

2

The Dirichlet process, related priors and posterior asymptotics

Subhashis Ghosal

Here we review the role of the Dirichlet process and related prior distributions in nonparametric Bayesian inference. We discuss construction and various properties of the Dirichlet process. We then review the asymptotic properties of posterior distributions. Starting with the definition of posterior consistency and examples of inconsistency, we discuss general theorems which lead to consistency. We then describe the method of calculating posterior convergence rates and briefly outline how such rates can be computed in nonparametric examples. We also discuss the issue of posterior rate adaptation, Bayes factor consistency in model selection and Bernstein–von Mises type theorems for nonparametric problems.

2.1 Introduction

Making inferences from observed data requires modeling the data-generating mechanism. Often, owing to a lack of clear knowledge about the data-generating mechanism, we can only make very general assumptions, leaving a large portion of the mechanism unspecified, in the sense that the distribution of the data is not specified by a finite number of parameters. Such nonparametric models guard against possible gross misspecification of the data-generating mechanism, and are quite popular, especially when adequate amounts of data can be collected. In such cases, the parameters can be best described by functions, or some infinite-dimensional objects, which assume the role of parameters. Examples of such infinite-dimensional parameters include the cumulative distribution function (c.d.f.), density function, nonparametric regression function, spectral density of a time series, unknown link function in a generalized linear model, transition density of a Markov chain and so on. The Bayesian approach to nonparametric inference,

Bayesian Nonparametrics, ed. Nils Lid Hjort, Chris Holmes, Peter Müller and Stephen G. Walker. Published by Cambridge University Press. © Cambridge University Press 2010.

however, faces challenging issues since construction of prior distribution involves specifying appropriate probability measures on function spaces where the parameters lie. Typically, subjective knowledge about the minute details of the distribution on these infinite-dimensional spaces is not available for nonparametric problems. A prior distribution is generally chosen based on tractability, computational convenience and desirable frequentist behavior, except that some key parameters of the prior may be chosen subjectively. In particular, it is desirable that a chosen prior is spread all over the parameter space, that is, the prior has large topological *support*. Together with additional conditions, large support of a prior helps the corresponding posterior distribution to have good frequentist properties in large samples. To study frequentist properties, it is assumed that there is a true value of the unknown parameter which governs the distribution of the generated data.

We are interested in knowing whether the posterior distribution eventually concentrates in the neighborhood of the true value of the parameter. This property, known as *posterior consistency*, provides the basic frequentist validation of a Bayesian procedure under consideration, in that it ensures that with a sufficiently large amount of data, it is nearly possible to discover the truth accurately. Lack of consistency is extremely undesirable, and one should not use a prior if the corresponding posterior is inconsistent. However, consistency is satisfied by many procedures, so typically more effort is needed to distinguish between consistent procedures. The speed of convergence of the posterior distribution to the true value of the parameter may be measured by looking at the smallest shrinking ball around the true value which contains posterior probability nearly one. It will be desirable to pick up the prior for which the size of such a shrinking ball is the minimum possible. However, in general it is extremely hard to characterize size exactly, so we shall restrict ourselves only to the rate at which a ball around the true value can shrink while retaining almost all of the posterior probability, and call this the *rate of convergence* of the posterior distribution. We shall also discuss adaptation with respect to multiple models, consistency for model selection and *Bernshtein–von Mises theorems*.

In the following sections, we describe the role of the *Dirichlet process* and some related prior distributions, and discuss their most important properties. We shall then discuss results on convergence of posterior distributions, and shall often illustrate results using priors related to the Dirichlet process. At the risk of being less than perfectly precise, we shall prefer somewhat informal statements and informal arguments leading to these results. An area which we do not attempt to cover is that of Bayesian survival analysis, where several interesting priors have been constructed and consistency and

rate of convergence results have been derived. We refer readers to Ghosh and Ramamoorthi (2003) and Ghosal and van der Vaart (2009) as general references for all topics discussed in this chapter.

2.2 The Dirichlet process

2.2.1 Motivation

We begin with the simplest nonparametric inference problem for an uncountable sample space, namely, that of estimating a probability measure (equivalently, a c.d.f.) on the real line, with independent and identically distributed (i.i.d.) observations from it, where the c.d.f. is completely arbitrary. Obviously, the classical estimator, the empirical distribution function, is well known and is quite satisfactory. A Bayesian solution requires describing a random probability measure and developing methods of computation of the posterior distribution. In order to understand the idea, it is fruitful to look at the closest parametric relative of the problem, namely the multinomial model. Observe that the multinomial model specifies an arbitrary probability distribution on the sample space of finitely many integers, and that a multinomial model can be derived from an arbitrary distribution by grouping the data in finitely many categories. Under the operation of grouping, the data are reduced to counts of these categories. Let (π_1, \dots, π_k) be the probabilities of the categories with frequencies n_1, \dots, n_k . Then the likelihood is proportional to $\pi_1^{n_1} \dots \pi_k^{n_k}$. The form of the likelihood matches with the form of the finite-dimensional Dirichlet prior, which has density † proportional to $\pi_1^{c_1-1} \dots \pi_k^{c_k-1}$. Hence the posterior density is proportional to $\pi_1^{n_1+c_1-1} \dots \pi_k^{n_k+c_k-1}$, which is again a Dirichlet distribution.

With this nice conjugacy property in mind, Ferguson (1973) introduced the idea of a Dirichlet process – a probability distribution on the space of probability measures which induces finite-dimensional Dirichlet distributions when the data are grouped. Since grouping can be done in many different ways, reduction to a finite-dimensional Dirichlet distribution should hold under any grouping mechanism. In more precise terms, this means that for any finite measurable partition $\{B_1, \dots, B_k\}$ of \mathbb{R} , the joint distribution of the probability vector $(P(B_1), \dots, P(B_k))$ is a finite-dimensional Dirichlet distribution. This is a very rigid requirement. For this to be true, the parameters of the finite-dimensional Dirichlet distributions need to be very special. This is because the joint distribution of $(P(B_1), \dots, P(B_k))$ should

† Because of the restriction $\sum_{i=1}^k \pi_i = 1$, the density has to be interpreted as that of the first $k-1$ components.

agree with other specifications such as those derived from the joint distribution of the probability vector $(P(A_1), \dots, P(A_m))$ for another partition $\{A_1, \dots, A_m\}$ finer than $\{B_1, \dots, B_k\}$, since any $P(B_i)$ is a sum of some $P(A_j)$. A basic property of a finite-dimensional Dirichlet distribution is that the sums of probabilities of disjoint chunks again give rise to a joint Dirichlet distribution whose parameters are obtained by adding the parameters of the original Dirichlet distribution. Letting $\alpha(B)$ be the parameter corresponding to $P(B)$ in the specified Dirichlet joint distribution, it thus follows that $\alpha(\cdot)$ must be an additive set function. Thus it is a prudent strategy to let α actually be a measure. Actually, the countable additivity of α will be needed to bring in countable additivity of the random P constructed in this way. The whole idea can be generalized to an abstract Polish space.

Definition 2.1 Let α be a finite measure on a given Polish space \mathfrak{X} . A random measure P on \mathfrak{X} is called a Dirichlet process if for every finite measurable partition $\{B_1, \dots, B_k\}$ of \mathfrak{X} , the joint distribution of $(P(B_1), \dots, P(B_k))$ is a k -dimensional Dirichlet distribution with parameters $\alpha(B_1), \dots, \alpha(B_k)$.

We shall call α the base measure of the Dirichlet process, and denote the Dirichlet process measure by \mathcal{D}_α .

Even for the case when α is a measure so that joint distributions are consistently specified, it still remains to be shown that the random set function P is a probability measure. Moreover, the primary motivation for the Dirichlet process was to exploit the conjugacy under the grouped data set up. Had the posterior distribution been computed based on conditioning on the counts for the partitioning sets, we would clearly retain the conjugacy property of finite-dimensional Dirichlet distributions. However, as the full data are available under the setup of continuous data, a gap needs to be bridged. We shall see shortly that both issues can be resolved positively.

2.2.2 Construction of the Dirichlet process

Naive construction

At first glance, because joint distributions are consistently specified, viewing P as a function from the Borel σ -field \mathcal{B} to the unit interval, a measure with the specified marginals can be constructed on the uncountable product space $[0, 1]^{\mathcal{B}}$ with the help of Kolmogorov's consistency theorem. Unfortunately, this simple strategy is not very fruitful for two reasons. First, the product σ -field on $[0, 1]^{\mathcal{B}}$ is not rich enough to contain the space of probability measures. This difficulty can be avoided by working with outer measures,

provided that we can show that P is a.s. countably additive. For a given sequence of disjoint sets A_n , it is indeed true that $P(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$ a.s. Unfortunately, the null set involved in the a.s. statement is dependent on the sequence A_n , and since the number of such sequences is uncountable, the naive strategy using the Kolmogorov consistency theorem fails to deliver the final result.

Construction using a countable generator

To save the above construction, we need to work with a countable generating field \mathcal{F} for \mathcal{B} and view each probability measure P as a function from \mathcal{F} to $[0, 1]$. The previously encountered measure theoretic difficulties do not arise on the countable product $[0, 1]^{\mathcal{F}}$.

Construction by normalization

There is another construction of the Dirichlet process which involves normalizing a *gamma process* with intensity measure α . A gamma process is an independent increment process whose existence is known from the general theory of *Lévy processes*. The gamma process representation of the Dirichlet process is particularly useful for finding the distribution of the mean functional of P and estimating of the tails of P when P follows a Dirichlet process on \mathbb{R} .

2.2.3 Properties

Once the Dirichlet process is constructed, some of its properties are immediately obtained.

Moments and marginal distribution

Considering the partition $\{A, A^c\}$, it follows that $P(A)$ is distributed as $\text{Beta}(\alpha(A), \alpha(A^c))$. Thus in particular, $E(P(A)) = \alpha(A)/(\alpha(A) + \alpha(A^c)) = G(A)$, where $G(A) = \alpha(A)/M$, a probability measure and $M = \alpha(\mathbb{R})$, the total mass of α . This means that if $X|P \sim P$ and P is given the measure \mathcal{D}_α , then the marginal distribution of X is G . We shall call G the *center measure*. Also, observe that $\text{Var}(P(A)) = G(A)G(A^c)/(M + 1)$, so that the prior is more tightly concentrated around its mean when M is larger, that is, the prior is more precise. Hence the parameter M can be regarded as the *precision parameter*. When P is distributed as the Dirichlet process with base measure $\alpha = MG$, we shall often write $P \sim \text{DP}(M, G)$.

Linear functionals

If ψ is a G -integrable function, then $E(\int \psi dP) = \int \psi dG$. This holds for indicators from the relation $E(P(A)) = G(A)$, and then the standard measure theoretic arguments extend this sequentially to simple measurable functions, nonnegative measurable functions and finally to all integrable functions. The distribution of $\int \psi dP$ can also be obtained analytically, but this distribution is substantially more complicated than beta distribution followed by $P(A)$. The derivation involves the use of a lot of sophisticated machinery. Interested readers are referred to Regazzini, Guglielmi and Di Nunno (2002) and references therein.

Conjugacy

Just as the finite-dimensional Dirichlet distribution is conjugate to the multinomial likelihood, the Dirichlet process prior is also conjugate for estimating a completely unknown distribution from i.i.d. data. More precisely, if X_1, \dots, X_n are i.i.d. with distribution P and P is given the prior \mathcal{D}_α , then the posterior distribution of P given X_1, \dots, X_n is $\mathcal{D}_{\alpha + \sum_{i=1}^n \delta_{X_i}}$.[†] To see this, we need to show that for any measurable finite partition $\{A_1, \dots, A_k\}$, the posterior distribution of $(P(A_1), \dots, P(A_k))$ given X_1, \dots, X_n is k -dimensional Dirichlet with parameters $\alpha(A_j) + N_j$, where $N_j = \sum_{i=1}^n \mathbb{1}\{X_i \in A_j\}$, the count for A_j , $j = 1, \dots, k$. This certainly holds by the conjugacy of the finite-dimensional Dirichlet prior with respect to the multinomial likelihood had the data been coarsened to only the counts N_1, \dots, N_k . Therefore, the result will follow if we can show that the additional information contained in the original data X_1, \dots, X_n is irrelevant as far as the posterior distribution of $(P(A_1), \dots, P(A_k))$ is concerned. One can show this by first considering a partition $\{B_1, \dots, B_m\}$ finer than $\{A_1, \dots, A_k\}$, computing the posterior distribution of $(P(B_1), \dots, P(B_m))$ given the counts of $\{B_1, \dots, B_m\}$, and marginalizing to the posterior distribution of $(P(A_1), \dots, P(A_k))$ given the counts of $\{B_1, \dots, B_m\}$. By the properties of finite-dimensional Dirichlet, this coincides with the posterior distribution of $(P(A_1), \dots, P(A_k))$ given the counts of $\{A_1, \dots, A_k\}$. Now making the partitions infinitely finer and applying the martingale convergence theorem, the final result is obtained.

Posterior mean

The above expression for the posterior distribution combined with the formula for the mean of a Dirichlet process imply that the posterior mean

[†] Of course, there are other versions of the posterior distribution which can differ on a null set for the joint distribution.

of P given X_1, \dots, X_n can be expressed as

$$\tilde{\mathbb{P}}_n = \mathbb{E}(P|X_1, \dots, X_n) = \frac{M}{M+n}G + \frac{n}{M+n}\mathbb{P}_n, \quad (2.1)$$

a convex combination of the prior mean and the empirical distribution. Thus the posterior mean essentially shrinks the empirical distribution towards the prior mean. The relative weight attached to the prior is proportional to the total mass M , giving one more reason to call M the precision parameter, while the weight attached to the empirical distribution is proportional to the number of observations it is based on.

Limits of the posterior

When n is kept fixed, letting $M \rightarrow 0$ may be regarded as making the prior imprecise or noninformative. The limiting posterior, namely $\mathcal{D}_{\sum_{i=1}^n \delta_{X_i}}$, is known as the Bayesian bootstrap. Samples from the Bayesian bootstrap are discrete distributions supported at only the observation points whose weights are Dirichlet distributed, and hence the *Bayesian bootstrap* can be regarded as a resampling scheme which is smoother than Efron's bootstrap. On the other hand, when M is kept fixed and n varies, the asymptotic behavior of the posterior mean is entirely controlled by that of the empirical distribution. In particular, the c.d.f. of $\tilde{\mathbb{P}}_n$ converges uniformly to the c.d.f. of the true distribution P_0 and $\sqrt{n}(\tilde{\mathbb{P}}_n - P_0)$ converges weakly to a Brownian bridge process. Further, for any set A , the posterior variance of $P(A)$ is easily seen to be $\mathcal{O}(n^{-1})$ as $n \rightarrow \infty$. Hence Chebyshev's inequality implies that the posterior distribution of $P(A)$ approaches the degenerate distribution at $P_0(A)$, that is, the posterior distribution of $P(A)$ is consistent at P_0 , and the rate of this convergence is $n^{-1/2}$. Shortly, we shall see that the entire posterior of P is also consistent at P_0 .

Lack of smoothness

The presence of the point masses δ_{X_i} in the base measure of the posterior Dirichlet process gives rise to some peculiar behavior. One such property is the total disregard of the topology of the sample space. For instance, if A is a set such that many observations fall close to it but A itself does not contain any observed point, then the posterior mean of $P(A)$ is smaller than its prior mean. Thus the presence of observations in the vicinity does not enhance the assessment of the probability of a set unless the observations are actually contained there. Hence it is clear that the Dirichlet process is somewhat primitive in that it does not offer any smoothing, quite unlike the characteristic of a Bayes estimator.

Negative correlation

Another peculiar property of the Dirichlet process is negative correlation between probabilities of any two disjoint sets. For a random probability distribution, one may expect that the masses assigned to nearby places increase or decrease together, so the blanket negative correlation attached by the Dirichlet process may be disappointing. This again demonstrates that the topology of the underlying space is not considered by the Dirichlet process in its mass assignment.

Discreteness

A very intriguing property of the Dirichlet process is the discreteness of the distributions sampled from it, even when G is purely nonatomic. This property also has its roots in the expression for the posterior of a Dirichlet process. To see why this is so, observe that a distribution P is discrete if and only if $P(x : P\{x\} > 0) = 1$. Now, considering the model $X|P \sim P$ and P given \mathcal{D}_α measure, the property holds if

$$(\mathcal{D}_\alpha \times P)\{(P, x) : P\{x\} > 0\} = 1. \quad (2.2)$$

The assertion is equivalent to

$$(G \times \mathcal{D}_{\alpha+\delta_x})\{(x, P) : P\{x\} > 0\} = 1 \quad (2.3)$$

as G is the marginal of X and the conditional distribution of $P|X$ is $\mathcal{D}_{\alpha+\delta_x}$. The last relation holds, since the presence of the atom at x in the base measure of the posterior Dirichlet process ensures that almost all random P sampled from the posterior process assigns positive mass to the point x . Thus the discreteness property is the consequence of the presence of an atom at the observation in the base measure of the posterior Dirichlet process.

The discreteness property of the Dirichlet process may be disappointing if one's perception of the true distribution is nonatomic, such as when it has a density. However, discreteness itself may not be an obstacle to good convergence properties of estimators, considering the fact that the empirical distribution is also discrete but converges uniformly to any true distribution.

Support

Even though only discrete distributions can actually be sampled from a Dirichlet process, the topological support of the Dirichlet measure \mathcal{D}_α , which is technically the smallest closed set of probability one, could be quite big. The support is actually characterized as all probability measures P^* whose supports are contained in that of G , that is,

$$\text{supp}(\mathcal{D}_\alpha) = \{P^* : \text{supp}(P^*) \subset \text{supp}(G)\}. \quad (2.4)$$

In particular, if G is fully supported, like the normal distribution on the line, then trivially every probability measure is in the support of \mathcal{D}_α . To see why this is true, first observe that any supported P^* must have $P^*(A) = 0$ if A is disjoint from the support of G , which implies that $G(A) = 0$ and so $P(A) = 0$ a.s. $[\mathcal{D}_\alpha]$. For the opposite direction, we use the fact that weak approximation will hold if probabilities of a fine partition are approximated well, and this property can be ensured by the nonsingularity of the Dirichlet distribution with positive parameters.

Self-similarity

Another property of the Dirichlet which distinguishes it from other processes is the self-similarity property described as follows. Let A be any set with $0 < G(A) < 1$, which ensures that $0 < P(A) < 1$ for almost all Dirichlet samples. Let $P|_A$ be the restriction of P to A , that is, the probability distribution defined by $P|_A(B) = P(A \cap B)/P(A)$, and similarly $P|_{A^c}$ is defined. Then the processes $\{P(A), P(A^c)\}$, $P|_A$ and $P|_{A^c}$ are mutually independent, and moreover $P|_A$ follows $\text{DP}(MG(A), G|_A)$. Thus the assertion says that at any given locality A , how mass is distributed within A is independent of how mass is distributed within A^c , and both mass distribution processes are independent of how much total mass is assigned to the locality A . Further, the distribution process within A again follows a Dirichlet process with an appropriate scale. The property has its roots in the connection between independent gamma variables and the Dirichlet distributed variable formed by their ratios: if X_1, \dots, X_k are independent gamma variables, then $X = \sum_{i=1}^k X_i$ and $(X_1/X, \dots, X_k/X)$ are independent. The self-similarity property has many interesting consequences, an important one being that a Dirichlet process may be generated by sequentially distributing mass independently to various subregions following a tree structure. The independence at various levels of allocation, known as the *tail-freeness* property, is instrumental in obtaining large weak support of the prior and weak consistency of posterior. In fact, the Dirichlet process is the only *tail-free process* where the choice of the partition does not play a role.

Limit types

When we consider a sequence of Dirichlet processes such that the center measures converge to a limit G , then there can be three types of limits:

- (i) if the total mass goes to infinity, the sequence converges to the prior degenerate at G ;

- (ii) if the total mass goes to a finite nonzero number M , then the limit is $\text{DP}(M, G)$;
- (iii) if the total mass goes to 0, the limiting process chooses a random point from G and puts the whole mass 1 at that sampled point.

To show the result, one first observes that tightness is automatic here because of the convergence of the center measures, while finite dimensionals are Dirichlet distributions, which converge to the appropriate limit by convergence of all mixed moments. The property has implications in two different scenarios: the Dirichlet posterior converges weakly to the Bayesian bootstrap when the precision parameter goes to zero, and converges to the degenerate measure at P_0 as the sample size n tends to infinity, where P_0 is the true distribution. Thus the entire posterior of P is weakly consistent at P_0 , and the convergence automatically strengthens to convergence in the Kolmogorov–Smirnov distance, much in the tone with the Glivenko–Cantelli theorem for the empirical distribution. The result is extremely intriguing in that no condition on the base measure of the prior is required; consistency holds regardless of the choice of the prior, even when the true distribution is not in the support of the prior. This is very peculiar in the Bayesian context, where having the true distribution in the support of the prior is viewed as the minimum condition required to make the posterior distribution consistent. The rough argument is that when the prior excludes a region, the posterior, obtained by multiplying the prior with the likelihood and normalizing, ought to exclude that region. In the present context, the family is undominated and the posterior is not obtained by applying the Bayes theorem, so the paradox is resolved.

Dirichlet samples and ties

As mentioned earlier, the Dirichlet process samples only discrete distributions. The discreteness property, on the other hand, is able to generate ties in the observations and is extremely useful in clustering applications. More specifically, the marginal joint distribution of n observations (X_1, \dots, X_n) from P which is sampled from $\text{DP}(M, G)$ may be described sequentially as follows. Clearly, $X_1 \sim G$ marginally. Now

$$X_2|P, X_1 \sim P \quad \text{and} \quad P|X_1 \sim \text{DP} \left(M + 1, \frac{M}{M + 1}G + \frac{1}{M + 1}\delta_{X_1} \right), \quad (2.5)$$

which implies, after eliminating P , that $X_2|X_1 \sim \frac{M}{M+1}G + \frac{1}{M+1}\delta_{X_1}$, that is, the distribution of X_2 given X_1 can be described as duplicating X_1 with probability $1/(M + 1)$ and getting a fresh draw from G with probability

$M/(M+1)$. Continuing this argument to X_n given X_1, \dots, X_{n-1} , it is clear that X_n will duplicate any previous X_i with probability $1/(M+n-1)$ and will obtain a fresh draw from G with probability $M/(M+n-1)$. Of course, many of the previous X_i are equal among themselves, so the conditional draw can be characterized as setting to θ_j with probability $n_j/(M+n-1)$, where the θ_j are distinct values of $\{X_1, \dots, X_{n-1}\}$ with frequencies n_j respectively, $j = 1, \dots, k$, and as before, a fresh draw from G with probability $M/(M+n-1)$:

$$X_n | X_1, \dots, X_{n-1} \sim \begin{cases} \delta_{\theta_j} & \text{with probability } \frac{n_j}{M+n-1} \quad j = 1, \dots, k \\ G & \text{with probability } \frac{M}{M+n-1}, \end{cases}$$

where k is the number of distinct observations in X_1, \dots, X_{n-1} and $\theta_1, \dots, \theta_k$ are those distinct values. Also observe that, since (X_1, \dots, X_n) are exchangeable, the same description applies to any X_i given X_j , $j = 1, \dots, i-1, i+1, \dots, n$. This procedure, studied in Blackwell and MacQueen (1973), is known as the generalized Pólya urn scheme. This will turn out to have a key role in the development of Markov chain Monte Carlo (MCMC) procedures for latent variables sampled from a Dirichlet process, as in Dirichlet mixtures discussed shortly.

Because of ties in the above description, the number of distinct observations, the total number of fresh draws from G including the first, is generally much smaller than n . The probabilities of drawing a fresh observation at steps $1, 2, \dots, n$ are $1, M/(M+1), \dots, M/(M+n-1)$ respectively, and so the expected number of distinct values K_n is

$$E(K_n) = \sum_{i=1}^n \frac{M}{M+i-1} \sim M \log \frac{n}{M} \quad \text{as } n \rightarrow \infty. \quad (2.6)$$

Moreover, one can obtain the exact distribution of K_n , and its normal and Poisson approximation, quite easily. The logarithmic growth of K_n induces sparsity that is often used in machine learning applications.

Sethuraman stick-breaking representation

The Dirichlet process $DP(M, G)$ also has a remarkable representation known as the Sethuraman (1994) representation:

$$P = \sum_{i=1}^{\infty} V_i \delta_{\theta_i}, \quad \theta_i \stackrel{\text{i.i.d.}}{\sim} G, \quad V_i = \left[\prod_{j=1}^{i-1} (1 - Y_j) \right] Y_i, \quad Y_i \stackrel{\text{i.i.d.}}{\sim} \text{Beta}(1, M). \quad (2.7)$$

Thus $P = Y_1\delta_{\theta_1} + (1 - Y_1)\sum_{i=2}^{\infty} V'_i\delta_{\theta_{i+1}}$, where $V'_i = [\prod_{j=2}^i(1 - Y_j)]Y_{i+1}$, so that

$$P =_d Y_1\delta_{\theta_1} + (1 - Y)P. \quad (2.8)$$

This distributional equation is equivalent to the representation (2.7), and can be used to derive various properties of the random measure defined by (2.7) and to generate such a process by MCMC sampling. The weights V_i attached to the points $\theta_1, \theta_2, \dots$ respectively may be thought of arising as a result of breaking a stick of unit length randomly in infinite fragments as follows. First break the stick at a location $Y_1 \sim \text{Beta}(1, M)$ and assign the mass Y_1 to a random point $\theta_1 \sim G$. The remaining mass $(1 - Y_1)$ is split in the proportion $Y_2 \sim \text{Beta}(1, M)$ and the net mass $(1 - Y_1)Y_2$ is assigned to a random point $\theta_2 \sim G$. This process continues infinitely many times to complete the assignment of the whole mass to countably many points. What is intriguing is that the resulting process is actually $\text{DP}(M, G)$. To get a rough idea why this is so, recall that for any random distribution P and $\theta \sim P$, the prior for P is equal to the mixture of the posterior distribution $P|\theta$ where θ follows its marginal distribution. In the context of the Dirichlet process, this means that $\mathcal{D}_\alpha = \int \mathcal{D}_{\alpha+\delta_\theta} dG(\theta)$. Now if P is sampled from $\mathcal{D}_{\alpha+\delta_\theta}$, then $P\{\theta\} \sim \text{Beta}(1, M)$ assuming that α is nonatomic. Thus the random P has a point mass at θ of random magnitude distributed as $Y \sim \text{Beta}(1, M)$. With the remaining probability, P is spread over $\{\theta\}^c$, and $P|_{\{\theta\}^c} \sim \text{DP}(M, G)$ independently of $P\{\theta\}$ by the self-similarity property of the Dirichlet process, that is $P|_{\{\theta\}^c} =_d P$. This implies that the $\text{DP}(M, G)$ satisfies the distributional equation (2.8), where $Y \sim \text{Beta}(1, M)$, $\theta \sim G$ and are mutually independent of P . The solution of the equation can be shown to be unique, so the process constructed through the stick-breaking procedure described above must be $\text{DP}(M, G)$.

Sethuraman's representation of the Dirichlet process has far reaching significance. First, along with an appropriate finite stage truncation, it allows us to generate a Dirichlet process approximately. This is indispensable in various complicated applications involving Dirichlet processes, where analytic expressions are not available, so that posterior quantities can be calculated only by simulating them from their posterior distribution. Once a finite stage truncation is imposed, for computational purposes, the problem can be treated essentially as a parametric problem for which general MCMC techniques such as Metropolis–Hastings algorithms and reversible jump MCMC methods can be applied. Another advantage of the sum representation is that new random measures can be constructed by changing the stick-breaking distribution from $\text{Beta}(1, M)$ to other possibilities. One

example is the *two-parameter Poisson–Dirichlet process* where actually the stick-breaking distribution varies with the stage. Even more significantly, for more complicated applications involving covariates, dependence can be introduced among several random measures which are marginally Dirichlet by allowing dependence in their support points θ , or their weights V or both.

Mutual singularity

There are many more interesting properties of the Dirichlet process, for example any two Dirichlet processes are mutually singular unless their base measures share same atoms; see Korwar and Hollander (1973). In particular, the prior and the posterior Dirichlet processes are mutually singular if the prior base measure is nonatomic. This is somewhat peculiar because the Bayes theorem, whenever applicable, implies that the posterior is absolutely continuous with respect to the prior distribution. Of course, the family under consideration is undominated, so the Bayes theorem does not apply in the present context.

Tail of a Dirichlet process

We end this section by mentioning the behavior of the tail of a Dirichlet process. Since $E(P) = G$, one may think that the tails of G and the random P are equal on average. However, this is false as the tails of P are much thinner almost surely. Mathematically, this is quite possible as the thickness of the tail is an asymptotic property. The exact description of the tail involves long expressions, so we do not present it here; see Doss and Sellke (1982). However, it may be mentioned that if G is standard normal, the tail of $P(X > x)$ is thinner than $\exp[-e^{x^2/2}]$ for all sufficiently large x a.s., much thinner than the original Gaussian tail. In a similar manner, if G is standard Cauchy, the corresponding random P has finite moment generating functions, even though the Cauchy distribution does not even have a mean.

2.3 Priors related to the Dirichlet process

Many processes constructed using the Dirichlet process are useful as prior distributions under a variety of situations. Below we discuss some of these processes.

2.3.1 Mixtures of Dirichlet processes

In order to elicit the parameters of a Dirichlet process $DP(M, G)$, as the center measure G is also the prior expectation of P , it is considered as

the prior guess about the parameter P . However, in practice, it is difficult to specify a distribution like $\text{Nor}(0, 1)$ as the prior guess; it is more natural to propose a parametric family with unknown parameters as one's guess about the data generating mechanism. In other words, the center measure contains additional hyperparameters which are then given somewhat flat prior distributions. The resulting process is thus a *mixture of Dirichlet process* (MDP) studied by Antoniak (1974).

Some of the properties of the MDP are quite similar to the Dirichlet. For instance, samples from an MDP are a.s. discrete. Exact expressions for prior mean and variance may be obtained by conditioning on the hyperparameter and finally integrating it out. However, the self-similarity and tail-freeness properties no longer hold for the MDP.

The posterior distribution based on an MDP prior is again MDP. To see this, observe that conditionally on the hyperparameter θ , the structure of a Dirichlet process, and hence its conjugacy property, is preserved. Thus the posterior is a mixture of these Dirichlet processes, although the posterior distribution of θ changes from its prior density $\pi(\theta)$ to the posterior density $\pi(\theta|\text{data})$. In many applications, the precision parameter in the MDP set-up is kept unchanged and the base measure G_θ admits a density g_θ . In this case, the posterior distribution can be found relatively easily and is given by

$$\pi(\theta|\text{data}) \propto \pi(\theta) \prod_{i=1}^n g_\theta(X_i), \quad (2.9)$$

provided that the data are actually sampled from a continuous distribution so that there are no ties among the observations. This is a consequence of the Blackwell–MacQueen urn scheme describing the joint density of (X_1, \dots, X_n) as $\prod_{i=1}^n g_\theta(X_i)$, assuming all the X are distinct, and the Bayes theorem.

2.3.2 Dirichlet process mixtures

While the MDP is a parametric mixture of “nonparametric priors,” a very different scenario occurs when one mixes parametric families nonparametrically. Assume that given a latent variable θ_i , the observations X_i follow a parametric density $\psi(\cdot, \theta_i)$, $i = 1, \dots, n$, respectively. The unknown quantities, unlike in the parametric inference, are not assumed to be equal, but appear, like random effects, from a distribution P . The resulting marginal density for any X_i is thus $f_P(x) = \int \psi(x; \theta) dP(\theta)$ and X_1, \dots, X_n are independent. Since P is not known and is completely unrestricted, a Dirichlet

process prior may be considered as an appropriate prior for P . This induces a prior for the density f_P known as the *Dirichlet process mixture* (DPM), and serves as an extremely useful Bayesian model for density estimation; see Ferguson (1983) and Lo (1984). The model is very rich under a variety of situations, for instance, if the kernel is a normal density with mean θ and scale σ converging to 0, since for any density $\int \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2/\sigma^2} f_0(\theta) d\theta \rightarrow f_0(x)$ in L_1 -distance.

It is possible to write down the expressions for the posterior mean and the posterior variance of the density $f_P(x)$, but the formulae contain an enormously large number of terms prohibiting any real use of them. Fortunately, computable expressions can be obtained by MCMC methods by simulating the latent variables $(\theta_1, \dots, \theta_n)$ from their posterior distribution by a scheme very similar to the Blackwell–MacQueen urn scheme, as studied by Escobar and West (1995) and many others. As before, we can describe the distribution of any θ_i given the other θ_j and all X_i . The scheme is structurally quite similar to the original Blackwell–MacQueen scheme. However, the presence of the extra X_i in the conditioning changes the relative weights and the distribution from where a fresh sample is drawn. More precisely, given θ_j , $j \neq i$, only conditioning by X_i matters, which weighs the selection probability of an old θ_j by $\psi(X_i; \theta_j)$, and the fresh draw by $\int \psi(X_i; \theta) dG(\theta)$, and a fresh draw, whenever obtained, is taken from the “baseline posterior” defined by $dG_b(\theta) \propto \psi(X_i; \theta) dG(\theta)$.

The kernel used in forming the DPM can be chosen in different ways depending on purpose. If density estimation on the line is the objective, one may use a location-scale kernel such as the normal density. On the half-line, gamma, log-normal and Weibull mixtures seem to be more appropriate. On the unit interval, mixtures of beta densities can be considered. Sometimes, special shapes can be produced by special types of mixtures. For instance, a decreasing density on the half-line is a scale mixture of uniform densities, so a prior on a decreasing density can be induced by the mixture model technique. A prior on symmetric strongly unimodal densities can be induced using normal scale mixtures.

2.3.3 Hierarchical Dirichlet processes

A curious process is obtained when one models observations X_{ij} coming from totally unknown distributions F_i , and the distributions F_1, \dots, F_k themselves, treated as unknown parameters, are sampled i.i.d. from a Dirichlet process whose center measure G is itself randomly sampled from a Dirichlet process. Since Dirichlet samples are discrete, the discreteness of G forces

F_1, \dots, F_k to share their atoms, and hence ties will be observed in the values of X even across different groups. This feature is often desirable in some genomic and machine learning applications. Because of the presence of two levels of Dirichlet process, the prior on F_1, \dots, F_k is known as the *hierarchical Dirichlet process*; see Teh, Jordan, Beal and Blei (2006).

2.3.4 Invariant and conditioned Dirichlet processes

In some applications, an unknown distribution needs to be moulded to satisfy certain invariance requirements, such as symmetry. Since the Dirichlet process supports all types of distributions, one needs to symmetrize a random distribution obtained from the Dirichlet process prior as in Dalal (1979). This technique can be used, for instance, in proposing a prior for the distribution of error, so that a location or regression parameter can be made identifiable. Another alternative is to constrain the distribution to have median zero. A prior for this was obtained in Doss (1985a) by conditioning the Dirichlet process to assign probability $\frac{1}{2}$ to $[0, \infty)$.

2.4 Posterior consistency

2.4.1 Motivation and implications

Now we turn our attention to asymptotic properties of the posterior distributions, and begin with a discussion on posterior consistency. Consider a sequence of statistical experiments parameterized by a possibly infinite-dimensional parameter θ taking values in a separable metric space. Let Π stand for the prior distribution on θ and $\Pi(\cdot|\text{data})$ stand for (a version of) the posterior distribution.

Definition 2.2 The posterior distribution is said to be *consistent* at a given θ_0 , or (θ_0, Π) is a consistent pair, if for any neighborhood V of θ_0 , $\Pi(\theta \notin V|\text{data}) \rightarrow 0$ (in probability or a.s.) as the size of the data tends to infinity when θ_0 is the true value of the parameter.

It may be noted that consistency depends on the choice of a version, but in dominated cases, there is essentially only one version that matters.

The importance of consistency stems from the desire to be able to identify correctly the data-generating mechanism when an unlimited supply of data is available. Even though this is purely a large sample property, an inconsistent posterior is often an indication of seriously incorrect inference, even for moderate sample sizes. Moreover, consistency can be shown to be

equivalent with agreement among Bayesians with different sets of priors; see Diaconis and Freedman (1986).

Consistency has several immediate connections with other properties. First, it may be observed that it is not necessary to check convergence for every possible neighborhood; it is enough to consider a class which forms a local sub-base for the topology, that is, a class whose finite intersections form a base for the topology at the true value. Consistency implies existence of a rate of convergence also, that is, a shrinking ball around the true value whose posterior probability tends to one. Consistency also implies existence of an estimator which is consistent in a frequentist sense. To construct such an estimator, one may look at the point such that a small ball around it has maximum posterior probability among all balls of equal radius. Then, since balls around the true parameter have posterior probability tending to one, and the chosen point, by definition, cannot be beaten in this game, we obtain two small balls both of which have posterior probability close to one. Such balls must intersect, since otherwise the total posterior probability would exceed one. Therefore the two center points, the true parameter value and the estimator, must be infinitesimally close. In other words, the estimator constructed by maximizing the posterior probability of a ball of small radius around it is consistent. If posterior consistency holds, then for convex parameter spaces such as the space of densities with the L_1 , Hellinger or some bounded metric on it which induces convex neighborhoods, the posterior mean gives another consistent estimator.

2.4.2 Doob's theorem

There is an extremely general theorem by Doob (1948), which essentially ensures consistency under any model where consistent estimators exist (such as when i.i.d. observations are available and the model is identifiable), provided we are happy to live with possible inconsistency on a null set with respect to the prior. While, at first glance, this may be a cheering general fact, the interpretation of nullity in view of the chosen prior is not satisfactory. It is easy for a null set to be topologically huge. An extreme possibility is exhibited by the prior degenerate at a point θ^* . In this case, consistency fails everywhere except for $\theta_0 = \theta^*$, yet this huge exceptional set is a null set in view of the prior considered. Thus, to study whether consistency holds in a given situation, it is important to give sufficient conditions on the true value of the parameter and the prior which ensure consistency. It may be noted that if the parameter space is countable, Doob's result

actually implies consistency everywhere provided all points receive positive prior mass.

2.4.3 Instances of inconsistency

For finite-dimensional parameter spaces, consistency is almost guaranteed, at least for well-behaved parametric families, if the prior density is positive in the neighborhoods of the true parameter. Surprisingly, consistency can fail in infinite-dimensional spaces for quite well-behaved models even for seemingly natural priors. In particular, the condition of assigning positive prior probabilities in usual neighborhoods of the true parameter is not at all sufficient to ensure consistency.

An interesting counterexample was constructed by Freedman (1963) in the context of estimating a discrete distribution on natural numbers, which is the simplest nonparametric estimation problem. Let the true distribution of observations be geometric with parameter $\frac{1}{4}$. Freedman constructed a prior which assigns positive probability to every weak neighborhood of the true distribution, but the posterior concentrates near a wrong value, the geometric distribution with parameter $\frac{3}{4}$. A more striking and counter-intuitive example was constructed more recently in Kim and Lee (2001) in the context of Bayesian survival analysis, where it was shown that among two priors for cumulative hazard function, both with mean equal to the true cumulative hazard, the one with larger prior spread achieves posterior consistency but the one which is more tightly spread leads to inconsistency.

Freedman's example is actually generic in the topological sense. If we look at all possible pairs of true parameter values and priors which lead to consistency, then the collection is extremely narrow when the size is measured topologically. A set F is called *meager* and considered to be topologically small if F can be expressed as a countable union of sets C_i , $i \geq 1$, whose closures \bar{C}_i have empty interior. Freedman (1963) showed that the collection of "good pairs" is meager in the product space.

Should this result scare a Bayesian into abandoning his approach? No. The reason is that we are only concerned about a relatively small collection of priors. Therefore, what happens to most priors does not bother us, as long as we can find a prior incorporating any available subjective features and the corresponding posterior distribution good frequentist properties of the posterior. However, the counterexample and result above warn against careless use of a prior and emphasize the need to prove theorems assuring posterior consistency under mild conditions on the true parameter and the prior.

2.4.4 Approaches to consistency

If the posterior has an explicit expression, it may be possible to prove consistency or rate of convergence by simple Chebyshev-type inequalities. This is often the case in Bayesian survival analysis, where posterior conjugacy holds, for instance, for priors described by a Lévy process. We have also seen that convergence properties of the Dirichlet process give rise to posterior consistency. However, these situations are very special and are not to be expected in all applications.

In the context of estimating a c.d.f., a reasonably general class of priors for which posterior consistency holds is given by the class of tail-free priors considered in Freedman (1963). For the weak topology, convergence can be assessed through convergence of probabilities of the sets in a sufficiently fine partition. Thus one can restrict attention to the finite-dimensional object given by the probability vector corresponding to this fine partition. Interestingly, the posterior distribution of this vector depends only on the counts of the corresponding cells by the tail-freeness property, so the problem reduces to that of estimating parameters in a multinomial distribution, for which consistency holds under the general conditions that the weak support of the tail-free process contains the true distribution. A particularly important tail-free class is given by the Pólya tree process; see Lavine (1992). In this case, the space is split binarily and each time mass is distributed to the left and the right parts according to an independent random variable following a beta distribution, whose parameters can vary freely while the remaining mass is assigned to the corresponding right portion. Such priors have generally large weak support, ensuring consistency.

Although the above result is very interesting, it is also somewhat restrictive in that it is applicable only to the problem of estimating a c.d.f., and only if a tail-free prior is used. The tail-freeness property is very delicate and can be easily destroyed by common operations like symmetrization or mixing. Indeed, inconsistency may occur in this way as Diaconis and Freedman (1986) showed.

2.4.5 Schwartz's theory

A more useful approach, due to Schwartz (1965), is obtained by putting appropriate size restrictions on the model and conditions on the support of the prior in the sense of Kullback–Leibler divergence. Below, we describe Schwartz's theory and its extensions along with some applications, especially to the density estimation problem.

For the general theory, we assume that the family is dominated. Let $p_{\theta,n}(X_1, \dots, X_n)$ stand for the joint density of observations and Π for the prior distribution. It is possible to let Π depend on n , but to keep ideas simple, we assume that the prior is fixed. Let θ_0 stand for the true value of the parameter. Then the posterior probability of any set B can be written as

$$\Pi(\theta \in B | X_1, \dots, X_n) = \frac{\int_B \frac{p_{\theta,n}(X_1, \dots, X_n)}{p_{\theta_0,n}(X_1, \dots, X_n)} d\Pi(\theta)}{\int \frac{p_{\theta,n}(X_1, \dots, X_n)}{p_{\theta_0,n}(X_1, \dots, X_n)} d\Pi(\theta)}. \quad (2.10)$$

To establish consistency, we let B be the complement of a neighborhood U of θ_0 and show that the above expression with $B = U^c$ goes to 0 as $n \rightarrow \infty$, either in P_{θ_0} -probability or P_{θ_0} a.s. A strategy that often works, especially when the observations are i.i.d., is to show that the numerator in (2.10) converges to zero exponentially fast like $e^{-\beta n}$ for some $\beta > 0$, while the denominator multiplied by $e^{\beta n}$ converges to infinity for all $\beta > 0$. Thus we need to give sufficient conditions to ensure these two separate assertions.

Below we assume that the observations are i.i.d. following a density p_θ with respect to a σ -finite measure ν ; we shall indicate later how to extend the result to independent non-identically distributed or even dependent observations.

Kullback–Leibler property

Note that the integrand in the denominator of (2.10) can be written as $e^{-n\Lambda_n(\theta, \theta_0)}$, and for large n , $\Lambda_n(\theta, \theta_0) := n^{-1} \sum_{i=1}^n \log(p_{\theta_0}(X_i)/p_\theta(X_i))$ behaves like the *Kullback–Leibler divergence* number given by $K(p_{\theta_0}; p_\theta) = \int p_{\theta_0} \log(p_{\theta_0}/p_\theta) d\nu$. Thus the integrand is at least as big as $e^{-2n\epsilon}$ for all sufficiently large n if $K(p_{\theta_0}; p_\theta) < \epsilon$, so the contribution of the part $A := \{\theta : K(p_{\theta_0}; p_\theta) < \epsilon\}$ alone to the integral in the denominator of (2.10) is at least $e^{-2n\epsilon} \Pi(\theta : K(p_{\theta_0}; p_\theta) < \epsilon)$. Since $\epsilon > 0$ can be chosen arbitrarily small, it is clear that the term in the denominator multiplied by $e^{n\beta}$ is exponentially big, provided that $\Pi(\theta : K(p_{\theta_0}; p_\theta) < \epsilon) > 0$ for all $\epsilon > 0$. The whole argument can easily be made rigorous by an application of Fatou’s lemma. Thus the last condition emerges as a key condition in the study of consistency, and will be referred to as Schwartz’s prior positivity condition or the *Kullback–Leibler property* of the prior, or the true parameter is said to be in the *Kullback–Leibler support* of the prior. Note that the condition essentially means that the prior should assign positive probability to any neighborhood of the true parameter, much in the spirit of the “obvious requirement” for consistency. However, the important matter here is that

the neighborhood is defined by nearness in terms of the Kullback–Leibler divergence, not in terms of the topology of original interest. In many parametric cases, regularity conditions on the family ensure that a Euclidean neighborhood is contained inside such a Kullback–Leibler neighborhood, so the usual support condition suffices. For infinite-dimensional families, the Kullback–Leibler divergence is usually stronger than the metric locally around θ_0 .

Bounding the numerator

To show that the numerator in (2.10) is exponentially small, a naive but straightforward approach would be to bound $\Lambda_n(\theta, \theta_0)$ uniformly over θ lying outside the given neighborhood U . For individual θ , the above quantity stays below a negative number by the law of large numbers and the fact that the Kullback–Leibler divergence is strictly positive. However, to control the integral, one needs to control $\Lambda_n(\theta, \theta_0)$ uniformly over U^c , which poses a tough challenge. If the parameter space is compact, a classical approach used by Wald, which bounds the log-likelihood ratio by a maximum of finitely many terms each of which is a sum of integrable i.i.d. variables with negative expectation, is useful. More modern approaches to bounding the log-likelihood ratio outside a neighborhood involve bounding bracketing entropy integrals, which is also a strong condition.

Uniformly consistent tests

Clearly, bounding an average by the maximum is not the best strategy. Schwartz’s ingenious idea is to link the numerator in (2.10) with the power of uniformly exponentially consistent tests for the hypothesis $\theta = \theta_0$ against $\theta \in U^c$. Under the existence of such a test, Schwartz (1965) showed that the ratio of the marginal density of the observation with θ conditioned to lie outside U to the true joint density is exponentially small except on a set with exponentially small sampling probability. This is enough to control the numerator in (2.10) as required.

Uniformly exponentially consistent tests, that is, tests for which both the type I and type II error probabilities go to zero exponentially fast, have been well studied in the literature. If two convex sets of densities C_1 and C_2 are separated by a positive distance at least ϵ in terms of the Hellinger distance, then one can construct a test for the pair of hypotheses $p_\theta \in C_1$ against $p_\theta \in C_2$ whose error probabilities decay like $e^{-cn\epsilon^2}$ for some universal constant $c > 0$. In this result, the sizes of C_1 and C_2 are immaterial; only convexity and their distance matter. Generally, U^c is not convex, so the result does not directly give an exponentially consistent test for testing

$\theta = \theta_0$ against the alternative U^c . However, we observe that if U^c can be covered by finitely many convex bodies C_1, \dots, C_k , each of which maintains a positive distance from θ_0 , then a uniformly exponentially consistent test ϕ_j is obtained for testing $\theta = \theta_0$ against C_j with both error probabilities bounded by $e^{-c'n}$. Then define a test $\phi = \max_j \phi_j$. Clearly, the power of ϕ at any point is better than the power of any of the ϕ_j . In particular, for any $j = 1, \dots, k$, if $p_\theta \in C_j$, then $E_\theta(1 - \phi) \leq E_\theta(1 - \phi_j)$. By the given construction, the latter term is already exponentially small uniformly over C_j . Thus the type II error probability is easily bounded. For the type I error probability, we can bound $E_{\theta_0} \phi \leq \sum_{j=1}^k E_{\theta_0} \phi_j \leq ke^{-c'n}$, establishing the required exponential bound.

The strategy works nicely in many parametric models where a uniformly exponentially consistent test for the complement of a very large ball can often be obtained directly, so that the remaining compact portion may be covered with a finite number of balls. In infinite-dimensional spaces, this is much harder. When the topology under consideration is the weak topology, U^c can be covered by finitely many convex sets. To see this, observe that a basic open neighborhood U of a true density p_0 is described by conditions on finitely many integrals $\{p : |\int \psi_j p - \int \psi_j p_0| < \epsilon_j, j = 1, \dots, k\}$, which can be written as $\cap_{j=1}^k \{p : \int \psi_j p < \int \psi_j p_0 + \epsilon_j\} \cap \cap_{j=1}^k \{p : \int \psi_j p_0 < \int \psi_j p + \epsilon_j\}$. Thus U^c is a finite union of sets of the form $\{p : \int \psi p \geq \int \psi p_0 + \epsilon\}$ or $\{p : \int \psi p_0 \geq \int \psi p + \epsilon\}$, both of which are convex and separated from p_0 .

Entropy and sieves

Unfortunately, the procedure runs into difficulty in infinite-dimensional spaces with stronger topologies, such as for density estimation with the Hellinger or L_1 -distance, unless the space of densities under consideration is assumed to be compact. For the space of density functions, the complement of a neighborhood cannot be covered by finitely many balls or convex sets, each maintaining a positive distance from the true one. However, not everything is lost, and much of the idea can be carried out with the help of a technique of truncating the parameter space, depending on the sample size. Observe that in the argument, the type II error probability will not be problematic whenever the final test is greater than individual tests, so it is only the type I error probability which needs to be properly taken care of. From the bound $ke^{-c'n}$, it is clear that one may allow k to depend on n , provided that its growth is slower than $e^{c'n}$, to spare an exponentially small factor.

To formalize the idea, let p denote the density function which itself is treated as the parameter. Let \mathcal{P} be a class of density functions where the

possible values of p lie and $p_0 \in \mathcal{P}$ stands for the true density function. For definiteness, we work with the Hellinger distance on \mathcal{P} , although the L_1 -distance may also be used. In fact, the two metrics define the same notion of convergence. Let $U = \{p : d(p, p_0) < \epsilon\}$ for some given $\epsilon > 0$. Let \mathcal{P}_n be a sequence of subsets of \mathcal{P} , also called a *sieve* (possibly depending on ϵ), gradually increasing to \mathcal{P} . Let N_n be the number of balls of size $\epsilon/2$ with center in \mathcal{P}_n , needed to cover \mathcal{P}_n . Any such ball which intersects U^c clearly maintains a distance at least $\epsilon/2$ from p_0 . Thus the type I and type II error probability for testing $p = p_0$ against any ball is bounded by $e^{-n\delta}$, where $\delta > 0$ depends on ϵ , and can be made explicit if desired. Then if $\log N_n < n\delta'$ for all n and $\delta' < \delta$, then by the discussion in the preceding paragraph, it follows that the numerator in (2.10) is exponentially small, and hence $\Pi(p \in U^c \cap \mathcal{P}_n | X_1, \dots, X_n) \rightarrow 0$. The number N_n is called the $\epsilon/2$ -covering number of \mathcal{P}_n with respect to the metric d , which is denoted by $N(\epsilon/2, \mathcal{P}_n, d)$, and its logarithm is known as the *metric entropy*. Thus a bound for the metric entropy of the sieve limited by a suitably small multiple of n implies that the posterior probability of U^c fades out unless it goes to the complement of the sieve \mathcal{P}_n . In order to complete the proof of posterior consistency, one must show that $\Pi(p \in \mathcal{P}_n^c | X_1, \dots, X_n) \rightarrow 0$ by other methods. This is sometimes possible by direct calculations. Note that \mathcal{P}_n^c has small prior probability, so we should expect it to have small posterior probability in some sense if the likelihood is bounded appropriately. Unfortunately, small prior probability of \mathcal{P}_n^c does not imply small posterior probability, since the likelihood may increase exponentially, enhancing the posterior probability. However, if the prior probability is exponentially small, then so is the posterior probability, under the Kullback–Leibler positivity condition. This follows quite easily by a simple application of Fubini’s theorem applied to the numerator of (2.10) with $B = \mathcal{P}_n^c$. Thus, to establish consistency, one just needs to construct a sieve \mathcal{P}_n such that $\log N(\epsilon/2, \mathcal{P}_n, d) < n\delta' < n\delta$, where δ is defined before, and $\Pi(\mathcal{P}_n^c)$ is exponentially small. In particular, the entropy condition will hold for a sieve \mathcal{P}_n with $\log N(\epsilon/2, \mathcal{P}_n, d) = o(n)$. The main ideas behind the result were developed through the works Schwartz (1965), Barron, Schervish and Wasserman (1999) and Ghosal, Ghosh and Ramamoorthi (1999a). It is also interesting to note that the approach through testing is “optimal.” This is because a result of Barron (see Theorem 4.4.3 of Ghosh and Ramamoorthi (2003)) shows that consistency with exponential speed holds if and only if one can find a sieve whose complement has exponentially small prior probability and a test which has exponentially small error probabilities on the sieve.

2.4.6 Density estimation

The above consistency theorem can be applied to derive posterior consistency in Bayesian density estimation using the commonly used priors such as the Dirichlet mixture or Gaussian processes.

Dirichlet mixtures

To establish the Kullback–Leibler property of a Dirichlet mixture of normal prior at a true density p_0 , one approximates p_0 by p_m defined as the convolution of p_0 truncated to $[-m, m]$ for some large m and the normal kernel with a small bandwidth. This convolution, which is itself a normal mixture, approximates p_0 pointwise as well as in the Kullback–Leibler sense under mild conditions on p_0 . Now a Kullback–Leibler neighborhood around p_m includes a set which can be described in terms of a weak neighborhood around p_0 truncated to $[-m, m]$. Since the Dirichlet process has large weak support, the resulting weak neighborhood will have positive probability, proving the Kullback–Leibler property. Indeed, the argument applies to many other kernels. To construct appropriate sieves, consider all mixture densities arising from all bandwidths $h > h_n$ and mixing distribution F with $F[-a_n, a_n] > 1 - \delta$, where $a_n/h_n < cn$ for some suitably small $c > 0$. Then the condition for exponentially small prior probability of the complement of the sieve holds if the prior on the bandwidth is such that $\Pi(h < h_n)$ is exponentially small, and the base measure α of the Dirichlet process assigns exponentially small mass to $[-a_n, a_n]^c$. These conditions can be met, for instance, if the prior for h^2 is inverse gamma and the base measure of the Dirichlet process of the mixing distribution is normal. Results of this kind were obtained in Ghosal, Ghosh and Ramamoorthi (1999a), Lijoi, Prünster and Walker (2005), Tokdar (2006) and Wu and Ghosal (2008).

Gaussian processes

A prior for density estimation on a compact interval I can also be constructed from a Gaussian process $\xi(t)$ by exponentiating and normalizing to $e^{\xi(t)} / \int_I e^{\xi(u)} du$. Assuming that the true density p_0 is positive throughout, it is easy to see that a Kullback–Leibler neighborhood of p_0 contains a set which is described by the uniform neighborhood in terms of $\xi(t)$ about a function $\xi_0(t)$ satisfying $p_0(t) = e^{\xi_0(t)} / \int_I e^{\xi_0(u)} du$. Now, for a Gaussian process, a continuous function $\xi_0(t)$ is in the support if and only if it belongs to the closure of the *reproducing kernel Hilbert space* (RKHS) of the Gaussian process. The Brownian motion and its integrals are Gaussian processes with large RKHS. Another possibility is to consider a Gaussian process with

kernel containing a scale which is allowed to assume arbitrarily large positive values (so that the prior is actually a mixture of Gaussian processes). Then under extremely mild conditions on the kernel, the overall support of the process includes all continuous functions. Further, sieves can easily be constructed using Borell's inequality or smoothness properties of Gaussian paths and maximal inequalities for Gaussian processes; see Ghosal and Roy (2006), Tokdar and Ghosh (2007), and van der Vaart and van Zanten (2007, 2008).

Pólya tree processes

To estimate the density of the observations using a Pólya tree prior, consider for simplicity, binary partitions used in the mass distribution in the tree structure obtained sequentially by the median, quartiles, octiles etc. of a density λ . Further assume that the parameters of the beta distributions used to split mass randomly are all equal within the same level of the tree, that is, say the parameters are all equal to a_m at level m . Then by a theorem of Kraft (1964), it follows that the random distributions generated by a Pólya tree admit densities a.s. if $\sum_{m=1}^{\infty} a_m^{-1} < \infty$. To establish the Kullback–Leibler property, one needs to strengthen the condition to $\sum_{m=1}^{\infty} a_m^{-1/2} < \infty$ and assume that the density λ has finite entropy in the sense $\int p_0(x) \log \lambda(x) dx < \infty$; see Ghosal, Ghosh and Ramamoorthi (1999b) for a proof. The Kullback–Leibler property implies posterior consistency with respect to the weak topology. However, since the sample paths of a Pólya tree lack appropriate smoothness, it is difficult to control the size of the space where the prior is essentially supported. Under quite strong growth conditions $a_m \sim 8^m$ on the parameters, appropriate sieves were constructed in Barron, Schervish and Wasserman (1999), giving consistency with respect to the Hellinger distance.

2.4.7 Semiparametric applications

Schwartz's consistency theory and its extensions lead to very useful consistency results in Bayesian semiparametric inference. Diaconis and Freedman (1986), (respectively, Doss (1985b)) give examples showing that inconsistency may occur when one estimates the location parameter θ in the location model $X = \theta + e$, $e \sim F$, using a symmetrized Dirichlet (respectively, Dirichlet conditioned to have mean zero) prior for F . To understand why this is happening, ignore the issue of symmetrization and represent the prior as a mixture of Dirichlet process with θ acting as a location parameter for the base measure G with p.d.f. g . Then it follows from (2.9) that the likelihood

for θ given X_1, \dots, X_n is $\prod_{i=1}^n g(X_i - \theta)$, so the Bayes estimator is similar to the maximum likelihood estimator (MLE) based on the above incorrect “parametric” likelihood, which may or may not give the correct result. Also observe that discreteness of Dirichlet samples prohibits the prior putting any mass in the Kullback–Leibler neighborhoods, so Schwartz’s theory does not apply there. However, positive results were obtained in Ghosal, Ghosh and Ramamoorthi (1999b) for a prior which leads to densities. Indeed, if the true error density f_0 is in the Kullback–Leibler support of the prior Π for the density f of F , then the true density of observations $f_0(\cdot - \theta_0)$ is in the support of the prior for the density of observations. Thus the Kullback–Leibler property is not destroyed by location shifts, unlike the fragile tail-freeness property of the Dirichlet process which is lost by symmetrization and location change. For instance, using an appropriate Pólya tree, Dirichlet mixture or Gaussian process prior for f , we can ensure that the distribution of X is consistently estimated in the weak topology. Now within a class of densities with fixed median, the map $(\theta, f) \mapsto f(\cdot - \theta)$ is both-way continuous with respect to the weak topology. Thus consistency for θ follows from the weak consistency for the density of the observations, which is obtained without directly constructing any test or sieves. This clearly shows the power of Schwartz’s theory, especially for semiparametric problems, where the infinite-dimensional part is usually not of direct interest.

2.4.8 Non-i.i.d. observations

The theory of posterior consistency can be extended to independent, non-identically distributed variables as well as to some dependent situations. First observe that the denominator in (2.10) can be tackled essentially in the same way when a law of large numbers is applicable to the summands appearing in the log-likelihood ratio. For independent, non-identically distributed random variables, this is possible by Kolmogorov’s strong law when variances of the log-likelihood ratio based on each observation can be controlled appropriately. For ergodic Markov processes, a law of large numbers is available too.

To control the numerator, one needs to construct appropriate tests against complements of neighborhoods for a given topology. Such tests have been constructed in the literature for applications such as linear regression with nonparametric error (Amewou-Atisso, Ghosal, Ghosh and Ramamoorthi 2003), binary regression with Gaussian process prior (Ghosal and Roy, 2006), normal regression with Gaussian process prior (Choi and Schervish, 2007),

estimation of the spectral density of a time series using Whittle likelihood (Choudhuri, Ghosal and Roy, 2004) and estimating the transition density of a Markov process using Dirichlet mixtures (Tang and Ghosal, 2007). Other important work on consistency includes Diaconis and Freedman (1993) and Coram and Lalley (2006) showing fine balance between consistency and inconsistency in binary regression with a prior supporting only piecewise constant functions. The last two works use direct computation, rather than Schwartz's theory, to prove consistency.

2.4.9 Sieve-free approaches

We end this section by mentioning alternative approaches to consistency which do not require the construction of sieves and uniformly exponentially consistent tests on them.

Martingale method

Consider the Hellinger distance on the space of densities of i.i.d. observations and assume that the Kullback–Leibler property holds. Then, by utilizing a martingale property of marginal densities, Walker (2004) showed that the posterior probability of a set A goes to 0 if the posterior mean of p , when the prior is restricted to A , is asymptotically a positive distance away from the true density p_0 . Now, when $A = U^c$, the result is not directly applicable as the posterior mean for the prior restricted to U^c may come close to p_0 . To obtain the required result, Walker (2004) covered U^c with countably many balls, and controlled both the size and prior probability of each ball. More precisely, Walker (2004) showed that if for any given $\epsilon > 0$, there is $0 < \delta < \epsilon$ such that each ball has diameter at most δ and the sum of the square root of the prior probabilities of these balls is finite, then the posterior is consistent. The argument also extends to Markov processes as shown by Ghosal and Tang (2006). A lucid discussion on the basis of consistency or inconsistency without referring to sieves is given by Walker, Lijoi and Prünster (2005).

Although the proof of consistency based on the martingale property is very interesting and one does not need to construct sieves, the approach does not lead to any new consistency theorem. This is because the condition on the diameter of each ball and the summability of square roots of prior probabilities imply existence of a sieve whose complement has exponentially small prior probability and which has $\epsilon/2$ -Hellinger metric entropy bounded by a small multiple of n , that is, the conditions of the consistency theorem obtained from the extension of Schwartz's theory discussed before, hold.

Power-posterior distribution

A remarkable finding of Walker and Hjort (2001) is that the posterior distribution is consistent only under the Kullback–Leibler property if the likelihood function is raised to a power $\alpha < 1$ before computing the “posterior distribution” using the Bayes formula. The resulting random measure can be called the *power-posterior distribution*. The above assertion follows quite easily by bounding the numerator in the Bayes theorem by Markov’s inequality and the denominator by Schwartz’s method using the Kullback–Leibler property as described before. The greatest advantage with this approach is that there is no need for any additional size constraint in the form of tests or entropy bounds, and hence there is no need to construct any sieves, provided that one is willing to alter the posterior distribution to make inference. However, usual MCMC methods may be difficult to adapt, especially for density estimation using Dirichlet mixture prior. The Kullback–Leibler property alone can lead to other desirable conclusions such as convergence of sequential predictive densities in relative entropy risk as shown by Barron (1999).

2.5 Convergence rates of posterior distributions*2.5.1 Motivation, description and consequences*

As mentioned in the introduction, in naive terms, the convergence rate is the size ϵ_n of the smallest ball centered about the true parameter θ_0 such that the posterior probability converges to one. In practice, we often just find one sequence ϵ_n such that the posterior probability of the ball of radius ϵ_n around θ_0 converges to one, so it may be more appropriate to term this “a rate of convergence.”

Definition 2.3 Let $X^{(n)}$ be data generated by a model $P_\theta^{(n)}$. We say that a sequence $\epsilon_n \rightarrow 0$ is the *convergence rate of the posterior distribution* $\Pi_n(\cdot|X^{(n)})$ at the true parameter θ_0 with respect to a pseudo-metric d if for any $M_n \rightarrow \infty$, we have that $\Pi_n(\theta : d(\theta, \theta_0) \geq M_n \epsilon_n) \rightarrow 0$ in $P_{\theta_0}^{(n)}$ probability.

Thus, by rate of convergence, we mean only up to a multiplicative constant, thus disregarding the constants appearing in the bound. At present, the available techniques do not guide us to the best possible constants. It is well known that the convergence rate in regular parametric families is $n^{-1/2}$, agreeing with the *minimax rate*, that is the best convergence rate for estimators. For infinite-dimensional models, the rate of convergence may

be $n^{-1/2}$ or slower. In many cases, the posterior convergence rate corresponding to well-known priors agrees with the minimax rate, possibly up to a logarithmic factor.

There are some immediate consequences of the posterior convergence rate. If the posterior converges at the rate ϵ_n , then as in the previous section, the estimator defined as the center of the ball of radius maximizing the posterior probability converges to the true parameter at rate ϵ_n in the frequentist sense. Since the convergence rate of an estimator cannot be faster than the minimax rate for the problem, it also follows that the posterior convergence rate cannot be better than the minimax rate. Thus achieving the minimax rate can be regarded as the ideal goal. For the special case of density estimation with the L_1 or the Hellinger metric defining the convergence rate, the posterior mean also converges at the same rate at which the posterior converges.

When the parameter space is equipped with the L_2 -norm and expressions for the posterior mean $\hat{\theta}$ and variance are explicitly available, it may be possible to derive the convergence rate from Chebyshev's inequality. When θ stands for the mean of an infinite-dimensional normal distribution and an appropriate conjugate normal prior is used, the convergence rate can be calculated easily by explicit calculations. In general, this seems to be difficult, so we need to develop a general theory along the lines of Schwartz's theory for posterior consistency for dominated families.

2.5.2 General theory

Let us first consider the i.i.d. case where observations $X_1, X_2, \dots \sim p$, and p is given possibly a sequence of prior Π . Let $\epsilon_n \rightarrow 0$ be the targeted rate. In density estimation problems, this is slower than $n^{-1/2}$, so we assume that $n\epsilon_n^2 \rightarrow \infty$. The basic ideas we use here were developed by Ghosal, Ghosh and van der Vaart (2000), and are similar to those used to study consistency. See also Shen and Wasserman (2001) and Walker, Lijoi and Prünster (2007) for alternative approaches involving somewhat stronger conditions. As in (2.10), we express the posterior probability of $B = \{p : d(p, p_0) \geq \epsilon_n\}$ as a ratio, and show that, under appropriate conditions, (i) the numerator is bounded by $e^{-c n \epsilon_n^2}$, where $c > 0$ can be chosen sufficiently large, while (ii) the denominator is at least $e^{-b n \epsilon_n^2}$.

The above two assertions hold under conditions which can be thought of as the quantitative analog of the conditions ensuring consistency. This is quite expected as a rate statement is a quantitative refinement of consistency.

Prior concentration rate

To take care of the second assertion, we replace the Kullback–Leibler positivity condition by

$$\Pi(B(p_0, \epsilon_n)) := \Pi\{p : K(p_0; p) \leq \epsilon_n^2, V(p_0; p) \leq \epsilon_n^2\} \geq e^{-b_1 n \epsilon_n^2}, \quad (2.11)$$

where $V(p_0; p) = \int p_0(\log(p_0/p))^2$. Thus apart from the fact that the description of the neighborhood also involves the second moment of the log-likelihood ratio, the condition differs from Schwartz's condition in requiring a minimum level of concentration of prior probability around the true density p_0 . Intuitively, the likelihood function at p with $K(p_0; p) \leq \epsilon_n^2$, apart from random fluctuations, is at least $e^{-n\epsilon_n^2}$. When the variance $V(p_0; p)$ is also smaller than ϵ_n^2 , it can be seen that the random fluctuations can change the lower bound only slightly, to $e^{-b_2 n \epsilon_n^2}$, except on a set of small probability. The prior probability of the set of such p is at least $e^{-b_1 n \epsilon_n^2}$ by (2.11), leading to assertion (ii).

Entropy and tests

To take care of assertion (i), we follow the testing approach of Schwartz. Note that, as the alternative $\{p : d(p, p_0) \geq \epsilon_n\}$ is getting close to the null $p = p_0$, it is not possible to test the pair with exponentially small type I and type II error probabilities uniformly. However, since we now have a better lower bound for the denominator, our purpose will be served if we can test with both type I and type II error probabilities bounded by $e^{-c n \epsilon_n^2}$, where c is larger than b appearing in assertion (ii). By the discussion in the previous section, such error probabilities are possible for convex alternatives which are separated from p_0 by at least ϵ_n in terms of the Hellinger distance.† To construct the final test, one needs to cover the alternative with balls of size $\epsilon_n/2$ and control their number to no more than $e^{c_1 n \epsilon_n^2}$, that is, satisfy the metric entropy condition $\log N(\epsilon_n/2, \mathcal{P}, d) \leq c_1 n \epsilon_n^2$. The smallest ϵ satisfying the inequality $\log N(\epsilon/2, \mathcal{P}, d) \leq n \epsilon^2$ appears in the classical theory of minimax rates in that the resulting rate ϵ_n determines the best possible rate achievable by an estimator. Therefore, if the prior concentration rate can be matched with this ϵ_n , the posterior will converge at the rate at par with the minimax rate.

Sieves

Of course, satisfying the entropy condition is not generally possible unless \mathcal{P} is compact, so we need to resort to the technique of sieves. As before, if

† Alternatively, the L_1 -distance can be used, and also the L_2 -distance if densities are uniformly bounded.

there exists a sieve \mathcal{P}_n such that

$$\log N(\epsilon_n/2, \mathcal{P}_n, d) \leq c_1 n \epsilon_n^2, \quad (2.12)$$

then, by replacing ϵ_n by a sufficiently large multiple $M\epsilon_n$, it follows that $\Pi(p \in \mathcal{P}_n : d(p, p_0) \geq \epsilon_n)$ converges to zero. To take care of the remaining part \mathcal{P}_n^c , the condition $\Pi(\mathcal{P}_n^c) \leq e^{-c_2 n \epsilon_n^2}$ suffices, completing the proof that the rate of convergence is ϵ_n .

2.5.3 Applications

The rate theorem obtained above can be applied to various combinations of model and prior.

Optimal rates using brackets

First we observe that optimal rates can often be obtained by the following technique of bracketing applicable for compact families. By an ϵ -bracketing of \mathcal{P} , we mean finitely many pairs of functions (l_j, u_j) , $l_j(\cdot) \leq u_j(\cdot)$, $d(u_j, l_j) < \epsilon$, known as ϵ -brackets, such that any $p \in \mathcal{P}$ is contained in one of these brackets. The smallest number of ϵ -brackets covering \mathcal{P} is called the ϵ -bracketing number of \mathcal{P} , denoted by $N_{[\cdot]}(\epsilon, \mathcal{P}, d)$. Let ϵ_n be the smallest number satisfying $\log N_{[\cdot]}(\epsilon, \mathcal{P}, d) \leq n\epsilon^2$. For all j , find an ϵ_j -bracketing and normalize its upper brackets to p.d.f.s. Now put the discrete uniform distribution on these p.d.f.s and mix these discrete uniform distributions according to a thick-tailed distribution λ_j on the natural numbers. Then the resulting prior automatically satisfies the metric entropy condition and prior concentration condition for the sequence $c\epsilon_n$ for some $c > 0$. Although the *bracketing entropy* $\log N_{[\cdot]}(\epsilon, \mathcal{P}, d)$ can be larger than the ordinary metric entropy $\log N(\epsilon, \mathcal{P}, d)$, often they are of equal order. In such cases, the construction leads to the optimal rate of convergence of the posterior distribution in the sense that the frequentist minimax rate is achieved. Since the minimax rate is unbeatable, this recipe of prior construction leads to the best possible posterior convergence rate. The construction can be extended to noncompact parameter spaces with the help of sieves.

As for specific applications of the bracketing techniques, consider the Hölder class of densities with smoothness α , which is roughly defined as the class of densities on a compact interval with α -many continuous derivatives. The bracketing entropy grows as $\epsilon^{-1/\alpha}$ in this case. This leads to the rate equation $n\epsilon^2 = \epsilon^{-1/\alpha}$, leading to the rate $n^{-\alpha/(2\alpha+1)}$, agreeing with the corresponding minimax rate. Another example is provided by the class

of monotone densities, whose bracketing entropy grows as ϵ^{-1} . The corresponding rate equation is $n\epsilon^2 = \epsilon^{-1}$, again leading to the minimax rate $n^{-1/3}$ for this problem.

Finite-dimensional models

The conditions assumed in the rate theorem are suboptimal in the sense that for parametric applications, or some other situation for which the calculation involves Euclidean spaces, the rate equations lead to the best rate only up to a logarithmic factor. It is possible to remove this extra undesirable factor by refining both the entropy condition and the condition on the concentration of prior probability. The entropy condition can be modified by considering the *local entropy* $\log N(\epsilon/2, \{p \in \mathcal{P}_n : \epsilon \leq d(p, p_0) \leq 2\epsilon\}, d)$, which is smaller in finite-dimensional models but is as large as the ordinary metric entropy in many nonparametric models. The condition on the prior is modified to

$$\frac{\Pi\{p : j\epsilon_n < d(p, p_0) \leq 2j\epsilon_n\}}{\Pi(B(p_0, \epsilon_n))} \leq e^{Knj^2\epsilon_n^2} \text{ for all } j. \quad (2.13)$$

With this modification, the posterior convergence rate in parametric families turns out to be $n^{-1/2}$.

Log-spline priors

The improved posterior convergence theorem based on the local entropy condition and (2.13) has a very important application in density estimation with log-spline priors. We form an exponential family of densities by a B-spline basis for α -smooth functions. A prior is induced on the densities through independent uniform priors on the coefficients. The exponential family based on B-splines approximates any α -smooth density within $J^{-\alpha}$, where J is the number of basis elements, so we need to keep increasing J with n appropriately. For a given J , the whole calculation can be done in \mathbb{R}^J . The rate equation is then essentially given by $J \sim n\epsilon_n^2$. However, since the convergence rate ϵ_n cannot be better than the rate of approximation $J^{-\alpha}$, the best trade-off is obtained by $J = J_n \sim n^{1/(1+2\alpha)}$ and $\epsilon_n \sim n^{-\alpha/(1+2\alpha)}$.

Applications of rate theorems to Dirichlet mixtures and Gaussian process priors are more involved.

Dirichlet mixtures

For the Dirichlet mixture of normal kernel, we consider two separate cases:

- (a) the *supersmooth* case when the true density itself is a mixture of normal with standard deviation lying between two positive numbers,

- (b) the *ordinary smooth* case when the true density is twice-continuously differentiable but need not itself be a mixture of normal.

The Dirichlet mixture prior used in the first case restricts the variation of the standard deviation of the normal kernel in between the two known bounds for it. Assume further that both the mixing distribution and the base measure of the Dirichlet process are compactly supported. Then one can approximate a normal mixture within ϵ by a finite mixture of normal with only $\mathbb{O}(\log \frac{1}{\epsilon})$ support points. Then the calculation essentially reduces to that in a simplex of dimension $N = \mathbb{O}(\log \frac{1}{\epsilon})$. Entropy of the N -simplex grows as ϵ^{-N} while the concentration rate of a Dirichlet distribution is $e^{-cN \log \frac{1}{\epsilon}}$. This shows that the Hellinger metric entropy grows as $\log^2 \frac{1}{\epsilon}$ and the concentration rate of the Dirichlet mixture prior in Kullback–Leibler neighborhoods of size ϵ is $e^{-c \log^2 \frac{1}{\epsilon}}$. Equating $n\epsilon^2 = \log^2 \frac{1}{\epsilon}$, the best rate of convergence $n^{-1/2} \log n$ is obtained. The compactness conditions assumed above are easy to relax with the consequence of a slight increase in the power of the logarithm. Interestingly, the convergence rate is nearly equal to the parametric convergence rate. Details are given in Ghosal and van der Vaart (2001).

For the ordinary smooth case, one needs to make the scale parameter close to zero with sufficiently high probability, so a sequence of priors for it can be constructed by scaling a fixed prior by some sequence σ_n . A twice-continuously differentiable density can be approximated by such a normal mixture up to σ_n^2 . In this case, the estimate of entropy is $\sigma_n^{-1} \log^2 \frac{1}{\epsilon}$ and the prior concentration rate is $e^{-c\sigma_n^{-1} \log^2 \frac{1}{\epsilon}}$. Equating $\sigma_n^{-1} \log^2 \frac{1}{\epsilon_n}$ with $n\epsilon_n^2$ subject to $\epsilon_n \geq \sigma_n$ gives the optimal frequentist rate $n^{-2/5}$ up to some logarithmic factors. Details are given in Ghosal and van der Vaart (2007b).

Gaussian processes

The rate of convergence for density estimation using a Gaussian process prior was calculated in van der Vaart and van Zanten (2008). In this case, sieves are constructed from Borell’s inequality and the prior concentration rate from small ball probability for Gaussian processes. When the true density function is α -smooth, van der Vaart and van Zanten (2008) showed that by using an integrated Brownian motion (or some other similar processes) whose sample paths are also α -smooth, the minimax rate $n^{-\alpha/(2\alpha+1)}$ is achieved. As mentioned in the last section, another way to construct Gaussian processes with large support is to rescale the covariance kernel of the process by a sequence c_n . For large c_n , the procedure can “pack up” the variation of the Gaussian process on a long interval into a Gaussian process

on the unit interval, resulting in rougher sample paths. Thus, starting with an infinitely smooth kernel, by the rescaling technique, one can approximate any continuous function. Indeed, it was shown in van der Vaart and van Zanten (2007) that this approximation holds in the right order, while entropies and prior concentration change in tune with the rescaling, resulting in the usual rate $n^{-\alpha/(2\alpha+1)}$ for α -smooth densities.

2.5.4 Misspecified models

The general theory discussed so far assumes that the true density belongs to the model, at least in a limiting sense. If the true density maintains a positive distance from the model, the model is called *misspecified*. Experience with parametric cases suggests that the posterior concentrates around the *Kullback–Leibler projection* of the true density, that is, the density within the model minimizing the Kullback–Leibler divergence from the true density. For general infinite-dimensional cases, a theory of posterior convergence rate for such misspecified models was developed in Kleijn and van der Vaart (2006). In analogy with the well-specified case discussed earlier, one needs to measure the prior concentration rate near the Kullback–Leibler projection and control the size of the sieve. It turns out that some different notion of covering, instead of ordinary metric entropy, is the appropriate concept of size in the misspecified case. The theory can be applied to several examples. An important conclusion from their work is that in the semiparametric linear regression model with unknown error density, the convergence rate of the posterior at the true value of the regression parameter does not suffer any loss if the error density is misspecified, such as a normal density instead of the correct double exponential density.

2.5.5 Non-i.i.d. extensions

Like the theory of consistency, the theory of the convergence rate can also be extended beyond the i.i.d. setup. In Ghosal and van der Vaart (2007a), rate theorems are derived for any general dependence and for any given (sequence of) metric d_n , where one can test the null hypothesis $\theta = \theta_0$ against balls of the type $\{\theta : d_n(\theta, \theta_1) < \xi\epsilon\}$, $d_n(\theta_1, \theta_0) > \epsilon$ and ξ is a universal constant, with both type I and type II error probabilities bounded by $e^{-cn\epsilon^2}$. In this case, the rate of convergence ϵ_n can again be characterized as the smallest solution of the entropy inequality

$$\sup_{\epsilon > \epsilon_n} \log N(\xi\epsilon, \{\theta : \epsilon < d_n(\theta, \theta_0) \leq 2\epsilon\}, d_n) \leq n\epsilon_n^2 \quad (2.14)$$

and meeting the condition on concentration probability in a Kullback–Leibler neighborhood of the joint density

$$\Pi\{\theta : K(p_{\theta_0}^n; p_{\theta}^n) \leq n\epsilon_n^2, \quad V(p_{\theta_0}^n; p_{\theta}^n) \leq n\epsilon_n^2\} \geq e^{-n\epsilon_n^2}. \quad (2.15)$$

Admittedly, the statement looks complicated, but substantial simplification is possible in the important special cases of independent, non-identically distributed (i.n.i.d.) variables and Markov processes. For i.n.i.d. variables, the Kullback–Leibler divergence measures can be replaced by the sum of individual divergence measures and the testing condition holds automatically if d_n is the root average squared Hellinger distance $d_n^2(\theta, \theta_0) = n^{-1} \sum_{i=1}^n d^2(p_{i,\theta}, p_{i,\theta_0})$, so entropies need to be evaluated when the distance is measured by d_n . For Markov processes, the root average squared Hellinger distance is given by $\int d^2(p_{\theta_1}(\cdot|x), p_{\theta_2}(\cdot|x))d\nu(x)$, where $p_{\theta}(\cdot|x)$ stands for the transition density and ν is a probability measure. Thus the square-root average squared Hellinger distance emerges as the canonical metric for i.n.i.d. or Markov models, playing the role of the Hellinger distance for i.i.d. observations. The non-i.i.d. extension is extremely useful in estimations involving nonparametric regression with fixed covariates, estimation of the spectral density of a time series using the Whittle likelihood and so on. Explicit convergence rates were obtained in Ghosal and van der Vaart (2007a) for these models and various choices of prior distributions. For instance, for a nonparametric normal regression model with nonrandom covariates, the rate $n^{-\alpha/(2\alpha+1)}$ is obtained using a suitable random B-spline series prior when the true regression function is α -smooth.

2.6 Adaptation and model selection

2.6.1 Motivation and description

Given the level of smoothness, we have seen in the last section that the minimax rate of convergence may be achieved using a suitable prior distribution. However, it is important to know the level of smoothness to construct the appropriate prior. For instance, in constructing a prior for α -smooth densities using an exponential family based on splines, the number of elements in the B-spline basis was chosen depending on the value of α . In particular, this implies that different priors are needed for different smoothness levels. In practice, the smoothness level of the target class is rarely known, so a prior appropriate for a hypothesized class will give suboptimal rate at a true density if it is actually smoother or coarser than the wrongly targeted class.

Therefore the question arises whether we can actually achieve the optimal rate by using a prior constructed without using the actual knowledge of smoothness. If such a prior exists, then the posterior is called *rate adaptive* or simply *adaptive* in short.

In classical statistics, estimators with optimal mean squared error have been constructed in the adaptive framework, that is, without knowing the correct smoothness level. The property can be considered as an *oracle property*, in the sense that the lack of knowledge of the smoothness does not diminish the performance of the estimator compared to the oracle, which uses the extra information on smoothness. Although in classical statistics, even the constant appearing in the limiting distribution may be matched with that of the oracle estimator, such a goal will not be achieved in the Bayesian framework since the best constants in our nonadaptive posterior convergence theorems are also unknown.

From the Bayesian point of view, it is natural to treat the unknown level of smoothness α as a parameter and put a prior distribution on it. Given the value of α , a prior for the class of densities may be obtained as before, using the spline based exponential family. Thus the resulting two-level hierarchical prior is actually a mixture of the spline series prior constructed before. Then the natural question is whether the resulting posterior distribution will converge at the right rate for all smoothness levels, or at least for a rich class of smoothness levels.

It can be seen that there is a strong connection between adaptation and the posterior probability attached to various smoothness levels. In the hierarchy of the models indexed by a smoothness parameter, classes of densities are nested, so a coarser level of smoothness spans a bigger class, increasing the size of the support of the prior. This corruption of support leads to larger entropy estimates slowing down the rate, unless the additional portion can be handled separately. Thus adaptation requires showing that the posterior probability of obtaining a coarser model converges to zero. The treatment of smoother models is somewhat different. Assuming that the correct smoothness level is chosen by the prior with positive probability, the contribution of the smoother priors in the mixture can be ignored for the purpose of lower bounding the prior concentration, at the expense of incorporating an irrelevant constant, which is eventually ignored for the purpose of calculating the rate. Clearly, smoother priors do not contribute to increasing the size of the support of the mixture prior. Thus, if the posterior asymptotically ignores models coarser than the true one, adaptation is expected to take place.

2.6.2 Infinite-dimensional normal models

The natural strategy for adaptation works in several models. We begin with one of the simplest cases, the infinite-dimensional normal model: $X_i \stackrel{\text{ind}}{\sim} \text{Nor}(\theta_i, n^{-1})$, $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots) \in \ell_2$, considered in Belitser and Ghosal (2003). The “smoothness” of $\boldsymbol{\theta}$ is measured by the behavior of the tail of the sequence $\theta_1, \theta_2, \dots$, and is defined as the largest q for which $\sum_{i=1}^{\infty} i^{2q} \theta_i^2 < \infty$. It is well known that the minimax rate of estimation of $\boldsymbol{\theta}$ on the subspace $\sum_{i=1}^{\infty} i^{2q} \theta_i^2 < \infty$ is $n^{-q/(2q+1)}$. The minimax rate is attained by the Bayes estimator with respect to the prior $\theta_i \sim \text{Nor}(0, i^{-(2q+1)})$, which we denote by Π_q . This assertion follows by direct calculations of posterior mean and variance and bounding posterior probabilities of deviations by Chebyshev’s inequality. However, the prior is clearly dependent on the unknown smoothness level q . We consider only countably many possible values of q and suppose that they do not accumulate at any point other than 0 or ∞ . Hence we can arrange the possible values of q in an increasing double sequence $\dots, q_{-1}, q_0, q_1, \dots$, where q_0 is the true value of smoothness. We attach positive weights λ_q to each possible value of q and mix to form the hierarchical prior $\sum_q \lambda_q \Pi_q$. By using explicit properties of normal likelihood, an upper bound for the posterior probability of each q can be obtained, leading to exponential-type decay $\Pi(q < q_0 | X_1, X_2, \dots) \leq e^{-cn^\delta}$ with $\delta > 0$. Here, the fact that the q values are separated by a positive distance plays a crucial role. Thus the role of $q < q_0$ is asymptotically negligible. To treat the case $q > q_0$, again due to positive separation, it is easily shown by Chebyshev’s inequality that the contribution of Π_q , $q > q_0$, to the posterior probability of $\mathcal{C}^c := \{\sum_{i=1}^{\infty} i^{2q_0} \theta_i^2 > B\}$ is negligible for large B . Thus what matters eventually is the contribution of Π_{q_0} , for which the correct rate $n^{-q_0/(2q_0+1)}$ is already in force, and that of Π_q , $q > q_0$, restricted to \mathcal{C} . The entropy of \mathcal{C} grows as ϵ^{-1/q_0} , while the rate of concentration of the mixture prior, due to the presence of the component Π_{q_0} in the convex combination, is at least $\lambda_{q_0} \exp(-c\epsilon^{-1/q_0})$. The last assertion is a consequence of tail behavior of random $\boldsymbol{\theta}$ from Π_{q_0} and for large N , a lower bound for the probability of a ball $\{\sum_{i=1}^N (\theta_i - \theta_{i0})^2 \leq \delta\}$ of the form e^{-cN} . Therefore the natural Bayesian strategy for adaptation works well in the infinite-dimensional normal model.

2.6.3 General theory of Bayesian adaptation

For countably many competing abstract models indexed by α , say, a general result was obtained in Ghosal, Lember and van der Vaart (2008) based on some earlier results by Ghosal, Lember and van der Vaart (2003), Huang

(2004) and Lember and van der Vaart (2007). We allow the index set A for α and the prior λ_α to depend on n , but we do not make n explicit in the notation. We work with the Hellinger distance for definiteness. Let $\epsilon_{n,\alpha}$ be the usual optimal rate in model α and β stand for the “true value” of α in the sense that the β th model is the best approximating model for the true density p_0 . Let $B_\beta(\epsilon)$ be the Kullback–Leibler neighborhood in the true model. For simplicity, we do not include sieves in the condition by assuming that the required entropy condition holds in the original model, but the result will be easily modified to the general case assuming that sieves have exponentially small prior probability $e^{-cn\epsilon_{n,\beta}^2}$. Essentially four basic conditions ensure adaptation.

- (i) For each model α , the local entropy condition $\log N(\epsilon/3, C_\alpha(\epsilon), d) \leq E_\alpha \epsilon_{n,\alpha}^2$ for all $\epsilon > \epsilon_{n,\alpha}$ holds, where $C_\alpha(\epsilon) = \{p : d(p, p_0) \leq 2\epsilon\}$ and E_α is a constant free of n .
- (ii) For coarser models $\alpha < \beta$ and any positive integer j , the net prior probabilities of $C_\alpha(j\epsilon_{n,\alpha})$ compare to that of $B_\beta(\epsilon_{n,\beta})$ by the condition

$$\frac{\lambda_\alpha \Pi_\alpha(C_\alpha(j\epsilon_{n,\alpha}))}{\lambda_\beta \Pi_\beta(B_\beta(\epsilon_{n,\beta}))} \leq \mu_\alpha e^{Lj^2 n \epsilon_{n,\alpha}^2}.$$

- (iii) For finer models $\alpha \geq \beta$, the net prior probabilities of $C_\alpha(j\epsilon_{n,\beta})$ compare to that of $B_\beta(\epsilon_{n,\beta})$ by the condition

$$\frac{\lambda_\alpha \Pi_\alpha(C_\alpha(j\epsilon_{n,\beta}))}{\lambda_\beta \Pi_\beta(B_\beta(\epsilon_{n,\beta}))} \leq \mu_\alpha e^{Lj^2 n \epsilon_{n,\beta}^2}.$$

- (iv) For a sufficiently large B , the total prior mass of a $B\epsilon_{n,\alpha}$ -ball in coarser models compared to the concentration rate in the true model is significantly small in that

$$\sum_{\alpha < \beta} \frac{\lambda_\alpha \Pi_\alpha(C_\alpha(B\epsilon_{n,\alpha}))}{\lambda_\beta \Pi_\beta(B_\beta(\epsilon_{n,\beta}))} = o(e^{-2n\epsilon_{n,\beta}^2});$$

here the constants μ_α satisfy $\sum \sqrt{\mu_\alpha} \leq e^{n\epsilon_{n,\beta}^2}$.

In the above, we restricted attention to nonparametric models. If parametric models with $n^{-1/2}$ -convergence rates are involved, then slight modifications of the conditions are necessary, as argued in Section 2.5.3, to avoid an undesirable logarithmic factor. In the present situation, the prior concentration level in ϵ -balls should be given by a power of ϵ , say ϵ^D and the right-hand side of condition (iv) should be replaced by $o(n^{-3D})$.

It may be seen that in order to satisfy the required conditions, one may control the weight sequence λ_α suitably, in particular, depending on n . The choice $\lambda_\alpha \propto \mu_\alpha e^{-cn\epsilon_{n,\alpha}^2}$ is sometimes fruitful.

2.6.4 Density estimation using splines

The above general result applies directly in the context of density estimation in Hölder α -classes with log-spline priors as discussed in the last section. Interestingly, a simple hierarchical prior obtained by using a single prior on α leads to the optimal rate $n^{-\alpha/(2\alpha+1)}$ only up to a logarithmic factor, as in Ghosal, Lember and van der Vaart (2003). The logarithmic factor was removed by using very special weights in Huang (2004), who also treated a similar problem for nonparametric regression using a wavelet basis. This result requires restricting to finitely many competing smoothness classes. For such a case, the same result can be obtained from the general theory of Ghosal, Lember and van der Vaart (2008) by using a sequence of weights $\lambda_\alpha \propto \prod_{\gamma < \alpha} (C\epsilon_{n,\gamma})^{J_{n,\gamma}}$ where, as before, $J_{n,\alpha} \sim n^{1/(1+2\alpha)}$ is the dimension of the optimal spline approximation model. Another possibility is to consider a discrete uniform prior on $\epsilon_{n,\alpha}$ -nets in Hölder classes with $\lambda_\alpha \propto \mu_\alpha e^{-cn\epsilon_{n,\alpha}^2}$.

As discussed in Section 2.6.1 and exhibited in the infinite normal model, it is to be expected that the posterior probability of models coarser than the true one combined together should be negligible. On the other hand, the posterior probability of the complement of a large multiple of $\epsilon_{n,\beta}$ -balls in smoother models is small. In particular, if the true density lies outside the closure of these models, then the posterior probability of smoother models also converges to zero.

2.6.5 Bayes factor consistency

The results become much more transparent when we consider just a pair of competing models. The conclusion is equivalent to the *consistency of the Bayes factor* of the smoother model relative to the coarser model – the Bayes factor goes to infinity if the true density belongs to the smoother model while the Bayes factor goes to zero otherwise. Since we allow the true density to lie outside the models, we interpret Bayes factor consistency in the following generalized sense: the Bayes factor converges to zero if the true density stays away from the smoother models by an amount larger than its convergence rate, while the Bayes factor converges to infinity if the true density is within the convergence rate of the smoother model.

The asymptotic behavior of Bayes factors has been studied by many authors in the parametric case, but only a few results are available in the nonparametric case. When the smaller model is a singleton and the prior in the larger model satisfies the Kullback–Leibler property, then Dass and Lee (2004) showed Bayes factor consistency. Their proof uses Doob’s consistency theorem and is heavily dependent on the assumption that there is a positive prior mass at the true density. Hence it is extremely difficult to generalize this result to composite null hypotheses. Another result was obtained by Walker, Damien and Lenk (2004), who showed that if the Kullback–Leibler property holds in one model and does not hold in another model, then the Bayes factor shows the same kind of dichotomy. This result helps only if the two models separate well, but such a situation is rare. What is commonly observed is nested families with differing convergence rates. In such a case, consistency of Bayes factors can be obtained from the calculation done in our result on adaptation. We can state the result roughly as follows.

Let $\epsilon_{n,1} > \epsilon_{n,2}$ be the two possible rates in the respective competing models and suppose, for simplicity, the models are given equal weight. Assume that the prior of the $\epsilon_{n,j}$ -size Kullback–Leibler neighborhood around the true density in model j is at least $e^{-n\epsilon_{n,j}^2}$, $j = 1, 2$, and further the prior probability of a large multiple of an $\epsilon_{n,1}$ -size Hellinger ball in model 1 is at most $o(e^{-3n\epsilon_{n,1}^2})$. Then the Bayes factor is consistent.

The result applies readily to some goodness-of-fit tests for parametric models against nonparametric alternatives; see Ghosal, Lember and van der Vaart (2008) for details.

2.7 Bernshtein–von Mises theorems

2.7.1 Parametric Bernshtein–von Mises theorems

The convergence rate of a posterior distribution asserts concentration at the true value of the parameter at a certain speed, but does not tell us the asymptotic shape of the posterior distribution. For smooth parametric families, a remarkable result, popularly known as the *Bernshtein–von Mises theorem*, says that the posterior distribution asymptotically looks like a normal distribution centered at the MLE with variance equal to the inverse of the Fisher information; see van der Vaart (1998) for example. As a consequence, the posterior distribution of $\sqrt{n}(\theta - \hat{\theta}_n)$ conditioned on the sample, where θ is the (random) parameter and $\hat{\theta}_n$ is the MLE of θ , approximately coincides with its frequentist distribution under the parameter value θ , where $\hat{\theta}_n$ car-

ries the randomness. This is a remarkable, and very mysterious result, since the interpretations of randomness in the two situations are quite different and the two quantities involved in the process are obtained from very different principles. From an application point of view, the importance of the Bernshtein–von Mises theorem lies in its ability to construct approximately valid confidence sets using Bayesian methods. This is very useful especially in complex problems since the sampling distribution is often hard to compute while the samples from the posterior distribution can be obtained relatively easily using various computational devices. The Bernshtein–von Mises phenomenon extends beyond smooth families; see Ghosal, Ghosh and Samanta (1995) for a precise description and a necessary and sufficient condition leading to the asymptotic matching in the first order. The prior, which needs only to have positive and continuous density at the true value of the parameter, plays a relatively minor role in the whole development.

2.7.2 Nonparametric Bernshtein–von Mises theorems

For infinite-dimensional cases, only very few results are available so far. Here, the validity of the phenomenon depends not only on the model, but also heavily on the prior. For estimating a c.d.f. F with a Dirichlet process prior, Lo (1983) showed that the posterior of $\sqrt{n}(F - \mathbb{F}_n)$ converges weakly to the F_0 -Brownian bridge process a.s. under F_0 , where \mathbb{F}_n stands for the empirical c.d.f. On the other hand, the well known Donsker theorem tells us that $\sqrt{n}(\mathbb{F}_n - F_0)$ converges weakly to the F_0 -Brownian bridge. Thus by the symmetry of the Brownian bridge, again we observe that the two limiting distributions coincide, leading to Bayesian matching. The result is extended in the multidimensional case in Lo (1986). Both results use the explicit structure of the Dirichlet process posterior and hence are difficult to extend to other priors. Recently, the result was substantially generalized to a wide class of Lévy process priors for cumulative hazard functions by Kim and Lee (2004) using fine properties of Poisson random measures and corresponding conjugacy results. The result concludes that the posterior distribution of $\sqrt{n}(H - \mathbb{H}_n)$, where H is the cumulative hazard function and \mathbb{H}_n is the Nelson–Aalen estimator, converges weakly to a Brownian bridge process a.s. under the true cumulative hazard H_0 . From classical survival analysis and the martingale central limit theorem, it is known that the same Brownian bridge process appears as the limit of $\sqrt{n}(\mathbb{H}_n - H_0)$ when H_0 is the true cumulative hazard function.

2.7.3 Semiparametric Bernshtein–von Mises theorems

For semiparametric problems, often only the parametric part θ is of interest. Thus the marginal posterior distribution of θ is especially of interest. One expects that, in spite of the possibly slower convergence rate of the posterior distribution of the nonparametric part η , the convergence rate for θ is still $n^{-1/2}$. Indeed, one can hope that the Bernshtein–von Mises phenomenon holds in the sense that the marginal posterior distribution of $\sqrt{n}(\theta - \hat{\theta}_n)$ is asymptotically normal and asymptotically coincides with the frequentist distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ a.s. under the distribution generated by (θ_0, η_0) for most pairs of true values (θ_0, η_0) of (θ, η) . A major barrier in deriving such a result is the fact that even the marginal posterior of θ is obtained through the posterior of the whole parameter (θ, η) , making it difficult to study unless explicit expressions are available. A semiparametric Bernshtein–von Mises theorem was obtained in Shen (2002), but it seems that the conditions are somewhat difficult to verify. More transparent results with mild and verifiable conditions are highly desirable. For the specific case of the Cox proportional hazard model, a semiparametric Bernshtein–von Mises theorem was obtained by Kim (2006).

2.7.4 Non-existence of Bernshtein–von Mises theorems

All the Bernshtein–von Mises results obtained thus far in infinite-dimensional models appear to be associated with convergence rate $n^{-1/2}$. In principle, there is no reason why a Bernshtein–von Mises theorem should be restricted to the $n^{-1/2}$ -convergence domain. However, certain negative results lead to the suspicion that the Bernshtein–von Mises theorem may not hold for slower convergence. The first result of this kind was obtained in Cox (1993), where the problem of estimating the signal in a white noise model was considered, and the signal was assumed to lie in a *Sobolev space*. The prior was obtained by expanding the function in a suitable series and putting independent normal priors on the coefficients. It was observed that the coverage probability of credible intervals constructed from the posterior distribution can converge to an arbitrarily small number with increasing sample size. Thus not only does the Bernshtein–von Mises theorem fail, but it fails in the worst possible way. The main problem seems to be the optimal trade-off used in smoothing, making the order of the bias the same as that of the standard deviation. For the infinite-dimensional normal model discussed in Section 2.6.2, it was also shown in Freedman (1999) that the frequentist and the Bayesian distributions of the L_2 -norm of the difference of the Bayes estimator and the parameter differ by an amount equal to the scale

of interest. This again leads to the conclusion that the frequentist coverage probability of a Bayesian credible set for the parameter can be infinitesimally small. This is disappointing since a confidence set must now be found using only frequentist methods and hence is harder to obtain. Interestingly, if a (sequence of) functional depends only on the first p_n coordinates where $p_n/\sqrt{n} \rightarrow 0$, then the Bernstein–von Mises theorem holds for that functional. Indeed, the posterior distribution of the entire p_n -vector consisting of the first p_n -coordinates centered by the MLE is approximated by the p_n -dimensional normal distribution which approximates the distribution of the MLE; see Ghosal (2000) for details.

2.8 Concluding remarks

The nonparametric Bayesian approach to solving problems is rapidly gaining popularity among practitioners as theoretical properties become increasingly better understood and computational hurdles are being removed. Nowadays, many new Bayesian nonparametric methods for complex models arising in biomedical, geostatistical, environmental, econometric and many other applications are being proposed. The purpose of the present article is to give a brief outline of the fundamental basis of prior construction and large sample behavior of the posterior distribution. Naturally, it has not been possible to cite every relevant paper in this article.

In this article, we have reviewed the properties of the Dirichlet process, other priors constructed using the Dirichlet process and asymptotic properties of the posterior distribution for nonparametric and semiparametric problems. We discussed a naive construction of the Dirichlet process, indicated measure theoretic difficulties associated with the approach and subsequently rectified the problem by working with a suitable countable generator. We also discussed a method of construction using a gamma process. We then discussed basic properties of the Dirichlet process such as expressions of prior mean and variance, interpretation of the parameters and posterior conjugacy with implications such as explicit expressions for posterior expectation and variance, and some peculiar consequences of the presence of point mass in the base measure for the posterior Dirichlet process. We further discussed the discreteness property, characterization of support, self-similarity, convergence, tail behavior and mutual singularity of the Dirichlet process. The joint distribution of samples drawn from a random probability measure obtained from a Dirichlet process was described by the Blackwell–MacQueen generalized Pólya urn scheme and its relation to MCMC sampling, clustering and distribution of distinct observations was discussed. We described

Sethuraman's stick-breaking representation of a Dirichlet process and its potential role in constructing new processes and computation involving Dirichlet processes.

The Dirichlet process leads to various new processes through some common operations. We discussed the role of a mixture of Dirichlet processes in eliciting a prior distribution. We then thoroughly discussed the importance of kernel smoothing applied to a Dirichlet process leading to Dirichlet mixtures. Computational techniques through MCMC sampling using the generalized Pólya urn structure were briefly outlined. We also mentioned the role of symmetrization and conditioning operations especially in semiparametric applications leading to respectively the symmetrized Dirichlet and conditioned Dirichlet processes.

We discussed consistency and its implications thoroughly. We started with a very general theorem by Doob which concludes consistency at almost all parameter values with respect to the prior measure. However, null sets could be huge and we discussed some prominent examples of inconsistency. In fact, we mentioned that this inconsistency behavior is more common than consistency in a topological sense if the prior distribution is completely arbitrarily chosen. We then discussed the tail-freeness property and its role in providing consistency. In particular, consistency follows for Pólya tree priors. We then presented the general formulation of Schwartz's theory of consistency and the role of the condition of prior positivity in Kullback–Leibler neighborhoods. We argued that a size condition on the model described by the existence of uniformly exponentially consistent tests plays a key role, which further reduces to entropy conditions for commonly used metrics. This is because a desired test can be constructed by covering the space with small balls guided by the bounds for metric entropy, finding appropriate tests against these balls and combining these basic tests. The role of sieves in compactifying a noncompact space is shown to be extremely important. The general result on consistency was applied to density estimation using Dirichlet mixtures or Gaussian process priors leading to very explicit conditions for consistency for these commonly used priors. We also argued that Schwartz's theory is an appropriate tool for studying posterior consistency in semiparametric problems. Further we indicated how Schwartz's theory can be extended to the case of independent nonidentically distributed and some dependent observations, with applications to estimating binary regression, spectral density and transition density using commonly used priors.

We next studied convergence rates of posterior distribution. We discussed the theory developed in Ghosal, Ghosh and van der Vaart (2000) refining Schwartz's theory for consistency. It was again seen that the concentration

of priors in Kullback–Leibler type neighborhoods and the growth rate of entropy functions determine the rate. We explicitly constructed priors using bracketing approximations or exponential families based on splines to achieve optimal rates. We further discussed the convergence rate of Dirichlet mixture priors and Gaussian priors under various setups. Extensions to misspecified or non-i.i.d. models were briefly indicated with applications.

The issue of adaptation was considered next. It was argued that mixing “optimal priors” by putting a prior distribution on the indexing parameter is a prudent strategy. We described conditions under which this natural Bayesian strategy works, and in particular showed that the adaptation takes place in an infinite-dimensional normal model and class of log-spline densities. We also discussed the connection between adaptation and model selection. Under the same conditions, we argued that the posterior probability of the true model increases to one leading to the consistency of the Bayes factor when only two models are present. Finally, we discussed the importance of the Bernshtein–von Mises theorem and mentioned some instances where the theorem holds true, and where it fails.

Although a lot of development has taken place in the last ten years, we still need to know much more about finer properties of the posterior distribution in various applications. Asymptotics can play a key role in separating a desirable approach from an undesirable one. It will be of great interest to compose a catalog of priors appropriate for applications having desirable posterior convergence properties.

References

- Amewou-Atisso, M., Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (2003). Posterior consistency for semiparametric regression problems. *Bernoulli*, **9**, 291–312.
- Antoniak, C. (1974). Mixtures of Dirichlet processes with application to Bayesian non-parametric problems. *Annals of Statistics*, **2**, 1152–74.
- Barron, A. R. (1999). Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In *Bayesian Statistics 6*, ed. J. M. Bernardo et al., 27–52. Oxford: Oxford University Press.
- Barron, A. R., Schervish, M. and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Annals of Statistics*, **27**, 536–61.
- Belitser, E. N. and Ghosal, S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Annals of Statistics*, **31**, 536–59.
- Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Annals of Statistics*, **1**, 353–55.
- Choi, T. and Schervish, M. (2007). Posterior consistency in nonparametric Bayesian problems using Gaussian process prior. *Journal of Multivariate Analysis*, **98**, 1969–87.

- Choudhuri, N., Ghosal, S. and Roy, A. (2004). Bayesian estimation of the spectral density of a time series. *Journal of the American Statistical Association*, **99**, 1050–59.
- Coram, M. and Lalley, S. P. (2006). Consistency of Bayes estimators of a binary regression. *Annals of Statistics*, **34**, 1233–69.
- Cox, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Annals of Statistics*, **21**, 903–23.
- Dalal, S. R. (1979). Dirichlet invariant processes and applications to nonparametric estimation of symmetric distribution functions. *Stochastic Processes and Their Applications*, **9**, 99–107.
- Dass, S. C. and Lee, J. (2004). A note on the consistency of Bayes factors for testing point null versus nonparametric alternatives. *Journal of Statistical Planning and Inference*, **119**, 143–52.
- Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates (with discussion). *Annals of Statistics*, **14**, 1–67.
- Diaconis, P. and Freedman, D. (1993). Nonparametric binary regression: a Bayesian approach. *Annals of Statistics*, **21**, 2108–37.
- Doob, J. L. (1948). Application of the theory of martingales. In *Le Calcul des Probabilités et ses Applications*, Colloques Internationaux du Centre National de la Recherche Scientifique, 13, 23–37. Paris: CNRS.
- Doss, H. (1985a). Bayesian nonparametric estimation of the median. I: Computation of the estimates. *Annals of Statistics*, **13**, 1432–44.
- Doss, H. (1985b). Bayesian nonparametric estimation of the median. II: Asymptotic properties of the estimates. *Annals of Statistics*, **13**, 1445–64.
- Doss, H. and Sellke, T. (1982). The tails of probabilities chosen from a Dirichlet process. *Annals of Statistics*, **10**, 1302–05.
- Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**, 577–88.
- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209–30.
- Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics*, ed. M. Rizvi, J. Rustagi and D. Siegmund, 287–302. New York: Academic Press.
- Freedman, D. (1963). On the asymptotic distribution of Bayes estimates in the discrete case. I. *Annals of Mathematical Statistics*, **34**, 1386–403.
- Freedman, D. (1999). On the Bernstein–von Mises theorem with infinite-dimensional parameters. *Annals of Statistics*, **27**, 1119–40.
- Ghosal, S. (2000). Asymptotic normality of posterior distributions for exponential families with many parameters. *Journal of Multivariate Analysis*, **74**, 49–69.
- Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1999a). Posterior consistency of Dirichlet mixtures in density estimation. *Annals of Statistics*, **27**, 143–58.
- Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1999b). Consistent semiparametric Bayesian inference about a location parameter. *Journal of Statistical Planning and Inference*, **77**, 181–93.
- Ghosal, S., Ghosh, J. K. and Samanta, T. (1995). On convergence of posterior distributions. *Annals of Statistics*, **23**, 2145–52.
- Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Annals of Statistics*, **28**, 500–31.
- Ghosal, S., Lember, J. and van der Vaart, A. W. (2003). On Bayesian adaptation. *Acta Applicandae Mathematica*, **79**, 165–75.

- Ghosal, S., Lember, J. and van der Vaart, A. W. (2008). Nonparametric Bayesian model selection and averaging. *Electronic Journal of Statistics*, **2**, 63–89.
- Ghosal, S. and Roy, A. (2006). Posterior consistency of Gaussian process prior for nonparametric binary regression. *Annals of Statistics*, **34**, 2413–29.
- Ghosal, S. and Tang, Y. (2006). Bayesian consistency for Markov processes. *Sankhyā*, **68**, 227–39.
- Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixture of normal densities. *Annals of Statistics*, **29**, 1233–63.
- Ghosal, S. and van der Vaart, A. W. (2007a). Convergence of posterior distributions for non iid observations. *Annals of Statistics*, **35**, 192–223.
- Ghosal, S. and van der Vaart, A. W. (2007b). Posterior convergence of Dirichlet mixtures at smooth densities. *Annals of Statistics*, **29**, 697–723.
- Ghosal, S. and van der Vaart, A. W. (2009). *Theory of Nonparametric Bayesian Inference*. Cambridge: Cambridge University Press, to appear.
- Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. New York: Springer-Verlag.
- Huang, T. Z. (2004). Convergence rates for posterior distribution and adaptive estimation. *Annals of Statistics*, **32**, 1556–93.
- Kim, Y. (2006). The Bernstein–von Mises theorem for the proportional hazard model. *Annals of Statistics*, **34**, 1678–700.
- Kim, Y. and Lee, J. (2001). On posterior consistency of survival models. *Annals of Statistics*, **29**, 666–86.
- Kim, Y. and Lee, J. (2004). A Bernstein–von Mises theorem in the nonparametric right-censoring model. *Annals of Statistics*, **32**, 1492–512.
- Kleijn, B. and van der Vaart, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Annals of Statistics*, **34**, 837–77.
- Korwar, R. and Hollander, M. (1973). Contributions to the theory of Dirichlet processes. *Annals of Statistics*, **1**, 706–11.
- Kraft, C. H. (1964). A class of distribution function processes which have derivatives. *Journal of Applied Probability*, **1**, 385–88.
- Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modeling. *Annals of Statistics*, **20**, 1222–35.
- Lember, J. and van der Vaart, A. W. (2007). On universal Bayesian adaptation. *Statistics and Decisions*, **25**, 127–52.
- Lijoi, A., Prünster, I. and Walker, S. G. (2005). On consistency of nonparametric normal mixtures for Bayesian density estimation. *Journal of the American Statistical Association*, **100**, 1292–96.
- Lo, A. Y. (1983). Weak convergence for Dirichlet processes. *Sankhyā, Series A*, **45**, 105–11.
- Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates. I: Density estimates. *Annals of Statistics*, **12**, 351–57.
- Lo, A. Y. (1986). A remark on the limiting posterior distribution of the multiparameter Dirichlet process. *Sankhyā, Series A*, **48**, 247–49.
- Regazzini, E., Guglielmi, A. and Di Nunno, G. (2002). Theory and numerical analysis for exact distributions of functionals of a Dirichlet process. *Annals of Statistics*, **30**, 1376–1411.
- Schwartz, L. (1965). On Bayes procedures. *Probability Theory and Related Fields*, **4**, 10–26.

- Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–50.
- Shen, X. (2002). Asymptotic normality of semiparametric and nonparametric posterior distributions. *Journal of the American Statistical Association*, **97**, 222–35.
- Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Annals of Statistics*, **29**, 687–714.
- Tang, Y. and Ghosal, S. (2007). Posterior consistency of Dirichlet mixtures for estimating a transition density. *Journal of Statistical Planning and Inference*, **137**, 1711–26.
- Teh, Y. W., Jordan, M. I., Beal, M. and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, **101**, 1566–81.
- Tokdar, S. T. (2006). Posterior consistency of Dirichlet location-scale mixture of normal density estimation and regression. *Sankhyā*, **67**, 90–110.
- Tokdar, S. T. and Ghosh, J. K. (2007). Posterior consistency of Gaussian process priors in density estimation. *Journal of Statistical Planning and Inference*, **137**, 34–42.
- van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.
- van der Vaart, A. W. and van Zanten, H. (2007). Bayesian inference with rescaled Gaussian process prior. *Electronic Journal of Statistics*, **1**, 433–48.
- van der Vaart, A. W. and van Zanten, H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Annals of Statistics*, **36**, 1435–63.
- Walker, S. G. (2004). New approaches to Bayesian consistency. *Annals of Statistics*, **32**, 2028–43.
- Walker, S. G., Damien, P. and Lenk, P. (2004). On priors with a Kullback–Leibler property. *Journal of the American Statistical Association*, **99**, 404–8.
- Walker, S. G. and Hjort, N. L. (2001). On Bayesian consistency. *Journal of the Royal Statistical Society, Series B*, **63**, 811–21.
- Walker, S. G., Lijoi, A. and Prünster, I. (2005). Data tracking and the understanding of Bayesian consistency. *Biometrika*, **92**, 765–78.
- Walker, S. G., Lijoi, A. and Prünster, I. (2007). On rates of convergence of posterior distributions in infinite-dimensional models. *Annals of Statistics*, **35**, 738–46.
- Wu, Y. and Ghosal, S. (2008). Kullback–Leibler property of kernel mixture priors in Bayesian density estimation. *Electronic Journal of Statistics*, **2**, 298–331.