm  $\|f - g\|_1 = \int |f(x) - g(x)| dx$  and the  $L_2$ -norm  $\|f - g\|_2 = \left( \int |f(x) - g(x)|^2 dx \right)^{1/2}$ . It may be recalled that

$$(2.4) \quad d^2(f, g) \leq \|f - g\|_1 \leq 2d(f, g)$$

for any two densities  $f$  and  $g$ . Further, if the densities  $f$  and  $g$  are uniformly bounded by  $M$ , say, then

$$(2.5) \quad \|f - g\|_2 \leq 2\sqrt{M}d(f, g).$$

We shall show that under mild conditions, the posterior distribution based on  $X_1, \dots, X_n$  concentrates on Hellinger neighborhoods of  $p_0$  of size a large multiple of  $(\log n)^\kappa / \sqrt{n}$ , where  $\kappa \geq 1$  depends on the context. This substantially strengthens the assertion of posterior consistency shown by Ghosal, Ghosh and Ramamoorthi (1999a). Note that since normal mixtures are uniformly bounded, by (2.4) and (2.5),  $(\log n)^\kappa / \sqrt{n}$  is an upper bound for the rate of convergence for the  $L_1$  and  $L_2$ -distances as well.

To compute the rate of convergence of the posterior distribution, we shall compute Hellinger metric (and bracketing) entropies of the space of Gaussian mixtures and estimate from below the prior probabilities of a Kullback-Leibler type neighborhood of  $P_0$ . The recently obtained general results of Ghosal, Ghosh and van der Vaart (2000) on the rate of convergence of posterior distributions then immediately give us the the desired rate  $(\log n)^\kappa / \sqrt{n}$ . The rates

for the MLE or sieve MLEs are obtained from the estimates of the bracketing entropies and the results of Wong and Shen (1995), van de Geer (1993) or Birgé and Massart (1998).

The methods of obtaining entropy and prior estimates for the three models above differ in details, although the essential ideas are similar. Nevertheless, we need to consider these three models separately.

Let  $\mathcal{P}$  be a class of densities on  $\mathbb{R}$  and  $d$  be a metric on it. Let  $D(\varepsilon, \mathcal{P}, d)$  stand for the  $\varepsilon$ -packing number defined to be the maximum number of points in  $\mathcal{P}$  such that the distance between each pair is at least  $\varepsilon$ . The  $\varepsilon$ -covering number  $N(\varepsilon, \mathcal{P}, d)$  is defined to be the minimum number of balls of radius  $\varepsilon$  needed to cover  $\mathcal{P}$  and is related to the packing number by the inequalities

$$(2.6) \quad N(\varepsilon, \mathcal{P}, d) \leq D(\varepsilon, \mathcal{P}, d) \leq N(\varepsilon/2, \mathcal{P}, d).$$

A set  $\mathcal{P}_\varepsilon$  with the property that any element of  $\mathcal{P}$  is within  $\varepsilon$  distance from an element of  $\mathcal{P}_\varepsilon$  will be referred to as an  $\varepsilon$ -net over  $\mathcal{P}$ . The  $\varepsilon$ -bracketing number  $N_{[]}(\varepsilon, \mathcal{P}, d)$  is defined to be the minimum number of brackets of size  $\varepsilon$  necessary to cover  $\mathcal{P}$ , where a bracket of size  $\varepsilon$  is a set of the form  $[l, u] = \{f : l(x) \leq f(x) \leq u(x) \text{ for all } x\}$  for nonnegative integrable functions  $l$  and  $u$  with  $l(x) \leq u(x)$  for all  $x$  and  $d(l, u) < \varepsilon$ . Clearly,

$$(2.7) \quad N(\varepsilon, \mathcal{P}, d) \leq N_{[]}(\varepsilon, \mathcal{P}, d).$$

The logarithm of the packing (or covering) number is often called the (metric) entropy and that of the bracketing number is called the bracketing entropy. For more details, we refer the readers to Kolmogorov and Tihomirov (1961) and van der Vaart and Wellner (1996).

In the next section, we estimate packing numbers and bracketing numbers of these classes of densities. These results are of independent interest as they provide useful “size estimates” of these important classes of densities, unavailable in the literature so far. Bounds for the bracketing numbers have a number of implications for the convergence rates of the MLE and sieve MLEs vide the results of Wong and Shen (1995), van de Geer (1993) and Birgé and Massart (1998). Using these results, we obtain, in Section 4, the convergence rates of the MLE and sieve MLEs.

Bounds on bracketing numbers also allow us to construct certain priors based on finite approximating sets such that the posterior distributions are guaranteed to converge at a certain rate depending on the bounds; see Section 3 of Ghosal, Ghosh and van der Vaart (2000).

To compute the posterior rate of convergence, we shall use Theorem 2.1 of Ghosal, Ghosh and van der Vaart (2000) stated below in a slightly different way to exactly suit our purpose.

Let  $p_0 \in \mathcal{P}$ , a class of densities and let  $P_0$  be the probability measure with density  $p_0$ . Put  $K(p_0, p) = \int \log(p_0/p) dP_0$ ,  $V(p_0, p) = \int (\log(p_0/p))^2 dP_0$ ,  $B(\varepsilon, p_0) = \{p : K(p_0, p) \leq \varepsilon^2, V(p_0, p) \leq \varepsilon^2\}$ . Henceforth,  $d$  will stand for the Hellinger distance.

**THEOREM 2.1.** *Let  $\Pi_n$  be a sequence of priors on  $\mathcal{P}$ . Suppose that for positive sequences  $\bar{\varepsilon}_n, \tilde{\varepsilon}_n \rightarrow 0$  with  $n \min(\bar{\varepsilon}_n^2, \tilde{\varepsilon}_n^2) \rightarrow \infty$ , constants  $c_1, c_2, c_3, c_4 > 0$  and sets  $\mathcal{P}_n \subset \mathcal{P}$ , we have*

$$(2.8) \quad \log D(\bar{\varepsilon}_n, \mathcal{P}_n, d) \leq c_1 n \bar{\varepsilon}_n^2,$$

$$(2.9) \quad \Pi_n(\mathcal{P} \setminus \mathcal{P}_n) \leq c_3 e^{-(c_2+4)n\bar{\varepsilon}_n^2},$$

$$(2.10) \quad \Pi_n(B(\tilde{\varepsilon}_n, p_0)) \geq c_4 e^{-c_2 n \tilde{\varepsilon}_n^2}.$$

*Then for  $\varepsilon_n = \max(\bar{\varepsilon}_n, \tilde{\varepsilon}_n)$  and a sufficiently large  $M > 0$ , the posterior probability*

$$(2.11) \quad \Pi_n(p : d(p, p_0) > M\varepsilon_n | X_1, \dots, X_n) \rightarrow 0$$

*in  $P_0^n$ -probability.*

In what follows, we show that the conditions of the above theorem are satisfied by  $\bar{\varepsilon}_n = (\log n)^\kappa/\sqrt{n}$  and  $\tilde{\varepsilon}_n = (\log n)/\sqrt{n}$  for the Dirichlet mixture prior for a suitable power  $\kappa \geq 1$ . For that, apart from the estimates of packing numbers for the class of normal mixtures, we further need to estimate the prior probability of  $B(\varepsilon, p_0)$ .

In Sections 4, we estimate this prior probability for the location mixture and obtain the convergence rate of the posterior using Theorem 2.1. Analogous results for the location-scale mixtures are presented in Section 5. Note that convergence of the posterior at rate  $\varepsilon_n$  also implies that there exist estimators, the posterior mean for instance, that converge at the rate  $\varepsilon_n$  in the frequentist sense. See Theorem 2.5 of Ghosal, Ghosh and van der Vaart (2000) and the discussion following that for details.

The symbol “ $\lesssim$ ” will be used throughout to denote inequality up to a constant multiple where the value of the constant is fixed within our set-up. The symbol  $a \sim b$  will stand for  $a/b \rightarrow 1$ .

**3. Entropy estimates.** As mentioned in the introduction and the last section, estimates of packing and bracketing numbers are essential for the computation of rates for both the posterior distribution and the MLE and sieve MLEs. In this section, we provide such estimates for the family of normal mixtures defined by (2.1), (2.2) and (2.3).

Let  $\|\cdot\|_\infty$  and  $\|\cdot\|_1$  stand for the supremum and the  $L_1$ -norm respectively and  $d$  for the Hellinger distance. Let  $\mathfrak{M}(S)$  stand for the set of all probability measures on a given set  $S$ .

**3.1. Location mixtures.** First, we consider location mixtures. The following theorem gives the estimates of entropies and bracketing entropies.

**THEOREM 3.1.** *Let  $\mathcal{F}_a^1 = \{p_{F,\sigma} : F \in \mathfrak{M}[-a, a], \underline{\sigma} \leq \sigma \leq \bar{\sigma}\}$ , where  $a \leq L(\log \frac{1}{\varepsilon})^\gamma$  and  $\gamma \geq \frac{1}{2}$ . Then for  $0 < \varepsilon < \frac{1}{2}$ ,*

$$(3.1) \quad \log N(\varepsilon, \mathcal{F}_a^1, \|\cdot\|_\infty) \lesssim \left(\log \frac{1}{\varepsilon}\right)^{2\gamma+1},$$

$$(3.2) \quad \log N_{[\cdot]}(\varepsilon, \mathcal{F}_a^1, \|\cdot\|_1) \lesssim \left(\log \frac{1}{\varepsilon}\right)^{2\gamma+1}$$

and

$$(3.3) \quad \log N_{[\cdot]}(\varepsilon, \mathcal{F}_a^1, d) \lesssim \left(\log \frac{1}{\varepsilon}\right)^{2\gamma+1}.$$

The key idea behind the proof of (3.1) is to get hold of a finitely supported mixing distribution with sufficiently restricted number of support points such that the corresponding normal mixture uniformly approximates a given normal mixture. Such a finitely supported mixing distribution may be found by matching a certain number of moments of the given mixing distribution with that of the finitely supported mixing distribution. This is done in the next lemma. The same idea will be used in the next section to estimate the prior probabilities of the Kullback-Leibler type balls. It may be mentioned here that the naive choice of the mixing distribution with equally spaced support points will need many more points for the same degree of approximation.

**LEMMA 3.1.** *Let  $0 < \varepsilon < \frac{1}{2}$  be given and  $|\sigma - \sigma'| < \varepsilon$ . For any probability measure  $F$  on an interval  $[-a, a]$ , where  $a \leq L(\log \frac{1}{\varepsilon})^\gamma$  and  $\gamma \geq \frac{1}{2}$  and  $L > 0$  are constants, there exists a discrete probability measure  $F'$  on  $[-a, a]$  with at most  $N \lesssim (\log \frac{1}{\varepsilon})^{2\gamma}$  support points in  $[-a, a]$  such that*

$$(3.4) \quad \|p_{F,\sigma} - p_{F',\sigma'}\|_\infty \lesssim \varepsilon.$$

**PROOF.** Since  $\|\phi_\sigma - \phi_{\sigma'}\|_\infty \lesssim |\sigma - \sigma'|$ , it easily follows that

$$(3.5) \quad \|p_{F,\sigma} - p_{F',\sigma'}\|_\infty \lesssim |\sigma - \sigma'|$$

for any probability measure  $F$ . For  $M = \max(2a, \sqrt{8\bar{\sigma}}(\log \frac{1}{\varepsilon})^{1/2})$ , we have for any probability  $F$  on  $[-a, a]$ ,

$$\begin{aligned} \sup_{|x| \geq M} |p_{F,\sigma}(x) - p_{F',\sigma'}(x)| &\leq 2\phi_\sigma(M-a) \\ &\leq 2\sigma^{-1}\phi(M/2\sigma) \\ &\leq 2^{1/2}\pi^{-1/2}\underline{\sigma}^{-1}\exp[-M^2/(8\bar{\sigma}^2)] \\ &\leq 2^{1/2}\pi^{-1/2}\underline{\sigma}^{-1}\varepsilon, \end{aligned}$$

so that

$$(3.6) \quad \sup_{|x| \geq M} |p_{F,\sigma}(x) - p_{F',\sigma'}(x)| \lesssim \varepsilon.$$

By Taylor's expansion of  $e^y$  and  $k! \geq k^k e^{-k}$ , we have for any  $y < 0$ ,  $k > 1$ ,

$$\left| e^y - \sum_{j=1}^{k-1} \frac{y^j}{j!} \right| \leq \frac{|y|^k}{k!} \leq \frac{(e|y|)^k}{k^k}.$$

Putting  $y = -x^2/2$  and expanding  $\phi_\sigma(x)$ , we obtain

$$(3.7) \quad \left| \phi_\sigma(x) - \sum_{j=0}^{k-1} \frac{(-1)^j \sigma^{-(2j+1)} x^{2j}}{\sqrt{2\pi} j!} \right| \leq \sigma^{-1} \frac{(e^{1/2} 2^{-1/2} \sigma^{-1} |x|)^{2k}}{\sqrt{2\pi} k^k},$$

and hence for any  $F \in \mathfrak{M}[-a, a]$ ,

$$(3.8) \quad \begin{aligned} & \sup_{|x| \leq M} |p_{F, \sigma}(x) - p_{F', \sigma}(x)| \\ & \leq \sup_{|x| \leq M} \left| \int \sum_{j=0}^{k-1} (2\pi)^{-1/2} \frac{(-1)^j \sigma^{-(2j+1)} (x-z)^{2j}}{j!} d(F - F')(z) \right| \\ & \quad + 2 \sup_{\substack{|x| \leq M \\ |z| \leq a}} \left| \phi_\sigma(x-z) - \sum_{j=0}^{k-1} (2\pi)^{-1/2} \frac{(-1)^j \sigma^{-(2j+1)} (x-z)^{2j}}{j!} \right| \\ & = \sup_{|x| \leq M} \left| \int \sum_{j=0}^{k-1} \sum_{l=0}^{2j} (2\pi)^{-1/2} \frac{(-1)^j \sigma^{-(2j+1)} \binom{2j}{l} x^{2j-l} z^l}{j!} d(F - F')(z) \right| \\ & \quad + 2 \sup_{\substack{|x| \leq M \\ |z| \leq a}} \left| \phi_\sigma(x-z) - \sum_{j=0}^{k-1} (2\pi)^{-1/2} \frac{(-1)^j \sigma^{-(2j+1)} (x-z)^{2j}}{j!} \right|. \end{aligned}$$

If

$$(3.9) \quad \int z^l dF(z) = \int z^l dF'(z), \quad l = 1, \dots, 2k - 2,$$

then the first term on the right hand side (RHS) of (3.8) vanishes. If  $|x| \leq M$  and  $|z| \leq a$ , then

$$(3.10) \quad |x - z| \leq M + a \leq \frac{3M}{2} \leq \max(3L, \sqrt{18\sigma}) \left( \log \frac{1}{\varepsilon} \right)^\gamma.$$

Therefore, with  $c = e^{1/2} 2^{-1/2} \sigma^{-1} \max(3L, \sqrt{18\sigma})$ , the second term on the RHS of (3.8) is bounded by a constant multiple of

$$(3.11) \quad \frac{(c(\log \frac{1}{\varepsilon})^\gamma)^{2k}}{k^k} = \exp[-k(\log k - 2 \log(c(\log \varepsilon^{-1})^\gamma))].$$

Clearly the bound decreases as  $k$  increases. If we choose  $k$  to be the smallest integer exceeding  $(1 + c^2)(\log \frac{1}{\varepsilon})^{2\gamma}$ , it follows that

$$(3.12) \quad \sup_{|x| \leq M} |p_{F, \sigma}(x) - p_{F', \sigma}(x)| \lesssim \varepsilon.$$

By Lemma A.1,  $F'$  can be chosen to be a discrete distribution on  $[-a, a]$  with at most  $N = 2k - 1$  support points. The result now follows by combining (3.5), (3.6) and (3.12).  $\square$

PROOF OF (3.1). Choose an  $\varepsilon$ -net  $\Sigma$  for  $[\underline{\sigma}, \bar{\sigma}]$ . Let  $\mathcal{J}$  be the set of all  $p_{F,\sigma}$  such that  $\sigma$  comes from  $\Sigma$  and  $F$  has at most  $N \leq D(\log \frac{1}{\varepsilon})^{2\gamma}$  support points in  $[-a, a]$ , where  $D$  is a constant. By Lemma 3.1,  $D$  can be so chosen that  $\mathcal{J}$  is an  $\varepsilon$ -net over  $\mathcal{F}_a^1$ . Thus an  $\varepsilon$ -net over  $\mathcal{J}$  is a  $2\varepsilon$ -net over  $\mathcal{F}_a^1$ . Choose and fix an  $\varepsilon$ -net  $\mathcal{S}$  over the  $N$ -dimensional simplex for the  $\ell_1$ -norm. By Lemma A.4 of the Appendix, this can be chosen in such a way that the cardinality of  $\mathcal{S}$  does not exceed  $(5/\varepsilon)^N$ . Let  $\mathcal{J}'$  be the set of all  $p_{F,\sigma} \in \mathcal{J}$  such that  $F$  is supported on  $0, \pm\varepsilon, \pm 2\varepsilon, \dots$  with weights coming from  $\mathcal{S}$  only. A given  $p_{F,\sigma}$  can be “projected” into  $\mathcal{J}'$  by first moving the point masses of  $F$  to the closest point in  $0, \pm\varepsilon, \pm 2\varepsilon, \dots$ , and next changing the vector of sizes of point masses to the closest vector in  $\mathcal{S}$ . The new  $p_{F,\sigma}$ 's obtained this way are respectively less than  $\varepsilon\|\phi'_\sigma\|_\infty$  and  $\varepsilon\|\phi_\sigma\|_\infty$  away from their starting points. Thus any  $p_{F,\sigma}$  is within distance  $\varepsilon\|\phi_\sigma\|_\infty + \varepsilon\|\phi'_\sigma\|_\infty \lesssim \varepsilon$  of some element of  $\mathcal{J}'$ . Now the cardinality of  $\mathcal{J}'$  can be bounded as

$$(3.13) \quad \#\mathcal{J}' \lesssim \frac{1}{\varepsilon} \times \left(\frac{2a}{\varepsilon}\right)^N \times \left(\frac{5}{\varepsilon}\right)^N = (10a)^N \varepsilon^{-(2N+1)},$$

and so for some constants  $c_1$  and  $c_2$ ,

$$(3.14) \quad \begin{aligned} & \log N(c_1\varepsilon, \mathcal{F}_a^1, \|\cdot\|_\infty) \\ & \leq N \log(10a) + (2N + 1) \left(\log \frac{1}{\varepsilon}\right) + c_2 \\ & \leq D \left(\log \frac{1}{\varepsilon}\right)^{2\gamma} \left(\log \left(10L \left(\log \frac{1}{\varepsilon}\right)^\gamma\right) + 2 \log \frac{1}{\varepsilon}\right) + \log \frac{1}{\varepsilon} + c_2 \\ & \lesssim \left(\log \frac{1}{\varepsilon}\right)^{2\gamma+1}. \end{aligned}$$

The result follows.  $\square$

In the transition from the  $L_\infty$  metric  $\|\cdot\|_\infty$  to the variation and the Hellinger metrics  $\|\cdot\|_1$  and  $d$ , the following lemma will be helpful.

LEMMA 3.2. *For any two probability measures  $F, F'$  on  $[-a, a]$  where  $a > 0$  is arbitrary and any  $\sigma, \sigma' \in [\underline{\sigma}, \bar{\sigma}]$ ,*

$$(3.15) \quad \begin{aligned} & \|p_{F,\sigma} - p_{F',\sigma'}\|_1 \\ & \lesssim \|p_{F,\sigma} - p_{F',\sigma'}\|_\infty \max \left\{ \sqrt{\log_+ \left(\frac{1}{\|p_{F,\sigma} - p_{F',\sigma'}\|_\infty}\right)}, a, 1 \right\}. \end{aligned}$$

PROOF. Since the  $L_1$ -norm between any two densities is bounded by 2, if  $\|p_{F,\sigma} - p_{F',\sigma'}\|_\infty \geq 1$ , the inequality

$$(3.16) \quad \|p_{F,\sigma} - p_{F',\sigma'}\|_1 \leq 2\|p_{F,\sigma} - p_{F',\sigma'}\|_\infty$$

holds trivially. We may therefore assume that  $\|p_{F,\sigma} - p_{F',\sigma'}\|_\infty < 1$ . For any  $T \geq 2a$  and  $|x| > T$ ,

$$(3.17) \quad p_{F,\sigma}(x) \leq \phi_\sigma(|x| - a) \leq \phi_\sigma(x/2) \leq \underline{\sigma}^{-1} e^{-x^2/8\bar{\sigma}^2}.$$

Hence  $\|p_{F,\sigma} - p_{F',\sigma'}\|_1$  is bounded by

$$\begin{aligned} & 2 \int_{|x|>T} \underline{\sigma}^{-1} e^{-x^2/8\bar{\sigma}^2} dx + 2T \|p_{F,\sigma} - p_{F',\sigma'}\|_\infty \\ & \lesssim e^{-T^2/8\bar{\sigma}^2} + T \|p_{F,\sigma} - p_{F',\sigma'}\|_\infty. \end{aligned}$$

For the choice

$$T = \max \left( \bar{\sigma} \sqrt{8 \log \left( \frac{1}{\|p_{F,\sigma} - p_{F',\sigma'}\|_\infty} \right)}, 2a \right),$$

the first term is bounded by  $\|p_{F,\sigma} - p_{F',\sigma'}\|_\infty$ , while the second term is bounded by a multiple of  $\|p_{F,\sigma} - p_{F',\sigma'}\|_\infty \max \left( \sqrt{\log \left( \frac{1}{\|p_{F,\sigma} - p_{F',\sigma'}\|_\infty} \right)}, a \right)$ . The result follows.  $\square$

Now we are ready to prove the remaining parts of Theorem 3.1.

PROOFS OF (3.2) AND (3.3). Let  $\varepsilon > 0$  be given and let  $\eta \leq \varepsilon$  to be chosen later. Note that clearly  $a \leq L(\log \frac{1}{\eta})^\gamma$ . Let  $f_1, \dots, f_N$  be an  $\eta$ -net for  $\|\cdot\|_\infty$  over  $\mathcal{F}_a^1$ . For any  $p_{F,\sigma} \in \mathcal{F}_a^1$ , we have

$$0 \leq p_{F,\sigma}(x) \leq \begin{cases} \underline{\sigma}^{-1} \phi(0), & \text{for all } x, \\ \underline{\sigma}^{-1} \phi \left( \frac{x}{2\bar{\sigma}} \right), & \text{if } |x| > 2a. \end{cases}$$

Thus

$$H(x) = \begin{cases} \underline{\sigma}^{-1} \phi \left( \frac{x}{2\bar{\sigma}} \right), & \text{if } |x| > 2a, \\ \underline{\sigma}^{-1} \phi(0), & \text{otherwise,} \end{cases}$$

is an envelope for  $\mathcal{F}_a^1$ . Construct brackets  $[l_i, u_i]$  by setting

$$l_i = \max(f_i - \eta, 0), \quad u_i = \min(f_i + \eta, H).$$

Then clearly  $\mathcal{F}_a^1 \subset \cup_i [l_i, u_i]$  and  $u_i - l_i \leq \min(2\eta, H)$ . Therefore for any  $B > 0$ ,

$$(3.18) \quad \int (u_i(x) - l_i(x)) dx \leq \int_{|x| \leq B} 2\eta dx + \int_{|x| > B} H(x) dx.$$

Choose  $B = \max(2L, \sqrt{8\bar{\sigma}})(\log \frac{1}{\eta})^\gamma$ . Then  $B \geq 2a$  and by Mill's ratio,

$$\int_{|x| > B} H(x) dx \lesssim H(B) \leq \eta.$$

The first term on the RHS of (3.18) is therefore bounded by a constant multiple of  $\eta(\log \frac{1}{\eta})^\gamma$ . Thus for some constant  $C$ ,

$$N_{[\cdot]} \left( C\eta \left( \log \frac{1}{\eta} \right)^\gamma, \mathcal{F}_a^1, \|\cdot\|_1 \right) \leq N.$$

By (3.1), we can choose  $\log N \lesssim (\log \frac{1}{\eta})^{2\gamma+1}$ . Choosing  $\eta$  the solution of  $C\eta(\log \frac{1}{\eta})^\gamma = \varepsilon$  and noting that  $\log \frac{1}{\eta} \sim \log \frac{1}{\varepsilon}$ , we obtain (3.2). Therefore by (2.4),

$$(3.19) \quad \log N_{[\cdot]}(\varepsilon, \mathcal{F}_a^1, d) \leq \log N_{[\cdot]}(\varepsilon^2, \mathcal{F}_a^1, \|\cdot\|_1) \lesssim \left( \log \frac{1}{\varepsilon} \right)^{2\gamma+1}.$$

**3.2. Location-scale mixtures.** We now present the analogous estimates of the metric and bracketing entropies for the class of location-scale mixtures defined by (2.2) and (2.3).

**THEOREM 3.2.** *Let  $\mathcal{F}_a^2 = \{p_{F,G} : F \in \mathfrak{M}[-a, a], G \in \mathfrak{M}[\underline{\sigma}, \bar{\sigma}]\}$ , where  $a \leq L(\log \frac{1}{\varepsilon})^\gamma$  and  $\gamma \geq \frac{1}{2}$ . Then for  $0 < \varepsilon < \frac{1}{2}$ ,*

$$(3.20) \quad \log N(\varepsilon, \mathcal{F}_a^2, \|\cdot\|_\infty) \lesssim \left( \log \frac{1}{\varepsilon} \right)^{2\gamma+1},$$

$$(3.21) \quad \log N_{[\cdot]}(\varepsilon, \mathcal{F}_a^2, \|\cdot\|_1) \lesssim \left( \log \frac{1}{\varepsilon} \right)^{2\gamma+1}$$

and

$$(3.22) \quad \log N_{[\cdot]}(\varepsilon, \mathcal{F}_a^2, d) \lesssim \left( \log \frac{1}{\varepsilon} \right)^{2\gamma+1}.$$

As in the family location mixtures, the key step in the proof is a uniform approximation by a discretely supported mixture with sufficiently restricted number of support points. The following result gives us such an approximation in the spirit of Lemma 3.1

**LEMMA 3.3.** *Let  $0 < \varepsilon < \frac{1}{2}$  be given. For any probability measure  $F \in \mathfrak{M}[-a, a]$ , where  $a \leq L(\log \frac{1}{\varepsilon})^\gamma$  and  $\gamma \geq \frac{1}{2}$  and  $L > 0$  are constants, and  $G \in \mathfrak{M}[\underline{\sigma}, \bar{\sigma}]$ , there exist discrete probability measures  $F'$  on  $[-a, a]$  and  $G'$  on  $[\underline{\sigma}, \bar{\sigma}]$  with at most  $N \lesssim (\log \frac{1}{\varepsilon})^{2\gamma}$  support points in  $[-a, a]$  and  $[\underline{\sigma}, \bar{\sigma}]$  respectively, such that*

$$(3.23) \quad \|p_{F,G} - p_{F',G'}\|_\infty \lesssim \varepsilon.$$

**PROOF.** For any choices of  $F' \in \mathfrak{M}[-a, a]$  and  $G' \in \mathfrak{M}[\underline{\sigma}, \bar{\sigma}]$ , we have, as in (3.6),

$$(3.24) \quad \sup_{|x| \geq M} |p_{F,G}(x) - p_{F',G'}(x)| \lesssim \varepsilon.$$

If probability distributions  $F'$  and  $G'$ , respectively on  $[-a, a]$  and  $[\underline{\sigma}, \bar{\sigma}]$ , are chosen to satisfy

$$(3.25) \quad \int z^l dF(z) = \int z^l dF'(z), \quad l = 1, \dots, 2k - 2$$

and

$$(3.26) \quad \int \sigma^{-(2j+1)} dG(\sigma) = \int \sigma^{-(2j+1)} dG'(\sigma), \quad j = 0, \dots, k - 1,$$

where  $k$  is the smallest integer exceeding  $(1+c^2)(\log \frac{1}{\varepsilon})^{2\gamma}$  and  $c$  is the constant  $e^{1/2}2^{-1/2}\underline{\sigma}^{-1} \max(3L, \sqrt{18\bar{\sigma}})$ , then by the estimate (3.7) and the arguments given in the proof of Lemma 3.1,  $\|p_{F,G} - p_{F',G'}\|_\infty \lesssim \varepsilon$ . Applying Lemma A.1 of the Appendix to  $z^l$ ,  $l = 0, \dots, 2k - 2$  and  $F$  on  $[-a, a]$  and to  $\sigma^{-(2j+1)}$ ,  $j = 1, \dots, k - 1$  and  $G$  on  $[\underline{\sigma}, \bar{\sigma}]$ , we may restrict the number of support points of  $F'$  and  $G'$  to  $N = 2k - 1 \lesssim (\log \frac{1}{\varepsilon})^{2\gamma}$ .  $\square$

PROOF OF THEOREM 3.2. Consider the class of densities  $\mathcal{J}$  consisting of all  $p_{F,G}$  where  $F$  and  $G$  have at most  $N \lesssim (\log \frac{1}{\varepsilon})^{2\gamma}$  support points from  $[-a, a]$  of the form  $0, \pm\varepsilon, \pm 2\varepsilon, \dots$ , and from  $[\underline{\sigma}, \bar{\sigma}]$  of the form  $\varepsilon, 2\varepsilon, \dots$  respectively, and the weight corresponding to each point of support of  $F$  and  $G$  comes from a chosen  $\varepsilon$ -net over the  $N$ -simplex for the  $\ell_1$ -norm. Then by Lemma 3.3) and the arguments given in the proof of (3.1),  $\mathcal{J}$  is a  $k\varepsilon$ -net over  $\mathcal{F}_a^2$  for some fixed constant  $k$ . The cardinality of  $\mathcal{J}$  is bounded by a constant times

$$\begin{aligned} & \left(\frac{\bar{\sigma} - \underline{\sigma}}{\varepsilon}\right)^N \times (f2a\varepsilon)^N \times \left(\frac{5}{\varepsilon}\right)^N \times \left(\frac{5}{\varepsilon}\right)^N \\ & = (50a(\bar{\sigma} - \underline{\sigma}))^N \varepsilon^{-4N}. \end{aligned}$$

Thus (3.20) follows. Now observe that as in Lemma 3.2, if  $F$  and  $F'$  are probability measures on  $[-a, a]$  and  $G$  and  $G'$  are probability measures on  $[\underline{\sigma}, \bar{\sigma}]$ , then

$$(3.27) \quad \begin{aligned} & \|p_{F,G} - p_{F',G'}\|_1 \\ & \lesssim \|p_{F,G} - p_{F',G'}\|_\infty \max \left\{ \sqrt{\log_+ \left( \frac{1}{\|p_{F,G} - p_{F',G'}\|_\infty} \right)}, a, 1 \right\}. \end{aligned}$$

The rest of the proof can be completed as in that of Theorem 3.1.  $\square$

The following result gives entropy estimates for the general location-scale mixtures defined by (2.3). Note that the entropy estimates are weaker than the corresponding bounds in Theorem 3.2.

THEOREM 3.3. Let  $\mathcal{F}_a^3 = \{p_H : H \in \mathfrak{M}([-a, a] \times [\underline{\sigma}, \bar{\sigma}])\}$ , where  $a \leq L(\log \frac{1}{\varepsilon})^\gamma$  and  $\gamma \geq \frac{1}{2}$ . Then for  $0 < \varepsilon < \frac{1}{2}$ ,

$$(3.28) \quad \log N(\varepsilon, \mathcal{F}_a^3, \|\cdot\|_\infty) \lesssim \left(\log \frac{1}{\varepsilon}\right)^{4\gamma+1},$$

$$(3.29) \quad \log N_{[\cdot]}(\varepsilon, \mathcal{F}_a^3, \|\cdot\|_1) \lesssim \left(\log \frac{1}{\varepsilon}\right)^{4\gamma+1}$$

and

$$(3.30) \quad \log N_{[\cdot]}(\varepsilon, \mathcal{F}_a^3, d) \lesssim \left(\log \frac{1}{\varepsilon}\right)^{4\gamma+1}.$$

To prove the theorem, we need the following analogue of Lemma 3.3.

LEMMA 3.4. *Let  $0 < \varepsilon < \frac{1}{2}$  be given. For any probability measure  $H \in \mathfrak{M}([-a, a] \times [\underline{\sigma}, \bar{\sigma}])$ , where  $a \leq L(\log \frac{1}{\varepsilon})^\gamma$  and  $\gamma \geq \frac{1}{2}$  and  $L > 0$  are constants, there exists a discrete probability measure  $H'$  on  $[-a, a] \times [\underline{\sigma}, \bar{\sigma}]$  with at most  $N \lesssim (\log \frac{1}{\varepsilon})^{4\gamma}$  support points in  $[-a, a] \times [\underline{\sigma}, \bar{\sigma}]$ , such that*

$$(3.31) \quad \|p_H - p_{H'}\|_\infty \lesssim \varepsilon.$$

PROOF. Applying Lemma A.1 of the Appendix to  $[-a, a] \times [\underline{\sigma}, \bar{\sigma}]$ , find discrete distributions  $H'$  on  $[-a, a] \times [\underline{\sigma}, \bar{\sigma}]$  with at most  $N = k(2k - 1) + 1$  support points such that

$$(3.32) \quad \int z^l \sigma^{-(2j+1)} dH(z) = \int z^l \sigma^{-(2j+1)} dH'(z),$$

for all  $l = 0, \dots, 2k - 2$  and  $j = 0, \dots, k - 1$ . The rest of the proof is almost identical to that of Lemma 3.3.  $\square$

PROOF OF THEOREM 3.3. Consider the class of densities  $\mathcal{J}$  consisting of all  $p_H$  where  $H$  has at most  $N$  support points from  $[-a, a] \times [\underline{\sigma}, \bar{\sigma}]$  of the form  $(\pm k\varepsilon, l\varepsilon)$ , where  $k, l = 0, 1, \dots$  respectively, and the weight corresponding to each point of support of  $H$  comes from a chosen  $\varepsilon$ -net over the  $N$ -simplex for the  $\ell_1$ -norm, where  $N \lesssim (\log \frac{1}{\varepsilon})^{4\gamma}$ . Then, as before,  $\mathcal{J}$  is a  $k\varepsilon$ -net over  $\mathcal{F}_a^3$  for some fixed constant  $k$  and the cardinality of  $\mathcal{J}$  is bounded by a constant times  $(\log \frac{1}{\varepsilon})^{4\gamma+1}$ , proving (3.28). The rest of the proof can be completed as before.  $\square$

**4. Maximum likelihood and sieve methods.** The estimate of  $\varepsilon$ -bracketing Hellinger entropy obtained in Theorems 3.1–3.3 allows us to compute an upper bound on the rate of convergence of the MLE and sieve MLEs using the results of Wong and Shen (1995). Alternatively, one can also apply Theorem 3.4.4 of van der Vaart and Wellner (1996).

Let  $\hat{p}$  be an MLE, that is, a measurable function of the observations taking values in  $\mathcal{P}$  such that

$$n^{-1} \sum_{i=1}^n \log \hat{p}(X_i) \geq \sup_{p \in \mathcal{P}} n^{-1} \sum_{i=1}^n \log p(X_i).$$

It is known that the MLE exists; see Lindsay (1995), Theorem 18.

In the following theorems, we consider three possible models  $\mathcal{P}$ , deriving from (2.1), (2.2) and (2.3), respectively. In model (2.1), let  $F$  be any probability measure supported on  $[-a, a]$  where  $a \lesssim (\log n)^\gamma$  and  $\sigma$  take any arbitrary value on the interval  $[\underline{\sigma}, \bar{\sigma}]$ . For model (2.2), we assume that  $F$  is as above and the distribution  $G$  of  $\sigma$  is supported on  $[\underline{\sigma}, \bar{\sigma}]$ . In model (2.3), the distribution  $H$  for  $(z, \sigma)$  is supported on  $[-a, a] \times [\underline{\sigma}, \bar{\sigma}]$ , where  $a$  is as above. Assume that the true  $F_0$  has compact support in case of model (2.1) and (2.2); for model (2.3),  $H_0$  is assumed to have compact support.

THEOREM 4.1. *Under the above set-up, for a sufficiently large constant  $M$ ,*

$$(4.1) \quad P_0(d(\hat{p}, p_0) > M\varepsilon_n) \lesssim e^{-c(\log n)^2},$$

where  $\varepsilon_n = (\log n)^{\max(\gamma, \frac{1}{2}) + \frac{1}{2}} / \sqrt{n}$  for models (2.1) and (2.2) while we have  $\varepsilon_n = (\log n)^{2\max(\gamma, \frac{1}{2}) + \frac{1}{2}} / \sqrt{n}$  for model (2.3), and  $c$  is a constant. In particular,  $\hat{p}$  converges to  $p_0$  in Hellinger distance at a rate  $\varepsilon_n$  in  $P_0$ -probability, and  $P_0$ -almost surely.

Clearly, the best rates  $\varepsilon_n = (\log n) / \sqrt{n}$  for models (2.1) and (2.2) and  $\varepsilon_n = (\log n)^{3/2} / \sqrt{n}$  for model (2.3), are obtained by choosing  $\gamma \leq \frac{1}{2}$ . Genovese and Wasserman (2000) considered convergence rates of sieve MLEs for location mixtures only, where again the scale is restricted to lie between two positive numbers, and obtained the much weaker rate  $n^{-1/6}(\log n)^{(1/6)+\delta}$  of convergence of the (sieve) MLE for some  $\delta > 0$ .

To prove the above theorem, we apply Theorem 2 of Wong and Shen (1995) by noting that

$$(4.2) \quad \int_0^{\varepsilon_n} \sqrt{\log N_{[]} (u, \mathcal{P}, d)} du \lesssim \sqrt{n} \varepsilon_n^2$$

for all the models with the appropriate  $\varepsilon_n$  by the estimates of the bracketing entropy shown in Theorems 3.1–3.3.

Note that although the true mixing measure  $F_0$  is supported on an interval  $[-k_0, k_0]$  (in case of (2.3),  $H_0$  is supported in  $[-k_0, k_0] \times [\underline{\sigma}, \bar{\sigma}]$ ), it is not necessary to know  $k_0$  since we can increase  $a$  to infinity. One, however, needs to know an interval where the possible values of  $\sigma$  will lie.

Since the MLE exists and converges at the desired rate, it is not necessary to restrict the maximization to a sieve if  $F_0$  has a known compact support. A suitable sieve may, however, give a simpler and equally efficient estimator. Priebe (1994) considered the sieve where the mixing distributions are finitely supported and argued that often it is possible to estimate the density with a relatively small number of normal components. It is known that [Lindsay (1995), Theorem 21] the MLE is a discrete distribution supported on at most  $n$  points. The following theorem shows that restricting the maximization to the sieve of all discrete distributions with at most  $C \log n$  support points, where  $C$  is a sufficiently large constant, we obtain the same rate of convergence. However, we believe it is a reasonable conjecture that the full MLE has of the

order of  $\log n$  support points. Thus the resulting estimators in Theorem 4.1 and 4.2 may not be different.

For model (2.1), let  $\hat{p}_k$  be the maximizer of the likelihood on  $\{p_{F,\sigma} : F = \sum_{j=1}^k p_j \delta_{z_j}, p_j \geq 0, \sum_{j=1}^k p_j = 1, z_j \in [-a, a]\}$ . Define  $\hat{p}_k$  in model (2.2) similarly by restricting the mixing distributions  $F$  and  $G$  to have at most  $k$  support points, while in model (2.3), let  $H$  to be supported on  $k^2$  points. The number  $k = k_n$  will be allowed to grow with  $n$ . To compute the rate of convergence of  $\hat{p}_k$ , we apply part (ii) of Theorem 4 of Wong and Shen (1995). Since the sieve, being a subset of the parameter space, already meets the required entropy condition (4.2), we only need to check the approximation properties of the sieve in the Kullback-Leibler sense.

**THEOREM 4.2.** *If  $k \geq C \log n$  for some sufficiently large  $C$ , then for some  $M$ ,*

$$(4.3) \quad P_0(d(\hat{p}_k, p_0) > M\varepsilon_n) \rightarrow 0,$$

where  $\varepsilon_n$  is as in Theorem 4.1.

To prove the theorem, we need the following result which bounds  $K(\cdot, \cdot)$  and  $V(\cdot, \cdot)$  of two normal mixtures in terms of their Hellinger distance. Although we state and prove the result only for location mixtures, exactly analogous results hold for location-scale mixtures as well. This result will be useful in the study of the convergence rate of the posterior distribution also.

**LEMMA 4.1.** *If  $F[-B, B] > \frac{1}{2}$  for some constant  $B$  and  $F^*$  is a probability measure satisfying  $F^*(z : |z| > t) \lesssim e^{-b't^2}$  for some constant  $b' > 0$ , then for  $\varepsilon = d(p_{F^*,\sigma^*}, p_{F,\sigma}) < \frac{1}{2}$ ,*

$$(4.4) \quad \begin{aligned} K(p_{F^*,\sigma^*}, p_{F,\sigma}) &\lesssim \varepsilon^2 \log \frac{1}{\varepsilon} \\ V(p_{F^*,\sigma^*}, p_{F,\sigma}) &\lesssim \varepsilon^2 \left(\log \frac{1}{\varepsilon}\right)^2. \end{aligned}$$

**PROOF.** Note that

$$p_{F^*,\sigma^*}(x) \leq \frac{1}{\sigma} \phi(0)$$

and if  $F[-B, B] > \frac{1}{2}$ ,

$$p_{F,\sigma}(x) \geq \frac{1}{\sigma} \int_{-B}^B \phi\left(\frac{x-z}{\sigma}\right) dF(z) \geq \frac{1}{2\sigma} \phi\left(\frac{|x|+B}{\sigma}\right).$$

Therefore for some  $c$  (depending on  $B$ ),

$$(4.5) \quad \frac{p_{F^*,\sigma^*}(x)}{p_{F,\sigma}(x)} \lesssim e^{cx^2},$$

and so for some  $\delta > 0$ ,

$$(4.6) \quad \int \left( \frac{p_{F^*,\sigma^*}(x)}{p_{F,\sigma}(x)} \right)^\delta p_{F^*,\sigma^*}(x) dx < \infty.$$

Applying Theorem 5 of Wong and Shen (1995), we get the result.  $\square$

Let us indicate the proof of Theorem 4.2 only for model (2.1). Given  $\varepsilon > 0$  and  $\eta \leq \varepsilon$  to be chosen later, Lemma 3.1 shows that there exists a discrete distribution  $F'_0$  supported on  $[-k_0, k_0]$  with  $N \lesssim \log \frac{1}{\eta}$  support points such that  $\|p_{F_0,\sigma_0} - p_{F'_0,\sigma_0}\|_\infty \lesssim \eta$ . By Lemmas 3.2 and 4.1, we obtain

$$K(p_{F_0,\sigma_0}, p_{F'_0,\sigma_0}) \lesssim \eta \left( \log \frac{1}{\eta} \right)^{3/2}, \quad V(p_{F_0,\sigma_0}, p_{F'_0,\sigma_0}) \lesssim \eta \left( \log \frac{1}{\eta} \right)^{5/2}.$$

Choosing  $\eta$  the solution of  $\eta^{1/2}(\log \frac{1}{\eta})^{5/4} = \varepsilon$  and noting that  $\log \frac{1}{\varepsilon} \sim \log \frac{1}{\eta}$ , we see that  $F'_0$  has  $N \lesssim (\log \frac{1}{\varepsilon})^{2\gamma}$  support points and

$$(4.7) \quad K(p_{F_0,\sigma_0}, p_{F,\sigma_0}) \lesssim \varepsilon^2, \quad V(p_{F_0,\sigma_0}, p_{F,\sigma_0}) \lesssim \varepsilon^2.$$

Now apply Theorem 4 of Wong and Shen (1995) with  $\varepsilon = \varepsilon_n$  and  $k \geq N$  to obtain

$$(4.8) \quad P_0(d(\hat{p}_k, p_0) > M\varepsilon_n) \lesssim e^{-n\varepsilon_n^2} + \frac{1}{n} \rightarrow 0.$$

Theorems 4.1 and 4.2 remain valid even if the true mixing distribution  $F_0$  (or  $H_0$  for model (2.3)) is not compactly supported, provided that it has sub-Gaussian tails. We choose sieves as before with  $a \lesssim (\log \frac{1}{\varepsilon})^{1/2}$ , and using Lemma 4.1, we can show that some element from the sieve approximates the true density in the Kullback-Leibler sense as in (4.7). We omit the details; a similar theorem for the posterior distribution will be proved in the next section.

Following the construction in Example 4 of Wong and Shen (1995) and using the estimate of the bracketing entropy, we can construct a sieve consisting of finitely many densities which also gives the desired rate of convergence of the sieve-MLE. To this end, consider the sieve  $\mathcal{P}_n = \{g_1, \dots, g_N\}$ , where  $N = N_{[]}(\varepsilon_n, \mathcal{P}, d)$ ,  $\varepsilon_n$  is the solution of the entropy equation

$$(4.9) \quad \log N_{[]}(\varepsilon, \mathcal{P}, d) \leq n\varepsilon^2,$$

$g_j = u_j / \int u_j$ ,  $j = 1, \dots, N$ , and  $[l_1, u_1], \dots, [l_N, u_N]$  is a Hellinger bracketing for  $\mathcal{P}$  of size  $\varepsilon_n$ . If we choose  $a \lesssim (\log n)^{1/2}$ , then  $\varepsilon_n \lesssim (\log n) / \sqrt{n}$  for models (2.1) and (2.2),  $\varepsilon_n \lesssim (\log n)^{3/2} / \sqrt{n}$  for model (2.3) and the sieve MLE  $\hat{p}$  maximizing the likelihood on  $\mathcal{P}_n$  satisfies (4.1).

The above sieve  $\mathcal{P}_n$  can also be used to construct a prior for which the posterior converges at rate  $\varepsilon_n$ . We follow the construction in Theorem 3.1 of Ghosal, Ghosh and van der Vaart (2000). Put the uniform distribution  $\Pi_j$  on  $\mathcal{P}_j$  and consider the prior  $\Pi = \sum_{j=1}^\infty \lambda_j \Pi_j$ , where  $\lambda_j > 0$ ,  $\sum_{j=1}^\infty \lambda_j = 1$  and

$\log \lambda_j^{-1} = O(\log j)$  as  $j \rightarrow \infty$ . Alternatively, for a sample of size  $n$ , simply consider the prior  $\Pi_n$ . Then Theorem 3.1 of Ghosal, Ghosh and van der Vaart (2000) implies that the posterior converges at the intended rate  $\varepsilon_n$ .  $\square$

**5. Posterior convergence: Location mixtures.** In this section, we consider model (2.1) and the Dirichlet mixture of normal prior described in Section 2. We shall write  $\Pi$  for the prior. The following theorem gives the rate of convergence of the posterior assuming that the true mixing distribution is compactly supported.

**THEOREM 5.1.** *Assume that the true mixing measure  $F_0$  has compact support, i.e.,  $F_0[-k_0, k_0] = 1$  for some  $k_0$ . If the prior for  $\sigma$  has a continuous and positive density on an interval containing  $\sigma_0$ , the base measure  $\alpha$  has a continuous and positive density on an interval containing  $[-k_0, k_0]$  and satisfies the tail condition*

$$(5.1) \quad \alpha(|z| > t) \lesssim e^{-b|t|^\delta} \quad \text{for all } t > 0$$

and for some constants  $b > 0$ ,  $\delta > 0$ , then for a sufficiently large constant  $M$ ,

$$(5.2) \quad \Pi \left( p : d(p, p_0) > M \frac{(\log n)^\kappa}{\sqrt{n}} \mid X_1, \dots, X_n \right) \rightarrow 0$$

in  $P_0^n$ -probability, where  $\kappa = \max(\frac{2}{\delta}, \frac{1}{2}) + \frac{1}{2}$ .

**REMARK 5.1.** In the above theorem, the best rate  $(\log n)/\sqrt{n}$  is obtained when  $\delta$  can be chosen to be 4 or more. For instance, a compactly supported base measure will give rise to this rate. For the commonly used normal base measure, the preceding theorem with  $\delta = 2$  yields the rate  $(\log n)^{3/2}/\sqrt{n}$ .

We verify the conditions of Theorem 2.1 with  $\bar{\varepsilon}_n = (\log n)^\kappa/\sqrt{n}$  and  $\tilde{\varepsilon}_n = (\log n)/\sqrt{n}$ . Since we are interested in the rate only and not in constants, we may replace the packing number by the covering number in (2.8) in view of (2.6). The estimates of the covering number is given by Theorem 3.1. It remains to obtain an estimate of the prior probability to satisfy (2.9) and (2.10). The following lemma bounds the variation distance between a discrete normal mixture and another normal mixture and is instrumental in bounding the prior probabilities.

**LEMMA 5.1.** *Let  $F^* = \sum_{j=1}^N p_j \delta_{z_j}$  be a probability measure with  $|z_j - z_k| > 2\varepsilon$  for all  $j \neq k$ . Then for any probability measure  $F$  on  $\mathbb{R}$ ,*

$$(5.3) \quad \|p_{F, \sigma} - p_{F^*, \sigma^*}\|_1 \lesssim \varepsilon + |\sigma - \sigma^*| + \sum_{j=1}^N |F[z_j - \varepsilon, z_j + \varepsilon] - p_j|.$$

**PROOF.** Because  $\|\phi_\sigma - \phi_{\sigma^*}\| \lesssim |\sigma - \sigma^*|$ , for any  $F$ , we have

$$(5.4) \quad \|p_{F, \sigma} - p_{F, \sigma^*}\|_1 \lesssim |\sigma - \sigma^*|.$$

Now

$$\begin{aligned}
 |p_{F, \sigma^*}(x) - p_{F^*, \sigma^*}(x)| &\leq \int_{z: |z-z_j| > \varepsilon \forall j} \phi_{\sigma^*}(x-z) dF(z) \\
 (5.5) \qquad &+ \sum_{j=1}^N \int_{|z-z_j| \leq \varepsilon} |\phi_{\sigma^*}(x-z) - \phi_{\sigma^*}(x-z_j)| dF(z) \\
 &+ \sum_{j=1}^N \phi_{\sigma^*}(x-z_j) |F[z_j - \varepsilon, z_j + \varepsilon] - p_j|.
 \end{aligned}$$

Since  $\phi_{\sigma^*}(\cdot)$  integrates to one and

$$\|\phi_{\sigma^*}(\cdot - z) - \phi_{\sigma^*}(\cdot - z_j)\|_1 \leq \sqrt{2/\pi\sigma^*} |z - z_j|,$$

we obtain from (5.5) and Fubini's theorem that

$$\begin{aligned}
 (5.6) \qquad &\|p_{F, \sigma^*} - p_{F^*, \sigma^*}\|_1 \\
 &\leq F(z : |z - z_j| > \varepsilon \forall j) + \sqrt{2/\pi\sigma^*} \sum_{j=1}^N \int_{|z-z_j| \leq \varepsilon} |z - z_j| dF(z) \\
 &+ \sum_{j=1}^N |F[z_j - \varepsilon, z_j + \varepsilon] - p_j|.
 \end{aligned}$$

To bound the first term on the RHS of (5.6), note that since the intervals  $(z_j - \varepsilon, z_j + \varepsilon)$ 's are disjoint and the  $p_j$ 's add up to 1,

$$\begin{aligned}
 (5.7) \qquad &F(z : |z - z_j| > \varepsilon \forall j) = 1 - \sum_{j=1}^N F[z_j - \varepsilon, z_j + \varepsilon] \\
 &\leq \sum_{j=1}^N |F[z_j - \varepsilon, z_j + \varepsilon] - p_j|.
 \end{aligned}$$

The second term on the RHS of (5.6) is bounded by  $\sqrt{2/\pi\sigma^*} \varepsilon$ . The result follows by combining these assertions.  $\square$

We are now ready to prove the theorem.

PROOF OF THEOREM 5.1. We may assume without loss of generality that  $\delta \leq 4$ . Given  $\eta > 0$  and  $a$  satisfying  $a \leq L(\log \frac{1}{\eta})^{2/\delta}$ , where  $L$  is a constant, set  $\mathcal{F}_{a, \eta}^1 = \{p_{F, \sigma} : F[-a, a] \geq 1 - \eta, \underline{\sigma} \leq \sigma \leq \bar{\sigma}\}$  and put, as in Theorem 3.1,  $\mathcal{F}_a^1 = \{p_{F, \sigma} : F[-a, a] = 1\}$ . We estimate the  $\eta$ -entropy of the class  $\mathcal{F}_{a, \eta}$  as  $\eta \rightarrow 0$ . First observe that, by Lemma A.3 with  $A = [-a, a]$ ,

$$(5.8) \qquad N(3\eta, \mathcal{F}_{a, \eta}^1, \|\cdot\|_1) \leq N(\eta, \mathcal{F}_a^1, \|\cdot\|_1).$$

Now (3.2), together with (2.7), implies that  $\log N(\eta, \mathcal{F}_a^{-1}, \|\cdot\|_1) \lesssim (\log \frac{1}{\eta})^{(4/\delta)+1}$ . This and (5.8) together imply that

$$(5.9) \quad \log N(\eta, \mathcal{F}_{a,\eta}^{-1}, \|\cdot\|_1) \lesssim \left(\log \frac{1}{\eta}\right)^{(4/\delta)+1}.$$

Therefore, by (2.4),

$$(5.10) \quad \log N(\eta, \mathcal{F}_{a,\eta}^{-1}, d) \leq \log N(\eta^2, \mathcal{F}_{a,\eta}, \|\cdot\|_1) \lesssim \left(\log \frac{1}{\eta}\right)^{(4/\delta)+1}.$$

Next, to estimate  $\Pi(\mathcal{F}_{a,\eta}^c)$ , note that, because  $F([-a, a]^c)$  has a beta distribution with parameters  $\alpha([-a, a]^c)$  and  $\alpha[-a, a]$ , Chebyshev’s inequality implies that

$$(5.11) \quad \Pi(F : F[-a, a] < 1 - \eta) \leq \frac{1}{\alpha(\mathbb{R})\eta} \alpha([-a, a]^c) \lesssim \eta^{-1} e^{-ba^\delta}.$$

We now estimate  $\Pi(B(\varepsilon, p_{F_0, \sigma_0}))$  as  $\varepsilon \rightarrow 0$ . First, for given  $0 < \varepsilon < \frac{1}{2}$ , by Theorem 3.1 applied to  $F = F_0$  and  $a = k_0$ , we can find a discrete distribution  $F'_0$  (depending on  $\varepsilon$ ) on  $[-k_0, k_0]$  supported on at most  $C_1 \log \frac{1}{\varepsilon}$ , points such that  $\|p_{F_0, \sigma_0} - p_{F'_0, \sigma_0}\|_\infty \lesssim \varepsilon$ , where  $C_1$  is a constant. Without loss of generality, we may assume that the support points of  $F'_0$  are at least  $2\varepsilon$ -separated. If not, take a maximal  $2\varepsilon$ -separated set in the support points of  $F'_0$ . Let  $F''_0$  be the discrete measure on this  $2\varepsilon$ -net with weights obtained by moving the masses in  $F'_0$  to the closest point in the support of  $F''_0$ . Then  $\|p_{F'_0, \sigma_0} - p_{F''_0, \sigma_0}\|_\infty \leq 2\varepsilon \|p_{F'_0, \sigma_0}\|_\infty$ , and hence we can replace  $F'_0$  by  $F''_0$ .

By Lemma 3.2, we have  $\|p_{F_0, \sigma_0} - p_{F'_0, \sigma_0}\|_1 \lesssim \varepsilon (\log \frac{1}{\varepsilon})^{1/2}$ . Represent  $F'_0$  as  $\sum_{j=1}^N p_j \delta_{z_j}$ , so that  $N \leq C_1 \log \frac{1}{\varepsilon}$ . Then by Lemma 5.1, for some constants  $d_1$  and  $d_2$ ,

$$(5.12) \quad \begin{aligned} & \left\{ (F, \sigma) : \|p_{F, \sigma} - p_{F_0, \sigma_0}\|_1 \leq d_1 \varepsilon \left(\log \frac{1}{\varepsilon}\right)^{1/2} \right\} \\ & \supset \{ (F, \sigma) : \|p_{F, \sigma} - p_{F'_0, \sigma_0}\|_1 \leq d_2 \varepsilon \} \\ & \supset \left\{ (F, \sigma) : \sum_{j=1}^N |F[z_j - \varepsilon, z_j + \varepsilon] - p_j| \leq \varepsilon, |\sigma - \sigma_0| \leq \varepsilon \right\}. \end{aligned}$$

Since the  $z_j$ ’s are in  $[-k_0, k_0]$ , the intervals  $[z_j - \varepsilon, z_j + \varepsilon]$  are contained in  $[-k_0 - 1, k_0 + 1]$ . Hence for any  $F$  satisfying  $\sum_{j=1}^N |F[z_j - \varepsilon, z_j + \varepsilon] - p_j| \leq \varepsilon$ , we have  $F[-k_0 - 1, k_0 + 1] > 1 - \varepsilon > \frac{1}{2}$ . By (2.4) and Lemma 4.1, we find that for any  $(F, \sigma)$  on the left hand side (LHS) of (5.12)

$$(5.13) \quad p_{F, \sigma} \in B \left( c\varepsilon^{1/2} \left(\log \frac{1}{\varepsilon}\right)^{5/4}, p_{F_0, \sigma_0} \right)$$

for some constant  $c$ . Therefore using the prior independence of  $F$  and  $\sigma$ , we obtain

$$(5.14) \quad \begin{aligned} & \Pi \left( B \left( c\varepsilon^{1/2} \left( \log \frac{1}{\varepsilon} \right)^{5/4}, p_{F_0, \sigma_0} \right) \right) \\ & \geq \Pi(|\sigma - \sigma_0| \leq \varepsilon) \Pi \left( \sum_{j=1}^N |F[z_j - \varepsilon, z_j + \varepsilon] - p_j| \leq \varepsilon \right). \end{aligned}$$

By the positivity and continuity of the prior density of  $\sigma$ , the first factor on the RHS of (5.14) is bounded below by a constant multiple of  $\varepsilon$ . To bound the second factor from below, we apply Lemma A.2 with  $N + 1$  sets  $A_j = [z_j - \varepsilon, z_j + \varepsilon]$ ,  $j = 1, \dots, N$  and  $A_{N+1} = (\cup_{j=1}^N A_j)^c$ . Clearly, if  $\varepsilon$  is sufficiently small, for some constant  $A$ ,

$$(5.15) \quad A\varepsilon \leq \alpha(A_j) \leq 1, \quad j = 1, \dots, N.$$

We may also assume without loss of generality that  $\alpha(A_{N+1}) \leq 1$ ; otherwise we subdivide  $A_{N+1}$  into a number of subsets each satisfying the required condition. Then  $N$  will be increased by only a number not depending on  $\varepsilon$ , not affecting the conclusion of that Lemma A.2, which gives a bound a multiple of  $\exp[-c'N \log \frac{1}{\varepsilon}] \geq \exp[-c''(\log \frac{1}{\varepsilon})^2]$  for some constants  $c'$  and  $c''$ . The first factor on the RHS of (5.14) can be absorbed into this. Thus we obtain

$$(5.16) \quad \Pi \left( B \left( c\varepsilon^{1/2} \left( \log \frac{1}{\varepsilon} \right)^{5/4}, p_{F_0, \sigma_0} \right) \right) \geq \bar{C} \exp \left[ -\bar{c} \left( \log \frac{1}{\varepsilon} \right)^2 \right].$$

for some constants  $\bar{C}$  and  $\bar{c}$ . Putting  $\varepsilon' = c\varepsilon^{1/2}(\log \frac{1}{\varepsilon})^{5/4}$  and noting that  $\log \frac{1}{\varepsilon} \sim \log \frac{1}{\varepsilon'}$ , we have

$$(5.17) \quad \Pi(B(\varepsilon', p_{F_0, \sigma_0})) \geq c_1 \exp \left[ -c_2 \left( \log \frac{1}{\varepsilon'} \right)^2 \right]$$

for some constants  $c_1$  and  $c_2$ .

It therefore follows from (5.17) that the sequence  $\tilde{\varepsilon}_n = (\log n)/\sqrt{n}$  satisfies the Condition (2.10) of Theorem 2.1. If we now choose  $\tilde{\varepsilon}_n = (\log n)^\kappa/\sqrt{n}$ , where  $\kappa = \frac{2}{\delta} + \frac{1}{2}$ ,  $a_n = L(\log \frac{1}{\tilde{\varepsilon}_n})^{2/\delta}$ ,  $\mathcal{P} = \{p_{F, \sigma} : F \in \mathfrak{M}(\mathbb{R}), \underline{\sigma} \leq \sigma \leq \bar{\sigma}\}$ ,  $\mathcal{P}_n = \mathcal{F}_{a_n, \tilde{\varepsilon}_n}^1$  and  $L > (4(c_2 + 4)/b)^{1/\delta}$ , then from (5.11), it follows that  $\Pi(\mathcal{P} \setminus \mathcal{P}_n)$  is bounded above by a multiple of  $\exp[-(c_2+4)n\tilde{\varepsilon}_n^2]$ . Therefore the Condition (2.9) holds for these choices of  $\tilde{\varepsilon}_n$  and  $\mathcal{P}_n$ . The estimate in (5.10) then shows that Condition (2.8) holds for the given choice of  $\tilde{\varepsilon}_n$ . The result now follows from Theorem 2.1.  $\square$

**REMARK 5.2.** An examination of the proof reveals that the existence, continuity and positivity of the density of the base measure are used only to guarantee that intervals of size  $\varepsilon$  have  $\alpha$ -measure at least of the order of  $\varepsilon$  as  $\varepsilon \rightarrow 0$ . This condition holds also if the absolutely continuous part of  $\alpha$  possesses a density that is bounded away from zero on the support of  $F_0$ . In particular,  $\alpha$  may contain point masses.

As mentioned in Section 2, Theorem 5.1 also implies that the posterior mean of the density, as a point estimator, converges to the true density at the same rate  $(\log n)^\kappa/\sqrt{n}$  in Hellinger distance; see Ghosal, Ghosh and van der Vaart [(2000), page 507]. Similar propositions hold for all the later theorems, although they will not be separately stated.

**COROLLARY 5.1.** *Under the conditions of Theorem 5.1, the posterior mean  $\hat{p}(x) = \int p(x)d\Pi(p|X_1, \dots, X_n)$  satisfies  $d(\hat{p}, p_0) = O_p((\log n)^\kappa/\sqrt{n})$ .*

The condition of compact support of the true mixing distribution  $F_0$  is not necessary; it suffices that  $F_0$  has sub-Gaussian tails provided a normal base measure is used.

**THEOREM 5.2.** *Assume that the true mixing measure has sub-Gaussian tails in the sense that for some  $c_0 > 0$ ,  $F_0(|z| > t) \lesssim e^{-c_0 t^2}$  for all  $t$ . If the prior for  $\sigma$  has continuous and positive density on an interval containing  $\sigma_0$  and the base measure  $\alpha$  is normal, then for a large enough constant  $M$ ,*

$$(5.18) \quad \Pi \left( p : d(p, p_0) > M \frac{(\log n)^{3/2}}{\sqrt{n}} \mid X_1, \dots, X_n \right) \rightarrow 0$$

in  $P_0^n$ -probability.

**PROOF.** A normal base measure  $\alpha$  satisfies the tail condition (5.1) in Theorem 5.1 with  $\delta = 2$ . The proof of Theorem 5.1 uses the compact support of  $F_0$  only to verify (2.10) of Theorem 2.1 with  $\tilde{\varepsilon}_n = (\log n)/\sqrt{n}$ . Thus for  $F_0$  with sub-Gaussian tails, we need only to verify (2.10) with this  $\tilde{\varepsilon}_n$ . Then it will follow as in the last theorem that the posterior converges at rate  $(\log n)^\kappa/\sqrt{n}$ , where  $\kappa = \max(\frac{2}{2}, \frac{1}{2}) + \frac{1}{2} = \frac{3}{2}$ .

Set  $B = 2(\int z^2 dF_0(z))^{1/2}$ . For a given  $0 < \varepsilon < \frac{1}{4}$ , let  $a = c_0^{-1/2}(\log \frac{1}{\varepsilon})^{1/2}$  and  $F_0^*$  be  $F_0$  restricted to  $[-a, a]$  and normalized. By Lemma A.3 of the Appendix,

$$(5.19) \quad \|p_{F_0^*, \sigma_0} - p_{F_0, \sigma_0}\|_1 \leq 2F_0([-a, a]^c) \lesssim e^{-c_0 a^2} = \varepsilon.$$

Find a discrete distribution  $F'_0$  on  $[-a, a]$  which matches the moments of  $F_0^*$  up to the order  $N$ , where  $N \lesssim \log \frac{1}{\varepsilon}$ . By Lemma A.1 of the Appendix,  $F'_0$  can be chosen to have at most  $N + 1$  support points. Represent  $F'_0 = \sum_{j=1}^{N+1} p_j \delta_{z_j}$ . Because  $F_0^*$  has smaller second moment than  $F_0$ , Chebyshev's inequality gives

$$F'_0([-B, B]^c) \leq B^{-2} \int z^2 dF'_0(z) = B^{-2} \int z^2 dF_0^*(z) \leq B^{-2} \int z^2 dF_0(z) = \frac{1}{4}.$$

We may also assume that, as before,  $|z_j - z_k| > 2\varepsilon$ ,  $j \neq k$ , increasing  $B$  to  $B + 1$ , if necessary, to satisfy  $F'_0([-B, B]^c) \leq \frac{1}{4}$ . Thus if  $F$  is such that

$$(5.20) \quad \sum_{j=1}^{N+1} |F[z_j - \varepsilon, z_j + \varepsilon] - p_j| < \varepsilon,$$

then  $F([-B, B]^c) < \frac{1}{4} + \varepsilon < \frac{1}{2}$ . Therefore, as in the proof of Theorem 5.1, Lemmas 4.1 and 5.1 imply that for some  $c > 0$ ,

$$(5.21) \quad B \left( c\varepsilon^{1/2} \left( \log \frac{1}{\varepsilon} \right)^{5/4}, p_{F_0, \varepsilon_0} \right) \supset \{ (F, \sigma) : |F[z_j - \varepsilon, z_j + \varepsilon] - p_j| \leq \varepsilon, |\sigma - \sigma_0| \leq \varepsilon \}.$$

As argued in the proof of Theorem 5.1, it is then enough to estimate the probability of (5.20). To this end, note that  $|z_j| \leq a \lesssim (\log \frac{1}{\varepsilon})^{1/2}$ , so that  $\alpha[z_j - \varepsilon, z_j + \varepsilon] \geq A\varepsilon \exp(-c' \log \frac{1}{\varepsilon}) \geq A\varepsilon^b$  for some constants  $A, c'$  and  $b$ . Now Lemma A.2 can be applied to conclude that (2.10) is satisfied by  $\tilde{\varepsilon}_n = (\log n)/\sqrt{n}$ .

**6. Posterior convergence: Location-scale mixtures.** Now consider the case when the true density is a location-scale mixture of normals and the prior is a Dirichlet location-scale mixture. We have the following theorems.

**THEOREM 6.1.** *Let  $F_0$  and  $\alpha$  be as in Theorem 5.1 and assume that  $G_0$  has support in  $(\underline{\sigma}, \bar{\sigma})$ . Let the base measure  $\beta$  of the Dirichlet prior for  $G$  have continuous and positive density on an interval containing  $\text{supp}(G_0)$  and have support contained in  $[\underline{\sigma}, \bar{\sigma}]$ . Then for a large enough constant  $M$ ,*

$$(6.1) \quad \Pi \left( p : d(p, p_0) > M \frac{(\log n)^\kappa}{\sqrt{n}} \mid X_1, \dots, X_n \right) \rightarrow 0$$

in  $P_0^n$ -probability, where  $\kappa = \max(\frac{2}{\delta}, \frac{1}{2}) + \frac{1}{2}$ .

The following analogue of Lemma 5.1 will be used.

**LEMMA 6.1.** *Let  $F^* = \sum_{j=1}^N p_j \delta_{z_j}$  be a probability measure on  $\mathbb{R}$  with  $|z_j - z_{j'}| > 2\varepsilon$  for  $j \neq j'$  and  $G^* = \sum_{k=1}^N q_k \delta_{\sigma_k}$  be a probability measure on  $[\underline{\sigma}, \bar{\sigma}]$  with  $|\sigma_k - \sigma_{k'}| > 2\varepsilon$  for  $k \neq k'$ . Then for any probability measures  $F$  on  $\mathbb{R}$  and  $G$  on  $[\underline{\sigma}, \bar{\sigma}]$ ,*

$$(6.2) \quad \|p_{F,G} - p_{F^*,G^*}\|_1 \lesssim \varepsilon + \sum_{j=1}^N |F[z_j - \varepsilon, z_j + \varepsilon] - p_j| + \sum_{k=1}^N |G[\sigma_k - \varepsilon, \sigma_k + \varepsilon] - q_k|.$$

**PROOF.** Write

$$\begin{aligned} p_{F,G}(x) - p_{F^*,G^*}(x) &= \int_{(z,\sigma): |z-z_j| > \varepsilon \forall j, \text{ or } |\sigma-\sigma_k| > \varepsilon \forall k} \phi_\sigma(x-z) dF(z) dG(\sigma) \\ &\quad + \sum_{j=1}^N \sum_{k=1}^N \int_{|z-z_j| \leq \varepsilon, |\sigma-\sigma_k| \leq \varepsilon} (\phi_\sigma(x-z) - \phi_{\sigma_k}(x-z)) dF(z) dG(\sigma) \end{aligned}$$

$$\begin{aligned}
 & + \sum_{j=1}^N \sum_{k=1}^N \int_{|z-z_j| \leq \varepsilon, |\sigma-\sigma_k| \leq \varepsilon} (\phi_{\sigma_k}(x-z) - \phi_{\sigma_k}(x-z_j)) dF(z) dG(\sigma) \\
 & + \sum_{j=1}^N \sum_{k=1}^N \phi_{\sigma_k}(x-z_j) (F[z_j - \varepsilon, z_j + \varepsilon] G[\sigma_k - \varepsilon, \sigma_k + \varepsilon] - p_j q_k).
 \end{aligned}$$

The integral of the first term on the RHS can be bounded by

$$\begin{aligned}
 & (F \times G)((z, \sigma) : |z - z_j| > \varepsilon \forall j, \text{ or } |\sigma - \sigma_k| > \varepsilon \forall k) \\
 & = 1 - \sum_{j=1}^N \sum_{k=1}^N F[z_j - \varepsilon, z_j + \varepsilon] G[\sigma_k - \varepsilon, \sigma_k + \varepsilon] \\
 & \leq \sum_{j=1}^N \sum_{k=1}^N |F[z_j - \varepsilon, z_j + \varepsilon] G[\sigma_k - \varepsilon, \sigma_k + \varepsilon] - p_j q_k| \\
 & \leq \sum_{j=1}^N |F[z_j - \varepsilon, z_j + \varepsilon] - p_j| + \sum_{k=1}^N |G[\sigma_k - \varepsilon, \sigma_k + \varepsilon] - q_k|.
 \end{aligned}$$

Note that  $\|\phi_{\sigma}(\cdot) - \phi_{\sigma'}(\cdot)\|_1 \lesssim |\sigma - \sigma'|$  and  $\|\phi_{\sigma}(\cdot - z) - \phi_{\sigma}(\cdot - z')\|_1 \lesssim |z - z'|$ , so the integrals of the second and the third terms are bounded by a multiple of  $\varepsilon$ . The proof is complete.  $\square$

**PROOF OF THEOREM 6.1.** The proof follows the trail of that of Theorem 5.1. In this case,  $\mathcal{P} = \{p_{F,G} : F \in \mathfrak{M}(\mathbb{R}), G \in \mathfrak{M}[\underline{\sigma}, \bar{\sigma}]\}$ .

By Lemma 3.3, find discrete distributions  $F_0^*$  and  $G_0^*$  on  $[-k_0, k_0]$  and an interval containing  $\text{supp}(G_0)$  on which  $\beta$  has a positive density, respectively, with at most  $N \lesssim \log \frac{1}{\varepsilon}$  support points such that  $\|p_{F_0, G_0} - p_{F_0^*, G_0^*}\| \leq \varepsilon$ . Represent  $F_0^*$  and  $G_0^*$  respectively as  $\sum_{j=1}^N p_j \delta_{z_j}$  and  $\sum_{k=1}^N q_k \delta_{\sigma_k}$ , where  $|z_j - z_{j'}| > 2\varepsilon, j \neq j', |\sigma_k - \sigma_{k'}| > 2\varepsilon, k \neq k'$ , without loss of generality. To estimate the prior probability of  $\|p_{F,G} - p_{F_0^*, G_0^*}\| \leq \varepsilon$ , observe that by Lemma 6.1, it suffices to estimate the probability of

$$(6.3) \quad \sum_{j=1}^N |F[z_j - \varepsilon, z_j + \varepsilon] - p_j| \leq \varepsilon \text{ and } \sum_{k=1}^N |G[\sigma_k - \varepsilon, \sigma_k + \varepsilon] - q_k| \leq \varepsilon,$$

In view of the prior independence of  $F$  and  $G$  and Lemma A.2, the prior probability of the above set is at least  $C \exp[-c(\log \frac{1}{\varepsilon})^2]$  for some constants  $C$  and  $c$ . By an obvious analogue of Lemma 4.1 and the arguments given in (5.12), the above set is contained in  $B(c\varepsilon^{1/2}(\log \frac{1}{\varepsilon})^{5/4}, p_0)$ . Now proceeding as in the proof of Theorem 5.1 and estimating the covering numbers by Theorem 3.2, we conclude that conditions of Theorem 2.1 are satisfied for  $\bar{\varepsilon}_n = (\log n)^{\frac{2}{\delta} + \frac{1}{2}} / \sqrt{n}, \tilde{\varepsilon}_n = (\log n) / \sqrt{n}$  and  $\mathcal{P}_n = \{p_{F,G} : F[-a_n, a_n] \geq 1 - \bar{\varepsilon}_n\}$ , where  $a_n = L(\log \frac{1}{\varepsilon_n})^{2/\delta}$ , and  $L$  is a sufficiently large constant.  $\square$

The following result gives the posterior convergence rate for the general location-scale mixture model defined by (2.3).

**THEOREM 6.2.** *Let  $H_0$  be a compactly supported probability measure in  $\mathbb{R} \times (\underline{\sigma}, \bar{\sigma})$  and let  $H$  be given a Dirichlet process prior with a base measure  $\gamma$  which has a positive and continuous density on a rectangle inside  $\mathbb{R} \times [\underline{\sigma}, \bar{\sigma}]$  containing  $\text{supp}(F_0)$  and satisfies the tail condition*

$$(6.4) \quad \gamma((z, \sigma) : |z| > t) \leq ce^{-b|t|^\delta} \quad \text{for all } t > 0.$$

*Then for a sufficiently large constant  $M$ ,*

$$(6.5) \quad \Pi \left( p : d(p, p_0) > M \frac{(\log n)^\kappa}{\sqrt{n}} \mid X_1, \dots, X_n \right) \rightarrow 0$$

*in  $P_0^n$ -probability, where  $\kappa = 2 \max(\frac{3}{\delta}, \frac{1}{2}) + \frac{1}{2}$ .*

**REMARK 6.1.** In the above theorem, the best rate  $(\log n)^{3/2}/\sqrt{n}$  is obtained when  $\delta$  can be chosen to be 6 or more, which is the case if the base measure is compactly supported. When the base measure is the product of a normal distribution with a distribution supported in  $[\underline{\sigma}, \bar{\sigma}]$  such that the density is positive on a rectangle containing  $\text{supp}(F_0)$ , we may take  $\delta = 2$  and so the rate  $(\log n)^{7/2}/\sqrt{n}$  is obtained.

**PROOF OF THEOREM 6.2.** The proof is similar to that of Theorem 6.1.

Here, the class of densities is  $\mathcal{P} = \{p_H : H \in \mathfrak{M}(\mathbb{R} \times [\underline{\sigma}, \bar{\sigma}])\}$ . We may assume without loss of generality that  $\delta \leq 6$ .

By Lemma 3.4, get a discrete distribution  $H_0^*$  on the support of  $H_0$  with at most  $N \lesssim (\log \frac{1}{\varepsilon})^2$  support points such that  $\|p_{H_0} - p_{H_0^*}\|_\infty \leq \varepsilon$ . Represent  $H_0^* = \sum_{j=1}^N r_j \delta_{(z_j, \sigma_j)}$ , where, without loss of generality, the sets  $[z_j - \varepsilon, z_j + \varepsilon] \times [\sigma_j - \varepsilon, \sigma_j + \varepsilon]$ ,  $j = 1, \dots, N$ , are disjoint. Similar to Lemma 6.1, the set  $\|p_H - p_{H_0^*}\|_1 \leq \varepsilon$ , contains

$$(6.6) \quad \sum_{j=1}^N |H([z_j - \varepsilon, z_j + \varepsilon] \times [\sigma_j - \varepsilon, \sigma_j + \varepsilon]) - r_j| \leq \varepsilon,$$

and, is contained in  $B(c\varepsilon^{1/2}(\log \frac{1}{\varepsilon})^{5/4}, p_0)$  for some  $c$ . Lemma A.2 now shows that the prior probability of the last set is at least  $C \exp[-c(\log \frac{1}{\varepsilon})^3]$  for some constants  $C$  and  $c$ .

This, together with Theorem 3.3 and the arguments similar to those used in the proofs of Theorems 5.1 and 6.1, imply that the conditions of Theorem 2.1 are satisfied for  $\bar{\varepsilon}_n = (\log n)^{\frac{6}{\delta} + \frac{1}{2}}/\sqrt{n}$ ,  $\tilde{\varepsilon}_n = (\log n)^{3/2}/\sqrt{n}$  and  $\mathcal{P}_n = \{p_H : H([-a_n, a_n] \times [\sigma_j - \varepsilon, \sigma_j + \varepsilon]) \geq 1 - \bar{\varepsilon}_n\}$ , where  $a_n = L(\log \frac{1}{\varepsilon_n})^{3/\delta}$ , and  $L$  is a sufficiently large constant.  $\square$

**REMARK 6.2.** The slower rate  $(\log n)^{\max(\frac{6}{\delta}, 1) + \frac{1}{2}}/\sqrt{n}$ , compared to the rate  $(\log n)^{\max(\frac{2}{\delta}, \frac{1}{2}) + \frac{1}{2}}/\sqrt{n}$  in Theorems 5.1 and 6.1, is due to the lack of product structure in the mixing measure. Therefore we had to match many more moments in Lemma 3.4 compared to Lemma 3.1 for model (2.1) and Lemma 3.3

for model (2.2). Hence the entropy estimate as well as the prior estimate are relatively inferior causing a weaker rate of convergence.

Clearly one can extend Theorems 6.1 and 6.2 when the true mixing measure is not compactly supported in the spirit of Theorem 5.2. These results are not separately stated here.

**7. Extensions and variations.** The method of proof used in Sections 3–5 also allows certain variations that are useful from practical aspects. Often it is sensible to allow the precision parameter  $m = \alpha(\mathbb{R})$  of the base measure of the Dirichlet process to grow with the sample size. As the following theorem shows, the conclusion of Theorem 5.1 is not affected provided that  $m$  does not grow too fast.

**THEOREM 7.1.** *Assume the set-up and conditions of Theorem 5.1 except that  $m = \alpha(\mathbb{R})$  varies with the sample size. If  $1 \lesssim m \lesssim \log n$ , then for a sufficiently large constant  $M$ ,*

$$\Pi \left( p : d(p, p_0) > M \frac{(\log n)^\kappa}{\sqrt{n}} \mid X_1, \dots, X_n \right) \rightarrow 0$$

in  $P_0^n$ -probability, where  $\kappa = \max(\frac{2}{\delta}, \frac{1}{2}) + \frac{1}{2}$ .

To prove the theorem, we use Theorem 2.1 with the described sequence of Dirichlet mixture priors and proceed as in the proof of Theorem 5.1. We follow the notations of Theorem 5.1. Clearly, the estimate in (5.11) remains valid since  $m$  is bounded below. Next, to estimate the Dirichlet probability in (5.14), subdivide  $(\cup_{j=1}^N [z_j - \varepsilon, z_j + \varepsilon])^c$  into  $k$  sets each having  $\alpha$ -measure at most 1, where  $k$  is the smallest integer greater than or equal to  $m$ , and apply Lemma A.2 with  $N$  replaced by  $N + k$  sets. Since  $m$  grows at most like  $\log n$ , the estimate is unaffected. At all the other places,  $m$  has no role. Therefore (5.2) holds.

The base measure of the Dirichlet process is usually specified up to certain parameters, which themselves are given a prior. The  $N(\mu, \tau)$  base measure, where  $\mu$  and  $\tau$  are also distributed according to some prior, is often used. Another possibility is to put a prior on the precision parameter  $\alpha(\mathbb{R})$  of the Dirichlet base measure  $\alpha$ . If the base measures corresponding to the different values of the hyperparameters satisfy the conditions on  $\alpha$  in Theorem 5.1 uniformly, then the results go through with minor modifications.

**THEOREM 7.2.** *Assume the set-up of Theorem 5.1. However, instead of the prior  $D_\alpha$  for  $F$ , consider a mixture of  $D_{\alpha_\theta}$  priors, where  $\theta$  is given an arbitrary prior and  $\alpha_\theta$  are base measures satisfying the following conditions:*

- (i)  $\alpha_\theta(\mathbb{R})$  is bounded above and below in  $\theta$ .
- (ii) There exist constants  $B$ ,  $b$  and  $\delta > 0$  such that  $\alpha_\theta(z : |z| > t) \leq B e^{-bt^\delta}$  for all  $t > 0$  and  $\theta$ .

(iii) Every  $\alpha_\theta$  has a density  $\alpha'_\theta$  such that for some  $\varepsilon > 0$ ,  $\alpha'_\theta(x) \geq \varepsilon$  for all  $\theta$  and  $x \in [-k_0, k_0]$ .

Then for a sufficiently large constant  $M$ ,

$$\Pi \left( p : d(p, p_0) > M \frac{(\log n)^\kappa}{\sqrt{n}} \mid X_1, \dots, X_n \right) \rightarrow 0$$

in  $P_0^n$ -probability, where  $\kappa = \max(\frac{2}{\delta}, \frac{1}{2}) + \frac{1}{2}$ .

To prove the above, one needs to check the bounds for prior estimates. The prior probability for any set under this hierarchical prior could be obtained by first conditioning on  $\theta$  and then integrating the possible values of the probability of that set given  $\theta$  with respect to the prior for  $\theta$ . Therefore, if the constants in the prior estimates in (5.11) and (5.17) can be chosen free of  $\theta$ , then the conclusion holds. The assumed conditions (i) and (ii) clearly imply that the estimate in (5.11) is uniform. From (i) and (iii), (5.15) and the arguments following that, it is easily seen that the estimate (5.17) is also uniform.

Conditions of the above theorem usually hold if the parametric family  $\alpha_\theta$  is “well behaved” and  $\theta$  has a compact range. However, conditions (ii) and (iii) are not expected to hold if the range of  $\theta$  is unbounded. Nevertheless, in certain situations, the conclusion may still hold even if the hyperparameters are not compactly supported. We consider the important special case where the base measure is  $N(\mu, \tau)$  and  $\mu$  is also given a normal prior.

**THEOREM 7.3.** *Assume the set-up and conditions of Theorem 5.1 and suppose that the base measure is  $N(\mu, \tau)$ , where  $\mu$  is given a  $N(\mu_0, A)$  prior, and  $\tau$  is either given, or has a compactly supported prior distribution. Then*

$$\Pi \left( p : d(p, p_0) > M \frac{(\log n)^\kappa}{\sqrt{n}} \mid X_1, \dots, X_n \right) \rightarrow 0$$

in  $P_0^n$ -probability, where  $\kappa = \max(\frac{2}{\delta}, \frac{1}{2}) + \frac{1}{2}$ .

To prove Theorem 7.3, we proceed as in the proof of Theorem 7.2. Since we are seeking only a lower bound in (5.17), we may restrict our attention to a compact interval for  $\mu$ 's, where a uniform estimate is available by the argument given in the last theorem. It therefore remains to consider (5.11). Let  $\Pi(\cdot|\mu)$  denote the prior given  $\mu$  and  $\Pi$  the overall prior. With the notations as in the proof of Theorem 5.1, we have for any  $C, a, \varepsilon$  and  $\eta$ ,

$$\Pi((\mathcal{F}_{a,\eta}^1)^c) \leq \Pi \left( |\mu| > C \log \frac{1}{\varepsilon} \right) + \sup \left\{ \Pi((\mathcal{F}_{a,\eta}^1)^c | \mu) : \mu \leq C \log \frac{1}{\varepsilon} \right\}.$$

If  $C$  is chosen sufficiently large,  $L > C$  and  $a = L \log \frac{1}{\varepsilon}$ , then both terms are bounded by  $\exp[-c(\log \frac{1}{\varepsilon})^2]$ , for some  $c$ . This is clearly true for the first term, while the same follows for the second term from the inequality

$$\alpha_\mu([-a, a]^c) \leq \alpha_0([-(L - C) \log \varepsilon^{-1}, (L - C) \log \varepsilon^{-1}]^c) \lesssim e^{-c(\log \frac{1}{\varepsilon})^2}$$

for all  $\mu \leq C \log \frac{1}{\varepsilon}$ . The result follows.

Similar remarks apply in the context of Theorems 5.2, 6.1 and 6.2.

## APPENDIX

LEMMA A.1. *Let  $K$  be a compact metric space and let  $\psi_1, \dots, \psi_N$  be continuous functions from  $K$  to  $\mathbb{R}$ . Then for any probability measure  $F_0$  on  $K$ , there exists a discrete probability measure  $F$  on  $K$  with at most  $N + 1$  support points such that*

$$(A.1) \quad \int \psi_j dF = \int \psi_j dF_0 \quad \text{for } j = 1, \dots, N.$$

PROOF. The set

$$C = \{(\psi_1(x), \dots, \psi_N(x)) : x \in K\}$$

is compact in  $\mathbb{R}^N$ . Therefore its convex hull  $\text{conv}(C)$  is also compact. Then for every probability measure  $F_0$  on  $K$ ,

$$(A.2) \quad v_0 = \left( \int \psi_1(x) dF_0(x), \dots, \int \psi_N(x) dF_0(x) \right)^T \in \text{conv}(C)$$

[see, e.g., Lemma 3 on page 74 of Ferguson (1967)]. Since  $\text{conv}(C) \subset \mathbb{R}^N$ , any element of  $\text{conv}(C)$  may be written as a convex combination of at most  $N + 1$  elements of  $C$  [see, e.g., Rudin (1973, page 73)]. Thus there exist  $\lambda_1, \dots, \lambda_{N+1} \geq 0$ ,  $\sum_{j=1}^{N+1} \lambda_j = 1$ ,  $x_1, \dots, x_{N+1} \in K$  such that

$$(A.3) \quad \begin{aligned} v_0 &= \sum_{j=1}^{N+1} \lambda_j (\psi_1(x_j), \dots, \psi_N(x_j))^T \\ &= \left( \int \psi_1(x) dF(x), \dots, \int \psi_N(x) dF(x) \right)^T, \end{aligned}$$

where  $F = \sum_{j=1}^{N+1} \lambda_j \delta_{x_j}$ .

The following estimate of the probability of a  $\ell_1$ -ball under a Dirichlet distribution appeared as Lemma 6.1 in Ghosal, Ghosh and van der Vaart (2000).

LEMMA A.2. *Let  $(X_1, \dots, X_N)$  be distributed according to the Dirichlet distribution on the unit  $\ell_1$ -simplex in  $\mathbb{R}^N$ ,  $N \geq 2$ , with parameters  $(m; \alpha_1, \dots, \alpha_N)$ , where  $A\varepsilon^b \leq \alpha_j \leq 1$  and  $\sum_{j=1}^N \alpha_j = m$  for some constant  $A$  and  $b$ . Let  $(x_1, \dots, x_N)$  be any point on the  $N$ -simplex. Then there exist positive constants  $c$  and  $C$  depending only on  $A$  and  $b$  such that for  $\varepsilon \leq 1/N$ ,*

$$(A.4) \quad P \left( \sum_{j=1}^N |X_j - x_j| \leq 2\varepsilon \right) \geq C \exp \left( -cN \log \frac{1}{\varepsilon} \right).$$

The following lemma helps truncate the domain of the mixing distribution.

Let  $(z, B) \mapsto \Psi(B|z)$  be a Markov kernel on arbitrary measurable spaces admitting densities  $\psi(\cdot|z)$ . Consider a mixture  $p_F(x) = \int \psi(x|z)dF(z)$ , where  $F$  is a probability measures on the domain  $Z$  of  $z$ .

LEMMA A.3. *Let  $F$  be an arbitrary probability measure on  $Z$  and  $F^*$  be its renormalized restriction to a subset  $A \subset Z$  with  $F(A) > 0$ , that is,  $F^*(B) = F(A \cap B)/F(A)$  for all  $B$ . Then*

$$(A.5) \quad \|p_{F^*} - p_F\|_1 \leq 2F(A^c).$$

Therefore, if  $\mathcal{F} = \{p_F : F(A) \geq 1 - \varepsilon\}$  and  $\mathcal{F}^* = \{p_F : F(A) = 1\}$ , then

$$(A.6) \quad N(3\varepsilon, \mathcal{F}, \|\cdot\|_1) \leq N(\varepsilon, \mathcal{F}^*, \|\cdot\|_1).$$

PROOF. For any  $x$ , we have

$$\begin{aligned} \int \psi(x|z)d(F^* - F)(z) &= \int_A \psi(x|z)dF(z) \left( \frac{1}{F(A)} - 1 \right) \\ &\quad - \int_{A^c} \psi(x|z)dF(z). \end{aligned}$$

Using the triangle inequality and next integrating over  $x$ , the LHS of (A.5) is bounded by

$$F(A) \left( \frac{1}{F(A)} - 1 \right) + F(A^c) = 2F(A^c).$$

For (A.6), note that any  $\varepsilon$ -net over  $\mathcal{F}^*$  is a  $3\varepsilon$ -net over  $\mathcal{F}$  by (A.5).

The following lemma is possibly known in the literature. However, the proof does not appear to be readily available, so we include a brief proof as well.

LEMMA A.4. *Let  $\Delta_N = \{(x_1, \dots, x_N) : x_i \geq 0, \sum_{i=1}^N x_i = 1\}$  be the unit  $\ell_1$ -simplex in  $\mathbb{R}^N$ ,  $N \geq 2$ . Then for  $\varepsilon \leq 1$ ,*

$$(A.7) \quad D(\varepsilon, \Delta_N, \|\cdot\|_1) \leq \left( \frac{5}{\varepsilon} \right)^{N-1}.$$

PROOF. For  $x \in \Delta_N$ , let  $x^*$  denote the vector of its first  $N - 1$  co-ordinates. Then  $x^*$  belongs to the set  $D_{N-1} = \{(y_1, \dots, y_{N-1}) : y_i \geq 0, \sum_{i=1}^{N-1} y_i \leq 1\}$ . The correspondence  $x \mapsto x^*$  is one-to-one and  $\|x_1 - x_2\|_1 \leq 2\|x_1^* - x_2^*\|_1$ . Let  $x_1, \dots, x_m \in \Delta_N$  such that  $\|x_i - x_j\|_1 > \varepsilon$ ,  $i, j = 1, \dots, N$ ,  $i \neq j$ . Then  $\|x_i^* - x_j^*\|_1 > \varepsilon/2$ ,  $i, j = 1, \dots, N$ ,  $i \neq j$ , and so  $B_{N-1}(x_i^*, \varepsilon/4)$ , the  $\ell_1$ -balls in  $\mathbb{R}^{N-1}$  of radius  $\varepsilon/4$  centered at  $x_i^*$ , are disjoint. The union of these balls is clearly contained in the set

$$(A.8) \quad \left\{ (y_1, \dots, y_{N-1}) : \sum_{i=1}^{N-1} |y_i| \leq (1 + \varepsilon/4) \right\}.$$

Let  $V_{N-1}$  be the volume of the unit  $\ell_1$ -ball in  $\mathbb{R}^{N-1}$ . Then the volume of the set in (A.8) is  $(1 + \varepsilon/4)^{N-1}V_{N-1} \leq (5/4)^{N-1}V_{N-1}$ , while the volume of each  $B(x_i^*, \varepsilon/4)$  is  $(\varepsilon/4)^{N-1}V_{N-1}$ . A volume argument now gives the desired bound.  $\square$

It is interesting to note that the technique of the proof applies to any unit  $\ell_p$ -ball in  $\mathbb{R}^N$ . By a similar argument, we get

$$(A.9) \quad D(\varepsilon, \{x : \|x\|_p \leq 1\}, \|\cdot\|_p) \leq \left(\frac{3}{\varepsilon}\right)^N$$

for all  $\varepsilon \leq 1$ .

**Acknowledgment.** Most of this work was done when the first author was affiliated with the Free University of Amsterdam.

## REFERENCES

- BANFIELD, J. and RAFTERY, A. (1993). Model based Gaussian and non-Gaussian clustering. *Biometrics* **49** 803–821.
- BARRON, A., SCHERVISH, M. and WASSERMAN, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27** 536–561.
- BIRGÉ, L. and MASSART, P. (1998). Minimum contract estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* **4** 329–375.
- ESCOBAR, M. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588.
- FERGUSON, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York.
- FERGUSON, T. S. (1983). Bayesian density estimation by mixtures of Normal distributions. In *Recent Advances in Statistics* (M. Rizvi, J. Rustagi and D. Siegmund, eds.) 287–302. Academic Press, New York.
- GEMAN, S. and HWANG, C. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.* **10** 401–414.
- GENOVESE, C. and WASSERMAN, L. (2000). Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.* **28** 1105–1127.
- GHOSAL, S., GHOSH, J. K. and RAMAMOORTHI, R. V. (1999a). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27** 143–158.
- GHOSAL, S., GHOSH, J. K. and RAMAMOORTHI, R. V. (1999b). Consistency issues in Bayesian Nonparametrics. In *Asymptotics, Nonparametrics and Time Series: A Tribute to Madan Lal Puri* (S. Ghosh, ed.) 639–668. Dekker, New York.
- GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531.
- GRENANDER, U. (1981). *Abstract Inference*. Wiley, New York.
- IBRAGIMOV, I. A. (2001). Estimation of analytic functions. In *State of the Art in Probability and Statistics. Festschrift for W. R. van Zwet*. IMS, Hayward, CA.
- IBRAGIMOV, I. A. and KHASHMINSKII, R. Z. (1982). An estimate of the density of a distribution belonging to a class of entire functions. *Theory Probab. Appl.* **27** 514–524 (in Russian).
- KOLMOGOROV, A. N. and TIHOMIROV, V. M. (1961).  $\varepsilon$ -entropy and  $\varepsilon$ -capacity of sets in function spaces. *Amer. Math. Soc. Transl. Ser. 2* **17** 277–364. [Translated from Russian (1959) *Uspekhi Mat. Nauk* **14** 3–86.]
- LINDSAY, B. (1995). *Mixture Models: Theory, Geometry and Applications*. IMS, Hayward, CA.
- LO, A. Y. (1984). On a class of Bayesian nonparametric estimates I: Density estimates. *Ann. Statist.* **12** 351–357.

- McLACHLAN, G. and BASFORD, K. (1988). *Mixture Models: Inference and Applications to Clustering*. Dekker, New York.
- PRIEBE, C. E. (1994). Adaptive mixtures. *J. Amer. Statist. Assoc.* **89** 796–806.
- ROBERT, C. (1996). Mixtures of distributions: inference and estimation. In *Markov Chain Monte Carlo in Practice* (W. Gilks, S. Richardson and D. Spiegelhalter, eds.) 441–464. Chapman and Hall, London.
- ROEDER, K. (1992). Semiparametric estimation of normal mixture densities. *Ann. Statist.* **20** 929–943.
- ROEDER, K. and WASSERMAN, L. (1997). Practical Bayesian density estimation using mixtures of normals. *J. Amer. Statist. Assoc.* **92** 894–902.
- RUDIN, W. (1987). *Real and Complex Analysis*, 3rd ed. McGraw-Hill, New York.
- SCHWARTZ, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* **4** 10–26.
- SHEN, X. and WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29** 687–714.
- SHEN, X. and WONG, W. H. (1994). Convergence rate of sieve estimates. *Ann. Statist.* **22** 580–615.
- VAN DE GEER, S. (1993). Hellinger consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21** 14–44.
- VAN DE GEER, S. (1996). Rates of convergence for the maximum likelihood estimator in mixture models. *J. Nonparametr. Statist.* **6** 293–310.
- VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.
- WASSERMAN, L. (1998). Asymptotic properties of nonparametric Bayesian procedures. *Practical Nonparametric and Semiparametric Bayesian Statistics. Lecture Notes in Statist.* **133** 293–304. Springer, New York.
- WEST, M. (1992). Modeling with mixtures. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 503–524. Oxford Univ. Press.
- WEST, M., MULLER, P. and ESCOBAR, M. D. (1994). Hierarchical priors and mixture models, with applications in regression and density estimation. In *Aspects of Uncertainty: A Tribute to D. V. Lindley* (P. R. Freeman and A. F. M. Smith, eds.) 363–386. Wiley, New York.
- WONG, W. H. and SHEN, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.* **23** 339–362.

SCHOOL OF STATISTICS  
376 FORD HALL  
224 CHURCH ST. SE  
UNIVERSITY OF MINNESOTA  
MINNEAPOLIS, MINNESOTA 55455  
E-MAIL: ghosal@stat.umn.edu

DIVISION OF MATHEMATICS AND  
COMPUTER SCIENCE  
FREE UNIVERSITY  
DE BOELELAAN 1081A  
1081 HV AMSTERDAM  
THE NETHERLANDS  
E-MAIL: aad@cs.vu.nl