CONVERGENCE RATES FOR DENSITY ESTIMATION WITH BERNSTEIN POLYNOMIALS

By Subhashis Ghosal

University of Minnesota

Mixture models for density estimation provide a very useful set up for the Bayesian or the maximum likelihood approach. For a density on the unit interval, mixtures of beta densities form a flexible model. The class of Bernstein densities is a much smaller subclass of the beta mixtures defined by Bernstein polynomials, which can approximate any continuous density. A Bernstein polynomial prior is obtained by putting a prior distribution on the class of Bernstein densities. The posterior distribution of a Bernstein polynomial prior is consistent under very general conditions. In this article, we present some results on the rate of convergence of the posterior distribution. If the underlying distribution generating the data is itself a Bernstein density, then we show that the posterior distribution converges at "nearly parametric rate" $(\log n)/\sqrt{n}$ for the Hellinger distance. If the true density is not of the Bernstein type, we show that the posterior converges at a rate $n^{-1/3}(\log n)^{5/6}$ provided that the true density is twice differentiable and bounded away from 0. Similar rates are also obtained for sieve maximum likelihood estimates. These rates are inferior to the pointwise convergence rate of a kernel type estimator. We show that the Bayesian bootstrap method gives a proxy for the posterior distribution and has a convergence rate at par with that of the kernel estimator.

1. Introduction. Mixture models, formed by convex combinations of densities from parametric families give us simple, well-behaved, flexible nonparametric classes of densities. Mixture models are used in various inference problems such as density estimation, clustering analysis and robust estimation; see for example, Lindsay (1995) and McLachlan and Basford (1988). On the real line, a mixture of normal densities is often used to model an unknown smooth density. From a Bayesian point of view, the mixture model gives us a very convenient set-up for density estimation in that one can induce a prior distribution on the densities simply by specifying a prior distribution on the mixing distribution. Early users of this approach were Ferguson (1983) and Lo (1984), who used a Dirichlet process prior for the mixing distribution and gave expressions for posterior expectations of functions. Ghosal, Ghosh and Ramamoorthi (1999) showed that a Dirichlet mixture of normals prior gives rise to a consistent posterior under general conditions for the weak topology and the variation distance. Ghosal and van der Vaart (2001) showed that the posterior converges at "nearly parametric rate" if the true density generating the observations is also a mixture of normals with standard deviations

---

bounded away from zero and infinity. Gibbs sampling techniques to compute the posterior mean and other posterior characteristics have been developed; see, for example, Escobar and West (1995) and the references therein.

While the normal mixture model is a very sensible choice for densities on the entire real line, its usefulness is rather limited when we consider densities on a known bounded interval, taken to be $(0, 1]$ without loss of generality. The normal kernel density estimate suffers from boundary effects at 0 and 1. In this case, the family of beta distribution forms a flexible two-parameter family of densities and mixtures of beta distributions form an appropriate mixture model. In fact, a mixture of only a relatively few beta densities can approximate any distribution on $(0,1]$. For a continuous probability distribution function F on $(0,1]$, the associated Bernstein polynomial

$$(1.1) \qquad B(x; k, F) = \sum_{j=0}^{k} F(j/k) \binom{k}{j} x^j (1-x)^{k-j}$$

converges uniformly to F as $k \to \infty$. Clearly, $B(x; k, F)$ is a mixture of beta distributions, since it has density

$$(1.2) \qquad b(x; k, F) = \sum_{j=1}^{k} (F(j/k) - F((j-1)/k)) \beta(x; j, k-j+1),$$

where $\beta(x; a, b)$ stands for the beta density

$$\beta(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}.$$

The uniform approximation property of the Bernstein polynomials motivated Vitale (1975) to study density estimates based on the Bernstein polynomials. Tenbusch (1994) extended this idea to multidimensional densities. Diaconis (1993) suggested that a prior on the class of densities on $(0,1]$ with a full topological support may be constructed using the approximating property of the Bernstein polynomials. Using this idea, Petrone (1999a, b) proposed the following hierarchical prior called the Bernstein polynomial prior: the density $f(\cdot)$ on $(0,1]$ is given by the following mixture of beta densities:

$$(1.3) \qquad f(x) = \sum_{j=1}^{k} w_{j,k} \beta(x; j, k-j+1),$$

where k has probability mass function $\rho(\cdot)$, and given k, $\mathbf{w}_k = (w_{1,k}, \ldots, w_{k,k})$ has a distribution $H_k(\cdot)$ on the k-dimensional simplex

$$\Delta_k = \left\{ (x_1, \ldots, x_k) : \ 0 \le x_j \le 1, \ j = 1, \ldots, k, \ \sum_{j=1}^{k} x_j = 1 \right\}.$$

Following Petrone (1999a, b), we shall call the right-hand side (RHS) of (1.3) a Bernstein density with parameters k and $\mathbf{w}_k$, and denote it by $b(x; k, \mathbf{w}_k)$;

here we slightly abuse the notation in (1.2). The class of all Bernstein densities of order k will be denoted $\mathscr{B}_k$. It is useful to note that

$$(1.4) \qquad\qquad f(x) = k \sum_{j=1}^{k} w_{j,k} p_{j,k}(x),$$

where $p_{j,k}(x) = \binom{k-1}{j-1} x^{j-1} (1-x)^{k-j}$. In particular, $f(x) \leq k$.

Petrone (1999a) showed that if for all k, $\rho(k) > 0$ and $\mathbf{w}_k$ has full support on Δ_k, then every distribution on (0,1] is in the weak support of the Bernstein polynomial prior, and every continuous distribution is in the topological support of the prior defined by the Kolmogorov–Smirnov distance.

The posterior mean, given k, is

$$(1.5) \quad E(f(x)|k, x_1, \ldots, x_n) = \sum_{j=1}^{k} E(w_{j,k}|x_1, \ldots, x_n)\beta(x; j, k - j + 1),$$

and the distribution of k is updated to $\rho(k|x_1, \ldots, x_n)$. Petrone (1999a, b) and Petrone and Wasserman (2001) discussed Markov chain Monte Carlo (MCMC) algorithms to compute the posterior expectations and carried out extensive simulations to show that the resulting density estimates work well. The MCMC algorithm is largely satisfactory, but sometimes the convergence could be slow. Petrone and Wasserman (2001) also suggested an alternative to the Bayes estimate by considering the average of the maximum likelihood estimate (MLE) for each k with respect to weights obtained by normalizing the BIC or AIC.

The issue of consistency of the posterior distribution of a Bernstein polynomial prior has been addressed by Petrone and Wasserman (2001). They showed that if for all k, $\rho(k) > 0$ and $\mathbf{w}_k$ has full support on Δ_k, then the posterior distribution is consistent at any continuous density f_0 on (0,1] for the weak topology. If further, the sequence of weights $\rho(k)$ satisfies a certain tail condition, then the posterior is consistent with respect to the Hellinger (equivalently, variation) metric. The main idea behind the proof of consistency is to show that the prior satisfies Schwartz's (1965) condition of positivity of the prior probabilities of every neighborhood of the true density f_0 defined by the Kullback–Leibler divergence. See Ghosal, Ghosh and Ramamoorthi (1999b) and Wasserman (1998) for recent reviews on consistency.

In this article, we obtain the rate of convergence of the posterior distribution for the Bernstein polynomial prior, under additional smoothness conditions. Note that the rate of convergence of a density estimate may be arbitrarily slow at a density which is merely continuous. Assuming more smoothness in the true density, we compute the concentration rate of the prior distribution on a Kullback–Leibler type neighborhood and then apply the general theory of posterior rate of convergence developed by Ghosal, Ghosh and van der Vaart (2000).

Since we study rate of convergence, we need a sufficiently tight lower bound on the posterior probability of a shrinking Kullback–Leibler type ball,

and therefore we will work with less general priors compared to Petrone and Wasserman (2001). Note that $\mathbf{w}_k$ is a random element on Δ_k, so a k-dimensional Dirichlet distribution $D(k; \alpha_{1, k}, \ldots, \alpha_{k, k})$ is a reasonable choice. We restrict our attention to the cases where $\alpha_{j, k}$ are bounded by some number M for all j and k. This assumption will be in force throughout. The following two special cases correspond to two important choices of $\alpha_{j, k}$'s that satisfy the required condition. In the first case, $\alpha_{j, k} = M\alpha(((j-1)/k, j/k])$, where M is a positive constant and α is a probability measure on $(0,1]$. This is equivalent to saying that given k, the density f has form (1.2) where F has the Dirichlet process distribution with parameter $M\alpha(\cdot)$. This prior has been named the Bernstein–Dirichlet prior by Petrone (1999a). Another choice, often thought to be noninformative, corresponds to $\alpha_{j, k} = 1$ for all j, k.

We show that if the true density is itself a Bernstein polynomial, then the posterior converges at $(\log n)/\sqrt{n}$ rate. If the true density is not of the Bernstein type, then the convergence rate, of course, cannot be this fast. We show that if the true density is continuously differentiable in $(0,1)$, bounded below and has a bounded second derivative, then the posterior distribution converges at the rate $n^{-1/3}(\log n)^{5/6}$. A slight improvement to $n^{-1/3}(\log n)^{1/3}$ is possible by considering a sequence of priors. It should be noted that the Bayes estimate, defined as the pointwise posterior expectation, also converges at the rate equal to that of convergence of posterior distribution; see the discussion following Theorem 2.5 of Ghosal, Ghosh and van der Vaart (2000). Similar rates are also obtained for the sieve maximum likelihood estimate. These rates are substantially slower than $n^{-2/5}$, the rate of convergence of the kernel type estimator of Vitale (1975). To overcome this drawback, we consider the Bayesian bootstrap method and show that the proxy posterior has the desired $n^{-2/5}$ convergence rate. Confidence intervals and bands are easy to obtain from the Bayesian bootstrap distribution.

The organization of the paper is as follows. In the next section, we study the convergence rate of the posterior distribution of a Bernstein polynomial prior. Convergence rate of the MLE is treated in Section 3. In Section 4, we present a result on the convergence rate of the Bayesian bootstrap. We shall use the symbol "$\lesssim$" to mean an inequality up to a constant multiple.

2. Convergence rate of posterior. Let $X_1, X_2, \ldots$ be independent observations from a density f on $(0,1]$. To estimate f, a Bernstein polynomial prior is put on f. We determine the rate of convergence of the posterior distribution. We apply the general theorem of Ghosal, Ghosh and van der Vaart (2000) on the rate of convergence of the posterior described below.

The Bernstein polynomials of order k have a uniform rate of approximation k^{-1} at smooth densities. More precisely, if $f(x)$ is a continuously differentiable probability density on $(0,1]$ with bounded second derivative,

$$(2.1) \qquad \sup_{0 < x \leq 1} |f(x) - b(x; k, F)| = O(k^{-1}),$$

where F is the distribution function corresponding to f. This property is well known and may be shown by observing that

$$b(x; k, F) = kE\left(\int_{J/k}^{(J+1)/k} f(z)\,dz\right),$$

where J has a binomial distribution with parameters $(k-1)$ and x, and then using the Taylor series expansion. See Lorenz (1953) for more details on the properties of the Bernstein polynomials.

For a distance d on a class of densities $\mathscr{F}$, let $D(\varepsilon, \mathscr{F}, d)$ stand for the ε-packing number defined to be the maximum number of points in $\mathscr{F}$ such that the distance between each pair is at least ε. We refer the readers to Kolmogorov and Tihomirov (1961) and van der Vaart and Wellner (1996) for details on packing numbers and related concepts. Let the true density $f_0 \in \mathscr{F}$, a class of densities and let P_0 be the probability measure with density f_0. Let $\|f - f_0\|_1$ stand for the L_1-distance and $h(f, f_0) = \|f^{1/2} - f_0^{1/2}\|_2$ stand for the Hellinger distance. Put $K(f_0, f) = \int \log(f_0/f)\,dP_0$, $V(f_0, f) = \int(\log(f_0/f))^2\,dP_0$, $N(\varepsilon, f_0) = \{f: K(f_0, f) \leq \varepsilon^2, \ V(f_0, f) \leq \varepsilon^2\}$. Let d stand for either the L_1 or the Hellinger distance. The following variation of Theorem 2.1 of Ghosal, Ghosh and van der Vaart (2000) will be useful. A similar result under stronger conditions has also been obtained by Shen and Wasserman (2001).

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

(2.2) $\log D(\tilde{\varepsilon}_n, \mathscr{F}_n, d) \leq c_1 n\bar{\varepsilon}_n^2,$

(2.3) $\Pi_n(\mathscr{F} \backslash \mathscr{F}_n) \leq c_3 e^{-(c_2+4)n\tilde{\varepsilon}_n^2},$

(2.4) $\Pi_n(N(\tilde{\varepsilon}_n, f_0)) \geq c_4 e^{-c_2 n\tilde{\varepsilon}_n^2}.$

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

(2.5) $\Pi_n(f: d(f, f_0) > M\varepsilon_n | X_1, \ldots, X_n) \to 0$

in P_0^n-probability.

Let $Q(k; \alpha_{1,k}, \ldots, \alpha_{k,k})$ denote the probability measure induced on $\mathscr{B}_k$ by assigning the Dirichlet distribution $D(k; \alpha_{1,k}, \ldots, \alpha_{k,k})$ to $\mathbf{w}_k$. The mixture $\sum_{k=1}^{\infty} \rho(k)Q(k; \alpha_{1,k}, \ldots, \alpha_{k,k})$ is considered as the prior on the density f.

Our first theorem shows the rate $n^{-1/2}\log n$ of convergence is obtained when the true density is actually a Bernstein density. This is analogous to Theorem 5.1 of Ghosal and van der Vaart (2001) for the case of normal mixtures.

THEOREM 2.2.   *Let the true density $f_0 = b(\cdot; k_0, \mathbf{w}_{k_0}^0)$ for some k_0 and $\mathbf{w}_{k_0}^0 \in \Delta_{k_0}$. Let $0 < \rho(k) \leq Be^{-\beta k}$ for some constants B and β. Then for a sufficiently*

large constant C,

$$(2.6) \qquad \Pi\left(f: d(f, f_0) > C\frac{\log n}{\sqrt{n}}\Big| X_1, \ldots, X_n\right) \to 0$$

in P_0^n-probability.

PROOF. Observe that when $f(x) = b(x; k_0, \mathbf{w}_{k_0})$, we have

$$\|f - f_0\|_1 \le \sum_{j=1}^{k_0} |w_{j, k_0} - w_{j, k_0}^0|.$$

Therefore, if $\sum_{j=1}^{k_0} |w_{j, k_0} - w_{j, k_0}^0| \le \varepsilon$, then $\|f - f_0\|_1 \le \varepsilon$, and so $h(f, f_0) \le \sqrt{\varepsilon}$.

Let $\mathbf{w}_{k_0}$ be such that

$$(2.7) \qquad \sum_{j=1}^{k_0} |w_{j, k_0} - w_{j, k_0}^0| \le \varepsilon, \qquad w_{1, k_0} > \varepsilon^4, w_{k_0, k_0} > \varepsilon^4.$$

Then

$$f(x) \ge w_{1, k_0} k_0 (1 - x)^{k_0 - 1} + w_{k_0, k_0} k_0 x^{k_0 - 1} > k_0 2^{-(k_0 - 1)} \varepsilon^4.$$

Applying Theorem 5 of Wong and Shen (1995), it then follows that

$$\max(K(f_0, f), V(f_0, f)) \lesssim \varepsilon\left(\log \frac{1}{\varepsilon}\right)^2.$$

Therefore, for some C_1,

$$N\left(C_1 \sqrt{\varepsilon} \log \frac{1}{\varepsilon}, f_0\right) \supset \{b(\cdot; k_0, \mathbf{w}_{k_0}): (2.7) \text{ holds}\}.$$

We estimate the probability of the sets in (2.7). By Lemma A.1 of the Appendix, these probabilities are bounded below by a multiple of $e^{-c \log(1/\varepsilon)}$, where c is a constant. Since k_0 is fixed and $\rho(k_0) > 0$, it follows that for some C_1,

$$\Pi\left(N\left(f_0, C_1 \sqrt{\varepsilon} \log \frac{1}{\varepsilon}\right)\right) \ge D_1 e^{-d_1 \log(1/\varepsilon)},$$

for constants D_1 and d_1. Putting $\eta = C_1 \sqrt{\varepsilon} \log(1/\varepsilon)$ and observing that $\log(1/\eta) \sim \log(1/\varepsilon)$, we have

$$\Pi(N(f_0, \eta)) \ge D_2 e^{-d_2 \log(1/\eta)},$$

for constants D_2 and d_2. Therefore $\tilde{\varepsilon}_n = n^{-1/2}(\log n)^{1/2}$ satisfies (2.4).

To check the first two conditions (2.2) and (2.3) in Theorem 2.1, we proceed as in the proof of Theorem 5 of Petrone and Wasserman (2001). Let $\mathcal{F}_n = \bigcup_{r=1}^{k_n} \mathcal{B}_r$, where k_n will be determined shortly. Then $D(\varepsilon, \mathcal{F}_n, d) \le \sum_{r=1}^{k_n} D(\varepsilon, \mathcal{B}_r, d)$. Since $\mathcal{B}_r$ is the class of all convex combinations of $r \le k_n$ fixed densities, its packing numbers can be obtained by relating them with the packing numbers of the r-dimensional simplex. To be more precise, we claim that for some absolute constant C,

$$(2.8) \qquad D(\varepsilon, \mathcal{B}_r, d) \le D(\varepsilon, \Delta_r, d') \le (C/\varepsilon)^r,$$

where d' is the ℓ_1-distance $\sum_{j=1}^{r}|x_j - y_j|$ on Δ_r if d is the L_1-distance on densities, and d' is the Hellinger distance $h'(\mathbf{x}, \mathbf{y}) = (\sum_{j=1}^{r}(\sqrt{x_i} - \sqrt{y_i})^2)^{1/2}$ on Δ_r if d is the Hellinger distance h on densities. If d is the L_1-distance, then the first relation in (2.8) follows by the triangle inequality while the second by the estimate of packing numbers of Δ_r for the ℓ_1-distance; see, for instance, Lemma A.4 of Ghosal and van der Vaart (2001). If d is the Hellinger distance on densities, then the two relations in (2.8) follow, for instance, respectively by Lemma 4 and Lemma 2 of Genovese and Wasserman (2000), and the relationship between packing and covering numbers. Therefore,

$$\log D(\varepsilon, \mathscr{F}_n, d) \le \log\left(\sum_{r=1}^{k_n}(C/\varepsilon)^r\right) \le k_n \log(C/\varepsilon) + \log k_n.$$

If we choose $L_1 \log(1/\tilde{\varepsilon}_n) \le k_n \le L_2 \log(1/\tilde{\varepsilon}_n)$ for constants L_1 and L_2, where $\tilde{\varepsilon}_n = n^{-1/2}(\log n)^{1/2}$ and $\bar{\varepsilon}_n = (\log n)/\sqrt{n}$, then $\log D(\bar{\varepsilon}_n, \mathscr{F}_n, d) \lesssim (\log(1/\tilde{\varepsilon}_n))^2$. As $(\log(1/\bar{\varepsilon}_n))^2 \lesssim n\bar{\varepsilon}_n^2$, (2.2) holds. Now,

$$\Pi(\mathscr{F}_n^c) \le \sum_{j=k_n}^{\infty} \rho(j) \lesssim e^{-\beta k_n} \le e^{-L_1\beta\,\log(1/\tilde{\varepsilon}_n)} \le e^{-Ln\tilde{\varepsilon}_n^2},$$

where L can be made as large as we want by choosing L_1 sufficiently large. It follows that condition (2.3) is satisfied, and hence the result follows.

REMARK 2.1. As remarked in Petrone and Wasserman (1999), an extended Bernstein prior, which is supported on extended Bernstein densities $\sum_{r=1}^{k}\sum_{j=1}^{r} w_{j,r}\beta(x; j, r - j + 1)$ instead of Bernstein densities, may also be considered. Note that extended Bernstein densities are nothing but mixtures of Bernstein densities. It can be easily checked that if the true density is of the extended Bernstein type and an extended Bernstein prior [see Petrone and Wasserman (2001)] is used, then the same rate $(\log n)/\sqrt{n}$ is obtained.

When the true density f_0 is not of the Bernstein type, the convergence rate would be naturally much slower. The following result shows that if the true density is continuously differentiable in $(0, 1)$, bounded away from 0 and has a bounded second derivative, then the posterior distribution converges at the rate $n^{-1/3}(\log n)^{5/6}$. This rate is somewhat slower than the $n^{-2/5}$ rate of Vitale's Bernstein polynomial density estimator. While our theorem only gives an upper bound for the rate of convergence of the posterior distribution, it is nevertheless an indication of a weaker rate. We believe that the obtained rate is essentially sharp except possibly for factor a power of $\log n$. The reason behind our belief is that Bernstein polynomials have a relatively weak approximation property (2.1), which, unlike the approximation by convolutions, is only proportional to bandwidth k^{-1}. This means that we must use very high degree Bernstein polynomials to approximate a general smooth density. Somewhat paradoxically, Vitale's estimator does not suffer from a rate deficiency since the variance of his estimator is small compared with analogous kernel estimators with the same bandwidth.

THEOREM 2.3.   *Let the true density f_0 be bounded away from 0 and have bounded second derivative. Consider a Bernstein polynomial prior for f satisfying the condition $B_1 e^{-\beta_1 j} \leq \rho(j) \leq B_2 e^{-\beta_2 j}$ for all j for some constants $B_1, B_2, \beta_1, \beta_2 > 0$. Then for a sufficiently large constant C,*

$$(2.9) \qquad \Pi(f \colon d(f, f_0) > C n^{-1/3} (\log n)^{5/6} | X_1, \ldots, X_n) \to 0$$

in P_0^n-probability.

PROOF.   For $k \geq 1$, define $f_k(x) = b(x; k, F_0)$, where F_0 is the cumulative distribution function for f_0. By (2.1), $\sup_{0 < x \leq 1} |f_0(x) - f_k(x)| = O(k^{-1})$, and so f_k is also uniformly bounded away from 0 for all large k. Note that we may also write $f_k(x) = b(x; k, \mathbf{w}_k^0)$, where $\mathbf{w}_k^0 = (w_{1,k}^0, \ldots, w_{k,k}^0)$ and

$$w_{j,k}^0 = \int_{(j-1)/k}^{j/k} f_0(x)\, dx = F_0(j/k) - F_0((j-1)/k), \qquad j = 1, \ldots, k.$$

Also observe that

$$|b(x; k, \mathbf{w}_k) - b(x; k, \mathbf{w}_k^0)| = \left| k \sum_{j=1}^{k} (w_{j,k} - w_{j,k}^0) \binom{k-1}{j-1} x^{j-1} (1-x)^{k-j-1} \right|$$

$$(2.10) \qquad\qquad \leq k \max_{1 \leq j \leq k} |w_{j,k} - w_{j,k}^0|$$

$$\leq k \sum_{j=1}^{k} |w_{j,k} - w_{j,k}^0|.$$

Therefore if $\|\mathbf{w}_k - \mathbf{w}_k^0\|_1 \leq \varepsilon^2$ and $d_1 \varepsilon^{-1} \leq k \leq d_2 \varepsilon^{-1}$ for some constants d_1 and d_2, then $\sup_{0 < x \leq 1} |f_0(x) - b(x; k, \mathbf{w}_k)| \leq D_1 \varepsilon$ for a constant D_1 and also $b(x; k, \mathbf{w}_k)$ is bounded away from 0 for sufficiently small ε. It therefore follows that for some constant D_2, $h(f_0, b(\cdot; k, \mathbf{w}_k)) \leq D_2 \varepsilon$ and so (8.6) of Ghosal, Ghosh and van der Vaart (2000) implies that $b(\cdot; k, \mathbf{w}_k) \in N(C_1 \varepsilon, f_0)$ for a constant C_1. Hence

$$N(C_1 \varepsilon, f_0) \supset \{b(\cdot; k, \mathbf{w}_k) : \|\mathbf{w}_k - \mathbf{w}_k^0\|_1 \leq \varepsilon^2\}.$$

If we choose k_n satisfying

$$(2.11) \qquad\qquad b_1 \left(\frac{n}{\log n} \right)^{1/3} \leq k_n \leq b_2 \left(\frac{n}{\log n} \right)^{1/3}$$

for some constants b_1 and b_2 and $\tilde{\varepsilon}_n = k_n^{-1}$, Lemma A.1 of the Appendix implies that for some constants C_3, C_4, D and d,

$$\Pi(N(C_1 \tilde{\varepsilon}_n, f_0)) \geq \rho(k_n) C_2 e^{-C_3 k_n \log(1/\tilde{\varepsilon}_n)}$$

$$\geq B_1 e^{-\beta_1 d_2 / \tilde{\varepsilon}_n} \times C_2 e^{-C_3 d_2 (1/\tilde{\varepsilon}_n) \log(1/\tilde{\varepsilon}_n)}$$

$$\geq D e^{-d(1/\tilde{\varepsilon}_n) \log(1/\tilde{\varepsilon}_n)}.$$

Hence $\tilde{\varepsilon}_n = n^{-1/3} (\log n)^{1/3}$ satisfies condition (2.4) of Theorem 2.1.

Let s_n be an integer satisfying

$$(2.12) \qquad L_1 \frac{1}{\tilde{\varepsilon}_n} \log \frac{1}{\tilde{\varepsilon}_n} \leq s_n \leq L_2 \frac{1}{\tilde{\varepsilon}_n} \log \frac{1}{\tilde{\varepsilon}_n}$$

for some constants L_1 and L_2. Then

$$L_1' n^{1/3} (\log n)^{2/3} \leq s_n \leq L_2' n^{1/3} (\log n)^{2/3},$$

where we may choose $L_1' = L_1/6$ and $L_2' = L_2/3$. Put $\mathscr{F}_n = \cup_{r=1}^{s_n} \mathscr{B}_r$. Then for constants B_3, B and L,

$$\Pi(\mathscr{F}_n^c) \leq \sum_{r=s_n+1}^{\infty} \rho(r) \leq B_3 e^{-\beta_2 s_n} \leq B e^{-L(1/\tilde{\varepsilon}_n) \log(1/\tilde{\varepsilon}_n)},$$

and L can be made arbitrarily large by choosing L_1 sufficiently large. As $(1/\tilde{\varepsilon}_n) \log(1/\tilde{\varepsilon}_n)$ and $n\tilde{\varepsilon}_n^2$ have the same order, (2.3) holds.

Now by the arguments given in the proof of Theorem 2.2, for some constants C and L_3, we have

$$\log D(\varepsilon, \mathscr{F}_n, d) \leq s_n \log \left(\frac{C}{\varepsilon} \right) + \log s_n$$

$$\leq L_2' n^{1/3} (\log n)^{2/3} \log \left(\frac{C}{\varepsilon} \right) + \log \left(L_2' n^{1/3} (\log n)^{2/3} \right)$$

$$\leq L_3 n^{1/3} (\log n)^{2/3} \log \frac{1}{\varepsilon}.$$

So (2.2) holds for the choice $\bar{\varepsilon}_n = n^{-1/3} (\log n)^{5/6}$. Hence the posterior converges at the rate $n^{-1/3} (\log n)^{5/6}$. $\square$

A slight improvement in the rate is possible by considering a sequence of priors. Here we choose a sequence of prior supported on the sieve $\mathscr{F}_n$ so that condition (2.3) becomes trivial. It will then allow us to choose $s_n = k_n$ in the proof of the last theorem and will yield the slightly stronger rate $n^{-1/3} (\log n)^{1/3}$. The result is stated below, but the proof is omitted.

Consider a sequence of priors

$$\Pi_n = \sum_{r=1}^{k_n} \rho_n(j) Q(r; \alpha_{1,r}^{(n)}, \ldots, \alpha_{r,r}^{(n)}),$$

where k_n is a sequence of integers tending to infinity and $\sum_{r=1}^{k_n} \rho_n(r) = 1$ for all n. Assume, as before, that for some constant M, $\alpha_{j,r}^{(n)} \leq M$ for all $j = 1, \ldots, r$ and sufficiently large n and r.

THEOREM 2.4.   *Let the true density f_0 be bounded below and have bounded second derivative. Consider a sequence of priors Π_n defined above satisfying the*

condition $\rho_n(j) \geq Be^{-bj}$ for some constants B, $b > 0$ for all n and j, and choose a sequence k_n satisfying the condition

$$(2.13) \qquad b_1\left(\frac{n}{\log n}\right)^{1/3} \leq k_n \leq b_2\left(\frac{n}{\log n}\right)^{1/3}$$

for some constants b_1 and b_2. Then for a sufficiently large constant C,

$$(2.14) \qquad \Pi(f: d(f, f_0) > Cn^{-1/3}(\log n)^{1/3}|X_1, \ldots, X_n) \to 0$$

in P_0^n-probability.

3. Sieved maximum likelihood estimate. The sieve method was introduced by Grenander (1981) and studied by many authors including Geman and Hwang (1982), van de Geer (1993, 1996), Shen and Wong (1994), Wong and Shen (1995) and Birgé and Massart (1998). This sieved MLE is obtained by maximizing the likelihood function on a suitable subset of the parameter space called the sieve. For Bernstein polynomial densities, the convergence rate of the sieve MLE is obtained below by applying Theorem 4 of Wong and Shen (1995). Alternatively, one can also apply Theorem 3.4.4 of van der Vaart and Wellner (1996). This refines the conclusion of Theorem 8 of Petrone and Wasserman (2001) from consistency to a rate of convergence statement under the stated conditions.

THEOREM 3.1. *Consider a sieve $\mathscr{F}_n = \cup_{r=1}^{k_n}\mathscr{B}_r$, the space of all Bernstein densities of the order k_n or less, where k_n is a sequence of integers tending to infinity. Let $\hat{f}_n$ maximize the likelihood on this sieve.*

If the true density $f_0 = b(\cdot; k_0, \mathbf{w}_{k_0}^0)$ for some k_0 and $w_{k_0}^0 \in \Delta_{k_0}$, that is, f_0 itself is a Bernstein density, then $d(\hat{f}_n, f_0) = O_p(k_n^{1/2}n^{-1/2}(\log n)^{1/2})$. In particular, the rate can be made arbitrarily close to $n^{-1/2}(\log n)^{1/2}$ by letting k_n grow arbitrarily slowly.

If f_0 is not of the Bernstein type, but is bounded away from 0 and has a bounded second derivative, then for the choice

$$c_1 n^{1/3}(\log n)^{-1/3} \leq k_n \leq c_2 n^{1/3}(\log n)^{-1/3},$$

where c_1 and c_2 are constants, we have $d(\hat{f}_n, f_0) = O_p(n^{-1/3}(\log n)^{1/3})$.

PROOF. In the first case, $\mathscr{F}_n$ contains the true density f_0 for sufficiently large n. Therefore it suffices to check (3.1) of Wong and Shen (1995) with ε a multiple of $k_n^{1/2}n^{-1/2}(\log n)^{1/2}$. As the ε-bracketing Hellinger entropy of $\mathscr{F}_n$ is bounded by a multiple of $k_n \log(1/\varepsilon)$ (see the arguments given in the proof of Theorem 2.2), the stated claim follows easily.

When f_0 is not of the Bernstein type, consider, as in the proof of Theorem 2.3, $f_k(x) = b(x; k, \mathbf{w}_k^0)$, where $\mathbf{w}_k^0 = (w_{1,k}^0, \ldots, w_{k,k}^0)$ and $w_{j,k}^0 = F_0(j/k) - F_0((j-1)/k)$. Then by the arguments similar to that in Theorem 2.3, $K(f_0, f_k) \leq Dk^{-1}$ for a constant D. Thus $\mathscr{F}_n$ approximates f_0 at the rate k_n^{-1} in the

Kullback–Leibler distance. As (3.1) of Wong and Shen (1995) holds for a multiple of $k_n^{1/2} n^{-1/2} (\log n)^{1/2}$, Theorem 4 of Wong and Shen (1995) implies that

$$d(\hat{f}_n, f_0) = O_p(\max(k_n^{-1}, k_n^{1/2} n^{-1/2} (\log n)^{1/2})).$$

For the stated choice of k_n, the two rates inside the maximum agree and the rate $n^{-1/3} (\log n)^{1/3}$ is obtained. $\square$

It is interesting to note that the obtained rate of convergence of the sieved MLE agrees with the corresponding rate of the posterior given by Theorem 2.4. This suggests that the suboptimality of the rate of convergence of the posterior is possibly not due to the inadequacy of the prior, but due to the slow rate of approximation in (2.1).

4. The Bayesian bootstrap of Bernstein densities. An attractive feature of the Bayesian approach is that the posterior distribution not only gives an estimate, it also gives us a probability distribution on the parameter space given the data, which may be viewed as an updated opinion about the parameter in the light of the data and can be used for many purposes such as for the construction of confidence intervals or prediction of the next observation. Unfortunately, in nonparametric problems, Bayesian and the maximum likelihood approach may suffer from suboptimality in the rate of convergence; see Ghosal, Ghosh and van der Vaart (2000) and Shen and Wasserman (2001). As these methods are model based, often they perform very well if the model is right and is not too large, but generally the methods will suffer if either the model is incorrect or the model size is too large. Further, both the methods are tied up with the Hellinger distance and the Kullback–Leibler divergence, which are difficult to bound, particularly for mixtures. Also, Bayesian methods can suffer from an additional difficulty due to the lack of sufficient prior probability in the neighborhoods of the true density, as it is generally difficult for a prior to assign substantial probabilities to every part of the entire parameter space in a nonparametric model.

The Bayesian bootstrap (BB), introduced by Rubin (1981), provides an alternative to the Bayesian method in that it also gives us a probability distribution for the parameter given the data, and so it can be used for constructing confidence intervals. On the one hand, it is easy to compute the BB distribution by simulations, whereas the convergence properties of the BB are also easy to study using the simple structure of the BB. It should be noted that the BB does not provide a different estimator, but it gives us a proxy for the posterior distribution which is roughly centered at a standard estimator. Below, we show that the the BB approach can be applied to the Bernstein polynomial density estimation problem to construct a posterior distribution of the density given the data so that the rate of convergence agrees with the convergence rate of Vitale's (1975) kernel type estimator. The BB approach has been used for standard kernel density estimation [Lo (1987)], who called it the smoothed BB. The treatment here is somewhat similar, and therefore, it will be only briefly described.

Let $\alpha(\cdot)$ be a finite measure (the possibility of the null measure is not ruled out) on $(0, 1]$. Let m_0 stand for the prior strength $\alpha((0, 1])$ and $\bar{\alpha} = \alpha/m_0$ if $m_0 > 0$. The density is modeled by (1.2) and a Dirichlet process prior $\mathscr{D}_\alpha$ with base measure α assigned to the mixing measure F. An honest Bayesian, believing that the observations are generated from $f(x) = b(x, k, F)$, first induces a prior on f from the prior for F and then computes the posterior distribution obeying the Bayes principle. A Bayesian bootstrapper, not restricted by Bayes principle, first computes the "posterior distribution of F" pretending that $X_1, \ldots, X_n$ are observations from the distribution F, and then induces it on $f(x) = b(x; k, F)$. Note that if the observations $X_1, X_2, \ldots$ had had the law F which is given a Dirichlet prior $\mathscr{D}_\alpha$, then $\mathscr{D}_{\alpha + \sum_{i=1}^n \delta_{X_i}}$ would have been the posterior distribution of F. Let the BB distribution of $f(\cdot)$ be denoted by $\Pi_n^{\text{BB}}(\cdot | X_1, \ldots, X_n)$. In the BB approach, it is also easy to incorporate one's prior opinion and the strength of this prior belief is $\alpha((0, 1])$. The null measure corresponds to the "noninformative prior." In fact, following Lo (1987), it would be more appropriate to call Π_n^{BB} the posterior smoothed Dirichlet process, while calling the special case corresponding to $\alpha = 0$ the smoothed BB. However, the idea behind these distributions is the same and they are asymptotically equivalent. Therefore, we will simply use the more convenient and familiar term, the Bayesian bootstrap.

Let F_n stand for the empirical distribution $n^{-1} \sum_{i=1}^n \delta_{X_i}$ and D_n^* be the empirical distribution of a bootstrap sample, that is, $n^{-1} \sum_{i=1}^n \delta_{X_i^*}$, where $X_1^*, \ldots, X_n^*$ are i.i.d. F_n. Let $D_{n, \alpha}$ denote a random probability distribution distributed as the Dirichlet process $\mathscr{D}_{\alpha + \sum_{i=1}^n \delta_{X_i}}$. When α is the null measure, $D_n = D_{n,0}$ is Rubin's BB.

Let $k = k_n$ vary with n. Then Vitale's (1975) estimator can be written as

$$(4.1) \qquad \hat{f}_n(x) = b(x; k_n, F_n).$$

The bootstrap distribution of f is defined as the distribution of $b(\cdot; k_n, D_n^*)$, while Π_n^{BB} is the distribution of $b(\cdot; k_n, D_{n,\alpha})$. Thus Π_n^{BB} is supported on the space of Bernstein polynomials densities of order k_n and has expectation given by

$$(4.2) \qquad \hat{f}_{n, \alpha}(x) = \frac{m_0}{m_0 + n} b(x; k_n, \bar{\alpha}) + \frac{n}{m_0 + n} \hat{f}_n(x).$$

Note that the estimator obtained from the BB approach is essentially the same as Vitale's estimator except for a shrinkage toward the prior mean. In fact, for Rubin's BB, the estimator is exactly equal to Vitale's estimator.

The following result shows that the BB distribution of $f(x)$ concentrates near $f_0(x)$ at the right rate $n^{-2/5}$ if k_n is chosen to be of the order $n^{2/5}$. Let F_0 stand for the cumulative distribution function of f_0.

THEOREM 4.1. *Assume that the true density is twice differentiable. Let α have a continuous density and let $c_1 n^{2/5} \leq k_n \leq c_2 n^{2/5}$ for some constants c_1*

and c_2. Then for $0 < x < 1$ and any sequence $M_n \to \infty$,

(4.3) $\qquad \Pi_n^{\mathrm{BB}}(n^{2/5}(f(x) - f_0(x)) \geq M_n | X_1, \ldots, X_n) \to 0$

in P_0^n-probability.

To prove the theorem, it suffices to show that the BB mean of $f(x)$ converges to $f_0(x)$ at the rate $n^{-2/5}$ in probability and the BB variance of $f(x)$ is $O_p(n^{-4/5})$. Abbreviate k_n by k. To prove the first assertion, note that the first term on the RHS of (4.2) is bounded by $m_0 k / (m_0 + n)$, while the second term is asymptotically equivalent to $\hat{f}_n(x)$. Vitale (1975) showed that

$$E(\hat{f}_n(x) - f_0(x))^2 = O(k^{-2} + k^{1/2}/n).$$

Note that the variance can be written as $\mathrm{var}\big(k \sum_{j=1}^k w_{j,k} p_{j,k}(x)\big)$ where $(w_{1,k}, \ldots, w_{k,k})$ has $D(k; \alpha_{1,k}^*, \ldots, \alpha_{k,k}^*)$ distribution and

$$\alpha_{j,k}^* = \alpha_{j,k} + \sum_{i=1}^n I\left\{ \frac{j-1}{k} < X_i \leq \frac{j}{k} \right\}.$$

Noting that

$$\mathrm{cov}(w_{j,k}, w_{l,k}) = \begin{cases} \alpha_{j,k}^*(m_0 + n - \alpha_{j,k}^*)/((m_0+n)^2(m_0+n+1)), & j = l, \\ -\alpha_{j,k}^* \alpha_{l,k}^*/((m_0+n)^2(m_0+n+1)), & j \neq l, \end{cases}$$

the BB variance of $f(x)$ is bounded by $k^2 n^{-2} \sum_{j=1}^k \alpha_{j,k}^* p_{j,k}^2(x)$, where $p_{j,k}(x)$ is as in (1.4). Since $E(\alpha_{j,k}^*) = \alpha_{j,k} + n(F_0(j/k) - F_0((j-1)/k))$ and

$$(F_0(j/k) - F_0((j-1)/k)) p_{j,k}^2(x) \sim k^{-3/2} \frac{f_0(x)}{2\sqrt{x(1-x)}}$$

by Vitale [(1975), pages 93–95], the BB variance is $O_p(k^{1/2}/n)$. When k is of order $n^{2/5}$, the orders of the variance and the square of bias agree and equal $n^{-4/5}$. The result follows.

The above result shows that the BB credible intervals for $f(x)$ will have length of the order $n^{-2/5}$. It is easy to construct these intervals with the help of simulations. The result extends immediately to densities at points $x_1, \ldots, x_m$ and credible confidence set for $(f(x_1), \ldots, f(x_m))$ may be obtained by exploiting the log-concavity of its BB density; see Choudhuri (1998) for details.

It is a natural to ask whether the credible interval obtained from the BB has asymptotically the right frequentist coverage. In Bernstein polynomial density estimation, or density estimation in general, the order of the bias matches the variability. Therefore, the credible interval, which is centered at the Bernstein polynomial density estimate, drifts away from $f_0(x)$ by an amount that has order equal to the scale of interest, resulting in a loss of confidence. To counter the effect of this bias, we will have to view the interval as a confidence interval for $b(x, k_n, F_0)$ only. Below, we indicate why the BB

credible interval is expected to have the right asymptotic frequentist coverage. Note that a confidence interval for $f_0(x)$ may be obtained from that of $b(x, k_n, F_0)$ by either estimating the asymptotic bias

$$\frac{f_0'(x)(1-2x) + f_0''(x)x(1-x)}{2k_n},$$

or using a conservative bound for the absolute value of bias, and subsequently increasing the interval.

First, we observe that Vitale's estimator $\hat{f}(x) = b(x, k_n, F_n)$ is asymptotically normal. To see this, use Komlós, Major and Tusnády's (1975) strong approximation of the empirical process; see Csörgő and Révész (1981) for a detailed account of strong approximation. Let k abbreviate k_n, which is of order $n^{2/5}$. The stochastic process $\sqrt{n}(F_n(x) - F_0(x))$ is uniformly approximated by a Brownian bridge $Z_n(F_0(x))$ on $[0, 1]$ almost surely, within an error of $n^{-1/2}\log n$. Note that we can write $Z_n(F_0(x)) = W_n(F_0(x)) - F_0(x)W_n(1)$, where W_n is a Brownian motion. Now

$$\frac{\sqrt{n}}{k^{1/4}}(b(x; k, F_n) - b(x; k, F_0))$$

$$= k^{3/4}\sum_{j=1}^{k}(W_n(j/k) - W_n((j-1)/k))p_{j,k}(x)$$

$$+ k^{3/4}W_n(1)\sum_{j=1}^{k}(F_0(j/k) - F_0((j-1)/k))p_{j,k}(x)$$

$$+ O(k^{3/4}n^{-1/2}\log n).$$

Since $W_n(1)$ is $N(0, 1)$ and

$$k\sum_{j=1}^{k}(F_0(j/k) - F_0((j-1)/k))p_{j,k}(x) = b(x; k, F_0) \to f_0(x),$$

the second term is $O_p(k^{-1/4})$, and the third term goes to zero as well. The first term is clearly normal with mean zero and variance

$$k^{3/2}\sum_{j=1}^{k}(F_0(j/k) - F_0((j-1)/k))p_{j,k}^2(x) \to \frac{f(x)}{2\sqrt{\pi x(1-x)}}$$

by Vitale (1975). Thus

$$(4.4) \qquad \frac{\sqrt{n}}{k^{1/4}}(b(x; k, F_n) - b(x; k, F_0)) \to_d N\left(0, \frac{f_0(x)}{2\sqrt{\pi x(1-x)}}\right),$$

and hence for $k \sim n^{2/5}$,

$$\frac{\sqrt{n}}{k^{1/4}}(\hat{f}_n(x) - f_0(x))$$

$$(4.5) \qquad \to_d N\left(f_0'(x)(1-2x) + f_0''(x)x(1-x), \frac{f_0(x)}{2\sqrt{\pi x(1-x)}}\right).$$

The last assertion strengthens the conclusion of Vitale's theorem.

Note that Lo's strong approximation theorem for BB [Lo (1987), Theorem 2.1] also gives a Gaussian process approximation $Z_n^*(x)$, which has the same law as $Z_n(x)$, for the normalized BB process $\sqrt{n}(D_{n,\alpha} - F_n)$, and hence we should similarly expect that

$$(4.6) \quad \frac{\sqrt{n}}{k^{1/4}}(b(x; k, D_{n,\alpha}) - b(x; k, F_n)) \to_d N\left(0, \frac{f_0(x)}{2\sqrt{\pi x(1-x)}}\right) \quad \text{a.s.}$$

A comparison of (4.4) and (4.6) leads to the conclusion that the BB credible interval has asymptotically the right frequentist coverage probability for the parameter $b(x, k, F_0)$. Unfortunately, (4.6) cannot be concluded from Lo's theorem, as the error of the strong approximation there is only known to be $O(n^{-1/4}(\log n)^{1/2}(\log\log n)^{1/4})$. Note that the case of Bernstein polynomial density estimation is somewhat different from that of symmetric kernels in that $b(x; k, F)$ and $b(x; k, G)$ differ by $n^{2/5}\sup\{|F(z) - G(z)|: 0 \le z \le 1\}$. In the latter case, one chooses the bandwidth of the order $n^{-1/5}$ so that $\int h^{-1}\psi((x-z)/h)dF(z)$ and $\int h^{-1}\psi((x-z)/h)\,dG(z)$, where ψ is the kernel, differ only by the order $n^{1/5}\sup\{|F(z) - G(z)|\}$. However, we still believe that (4.6) is true. Note that the bootstrap is asymptotically equivalent to the BB by Lo's theorem. It should be noted that the BB distribution of $f(x)$ is completely different from a Bayesian's actual posterior distribution. Indeed, Theorems 2.2, 2.3 and 4.1 suggest that these two distributions have very different convergence properties.

Similarly, if the global deviation properties of the Vitale's estimator and the BB are similar, then the conjecture above could be strengthened to confidence bands. To be precise, we shall need the results analogous to Theorem 3.1 of Bickel and Rosenblatt (1973) and Theorem 5.1 of Lo (1987) for Bernstein polynomial density estimation. The former, in particular, will give a convergence rate for Vitale's estimator in the supremum norm.

APPENDIX

The following lemma generalizes the estimate of a Dirichlet probability given by Lemma 6.1 of Ghosal, Ghosh and van der Vaart (2000).

LEMMA A.1.   *Let* $(X_1, \ldots, X_N)$ *be distributed according to the Dirichlet distribution on the unit* ℓ_1-*simplex in* $\mathbb{R}^N$, $N \ge 2$, *with parameters* $(m; \alpha_1, \ldots, \alpha_N)$, *where* $A\varepsilon^b \le \alpha_j \le M$ *and* $\sum_{j=1}^N \alpha_j = m$ *for some constant* A, b, m *and* $M \ge 1$. *Let* $(x_1, \ldots, x_N)$ *be any point on the* N-*simplex. Then there exist positive constants* c *and* C *depending only on* A, M, m *and* b *such that for* $\varepsilon \le 1/(MN)$,

$$(A.1) \quad P\left(\sum_{j=1}^N |X_j - x_j| \le 2\varepsilon, \; X_j > \varepsilon^4 \text{ for all } j\right) \ge Ce^{-cN\log(1/\varepsilon)}.$$

To prove, first assume that $M = 1$, and proceed as in the proof of Lemma 6.1 of Ghosal, Ghosh and van der Vaart (2000). As shown there, $|X_j - x_j| < \varepsilon^2$

for all $j = 1, \ldots, N - 1$, implies that $X_N > \varepsilon^2 > \varepsilon^4$ and $\sum_{j=1}^{N} |X_j - x_j| \leq 2\varepsilon$. Therefore, the set on the left hand side (LHS) of (A.1) contains

$$\{|X_j - x_j| \leq \varepsilon^2, \ X_j > \varepsilon^4, \ j = 1, \ldots, N - 1\}$$

and hence the probability on the LHS of (A.1) is bounded below by

$$\frac{\Gamma(m)}{\prod_{j=1}^{N} \Gamma(\alpha_j)} \prod_{j=1}^{N-1} \int_{\max(x_j - \varepsilon^2, \varepsilon^4)}^{\min(x_j + \varepsilon^2, 1)} y_j^{\alpha_j - 1} \, dy_j.$$

Each interval of integration contains an intervals of length at least $\varepsilon^2 - \varepsilon^4 > \varepsilon^2/2$. The proof follows as before.

For a general M, we may assume without loss of generality that M is an integer. For each $j = 1, \ldots, N$, consider an auxiliary independent randomization that splits X_j into $X_{j,1}, \ldots, X_{j,M}$ according to the Dirichlet distribution $D(M; (\alpha_j/M), \ldots, (\alpha_j/M))$. Then the joint distribution of the whole collection $\{X_{j,k}: j = 1, \ldots, N, \ k = 1, \ldots, M\}$ is Dirichlet satisfying the conditions of the lemma with $M = 1$ and N replaced by MN. Clearly, the set on the LHS of (A.1) contains

$$\left\{ \sum_{j=1}^{N} \sum_{k=1}^{M} \left| X_{j,k} - \frac{x_j}{M} \right| \leq 2\varepsilon, \ X_{j,k} > \varepsilon^4, \ j = 1, \ldots, N - 1, \ k = 1, \ldots, M \right\}.$$

The result follows from the special case.

REFERENCES

BICKEL, P. J. and ROSENBLATT, M. (1973). On some global measures of the deviations of density estimates. *Ann. Statist.* **1** 1071–1095.

BIRGÉ, L. and MASSART, P. (1998). Minimum contract estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* **4** 329–375.

CHOUDHURI, N. (1998). Bayesian bootstrap credible sets for multidimensional mean functional. *Ann. Statist.* **26** 2104–2127.

CSÖRGÖ, M. and RÉVÉSZ, R. (1981). *Strong Approximations in Probability and Statistics.* Academic Press, New York.

DIACONIS, P. (1993). Seminar at the University of Minnesota.

ESCOBAR, M. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588.

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 non-parametrics. 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.

GHOSAL, S. and VAN DER VAART, A. W. (2001). Entropies and rates of convergence of maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29** 1233–1263.

GRENANDER, U. (1981). *Abstract Inference*. Wiley, New York.

KOLMOGOROV, A. N. and TIHOMIROV. V. M. (1961). ε-entropy and ε-capacity of sets in function spaces. *Amer. Math. Soc. Transl. Ser. 2* **17** 277–364. [Translated from Russian: *Uspekhi Mat. Nauk* **14** (1959) 3–86.]

KOMLÓS, J., MAJOR, P. and TUSNÁDY, G. (1975). An approximation of partial sums of independent R. V.'s and the sample DF. I. *Z. Wahrsch. Verw. Gebiete* **32** 111–131.

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.

LO, A. Y. (1987). A large sample study of the Bayesian bootstrap. *Ann. Statist.* **15** 360–375.

LORENZ, G. G. (1953). *Bernstein Polynomials*. Univ. Toronto Press.

MCLACHLAN, G. and BASFORD, K. (1988). *Mixture Models: Inference and Applications to Clustering*. Dekker, New York.

PETRONE, S. (1999a). Random Bernstein polynomials. *Scand. J. Statist.* **26** 373–393.

PETRONE, S. (1999b). Bayesian density estimation using Bernstein polynomials. *Canad. J. Statist.* **26** 373–393.

PETRONE, S. and WASSERMAN, L. (2001). Consistency of Bernstein polynomial posteriors. *J. Roy. Statist. Soc. Ser. B.* To appear.

RUBIN, D. (1981). The Bayesian bootstrap. *Ann. Statist.* **9** 130–134.

SCHWARTZ, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* **4** 10–26.

SHEN, X. and WASSERMAN, L. (2001). Rates of convesrgence 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.

TENBUSCH, A. (1994). Two-dimensional Bernstein polynomial density estimators. *Metrika* **41** 233–253.

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.

VITALE, R. A. (1975). A Bernstein polynomial approach to density estimation. In *Statistical Inference and Related Topics* (M. L. Puri, ed.) **2** 87–100. Academic Press, 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.

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