Aust. N. Z. J. Stat. **49**(3), 2007, 209–219 doi: 10.1111/j.1467-842X.2007.00476.x

ON CONVERGENCE RATES FOR NONPARAMETRIC POSTERIOR DISTRIBUTIONS

Antonio Lijoi[1], Igor Prünster[2] and Stephen G. Walker[3]*

University of Pavia, University of Turin and University of Kent

Summary

Rates of convergence of Bayesian nonparametric procedures are expressed as the maximum between two rates: one is determined via suitable measures of concentration of the prior around the "true" density f_0, and the other is related to the way the mass is spread outside a neighborhood of f_0. Here we provide a lower bound for the former in terms of the usual notion of prior concentration and in terms of an alternative definition of prior concentration. Moreover, we determine the latter for two important classes of priors: the infinite–dimensional exponential family, and the Pólya trees.

Key words: Chi-squared distance; Hellinger consistency; Posterior consistency; Posterior distribution; Rates of convergence.

1. Introduction

The Bayesian nonparametric methodology has undergone a series of criticisms from a frequentist point of view in the last two decades. The first, set forth in Diaconis & Freedman (1986a,b), concerned possible "frequentist" inconsistency of Bayesian nonparametric procedures. The "frequentist" or "what if" approach consists in assuming a "true" fixed distribution F_0 and checking whether the sequence of posterior distributions accumulates in suitable neighbourhoods of F_0. Up to the appearance of these papers, it was commonly believed that Bayes estimates would always be consistent for distributions in the support of the prior distribution. Consequently, their example of inconsistency, related to the estimation of a location parameter with a Dirichlet process prior, had a remarkable impact. Whereas the issue of weak consistency can be fixed by resorting to the "Kullback–Leibler support condition", due to Schwartz (1965), the question whether Bayesian nonparametric priors are strongly or Hellinger consistent still had to be answered: indeed, within the context of density estimation, the correct notion of convergence is the strong one. Given the criticism from a frequentist perspective, it appeared natural to rely on frequentist tools for facing it: indeed, Barron, Schervish & Wasserman (1999), Ghosal, Ghosh & Ramamoorthi (1999), and Petrone & Wasserman (2002) achieved their results by resorting to uniformly consistent tests, combined with the construction of suitable sieves and computation of metric entropies. A novel

*Author to whom correspondence should be addressed.

[1] Dipartimento Economia Politica e Metodi Quantitatavi, Università degli Studi di Pavia, via San Felice 5, 27100 Pavia and CNR–IMATI, Milano, Italy.

[2] Dipartimento Statistica e Matematica Applicata, Collegio Carlo Alberto and ICER, Università degli Studi di Torino, Piazza Arbarello 8, 10122 Torino, Italy.

[3] Institute of Mathematics, Statistics and Actuarial Science, University of Kent, Kent CT2 7NZ, UK.
e-mail: S.G.Walker@kent.ac.uk

Acknowledgements. A. Lijoi and I. Prünster were partially supported by MiUR, grants 2006/134525 and 2006/133449, respectively. The work of S.G. Walker is financed by an EPSRC Advanced Research Fellowship.

and genuinely Bayesian method for solving the issue can be found in Walker (2004), where a simple sufficient condition in terms of summability of prior probabilities is provided. This method has been fruitfully applied to nonparametric normal mixtures in Lijoi, Prünster & Walker (2005). Reasons for which Bayesian consistency should be faced with Bayesian tools are further investigated and explained in Walker, Lijoi & Prünster (2005).

Having successfully settled the issue of consistency, a second set of criticisms was directed at rates of convergence: whereas many results on rates of convergence are known in a classical nonparametric framework, within the Bayesian counterpart no such results are available and, even if they were, the resulting rates would be worse than the frequentist ones. First results on rates of convergence within infinite–dimensional models were provided by Shen & Wasserman (2001), Ghosal, Ghosh & van der Vaart (2000), Ghosal & van der Vaart (2001), and Ghosal (2001) relying upon a frequentist machinery and, indeed, they are often "suboptimal" with respect to those arising from classical sieve maximum likelihood estimators. The ultimate rate of convergence they achieve depends on two quantities: the prior mass assigned to suitable neighbourhoods of F_0, and the "entropy rate". A new approach for the determination of rates in Bayesian infinite–dimensional models, based on the consistency condition of Walker (2004), has been developed by Walker, Lijoi & Prünster (2007), where an alternative general theorem for the determination of rates is provided. This has allowed the authors to improve on known rates for Bernstein polynomials and mixtures of the Dirichlet process models, matching them with the frequentist rate of convergence for the sieve maximum likelihood estimator (MLE).

Starting from the well–known fact that the posterior rate of convergence arises as the maximum of two rates, that we denote as η_n and ϕ_n, here we examine in some detail each of them separately. We evaluate a lower bound for ϕ_n by exploiting its connection with concentration rates. In doing this, we consider a common notion of concentration typically used in the literature, and an alternative description of concentration expressed in terms of the χ^2 distance on the space of densities. As can be seen in Section 2, the lower bounds in the two cases essentially coincide with $\sqrt{(\log n)/n}$, thus suggesting a slower overall convergence if compared with the parametric case, where a rate of $n^{-1/2}$ is typically achieved. Moreover, in Section 3 we determine η_n for two popular classes of nonparametric priors, i.e. the infinite–dimensional exponential family and the Pólya trees. A discussion on the relation between ϕ_n and η_n is developed in Section 4. Proofs are postponed to the Appendix.

2. A lower bound for the concentration rate

2.1. Background and notation

Some further background and notation are introduced. Consider a sequence of observations $(X_n)_{n\geq 1}$ each taking values in some Polish space $\mathbb{X}$ endowed with a σ-algebra $\mathscr{X}$. If $\mathbb{F}$ indicates the space of probability density functions with respect to some measure λ on $\mathbb{X}$, the Hellinger metric h on $\mathbb{F}$ is defined by

$$h(f,g)=\left(\int_{\mathbb{X}}\left(\sqrt{f(x)}-\sqrt{g(x)}\right)^2\lambda(\mathrm{d}x)\right)^{1/2}$$

for any f and g in $\mathbb{F}$, and set $\mathscr{F}$ to be the Borel σ-algebra of $\mathbb{F}$. It is well–known that convergence in this metric is equivalent to convergence in the L_1 distance, the latter being defined by

$$d_1(f,g) = \int_{\mathbb{X}} |f(x) - g(x)| \, \lambda(\mathrm{d}x)$$

for any f and g in $\mathbb{F}$. Indeed, the following inequality holds true:

$$h^2(f,g) \leq d_1(f,g) < h(f,g).$$

Moreover, it is worth noting that d_1 is essentially the total variation distance on the space of probability measures, i.e. $d_1(f,g) = 2\, d_{TV}(F,G) = 2 \, \sup_A |F(A) - G(A)|$,where F and G are probability measures having densities f and g, respectively. Finally, two other stronger notions of divergence between densities to be considered are the Kullback–Leibler divergence and χ^2 distance. The former is defined by

$$d_{KL}(f,g) = \int_{\mathbb{X}} g(x) \, \log\left(\frac{g(x)}{f(x)}\right) \lambda(\mathrm{d}x),$$

whereas the latter is given by

$$d_2(f,g) = \int_{\mathbb{X}} \frac{g^2(x)}{f(x)} \, \lambda(\mathrm{d}x) - 1.$$

They are stronger in the sense that convergence with respect to (w.r.t.) d_2 implies convergence w.r.t. d_{KL} which, in turn, entails convergence w.r.t. h.

In a Bayesian nonparametric framework, $\mathbb{F}$ can be seen as an infinite–dimensional parameter space. Hence, if Π stands for a prior distribution on $(\mathbb{F}, \mathscr{F})$, the posterior distribution, given the observations $(X_1, \ldots, X_n)$, coincides with

$$\Pi_n(B) = \frac{\int_B \prod_{i=1}^n f(X_i) \, \Pi(\mathrm{d}f)}{\int_{\mathbb{F}} \prod_{i=1}^n f(X_i) \, \Pi(\mathrm{d}f)}$$

for all B in $\mathscr{F}$. In order to check consistency and related rates of convergence of Bayesian procedures according to the "frequentist" approach, one assumes that there exists a "true" density function f_0 such that the observations X_n are i.i.d. from f_0. A sequence of posterior distributions Π_n is said to be Hellinger consistent at f_0, if the posterior mass on sets of the type $A_\varepsilon = \{f\colon\ h(f, f_0) > \varepsilon\}$ becomes negligible as the sample size n increases. Having established consistency at f_0, the next step consists of determining the rate of convergence of Π_n to a point–mass at f_0. This issue can be formalized as the problem of finding a sequence $(\varepsilon_n)_{n\geq 1}$ such that $\varepsilon_n \downarrow 0$ and

$$\Pi_n(\{f \in \mathbb{F} : \ h(f, f_0) > M\varepsilon_n\}) \to 0 \tag{1}$$

for some constant $M > 0$, as $n \to \infty$. The above displayed convergence is understood as convergence in F_0^∞–probability, where F_0 denotes the probability distribution associated with f_0 and F_0^∞ is the distribution of the whole sequence $(X_n)_{n\geq 1}$ which makes the observations X_n i.i.d. from F_0. In the literature, there currently are two general approaches for tackling the problem as described in (1). Both arrive at the conclusion that the rate of convergence is given by

$$\varepsilon_n = \max\{\eta_n, \phi_n\}, \tag{2}$$

where the determining rates η_n and ϕ_n arise from two conceptually different sets of conditions. In the first approach, due to Ghosal *et al.* (2000), η_n depends on the growth rate of the Hellinger metric entropy whereas, in the second, due to Walker *et al.* (2007), η_n is determined by a condition expressed in terms of the sum of square roots of prior probabilities. A deeper discussion on η_n is postponed to Section 3. On the other hand, both approaches share the same condition for the evaluation of ϕ_n: as we will shortly see, ϕ_n admits a straightforward interpretation in terms of a prior concentration rate. In the following Subsection, we provide a new alternative procedure for the evaluation of ϕ_n and a lower bound for it.

2.2. The lower bound

Two notions are introduced which will be used to evaluate prior concentration around f_0 and, consequently, ϕ_n. The first one, not considered before, is defined with respect to the χ^2– distance d_2, and will be denoted by

$$\pi(\beta_n) = \Pi(\{f \in \mathbb{F} : d_2(f, f_0) < \beta_n\}), \tag{3}$$

where $(\beta_n)_{n\geq 1}$ is a sequence of positive numbers with $\beta_n \to 0$ as $n \to \infty$. In the sequel, when dealing with the χ^2–distance we will require f_0 to be in the χ^2–support of the prior. This entails that $\pi(\beta) > 0$ for all $\beta > 0$. Standard arguments, see for example Barron *et al.* (1999), can be used to show that χ^2–sets are measurable. The other notion, introduced in Wong & Shen (1995), is based on a combination of the Kullback–Leibler divergence and the $L_2(F_0)$–norm of $\log(f_0/f)$. If $V(f, f_0) = \int (\log(f_0(x)/f(x)))^2\, f_0(x)\lambda(\mathrm{d}x)$, a neighbourhood in terms of which concentration is measured can be defined as

$$B(\delta_n, f_0) = \{f \in \mathbb{F} : d_{KL}(f, f_0) \leq \delta_n^2, \quad V(f, f_0) \leq \delta_n^2\}. \tag{4}$$

Hence, $\Pi(B(\delta_n, f_0))$ can be used to define a measure of prior concentration around f_0. Let us now set

$$I_n = \int_{\mathbb{F}} R_n(f)\, \Pi(\mathrm{d}f),$$

where $R_n(f) = \prod_{i=1}^n f(X_i)/f_0(X_i)$. The interest in I_n is due to the fact that it coincides with the denominator of the posterior, rewritten in terms of $R_n(f)$. Hence, when one aims at establishing both consistency and convergence rates, it is essential to achieve an appropriate lower bound for I_n. Indeed, the sequence $(\phi_n)_{n\geq 1}$ defining the overall rate in (2) must be such that $\phi_n \to 0$, $n\phi_n^2 \to +\infty$ and $I_n > e^{-n\phi_n^2}$ in F_0^∞–probability. This important aspect is dealt with in the following lemma, which relates ϕ_n with β_n in (3).

Lemma 1. *Assume that f_0 is in the χ^2–support of Π and let $(\beta_n)_{n\geq 1}$ be a positive sequence such that $\beta_n \to 0$. Then $I_n > e^{-n\phi_n^2}$ in F_0^∞–probability if*

$$\mathrm{e}^{-n\phi_n^2}\, \frac{(1+\beta_n)^n}{\pi(\beta_n)} \to 0. \tag{5}$$

An implication of the statement of Lemma 1 is the fact that the sufficient condition (5) leads to an interesting connection between ϕ_n and β_n. Indeed, if $\beta_n \downarrow 0$ is such that

$$\pi(\beta_n) > \exp(-cn\log(1+\beta_n)) \tag{6}$$

for some $c > 0$ for all large n, then β_n can be termed the χ^2 *prior concentration rate*. From (5), if β_n is the χ^2 prior concentration rate, then one can take

$$\phi_n = \kappa\sqrt{-\frac{\log\pi(\beta_n)}{n} + \log(1+\beta_n)} < \kappa\sqrt{(1+c)\log(1+\beta_n)}, \tag{7}$$

for some $\kappa > 1$, in (2). Clearly, one has $\phi_n \to 0$ and $n\phi_n^2 \to \infty$. It is possible to deduce a lower bound for the χ^2 prior concentration rate which, combined with (5), yields a lower bound for ϕ_n when ϕ_n is measured in terms of the χ^2 prior concentration rate. Recall that Π does not have point masses in $\mathbb{F}$ if $\Pi(\{f\}) = 0$ for any f in $\mathbb{F}$, a natural requirement in a nonparametric setting.

Theorem 1. *Suppose Π does not have point masses in $\mathbb{F}$ and f_0 is in the χ^2–support of Π. Then $\beta_n > C\,(\log n)^{-\alpha}\, n^{\tau-1}$, for $\alpha > 1$ and $\tau \in (0,1)$, and*

$$\phi_n > C'\sqrt{\frac{\log n}{n}}$$

for n large enough.

Hence, if one evaluates ϕ_n in terms of the χ^2 prior concentration rate, a lower bound for the overall rate ε_n is essentially equal to $\sqrt{(\log n)/n}$.

Another approach for evaluating ϕ_n is based on the concentration on balls around f_0, defined as in (4). If

$$\Pi\{B(\phi_n, f_0)\} \geq \exp\big(-Cn\phi_n^2\big), \tag{8}$$

then ϕ_n is termed the prior concentration rate without any further specification since it is the one most widely used. The usefulness of this type of concentration can be seen from some recent contributions in the literature, such as, e.g., Shen & Wasserman (2001) and Ghosal *et al.* (2000), where it is shown that, if ϕ_n is the prior concentration rate according to (8), then $I_n > \mathrm{e}^{-cn\phi_n^2}$ in F_0^∞–probability, for some positive constant c. Analogously to what has been done for the χ^2 prior concentration rate, we can provide a lower bound for ϕ_n when this is evaluated in terms of the usual prior concentration rate.

Theorem 2. *Suppose Π does not have point masses in $\mathbb{F}$ and $\Pi(B(\varepsilon, f_0)) > 0$ for any $\varepsilon > 0$. Then $\phi_n \geq C'\sqrt{(\log n)/n}$ for n large enough.*

The above statement yields a similar conclusion to Theorem 1 in terms of a lower bound for the overall rate ε_n. Indeed, if this is evaluated in terms of the usual prior concentration rate, we have that the rate of convergence cannot be faster than $\sqrt{(\log n)/n}$. Hence, using two different notions of concentration around the true f_0 yields the same lower bound for ε_n. This finding suggests that a nonparametric Bayesian procedure cannot achieve the parametric rate of convergence of $n^{-1/2}$. This is not a surprising fact since, in the nonparametric case, one has to spread the prior mass on an infinite–dimensional parameter space.

3. On the determination of η_n

In the previous section we focussed on the evaluation of ϕ_n in (2), whereas the present Section is devoted to investigating η_n. In doing this we undertake the approach set forth in Walker *et al.* (2007). To this end, take the set $A_{\eta_n} = \{f : h(f, f_0) > \eta_n\}$ and consider a covering $\{A_{n,j} : j = 1, 2, \ldots\}$ of A_{η_n}, where each $A_{n,j}$ has radius, w.r.t. the Hellinger distance, $\tau_n \in (0, \eta_n)$. Consequently, define $K_{\eta_n} = \sum_{j\geq 1} \Pi(A_{n,j})^{1/2}$. From Walker *et al.* (2007), we recall the general theorem to be exploited for the determination of the rate of convergence.

Theorem 3. *Suppose* $\eta_n, \phi_n \to 0$ *and* $n\eta_n^2, n\phi_n^2 \to +\infty$ *and*

(i) $\mathrm{e}^{-n\eta_n^2/16}\, K_{\eta_n} \to 0$,

(ii) *for some* $C > 0$, $\Pi(B(\phi_n, f_0)) \geq \exp(-Cn\phi_n^2)$.

Then $\Pi_n(A_{\phi_n}) \to 0$ *in* F_0^∞*–probability when* $\eta_n \leq C'\phi_n$ *for some small enough* $C' > 0$.

Note that condition (ii) above, previously referred to as the usual prior concentration rate, can be replaced by the condition in terms of the χ^2 prior concentration rate, as seen in the previous Section. We now focus on condition (i) and determine η_n for two important examples: the infinite–dimensional exponential family studied in Leonard (1978) and Lenk (1988, 1991), and the Pólya tree priors investigated in Lavine (1992) and Mauldin *et al.* (1992). See Müller & Quintana (2004) for references on various applications of these classes of priors. It has to be remarked that, to date, nothing is known for both priors in terms of rates of convergence.

3.1. Infinite–dimensional exponential family

A random density function belongs to the infinite–dimensional exponential family if it can be represented as

$$f(x) = \exp\left(\sum_{j=1}^{\infty} \theta_j \phi_j(x) - c(\Theta)\right),$$

where the $\{\theta_j\}$ are independent normal random variables with zero means and variances $\{\sigma_j^2\}$, the $\{\phi_j\}$ are an orthonormal basis on [0, 1] and $c(\Theta)$ is the normalizing constant. Conditions for consistency of such a model have been derived in Barron *et al.* (1999) and by Walker (2004). In the latter paper it has been shown that

$$K_\delta \leq \prod_{j=1}^{\infty} \left(1 + \psi_j/\delta^{3/2}\right)$$

for positive $\{\psi_j\}$, which depend on the variances $\{\sigma_j^2\}$. The sum of the ψ_js to a finite number is sufficient for $K_\delta < +\infty$ for all $\delta > 0$, which implies consistency. See Walker (2004) for details.

Consider condition (i) of Theorem 3. The rate sequence η_n is derived. Putting $\psi_j = \mathrm{e}^{-j}$ and denoting by $[x]$ the integer part of $x > 0$, we have

$$\begin{aligned}\log K_\delta &\le \sum_{j=1}^{\infty} \log(1 + \exp(-j + [-1.5 \log \delta])) \\ &= \sum_{j=1}^{[-1.5 \log \delta]} \log(1 + \exp(-j + [-1.5 \log \delta])) \\ &\quad + \sum_{j=[-1.5 \log \delta]+1}^{\infty} \log(1 + \exp(-j + [-1.5 \log \delta])).\end{aligned}$$

The second sum is clearly bounded by $\sum_{j=[-1.5 \log \delta]+1}^{\infty} \exp(-j + [-1.5 \log \delta])$ which is easily seen to be bounded by a finite number $\log M$ not depending on δ. For the first sum, it can be bounded by $\sum_{j=1}^{[-1.5 \log \delta]} (\log(\delta^{3/2} + e^{-j}) - 1.5 \log \delta)$ and note that, for small enough δ, one has $\delta^{3/2} + e^{-j} < 1$. Hence, it follows that $\log K_\delta < (-1.5 \log \delta)^2 + \log M$ and so $K_\delta < M \exp(9/4(\log(1/\delta))^2)$. Now taking any positive sequence $(\eta_n)_{n\ge1}$, we have

$$K_{\eta_n} < M \exp(9/4\big(\log(1/\eta_n))^2\big)$$

and then it can be seen that Condition (i) of Theorem 3 is satisfied with

$$\eta_n = C \frac{\log n}{\sqrt{n}}.$$

3.2. Pólya tree priors

Pólya tree priors are random densities defined according to a suitable tree of nested partitions of the interval [0, 1]. Here we do consider binary partitions, so that at level k the sets $B_{k,j}$, with $j = 1, \ldots, 2^k$, partition [0, 1]. To each of these sets associate a random variable $\theta_{k,j}$ having a Beta(a_k, a_k) distribution restricted to $B_{k,j}$ when j is odd, whereas $\theta_{k,j} = 1 - \theta_{k,j-1}$ when j is even. If $\sum_{k\ge1} a_k^{-1} < \infty$, a random density function, with respect to the Lebesgue measure on [0, 1], is defined by

$$\tilde{f}(x) = \lim_{k\to\infty} 2^k \prod_{j=1}^{k} \theta_{k,j(x)},$$

where $j(x)$ identifies the specific set, at level k within the tree of partitions, where x lies. Sufficient conditions for consistency have been provided by Barron *et al.* (1999) and significantly improved in Walker (2004). Such a definition of Pólya trees yields some analogy with the infinite–dimensional exponential family in terms of the derivation of K_δ. Indeed, if we set $b_j \approx \delta_j/4$, $\delta_j = \gamma_j\, \delta$, $(\gamma_j)_{j\ge1}$ a positive sequence such that $\sum_{j\ge1} \gamma_j < \infty$, from Walker (2004) one can deduce that

$$\begin{aligned}\log K_\delta &\le \sum_{j\ge1} \log\left(1 + \frac{M\, a_j^{1/4}\big(1 - 4b_j^2\big)^{a_j/2-1/2}}{\sqrt{\delta_j}}\right) \le \sum_{j\ge1} \log\left(1 + \frac{M'\, a_j^{1/4}\, \mathrm{e}^{-2a_j b_j^2}}{\sqrt{\delta_j}}\right) \\ &\le \sum_{j\ge1} 2^{j-1} \log\left(1 + \frac{\psi_j}{\delta^{5/2}}\right)\end{aligned}$$

for some positive constants M and M', and the last inequality follows from $\exp(-\lambda\delta^2) < 1/(\lambda\delta^2)$ for all $\lambda > 0$. Finally, for the sake of simplicity we fixed $\psi_j = C'\, a_j^{-3/4}\, \gamma_j^{-5/2}$. If we, now, take $\psi_j = \exp(-e^j)$,

$$\log K_\delta \le \sum_{j=1}^{\infty} 2^{j-1} \log(1 + \exp(-(e^j + 2.5 \log \delta))). \tag{9}$$

Note that the specific choice of ψ_j suggests that $a_j = \exp\,(4\,\mathrm{e}^j/3)\,\gamma_j^{-10/3}$. As for the infinite–dimensional exponential family, we split the right hand side of (9) into two sums: one from $j = 1$ to $j = [\log(-2.5 \log \delta)]$ and the other from $j = [\log(-2.5 \log \delta)] + 1$ to $+\infty$. The latter is bounded by a finite number $\log C$ for all $\delta > 0$, whereas for the first term we can write $\sum_{j=1}^{[\log(-2.5\log\delta)]} 2^{j-1}(\log(\exp(-e^j) + \delta^{5/2}) - 2.5\log\delta)$. For small enough δ this is bounded by $2^{\log(-2.5\log\delta)}\,(-2.5 \log \delta)$ and so $\log K_\delta < (-2.5 \log \delta)^2 + \log C$. Hence we arrive to the conclusion that Condition (i) of Theorem 3 is satisfied with

$$\eta_n = C\frac{\log n}{\sqrt{n}}.$$

4. Discussion

A natural question at this point is which of the two rates, ϕ_n and η_n, in (2) dominates in determining the overall rate of convergence. Recall that ϕ_n stems from the evaluation of a suitable concentration rate and η_n can be either determined via (i) in Theorem 3 or via the growth rate of the metric entropy according to the approach set forth in Ghosal *et al.* (2000). It is our opinion that ϕ_n should dominate. On the one side this statement is supported by intuition. On the other side, the behaviour of specific priors seems to suggest a sort of general dominance. For example, for the Bernstein polynomial prior (Petrone, 1999a,b), Ghosal (2001) showed that, if f_0 is a Bernstein polynomial itself, $\phi_n = \sqrt{(\log n)/n}$. But, in the more interesting case in which f_0 is just required to satisfy some suitable regularity conditions and needs not be of Bernstein type, then $\phi_n = (\log n)^{1/3}/n^{1/3}$. Moreover, as noted in Walker *et al.* (2007), if the weights in the Bernstein representation decay sufficiently fast, one can attain $\eta_n = \log n/\sqrt{n}$. Hence, in this case, ϕ_n clearly dominates over η_n and determines the rate of convergence. As for the two priors examined in the previous Section, we obtained an estimate of η_n which does not depend on the specific f_0 and is close to the lower bound for ϕ_n. If one allows f_0 to be reasonably general, intuition suggests that the prior concentration increases (as also happens for the Bernstein case) and one expects ϕ_n to dominate over $\eta_n = \log n/\sqrt{n}$.

A final remark concerns a comparison with the frequentist sieve maximum likelihood estimator. For example, if the prior is a normal mixture of the Dirichlet process (Lo, 1984), in Walker *et al.* (2007) it is shown, under suitable conditions, that ϕ_n coincides with η_n and the rate of convergence is $\log n/\sqrt{n}$. This is also the best rate of convergence achievable by the sieve MLE, as shown by Ghosal & van der Vaart (2001). Moreover, in the Bernstein case, the above mentioned rate $\varepsilon_n = (\log n)^{1/3}/n^{1/3}$ coincides with the one of the corresponding sieve MLE determined by Ghosal (2001). It might not be a coincidence that, in cases where one is able to show that the concentration rate dominates, the posterior rate of convergence is the same as the rate of convergence for the frequentist sieve MLE.

5. Appendix

A1. Proof of Lemma 1. Define

$$I_{n,A} = \int_A R_n(f)\,\Pi(\mathrm{d}f).$$

Then

$$F_0^\infty\big(I_n < e^{-n\delta_n^2}\big) = F_0^\infty\big(I_n^{-1} > e^{n\delta_n^2}\big) < \mathrm{e}^{-n\delta_n^2}\,E_0\big(I_n^{-1}\big),$$

where E_0 denotes the expectation computed w.r.t. the probability distribution F_0^∞. Now it is easy to check that

$$\frac{I_{n+1,A}}{I_{n,A}} = \frac{f_{n,A}(X_{n+1})}{f_0(X_{n+1})},$$

where $f_{n,A}(x) = \int_A f(x)\,\Pi_n(\mathrm{d}f)/\Pi_n(A)$ is, for any A in $\mathscr{F}$, the predictive density with the posterior restricted and normalized to A. Therefore,

$$E_0\big(I_{n+1,A}^{-1} \mid \mathscr{F}_n\big) = I_{n,A}^{-1}\,(1 + d_2(f_{n,A}, f_0)).$$

Here $\mathscr{F}_n = \sigma(X_1, \ldots, X_n)$. Due to the convexity of the d_2 metric, setting $A = \{f \in \mathbb{F} : d_2(f, f_0) < \beta_n\}$ yields

$$E_0\big(I_{n+1,A}^{-1} \mid \mathscr{F}_n\big) < I_{n,A}^{-1}\,(1 + \beta_n),$$

and hence

$$E_0\big(I_{n,A}^{-1}\big) < \frac{(1+\beta_n)^n}{\pi(\beta_n)}.$$

This follows because $I_{0,A} = \pi(\beta_n)$. Using the obvious inequality that $I_n^{-1} < I_{n,A}^{-1}$, we have that

$$E_0\big(I_n^{-1}\big) < \frac{(1+\beta_n)^n}{\pi(\beta_n)}.$$

Consequently,

$$F_0^\infty\big(I_n < e^{-n\delta_n^2}\big) \to 0$$

when (5) holds true. ■

A.2. Proof of Theorem 1 and Theorem 2. The proofs of both results require the following preliminary lemma.

Lemma 2. *Assume the prior* Π *does not have point masses on* $\mathbb{F}$. *If* $\beta_n \downarrow 0$, *then*

$$\limsup_n \beta_n^{-1}\pi(\beta_n) < +\infty. \tag{10}$$

Analogously, one finds that $\limsup_n \delta_n^{-2}\,\Pi(B(\delta_n, f_0)) < \infty$.

Proof. We have that

$$\pi(\beta_n) < \Pi(\{f : h(f, f_0) < \beta_n\}) < \Pi(\{f : |F(A) - F_0(A)| < \beta_n\})$$

for any set A in $\mathscr{X}$. The fact that Π does not have point masses implies that the probability distribution of $F(A)$ is either absolutely continuous w.r.t. the Lebesgue measure on $[0, 1]$ or it is singular. In the first case, if g is the density function of $F(A)$, choose A such that $g(F_0(A)) < +\infty$ and g is continuous at $F_0(A)$. Then

$$\Pi(\{f : |F(A) - F_0(A)| < \beta_n\}) = \int_{F_0(A)-\beta_n}^{F_0(A)+\beta_n} g(s)\,\mathrm{d}s$$

and

$$\lim_n \beta_n^{-1} \int_{F_0(A)-\beta_n}^{F_0(A)+\beta_n} g(s)\,\mathrm{d}s = 2g(F_0(A)) < +\infty.$$

If the distribution of $F(A)$ is singular, then its density is, almost everywhere, 0 and (10) is trivially true. The statement for $\delta_n^{-2}\Pi(B(\delta_n, f_0))$ follows immediately due to the analogous inclusion of sets used at the beginning of this proof. ■

Proof of Theorem 1. From Lemma 1 it follows that there exists a constant K such that $\pi(\beta_n) \leq K\beta_n$ for n large enough. Now, if we allow β_n to be such that

$$\pi(\beta_n) > \exp(-cn \log(1 + \pi(\beta_n)/K)), \tag{11}$$

then we have

$$\pi(\beta_n) > \exp(-cn \log(1 + \beta_n)),$$

i.e. β_n is the χ^2 prior concentration rate. Hence, by (7) one has

$$\phi_n = \kappa\sqrt{-\log(\pi(\beta_n))/n + \log(1 + \beta_n)}.$$

The fastest possible $\pi(\beta_n)$ consistent with inequality (11) is given by $(\log n)^{-\alpha}\ n^{\tau-1}$ for $\alpha > 0$ and $\tau \in (0, 1)$. By Lemma 2, for large enough n one has $\beta_n > K(\log n)^{-\alpha}\ n^{\tau-1}$. The lower bound for ϕ_n, at this point, is immediate. ■

Proof of Theorem 2. By virtue of Lemma 2 we have $\Pi(B(\phi_n, f_0)) < M\ \phi_n^2$ for some positive constant M. Hence the inequality (8) is valid if

$$\Pi(B(\phi_n, f_0)) > \exp(-C'n\Pi(B(\phi_n, f_0))).$$

For this we can take $\Pi(B(\phi_n, f_0)) = (\log n)^{\beta}/n^{\alpha}$ and yet the best possible rate is obtained with $\alpha = \beta = 1$. So, $\phi_n > M'\sqrt{(\log n)/n}$ for some positive constant M'. ■

References

Barron, A., Schervish, M. J. & Wasserman, L. (1999). The consistency of distributions in nonparametric problems. *Ann. Statist.* **27**, 536–561.

Diaconis, P. & Freedman, D. (1986a). On the consistency of Bayes estimates. *Ann. Statist.* **14**, 1–26.

DIACONIS, P. & FREEDMAN, D. (1986b). On inconsistent Bayes estimates of location. *Ann. Statist.* **14**, 68–87.

GHOSAL, S. (2001). Convergence rates for density estimation with Bernstein polynomials. *Ann. Statist.* **29**, 1264–1280.

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

GHOSAL, S., GHOSH, J. K. & RAMAMOORTHI, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27**, 143–158.

GHOSAL, S., GHOSH, J. K. & VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28**, 500–531.

LAVINE, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.* **20**, 1222–1235.

LENK, P. J. (1988). The logistic normal distribution for Bayesian, nonparametric, predictive densities. *J. Amer. Statist. Assoc.* **83**, 509–516.

LENK, P. J. (1991). Towards a practicable Bayesian nonparametric density estimator. *Biometrika* **78**, 531–543.

LEONARD, T. (1978). Density estimation, stochastic processes and prior information. *J. Roy. Statist. Soc., Ser. B* **40**, 113–146.

LIJOI, A., PRÜNSTER, I. & WALKER, S. G. (2005). On consistency of nonparametric normal mixtures for Bayesian density estimation. *J. Amer. Statist. Assoc.* **100**, 1292–1296.

LO, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Statist.* **12**, 351–357.

MAULDIN, R. D., SUDDERTH, W. D. & WILLIAMS, S. C. (1992). Pólya trees and random distributions. *Ann. Statist.* **20**, 1203–1221.

MÜLLER, P. & QUINTANA, F. A. (2004). Nonparametric Bayesian data analysis. *Statist. Science* **19**, 95–110.

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

PETRONE, S. (1999b). Bayesian density estimation using Bernstein polynomials. *Canad. J. Statist.* **27**, 105–126.

PETRONE, S. & WASSERMAN, L. (2002). Consistency of Bernstein polynomial posteriors. *J. Roy. Stat. Soc. B* **64**, 79–100.

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

SHEN, X. & WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29**, 687–714.

WALKER, S. G. (2004). New approaches to Bayesian consistency. *Ann. Statist.* **32**, 2028–2043.

WALKER, S. G., LIJOI, A. & PRÜNSTER, I. (2005). Data–tracking and the understanding of Bayesian consistency. *Biometrika* **92**, 765–778.

WALKER, S.G., LIJOI, A. & PRÜNSTER, I. (2007). On rates of convergence for posterior distributions in infinite dimensional models. *Ann. Statist.* **35**, (forthcoming).

WONG, W. H. & SHEN, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.* **23**, 339–362.