Bernoulli **4**(4), 1998, 411–444

Consistency of Bayes estimates for nonparametric regression: normal theory

PERSI W. DIACONIS[1] and DAVID FREEDMAN[2*]

[1]*Mathematics Department and ORIE, Cornell University, Ithaca NY 14853, USA*
[2]*Statistics Department, University of California, Berkeley CA 94720, USA.*
E-mail: freedman@stat.berkeley.edu

Performance characteristics of Bayes estimates are studied. More exactly, for each subject in a data set, let ξ be a vector of binary covariates and let Y be a normal response variable, with $\mathrm{E}\{Y|\xi\} = f(\xi)$ and $\mathrm{var}\{Y|\xi\} = 1$. Here, f is an unknown function to be estimated from the data; the subjects are independent and identically distributed. Define a prior distribution on f as $\sum_k w_k \pi_k / \sum_k w_k$, where π_k is standard normal on the set of f which only depend on the first k covariates and $w_k > 0$ for infinitely many k. Bayes estimates are consistent for all f. On the other hand, if the π_k are flat, inconsistency is the rule.

Keywords: Consistency; Bayes estimates; model selection; binary regression

1. Introduction

Consider a sequence of independent pairs (Y_1, ξ_1), (Y_2, ξ_2), ... Given ξ_i, suppose Y_i is normally distributed with conditional mean $f(\xi_i)$ and conditional variance 1. Thus, $Y_i = f(\xi_i) + \varepsilon_i$, where $\varepsilon_i \sim N(0, 1)$. Here, f is an unknown function to be estimated from the data. A Bayesian approach postulates that f lies in some class of functions Θ and puts a prior distribution π on Θ. This generates a posterior distribution $\tilde{\pi}_n$, namely, the conditional law of the regression function f given the data (Y_1, ξ_1), (Y_2, ξ_2), ..., (Y_n, ξ_n). The prior π is said to be *consistent* for f if $\tilde{\pi}_n$ converges to point mass at f almost surely as $n \to \infty$.

When Θ is finite-dimensional, π will be consistent for any f in the support of π; of course, some additional regularity conditions are needed, but normality is not involved. If Θ is infinite-dimensional, the situation is quite different, and inconsistency is the rule rather than the exception. See, for instance, Freedman (1963; 1965) and Diaconis and Freedman (1988). This paper continues the story for normal models. We show that for conventional hierarchical normal priors, consistency obtains – provided the data are independent with a common normal distribution. We use a nested increasing family of finite-dimensional models with the usual normal prior in each dimension, but the dimension is itself a hyperparameter with its own (discrete) prior. So far, the priors under discussion have been

*To whom correspondence should be addressed.

1350–7265

proper; and technical conditions are given below. With flat priors, inconsistency is the rule, even under our regularity conditions.

In previous papers (Diaconis and Freedman 1993; 1995), we looked at nonparametric binary regression. There, natural priors were generally seen to give consistent estimates; but, under some circumstances, the estimates were inconsistent. This seemed quite mysterious, at least to us. We now have a heuristic understanding of the basic reason for inconsistency – to be explained below – and the present paper is a first test of that heuristic. The following paragraphs explain the background in more detail, and the heuristic. We also give a brief literature review on nonparametric Bayesian regression and consistency theorems.

1.1. Binary regression

First we summarize results from Diaconis and Freedman (1993; 1995). There is a binary response variable Y, which is related to a covariate ξ:

$$P\{Y = 1|\xi\} = f(\xi). \tag{1.1}$$

The problem is to estimate f from the data.

Following de Finetti (1959; 1972), we think of ξ as a sequence of 0s and 1s. Sequence space is given the usual product topology, and the parameter space Θ is the set of measurable functions f from sequence space to [0, 1]. The L_2 topology is installed on Θ, relative to coin-tossing measure λ in sequence space. A basic neighbourhood of $f \in \Theta$ is

$$\mathrm{N}(f, \varepsilon) = \left\{ g: \int (g - f)^2 \, \mathrm{d}\lambda < \varepsilon \right\}. \tag{1.2}$$

We will consider a prior π on Θ, with posterior $\tilde{\pi}_n$. Then π is *consistent* at f provided $\tilde{\pi}_n\{\mathrm{N}(f, \varepsilon)\} \to 1$ almost surely, for all positive ε.

The next step is to define the hierarchical priors on Θ. Begin with a prior π_k supported on the class of functions f that depend only on the first k coordinates, or bits, in ξ. Under π_0, the function f does not depend on ξ at all. Under π_1, f depends only on ξ_1. And so forth. Then treat k as an unknown 'hyperparameter', putting prior weight w_k on k. We refer to k as the *theory index*; theory k says that $f(x)$ depends only the first k bits of x; and w_k is a *theory weight*. Our prior is of the form

$$\pi = \sum_{k=0}^{\infty} w_k \pi_k \Big/ \sum_{k=0}^{\infty} w_k, \tag{1.3a}$$

where

$$w_k > 0 \text{ for all } k \text{ and } \sum_{k=0}^{\infty} w_k < \infty. \tag{1.3b}$$

To complete the description of the prior, π_k must be specified. According to π_k, only the first k bits in ξ matter, so f depends only on $\xi_1, \ldots, \xi_k$. Thus, π_k is determined by specifying the joint distribution of the 2^k possible values for f. More crudely, π_k involves

2^k free parameters – the possible values of f on its intervals of constancy. For now, we take these parameters to be independent and uniformly distributed over [0, 1].

We turn now to the data. For technical reasons, it is simplest to consider 'balanced' data, as in Diaconis and Freedman (1993); more conventional sampling plans are discussed in Diaconis and Freedman (1995). At stage n, there are 2^n subjects. Each has a covariate sequence; the first n bits of these covariate sequences cover all possible patterns of length n; each pattern appears once and only once. The remaining bits from $n+1$ onwards are generated by coin tossing. Given the covariates, response variables are generated from (1.1); the response of subject i depends only on the covariates for that subject. The preliminaries are now finished, and we can state a theorem. (The present paper will make the extension from binary data to normal data in Section 2.)

Theorem 1.1. *With nonparametric binary regression, balanced data, and a hierarchical uniform prior:*

(a) π *is consistent at* f *unless* $f \equiv \frac{1}{2}$;

(b) *Suppose* $f \equiv \frac{1}{2}$. *Then* π *is consistent at* f *provided that for some* $\delta > 0$, *for all sufficiently large* n,

$$\sum_{k=n}^{\infty} w_k < 2^{-n(\frac{1}{2}+\delta)}.$$

On the other hand, π *is inconsistent at* f *provided that for some* $\delta > 0$, *for infinitely many* n,

$$\sum_{k=n}^{\infty} w_k > 2^{-n(\frac{1}{2}-\delta)}.$$

The surprising part of this theorem is the inconsistency result in (b). Suppose the data are generated by tossing a fair coin, so $f \equiv \frac{1}{2}$. Theory 0 is true: f does not depend on ξ at all. You do not know that, and allow theories of finite but arbitrary complexity in your prior, according to (1.3). In the face of all these other theories, the posterior loses faith in theory 0 – the curse of dimensionality strikes again.

Regression is a natural problem, hierarchical priors are often used, and the one defined by (1.3) charges every weak star neighbourhood of the parameter space Θ. Still, inconsistency may result. In high-dimensional problems, little can be taken for granted. 'Rational use of additional information' is not a slogan to be adopted without reflection.

1.2. Why inconsistency?

What is the root cause of the inconsistency? Suppose $f \equiv \frac{1}{2}$, so the data result from coin tossing, and the covariates do not matter. Thus, theory 0 is the truth. The statistician does not know this, however, and high-order theories may be deceptively attractive because they have many parameters.

However, the 'curse of dimensionality' only strikes under some circumstances. When? To make this a little clearer, consider a design of order n, so there are 2^n subjects. According

to theory n, the response of each subject is determined by the toss of a coin, where the probability is uniform on [0, 1]. Now one toss of a coin with a uniformly distributed random p is just like one toss of a fair coin – you get heads with probability $\frac{1}{2}$ and tails with probability $\frac{1}{2}$. Thus, theory n competes with theory 0. Indeed, the predictive probability of the data under theory n is

$$\pi_n\{\text{data}\} = \frac{1}{2^{2^n}}.$$

Let S be the sum of the response variables – the total number of heads. Under theory 0, the predictive probability of the data is

$$\pi_0\{\text{data}\} = \left[(2^n + 1)\binom{2^n}{S}\right]^{-1} \approx \frac{\sqrt{\pi/2}}{2^{n/2}}\pi_n\{\text{data}\} \tag{1.4}$$

because $S \approx 2^n/2$. Thus,

$$\pi_n\{\text{data}\} \approx \text{const.}\, 2^{n/2}\pi_0\{\text{data}\}. \tag{1.5}$$

The prior π is a mixture $\sum_{k=0}^{\infty} w_k\pi_k / \sum_{k=0}^{\infty} w_k$. The posterior is a similar mixture, the posterior weight on theory k being w_k times the predictive probability of the data under π_k. If $f \equiv \frac{1}{2}$, then, it is the theory weights w_k that decide consistency. If w_k declines rapidly, for example, $w_k = 1/2^k$, the weight on theory n compensates for the factor $2^{n/2}$ in (1.5); and the prior is consistent at $f \equiv \frac{1}{2}$. On the other hand, if w_k declines slowly, for example, $w_k = 1/(k+1)^2$, the factor $2^{n/2}$ dominates, and inconsistency is the result.

The heart of the problem seems to be that a mixture of Bernoulli variables is again Bernoulli. Our heuristic, then, is that consistency obtains when mixing leads outside the basic parametric family. For example, suppose the response variable takes three values, 0, 1 and 2; and, given the covariates ξ, the response is distributed as the number of heads when an $f(\xi)$ coin is tossed twice. A mixture of Bin$(2, p)$ variables cannot be Bin$(2, p)$; the heuristic suggests that Bayes estimates will be consistent.

To prove this kind of theorem in any degree of generality, we would need to impose smoothness conditions like those which underlie the usual asymptotics of maximum likelihood estimates, including the Bernstein–von Mises theorem. We would also need integrability conditions of the kind which underlie the usual theory of entropy bounds. The second set of conditions would enable us to localize the problem, and the first set would enable us to make local estimates. Rather than pursue such technical issues here, we prove a theorem for normal response variables – which is difficult enough. Consistency obtains, according to our heuristic, because a mixture of $N(\mu, 1)$ variates cannot be $N(\mu, 1)$. The theorem is stated in Section 2, and proved in later sections. A second theorem shows that Bayesian regression gets the order of the model right – if the model is of finite order. Inconsistencies arising from flat priors are also discussed, and an extended example is given.

1.3. Literature review on Bayesian regression

Roughly speaking, one observes

$$Y_i = f(t_i) + \varepsilon_i, \tag{1.6}$$

with f in some class of functions, t_i in an interval (say), and ε_i independent and identically distributed (i.i.d.) errors. A prior is assumed for f, a posterior is computed, and the posterior mean is used to estimate f. Typically, the t_i are taken as deterministic; our t_i are random.

The earliest reference we know is Poincaré (1896). He used a Gaussian prior of the form $f(t) \sim \sum_i X_i t^i$ for t in $[-1, 1]$, the X_i being independent Gaussian variables with mean 0 and variances tending to 0. He assumed (1.6) with $\varepsilon_i = 0$. Invoking the 'method of causes' – the classical phrase for Bayes' theorem – he computed the posterior mean of $f(t)$ given $f(t_i)$ for $i = 1, \ldots, n$; his t_i were deterministic. Poincare's beautiful calculations are equivalent to what we now call the theory of 'reproducing kernel Hilbert spaces'.

The subsequent history of Bayesian regression is traced in Diaconis (1988) and Traub *et al*. (1988). There is closely related work on sieves and on model selection; see Geman and Hwang (1982), Shibata (1981; 1986) or Stone (1982). Hierarchical priors for regression in finite-dimensional settings go back to Lindley and Smith (1972).

The simplest possible regression model has a constant mean function. That is the location problem: $Y_i = \mu + \varepsilon_i$, where μ is an unknown constant and the errors ε_i are i.i.d. Diaconis and Freedman (1988) studied nonparametric priors on μ and on the law of the errors; see also Doss (1984; 1985a; 1985b). Some natural priors lead to inconsistent estimates, while other priors give consistent results.

Nonparametric Bayesian regression also connects with the theory of splines (Kimeldorf and Wahba 1970; Kohn and Ansley 1987); for a recent survey, see Wahba (1990). Cox (1993) has an elegant mathematical treatment. He begins with the model (1.6) on [0, 1], say, where f is confined by assumption to a given smoothness class (that is, a Sobolev space). He specifies a Gaussian prior by the Karhunen–Loeve representation,

$$f(t) \sim \sum_i a_i X_i g_i(t), \qquad t \in [0, 1].$$

The X_i are i.i.d. $N(0, 1)$, $\sum_i a_i^2 < \infty$, and the g_i are an orthonormal basis in a suitable Hilbert space – a set-up rather similar to Poincaré's.

Cox computes the posterior for f given $f(t_i) + \varepsilon_i$ for $i = 1, \ldots, n$, the ε_i being i.i.d. $N(0, 1)$. He shows that in this infinite-dimensional setting, the Bernstein–von Mises theorem does not apply: the posterior distribution (centred at the mean) may be radically different from the frequentist distribution of the Bayes estimates (centred at truth). His set-up differs from ours in several ways (recall that f is the true mean function in the sampling model that governs the data). First, his f is L_2 and smooth; our f is only L_2. Second, his prior is different; indeed, it is probably orthogonal to ours. Third, his t_is are deterministic and equally spaced, rather than random. That all said, an interesting heuristic connection between his results and ours can be made via wavelet theory – as pointed out by a very helpful referee. That discussion continues in Section 11 below. There is a similar connection with the Gaussian white noise model, which is discussed in Brown and Low (1996) and Donoho (1994). Also see Diaconis and Freedman (1997).

1.4. Literature review on consistency of Bayes estimates

Frequentist properties of Bayes rules have been studied since Laplace (1774), who showed that in smooth, finite-dimensional problems, the posterior concentrates in a neighbourhood of

the maximum likelihood estimates. Modern versions of the result can be found in Bernstein (1934), von Mises (1964), Johnson (1967; 1970), LeCam (1982) or Ghosh *et al*. (1982). These results hold for almost all data sequences. In very simple settings, we obtained bounds that hold for all sequences (Diaconis and Freedman 1990).

Freedman (1963) considered nonparametric Bayes procedures, with a counterexample: there is a prior supported on all of the parameter space, whose posterior converges almost surely to the wrong answer. That paper introduced the Dirichlet and tail-free priors, and showed them to be consistent. For reviews, see Ferguson (1974) or Diaconis and Freedman (1988). Also see Schwartz (1965).

LeCam (1953) proved a version of what has come to be known as the Bernstein–von Mises theorem; see also LeCam and Yang (1990). LeCam's theorems were almost sure results, with respect to the true underlying measure that had generated the data; and he proved convergence in total variation norm. Previous authors had demonstrated only convergence of distribution functions, in probability. Furthermore, LeCam seems to have been the first to condition on all the data, not just summary statistics (such as the mean). For more discussion, see Pollard *et al*. (1997).

Efforts are now under way to develop a unified theory for consistency of Bayes estimates in the infinite-dimensional case: see Bunke and Milhaud (1994), Ghosal *et al*. (1997), Shen (1996) and Barron *et al*. (1997). So far, the results are somewhat fragmentary; we do not think our examples are covered by such theories.

2. The formal set-up

The set-up is virtually identical to that for the binary case, except that the response variables are normal; details are repeated for ease of reference. The covariates ξ are a sequence of 0s and 1s, sequence space is given the product topology, and the parameter space Θ is the set of L_2 functions f from sequence space to $(-\infty, \infty)$. The L_2 topology is installed on Θ, relative to coin-tossing measure λ^∞ in sequence space C_∞. A basic neighbourhood of $f \in \Theta$ is the 'δ-ball'

$$\mathrm{N}(f, \delta) = \left\{ g \in L_2 : \int_{C_\infty} (g - f)^2 \, \mathrm{d}\lambda^\infty < \delta^2 \right\}. \tag{2.1}$$

Consider a prior π on Θ, with posterior $\tilde{\pi}_n$. Then π is consistent for f provided $\tilde{\pi}_n\{\mathrm{N}(f, \delta)\} \to 1$ almost surely, for all positive δ.

The prior π_k is supported on the class of functions f such that $f(x)$ depend only on the first k coordinates in $x = (x_1, x_2, \ldots)$. Thus, π_k is determined by specifying the joint distribution of the 2^k possible values for f. These are independent $N(0, 1)$ variables; we refer to π_k as 'standard normal'. We put prior weight w_k on k, so our prior is of the form

$$\pi = \sum_{k=0}^{\infty} w_k \pi_k \Big/ \sum_{k=0}^{\infty} w_k, \tag{2.2a}$$

where

$$w_k > 0 \text{ for infinitely many } k \text{ and } \sum_{k=0}^{\infty} w_k < \infty. \tag{2.2b}$$

(If $w_k > 0$ for finitely many k, then π would be a conventional hierarchical normal prior.)

Turn now to the data, which are 'balanced' in the sense of Diaconis and Freedman (1993). At stage n, there are 2^n subjects, indexed by t. Each has a response variable $Y(t)$ and a covariate sequence $\xi(t)$. The first n bits of the covariate sequences cover all possible patterns of length n; each pattern appears once and only once. The remaining bits from $n+1$ onwards are generated by coin tossing. Given the covariates, response variables are independent normals, with variance 1. The conditional mean response of subject t depends only on the covariate string for that subject, through the function f:

$$\text{Given the covariates, the response variables are independent across subjects, normally distributed, with common variance 1 and } \mathrm{E}\{Y(t)|\xi\} = f[\xi(t)]. \tag{2.3}$$

This completes the set-up. The main theorems can now be stated.

Theorem 2.1. *Suppose the design is balanced, and normal in the sense of (2.3). Suppose the prior π is hierarchical in the sense of (2.2), and the π_k are standard normal. Then π is consistent for all $f \in L_2$.*

Let Θ_k be the class of functions f which depend only on the first k bits of the argument x; these increase with k. Recall that $\tilde{\pi}_n$ is the posterior given the data at stage n.

Theorem 2.2. *Suppose the design is balanced, and normal in the sense of (2.3). Suppose the prior π is hierarchical, and the π_k are standard normal. If $f \in \Theta_k$ and $w_k > 0$, then $\tilde{\pi}_n\{\Theta_k\} \to 1$ almost surely as $n \to \infty$.*

Theorem 2.1 demonstrates consistency, while Theorem 2.2 says that the Bayesian gets the order of a finite model right, at least if there is positive prior mass on the right order. This is a bit surprising, because many selection algorithms estimate models that are too complex; for instance, see Breiman and Freedman (1983).

We turn now to improper priors; π_k is 'flat' if the joint distribution of $\{\theta_s: s \in C_k\}$ is Lebesgue measure on 2^k-dimensional Euclidean space. With flat priors, consistency will obtain only if the weights w_k decay at a very rapid rate, as in (2.4a); condition (2.4b), satisfied if $w_k = 1/k^2$ or $w_k = 1/2^k$ or $w_k = 1/k!$, ensures inconsistency:

$$\limsup_{n\to\infty} \frac{1}{2^n} \log \left[\sum_{k=n}^{\infty} w_k\right] < -\tfrac{1}{2}\log(2\pi e) \tag{2.4a}$$

$$\limsup_{n\to\infty} \frac{1}{2^n} \log \left[\sum_{k=n}^{\infty} w_k\right] > -\tfrac{1}{2}\log(2\pi e). \tag{2.4b}$$

Theorem 2.3. *Suppose the design is balanced, and normal in the sense of (2.3). Suppose the prior π is hierarchical, and the π_k are flat. Then π is consistent for all $f \in L_2$ if (2.4a) holds. If (2.4b) holds, π is inconsistent for all $f \in L_2$.*

Theorem 2.3 will be proved in Section 8 and Section 9 gives an example. Until then, the π_k will be normal. Flat priors present definitional problems, to be discussed in Section 10.

3. Proofs: the preliminaries

We will compute the predictive probability density of the data, under theory k; then the posterior. Where possible, we follow the notation and arguments in Diaconis and Freedman (1993). To review briefly, let C_k be the set of strings of 0s and 1s of length k. There are 2^k strings $s \in C_k$. Let C_∞ be the set of infinite sequences of 0s and 1s, in the product topology and product σ-field. Let λ^∞ be coin-tossing measure on C_∞. The parameter space Θ consists of all L_2 functions from C_∞ to $(-\infty, \infty)$; functions that are equal almost everywhere are identified. We endow Θ with the L_2 metric and the Borel σ-field generated by the balls (2.1). Of course, Θ is complete separable metric. As previously defined,

$$\Theta_k \text{ is the closed set consisting of all } f \in \Theta \text{ such that } f(x) \text{ depends only on the first } k \text{ coordinates of } x \in C_\infty. \tag{3.1a}$$

If $f \in \Theta_k$, then

$$f(x_1, \ldots x_k, x_{k+1}, x_{k+2}, \ldots) = \theta_s(f), \text{ where } s = (x_1, \ldots, x_k) \in C_k. \tag{3.1b}$$

The probability π_k on Θ concentrates on Θ_k and makes θ_s independent and $N(0, 1)$ as s varies over C_k.

All random variables are defined on some probability triple $(\Omega, \mathscr{F}, P_f)$, where $f \in \Theta$. At stage n, we have 2^n independent subjects indexed by $t \in C_n$, with response variables $Y(t)$ and covariate strings $\xi(t)$, forming a balanced design of order n. In particular, (2.3) holds. Furthermore, $\xi_i(t) = t_i$ for $1 \leqslant i \leqslant n$; for $i > n$ the variables $\xi_i(t)$ are independent, each being 0 or 1 with probability $\frac{1}{2}$: these are the 'balance' conditions. Here, $\xi_i(t)$ is the ith bit in the covariate sequence for subject t. To ease the notation, we sometimes write Y_t for $Y(t)$ or ξ_t for $\xi(t)$.

As usual, π_k can be extended to a probability on $\Theta \times \Omega$, by the formula

$$\pi_k(A \times B) = \int_A P_f\{B\}\pi_k\{\mathrm{d}f\}. \tag{3.2}$$

In this formula, A is a measurable subset of Θ and B is a measurable subset of Ω; $f \to P_f\{B\}$ is measurable because

$$f \to \int \prod_{t \in C_n} g_t(Y_t)\,\mathrm{d}P_f \tag{3.3}$$

is continuous for bounded continuous functions g_t.

Fix k and n. The response variables Y_t: $t \in C_n$ have a joint probability density – the *predictive probability density* – with respect to π_k. This density will be denoted ρ_{kn}, and viewed as a function of 2^n real variables Y_t: $t \in C_n$.

Lemma 3.1. *For a balanced normal design of order n and the standard normal prior π_k, the predictive probability density ρ_{kn} may be computed as follows:*

(a) *If $k \leqslant n$, then* $\log \rho_{kn} = a_n - b_{kn} - c_{kn} + q_{kn}$.
(b) *If $k > n$, then* $\log \rho_{kn} = log\, \rho_{nn}$.

In these formulae,

$$a_n = \tfrac{1}{2} 2^n \log(1/2\pi) - \tfrac{1}{2} \sum_{t \in C_n} y_t^2; \tag{3.4a}$$

$$b_{kn} = 2^k (n-k) \tfrac{1}{2} \log 2; \tag{3.4b}$$

$$c_{kn} = 2^k \tfrac{1}{2} \log\left(1 + \frac{1}{2^{n-k}}\right); \tag{3.4c}$$

$$q_{kn} = \tfrac{1}{2} d_{kn} 2^n \frac{1}{2^k} \sum_{s \in C_k} \bar{y}_s^2; \tag{3.4d}$$

$$d_{kn} = \frac{2^{n-k}}{2^{n-k}+1}; \tag{3.4e}$$

$$\bar{y}_s = \frac{1}{2^{n-k}} \sum_{t \in C_n} \{y_t\text{: } t \text{ extends } s\} \text{ for } s \in C_k. \tag{3.4f}$$

Proof. (a). Fix $k \leqslant n$. For $s \in C_k$, let

$$V_s = \{(Y_t, \xi_t)\text{: } t \in C_n \text{ and } t \text{ extends } s\}.$$

Each V_s is a 2^{n-k}-tuple of pairs of random variables. Recall θ_s from (3.1). Recall that π_k was extended to $\Theta \times \Omega$ by (3.2).

$$\text{Relative to } \pi_k, \text{ as } s \text{ ranges over } C_k, \text{ the pairs } (V_s, \theta_s) \text{ are i.i.d.} \tag{3.5}$$

Consequently, the general case in claim (a) follows from the case $k = 0$. The latter is a routine calculation. Abbreviate $m = 2^n$. Let ϕ_v be the normal density with mean 0 and variance v. Let $\exp(x) = e^x$, $\alpha = (1/2\pi)^{m/2}$, $\beta = \exp\{-\frac{1}{2}\sum_{t \in C_n}(y_t - \bar{y})^2\}$, $\bar{y} = 1/m \sum_{t \in C_n} y_t$. Write $*$ for convolution. Then

$$\rho_{0n} = \alpha \int \exp\left\{ -\tfrac{1}{2} \sum_{t \in C_n} (y_t - \theta)^2 \right\} \phi_1(\theta)\, d\theta$$

$$= \alpha\beta \int \exp\{-\tfrac{1}{2} m(\overline{y} - \theta)^2\} \phi_1(\theta)\, d\theta$$

$$= \alpha\beta \left(\frac{2\pi}{m}\right)^{1/2} (\phi_{1/m} * \phi_1)(\overline{y})$$

$$= \alpha\beta \left(\frac{2\pi}{m}\right)^{1/2} \phi_{1+1/m}(\overline{y})$$

$$= \alpha\beta \left(\frac{1}{m} \frac{1}{1 + \frac{1}{m}}\right)^{1/2} \exp\left\{ -\tfrac{1}{2}\overline{y}^2 \Big/ \left(1 + \frac{1}{m}\right) \right\}.$$

To verify the formula for q_{0n}, combine the last equation with β:

$$\sum_{t \in C_n} (y_t - \overline{y})^2 = \sum_{t \in C_n} y_t^2 - m\overline{y}^2$$

and

$$m - \frac{1}{1 + \frac{1}{m}} = m \frac{m}{m+1}.$$

This completes the proof of claim (a), and (b) is routine. □

Let $\tilde{\pi}_{kn}$ be the posterior distribution of f, computed relative to π_k, given the data from a balanced normal design of order n. Lemma 3.2 computes this posterior for $k \leqslant n$; and Lemma 3.3 does the job for $k > n$. Clearly, $\tilde{\pi}_{kn}$ concentrates on Θ_k, as defined in (3.1).

As a notational principle, the functions defined in (3.4) will be denoted by capital letters, when evaluated at $\{Y_t\}$ rather than $\{y_t\}$. The following definitions will be used throughout.

Definition 3.1. *For $k \leqslant n$ and $s \in C_k$, let $\overline{Y}_s$ be the average of Y_t over t such that $t \in C_n$ is an extension of s. For $x \in C_\infty$ and $\omega \in \Omega$, let $\overline{Y}_{kn}(x, \omega) = \overline{Y}_s(\omega)$, where $s \in C_k$ gives the first k bits of x. And let $\overline{Y}(\omega) = (1/2^n) \sum_{t \in C_n} Y_t(\omega)$.*

In other terms, $\overline{Y}_{kn}(x)$ is obtained by averaging the Y_t such that $x_i = t_i$ for $1 \leqslant i \leqslant k$. This function depends, of course, on ω; however, for each $s \in C_k$, $x \to \overline{Y}_{kn}(x, \omega)$ is constant on

$$\langle s \rangle = \{x \in C_\infty : x_i = s_i \text{ for } 1 \leqslant i \leqslant k\}.$$

Recall θ_s from (3.1) and d_{kn} from (3.4).

Lemma 3.2. *Suppose $k \leqslant n$ and π_k is standard normal. According to the posterior $\tilde{\pi}_{kn}$, given data from a balanced normal design of order n, the parameters θ_s are conditionally independent as s ranges over C_k, and θ_s is normal:*

$$E\{\theta_s | Y_t : t \in C_n\} = d_{kn}\overline{Y}_s$$

$$\text{var}\{\theta_s | Y_t : t \in C_n\} = 1 - d_{kn} = \frac{2^k}{2^n + 2^k}.$$

Proof. In view of (3.5), only the case $k = 0$ needs to be argued, and this is routine. Relative to π_0, there is only one parameter, θ; this is $N(0, 1)$. Given θ, the Y_t are independent $N(\theta, 1)$. Abbreviate $m = 2^n$, and $\overline{Y}$ for the mean of the Ys. Unconditionally, $\overline{Y}$ is $N(0, 1 + [1/m])$. Furthermore, $\text{cov}(\theta, \overline{Y}) = 1$ and $r^2(\theta, \overline{Y}) = m/(m+1)$. The balance of the argument is omitted. □

Remark. If $k \leqslant n$, $\{Y_t : t \in C_n\}$ and $\{\xi_t : t \in C_n\}$ are independent relative to π_k.

Recall that subjects are indexed by $t \in C_n$, and subject t has covariate string $\xi(t)$, with $\xi_1(t) = t_1, \ldots, \xi_n(t) = t_n$. If $k > n$, there are 2^k parameters, but only 2^n observations: some parameters are 'observed', others are not. More formally, $s \in C_k$ is observed if there is a $t = t_s \in C_n$ with $\xi_i(t) = s_i$ for $1 \leqslant i \leqslant k$. The set S of observed s is random, for S depends on the covariates. If $k > n$, $\{Y_t : t \in C_n\}$ and $\{\xi_t : t \in C_n\}$ are conditionally independent relative to π_k, given the set S of observed indices s, and the covariates $\xi_{k+1}(t_s), \ldots, \xi_n(t_s)$ for $s \in S$.

Lemma 3.3. *Suppose $k > n$ and π_k is standard normal. According to the posterior $\tilde{\pi}_{kn}$, given data from a balanced normal design of order n, the parameters θ_s are independent as s ranges over C_k, and θ_s is normal. If s is unobserved, θ_s is conditionally $N(0, 1)$. If s is observed,*

$$E\{\theta_s | Y_t : t \in C_n\} = \tfrac{1}{2}\overline{Y}_s$$

$$\text{var}\{\theta_s | Y_t : t \in C_n\} = \tfrac{1}{2}.$$

Turn now to the posterior $\tilde{\pi}_n$, computed relative to the hierarchical prior π defined in (2.2). The 'theory index' k in (2.2) is a parameter which has a posterior distribution relative to π. This will now be computed. Let

$$\tilde{w}_{kn} = w_k R_{kn}; \tag{3.6}$$

following our general notational principles, R_{kn} is the predictive density ρ_{kn} evaluated at the data $\{Y_t\}$; see (3.4). The posterior probability of theory k is

$$\tilde{w}_{kn} \Big/ \sum_{k=0}^{\infty} \tilde{w}_{kn}. \tag{3.7}$$

Then $\tilde{\pi}_n$ is a mixture of the posteriors $\tilde{\pi}_{kn}$, with weights $\tilde{w}_{kn}$; the latter will be called *posterior theory weights*. These (slightly informal) arguments prove the following:

Lemma 3.4. *Suppose π is a hierarchical prior, and the π_k are standard normal. Given the data from a balanced normal design of order n, the posterior is*

$$\tilde{\pi}_n = \sum_{k=0}^{\infty} \tilde{w}_{kn}\tilde{\pi}_{kn} \Big/ \sum_{k=0}^{\infty} \tilde{w}_{kn}.$$

For $k \leqslant n$, the posteriors $\tilde{\pi}_{kn}$ were computed in Lemma 3.2; for $k > n$, these posteriors were computed in Lemma 3.3.

4. Estimating the posteriors $\tilde{\pi}_{kn}$

The idea of the proof is simple, although details are quite tedious. We estimate the predictive probabilities R_{kn}, and show that the posterior concentrates on ks which are considerably smaller than n. In that range, the posteriors $\tilde{\pi}_{kn}$ concentrate near their mean functions.

We turn now to rigour. Recall (3.4) and Definition 3.1. In particular, for $k \leqslant n$ and $s \in C_k$, $\overline{Y}_s$ is the average of Y_t over t such that t is an extension of s. And $\overline{Y}_{kn}$ has domain $C_\infty \times \Omega$.

Lemma 4.1. *Fix $k \leqslant n$. We have data from a balanced normal design of order n, and a standard normal prior π_k. For all $\omega \in \Omega$,*

$$\int_\Theta \int_{C_\infty} [g(x) - d_{kn}\overline{Y}_{kn}(x)]^2 \lambda^\infty(\mathrm{d}x)\tilde{\pi}_{kn}(\mathrm{d}g) = \frac{2^k}{2^n + 2^k}.$$

Proof. The posterior $\tilde{\pi}_{kn}$ concentrates on Θ_k, as defined in (3.1). Then use Lemma 3.2. □

Definition 4.1. *Let g be an L_2 function on C_∞. Then $\overline{g}_s$ is the average of $g(\xi_t)$ over $t \in C_n$ that extend $s \in C_k$. The domain may change to C_∞ as follows: $\overline{g}_{kn}(x) = \overline{g}_s$ when $x \in C_\infty$ extends s. These functions are random, because they depend on the covariates. To emphasize that dependence, we may write $\overline{g}_{kn}(x, \omega)$. Let $\overline{\zeta}_s$ be the average over t that extend s of $\zeta_t = Y_t - f(\xi_t)$.*

Lemma 4.1 showed that $\tilde{\pi}_{kn}$ concentrates near its mean function $d_{kn}\overline{Y}_{kn}$. Next, we show that $\overline{Y}_{kn}$ can be well approximated by $\overline{f}_{kn}$, as in Definition 4.1.

For $t \in C_n$, let a_t be real. As is easily verified.

$$k \to \frac{1}{2^k} \sum_{s \in C_k} \left(\frac{1}{2^{n-k}} \sum \{a_t\colon t \text{ extends } s\} \right)^2 \text{ is monotone non-decreasing in } k, \text{ for } 0 \leqslant k \leqslant n. \tag{4.1}$$

The following are immediate; calculations are relative to P_f.

As t ranges over C_n, the pairs (Y_t, ξ_t) are independent. (4.2a)

$$\zeta_t = Y_t - f(\xi_t) \text{ is } N(0, 1). \tag{4.2b}$$

$$\{\zeta_t\colon t \in C_n\} \text{ is independent of } \{\xi_t\colon t \in C_n\}. \tag{4.2c}$$

Corollary 4.1. *Fix $k \leqslant n$. With a balanced normal design of order n,*

$$\int_\Omega \int_{C_\infty} (\overline{Y}_{kn}(x,\omega) - \overline{f}_{kn}(x,\omega))^2 \lambda^\infty(\mathrm{d}x) P_f(\mathrm{d}\omega) = \frac{1}{2^{n-k}}.$$

Proof. Clearly, $\overline{Y}_{kn}(x) - \overline{f}_{kn}(x) = \overline{\zeta}_s$, where s is the first k bits of x. Then

$$\int_{C_\infty} (\overline{Y}_{kn}(x,\,\omega) - \overline{f}_{kn}(x,\,\omega))^2 \lambda^\infty(\mathrm{d}x) = \frac{1}{2^k} \sum_{s \in C_k} \overline{\zeta}_s^2. \tag{4.3}$$

Now use (4.2). □

Proposition 4.1. *Fix $A > 2$ and $\delta > 0$. For P_f-almost all ω, for all sufficiently large n, for all $k < n - A \log n$,*

$$\int_{C_\infty} [\overline{Y}_{kn}(x,\omega) - \overline{f}_{kn}(x,\omega)]^2 \lambda^\infty(\mathrm{d}x) < \delta^2.$$

Proof. Use (4.1) with $a_t = Y_t - f(\xi_t)$, to see that only the maximal k in the given range needs to be considered. For that k, use Chebyshev's inequality with Corollary 4.1 to estimate the variance; the Borel–Cantelli lemma completes the proof. □

Proposition 4.1 shows that $\overline{Y}_{kn}(\cdot,\,\omega) - \overline{f}_{kn}(\cdot,\,\omega) \to 0$ in L_2 as $n \to \infty$, uniformly in $k < n - A \log n$, for almost all ω. We need a similar but weaker estimate for $k \leqslant n - B$, as given in Proposition 4.2. Indeed, convergence to 0 cannot be obtained: when $k = n - B$, there are only 2^B terms in each average $\overline{Y}_s(\omega)$: $s \in C_k$. Lemma 4.2 is nearly standard; only the case $d = 1$ needs to be verified, and that follows by considering the Laplace transform.

Lemma 4.2. *Let m and d be positive integers. For $i = 1, \ldots, m$, let X_i^2 be independent χ^2 variables, with d degrees of freedom. Fix $\varepsilon > 0$. There is a $\rho = \rho(\varepsilon) < 1$ such that*

$$P\left\{\left|\sum_{i=1}^m (X_i^2 - d)\right| > md\varepsilon\right\} < \rho^{md}.$$

Proposition 4.2. *Fix $\delta > 0$. There is a $B = B(\delta) < \infty$ so large that for P_f-almost all ω, for all sufficiently large n, for all $k \leqslant n - B$,*

$$\Delta_{kn}(\omega) = \int_{C_\infty} [\overline{Y}_{kn}(x,\omega) - \overline{f}_{kn}(x,\omega)]^2 \lambda^\infty(\mathrm{d}x) < \delta^2.$$

Proof. Again, we need only prove this for $k = n - B$; see (4.1). With that choice of k, Δ_{kn} is distributed as

$$\left(\sum_{i=1}^{2^{n-B}} X_i^2\right) \Big/ 2^n,$$

the X_i^2 being independent χ^2 variables with 1 degree of freedom; see (4.2) and (4.3). Fix $\varepsilon > 0$; fix B so large that

$$2^{-B}(1+\varepsilon) < \delta^2.$$

Then use Lemma 4.2: for P_f-almost all ω, for all sufficiently large n,

$$\Delta_{kn}(\omega) < 2^{n-B}(1+\varepsilon)/2^n = 2^{-B}(1+\varepsilon) < \delta^2. \qquad \square$$

Let g be an L_2 function on C_∞. By definition,

$$g_k(s) = \int_{C_\infty} g(sw)\lambda^\infty(\mathrm{d}w) \quad \text{and} \quad g_k(x) = g_k(x_1, \ldots, x_k). \tag{4.4}$$

These are deterministic functions on C_k and C_∞, respectively; $g_k(x)$ is well defined for all x, even though g may only be defined almost everywhere. By the usual martingale theorems,

$$g_k \to g \text{ almost everywhere and in } L_2 \text{ as } k \to \infty, \text{ relative to } \lambda^\infty. \tag{4.5a}$$

$$\text{If } j < k, \text{ then } \int g_j^2 < \int g_k^2 \text{ unless } g_j = g_k. \tag{4.5b}$$

The next step is to show that $\overline{f}_{kn}$ can be well approximated by f_k. We begin with a version of the strong law. Recall that P_f makes the ξ_t independent as t ranges over C_n. Furthermore, if $t \in C_n$, then $\xi_i(t) = t_i$ for $1 \leqslant i \leqslant n$; for $i > n$ the $\xi_i(t)$ are independent, each taking the values 0 or 1 with probability $\frac{1}{2}$. The ξ_t are independent but not identically distributed. Indeed,

$$\begin{aligned} &\text{The } P_f\text{-law of } \xi_t \text{ is just the } \lambda^\infty\text{-law of } x, \text{ given that} \\ &\quad x \in \langle t \rangle = \{x \in C_\infty | x_i = t_i \text{ for } 1 \leqslant i \leqslant n\}. \end{aligned} \tag{4.6}$$

Theorem 4.1. *Suppose h is a measurable function on C_∞. If h is L_1 with respect to λ^∞, then*

$$\frac{1}{2^n} \sum_{t \in C_n} h(\xi_t) \to \int_{C_\infty} h(x)\lambda^\infty(\mathrm{d}x)$$

as $n \to \infty$, with P_f-probability 1.

Proof. This is proved by a standard truncation argument, as in Feller (1968, p. 247). In more detail, let $h' = h$ provided $|h| \leqslant 2^n$, else let $h' = 0$; the dependence on n is not shown. We claim that

$$P_f\text{-almost surely, for all sufficiently large } n, \text{ for all } t \in C_n,\ h'(\xi_t) = h(\xi_t). \tag{4.7}$$

Indeed, the P_f-probability of the complementary event is at most

$$\sum_{t\in C_n} P_f\{|h(\xi_t)|<2^n\} = \sum_{t\in C_n} \lambda^\infty\{|h(x)|>2^n | x_i = t_i \text{ for } 1\leqslant i\leqslant n\}$$

$$= 2^n\lambda^\infty\{|h(x)|>2^n\},$$

where the first equality holds by (4.6). Now

$$\sum_{n=1}^\infty 2^n\lambda^\infty\{|h(x)|>2^n\} = \sum_{n=1}^\infty 2^n \sum_{m=n}^\infty \lambda^\infty\{2^m<|h(x)|\leqslant 2^{m+1}\}$$

$$= \sum_{m=1}^\infty \sum_{n=1}^m 2^n\lambda^\infty\{2^m<|h(x)|\leqslant 2^{m+1}\}$$

$$< 2\sum_{m=1}^\infty 2^m\lambda^\infty\{2^m<|h(x)|\leqslant 2^{m+1}\}$$

$$< 2\int_{C_\infty} |h|\,\mathrm{d}\lambda^\infty < \infty.$$

The Borel–Cantelli lemma completes the proof of (4.7).

For $t\in C_n$, let $m_t = 2^n\int_{\langle t\rangle} h'\,\mathrm{d}\lambda^\infty$, where $\langle t\rangle$ is the set of $x\in C_\infty$ with $x_i = t_i$ for $1\leqslant i\leqslant n$. By (4.6), $m_t = \int_\Omega h'(\xi_t)\,\mathrm{d}P_f$. By dominated convergence,

$$\frac{1}{2^n}\sum_{t\in C_n} m_t = \int_{C_\infty} h'\,\mathrm{d}\lambda^\infty \to \int_{C_\infty} h\,\mathrm{d}\lambda^\infty \qquad \text{as } n\to\infty. \tag{4.8}$$

We claim

$$\frac{1}{2^n}\sum_{t\in C_n}[h'(\xi_t) - m_t] \to 0 \text{ as } n\to\infty,\ P_f\text{-almost surely.} \tag{4.9}$$

To prove (4.9), fix $\varepsilon>0$ and use (4.6):

$$P_f\left\{\left|\frac{1}{2^n}\sum_{t\in C_n}[h'(\xi_t) - m_t]\right| > \varepsilon\right\} \leqslant \frac{1}{\varepsilon^2}\frac{1}{4^n}\sum_{t\in C_n} \operatorname{var}\{h'(\xi_t)\}$$

$$\leqslant \frac{1}{\varepsilon^2}\frac{1}{4^n}\sum_{t\in C_n}\int_\Omega h'(\xi_t)^2\,\mathrm{d}P_f$$

$$\leqslant \frac{1}{\varepsilon^2}\frac{1}{4^n}\sum_{t\in C_n} 2^n\int_{\langle t\rangle} h'(x)^2\lambda^\infty(\mathrm{d}x)$$

$$= \frac{1}{\varepsilon^2}\frac{1}{2^n}\sum_{t\in C_n}\int_{\{|h(x)|\leqslant 2^n\}} h(x)^2\lambda^\infty(\mathrm{d}x)$$

Furthermore,

$$\sum_{n=1}^{\infty} \frac{1}{2^n} \sum_{t \in C_n} \int_{\{|h(x)| \leqslant 2^n\}} h(x)^2 \lambda^{\infty}(\mathrm{d}x) = A + B,$$

where

$$A = \sum_{n=1}^{\infty} \frac{1}{2^n} \int_{\{|h(x)| \leqslant 1\}} h(x)^2 \lambda^{\infty}(\mathrm{d}x) < \infty$$

and

$$B = \sum_{n=1}^{\infty} \frac{1}{2^n} \sum_{m=1}^{n} \int_{\{2^{m-1} < |h(x)| \leqslant 2^m\}} h(x)^2 \lambda^{\infty}(\mathrm{d}x).$$

We now estimate B, as follows:

$$B = \sum_{m=1}^{\infty} \sum_{n=m}^{\infty} \frac{1}{2^n} \int_{\{2^{m-1} < |h(x)| \leqslant 2^m\}} h(x)^2 \lambda^{\infty}(\mathrm{d}x)$$

$$< 2 \sum_{m=1}^{\infty} \frac{1}{2^m} \int_{\{2^{m-1} < |h(x)| \leqslant 2^m\}} h(x)^2 \lambda^{\infty}(\mathrm{d}x)$$

$$< 2 \int |h(x)| \lambda^{\infty}(\mathrm{d}x) < \infty.$$

The Borel–Cantelli lemma completes the proof of (4.9). Relations (4.7)–(4.9) prove the theorem. □

Remark. Let $W_n = \{\xi_t\colon\ t \in C_n\}$, a set of 2^n random variables. The joint distribution of W_n, as n varies, does not matter in Theorem 4.1.

We return to the idea of approximating $\overline{f}_{kn}$ by f_k; on the former, see Definition 4.1; the latter is defined in (4.4).

Proposition 4.3. *For P_f-almost all ω, as $n \to \infty$,*

$$\max_{0 \leqslant k \leqslant n} \int_{C_\infty} [\overline{f}_{kn}(x, \omega) - f_k(x)]^2 \lambda^{\infty}(\mathrm{d}x) \to 0.$$

Proof. Use (4.1) with $a_t = Y_t - f_n(t)$, to see the maximum is attained for $k = n$. Write $\|\cdot\|$ for the L_2 norm relative to λ^{∞}. Fix $\delta > 0$. Using (4.5), choose j so large that $\|f - g\| < \delta$, where $g = f_j$ depends only on the first j bits of x.

We claim that

$$\overline{g}_{nn}(x, \omega) = g_n(x) \text{ for all } x \in C_\infty \text{ and all } \omega \in \Omega, \text{ provided } n > j. \tag{4.10}$$

The only difficulty here is the notation. Fix x. Let $t = (x_1, \ldots, x_n) \in C_n$. The left-hand side of (4.10) is $g(\xi_t(\omega))$, by Definition 4.1. By the balance conditions, $\xi_t(\omega) = x_1 \ldots x_n e_1 e_2 \ldots$,

where $e_i = 0$ or 1 depending on ω. However, $g(x) = f_j(x)$ only depends on the first j bits of x, by (4.4). So, the left-hand side of (4.10) boils down to $f_j(x_1, \ldots, x_j)$. For future reference,

$$\overline{g}_{nn}(x, \omega) = f_j(x) \text{ for all } x \in C_\infty \text{ and all } \omega \in \Omega, \text{ provided } n > j. \tag{4.11}$$

The right-hand side of (4.10) is $E_\lambda\{g|x_1, \ldots, x_n\}$, the expectation being taken relative to λ^∞, by definition (4.4). However, $g = f_j$ only depends on $x_1, \ldots, x_j$. So, the right-hand side of (4.10) is also $f_j(x_1, \ldots, x_j)$. This completes the proof of (4.10).

For all ω and all $n > j$,

$$\begin{aligned}\|\overline{f}_{nn}(\cdot, \omega) - f_n(\cdot)\| \leqslant \|\overline{f}_{nn}(\cdot, \omega) - \overline{g}_{nn}(\cdot, \omega)\| \\ + \|\overline{g}_{nn}(\cdot, \omega) - g_n(\cdot)\| + \|g_n(\cdot) - f_n(\cdot)\|.\end{aligned} \tag{4.12}$$

The middle term on the right-hand side of (4.12) vanishes, by (4.10). The last term may be recognized as $\|f_j(\cdot) - f_n(\cdot)\|$, whose limit as $n \to \infty$ is less than δ, by construction. The square of the first term, by Definition 4.1 and (4.11), is

$$\frac{1}{2^n} \sum_{t \in C_n} [f(\xi_t) - f_j(\xi_t)]^2,$$

whose P_f-almost sure limit as $n \to \infty$ is, by Theorem 4.1,

$$\int_{C_\infty} [f(x) - f_j(x)]^2 \lambda^\infty(\mathrm{d}x) < \delta^2,$$

again by construction. In short, for P_f-almost all ω,

$$\limsup_{n \to \infty} \|\overline{f}_{nn}(\cdot, \omega) - f_n(\cdot)\| < 2\delta. \tag{4.13}$$ □

Corollary 4.2. *Fix $\delta > 0$. There is a $B = B(\delta) < \infty$ so large that for P_f-almost all ω, for sufficiently large n, for all $k \leqslant n - B$,*

(a) $\int_{C_\infty} [\overline{Y}_{kn}(x, \omega) - f_k(x)]^2 \lambda^\infty(\mathrm{d}x) < \delta^2$,
(b) $\int_{C_\infty} [d_{kn} \overline{Y}_{kn}(x, \omega) - f_k(x)]^2 \lambda^\infty(\mathrm{d}x) < \delta^2$.

Proof. Claim (a) is immediate from Propositions 4.2 and 4.3. Then (b) follows. Indeed, d_{kn} is uniformly close to 1 by definition (3.4); and $\int f_k^2 \leqslant \int f^2$ by (4.4). □

Corollary 4.3. *Fix $\delta > 0$. There is a $B = B(\delta) < \infty$ so large that for P_f-almost all ω, for all sufficiently large n, for all $k \leqslant n - B$, $\tilde{\pi}_{kn}\{\mathrm{N}(f_k, \delta)\} > 1 - \delta$; the δ-ball* N *was defined in (2.1).*

This is immediate from Lemma 4.1 and Corollary 4.2b, by the triangle inequality. Corollary 4.3 completes our discussion of the posteriors $\tilde{\pi}_{kn}$, and we turn to the posterior theory weights $\tilde{w}_{kn}$.

5. Estimating the theory weights $\tilde{w}_{kn}$

As we will show, the posterior theory weights $\tilde{w}_{kn}$ tend to concentrate on theories k with $k < n - B$. The $\tilde{w}_{kn}$ are computed from the predictive probability densities R_{kn}; see (3.6). The R_{kn} in turn are driven by the quadratic Q_{kn}; see (3.4). According to our notation, R_{kn} is just ρ_{kn}, with $\{Y_t\}$ in place of $\{y_t\}$; likewise for Q_{kn} and q_{kn}. The first lemma is useful, if superficial.

Lemma 5.1. *Suppose g and h are L_2 functions. Then*

$$|\|g\|^2 - \|h\|^2| \leq \|g - h\| \times [2\|h\| + \|g - h\|].$$

Lemma 5.2. *Fix $\delta > 0$. There is a $B = B(\delta) < \infty$ so large that for P_f-almost all ω, for all sufficiently large n, for all $k \leq n - B$,*

(a) $|\|Y_{kn}(\cdot, \omega)\|^2 - \int f_k^2| < \delta$
(b) $|Q_{kn}(\omega) - \frac{1}{2} d_{kn} 2^n \int f_k^2| < \delta 2^n$.

Proof. Only claim (a) needs to be argued. By Lemma 5.1 and (4.5),

$$\|\overline{Y}_{kn}(\cdot, \omega)^2 - f_k^2\| \leq \|\overline{Y}_{kn}(\cdot, \omega) - f_k\| \times [2\|f\| + \|\overline{Y}_{kn}(\cdot, \omega) - f_k\|].$$

Finally, $\|\overline{Y}_{kn}(\cdot, \omega) - f_k\|$ is small, by Corollary 4.2. □

We must now consider theories with indices near n.

Lemma 5.3. *Fix $j = 0, 1, \ldots$. Let $k = n - j$. Let*

$$\Xi_{kn}(\omega) = \frac{1}{2^k} \sum_{s \in C_k} \overline{Y}_s(\omega)^2.$$

For P_f-almost all ω,

$$\Xi_{kn}(\omega) \to \int_{C_\infty} f(x)^2 \lambda^\infty(\mathrm{d}x) + \frac{1}{2^j} \qquad \text{as } n \to \infty.$$

Proof. Recall Definition 4.1. Then

$$\overline{Y}_s = \overline{f}_s + \overline{\xi}_s,$$

where the terms on the right are independent and $\overline{\xi}_s$ is $N(0, 1/2^j)$; see (4.2). Then Ξ_{kn} may be rewritten as

$$\int_{C_\infty} \overline{f}_{kn}(x, \cdot)^2 \lambda^\infty(\mathrm{d}x) + \frac{2}{2^k} \sum_{s \in C_k} \overline{f}_s \overline{\xi}_s + \frac{1}{2^k} \sum_{s \in C_k} \overline{\xi}_s^2. \tag{5.1}$$

In view of Lemma 5.1 and Proposition 4.3, the first term in (5.1) is $\int f_k(x)^2 + o(1)$,

almost surely. But $\int f_k^2 = \int f^2 + o(1)$, by (4.5). By Lemma 4.2, the last term in (5.1) converges almost surely to $1/2^j$.

Given the covariates, the middle term in (5.1) is normal with conditional mean 0 and conditional variance

$$\frac{4}{2^{j+k}}\frac{1}{2^k}\sum_{s\in C_k}\overline{f}_s^2 = \frac{4}{2^n}\frac{1}{2^k}\sum_{s\in C_k}\overline{f}_s^2 < \frac{4}{2^n}\frac{1}{2^n}\sum_{t\in C_n} f(\xi_t)^2;$$

the equality holds because $n = j + k$, and the inequality holds by (4.1). Consequently, the middle term in (5.1) has unconditional mean 0, and unconditional variance bounded above by $4\int f^2/2^n$; it tends to 0 almost surely, by Chebyshev and Borel–Cantelli. □

Lemma 5.4. *Let* $j = 1, 2, \ldots$.

(a) $j(2^j + 1)/2^j$ *increases with* j.
(b) $(j/2^j)\log 2 > 1/(2^j + 1)$.

Proof. (a). Fix j. We must show that

$$(j+1)\frac{2^{j+1}+1}{2^{j+1}} > j\frac{2^j+1}{2^j},$$

which boils down to $2^{j+1} + 1 > j$.

(b). By (a), the case $j = 1$ is critical; but $\log 2 > \frac{2}{3}$. □

Lemma 5.5. *Fix* $j = 0, 1, \ldots$. *Let* $k = n - j$. *For* P_f*-almost all* ω,

$$\limsup_{n\to\infty}\frac{2}{2^n}[Q_{kn}(\omega) - b_{kn} - c_{kn}] < \int_{C_\infty} f(x)^2\lambda^\infty(\mathrm{d}x).$$

Proof. The notation is defined in (3.4). To begin with, by Lemma 5.3,

$$\frac{2Q_{kn}}{2^n} = \frac{2^j}{2^j+1}\frac{1}{2^k}\sum_{s\in C_k}\overline{Y}_s(\omega)^2 \to \frac{2^j}{2^j+1}\left(\int f^2 + \frac{1}{2^j}\right).$$

The Case $j = 0$. Now $2Q_{kn}/2^n \to \frac{1}{2}\int f^2 + \frac{1}{4}$, $b_{kn} = 0$, and $2c_{kn}/2^n = \log 2$. But $\log 2 > \frac{1}{4}$.

The Case $j > 0$. Now $k = n - j$, $2b_{kn}/2^n = (j/2^j)\log 2$, $c_{kn} > 0$, and the result follows from Lemma 5.4(b). □

We are close to proving Theorem 2.1. The next lemma establishes that early theories become implausible, as the data come in. Recall that R_{kn} is the predictive density evaluated at $\{Y_t\}$.

Lemma 5.6. *Fix* j, k *with* $j < k$ *and* $f_j \neq f_k$. *Then* $R_{jn}/R_{kn} \to 0$ *as* $n \to \infty$, P_f*-almost surely.*

Proof. Fix δ with $0 < \delta < (\int f_k^2 - \int f_j^2)/10$; see (4.5). By Lemma 5.2, for almost all ω, for all sufficiently large n,

$$Q_{jn}(\omega) < \left(\tfrac{1}{2} d_{jn} \int f_j^2 + \delta\right) 2^n$$

$$Q_{kn}(\omega) > \left(\tfrac{1}{2} d_{kn} \int f_k^2 - \delta\right) 2^n.$$

Because $d_{in} \to 1$ as $n \to \infty$ for $i = j$ or k, for all large n,

$$Q_{jn}(\omega) < \left(\tfrac{1}{2} \int f_j^2 + 2\delta\right) 2^n$$

$$Q_{kn}(\omega) > \left(\tfrac{1}{2} \int f_k^2 - 2\delta\right) 2^n.$$

So $Q_{kn}(\omega) - Q_{jn}(\omega) > \delta 2^n$ for n large. On the other hand, $b_{kn} - b_{jn} = O(n)$ and $c_{kn} - c_{jn} = o(1)$ as $n \to \infty$. The upshot is that $\log R_{kn} - \log R_{jn} \to \infty$ as $n \to \infty$; see (3.4). □

Recall the hierarchical prior π from (2.2). Given the data, the posterior probability on theory j is $\tilde{\pi}_n\{j\}$, as computed in (3.7).

Corollary 5.1. *Fix j, k with $j < k$, $f_j \neq f_k$, and $w_k > 0$. Then $\tilde{\pi}_n\{j\}/\tilde{\pi}_n\{k\} \to 0$ as $n \to \infty$, P_f-almost surely.*

Lemma 5.6 showed that early theories become untenable; Lemma 5.7 shows that theories n, $n-1$, . . . and so forth are also quite implausible, a posteriori.

Lemma 5.7. *Fix $j = 0, 1, \ldots$. For any k, fixed but sufficiently large, $R_{n-j,n}/R_{k,n} \to 0$ as $n \to \infty$, P_f-almost surely.*

Proof. By Lemma 5.5, there is a small positive $\varepsilon = \varepsilon(j)$ such that, for all sufficiently large n, for almost all ω,

$$Q_{n-j,n}(\omega) - b_{n-j,n} - c_{n-j,n} < \tfrac{1}{2} 2^n \left(\int_{C_\infty} f(x)^2 \lambda^\infty(\mathrm{d}x) - \varepsilon\right).$$

By (4.5), we can fix a large k with

$$\int f_k^2 > \int f^2 - \tfrac{1}{3}\varepsilon.$$

By Lemma 5.2, for all sufficiently large n, for almost all ω,

$$Q_{k,n}(\omega) > \tfrac{1}{2} 2^n \left(\int_{C_\infty} f(x)^2 \lambda^\infty(\mathrm{d}x) - \tfrac{1}{2}\varepsilon\right);$$

this uses $\lim_{n\to\infty} d_{kn} = 1$. As before,

$$b_{k,n} + c_{k,n} = O(n) \qquad \text{as } n \to \infty,$$

so $\log R_{k,n}(\omega) - \log R_{n-j,n}(\omega) > \tfrac{1}{4}\varepsilon 2^n$ for n large enough; see (3.4). □

Corollary 5.2. *Fix $B < \infty$. For any k sufficiently large, $\tilde{\pi}_n\{j: j \geqslant n - B\}/\tilde{\pi}_n\{k\} \to 0$ as $n \to \infty$, P_f-almost surely.*

Proof. This is immediate from Lemma 5.7, with Lemma 3.1(b) to handle $j > n$. □

If f is *finitary*, that is, $f = f_k$ for some k, then weight concentrates on the minimal k with $f = f_k$ and $w_k > 0$; that case will be handled in the next section. Otherwise, the posterior weight on any particular k tends to 0.

We are now ready to prove Theorem 2.1, under the side condition

$$f \equiv f_j \text{ for no } j, \tag{5.3}$$

i.e. where f is not finitary.

Fix j. There is a $k > j$ such that $w_k > 0$ and $f_k \neq f_j$. This uses (5.3), (4.5) and the assumption (2.2) that $w_k > 0$ for arbitrarily large k. Theories in the range 0 to j become unlikely relative to theory k, by Corollary 5.1. Furthermore, theories in the range $[n - B, \infty)$ become unlikely by Corollary 5.2. In short, posterior mass concentrates on theories k with $j < k < n - B$, where j and B are any large positive integers. For k in that range, $\tilde{\pi}_{kn}$ concentrates near f_k by Corollary 4.3; and f_k is close to f. The L_2 metric is used throughout.

This completes the proof of Theorem 2.1 under the side condition (5.3). Lemma (3.10) in Diaconis and Freedman (1993) can be used to obtain bounds.

6. Finitary *f*

Suppose $f = f_k$ for some k; let k_0 be the least such k. We must prove Theorems 2.1 and 2.2. Let k_1 be the least $k \geqslant k_0$ with $w_k > 0$; then $f_{k_1} = f$. If $j < k_0$, then $\tilde{\pi}_n\{j\}/\tilde{\pi}\{k_1\} \to 0$ as $n \to \infty$, P_f-almost surely, by Corollary 5.1; if $k_0 \leqslant j < k_1$, then $\tilde{\pi}_n\{j\} = 0$. Theorem 2.1 follows, by Corollary 4.3 and Corollary 5.2. However, posterior mass does not drift towards larger and larger theories.

Theorem 2.2 follows from Corollary 5.2 and the next result.

Proposition 6.1. *For any B sufficiently large,*

$$\tilde{\pi}_n\{j: k_1 < j \leqslant n - B\}/\tilde{\pi}_n\{k_1\} \to 0$$

as $n \to \infty$, P_f-almost surely.

The proof of Proposition 6.1 is deferred. Only the case $k_1 = 0$ needs to be argued: we can assume that, for some constant c,

$$w_0 > 0 \text{ and } f \equiv c. \tag{6.1}$$

Since $f \equiv c$,

$$Q_{kn} = \tfrac{1}{2} d_{kn} \left\{ 2^n c^2 + 2c \sum_{t \in C_n} \zeta_t + \Xi_{kn} \right\}, \tag{6.2}$$

where

$$\Xi_{kn} = \sum_{s \in C_k} 2^{n-k} \overline{\zeta}_s^2 \tag{6.3}$$

and $\overline{\zeta}_s$ is the average of $\zeta_t = Y_t - f(\xi_t)$ over $t \in C_n$ that extend $s \in C_k$. In particular, by (4.2),

$$\Xi_{kn} \text{ is } \chi^2 \text{ with } 2^k \text{ degrees of freedom.} \tag{6.4}$$

The next lemma is elementary; see also (5.17) in Diaconis and Freedman (1993). The notation is laid out in (3.4).

Lemma 6.1.

(a) *Fix n. Then $k \to d_{kn}$ is monotone decreasing for $k = 0, 1, \ldots, n$.*
(b) *Fix n. Then $k \to 2^k(n-k)$ is monotone increasing for $k = 0, 1, \ldots, n-2$.*
(c) *Fix k. Then $c_{kn}/b_{kn} \to 0$ as $n \to \infty$.*

Recall Ξ_{kn} from (6.3).

Lemma 6.2. *Fix $\varepsilon > 0$. Let $B = 2/\varepsilon$. Almost surely, for all sufficiently large n, $\Xi_{kn} < \varepsilon 2^k (n-k)$ for all $k \leqslant n - B$.*

Proof. Let $n' = \log n$. Consider first the k with $n' < k \leqslant n - B$. Then $\varepsilon 2^k (n-k) \geqslant 2^{k+1}$, because $n - k \geqslant B$. By Lemma 4.2 and (6.4), for $\rho < 1$,

$$\sum_{n=1}^{\infty} \sum_{k=n'}^{n-B} P_f\{\Xi_{kn} > 2^{k+1}\} < \sum_{n=1}^{\infty} \sum_{k=n'}^{\infty} \rho^{2^k}. \tag{6.5}$$

Of course, $2^{k+1} \geqslant 2^k + 1$. Thus,

$$\sum_{k=n'}^{\infty} \rho^{2^k} < \rho^{2^{n'}} \Big/ (1-\rho) = \rho^{n^{\log 2}}/(1-\rho);$$

and the sum in (6.5) is finite.

Consider next the k with $k \leqslant n'$. Now $2^k \leqslant n^{\log 2}$; and $2^k(n-k) \geqslant n$, the value at $k = 0$, by Lemma 6.1(b). Let n^* be the greatest integer with $n^* \leqslant n^{\log 2}$. We have

$$\begin{aligned} P_f\{\Xi_{kn} > \varepsilon 2^k (n-k)\} &< P\{\chi^2_{n^*} > \varepsilon n\} \\ &= O(n^*/n^2) = O(n^{\log 2}/n^2) \end{aligned}$$

by Chebyschev's inequality: $\text{var}\{\chi^2_{n^*}\} = 2n^*$, and $E\{\chi^2_{n^*}\} = n^* = o(n)$. Then

$$\sum_{n=1}^{\infty}\sum_{k=1}^{n'} P_f\{\Xi_{kn} > \varepsilon n\} < \sum_{n=1}^{\infty} n' P\{\chi^2_{n^*} > \varepsilon n\}$$

$$< \text{const.} \sum_{n=1}^{\infty} (\log n) n^{\log 2}/n^2$$

$$< \infty.$$

The Borel–Cantelli lemma completes the proof. □

The next result is easily proved using Chebyshev's inequality; of course, much better estimates are available.

Lemma 6.3. *Let $S_n = \sum_{t\in C_n}\zeta_t$. Fix $\varepsilon > 0$. For all sufficiently large n, for P_f-almost all ω, $|S_n(\omega)| < 2^n\varepsilon$.*

Corollary 6.1. *Fix $\varepsilon > 0$. For all sufficiently large n, for all $j < n$, for P_f-almost all ω, $|S_n(\omega)|/(2^{n-j}+1) < \varepsilon b_{jn}$.*

Lemma 6.4. *Fix ε, with $0 < \varepsilon < 1/10$. Let $B = 2/\varepsilon$. Condition (6.1) is in force. Almost surely, for all sufficiently large n, for all j with $1 \leqslant j \leqslant n - B$,*

$$\log R_{jn} - \log R_{1n} < -n/20.$$

Proof. We evaluate $\log R_{jn} - \log R_{1n}$ by Lemma 3.1, as

$$b_{1n} - b_{jn} + c_{1n} - c_{jn} + Q_{jn} - Q_{1n} < A + B + C + D,$$

where

$$A = (1+\varepsilon)b_{1n} - b_{jn}$$

$$B = \tfrac{1}{2}(d_{jn} - d_{1n})2^n c^2$$

$$C = \tfrac{1}{2}[(d_{jn} - 1) - (d_{1n} - 1)]2cS_n$$

$$D = \tfrac{1}{2}d_{jn}\Xi_{jn} - \tfrac{1}{2}d_{1n}\Xi_{1n};$$

c_{1n} was estimated by Lemma 6.1(c), $-c_{jn} < 0$ was dropped, and Q was evaluated by (6.2).

Now B can be dropped; indeed, $B < 0$ by Lemma 6.1(a). In C, $1 - d_{in} = 1/(2^{n-i}+1)$ for $i = 1$ or j. In D, $-\frac{1}{2}d_{1n}\Xi_{1n} < 0$ can be dropped.

The upper bound becomes

$$(1+\varepsilon)b_{1n} - b_{jn} + |c||S_n|/(2^{n-j}+1) + |c||S_n|/(2^{n-1}+1) + \tfrac{1}{2}d_{jn}\Xi_{jn}.$$

The two terms involving S_n can be bounded above using Corollary 6.1; the last term can be

bounded above, using Lemma 6.2, to get $(\varepsilon b_{jn})/(\log 2) < 2\varepsilon b_{jn}$. We have an upper bound for $\log R_{jn} - \log R_{1n}$ of

$$(1 + 2\varepsilon)b_{1n} - (1 - 3\varepsilon)b_{jn} \leqslant (1 + 2\varepsilon)b_{1n} - (1 - 3\varepsilon)b_{2n}$$

by Lemma 6.1(b). Now use (3.4b) to get an upper bound of the form

$$(\log 2)[(-1 + 8\varepsilon)n + 6]. \qquad \square$$

Proposition 6.1 is an immediate consequence.

7. Possible generalizations

According to our priors π_k, the 2^k possible values θ_s for f were independent $N(0, 1)$ variables. Of course, $N(\mu, \sigma^2)$ would suffice. Furthermore, the problem can be broken down and handled separately on each of the 2^k pieces in C_k. In other words, according to π_j, the mean and variance of θ_s can depend on $s_1, \ldots, s_k$, provided $j > k$. Presumably, some sort of limiting argument is feasible, so that moderately general prior means and variances can be accommodated. Other possible generalizations are discussed in Diaconis and Freedman (1995). For example, the θ_s might taken as independent, with densities subject to uniform boundedness and smoothness conditions, as well as decay rates at $\pm\infty$. Another promising class of priors π_k is given by Ylvisaker (1987); these have some built-in smoothness.

8. Flat priors

Recall that π_k is 'flat' if the joint distribution of $\{\theta_s: s \in C_k\}$ is Lebesgue measure on 2^k-dimensional Euclidean space. We prove Theorem 2.3 by showing how to modify previous arguments. The predictive density, evaluated at the data, may be computed for $k \leqslant n$ as

$$\log R_{kn} = A_n - b_{kn} + c_k + Q_{kn}, \tag{8.1}$$

where (as before)

$$A_n = \tfrac{1}{2}2^n \log(1/2\pi) - \tfrac{1}{2}\sum_{t \in C_n} Y_t^2, \tag{8.2a}$$

$$b_{kn} = \tfrac{1}{2}2^k(n - k)\log 2. \tag{8.2b}$$

With flat priors,

$$c_k = \tfrac{1}{2}2^k \log 2\pi, \tag{8.2c}$$

$$Q_{kn} = \tfrac{1}{2}2^n \frac{1}{2^k}\sum_{s \in C_k} \overline{Y}_s^2. \tag{8.2d}$$

Although $R_{nn} = 1$, the representation (8.1) is more convenient for present purposes.

For $k > n$, there is a definitional problem, since R_{kn} must be infinite on sets of positive Lebesgue measure. For instance, take $k = 0$ and $n = 1$; suppose the first bit in x is 0. There

are two parameters, θ_0 and θ_1, both subject to Lebesgue measure; the first is observed, the second unobserved. The 'preditive measure' or 'marginal measure' of $\{Y \in A\}$ is

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_A \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}(y-\theta_0)^2} dy\, d\theta_0\, d\theta_1 = \infty.$$

Since this predictive measure assigns infinite mass to any set of positive Lebesgue measure, the usual disintegrations (and definitions of conditional measures) do not make much sense. See Section 10 below.

The simplest way around this definitional issue is to treat each design as a separate inference problem with its own prior. (Recall that the 'objective' part of the model does not require any particular specification of joint distributions across n.) At stage n, the prior puts weight w_k on theory $k < n$, and weight 0 on theory $k > n$. For theory n, we can keep the weight at w_n, or set it to $\sum_{k=n}^{\infty} w_k$; the latter seems to make the algebra a little easier. Thus, our convention is the following:

With flat priors, at stage n, the prior weight on theory $k < n$ remains w_k; the prior weight on theory $k > n$ is set to 0; the prior weight on theory n is set to be $\sum_{k=n}^{\infty} w_k$. (8.3)

We could also allow the Bayesian to ignore unobserved parameters when calculating predictive distributions and posteriors: the posterior distribution of an unobserved parameter stays flat. Arguments and results are essentially unchanged. Related papers that use improper priors include Kohn and Ansley (1987) and Wahba (1990).

We now estimate Q_{kn}; bounds are organized to prove the inconsistency result, but are modified later to make the consistency arguments.

Lemma 8.1. *Suppose the design is balanced, and normal in the sense of (2.3). Suppose the prior π is hierarchical, and the π_k are flat. For any small positive δ, there is a B finite but so large that, for P_f-almost all ω, for all sufficiently large n,*

(a) $Q_{kn}(\omega) < \frac{1}{2}2^n(\int f^2 + \delta) + \frac{1}{2}2^k$, *for* $k = 0, 1, \ldots, n-1, n$,
(b) $Q_{kn}(\omega) > \frac{1}{2}2^n(\int f^2 - \delta) + \frac{1}{2}2^k$, *for* $B \leqslant k \leqslant n$.

Proof. Fix $\delta > 0$. Use Lemma 5.2 for $k \leqslant n - B$, and Lemma 5.3 for $k > n - B$. □

Lemma 8.2. *Let*

$$\alpha_{kn} = -b_{kn} + c_k + 2^k/2 = \tfrac{1}{2}2^k\{(k-n)\log 2 + \log 2\pi + 1\}.$$

Then

(a) $\alpha_{kn} < 0$ *for* $k = 0, 1, \ldots, n-5$,
(b) $\alpha_{kn} > 0$ *for* $k = n-4, \ldots, n$,
(c) $\alpha_{nn} = \frac{1}{2}2^n \log 2\pi e$.

Remark. $k \to \alpha_{kn}$ is convex, monotone decreasing for $k = 0, 1, \ldots, n-6$, and monotone increasing for $k = n-6, \ldots, n-1, n$.

Corollary 8.1. *For any small positive δ, for P_f-almost all ω, for all sufficiently large n, for all $k \leqslant n-5$,*

$$\log R_{kn}(\omega) - \log R_{nn}(\omega) < -\tfrac{1}{2}2^n(\log 2\pi e - 2\delta).$$

Proof. For $k \leqslant n-5$,

$$\log R_{kn}(\omega) < \tfrac{1}{2}2^n\left(\int f^2 + \delta\right) + \alpha_{kn} + A_n$$

$$< \tfrac{1}{2}2^n\left(\int f^2 + \delta\right) + A_n;$$

the first inequality comes from Lemma 8.1(a) and the definitions; the second, from Lemma 8.2(a). On the other hand, by Lemma 8.2(b,c),

$$\log R_{nn}(\omega) > \tfrac{1}{2}2^n\left(\int f^2 - \delta\right) + \alpha_{nn} + A_n. \qquad \square$$

Corollary 8.2. *With data from a balanced normal design of order n, a hierarchical prior, and flat π_k, along subsequences of n for which*

$$\lim \frac{1}{2^n} \log\left(\sum_{k=n}^{\infty} w_k\right) > -\tfrac{1}{2}\log 2\pi e,$$

posterior mass concentrates on theories k with $k \geqslant n-4$, almost surely. Convention (8.3) is in force, so $k \leqslant n$.

Proof. Fix $\delta > 0$ so that for all sufficiently large n in the subsequence,

$$\sum_{k=n}^{\infty} w_k > \exp\left\{-\tfrac{1}{2}2^n(\log 2\pi e - 3\delta)\right\}. \tag{8.4}$$

The total posterior weight on theories 0 to $n-5$ is by Corollary 8.1 at most

$$\left(\sum_k w_k\right) \exp\left\{-\tfrac{1}{2}2^n(\log 2\pi e - 2\delta)\right\} R_{nn}(\omega).$$

The total posterior weight on theories k with $k \geqslant n$ is by (8.4) at least

$$\exp\left\{-\tfrac{1}{2}2^n(\log 2\pi e - 3\delta)\right\} R_{nn}(\omega).$$

Comparing the last two expressions completes the proof. $\square$

If $k \geqslant n-4$, there are at most 2^4 observations per parameter, so the posterior remains diffuse, and there is inconsistency. Lemmas 8.3–8.4 and (8.5) make this precise, and complete the proof of the inconsistency assertion in Theorem 2.3. Clearly,

$$\text{if } k \leqslant n, \tilde{\pi}_{kn} \text{ is a proper probability measure, making } \{\theta_s : s \in C_k\} \text{ independent } N(\overline{Y}_s, 1/2^{n-k}). \tag{8.5}$$

The next result applies a bit more generally. To state it, let Pr be a joint distribution for $\{\theta_s\colon s \in C_k\}$, making them independent $N(\mu_s, \sigma^2)$; the μ_s may be any real numbers. We can view Pr as a probability distribution on $h \in L_2$, as follows: Pr concentrates on Θ_k, the set of h that depend only on the first k bits of x; and the Pr law of $\{h(sx_{k+1}x_{k+2}\ldots)\colon s \in C_k\}$ is just the Pr law of $\{\theta_s\colon s \in C_k\}$. If $g \in L_2$, the δ-ball $N(g, \delta)$ around g was defined in (2.1).

Lemma 8.3. $\Pr\{N(g,\delta)\}$ *is maximized when g is piecewise constant, being* μ_s *on the x that extend s.*

Proof. The leading special case is $k = 0$ and $\sigma^2 = 1$. Let U be $N(\mu, 1)$. Then

$$\Pr\{N(g, \delta)\} = \Pr\left\{\int_{C_\infty} [U - g(x)]^2 \lambda^\infty(\mathrm{d}x) < \delta^2\right\}.$$

Of course,

$$\int (U - g)^2 = \left(U - \int g\right)^2 + \int \left(g - \int g\right)^2$$

is minimized when $g \equiv c$, and then

$$\int (U - g)^2 = (U - c)^2$$

is stochastically smallest when $c = E(U)$. □

Lemma 8.4. *Fix* $\delta > 0$ *with* $\delta^2 < 1/2^{B+1}$. *If* $n - B \leqslant k \leqslant n$, *the* $\tilde{\pi}_{kn}$*-mass of any* δ*-ball tends to* 0 *as* $n \to \infty$.

Proof. Let Ξ be χ^2 with 2^k degrees of freedom. By (8.4) and (8.5), the posterior mass in question is bounded above by $P\{\Xi < \delta^2 2^n\} < P\{\Xi < 2^k/2\}$, because $\delta^2 2^n < 2^k/2$. Then use Lemma 4.2. □

In particular, theories in the range $[n-4, n]$ cannot have posteriors concentrated near the true f – or anywhere else, for that matter. This completes the proof of inconsistency, and we now sketch the argument for consistency.

Fix k. For any $\delta > 0$, for all sufficiently large n, almost surely,

$$\log R_{kn}(\omega) > \tfrac{1}{2} 2^n \left(\int f_k^2 - \frac{\delta}{2}\right) + A_n. \tag{8.6}$$

This is obvious from (8.1) and Lemma 5.2. Recall α_{kn} from Lemma 8.2. Let

$$\alpha_j = \frac{1}{2^n}\alpha_{n-j,n} = \tfrac{1}{2}\frac{1}{2^j}(-j\log 2 + \log 2\pi\mathrm{e}). \tag{8.7}$$

This is negative for $j \geqslant 5$; see Lemma 8.2(a). Fix B, with $5 < B < \infty$. Then fix $\delta > 0$ so small that $\alpha_j < -2\delta$ for $5 \leqslant j \leqslant B$; choose k so large that $w_k > 0$ and $\int f_k^2 > \int f^2 - \delta/2$.

We claim that, almost surely, for all sufficiently large n,

theories in the range $[n - B, n - 5]$ are negligible a posteriori, relative to theory k. (8.8)

Indeed, theories in the range $[n - B, n - 5]$ have total posterior weight bounded above (almost surely, for all sufficiently large n) by

$$\left(\sum_i w_i\right) \exp\left\{2^n\left[\tfrac{1}{2}\left(\int f^2 + \delta\right) - 2\delta\right] + A_n\right\}; \tag{8.9}$$

see Lemma 8.1(a), and use the definition of the αs. On the other hand, theory k has posterior weight bounded below by

$$w_k \exp\left\{2^n\left[\tfrac{1}{2}\left(\int f^2 - \delta\right)\right] + A_n\right\}; \tag{8.10}$$

see (8.6). Comparing (8.9) and (8.10) proves (8.8).

Use condition (2.4a) to choose $\delta > 0$ so that, for all sufficiently large n,

$$\sum_{i=n}^{\infty} w_i < \exp\{-2^n[\tfrac{1}{2}\log 2\pi e + 32\delta]\}. \tag{8.11}$$

Choose k so large that $w_k > 0$ and $\int f_k^2 > \int f^2 - \delta/2$, for the new δ. Recall (8.3). We claim that

theory n is negligible a posteriori, relative to theory k. (8.12)

Indeed, Q_{nn} was bounded in Lemma 8.1. So, the total posterior weight on theory n is bounded above by

$$\left(\sum_{i=n}^{\infty} w_i\right) \exp\left\{2^n\left[\tfrac{1}{2}\left(\int f^2 + \delta\right) + \tfrac{1}{2}\log 2\pi e\right] + A_n\right\} < \exp\left\{2^n\left[\tfrac{1}{2}\left(\int f^2 + \delta\right) - 32\delta\right] + A_n\right\}. \tag{8.13}$$

Compare (8.13) with (8.10) – based on the new δ – to prove (8.12). The factor of 32 is quite generous here, but will be needed below.

With the same δ and k, we claim that,

for $j = 1, 2, 3, 4$, theory $n - j$ is negligible a posteriori, relative to theory k. (8.14)

Indeed,

$$w_{n-j} < \sum_{i=n-j}^{\infty} w_i < \exp\{-2^{n-j}[\tfrac{1}{2}\log 2\pi e + 32\delta]\}.$$

The posterior weight on theory $n - j$ is bounded above by

$$w_{n-j}\exp\left\{2^n\left[\tfrac{1}{2}\left(\int f^2+\delta\right)+\alpha_j\right]+A_n\right\}$$

$$<\exp\left\{2^n\left[\tfrac{1}{2}\left(\int f^2+\delta\right)+\alpha_j-\frac{1}{2^j}(\tfrac{1}{2}\log 2\pi e+32\delta)\right]+A_n\right\}$$

$$<\exp\left\{2^n\left[\tfrac{1}{2}\left(\int f^2+\delta\right)-\frac{1}{2^j}32\delta\right]+A_n\right\}$$

$$<\exp\left\{2^n\left[\tfrac{1}{2}\left(\int f^2+\delta\right)-2\delta\right]+A_n\right\}$$

because $\alpha_j<\alpha_0/2^j<(\log 2\pi e)/2^{j+1}$ and $2^j\leqslant 16$ (see (8.7)). Comparison with (8.10) proves (8.14). Combining (8.8) with (8.14) shows that posterior mass concentrates on theories k with $k<n-B$, and consistency follows as in the proof of Theorem 2.1.

9. An example

This section gives an example with flat priors and rapidly decreasing theory weights; the idea is to show that posterior mass can accumulate on theories n or $n-1$, whatever the time f may be. Suppose $w_k=0$ for odd k; while $w_k=\exp\{-C2^k\}$ for even k, where C is a positive constant. Clearly,

$$\lim\frac{1}{2^n}\log\left(\sum_{k=n}^{\infty}w_k\right)=\begin{cases}-C & \text{as even } n\to\infty\\ -2C & \text{as odd } n\to\infty.\end{cases}$$

If $C>\alpha_0=\frac{1}{2}\log 2\pi e$, there is consistency. If $0<C<\alpha_0$, inconsistency obtains. More interesting is this. Suppose

$$C>\tfrac{1}{3}\alpha_0+\tfrac{1}{6}\log 2, \tag{9.1a}$$

$$C<\alpha_0-\tfrac{1}{6}\log 2. \tag{9.1b}$$

We claim that:

$$\text{as odd } n\to\infty, \text{ posterior mass concentrates on theory } n-1; \tag{9.2a}$$

$$\text{as even } n\to\infty, \text{ posterior mass concentrates on theory } n. \tag{9.2b}$$

Only (9.2a) will be argued. Consider the odd n. Theories $k\leqslant n-5$ are negligible, by Corollary 8.2; theories $n-2$ and $n-4$ have prior mass 0. At stage n, by convention (8.3), theory n is given prior mass $w_{n+1}+w_{n+3}+\ldots$; indeed, $w_n=w_{n+2}=\ldots=0$. Thus, only theories $n-3$, $n-1$, and n are in contention. The posterior theory weights can be computed from (3.6) and (8.1)–(8.2), with Lemma 8.1 to estimate their magnitudes.

Let $K_n=A_n+\frac{1}{2}2^n\int f^2$. For any $\delta>0$, almost surely, for all sufficiently large odd n:

$$\text{the posterior weight on theory } n-3 \text{ is bounded above by } \exp\{K_n+2^n(C_3+\delta)\},$$

$$\text{where } C_3=(-C+\alpha_0-\tfrac{3}{2}\log 2)/2^3; \tag{9.3a}$$

the posterior weight on theory $n-1$ is bounded below by $\exp\{K_n + 2^n(C_1 - \delta)\}$,

where $C_1 = (-C + \alpha_0 - \frac{1}{2}\log 2)/2$; (9.3b)

the posterior weight on theory n is bounded above by $2\exp\{K_n + 2^n(D + \delta)]$,

where $D = -2C + \alpha_0$. (9.3c)

The factor of 2 in (9.3c) results from the estimate

$$\sum_{i=m}^{\infty} \exp\{-C2^{2i}\} < 2\exp\{-C2^{2m}\},$$

which holds because $e^{-4C_0} < e^{-C_0}/2$ when $C_0 > \frac{1}{3}\log 2$. It remains only to check that $C_1 > D$ and $C_1 > C_3$, which follow from (9.1a) and (9.1b), respectively.

10. A definitional issue with flat priors

You are about to observe independent normal variables X and Y. Both have variance 1. Theory 1 is that X and Y have the same mean, θ; there is a flat prior on θ. Theory 2 is that X has mean θ and Y has mean ψ; there is a flat prior on the pair (θ, ψ). To adjudicate between the two theories, you put prior mass 0.5 on each, observe (X, Y), and compute the posterior. But now suppose Y is not observed. Theory 2 has an infinite marginal 'density' for X; surely, that cannot tip the balance for theory 2. In this section, we review the calculus, and suggest a 'partial Bayes rule' convention: basically, the idea is to ignore Y and the prior on its parameter. That makes theories 1 and 2 agree on the observables, as seems sensible: X is $N(\theta, 1)$ and θ is uniform.

Let X be $N(\theta, 1)$ and let Y be $N(\psi, 1)$. Suppose X and Y are independent. Let λ be Lebesgue measure on the line. A Bayesian might assume a flat prior π for (θ, ψ), that is, $\pi = \lambda^2$. Let μ be the joint distribution of (θ, ψ, X, Y): if A, B, C, D are linear Borel sets, then

$$\mu\{\theta \in A \wedge \psi \in B \wedge X \in C \wedge Y \in D\} = \int_{\psi\in B}\int_{\theta\in A}\int_{y\in D}\int_{x\in C} f(x-\theta)f(y-\psi)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}\theta\,\mathrm{d}\psi. \tag{10.1}$$

Of course, the 'predictive' or 'marginal' law of (X, Y) relative to μ is λ^2. Given X and Y, the posterior law of θ, ψ is that of two independent $N(X, 1)$ and $N(Y, 1)$ variables. Indeed, let $Q_{xy}\{\mathrm{d}\theta, \mathrm{d}\psi\}$ be the proposed conditional. By Fubini's theorem,

$$\mu\{\theta \in A \wedge \psi \in B \wedge X \in C \wedge Y \in D\} = \int_{y\in D}\int_{x\in C} Q_{xy}\{A \times B\}\mu_0\{\mathrm{d}x, \mathrm{d}y\} \tag{10.2}$$

where μ_0 is the marginal law of X and Y, namely, λ^2. The 'disintegration' (10.2) makes rigorous the idea of the posterior.

That much is staightforward. Now suppose that Y is not observed. Suddenly, there is a

definitional crisis: the marginal law of X assigns infinite mass to any set of positive Lebesgue measure. Thus, it seems impossible to define the posterior distribution of θ, ψ given X by means of the usual disintegration formulas. For related calculations, see Eaton (1992).

There is a natural convention to make:

(a) the predictive law of X is uniform; and
(b) the posterior law of θ, ψ given X is this: θ is $N(X, 1)$, ψ is uniform, and the two are independent.

With these conventions, the inconsistency results of Sections 8 and 9 go through; only minor changes are needed in the arguments. Eliminating the weights on complex theories (k of order n or larger) tends to speed up the rate of convergence for proper priors; eliminating the prior mass beyond $n - 5$ does wonders even for flat-prior Bayesians. Thus, the convention followed in Section 8 seems more favourable to the Bayesians than the convention proposed here; even so, inconsistency is the result.

11. Bayesian regression, splines and wavelets

This section sketches a heuristic connection between our results and those in Cox (1993), via wavelet theory. Let $\{f_{jk}: k = 1, 2, \ldots, 2^j\}$ index the Haar wavelet functions of level j. Our covariates takes values in coin-tossing space, which is, of course, isomorphic to the unit interval. Thus, our prior can in principle be viewed as the distribution of

$$\sum_{j=0}^{\infty} \sum_{k=1}^{2^j} X_{jk} f_{jk}.$$

Each X_{jk} is a mixture of normal variates with mean 0, and the X_{jk} are uncorrelated. We may consider replacing X_{jk} by Z_{jk}, where the Z_{jk} are independent, normal, and $\mathrm{var}(Z_{jk}) = \mathrm{var}(X_{jk})$; the latter depends on j not k. Now

$$\sum_{j=0}^{\infty} \sum_{k=1}^{2^j} Z_{jk} f_{jk}$$

defines a prior of the kind studied by Cox.

This connection is interesting, but somewhat formal – because the law of $\{Z_{jk}\}$ is quite different from the law of $\{X_{jk}\}$. In particular, we do not see how to derive our results from his – or his from ours. Nor do we see how do derive consistency and inconsistency results of the kind we have previously demonstrated from wavelet theory. Cox's main result shows that, in his set-up, Bayesian confidence sets do not have good frequentist coverage probability, but that does not establish inconsistency in our sense, because the distance from the posterior mean to the true parameter is not bounded from below. Likewise, his estimates do not imply consistency, at least directly. However, calculations like those in his paper

should establish consistency, at least in his ℓ_2 set-up. For more discussion, see Diaconis and Freedman (1997).

Acknowledgements

We thank Peter Bickel and Joe Eaton for helpful comments. We are also indebted to a very helpful (but anonymous) referee.

References

Barron, A., Schervish, M.J. and Wasserman, L. (1997) The consistency of posterior distributions in nonparametric problems. Technical report, Statistics Department, Yale University.

Bernstein, S. (1934) *Theory of Probability.* Moscow.

Breiman, L. and Freedman, D. (1983) How many variables should be entered in a regression equation? *J. Amer. Math. Soc*., **78**, 131–136.

Brown, L. and Low, M. (1996) Asymptotic equivalence of non parametric regression and white noise. *Ann. Statist*., **24**, 2384–2398.

Bunke, O. and Milhaud, X. (1994) Asymptotic behavior of Bayes estimates under possibly incorrect models. Discussion Paper 24, Mathematics Institute, Humboldt University, Berlin.

Cox, D. (1993) An analysis of Bayesian inference for nonparametric regression. *Ann. Statist*., **21**, 903–923.

de Finetti, B. (1959) La probabilità, la statistica, nei rapporti con l'induzione, secondo diversi punti di vista. Centro Internazionale Matematica Estivo Cremonese, Rome. English translation in de Finetti (1972).

de Finetti, B. (1972) *Probability, Induction, and Statistics*. New York: Wiley.

Diaconis, P. (1988). Bayesian numerical analysis. In S.S. Gupta and J.O. Berger (eds), *Statistical Decision Theory and Related Topics IV,* Vol. 1, pp. 163–177.

Diaconis, P. and Freedman, D. (1988) On the consistency of Bayes estimates (with discussion). *Ann. Statist*., **14**, 1–67.

Diaconis, P. and Freedman, D. (1990) On the uniform consistency of Bayes estimates for multinomial probabilities. *Ann. Statist*., **18**, 1317–1327.

Diaconis, P. and Freedman, D. (1993) Nonparametric binary regression: a Bayesian approach. *Ann. Statist*., **21**, 2108–2137.

Diaconis, P. and Freedman, D. (1995) Nonparametric binary regression with random covariates. *Polish J. Math. Statist*., **15**, 243–273.

Diaconis, P. and Freedman, D. (1997) On the Bernstein–von Mises theorem for infinite dimensional parameters. Technical report no. 492, Department of Statistics, University of California, Berkeley.

Donoho, D.L. (1994) Asymptotic minimax risk for sup-norm loss: solution via optimal recovery. *Probab. Theory Related Fields*, **99**, 145–170.

Doss, H. (1984) Bayesian estimation in the symmetric location problem. *Z. Wahrscheinlichkeitstheorie*, **68**, 127–147.

Doss, H. (1985a) Bayesian nonparametric estimation of the median: Part I: Computation of the estimates. *Ann. Statist*., **13**, 1432–1444.

Doss, H. (1985b) Bayesian nonparametric estimation of the median. Part II: Asymptotic properties of the estimates. *Ann. Statist*., **13**, 1445–1464.

Eaton, M.L. (1992) A statistical diptych: admissible inferences – recurrence of symmetric Markov chains. *Ann. Statist*., **20**, 1147–1179.

Ferguson, T. (1974) Prior distributions on spaces of probability measures. *Ann. Statist*., **2**, 615–629.

Feller, W. (1968) *An Introduction to Probability and Its Applications*, Vol. I, 3rd edn. New York: Wiley.

Freedman, D. (1963) On the asymptotic behavior of Bayes estimates in the discrete case. *Ann. Math. Statist*., **34**, 1386–1403.

Freedman, D. (1965) On the asymptotic behavior of Bayes estimates in the discrete case II. *Ann. Math. Statist*., **36**, 454–456.

Geman, S. and Hwang, C.R. (1982) Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist*., **10**, 401–414.

Ghosal, S., Ghosh, J.K. and Ramamoorthi, R.V. (1997) Consistency issues in Bayesian nonparametrics. Technical Report, Indian Statistical Institute.

Ghosh, J.K., Sinha, B.K. and Joshi, S.N. (1982) Expansions for posterior probability and integrated Bayes risk. In S.S. Gupta and J.O. Berger (eds), *Statistical Decision Theory and Related Topics III*, Vol. 1, pp. 403–465. New York: Academic Press.

Johnson, R. (1967) An asymptotic expansion for posterior distributions. *Ann. Math. Statist*., **38**, 1899–1906.

Johnson, R. (1970) Asymptotic expansions associated with posterior distributions. *Ann. Math. Statist*., **41**, 851–864.

Kimeldorf, G. and Wahba, G. (1970) A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist*., **41**, 495–502.

Kohn, R. and Ansley, C. (1987) A new algorithm for spline smoothing and interpolation based on smoothing a stochastic process. *SIAM J. Sci. Statist. Comput*., **8**, 33–48.

Laplace, P.S. (1774) Memoire sur la probabilité des causes par les évènements. Memoires de mathématique et de physique présentés à l'Académie Royale des Sciences, par divers savants, et lûs dans ses assemblées 6. Reprinted in Laplace's *Oeuvres Complètes*, Vol. 8, pp. 27–65. English translation by S. Stigler (1986) *Statist. Sci*., **1**, 359–378.

LeCam, L. (1953) On some asymptotic properties of the maximum likelihood estimates and related Bayes estimates. *Univ. Calif. Publ. Statist*., **1**, 277–330.

LeCam, L. (1982) On the risk of Bayes estimates. In S.S. Gupta and J.O. Berger (eds), *Statistical Decision Theory and Related Topics III*, Vol. 2, pp. 121–138. New York: Academic Press.

LeCam, L. and Yang, G. (1990) *Asymptotics in Statistics: Some Basic Concepts*. New York: Springer-Verlag.

Lindley, D.V. and Smith, A.F.M. (1972) Bayes estimates for the linear model. *J. Roy. Statist. Soc. Ser. A/B*, **67**, 1–19.

Poincaré, H. (1896) *Calcul des probabilités*. Paris: G. Carré.

Pollard, D., Torgersen, E. and Yang, G. (eds) (1997) *Festschrift for Lucien LeCam*. New York: Springer-Verlag.

Schwartz, L. (1965) On Bayes procedures. *Z. Wahrscheinlichkeitstheorie verw. gebiete*, **4**, 10–26.

Shen, X. (1996) On the properties of Bayes procedures in general parameter spaces. Technical Report, Statistics Department, Ohio State University.

Shibata, R. (1981) An optimal selection of regression variables. *Biometrika*, **68**, 45–54.

Shibata, R. (1986) Consistency of model selection and parameter estimation. In J. Gani and M.B. Priestley (eds), *Essays in Time Series and Allied Processes*. *J. Appl. Probab*., **23A**.

Stone, C. (1982) Optimal global rates of convergence for nonparametric regression. *Ann. Statist*., **10**, 1040–1053.

Traub, J.F., Wasilkowski, G.W. and Wozniakowski, H. (1988) *Information-Based Complexity*. Boston: Academic Press.

von Mises, R. (1964) *Mathematical Theory of Probability and Statistics* (ed. H. Geiringer). New York: Academic Press.

Wahba, G. (1990) *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.

Ylvisaker, D. (1987) Prediction and design. *Ann. Statist*., **15**, 1–19.

Received July 1994 and revised May 1997.