Posterior consistency of logistic Gaussian process priors in density estimation

Surya T. Tokdar[a,*], Jayanta K. Ghosh[a,b]

[a]*Department of Statistics, Purdue University, USA*
[b]*Division of Theoretical Statistics and Mathematics, Indian Statistical Institute, India*

Abstract

We establish weak and strong posterior consistency of Gaussian process priors studied by Lenk [1988. The logistic normal distribution for Bayesian, nonparametric, predictive densities. J. Amer. Statist. Assoc. 83 (402), 509–516] for density estimation. Weak consistency is related to the support of a Gaussian process in the sup-norm topology which is explicitly identified for many covariance kernels. In fact we show that this support is the space of all continuous functions when the usual covariance kernels are chosen and an appropriate prior is used on the smoothing parameters of the covariance kernel. We then show that a large class of Gaussian process priors achieve weak as well as strong posterior consistency (under some regularity conditions) at true densities that are either continuous or piecewise continuous.

Keywords: Gaussian process; Logistic transformation; Nonparametric density estimation; Posterior consistency; Sup-norm support

1. Introduction

Logistic Gaussian process priors for Bayesian nonparametric density estimation were introduced and studied by Leonard (1978) and Lenk (1988, 1991). Lenk (1988) showed that the posterior has an elegant description through a conjugacy class of generalized logistic Gaussian processes. Different numerical approaches for calculating the Bayes estimate were also proposed by Lenk (1988, 1991). Compared to the Dirichlet mixtures of normals, which are currently the most popular as well as the most studied priors for densities, logistic Gaussian process priors have greater flexibility in modeling smoothness through covariance. Somewhat different Gaussian process priors appear in the works of Kimeldorf and Wahba (1970), Wahba (1978), which relate to estimation of integrated mean square risk of splines. See also Gu and Qiu (1993). However, the connection of these priors with the priors of Lenk is not yet fully explored.

In this paper, we initiate a theoretical study by examining weak and strong posterior consistency of the logistic Gaussian process prior. Posterior consistency is discussed in Ghosal et al. (1999), Barron et al. (1999), Ghosh and

* Corresponding author.
E-mail address: stokdar@stat.purdue.edu (S.T. Tokdar).

Ramamoorthi (2003) etc. It is well-known that weak and strong consistency of the posterior imply weak and strong consistency of the Bayes estimates (see Ghosh and Ramamoorthi, 2003, Proposition 4.2.1).

Our study of consistency, more specifically, Theorems 4.1–4.6 have helped us in identifying a novel and fast way to compute the posterior under a logistic Gaussian process prior. The details of this work will be reported elsewhere.

The prior is formally introduced in Section 2 for the space of densities on a bounded interval of $\mathbb{R}^d$, $d \geqslant 1$. Section 3 details the basic concepts and results about weak and strong posterior consistency. In Section 4, Theorem 4.1 relates weak consistency of a logistic Gaussian process prior to the sup-norm support of the underlying Gaussian process. A precise and useful characterization of this sup-norm support for a general class of Gaussian processes is obtained in the subsequent theorems. In Section 5, we obtain sufficient conditions required for strong consistency to hold. It is worth noting that when $d > 1$, conditions for strong consistency differ significantly from those in the case of $d = 1$. For the higher dimensions, certain differentiability conditions of the underlying process are required and the proof works through a different sieve, vide Van der Vaart and Wellner (1996), which first came to our notice from the paper of Ghosal and Roy (2005).

2. Logistic Gaussian process priors

As indicated before, we shall focus only on densities supported on a fixed bounded interval I in $\mathbb{R}^d$ for some $d \geqslant 1$. Without loss of generality we take $I = [0, 1]^d$. Denote by $w \mapsto f_w$ the logistic transformation from the space of functions on I to the space of densities (w.r.t the Lebesgue measure) on I given by,

$$f_w(t) = \frac{e^{w(t)}}{\int_I e^{w(s)}\, ds}, \quad t \in I, \tag{1}$$

whenever the integral exists.

Consider a fixed function $\mu(\cdot)$ on I and a family of covariance functions $\{\sigma_\beta(\cdot,\cdot)\}$ on $I \times I$ that depends on a finite dimensional parameter β. Take a probability distribution H on the space of β. Let $GP_I(0, \sigma)$ denote the distribution of a separable mean zero Gaussian process on I with covariance $\sigma(\cdot,\cdot)$. Assume that,

$$\forall \beta \in \text{support}(H), \quad W \sim GP_I(0, \sigma_\beta) \Rightarrow \int_I e^{\mu(s)+W(s)}\, ds < \infty \text{ a.s.} \tag{2}$$

This assumption allows us to model a random density f on I in the following way:

$$f \mid W, \beta = f_{\mu+W},$$

$$W \mid \beta \sim GP_I(0, \sigma_\beta),$$

$$\beta \sim H.$$

The process $f_{\mu+W}(t)$ realizes its values in the space of densities on I, thus inducing a prior on this space. We shall call this a logistic Gaussian process prior and denote it by Π.

The choice of a bounded interval avoids integrability problems. In this case, (2) is true whenever W admits almost surely continuous and hence bounded sample paths under $GP_I(0, \sigma_\beta)$. Such a condition is rather easily satisfied by many σ_β. In principle, we could define a process on an unbounded set and study the conditions required in defining (1). In this paper we concentrate on the bounded case for technical convenience. A major effort would be required to extend the results to the unbounded case.

The density f_μ sort of captures the central path of the process $f_{\mu+W}$. This can be seen from the identity $E(\log f_{\mu+W}(t)|\beta) = \mu(t) + const.$, whose logistic transform is nothing but f_μ. This enables one to appropriately elicit the parameter μ in presence of prior knowledge. One default choice is $\mu \equiv 0$ which produces the uniform density as the prior guess.

A simple way to choose the family σ_β is the following. Let $\sigma_0(\cdot,\cdot)$ be a fixed covariance function on $\mathbb{R}^d \times \mathbb{R}^d$ and let $\beta \in (\mathbb{R}^+)^d$. Then,

$$\sigma_\beta(s,t) = \sigma_0(\beta s, \beta t), \quad s,t \in I$$

is a covariance function on $I \times I$. Here, for $d > 1$, βs is the vector of coordinatewise products of β and s. What makes this formulation appealing is that,

$$W \sim GP_I(0, \sigma_\beta) \iff W(\cdot) \stackrel{\mathscr{D}}{=} W_0(\beta \cdot) \quad \text{with } W_0 \sim GP_{\beta I}(0, \sigma_0). \tag{3}$$

Hence small β results in smooth sample paths of f and large β produces oscillating sample paths. In other words, β acts like a (inverted) smoothing window in this model. The base covariance kernel σ_0 determines the degree of differentiability of the sample paths of W and can be selected appropriately to reflect prior expectations.

3. Basics of consistency

Suppose independent observations $X_1, \ldots, X_n$ are available from a density f_0 belonging to some space of densities $\mathscr{F}$. Let Π be a prior distribution on $\mathscr{F}$. The notion of consistency of the posterior $\Pi(\cdot|X_1, \ldots, X_n)$ at f_0 is formalized by the following two definitions, which differ only in terms of the topology on $\mathscr{F}$ under consideration.

Definition (*weak consistency*). A prior Π on $\mathscr{F}$ is said to achieve weak posterior consistency at f_0 if for any weak neighborhood U of f_0, $\Pi(U|X_1, \ldots, X_n) \to 1$ almost surely under P_{f_0}.

Definition (*strong consistency*). A prior Π on $\mathscr{F}$ is said to achieve strong posterior consistency at f_0 if for any L_1-neighborhood U of f_0, $\Pi(U|X_1, \ldots, X_n) \to 1$ almost surely under P_{f_0}.

For weak consistency, an elegant sufficient condition was derived in Schwartz (1965) in terms of a Kullback–Leibler (KL) support condition on Π and f_0. We give the details below.

Definition (*KL support*). Let $K(f, g)$ denote the KL divergence $\int f \log(f/g)$ between any two densities f and g. An $f_0 \in \mathscr{F}$ is said to be in the KL support of Π if

$$\forall \varepsilon > 0, \quad \Pi(f : K(f_0, f) < \varepsilon) > 0.$$

We would use the notation $f_0 \in KL(\Pi)$ to mean that f_0 is in the KL support of Π.

Theorem 3.1 (*Schwartz*). *If $f_0 \in KL(\Pi)$, then Π achieves weak posterior consistency at f_0.*

Remark. It is natural that for any kind of posterior consistency to hold, the true f_0 should belong to some sort of support of the prior. Otherwise, the posterior probability near f_0 would be always zero. Theorem 3.1 says that even for weak consistency one requires this condition in a fairly strong form, namely, f_0 is in the KL support of Π.

For strong consistency one needs more than just having $f_0 \in KL(\Pi)$. The following theorem from Ghosal et al. (1999, Theorem 2) gives a precise sufficient condition using metric entropy. We first provide with the definition of this.

Definition. Let $(\mathscr{F}, d)$ be a metric space. For any set $G \subset \mathscr{F}$ and any $\delta > 0$, the metric entropy $J(\delta, G, d)$ is defined as the logarithm of the minimum $k \geqslant 1$ for which there exist $g_1, \ldots, g_k \in \mathscr{F}$ such that $G \subset \bigcup_{i=1}^{k} \{f : d(f, g_i) < \delta\}$.

In the following we would use $\|\cdot\|_1$ to denote the L_1 norm on the space of densities on I.

Theorem 3.2. *Suppose for all $\varepsilon > 0$ there exist $\delta < \varepsilon$, $b < \varepsilon^2/2$, $c_0, c_1 > 0$ and sets F_n such that for all large n,*

(a) $\Pi(F_n^c) < c_1 e^{-nc_0}$ *and*
(b) $J(\delta, F_n, \|\cdot\|_1) < nb$.

Then, Π achieves strong posterior consistency at any $f_0 \in KL(\Pi)$.

Remark. The assumption in Theorem 3.2 is a kind of regularity condition on the prior. It identifies a relatively small set F_n outside which the prior puts exponentially small probability.

4. Weak consistency of logistic Gaussian process priors

We start by exploring the relationship between the processes W and $f_{\mu+W}$ in an attempt to find the KL support of Π. The following simple theorem is crucial. In the subsequent sections, $\|\cdot\|_\infty$ would denote the sup-norm on functions over I.

Theorem 4.1. *For any two functions $w_1(t)$ and $w_2(t)$ on I,*

$$\|w_1 - w_2\|_\infty < \varepsilon \Rightarrow \left\| \log \frac{f_{\mu+w_1}}{f_{\mu+w_2}} \right\|_\infty < 2\varepsilon.$$

Proof. Since $w_2(t) - \varepsilon < w_1(t) < w_2(t) + \varepsilon$ for all $t \in I$, it follows that,

$$e^{-\varepsilon} e^{\mu(t)+w_2(t)} < e^{\mu(t)+w_1(t)} < e^{\varepsilon} e^{\mu(t)+w_2(t)} \quad \forall t \in I$$

and hence,

$$e^{-\varepsilon} \int_I e^{\mu(t)+w_2(t)}\,dt < \int_I e^{\mu(t)+w_1(t)}\,dt < e^{\varepsilon} \int_I e^{\mu(t)+w_2(t)}\,dt.$$

Therefore,

$$e^{-2\varepsilon} f_{\mu+w_2}(t) < f_{\mu+w_1}(t) < e^{2\varepsilon} f_{\mu+w_2}(t) \quad \forall t \in I$$

from which the result follows easily. □

An immediate consequence of this result is that $\|w_1 - w_2\|_\infty < \varepsilon$ implies that $K(f_{\mu+w_1}, f_{\mu+w_2}) < 2\varepsilon$. Therefore, one can reformulate the condition of Theorem 3.1 as

$$f_0 = f_{\mu+w_0} \quad \text{with some } w_0 \text{ satisfying } \forall \varepsilon > 0\ \Pr(\|W - w_0\|_\infty < \varepsilon) > 0. \tag{4}$$

However, such a representation of f_0 is possible only if f_0 is strictly positive on I. The stronger form stated above would be more useful to address f_0 that may touch zero at some points (see Theorem 4.6 and the remarks following it).

The reformulation given in (4) suggests that one should study the sup-norm support of the process W. It is worth pointing out that up to this point we do not need W to be a Gaussian process.

To obtain a precise characterization of this sup-norm support we would require the Gaussian assumption to a large extent. The following theorem gives the key result in this direction.

Theorem 4.2. *Define a set of functions on I as,*

$$\mathscr{A} = \left\{ w = \sum_{i=1}^{k} a_i \sigma_\beta(t_i^*, \cdot) \text{ for some } \beta \in \text{support}(H),\ k \geqslant 1,\ a_i \in \mathbb{R},\ t_i^* \in I \right\}$$

and let $\bar{\mathscr{A}}$ denote its sup-norm closure. Assume

(A1) $\exists M, m > 0$ *such that* $m \leqslant \sigma_0(t,t) \leqslant M, \forall t \in (\mathbb{R}^+)^d$.
(A2) $\exists C > 0, q > 0$ *such that* $[\sigma_0(t,t) + \sigma_0(s,s) - 2\sigma_0(t,s)]^{1/2} \leqslant C\|s-t\|^q\ \forall s, t \in (\mathbb{R}^+)^d$
(A3) *For any $n \geqslant 1$ and any $t_1, \ldots, t_n \in (\mathbb{R}^+)^d$, $\Sigma = ((\sigma(t_i, t_j)))$ is nonsingular.*

Then,

$$w_0 \in \bar{\mathscr{A}} \iff \forall \varepsilon > 0, \quad \Pr(\|W - w_0\|_\infty < \varepsilon) > 0.$$

After proving this result we have found out from a referee that the result about Gaussian processes is known. In view of this we offer only a brief plausibility argument of the if part in the Appendix. This is the part needed for posterior consistency. We choose to omit the plausibility argument in the less important converse direction. Full details are available from us on request.

Remark. The quantity on the left-hand side of (A2) is nothing but the canonical metric $d(s,t)=[Var(W_0(s)-W_0(t))]^{1/2}$ on the index set induced by the process $W_0 \sim GP(0,\sigma_0)$. The Lipschitz condition in (A2) relates this metric to the Euclidean distance on the index set. This condition produces strong bounds on the oscillations of W_0 and ensures that it admits continuous sample paths almost surely.

Remark. A number of commonly used covariance functions satisfy the assumptions stated in the above theorem. For example, an easy way to satisfy (A3) for $d=1$, is to put $\sigma_0(s,t)=\phi(s-t)$ where ϕ is the characteristic function of some symmetric probability density. Such stationary covariance kernels can be easily generated by taking $\phi(h)=\exp(-h^2)$ or $\phi(h)=1/(1+h^2)$. For this special case of stationary σ_0, the condition in (A2) reduces to $\sqrt{1-\phi(h)}\leqslant c|h|$. This makes the verification straightforward.

Theorem 4.2 underlines the necessity to understand the set $\bar{\mathscr{A}}$ better. We would do so for some specific covariance functions arising from both stationary and nonstationary processes. It turns out that if support$(H)=(\mathbb{R}^+)^d$, then in most of the cases the set $\bar{\mathscr{A}}$ equals $C(I)$—the set of all continuous functions on I.

For the following theorems we would use the notation $\mathscr{A}_{\sigma_0,H}$ to emphasize the dependence of the set $\mathscr{A}$ on the particular σ_0 and H that define it. Theorem 4.3 deals with the case when the underlying process is a Brownian motion with a random shift. Theorems 4.4 and 4.5 cover the broad category of stationary covariance functions like $\exp(-\sum|t_i-s_i|^\gamma)$, $1/\prod_i(1+|t_i-s_i|^k)$, etc.

Theorem 4.3. *Suppose $d=1$ and $\sigma_0(t,s)=1+\min(t,s)$, then $\bar{\mathscr{A}}_{\sigma_0,H}=C(I)$.*

Proof. Observe that any function $f_{a,b}$ of the form

$$f_{a,b}(t)=\begin{cases}0, & t<a,\\ \dfrac{t-a}{b-a}, & a\leqslant t<b,\\ 1, & b\leqslant t,\end{cases} \tag{5}$$

for some $0\leqslant a<b\leqslant 1$, admits the representation,

$$f_{a,b}(t)=\frac{\sigma_0(t,b)-\sigma_0(t,a)}{b-a} \tag{6}$$

and that any piecewise linear continuous function f with knots at $\{0=t_0<t_1<\cdots<t_k=1\}$ can be expressed as the linear combination

$$f(t)=f(t_0)\sigma_0(t_0,t)+\sum_{i=1}^{k}(f(t_i)-f(t_{i-1}))f_{t_{i-1},t_i}(t). \tag{7}$$

Since the collection of piecewise linear continuous functions forms a dense subset of $C(I)$, the proof is complete. □

Theorem 4.4. *Take $d=1$ and* support$(H)=\mathbb{R}^+$. *Suppose σ_0 can be written as $\sigma_0(t,s)=\phi(t-s)$ for some continuous, nowhere zero, symmetric density function ϕ on $\mathbb{R}$. Then $\bar{\mathscr{A}}_{\sigma_0,H}=C(I)$.*

Proof. For $h>0$, define $\phi_h(x)=(1/h)\phi(x/h)$ and let

$$\mathscr{A}_\phi=\left\{w(t)=\sum_{i=1}^{k}a_i\phi_h(t,t_i^*)\ \text{for some}\ h>0,\ k\in\mathbb{N},\ a_i\in\mathbb{R},\ t_i^*\in I\right\}. \tag{8}$$

It is straightforward that $\mathscr{A}_{\sigma_0,H}=\mathscr{A}_\phi$. Now if $f=g*\phi_h$ for some continuous function g supported on I, where f_1*f_2 denotes the convolution of two functions, then $f\in\bar{\mathscr{A}}_\phi$. This follows from the approximations given by the

Riemann sums of the integral in the convolution. Now take any arbitrary continuous function f_0 on I. Fix an $\varepsilon > 0$ and consider the function

$$f_1(t) = f_0(t) - f_0(0)\frac{\phi_{h_0}(t)}{\phi_{h_0}(0)} - f_0(1)\frac{\phi_{h_0}(t-1)}{\phi_{h_0}(0)}, \tag{9}$$

where $h_0 > 0$ is suitably chosen to ensure that

$$\max(|f_1(0)|, |f_1(1)|) < \varepsilon/12. \tag{10}$$

Using continuity of f_1 we can find a $\delta > 0$ such that

$$\sup_{|t-s|<\delta} |f_1(t) - f_1(s)| < \varepsilon/12. \tag{11}$$

Take $h > 0$ such that

$$\int_{-\delta}^{\delta} \phi_h(x)\,\mathrm{d}x \geqslant 1 - \varepsilon/(12M), \tag{12}$$

where $M = \sup_{t\in I}|f_1(t)| < \infty$. Take $f_2 = f_1 * \phi_h$. Then for any $t \in I$,

$$\begin{aligned}
|f_1(t) - f_2(t)| &= \left|f_1(t) - \int_0^1 f_1(x)\phi_h(t-x)\,\mathrm{d}x\right| \\
&= \left|f_1(t) - \int_0^1 f_1(t)\phi_h(t-x)\,\mathrm{d}x + \int_0^1 (f_1(t) - f_1(x))\phi_h(t-x)\,\mathrm{d}x\right| \\
&\leqslant |f_1(t)|\left(1 - \int_0^1 \phi_h(t-x)\,\mathrm{d}x\right) + \int_0^1 |f_1(t) - f_1(x)|\phi_h(t-x)\,\mathrm{d}x \\
&\leqslant |f_1(t)|\left(1 - \int_0^1 \phi_h(t-x)\,\mathrm{d}x\right) \\
&\quad + \int_{I\cap[t-\delta,t+\delta]} |f_1(t) - f_1(x)|\phi_h(t-x)\,\mathrm{d}x \\
&\quad + \int_{I\cap[t-\delta,t+\delta]^c} |f_1(t) - f_1(x)|\phi_h(t-x)\,\mathrm{d}x
\end{aligned} \tag{13}$$

The first term of (13) is smaller than $\varepsilon/6$ by (10) for $t \in [0, \delta) \cup (1-\delta, 1]$ and by (12) for $t \in [\delta, 1-\delta]$. The second and the third terms of (13) are always less than $\varepsilon/6$ by (11) and (12), respectively. From this the result follows. □

Theorem 4.5. *Suppose* $d > 1$, $\mathrm{support}(H) = (\mathbb{R}^+)^d$ *and*

$$\sigma_0(t,s) = \sigma_0^{(1)}(t_1,s_1)\sigma_0^{(2)}(t_2,s_2)\cdots\sigma_0^{(d)}(t_d,s_d) \tag{14}$$

for some functions $\sigma_0^{(i)}(t,s)$, $1 \leqslant i \leqslant d$, *on* $[0,1]\times[0,1]$. *Assume* $\bar{\mathscr{A}}_{\sigma_0^{(i)},H_i} = C[0,1]$ *for each* i *where* H_i *is the marginal distribution of* β_i *under* H. *Then* $\bar{\mathscr{A}}_{\sigma_0,H} = C(I)$.

Proof. By Stone–Weierstrass theorem, the collection of functions

$$\mathscr{B} = \left\{ f(t_1,\ldots,t_d) = \sum_{k=1}^{n} g_{k1}(t_1)g_{k2}(t_2)\cdots g_{kd}(t_d) : n \geqslant 1,\ g_{ki} \in C[0,1] \right\} \tag{15}$$

forms a dense set in $C(I)$. But, since $\mathrm{support}(H) = (\mathbb{R}^+)^d$, $\mathscr{A}_{\sigma_0,H}$ itself is a dense subset of $\mathscr{B}$ and hence the result follows. □

Remark. It is claimed in Lenk (1988) that for any $Z \sim GP_I(0,\sigma)$ and any integrable function g on I, $\Pr(\|Z - g\|_1 < \varepsilon) > 0$ for all $\varepsilon > 0$. Our results in this section suggest that this is not true in general but holds for many commonly used covariance kernels.

A lot can be gained whenever the sup-norm support of W is identified as $C(I)$. The following theorem implies that for any such logistic Gaussian process prior Π, any continuous density function f_0 belongs to the KL support. In particular, this covers the case when $f_0 = \text{Beta}(a, b)$ with $a, b \geqslant 1$. The proof needs Theorem 4.1 in its full force since the representation (4) may not always apply.

Theorem 4.6. *Suppose σ_0 satisfies the assumptions* (A1)–(A3) *and that $\bar{\mathscr{A}} = C(I)$. Also assume that $\mu(\cdot)$ is continuous. Then any continuous density function f_0 on I satisfies $f_0 \in KL(\Pi)$.*

Proof. For any $\varepsilon > 0$ take $\delta > 0$ such that $\log(1+\delta) < \varepsilon/2$. Define f_1 as

$$f_1(t) = \frac{f_0(t) + \delta}{1+\delta}, \quad t \in I.$$

Then f_1 is continuous and strictly positive on I. Therefore, $w_1(\cdot) = \log f_1(\cdot) - \mu(\cdot) \in \bar{\mathscr{A}}$. Observe that,

$$\begin{aligned} K(f_0, f_{\mu+W}) &= \int_I f_0(t) \log \frac{f_0(t)}{f_1(t)} + \int_I f_0(t) \log \frac{f_1(t)}{f_{\mu+W}(t)}\, \mathrm{d}t \\ &= \int_I f_0(t) \log \frac{f_0(t)}{f_0(t)+\delta} + \log(1+\delta) + \int_I f_0(t) \log \frac{f_{\mu+w_1}(t)}{f_{\mu+W}(t)}\, \mathrm{d}t \leqslant \frac{\varepsilon}{2} + \left\| \log \frac{f_{\mu+w_1}}{f_{\mu+W}} \right\|_\infty. \end{aligned}$$

Therefore, by Theorem 4.1,

$$\Pr(K(f_0, f_{\mu+W}) < \varepsilon) \geqslant \Pr\left(\left\| \log \frac{f_{\mu+w_1}}{f_{\mu+W}} \right\|_\infty < \frac{\varepsilon}{2} \right) \geqslant \Pr\left(\|W - w_1\|_\infty < \frac{\varepsilon}{4} \right) > 0$$

since $w_1 \in \bar{\mathscr{A}}$. This proves the result as $\varepsilon > 0$ is arbitrary. □

Remark. A similar result can be proved when f_0 is a uniform density on some compact subinterval K of I. Here again, for any $\varepsilon > 0$, we construct a strictly positive continuous density f_1 on I for which $K(f_0, f_1) < \varepsilon/2$. But the construction is a little more involved and uses Urysohn's lemma to obtain an intermediate $\tilde{f}_0$ that is continuous and close to f_0.

Remark. The above two results can be extended to the case when f_0 is a finite mixture of densities that are either continuous on I or uniform on a subinterval. Such finite mixtures cover the large class of piecewise continuous densities on I when $d = 1$.

5. Strong consistency of logistic Gaussian process priors

For strong consistency results, we simply produce F_n's that satisfy the regularity condition of Theorem 3.2. When $d = 1$, we would use $F_n = \{f_{\mu+W} : W \in S_n\}$ where,

$$S_n = \left\{ w : \sup_{|s-t|<1/n} |w(s) - w(t)| < \varepsilon/12 \right\}$$

That such F_n's satisfy the requirements of Theorem 3.2 can be assessed using the following result.

Theorem 5.1. *Let σ_0 satisfy the assumptions* (A1)–(A3) *and suppose $\Pr(\beta > n^{q/2}) < \exp(-cn)$ for all large n for some fixed $c > 0$. Then $J(\varepsilon, F_n, \|\cdot\|_1) < nb$ and $\Pi(F_n^c) < A\exp(-na)$ for some A, a, b.*

Proof. A simple calculation along the line of Theorem 4.1 shows that $\|w_1 - w_2\|_\infty < \varepsilon/4 \Rightarrow \|f_{\mu+w_1} - f_{\mu+w_2}\|_1 < \varepsilon$ for small enough $\varepsilon > 0$. Therefore, $J(\varepsilon, F_n, \|\cdot\|_1) \leqslant J(\varepsilon/4, S_n, \|\cdot\|_\infty)$.

Fix an $n \geqslant 1$ and let $t_j = j/n$, $0 \leqslant j \leqslant n$. Define,

$$A_n = \{m = (m_0, \ldots, m_n) \in \mathbb{Z}^{n+1} : m_0 = 0, |m_{j+1} - m_j| \leqslant 1, j \geqslant 0\}$$

and let C_n denote the set of w such that $w(0) = 0$, $w(t_j) = m_j\varepsilon/12$ for $1 \leqslant j \leqslant n$ for some $m \in A_n$ and w is linear in every $[t_j, t_{j+1}]$. A standard argument produces that C_n forms an $\varepsilon/4$-net of S_n. From this we conclude $J(\varepsilon/4, S_n, \|\cdot\|_\infty) \leqslant \log(\#C_n) = \log(\#A_n) = n\log 3$. Note that the constant $\log 3$ could be changed to any constant b by redefining F_n with a suitable scaling on n.

To prove the other statement, note that it suffices to bound the probability of F_n^c uniformly over $\beta < n^{q/2}$. It can be argued using Borell's inequality (also see Van der Vaart and Wellner, 1996, Proposition A.27) that for any $\beta < n^{q/2}$,

$$\Pr\left(\sup_{|s-t|<1/n} |W(s) - W(t)| > \varepsilon/12 | \beta\right) \leqslant A\exp(-a'/\sigma_n^2)$$

from some constants A, a' where

$$\sigma_n^2 := \sup_{|s-t|<1/n} E(W(s) - W(t)\beta)^2 \leqslant (c\beta/n)^2 \leqslant c^2/n.$$

From this the result follows easily. □

Remark. For $d > 1$ the sieve S_n defined above fails as its entropy shoots up to n^d. But one can construct a smaller sieve with large probability using existence of higher order derivatives of W. We briefly overview the structure of these sieves as presented in Ghosal et al. (2003). Suppose there are numbers β_n, $M_n \to \infty$ and a positive integer α such that,

$$\Pr\left(\max_j \beta > \beta_n\right) < \mathrm{e}^{-cn} \quad \text{for some } c > 0,$$

$$M_n^2\beta_n^{-2\alpha} \geqslant b_1 n \text{ and } M_n^{d/\alpha} = o(n) \quad \text{for some } b_1 > 0$$

and for each $t \in I$, the function $\sigma_0(t, \cdot)$ admits continuous partial derivatives up to order $2\alpha + 2$. Define,

$$S_n = \{\tilde{w} : \|D^q(w)\|_\infty < M_n, |q| < \alpha\},$$

where for $q \in \{0, 1, 2, \ldots\}^d$, $|q| = \sum q_j$ and $D^q(w)$ stands for the partial derivative $(\partial^{|q|}/\partial^{q_1}t_1 \cdots \partial^{q_d}t_d)w(t_!, \ldots, t_d)$. Then S_n satisfies,

$$J(\delta_1, S_n, \|\cdot\|_\infty) \leqslant KM_n^{d/\alpha}\delta_1^{-d/\alpha}, \quad \Pr(\tilde{W} \notin S_n) \leqslant A\mathrm{e}^{-bn} \tag{16}$$

for some K, A, b.

Acknowledgements

We thank Prof. Subhashis Ghosal for some constructive comments including the suggestion that we explore extending our results to higher dimensions. We also thank an Associate Editor and the referees for helping us improve our exposition.

Appendix A

A plausibility argument for the if part of Theorem 4.2: To keep the argument simple we only consider the case $w_0 \in \mathscr{A}$. Note that one can write such a w_0 as $w_0 = \sum_{i=1}^k a_i\sigma_{\beta_0}(t_i^*, \cdot)$ for some $k \geqslant 1$, $a_i \in \mathbb{R}$, $t_i^* \in I$ and $\beta_0 \in \text{support}(H)$. It follows from the representation (3), the assumption (A2) on σ_0 and the fact that $\beta_0 \in \text{support}(H)$, that it is enough to prove $\Pr(\|W - w_0\|_\infty < \varepsilon | \beta_0) > 0$.

First, choose a fine grid $\{t_1, \ldots, t_m\}$ covering I that includes the points t_i^*. The prior probability of W and w_0 differing by less than ε at these grid points is positive by (A3).

The conditional distribution of W given $W_m = (W(t_1), \ldots, W(t_m))$ is a Gaussian process with covariance free of W_m. Let μ_{W_m} denote the mean of this conditional process. Then, for a fine grid, the oscillations of the centered conditional process $W - \mu_{W_m}$ can be suitably bounded. This makes the conditional process put positive mass on sample paths

which are within ε distance of μ_{W_m}. The proof of this is somewhat technical, drawing upon the theory of a.s. continuity and boundedness of sample paths for a GP (see Adler, 1990 and Van der Vaart and Wellner, 1996, Corollary 2.2.8).

It remains to handle the conditional mean μ_{W_m} when the vector W_m is close to $w_{0,m} = (w_0(t_1), \ldots, w_0(t_m))$. If this function is Lipschitz, then the condition $|W(t) - w_0(t)| < \varepsilon$ at the grid points and the fineness of the grid would ensure $\|\mu_{W_m} - w_0\|_\infty < \varepsilon$. Unfortunately, the Lipschitz property is hard to show since the conditional mean involves inverse of a high dimensional matrix. It is at this point that the assumption $w_0 \in \mathscr{A}$ comes handy. An easy direct calculation shows that

$$\mu_{w_{0,m}}(\cdot) = w_0(\cdot) \tag{A.1}$$

and hence its Lipschitz property follows from that of w_0 (which is Lipschitz by (A2), (A3)). A little more work shows that μ_{W_m} is Lipschitz when $\max_j |W(t_j) - w_0(t_j)| < \varepsilon$.

References

Adler, R.J., 1990. An introduction to continuity, extrema, and related topics for general Gaussian processes. IMS Lecture Notes—Monograph Series.

Barron, A., Schervish, M., Wasserman, L., 1999. The consistency of posterior distributions in nonparametric problems. Ann. Statist. 27, 536–561.

Ghosal, S., Roy, A., 2005. Posterior consistency of Gaussian process prior for nonparametric binary regression. Preprint.

Ghosal, S., Ghosh, J.K., Ramamoorthi, R.V., 1999. Posterior consistency of Dirichlet mixtures in density estimation. Ann. Statist. 27, 143–158.

Ghosh, J.K., Ramamoorthi, R.V., 2003. Bayesian Nonparametrics. Springer, New York.

Gu, C., Qiu, C., 1993. Smoothing spline density estimation. Ann. Statist. 21 (1), 217–234.

Kimeldorf, G., Wahba, G., 1970. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. Ann. Math. Statist. 41 (2), 495–502.

Lenk, P.J., 1988. The logistic normal distribution for Bayesian, nonparametric, predictive densities. J. Amer. Statist. Assoc. 83 (402), 509–516.

Lenk, P.J., 1991. Towards a practicable Bayesian nonparametric density estimator. Biometrika 78 (3), 531–543.

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

Schwartz, L., 1965. On Bayes procedures. Z. Wahr. Verw. Gebiete. 4, 10–26.

Van der Vaart, A.W., Wellner, J.A., 1996. Weak convergence and empirical processes. Springer, New York.

Wahba, G., 1978. Improper priors, spline smoothing and the problem of guarding against model errors in regression. J. Roy. Statist. Soc. Ser. B Statist. Methodol. 40 (3), 364–372.