ISSN 2279-9362

Collegio Carlo Alberto

Posterior asymptotics in the supremum L_1 norm for conditional density estimation

Pierpaolo De Blasi
Stephen G. Walker

No. 475
December 2016

Carlo Alberto Notebooks

www.carloalberto.org/research/working-papers

Posterior asymptotics in the supremum L_1 norm for conditional density estimation

Pierpaolo De Blasi *†
University of Torino and Collegio Carlo Alberto, Italy

Stephen G. Walker ‡
University of Texas at Austin, USA

Abstract

In this paper we study posterior asymptotics for conditional density estimation in the supremum L_1 norm. Compared to the expected L_1 norm, the supremum L_1 norm allows accurate prediction at any designated conditional density. We model the conditional density as a regression tree by defining a data dependent sequence of increasingly finer partitions of the predictor space and by specifying the conditional density to be the same across all predictor values in a partition set. Each conditional density is modeled independently so that the prior specifies a type of dependence between conditional densities which disappears after a certain number of observations has been observed. The rate at which the number of partition sets increases with the sample size determines when the dependence between pairs of conditional densities is set to zero and, ultimately, drives posterior convergence at the true data distribution.

Key words and phrases: Nonparametric Bayesian inference, Posterior asymptotics, Conditional density estimation, Regression tree model.

1 Introduction

For (Y, X) two random variables with continuous distribution on the product space $\mathbb{R} \times \mathbb{X}$, we consider nonparametric Bayesian estimation of the conditional density $f(y|x)$ based on an iid sample from (Y, X). Let Π define a prior distribution on the space $\mathcal{F}$ of conditional densities and $(Y, X)_{1:n} = (Y_1, X_1), \ldots, (Y_n, X_n)$ denote the sample from the joint density $f_0(y|x)q(x)$, where $q(x)$ is the marginal density of the covariate X. From an asymptotic

*address for correspondence: pierpaolo.deblasi@unito.it

†Supported by the European Research Council (ERC) through St G N-BNP 306406

‡Supported by U. S. National Science Foundation through grant DMS-1506879

point of view, it is desirable to validate posterior estimation by establishing that the posterior distribution accumulates in suitably defined neighborhoods of $f_0(y|x)$ as $n \to \infty$, that is $\Pi\big(f \in \mathcal{F} :\ d(f, f_0) > \epsilon_n | (Y, X)_{1:n}\big) \to 0$, where $d(\,\cdot\,, \cdot\,)$ is a loss function on $\mathcal{F}$ and ϵ_n is the posterior converge rate. The choice of the loss function is an important issue, the literature on Bayesian asymptotics being mainly restricted to the expected L_1 norm,

$$d(f_1, f_2) = \int_{\mathbb{X}} \|f_1(\cdot|x) - f_2(\cdot|x)\|_1 \, q(x) \mathrm{d}x,$$

where $\|f_1(\cdot|x) - f_2(\cdot|x)\|_1 = \int_{\mathbb{R}} |f_1(y|x) - f_2(y|x)| \mathrm{d}y$ is the L_1 norm on the response space $\mathbb{R}$. The convenience of working with the expected L_1 norm is that general convergence theorems for density estimation can be easily adapted. Its use, although in many ways natural, may not always be appropriate. Posterior concentration relative to such loss justifies confidence that, for a new random sample of individuals with covariates distributed according to $q(x)$, the responses will be reasonably well-predicted by conditional density function samples from the posterior, but it would not justify similar confidence at a fixed chosen, rather than sampled, x^*. For this, posterior concentration in the supremum L_1 norm would be required, namely under the loss

$$\|f_1 - f_2\|_{1,\infty} := \sup_{x \in \mathbb{X}} \|f_1(\cdot|x) - f_2(\cdot|x)\|_1.$$

This would then justify the use of the posterior predictive conditional density,

$$f_n(y|x^*) := \textstyle\int_{\mathcal{F}} f(y|x^*) \, \Pi(\mathrm{d}f | (Y, X)_{1:n})$$

to make inference on $f_0(y|x^*)$. Note that the supremum L_1 norm induces a stronger metric compared to the expected L_1 norm, so derivation of posterior convergence rates is expected to be harder: ultimately, one needs to model an entire density $f(y|x)$ accurately at y, and for all x.

Popular Bayesian models for conditional density estimation typically specify a dependence structure between $f(\cdot|x)$ and $f(\cdot|x')$ which is convenient for small to moderate sample sizes since it allows borrowing of information. However, from an asymptotic point of view, an over-strong dependence structure might not be desirable. To discuss this point, if the posterior eventually puts all the mass on f_0 then clearly the correlation between $f(\cdot|x)$ and $f(\cdot|x')$ is zero. Hence, there is a decay to 0 of the dependence as the sample size increases. But this decay needs to be carefully managed, as we shall see with the model we study. Our solution is to allow the dependence between $f(\cdot|x)$ and $f(\cdot|x')$ to exist up to a finite sample size, depending on $|x - x'|$, and then fall to become 0, once there is enough data locally to evaluate each $f(\cdot|x)$ accurately. To this purpose, we consider the model

$$f(y|x) = \sum_{j=1}^{N_n} f_j(y) \mathbb{1}_{C_{nj}}(x), \quad f_j \overset{\text{ind}}{\sim} \tilde{\Pi},\ j = 1, \ldots, N_n \tag{1}$$

where $\mathbb{1}_A(\cdot)$ is the indicator function on the set A, the sets C_{nj}, $j = 1, \ldots, N_n$, form a sample size dependent partition of the covariate space $\mathbb{X}$ and each $f_j(y)$ is a density function on $\mathbb{R}$, modeled independently with a nonparametric prior $\tilde{\Pi}$. We will occasionally refer to the prior distribution of f_j as $\tilde{\Pi}_j$, where it is implicitly assumed that $\tilde{\Pi}_j$, $j = 1, \ldots, N_n$, are identical copies of the same nonparametric prior $\tilde{\Pi}$. Our preferred choice for $\tilde{\Pi}$ is a Dirichlet process location mixture of normal densities, see Section 2.2, although other choices can be made. Note that, since the conditional density is set to be the same across all $x \in C_{nj}$, $f_j(y)$ also corresponds to the marginal density of Y when X is restricted to lie in C_{nj}. As we are going to let N_n depend on n, the prior (1) is sample size dependent, and will be denoted by Π_n. Specifically, we take $\mathbb{X}$ to be a bounded set and let N_n increase to ∞ as $n \to \infty$ and C_{nj}, $j = 1, \ldots, N_n$, to form a finer and finer partition of $\mathbb{X}$ such that

$$|C_{nj}| = N_n^{-1}, \tag{2}$$

where $|A|$ is the Lebesgue measure of A. For example, when $\mathbb{X} = [0, 1]$, we define $C_{nj} = [(j-1)/N_n, j/N_n]$ and in fact in this paper we will focus on this case. So the key is that while x and x' are both in C_{nj} they share a common $f(\cdot|x) = f(\cdot|x')$. However, after some sample size n which determines N_n the two densities separate and become independent. Consequently, the borrowing of strength is a $0-1$ phenomenon, rather than a gradual decay. Model (1) bears similarity to the Bayesian regression tree model proposed by Chipman et al. (1998), which is based on a constructive procedure that randomly divides the predictor space sequentially and then generates the conditional distribution from base models within the blocks. See Ma (2012) for a recent nonparametric extension. In our case the partitioning is non random and depends on the data only through the sample size n.

Given the model (1)–(2), the goal is to find the rate at which N_n should grow in terms of n so that the posterior accumulates in sup-L_1 neighborhoods of $f_0(y|x)$, according to:

$$\Pi_n\big\{f : \ \|f - f_0\|_{1,\infty} > \epsilon_n | (Y, X)_{1:n}\big\} \to 0. \tag{3}$$

We will assume throughout that the marginal density $q(x)$ of the covariate X is bounded away from zero so that there are approximately n/N_n observations to estimate the conditional density $f_0(y|x)$ for x in each block C_{nj}. If N_n grows too fast then there are not sufficient observations per bin to estimate the density at x accurately; whereas if N_n grows too slowly there are too many observations from densities with x' which are too far from x, again making the density at x inaccurate. It is expected that N_n is determined by the prior $\tilde{\Pi}$ and the regularity of the true conditional density f_0. More precisely, our results hold under two main conditions. First, $\tilde{\Pi}$ needs to satisfy a summability condition of prior probabilities over a suitably defined partition of the space of marginal densities of Y. This requires the existence

of a high mass - low entropy sieve on the support of $\tilde{\Pi}$. Second, f_0 has to satisfy a type of Lipschitz continuity measured by the Kullback-Leibler divergence of $f_0(y|x)$ and $f_0(y|x')$ for x and x' close. We then show that posterior convergence (3) holds for $N_n\epsilon_n^2 \to \infty$, $n\epsilon_n^4 \to \infty$ and $N_n\tilde{\epsilon}^2_{n/N_n} = o(\epsilon_n^2)$, where $\tilde{\epsilon}_m$, $m \to \infty$, is an upper bound to the prior rates attained by $\tilde{\Pi}$ at the marginal density of Y when X is restricted to lie in C_{nj}. Hence the prior rates $\tilde{\epsilon}_m$ ultimately determines the posterior convergence rate ϵ_n. In section 2 we obtain a best rate of $n^{-1/6}$ and acknowledge that this is a first step in the direction of finding optimal sup-L_1 rates for large classes of density. This may be sub-optimal and may arise as an artefact of the model which does not entertain smoothness in a reasonable way. However, the zero-one dependence seems to us important to be able to make mathematical progress and a more smooth form of dependence appears overly complicated to work with.

We end this introduction with a review of asymptotic results for Bayesian nonparametric inference on conditional distributions. In nonparametric normal regression, i.e. when $Y = g(X)+\epsilon$ with $\epsilon \sim N(0, \sigma^2)$, the aim is typically to estimate the regression function $g(x)$ with respect to the L_p norm on the space of functions $\mathbb{X} \to \mathbb{R}$. In the case of fixed design and known error variance, which corresponds to the celebrated Gaussian white noise model, the sup-L_1 norm $\|\cdot\|_{1,\infty}$ is equivalent to the supremum norm $\|g\|_\infty = \sup_x |g(x)|$ in the space of regression functions and optimal posterior convergence rates are derived in Ginè and Nickl (2011); Yoo and Ghosal (2016) by using conjugate Gaussian priors and by Castillo (2014); Hoffman et al. (2015) for nonconjugate priors. In all three papers, the prior on g is defined via independent product priors on the coordinates of g onto a wavelet orthogonal basis. Such use of independent priors on the coefficients of a multirelsolution analysis is, to some extent, similar to modeling the conditional densities independently on each sets C_{nj} as in (1). In particular, the technique set forth in Castillo (2014) consists of replacing the commonly used testing approach by tools from semiparametric Benstein-von Mises results and it has been successful to obtain rates in the sup-norm in density estimation on a compact domain by using log-density priors and random dyadic histograms. In the case of random design and unknown error variance, Shively et al. (2009) obtain posterior consistency with respect to neighborhoods of the type $\{\|g - g_0\|_\infty \leq \epsilon,\ |\sigma/\sigma_0 - 1| \leq \epsilon\}$ under a monotonicity constraint on $g(\cdot)$. Consistency under the expected L_1 norm is considered in nonparametric binary regression by Ghosal and Roy (2006) and in multinomial logistic regression by De Blasi et al. (2010). More generally, Bayesian nonparametric models for conditional density estimation follow two main approaches: (i) define priors for the joint density and then use the the induced conditional density for inference; (ii) construct conditional densities without specifying the marginal distribution of the predictors. Posterior asymptotics is studied by Tokdar et al. (2010) under the first approach and by Pati et al. (2013); Norets and Pati (2014); Shen and Ghosal (2016) under the second approach.

In all the aforementioned papers, convergence is defined with respect to the expected L_1 norm. Tang and Ghosal (2007) study posterior consistency for estimation of the transition density in a nonlinear autoregressive model with respect to both expected and supremum L_1 norm, however for the latter some restrictive assumptions on the true transition density are imposed. Finally, Xiang and Walker (2013) consider the supremum L_1 norm in conditional density estimation with fixed designs of predictors. Compared to the latter paper, the challenge in our study is taking the argument from a finite setting to an uncountable setting.

The rest of the paper is organized as follows. Section 2 presents the main result and sufficient conditions to effect it, see Theorem 2.1. These are illustrated in the case of the prior $\tilde{\Pi}$ on marginal densities being a Dirichlet process location mixture of normal densities. The existence of a high mass - low entropy sieve on the support of $\tilde{\Pi}$ is established in Proposition 2.1. The proof of Theorem 2.1 is reported in Section 3, where we deal, as is the norm, with the numerator and denominator of the posterior separately, see Proposition 3.1. Section 4 presents an illustration of the type of f_0 which meets the aforementioned condition of Lipschitz continuity in the Kullback-Leibler divergence and also discusses the role of the condition of $q(x)$ bounded away from zero and an alternative derivation of sup-L_1 rates from expected L_1 rates. Some proofs and a technical lemma are deferred to the Appendix.

Notation and conventions. The following notation will be used throughout the article. For X a random variable with distribution P, the expectation of the random variable $g(X)$ is denoted by Pg and its sample average as $\mathbb{P}_n g = n^{-1}\sum_{i=1}^n g(X_i)$, according to the conventions used in empirical process theory. This applies to probability measures P defined on $\mathbb{R}$, $\mathbb{X}$ or $\mathbb{R}\times\mathbb{X}$. The frequentist (true) distribution of the data (Y, X) is denoted P_0, i.e. $P_0(\mathrm{d}y, \mathrm{d}x) = f_0(y|x)q(x)\mathrm{d}y\,\mathrm{d}x$, with E_0 denoting expectation with respect to P_0. The dependence of N_n and C_{nj} on n is silent and is dropped in the notation, and, unless explicitly stated, the predictor space is $\mathbb{X} = [0, 1]$. The space of conditional densities $f(y|x)$ is denoted $\mathcal{F} = \{f : \mathbb{X}\times\mathbb{R} \to \mathbb{R}_+ : \int_{\mathbb{R}} f(y|x)\mathrm{d}y = 1 \ \forall x\}$, while the space of densities on $\mathbb{R}$ is denoted $\tilde{\mathcal{F}} = \{f : \mathbb{R} \to \mathbb{R}_+ : \int_{\mathbb{R}} f(y)\mathrm{d}y = 1\}$. For $f, g \in \tilde{\mathcal{F}}$, the Hellinger distance between f and g is denoted as $H(f, g) = [\int(\sqrt{f}-\sqrt{g})^2]^{1/2}$. All integrals are to be intended with respect to a common dominating measure, e.g. the Lebesgue measure. For real valued sequences a_n, b_n, $a_n \lesssim b_n$ means there exists a positive constant C such that $a_n \leq Cb_n$ for all n sufficiently large; and $a_n \asymp b_n$ means $0 < \liminf_{n\to\infty}(a_n/b_n) < \limsup_{n\to\infty}(a_n/b_n) < \infty$. For any $\beta > 0, \tau_0 \geq 0$ and a nonnegative function L on $\mathbb{R}$, define the locally β-Hölder class with envelope L, denoted $\mathcal{C}^{\beta,L,\tau_0}(\mathbb{R})$, to be the set of all function on $\mathbb{R}$ with derivatives $f^{(j)}$ of all orders up to $r = \lfloor\beta\rfloor$, and for every $k \leq r$ satisfying

$$|f^{(k)}(x) - f^{(k)}(y)| \leq L(x)\mathrm{e}^{\tau_0(x-y)^2}|x-y|^{\beta-r},$$

for all $x, y \in \mathbb{R}$, cf definition in Shen et al. (2013). For $f \in \tilde{\mathcal{F}}$, define the K-L neighborhood of f as

$$B(\epsilon, f) = \Big\{ g \in \tilde{\mathcal{F}} : \ \int f \log(f/g) \leq \epsilon^2, \ \int f [\log(f/g)]^2 \leq \epsilon^2 \Big\}.$$

2 Main result

2.1 Posterior convergence theorem

Let $A_{\epsilon_n} \subset \mathcal{F}$ be the complement of an ϵ_n-ball around $f_0(y|x)$ with respect to the supremum L_1 norm as in (3), where ϵ_n is a positive sequence such that $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$. We are interested in the sequence of posterior distributions $\Pi_n(A_{\epsilon_n}|(Y,X)_{1:n})$ going to zero in probability with respect to the true data distribution P_0. We make the following assumptions on P_0. First we assume that the marginal density $q(x)$ is bounded away from 0,

$$\inf_{x \in [0,1]} q(x) > 0 \tag{4}$$

This implies that, under (2), as N increases, the expected number of X_i in C_j is $nQ(C_j) \asymp n/N$ for each j, where Q is the distribution associated to $q(x)$. See the discussion in Section 4 about relaxing condition (4). Second, we assume that the conditional density $f_0(y|x)$ is regular in that it satisfies the following form of Lipschitz continuity in terms of Kullback-Leibler type divergences: for $L > 0$, and $\gamma > 0$,

$$\int_{\mathbb{R}} f_0(y|x) \left(\log \frac{f_0(y|x)}{f_0(y|x')} \right)^r \mathrm{d}y \leq L|x - x'|^2, \quad r = 1, 2 \tag{5}$$

for all x, x' with $|x - x'| < \gamma$. See Section 4 for a discussion.

As for the prior Π_n, recall that it defines a distribution on $\mathcal{F}$ induced by the product of N independent priors $\tilde{\Pi}_j$ on $\tilde{\mathcal{F}}$ for each marginal density f_j, cfr (1). Each f_j is estimating the marginal density of Y when X is restricted to lie in C_j,

$$f_{0,j}(y) = Q(C_j)^{-1} \int_{C_j} f_0(y|x) q(x) \mathrm{d}x,$$

on the basis of approximately $m = n/N$ observations (Y_i, X_i) such that $X_i \in C_j$, cfr. (2) and (4). Note that, although not explicit in the notation, $f_{0,j}(y)$ depends on n through C_j via (2). We make use of a sieve, that is we postulate the existence of a sequence of sub models, say $\{\tilde{\mathcal{F}}_m, \, m \geq 1\}$, such that $\tilde{\mathcal{F}}_m \uparrow \tilde{\mathcal{F}}$. Moreover, for $\bar{\epsilon}_m$ a positive sequence such that $\bar{\epsilon}_m \to 0$ and $m\bar{\epsilon}_m^2 \to \infty$, and $(\tilde{A}_{mi})_{i \geq 1}$ Hellinger balls of radius $\bar{\epsilon}_m$ with $\tilde{\mathcal{F}}_m \subseteq \bigcup_i \tilde{A}_{mi}$, we assume that the prior $\tilde{\Pi}$ satisfies

$$\tilde{\Pi}(\tilde{\mathcal{F}}_m^c) \lesssim \exp\{-(C+4) m \bar{\epsilon}_m^2\} \tag{6}$$

$$\sum\nolimits_{i \geq 1} \tilde{\Pi}(\tilde{A}_{mi})^{1/2} \exp\{-c\, m \bar{\epsilon}_m^2\} \to 0 \tag{7}$$

for any $c, C > 0$. A key difference with similar sufficient conditions for posterior convergence in density estimation, such as equations (8) and (9) in Shen et al. (2013), or equations (37) and (39) in Kruijer et al. (2010), is that the same sequence $\bar{\epsilon}_m$ is used in (6) and (7). Finally, we rely on the prior rates of $\tilde{\Pi}$ at $f_{0,j}$. We denote by $P_{0,j}$ the probability distribution associated with $f_{0,j}(y)$. For $j = 1, \ldots, N$, let $B(\epsilon, f_{0,j})$ define a KL-neighborhood of $f_{0,j}$,

$$B(\epsilon, f_{0,j}) = \left\{ f : \; P_{0,j} \log(f_{0,j}/f) \leq \epsilon^2, \; P_{0,j}[\log(f_{0,j}/f)]^2 \leq \epsilon^2 \right\},$$

and assume that for $m \to \infty$ and a sequence $\tilde{\epsilon}_m \to 0$ such that $m\tilde{\epsilon}_m^2 \to 0$,

$$\tilde{\Pi}(B(\tilde{\epsilon}_m, f_{0,j})) \geq \mathrm{e}^{-C' m \tilde{\epsilon}_m^2} \quad \text{for } j = 1, \ldots, N \tag{8}$$

for some constant $C' > 0$. We are now ready to state the general convergence result which expresses the posterior convergence rate ϵ_n in terms of N and the prior rates $\tilde{\epsilon}_m$ and $\bar{\epsilon}_m$.

Theorem 2.1. *Let the assumptions above prevail. Also, assume that N and ϵ_n satisfy*

$$N\epsilon_n^2 \to \infty, \quad n\epsilon_n^4 \to \infty, \tag{9}$$

and that (6), (7) and (8) hold for $m = n/N$ and

$$\bar{\epsilon}_{n/N} = \epsilon_n/2, \quad N\tilde{\epsilon}_{n/N}^2 = o(\epsilon_n^2). \tag{10}$$

Then $\Pi_n(A_{\epsilon_n} | (Y, X)_{1:n}) \to 0$ in P_0-probability.

Note that the first condition in (9) imposes a restriction on how slow N can grow in n, while the second condition in (10) typically induces a restriction on how fast N can grow. See the illustration in Section 2.3.

2.2 Prior specification

In this section we show which combination of N and ϵ_n yields posterior convergence when the prior $\tilde{\Pi}$ in (1) is set to be a Dirichlet process location mixture of normal densities. Specifically, a density from $\tilde{\Pi}$ is given by

$$f_{F,\sigma}(y) = \int_{\mathbb{R}} \phi_\sigma(y - \mu) \mathrm{d}F(\mu), \quad F \sim \mathrm{DP}(\alpha F^*), \; \sigma \sim G, \tag{11}$$

where $\phi_\sigma(x)$ is the normal density with mean zero and variance σ^2, α is a positive constant, F^* is a probability distribution on $\mathbb{R}$ and G is a probability distribution on $\mathbb{R}_+$. Asymptotic properties of model (11) in density estimation have been extensively studied in Ghosal and van der Vaart (2001, 2007); Lijoi et al. (2005); Walker et al. (2007); Shen et al. (2013). The following result on prior rates is adapted from Theorem 4 of Shen et al. (2013). Let $f \in \tilde{\mathcal{F}}$ satisfy

(a1) $f \in \mathcal{C}^{\beta, L, \tau_0}(\mathbb{R})$;

(a2) $\int(|f^{(k)}(y)|/f(y))^{(2\beta+\epsilon)/k}f(y)\mathrm{d}y < \infty$, $k \le \lfloor\beta\rfloor$, and $\int(L(y)/f(y))^{(2\beta+\epsilon)/\beta}f(y)\mathrm{d}y < \infty$ for some $\epsilon > 0$;

(a3) f has exponentially decreasing tails.

As for the prior (11), let

(b1) F^* admits a positive density function on $\mathbb{R}$ with sub-Gaussian tails.

(b2) For the prior G, σ^{-2} has a gamma distribution.

Then, for some $C > 0$ and all sufficiently large m,

$$\tilde{\Pi}\big(B(\tilde{\epsilon}_m, f)\big) \ge \mathrm{e}^{-Cm\tilde{\epsilon}_m^2}, \quad \tilde{\epsilon}_m = m^{-\beta/(2+2\beta)}(\log m)^t, \tag{12}$$

for some positive constant t depending on the tails of f and on β. Note that $m^{-\beta/(2+2\beta)}$ is slower than the minimax rate for β-Hölder density, due to the use of the gamma prior on σ^{-2} instead of on σ^{-1}. In fact, the latter has too heavy tail behavior for Proposition 2.1 below to hold. When f itself is of mixture form, i.e. $f(y) = \int \phi_{\sigma_0}(y-\mu)\mathrm{d}F_0(\mu)$ for some σ_0 and F_0 with sub-Gaussian tails, Ghosal and van der Vaart (2001) have proven that (12) holds for $\tilde{\epsilon}_m = m^{-1/2}\log m$.

Finally, we state the following result which relies on entropy calculations in Shen et al. (2013) and on techniques in Walker et al. (2007). See the Appendix for a proof.

Proposition 2.1. *Under (b1)-(b2), there exists a family of subsets* $\{\tilde{\mathcal{F}}_m, m \ge 1\}$, $\tilde{\mathcal{F}}_m \uparrow \tilde{\mathcal{F}}$, *and Hellinger balls* $(\tilde{A}_{mi})_{i\ge1}$ *of radius* $\bar{\epsilon}_m$ *with* $\tilde{\mathcal{F}}_m \subseteq \bigcup_i \tilde{A}_{mi}$, *such that* (6) *and* (7) *hold with*

$$\bar{\epsilon}_m = m^{-\gamma}(\log m)^t$$

for any $\gamma \in (0, 1/2)$ *and* $t > 0$.

2.3 Convergence rates

Assume that, for each j (and n), $f_{0,j}$ satisfies (a1)-(a2)-(a3) so that (8) holds for $\tilde{\epsilon}_m = m^{-\beta/(2+2\beta)}(\log m)^t$. Now let $\epsilon_n = n^{-\eta}(\log n)^r$ and $N = n^\alpha$ with η, r, α positive constants to be determined later. The first condition in (9), $N\epsilon_n^2 \to \infty$, is implied by $\alpha - 2\eta > 0$ while the second condition in (9), $n\epsilon_n^4 \to \infty$, is satisfied by $1 - 4\eta \ge 0$, hence, putting them together we have

$$0 < 2\eta \le \min\{\alpha, 1/2\}. \tag{13}$$

The first condition in (10) is satisfied under (13) because of Proposition 2.1 as we can find $t > 0$ and $\gamma \in (0, 1/2)$ for $\bar{\epsilon}_{n/N} = \epsilon_n/2$ to hold. Finally, the second condition in (10), $N\tilde{\epsilon}_{n/N}^2 = o(\epsilon_n^2)$, holds for $r > t$ and

$$0 < 2\eta \le \frac{2\beta}{2+2\beta}(1-\alpha) - \alpha. \tag{14}$$

Note that we need the right hand side to be positive, hence $\alpha < \beta/(1+2\beta) < 1/2$. Putting inequalities (13) and (14) together we have $r > t$,

$$0 < 2\eta \leq \min\left\{\alpha, \frac{1}{2}, \frac{2\beta}{2+2\beta}(1-\alpha) - \alpha\right\}.$$

The solution to this problem which maximizes the value of η is given by $\eta = \frac{\alpha}{2}$, $\alpha = \frac{\beta}{2+3\beta}$, so that the posterior convergence rate is bounded by

$$\epsilon_n \asymp (\log n)^r n^{-\beta/2(2+3\beta)}.$$

Note that the posterior rate is not adaptive in that the number of partition sets $N = n^{\beta/(2+3\beta)}$ depends on β. In Norets and Pati (2014) and Shen and Ghosal (2016), convergence rates under the expected-L_1 norm have been derived under the assumption of β-Hölder smoothness of the conditional density $f(y|x)$ both in y and x. The rate is $n^{-\beta/(2\beta+d+1)}$ for d the dimension of the covariate, clearly faster than the one obtained above. In a classical setting, Efromovich (2007) found the minimax rate to be $n^{-\beta/(2\beta+2)}$ for $d = 1$ and under the L_2 norm on the product space $\mathbb{R} \times [0, 1]$. To the best of our knowledge, the minimax rate of convergence for conditional densities with respect to the sup-L_1 loss is not yet known for any suitable large class, and certainly not for the class of conditional densities considered here, but it may be reasonable to expect that it should be the same up to a log factor. So, while our rate appears "slow", it is to be remembered that this is with respect to the supremum L_1 norm and hence a benchmark has been set.

3 Proofs

In this section we proceed to the proof of Theorem 2.1. Write

$$\Pi_n(A_{\epsilon_n}|(Y,X)_{1:n}) = D_n^{-1} \int_{A_{\epsilon_n}} R_n(f)\mathrm{d}\Pi_n(f),$$

where $R_n(f) = \prod_{i=1}^n f(Y_i|X_i)/f_0(Y_i|X_i)$ and $D_n = \int_{\mathcal{F}} R_n(f)\,\Pi(\mathrm{d}f)$. As is customary in Bayesian asymptotics, we deal with the numerator and denominator separately. Let $\tilde{\mathcal{F}}_m$ be as in (6) and (7) such that $\tilde{\mathcal{F}}_m \uparrow \tilde{\mathcal{F}}$ as $m \to \infty$. They induce a sequence of increasing subsets of the space of conditional densities $\mathcal{F}$ given by

$$\mathcal{F}_n = \left\{f \in \mathcal{F} : f(y|x) = \textstyle\sum_{j=1}^N f_j(y)\mathbb{1}_{C_j}(x),\ f_j \in \tilde{\mathcal{F}}_{n/N}, j = 1, \ldots, N\right\} \tag{15}$$

It is sufficient to show that the posterior accumulates in $A_{\epsilon_n}^c \cap \mathcal{F}_n$ provided the prior probability $\tilde{\Pi}(\tilde{\mathcal{F}}_{n/N}^c)$ decreases sufficiently fast to 0 as $n \to \infty$. Reasoning as in Walker (2004); Walker et al. (2007), let (A_{il}) be a two-dimensional array of subsets of $\mathcal{F}_n$ such that $A_{\epsilon_n} \cap \mathcal{F}_n = \bigcup_{jl} A_{jl}$, and denote $L_{njl}^2 = \int_{A_{jl}} R_n(f)\Pi_n(\mathrm{d}f)$. The following result is easily proved.

Proposition 3.1. *Let $N \to \infty$ as $n \to \infty$ such that $N \log N = o(n\epsilon_n^2)$ and assume that, for some constants $c, C > 0$,*

$$\tilde{\Pi}_j(\tilde{\mathcal{F}}_{n/N}^c) \leq \exp\{-c(C+4)n\epsilon_n^2/N\}, \tag{16}$$

$$P_0\big(\textstyle\sum_{j,l} L_{njl} < \exp\{-c(C+2)n\epsilon_n^2/N\}\big) \to 1, \tag{17}$$

$$P_0\big(D_n \geq \exp\{-c(C+2)n\epsilon_n^2/N\}\big) \to 1. \tag{18}$$

Then $\Pi_n(A_{\epsilon_n}|(Y,X)_{1:n}) \to 0$ in P_0-probability.

Proof. Without loss of generality, we set $c = 1$. Reasoning as in the proof of Theorem 2.1 in Ghosal et al. (2000), by Fubini's theorem and the fact that $P_0(f/f_0) \leq 1$,

$$E_0\Big(\textstyle\int_{\mathcal{F}_n^c} R_n(f)\mathrm{d}\Pi_n(f)\Big) \leq \Pi_n(\mathcal{F}_n^c).$$

Next

$$\begin{aligned}\Pi_n(\mathcal{F}_n^c) &= 1 - \Pi_n(\mathcal{F}_n) = 1 - \textstyle\prod_{j=1}^N \tilde{\Pi}_j(\tilde{\mathcal{F}}_{n/N}) = 1 - \big(1 - \tilde{\Pi}(\tilde{\mathcal{F}}_{n/N}^c)\big)^N \\ &\leq 1 - \big(1 - \exp\{-(C+4)n\epsilon_n^2/N\}\big)^N \leq N\exp\{-(C+4)n\epsilon_n^2/N\}.\end{aligned}$$

Let A_n be the event that $D_n \geq \exp\{-(C+2)n\epsilon_n^2/N\}$. By (18), $P_0(A_n) \to 1$, then

$$\begin{aligned}E_0[\Pi_n(\mathcal{F}_n^c|(Y,X)_{1:n})] &\leq E_0[\Pi_n(\mathcal{F}_n^c|(Y,X)_{1:n})\mathbb{1}_{A_n}] + P_0(A_n^c) \\ &\leq N\exp\{-(C+4)n\epsilon_n^2/N\}\exp\{(C+2)n\epsilon_n^2/N\} + o(1) \\ &= \exp\{-2n\epsilon_n^2/N + \log N\} + o(1) \to 0\end{aligned}$$

for n sufficiently large since, by assumption, $\log N = o(n\epsilon_n^2/N)$ and $n\epsilon_n^2/N \to \infty$ as $n \to \infty$. Therefore it is sufficient to prove that $E_0[\Pi_n(A_{\epsilon_n} \cap \mathcal{F}_n|(Y,X)_{1:n})] \to 0$. Now let B_n be the event that $\sum_{j,l} L_{njl} < \exp\{-(C+2)n\epsilon_n^2/N\}$. By using the inequality

$$\Pi_n(A_{\epsilon_n} \cap \mathcal{F}_n|(Y,X)_{1:n}) \leq D_n^{-1/2}\textstyle\sum_{j,l} L_{njl},$$

and $P_0(B_n) \to 1$, cfr (17), it follows that

$$\begin{aligned}&E_0[\Pi_n(A_{\epsilon_n} \cap \mathcal{F}_n|(Y,X)_{1:n})] \\ &\quad\leq E_0[\Pi_n(A_{\epsilon_n} \cap \mathcal{F}_n|(Y,X)_{1:n})\mathbb{1}_{A_n \cap B_n}] + P_0(A_n^c \cup B_n^c) \\ &\quad\leq E_0[D_n^{-1/2}\textstyle\sum_{j,l} L_{njl}\mathbb{1}_{A_n \cap B_n}] + o(1) \\ &\quad\leq \exp\{-(C+2)n\epsilon_n^2/N\}\exp\{(C+2)n\epsilon_n^2/(2N)\} + o(1) \\ &\quad= \exp\{-(C+2)n\epsilon_n^2/(2N)\} + o(1) \to 0.\end{aligned}$$

The proof is then complete. □

In order to prove Theorem 2.1, we proceed to the verification of the conditions of Proposition 3.1 under the hypothesis made. We start with (16) and (17). Recalling that the Hellinger and the L_1 distances induce equivalent topologies in $\tilde{\mathcal{F}}$, without loss of generality, we replace the L_1 norm in (3) with the Hellinger distance and define

$$A_{\epsilon_n} = \left\{ f \in \mathcal{F} : \sup_{x \in \mathbb{X}} H(f_0(\cdot|x), f(\cdot|x)) > \epsilon_n \right\}$$

with $\epsilon_n \to 0$ such that $n\epsilon_n^2 \to \infty$. According to (15), under model (1) we have that $A_{\epsilon_n} \cap \mathcal{F}_n = \bigcup_{j=1}^N A_j$ where $A_j = \{f(y|x) = \sum_{j=1}^N f_j(y)\mathbb{1}_{C_j}(x) : f_j \in \tilde{A}_j\}$ and

$$\tilde{A}_j = \left\{ f \in \tilde{\mathcal{F}}_{n/N} : \sup_{x \in C_j} H(f_0(\cdot|x), f) > \epsilon_n \right\}.$$

For each $j = 1, \ldots, N$, we can further cover $\tilde{A}_j$ into Hellinger balls of radius $\epsilon_n/2$ and centered on $f_{jl} \in \tilde{A}_j$,

$$\tilde{A}_{jl} = \left\{ f \in \tilde{A}_j : H(f, f_{jl}) < \epsilon_n/2 \right\}.$$

so $A_j \subset \bigcup_{l \geq 1} A_{jl}$ where $A_{jl} = \{f(y|x) = \sum_{j=1}^N f_j(y)\mathbb{1}_{C_j}(x) : \ f_j \in \tilde{A}_{jl}\}$. Now consider $L_{njl}^2 = \int_{A_{jl}} R_n(f)\Pi_n(\mathrm{d}f)$, from which we have

$$E_0\left(L_{njl} | (Y,X)_{1:n-1}, X_n \right) \leq L_{n-1jl}\left\{ 1 - \inf_{f_j \in \tilde{A}_{jl}} \frac{1}{2} H^2\big(f_0(\cdot|X_n), f_j\big)\mathbb{1}_{C_j}(X_n) \right\}.$$

A lower bound for $\inf_{x \in C_j} H^2(f_0(\cdot|x), f_j)$ is readily derived by using assumption (5). By (2) and $H^2(f,g) \leq \int f \log(f/g)$,

$$\sup_{x,x' \in C_j} H(f_0(\cdot|x), f_0(\cdot|x')) \leq \sqrt{L}/N, \quad \forall j \tag{19}$$

for n (and N) large enough. Now define $x' \in C_j$ as the x value which maximizes $H(f_0(\cdot|x), f_{jl}(\cdot))$, i.e.

$$x' = \arg\max_{x \in C_j} H(f_0(\cdot|x), f_{jl}(\cdot)).$$

Such a maximum exists since $x \mapsto f(\cdot|x)$ is Hellinger continuous by (5), and C_j can be taken as a closed interval in $[0,1]$. Since $f_{jl} \in \tilde{A}_j$, $H(f_0(\cdot|x'), f_{jl}(\cdot)) > \epsilon_n$ and so for $x \in C_j$ and $f_j \in \tilde{A}_{jl}$, we have

$$\begin{aligned} H(f_0(\cdot|x), f_j(\cdot)) &\geq H(f_0(\cdot|x'), f_j(\cdot)) - H(f_0(\cdot|x), f_0(\cdot|x')) \\ &\geq H(f_0(\cdot|x'), f_{jl}(\cdot)) - H(f_j(\cdot), f_{jl}(\cdot)) - \sqrt{L}/N \\ &\geq \epsilon_n - \epsilon_n/2 - \sqrt{L}/N = \epsilon_n/2 - \sqrt{L}/N \end{aligned}$$

using a further application of the triangle inequality. Thus, conditioning on the sample size $n_j = \sum_{i=1}^n \mathbb{1}_{C_j}(X_i)$,

$$E_0(L_{njl}|n_j) \le \left([1 - \tfrac{1}{2}(\epsilon_n/2 - \sqrt{L}/N)^2\right)^{n_j} \Pi_n(A_{jl})^{1/2}.$$

Since $X_1, \ldots, X_n$ is an i.i.d. sample from $q(x)$, $n_j \sim \text{binom}(n, Q(C_j))$. It is easy to check, by using the formula of the probability generating function of the binomial distribution, that

$$\begin{aligned} E_0\Big\{\left(1 - \tfrac{1}{2}(\epsilon_n/2 + \sqrt{L}/N)^2\right)^{n_j}\Big\} &= \Big\{1 - \tfrac{1}{2}(\epsilon_n/2 - \sqrt{L}/N)^2 Q(C_j)\Big\}^n \\ &\le \exp\big\{ - \tfrac{1}{2}(\epsilon_n/2 - \sqrt{L}/N)^2 n Q(C_j)\big\}, \end{aligned}$$

where the last inequality holds since $\log(1-x) < -x$. Hence

$$E_0(L_{njl}) \le \exp\big\{ - \tfrac{1}{2}(\epsilon_n/2 - \sqrt{L}/N)^2 n Q(C_j)\big\} \Pi_n(A_{jl})^{1/2}.$$

The first condition in (9) implies that for n (and N) sufficiently large, $\epsilon_n/2 - \sqrt{L}/N > \epsilon_n/4$. Also, under (2) and (4),

$$Q(C_j) \ge q/N, \;\; q := \inf_{x \in [0,1]} q(x)$$

so that

$$E_0(L_{njl}) \le \exp\{-(\epsilon_n^2/32) qn/N\} \Pi_n(A_{jl})^{1/2}.$$

It follows that, for any $d > 0$,

$$P_0\Big(\sum_{j=1}^N \sum_{l \ge 1} L_{njl} > \mathrm{e}^{-d\frac{qn}{N}}\Big) \le \exp\Big\{ - (\epsilon_n^2/32 - d)\frac{qn}{N}\Big\} \sum_{j=1}^N \sum_{l \ge 1} \Pi_n(A_{jl})^{1/2}.$$

Consider now that the $\tilde{A}_{jl}$ can be the same sets for each j so to form a covering $\{\tilde{A}_l\}_{l \ge 1}$ of $\tilde{\mathcal{F}}_{n/N}$ in terms of Hellinger balls of radius $\epsilon_n/2$. Hence $\Pi_n(A_{jl}) = \tilde{\Pi}(\tilde{A}_l)$. We then have

$$P_0\Big(\sum_{j=1}^N \sum_{l \ge 1} L_{njl} > \mathrm{e}^{-d\frac{qn}{N}}\Big) \le N \exp\Big\{ - (\epsilon_n^2/32 - d)\frac{qn}{N}\Big\} \sum_{l \ge 1} \tilde{\Pi}(\tilde{A}_l)^{1/2}. \tag{20}$$

Hence, taking $d = \epsilon_n^2/64$ in (20),

$$P_0\Big(\sum_{j=1}^N \sum_{l \ge 1} L_{njl} > \exp\Big\{ - \frac{\epsilon_n^2}{64}\frac{qn}{N}\Big\}\Big) \le N \exp\Big\{ - \frac{\epsilon_n^2}{64}\frac{qn}{N}\Big\} \sum_{l \ge 1} \tilde{\Pi}(\tilde{A}_l)^{1/2}.$$

Set $c = q/(64(C+2))$ in (17) for C to be determined later. For $m = n/N$ and the first condition in (10), (6) implies that $\tilde{\Pi}(\tilde{\mathcal{F}}^c_{n/N}) \lesssim \exp\{-(C'+4)n\epsilon_n^2/(4N)\}$ for any C', so that C' can be chosen to have (16) satisfied

for c and C above. Also, $\sum_{l\geq 1}\tilde{\Pi}(\tilde{A}_l)^{1/2}) = o(\mathrm{e}^{(q/64)n\epsilon_n^2/N})$ by the (7), and $N\log N = o(n\epsilon_n^2)$ by condition (9) as long as $N^2 = o(n)$, cfr Section 2.3. Hence (17) holds.

We now aim at establishing that (18) of Proposition 3.1 holds for the same C and c found before. To begin with, recall the definition of $f_{0,j}(y)$ as the marginal density of Y when X is restricted to lie in C_j, and let $P_{0,j}$ be the probability distribution associated to $f_{0,j}(y)$. Recall also that $n_j = \sum_{i=1}^n \mathbb{1}_{C_j}(X_i)$ and, using the notation $\mathcal{I}_j = \{i : X_i \in C_j\}$, we have $n_j = \#(\mathcal{I}_j)$, so we write

$$R_n(f) = \exp\Big\{ -\sum_{j=1}^N \sum_{i\in\mathcal{I}_j} \log \frac{f_0(Y_i|X_i)}{f(Y_i|X_i)} \Big\}$$
$$= \exp\Big\{ -\sum_{j=1}^N \sum_{i\in\mathcal{I}_j} \log \frac{f_0(Y_i|X_i)}{f_{0,j}(Y_i)} - \sum_{j=1}^N \sum_{i\in\mathcal{I}_j} \log \frac{f_{0,j}(Y_i)}{f_j(Y_i)} \Big\}.$$

Hence D_n, the denominator of $\Pi_n(A_{\epsilon_n}|(Y,X)_{1:n})$, is given by

$$D_n = \exp\Big\{ -\sum_{j=1}^N \sum_{i\in\mathcal{I}_j} \log \frac{f_0(Y_i|X_i)}{f_{0,j}(Y_i)} \Big\} \prod_{j=1}^N \int_{\tilde{\mathcal{F}}} \prod_{i\in\mathcal{I}_j} \frac{f(Y_i)}{f_{0,j}(Y_i)}\, \tilde{\Pi}_j(\mathrm{d}f) \qquad (21)$$

where we have made use of the independence among the N priors $\tilde{\Pi}_j$. We need to deal with the two parts of (21) separately. As for the term inside the curvy brackets, a key ingredient is the control on the Kullback-Leibler divergence between neighboring conditional densities in (5), see Lemma .1 in the Appendix for an intermediate result. Lemma .1 allows to establish the rate at which $n^{-1}\sum_{j=1}^N \sum_{i\in\mathcal{I}_j} \log f_0(y_i|x_i)/f_{0,j}(y_i)$ goes to zero, as stated in the following proposition.

Proposition 3.2. *Under (5), for d_n and N such that $nd_n^2 \to \infty$ and $d_n N \to \infty$,*

$$P_0\Big\{ \sum_{j=1}^N \sum_{i\in\mathcal{I}_j} \log \frac{f_0(Y_i|X_i)}{f_{0,j}(Y_i)} < \frac{d_n n}{N} \Big\} \to 1. \qquad (22)$$

See the Appendix for a proof. We now deal with the second term in (21). Note that $\{Y_i : i \in \mathcal{I}_j\}$ can be considered as i.i.d. replicates from $f_{0,j}$, the marginal density of Y when X is restricted to C_j. We next rely on prior rate $\tilde{\epsilon}_m$ of $\tilde{\Pi}(\mathrm{d}f)$ at $f_{0,j}$ in (8).

Proposition 3.3. *Under (8), as $n \to \infty$ and $\delta > 0$,*

$$P_0\Big\{ \prod_{j=1}^N \int_{\tilde{\mathcal{F}}} \prod_{i\in\mathcal{I}_j} \frac{f(Y_i)}{f_{0,j}(Y_i)}\, \tilde{\Pi}_j(\mathrm{d}f) \geq \exp\Big(-(C'+1+\delta)n\tilde{\epsilon}_{n/N}^2 \Big) \Big\} \to 1. \quad (23)$$

See the Appendix for a proof. Putting Propositions 3.2 and 3.3 together we obtain that

$$P_0\Big\{D_n \geq \exp\Big(-d_n n/N - (C'+1+\delta)n\tilde{\epsilon}_{n/N}^2\Big)\Big\} \to 1$$

for any $\delta > 0$, d_n and N such that $d_n N \to \infty$ and $nd_n^2 \to \infty$. Hence, for (18) to be satisfied with $C = C'$ and $c^{-1} = 64(C+2)$, we need

$$N\tilde{\epsilon}_{n/N}^2 \leq \epsilon_n^2/64(C+2)$$

for sufficiently large N upon setting $d_n = (1-\delta)\epsilon_n^2/64(C+2)$. This is implied by (10). Also the hypothesis of Proposition 3.2 are satisfied for this choice of d_n because of the two conditions in (9). The proof is then complete.

4 Discussion

4.1 Control on the Kulback-Leibler divergence between neighboring conditional densities

Here we provide two examples assuming forms for $f_0(y|x)$ that satisfy (5).

EXAMPLE 1. Assume that the true conditional density corresponds to a regression,

$$Y = g(X) + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

with known variance, say $\sigma = 1$. Then $f_0(y|x) = \phi(y - g_0(x))$. Assume that g_0 is (locally) Lipschitz, $|g_0(x) - g_0(x')| \leq l|x - x'|$ for some $l > 0$. Since

$$\int \phi(y-\mu_1)\log\frac{\phi(y-\mu_1)}{\phi(y-\mu_2)}\mathrm{d}y = \frac{(\mu_1-\mu_2)^2}{2},$$

$$\text{and} \quad \int \phi(y-\mu_1)\left(\log\frac{\phi(y-\mu_1)}{\phi(y-\mu_2)}\right)^2\mathrm{d}y = \frac{(\mu_1-\mu_2)^4}{4} + (\mu_1-\mu_2)^2,$$

the Lipschitz condition on $g_0(x)$ implies

$$\int f_0(y|x)\log\frac{f_0(y|x)}{f_0(y|x')}\mathrm{d}y \leq \frac{l^2}{2}|x-x'|^2$$

and

$$\int f_0(y|x)\left(\log\frac{f_0(y|x)}{f_0(y|x')}\right)^2\mathrm{d}y < \frac{l^4}{4}|x-x'|^4 + l^2|x-x'|^2,$$

so that Assumption (5) is satisfied for $L = l^4/4 + l^2$ for $|x - x'| < 1$. □

EXAMPLE 2. Here we consider the true conditional density as a mixture of normal densities with predictor-dependent weights given by

$$f_0(y|x) = \sum_{k=1}^{M} w_k(x)\,\phi_\sigma(y-\mu_k)$$

where M can be infinity and $\sum_{j=1}^{M} w_j(x) = 1$ for any x. Then the marginal density of Y when X is restricted to lie in C_j is

$$f_{0,j}(y) = \sum_{k=1}^{M} w_{jk}\, \phi_\sigma(y - \mu_k), \quad w_{jk} = \frac{1}{Q(C_j)} \int_{C_j} w_k(x)\, q(x) \mathrm{d}x,$$

so that the nearly parametric prior rate $\tilde{\epsilon}_m = m^{-1/2} \log m$ is achieved by the prior (11) of Section 2.2.

Our aim is to confirm Assumption (5), or to find conditions under which it holds. Thus we require

$$T = \int \frac{\big(f_0(y|x) - f_0(y|x')\big)^2}{f_0(y|x')} \mathrm{d}y \le L|x - x'|^2$$

for $|x - x'|$ small for some universal constant L. In fact, T is an upper bound for the left hand side of (5) for both $r = 1$ and $r = 2$ (use simple algebra together with $\log z \le z - 1$ and $4(\log z)^2 \le (1/z - z)^2$). Now

$$f_0(y|x') - f_0(y|x) = \sum_{k=1}^{M} w_k(x')\, \phi_\sigma(y - \mu_k)\, \{1 - w_k(x)/w_k(x')\}$$

and so if, for some $c > 0$,

$$\sup_k |1 - w_k(x')/w_k(x)| \le c|x - x'| \tag{24}$$

then $\big(f_0(y|x) - f_0(y|x')\big)^2 < c^2\, |x - x'|^2\, f_0(y|x')^2$ and so $T < L\, |x - x'|^2$ for $L = c^2$. If M is finite, then a weaker condition is sufficient. In fact, if, for some $c > 0$,

$$\sum_{k=1}^{M} w_k(x)\, \{1 - w_k(x')/w_k(x)\}^2 < c\, |x - x'|^2 \tag{25}$$

then, by using Cauchy-Schwartz inequality,

$$T < c\, |x - x'|^2 \int \frac{\sum_{k=1}^{M} w_k(x')\, [\phi_\sigma(y - \mu_k)]^2}{\sum_{k=1}^{M} w_k(x')\, \phi_\sigma(y - \mu_k)} \mathrm{d}y < c\, M\, |x - x'|^2$$

and so $T < C\, |x - x'|^2$ for $C = c\, M$.

In summary, if $M = \infty$ we require (24); whereas if $M < \infty$ then we need (25). Let us investigate the former as it is more stringent. A general form for normalized weights is given by

$$w_k(x) = \Lambda(x)^{-1} w_k\, H_k(x)$$

where

$$\sum_{k \ge 1} w_k = 1, \quad H_k(x) = \exp\{-\phi|x - z_k|\} \quad \text{and} \quad \Lambda(x) = \sum_{k \ge 1} w_k\, H_k(x)$$

for some sequence $(z_k)_{k \ge 1} \in (0, 1)$ and some $\phi > 0$. Then it is straightforward to show that, for all k, x and x', with $|x - x'| \le 1/N$,

$$\mathrm{e}^{-3\phi|x - x'|} \le w_k(x')/w_k(x) \le e^{3\phi|x - x'|}$$

and hence for some constant $c > 0$, $\sup_k |1 - w_k(x')/w_k(x)| < c\, |x - x'|$ as required. □

4.2 Covariate distribution

Here we discuss the assumption (4) of $q(x)$ bounded away from 0. Allowing the density to tend to 0, for example at the boundary of $[0,1]$ would be an interesting extension. It is not difficult to check that the same posterior convergence rate in the sup-L_1 norm of Theorem 2.1 will hold true by re-defining $\sup_{x\in[0,1]}$ to $\sup_{x\in D}$, where $D = \{x : q(x) > q\}$ for some arbitrarily small $q > 0$. However this would require some previous knowledge of the covariate distribution. In practice, one option is to set the partition sets C_j in a data driven way such that $n_j = \sum_{i=1}^n \mathbb{1}_{C_j}(X_i) \asymp n/N$ as $n \to \infty$, e.g. by using an empirical estimate $\mathbb{Q}_n$ of the covariate distribution. This would work fine with the proof of (16) and (17) in Section 3 but in the use of assumption (5) to establish the bound in (19). To illustrate the point, if $q(x) \sim x^\tau$ as $x \to 0$ for $\tau > 0$ and C_j is set such that $\mathbb{Q}_n(C_j) = 1/N$, then it is not difficult to show that $|C_1| \asymp 1/N^{1/(1+\tau)}$ as $n \to \infty$, in contrast with (2), so that the upper bound in (19) would be $1/N^{1/(1+\tau)}$ instead of $1/\sqrt{N}$. A close inspection of the arguments used in the proof of Theorem 2.1 reveals that the first condition in (9) should be replaced by $N\epsilon_n^{2\wedge(1+\tau)}$, which, in turn, would yield a worse convergence rate ϵ_n when $\tau > 1$, cfr. calculations in Section 2.3. This question is of interest and left for future work.

4.3 Alternative derivation of posterior convergence rates

An associate editor, whom we thank for the suggestion, has asked whether an alternative strategy would work for deriving posterior rates in the sup-L_1 norm from rates in the integrated L_1 norm. The idea is to use the representation of the conditional density $f(y|x)$ as a function of x in the Haar basis of $L^2[0,1]$. To set the notation, define $\phi(x) = \mathbb{1}_{(0,1)}(x)$, $\psi(x) := \psi_{0,0}(x) = \mathbb{1}_{(0,1/2)}(x) - \mathbb{1}_{(1/2,1)}(x)$, $\phi_{\ell,k}(x) = 2^{\ell/2}\phi(2^\ell x - k)$ and $\psi_{\ell,k}(x) = 2^{\ell/2}\psi(2^\ell x - k)$ for any integer ℓ and $0 \le k < 2^\ell$. Consider the regular dyadic partition of $[0,1]$ given by intervals $C_{nk} = (k2^{-L_n}, (k+1)2^{-L_n})$ so that $N_n = 2^{L_n}$. For $g \in L^2[0,1]$, let $K_j(g)$ be the orthogonal projection of g onto the subspace generated by the linear span of $\{\phi_{j,k},\, 0 \le k < 2^j\}$. By construction, the conditional density $f(y|x)$ in (1) coincides with $K_{L_n}(f(y|\cdot))$ so that, for any y,

$$f(y|x) - f_0(y|x) = \sum_{k=0}^{2^{L_n}-1} \langle f(y|\cdot) - f_0(y|\cdot), \phi_{L_n,k}\rangle \phi_{L_n,k}(x) \\ - \sum_{\ell \ge L_n} \sum_{k=0}^{2^\ell - 1} \langle f_0(y|\cdot), \psi_{\ell,k}\rangle \psi_{\ell,k}(x)$$

where $\langle\cdot,\cdot\rangle$ is the inner product in $L^2[0,1]$. By the localization property of the Haar basis, $\|\sum_k \phi_{\ell,k}\|_\infty = \|\sum_k \psi_{\ell,k}\|_\infty = 2^{\ell/2}$, and by standard

arguments one obtains the bound

$$\|f(y|\cdot) - f_0(y|\cdot)\|_\infty \leq 2^{L_n/2}\|f(y|\cdot) - f_0(y|\cdot)\|_2 + R_n(y),$$

where $R_n(y) = \sum_{\ell \geq L_n} 2^{\ell/2} \max_k |\langle f_0(y|\cdot), \psi_{\ell,k}\rangle|$ is related to the approximation property of projection kernel estimate $K_{L_n}(f_0(y|\cdot))$. Consider now the sup-L_1 norm. Exchange the sup with the integral sign to get

$$\|f(\cdot|x) - f_0(\cdot|x)\|_{1,\infty} \leq \int_{\mathbb{R}} \|f(y|\cdot) - f_0(y|\cdot)\|_\infty \mathrm{d}y$$

by an application of Minkowski inequality for integrals. Then, the bound above yields

$$\|f(\cdot|x) - f_0(\cdot|x)\|_{1,\infty} \leq 2^{L_n/2} \int_{\mathbb{R}} \|f(y|\cdot) - f_0(y|\cdot)\|_2 \mathrm{d}y + \int_{\mathbb{R}} R_n(y)\mathrm{d}y$$

In the case that the conditional densities are uniformly bounded in x, one can rely on an inequality between L_2 and L_1 norms to get, for a positive constant c,

$$\begin{aligned}\|f(\cdot|x) - f_0(\cdot|x)\|_{1,\infty} &\leq c2^{L_n/2} \int_{\mathbb{R}} \|f(y|\cdot) - f_0(y|\cdot)\|_1 \mathrm{d}y + \int_{\mathbb{R}} R_n(y)\mathrm{d}y \\ &= c2^{L_n/2} \int_0^1 \|f(\cdot|x) - f_0(\cdot|x)\|_1 \mathrm{d}x + \int_{\mathbb{R}} R_n(y)\mathrm{d}y\end{aligned}$$

that is the sup-L_1 norm is bounded by $2^{L_n/2}$ times the integrated L_1 norm plus an approximation term that depends on f_0. If $f_0(y|x)$ is Hölder smooth of level β in x, then

$$\sup_{\ell,k} 2^{\ell(1/2+\beta\wedge 1)} |\langle f_0(y|\cdot), \psi_{\ell,k}\rangle| < \infty,$$

so that $R_n(y) \lesssim 2^{-(\beta\wedge 1)L_n}$. If one further assumes that the bound above depends on y, say $R(y)$, and that $\int_{\mathbb{R}} R(y)\mathrm{d}y < \infty$, then a posterior convergence rate ϵ_n in the integrated L_1 norm would imply a posterior convergence rate $2^{L_n/2}\epsilon_n \vee 2^{-(\beta\wedge 1)L_n}$ in the sup-L_1 norm. Note that the poor approximation properties of the Haar basis for very smooth functions pose a limit to the rate that can be achieved. Still, it is of interest to investigate whether such a rate could improve upon the rate obtained in Section 2.3 for some regularities. This will be studied elsewhere.

Appendix

Proof of Proposition 2.1. Define the entropy of $\mathcal{G} \subset \tilde{\mathcal{F}}$ with respect to the metric d to be $\log N(\epsilon, \mathcal{G}, d)$ where $N(\epsilon, \mathcal{G}, d)$ is the minimum integer N for which there exists $f_1, \ldots, f_N \in \tilde{\mathcal{F}}$ such that $\mathcal{G} \subset \bigcup_{j=1}^N \{f : d(f, f_j) < \epsilon\}$. By hypothesis, the prior measure F^* satisfies $F^*[-a, a]^c \lesssim \mathrm{e}^{-ba^\tau}$ for some

$b > 0$, $\tau > 2$ and $|a|$ sufficiently large. Let G be the distribution of the square root of an inverse gamma random variable with shape parameter c_3 and rate parameter c_1. Define

$$\mathcal{F}_{a,\underline{\sigma},\bar{\sigma}} = \{f_{F,\sigma} : \ F[-a,a] = 1,\ \underline{\sigma} \le \sigma \le \bar{\sigma}\}$$
$$\mathcal{F}_{a,\eta,\underline{\sigma},\bar{\sigma}} = \{f_{F,\sigma} : \ F[-a,a] \ge 1-\eta,\ \underline{\sigma} \le \sigma \le \bar{\sigma}\}$$

Combining Lemma A.3 in Ghosal and van der Vaart (2001) and Lemma 3 in Ghosal and van der Vaart (2007),

$$\begin{aligned}\log N(\eta, \mathcal{F}_{a,\eta/3,\underline{\sigma},\bar{\sigma}}, \|\cdot\|_1) &\le \log N(\eta, \mathcal{F}_{a,\underline{\sigma},\bar{\sigma}}, \|\cdot\|_1) \\ &\lesssim \log\left(\frac{\bar{\sigma}}{\eta\underline{\sigma}}\right) + \left(\frac{a}{\underline{\sigma}} \vee 1\right)\left(\log\frac{1}{\eta}\right)\left[\log\left(\frac{a}{\eta\underline{\sigma}} + 1\right) + \log\frac{1}{\eta}\right].\end{aligned}$$

For each n, let $\sigma_n = (n\epsilon_n^2)^{-1/2}$, and $a_n = \sigma_n^{-1}(\log n)^{-3}$. Define

$$\begin{aligned}\mathcal{B}_{n,0} &= \{f_{F,\sigma} : \ F[-a_n,a_n] \ge 1-\epsilon_n^2/3,\ \sigma_n \le \sigma \le \sigma_n(1+\epsilon_n^2)^n\}, \\ \mathcal{B}_{n,j} &= \{f_{F,\sigma} : \ F[-(j+1)a_n,(j+1)a_n] \ge 1-\epsilon_n^2/3, \\ &\qquad F[-ja_n,ja_n] < 1-\epsilon_n^2/3,\ \sigma_n \le \sigma \le \sigma_n(1+\epsilon_n^2)^n\}, \quad j \ge 1 \\ \tilde{\mathcal{F}}_n &= \{f_{F,\sigma} : \ \sigma_n \le \sigma \le \sigma_n(1+\epsilon_n^2)^n\}\end{aligned}$$

It is clear that $\tilde{\mathcal{F}}_n \uparrow \tilde{\mathcal{F}}$ as $n \to \infty$ and $\tilde{\mathcal{F}}_n = \bigcup_{j\ge 0}\mathcal{B}_{n,j}$. By standard calculations,

$$\begin{aligned}\Pi(\tilde{\mathcal{F}}_n^c) &= G(\sigma < \sigma_n) + G(\sigma > \sigma_n(1+\epsilon_n^2)^n) \\ &\lesssim \mathrm{e}^{-c_1\sigma_n^{-2}/2} + \sigma_n^{-2c_3}(1+\epsilon_n^2)^{-2c_3 n}, \\ &= \exp\{-\tfrac{c_1}{2}n\epsilon_n^2\} + (n\epsilon_n^2)^{c_3}\exp\{-2c_3 n\log(1+\epsilon_n^2)\} \\ &\le \exp\{-\tfrac{c_1}{2}n\epsilon_n^2\} + (n\epsilon_n^2)^{c_3}\exp\{-c_3 n\epsilon_n^2\} \\ &\lesssim \exp\{-(C+4)n\epsilon_n^2\}\end{aligned}$$

for some $C > 0$ by choosing c_1 and c_3 sufficiently large. Next, define

$$\mathcal{B}_{n,0,k} = \{f_{F,\sigma}\ F[-a_n,a_n] \ge 1-\epsilon_n^2/3,\ \sigma_n(1+\epsilon_n^2)^{k-1} \le \sigma \le \sigma_n(1+\epsilon_n^2)^k\}$$

so that $\mathcal{B}_{n,0} = \bigcup_{k=1}^n \mathcal{B}_{n,0,k}$. Finally, let $K_{n,\epsilon_n} := \sum_{i\ge 1}\Pi(A_{ni})^{1/2}$ for $(A_{n,i})_{i\ge 1}$ the Hellinger balls of radius ϵ_n that covers $\tilde{\mathcal{F}}_n$. Following Walker et al. (2007),

$$K_{n,\epsilon_n} \le \sum_{k=1}^n N(\epsilon_n, \mathcal{B}_{n,0,k}, H)\Pi(\mathcal{B}_{n,0,k})^{\frac{1}{2}} + \sum_{j\ge 1} N(\epsilon_n, \mathcal{B}_{n,j}, H)\Pi(\mathcal{B}_{n,j})^{\frac{1}{2}}. \quad (26)$$

The goal is to show that the two sums in the right hand sides do not grow to ∞ faster than $\mathrm{e}^{cn\epsilon_n^2}$ for any $c > 0$. As for the second sum in (26), because of the inequality $H(f,g)^2 \le \|f-g\|_1$, $N(\epsilon_n, \mathcal{B}, H) \le N(\epsilon_n^2, \mathcal{B}, \|\cdot\|_1)$, so

that, for $j \geq 1$, $\mathcal{B}_{n,j} \subset \mathcal{F}_{(j+1)a_n, \epsilon_n^2/3, \sigma_n, \sigma_n(1+\epsilon_n^2)^n}$ and the entropy calculations above yield

$$\begin{aligned}
&\log N(\epsilon_n, \mathcal{B}_{n,j}, H) \\
&\lesssim \log \frac{3(1+\epsilon_n^2)^n}{\epsilon_n^2} + \left(\frac{(j+1)a_n}{\sigma_n} \vee 1 \right) \left(\log \frac{3}{\epsilon_n^2} \right) \left(\log \left(\frac{3(j+1)a_n}{\epsilon_n^2 \sigma_n} + 1 \right) + \log \frac{3}{\epsilon_n^2} \right) \\
&\lesssim \log \frac{(1+\epsilon_n^2)^n}{\epsilon_n^2} + \frac{(j+1)a_n}{\sigma_n} \left(\log \frac{(j+1)a_n}{\epsilon_n^2 \sigma_n} \right)^2.
\end{aligned}$$

An upper bound on the prior probability $\Pi(\mathcal{B}_{n,j})$, $j \geq 1$, is given by

$$\begin{aligned}
\Pi\big(F([-ja_n, ja_n]^c) > \epsilon_n^2/3\big) &\leq \frac{3}{\epsilon_n^2} \mathrm{E}\big(F([-ja_n, ja_n]^c)\big) \\
&= \frac{3}{\epsilon_n^2} F^*([-ja_n, ja_n]^c) \lesssim \frac{3}{\epsilon_n^2} \mathrm{e}^{-b(ja_n)^\tau}
\end{aligned}$$

where we used Markov inequality and the assumption on the tail of F^*. Hence, for some large constant $C' > 0$,

$$\begin{aligned}
&\sum_{j \geq 1} N(\epsilon_n, \mathcal{B}_{n,j}, H) \Pi(\mathcal{B}_{n,j})^{1/2} \\
&\leq \sum_{j \geq 1} \exp \left\{ C' \left[\log \frac{(1+\epsilon_n^2)^n}{\epsilon_n^2} + \frac{(j+1)a_n}{\sigma_n} \left(\log \frac{(j+1)a_n}{\epsilon_n^2 \sigma_n} \right)^2 \right] + \frac{1}{2} \log \frac{1}{\epsilon_n^2} - \frac{1}{2} b(ja_n)^\tau \right\} \\
&\leq \frac{\mathrm{e}^{C' n \epsilon_n^2}}{\epsilon_n^{2C'+1}} \sum_{j \geq 1} \exp \left\{ C' \frac{(j+1)a_n}{\sigma_n} \left(\log \frac{(j+1)a_n}{\epsilon_n^2 \sigma_n} \right)^2 - \frac{1}{2} b(ja_n)^\tau \right\}
\end{aligned}$$

Recalling the definition of σ_n and a_n, we have

$$\begin{aligned}
&\sum_{j \geq 1} N(\epsilon_n, \mathcal{B}_{n,j}, H) \Pi(\mathcal{B}_{n,j})^{1/2} \\
&\leq \frac{\mathrm{e}^{C' n \epsilon_n^2}}{\epsilon_n^{2C'+1}} \sum_{j \geq 1} \exp \left\{ C'(j+1)\sigma_n^{-2} (\log n)^{-3} \left(\log \frac{(j+1)\sigma_n^{-2}(\log n)^{-3}}{\epsilon_n^2} \right)^2 - \frac{1}{2} b(ja_n)^\tau \right\} \\
&= \frac{\mathrm{e}^{C' n \epsilon_n^2}}{\epsilon_n^{2C'+1}} \sum_{j \geq 1} \exp \left\{ C'(j+1)a_n^2 (\log n)^3 \left\{ \log(j+1) + \log[n(\log n)^{-3}] \right\}^2 - \frac{1}{2} b(ja_n)^\tau \right\} \\
&\leq \frac{\mathrm{e}^{C' n \epsilon_n^2}}{\epsilon_n^{2C'+1}} \sum_{j \geq 1} \exp \left\{ 2C'(j+1)a_n^2 (\log n)^3 \left[\log^2(j+1) + (\log n)^2 \right] - \frac{1}{2} b(ja_n)^\tau \right\}
\end{aligned}$$

where we used $(a+b)^2 \leq 2(a^2+b^2)$,

$$\leq \frac{\mathrm{e}^{C' n \epsilon_n^2}}{\epsilon_n^{2C'+1}} \sum_{j \geq 1} \exp \left\{ 2C' \left[(j+1) \log^2(j+1) a_n^2 (\log n)^3 + (j+1) a_n^2 (\log n)^5 - \frac{b}{4C'} (ja_n)^\tau \right] \right\}$$

Since $\tau > 2$, a_n^τ grows faster than $n\epsilon_n^2$ and $a_n^2(\log n)^5$, so that the first summands vanish to 0. We consider next the sum for $j \geq J$ and some J sufficiently large.

$$\begin{aligned}
&\frac{\mathrm{e}^{C'n\epsilon_n^2}}{\epsilon_n^{2C'+1}} \sum_{j\geq J} \exp\left\{2C'\left[(j+1)\log^2(j+1)a_n^2(\log n)^3 + (j+1)a_n^2(\log n)^5 - \frac{b}{4C'}(ja_n)^\tau\right]\right\} \\
\leq &\frac{\mathrm{e}^{C'n\epsilon_n^2}}{\epsilon_n^{2C'+1}} \sum_{j\geq J} \exp\left\{2C'\left[(1+\delta)(j+1)\log^2(j+1)a_n^2(\log n)^5 - \frac{b}{4C'}(ja_n)^\tau\right]\right\}
\end{aligned}$$

where we set $\delta = 1/\log^2(J+1)$,

$$\leq \frac{\mathrm{e}^{C'n\epsilon_n^2}}{\epsilon_n^{2C'+1}} \sum_{j\geq J} \exp\left\{2C'\left[(1+\delta)(j+1)\log^2(j+1) - \frac{b}{4C'}j^\tau\right]a_n^\tau\right\}$$

Now choose J large enough such that $(1+\delta)(j+1)\log^2(j+1) - (b/4C')j^\tau \leq -(j-1)^2$ for $j \geq J$. So we get the upper bound

$$\frac{\mathrm{e}^{C'n\epsilon_n^2}}{\epsilon_n^{2C'+1}} \sum_{j\geq J} \exp\left\{-2C'(j-1)^2 a_n^\tau\right\} \leq \frac{\mathrm{e}^{C'n\epsilon_n^2}}{\epsilon_n^{2C'+1}} \sum_{j\geq 1} \mathrm{e}^{-2C'ja_n^\tau} = \frac{\mathrm{e}^{C'n\epsilon_n^2}}{\epsilon_n^{2C'+1}} \frac{\mathrm{e}^{-2C'a_n^\tau}}{1-\mathrm{e}^{-2C'a_n^\tau}}$$

which vanishes to 0 as $n \to \infty$.

As for the first sum in (26), since $\mathcal{B}_{n,0,k} \subset \mathcal{F}_{a_n,\epsilon_n^2/3,\sigma_n(1+\epsilon_n^2)^{k-1},\sigma_n(1+\epsilon_n^2)^k}$,

$$\begin{aligned}
&\log N(\epsilon_n, \mathcal{B}_{n,0,k}, H) \\
\lesssim &\log\frac{1+\epsilon_n^2}{\epsilon_n^2} + \left(\frac{a_n}{\sigma_n(1+\epsilon_n^2)^{k-1}} \vee 1\right)\log\frac{3}{\epsilon_n^2}\left(\log\left(\frac{3a_n}{\epsilon_n^2\sigma_n(1+\epsilon_n^2)^{k-1}}+1\right)+\log\frac{3}{\epsilon_n^2}\right) \\
\lesssim &\begin{cases}
\dfrac{a_n}{\sigma_n(1+\epsilon_n^2)^{k-1}}\left(\log\dfrac{a_n}{\epsilon_n^2\sigma_n(1+\epsilon_n^2)^{k-1}}\right)^2, & \sigma_n(1+\epsilon_n^2)^{k-1} < a_n, \\
\left(\log\dfrac{1}{\epsilon_n^2}\right)^2, & \sigma_n(1+\epsilon_n^2)^{k-1} > a_n
\end{cases}
\end{aligned}$$

An upper bound on the prior probability on the $\mathcal{B}_{n,0,k}$ is found by direct calculation:

$$\begin{aligned}
\Pi(\mathcal{B}_{n,0,k}) &\leq G(\sigma_n(1+\epsilon_n^2)^{k-1} \leq \sigma \leq \sigma_n(1+\epsilon_n^2)^k) \\
&= G(\sigma_n^{-1}(1+\epsilon_n^2)^{-2k} \leq \sigma^{-2} \leq \sigma_n^{-2}(1+\epsilon_n^2)^{-2(k-1)}) \\
&= \int_{\sigma_n^{-2}(1+\epsilon_n^2)^{-2k}}^{\sigma_n^{-1}(1+\epsilon_n^2)^{-2(k-1)}} y^{c_3-1}\mathrm{e}^{-c_1 y}\mathrm{d}y \\
&\lesssim \sigma_n^{-2(c_3-1)}(1+\epsilon_n^2)^{-2(c_3-1)k}\exp\{-c_1\sigma_n^{-2}(1+\epsilon_n^2)^{-2k}\}
\end{aligned}$$

see proof of Theorem 2 in Kruijer et al. (2010). Hence

$$\sum_{k=1}^{n} N(\epsilon_n, \mathcal{B}_{n,0,k}, H)\Pi(\mathcal{B}_{n,0,k})^{1/2} = \sum_{\sigma_n(1+\epsilon_n^2)^{k-1} \le a_n} N(\epsilon_n, \mathcal{B}_{n,0,k}, H)\Pi(\mathcal{B}_{n,0,k})^{1/2}$$
$$+ \sum_{\sigma_n(1+\epsilon_n^2)^{k-1} > a_n} N(\epsilon_n, \mathcal{B}_{n,0,k}, H)\Pi(\mathcal{B}_{n,0,k})^{1/2} = I_{n,1} + I_{n,2}.$$

As for $I_{n,2}$, for any c,

$$I_{n,2} \le nN(\epsilon_n, \mathcal{B}_{n,0,k}, H) \lesssim \exp\left\{C'\left(\log\frac{1}{\epsilon_n^2}\right)^2 + \log n\right\} = o(\mathrm{e}^{-cn\epsilon_n^2}).$$

We prove next that also $I_{n,1}\mathrm{e}^{-cn\epsilon_n^2} \to 0$ for any c small. The sum extends over $k \ge 1$ such that $(1+\epsilon_n^2)^{k-1} \le \sigma_n^{-2}(\log n)^{-3}$, that is

$$k - 1 \le \frac{\log(n\epsilon_n^2(\log n)^{-3})}{\log(1+\epsilon_n^2)}; \quad k \le c'\epsilon_n^{-2}\log n$$

for some constant $c' > 0$. $N(\epsilon_n, \mathcal{B}_{n,0,k}, H)\Pi(\mathcal{B}_{n,0,k})^{1/2}$ is bounded by

$$\exp\left\{C'\frac{a_n}{\sigma_n(1+\epsilon_n^2)^{k-1}}\left(\log\frac{a_n}{\epsilon_n^2\sigma_n(1+\epsilon_n^2)^{k-1}}\right)^2\right\}$$
$$\times\, \sigma_n^{-(c_3-1)}(1+\epsilon_n^2)^{-(c_3-1)k}\exp\{-\tfrac{c_1}{2}\sigma_n^{-2}(1+\epsilon_n^2)^{-2k}\}$$

by writing $s_{n,k} = \sigma_n(1+\epsilon_n^2)^{k-1}$,

$$= \exp\left\{C'\frac{a_n}{s_{n,k}}\left(\log\frac{a_n}{\epsilon_n^2 s_{n,k}}\right)^2\right\} s_{n,k}^{-(c_3-1)}(1+\epsilon_n^2)^{-(c_3-1)}\exp\{-\tfrac{c_1}{2}s_{n,k}^{-2}(1+\epsilon_n^2)^{-2}\}$$

since $1 \le (1+\epsilon_n^2) \le 2$,

$$\le \exp\left\{C'\frac{a_n}{s_{n,k}}\left(\log\frac{a_n}{\epsilon_n^2 s_{n,k}}\right)^2 - \frac{c_1}{4}s_{n,k}^{-2} - (c_3-1)\log s_{n,k}\right\}$$

since $\sigma_n \le s_{n,k} \le a_n$, for some $c' > 0$,

$$\le \exp\left\{C'\frac{a_n}{s_{n,k}}\left(\log\frac{a_n}{\epsilon_n^2 s_{n,k}}\right)^2 - \frac{c_1}{4}s_{n,k}^{-2} + c'\log n\right\}$$

since $a_n/(\epsilon_n^2 s_{n,k}) \le a_n/(\epsilon_n^2\sigma_n) = \sigma_n^{-2}(\log n)^{-3}$,

$$\le \exp\left\{C'\frac{a_n\log^2 n}{s_{n,k}} - \frac{c_1}{4}s_{n,k}^{-2} + c'\log n\right\}$$
$$= \exp\left\{C'\frac{\sigma_n^{-2}(\log n)^{-1}}{(1+\epsilon_n^2)^{k-1}} - \frac{c_1}{4}\frac{\sigma_n^{-2}}{(1+\epsilon_n^2)^{2(k-1)}} + c'\log n\right\}$$
$$= \exp\left\{C'\frac{n\epsilon_n^2(\log n)^{-1}}{(1+\epsilon_n^2)^{k-1}} - \frac{c_1}{4}\frac{n\epsilon_n^2}{(1+\epsilon_n^2)^{2(k-1)}} + c'\log n\right\}$$
$$= \exp\left\{\frac{n\epsilon_n^2}{(1+\epsilon_n^2)^{k-1}}\left(C'(\log n)^{-1} - \frac{c_1}{4}\frac{1}{(1+\epsilon_n^2)^{k-1}}\right) + c'\log n\right\}.$$

The last display is bounded by a multiple of $\exp\{C' n\epsilon_n^2 (\log n)^{-1} + c' \log n\}$, hence the sum of $N(\epsilon_n, \mathcal{B}_{n,0,k}, H)\Pi(\mathcal{B}_{n,0,k})^{1/2}$ over $k = 1, \ldots, \epsilon_n^{-2} \log n$ is bounded by a multiple of $\exp\{C' n\epsilon_n^2 (\log n)^{-1} + c'' \log n\}$ which increases at a slower rate than $\exp\{cn\epsilon_n^2\}$ for any c. The proof is complete. □

Proof of Proposition 3.2. Write $I_{n,j} = n_j^{-1} \sum_{i \in \mathcal{I}_j} \log[f_0(y_i|x_i)/f_{0,j}(y_i)]$ with $I_{n,j}$ define to be 0 if $n_j = 0$. The goal is to show that $P_0(\sum_{j=1}^N n_j I_{n,j} \geq d_n\, n/N) \to 0$ as $n \to \infty$. To this aim, consider that

$$P_0\Big(\sum_{j=1}^N n_j I_{n,j} \geq d_n \frac{n}{N}\Big) = \sum_{n_{1:N}} P_0\Big(\sum_{j=1}^N n_j I_{n,j} \geq d_n \frac{n}{N} \Big| n_{1:N}\Big)\; P_0(n_{1:N}) \tag{27}$$

where we used $n_{1:N}$ as short hand notation for $n_1, \ldots, n_N$. We focus next on the conditional probability in the right hand side of (27), aiming at finding an upper bound that goes to zero as $n \to \infty$ uniformly in $n_{1:N}$.

Note that each $I_{n,j}$ can be written as the sample mean of n_j realizations $\log[f_0(y_i|x_i)/f_{0,j}(y_i)]$ for $i \in \mathcal{I}_j = \{i : x_i \in C_j\}$ with respect to the sample distribution P_0. We have

$$\begin{aligned} E_0\big(I_{n,j} \,|\, n_{1:N}\big) &= E_0\Big(\log \frac{f_0(Y|X)}{f_{0,j}(Y)} \,\Big|\, X \in C_j\Big) \\ &= \frac{1}{Q(C_j)} \int_{C_j} \int_{\mathbb{R}} \log \frac{f_0(y|x)}{f_{0,j}(y)} f_0(y|x)\mathrm{d}y\, q(x)\mathrm{d}x \\ &\leq \sup_{x \in C_j} \int_{\mathbb{R}} f_0(y|x) \log \frac{f_0(y|x)}{f_{0,j}(y)} \mathrm{d}y \lesssim N^{-2} \end{aligned}$$

where the last inequality follows from an application of Lemma .1. Also,

$$\begin{aligned} \mathrm{Var}\big(I_{n,j} \,|\, n_{1:N}\big) &= \frac{1}{n_j} \mathrm{Var}\Big(\log \frac{f_0(Y|X)}{f_{0,j}(Y)} \,\Big|\, X \in C_j\Big) \\ &\leq \frac{1}{n_j} E_0\Big\{\Big(\log \frac{f_0(Y|X)}{f_{0,j}(Y)}\Big)^2 \,\Big|\, X \in C_j\Big\} \\ &= \frac{1}{n_j} \frac{1}{Q(C_j)} \int_{C_j} \int_{\mathbb{R}} f_0(y|x) \Big(\log \frac{f_0(y|x)}{f_{0,j}(y)}\Big)^2 \mathrm{d}y\, q(x)\mathrm{d}x \\ &\leq \frac{1}{n_j} \sup_{x \in C_j} \int_{\mathbb{R}} f_0(y|x) \Big(\log \frac{f_0(y|x)}{f_{0,j}(y)}\Big)^2 \mathrm{d}y \lesssim \frac{1}{n_j N^2}, \end{aligned}$$

using again Lemma .1. Given the independence of $I_{n,j}$ across j conditional on $n_{1:N}$, we also have

$$E_0\big(\sum_{j=1}^N n_j I_{n,j} \,|\, n_{1:N}\big) \lesssim n/N^2, \quad \mathrm{Var}\big(\sum_{j=1}^N n_j I_{n,j} \,|\, n_{1:N}\big) \lesssim n/N^2.$$

Now write the conditional probability in (27) as

$$P_0\Big(\sum_{j=1}^N n_j I_{n,j} - E_0\Big[\sum_{j=1}^N n_j I_{n,j} \,|\, n_{1:N}\Big] \geq d_n\, n/N - E_0\Big[\sum_{j=1}^N n_j I_{n,j} \,|\, n_{1:N}\Big] \,|\, n_{1:N}\Big)$$

which is upper bounded by

$$P_0\Big(\sum_{j=1}^N n_j I_{n,j} - E_0\Big[\sum_{j=1}^N n_j I_{n,j} \,|\, n_{1:N}\Big] \geq d_n\, n/(2N) \,|\, n_{1:N}\Big)$$

for sufficiently large n, since $N^{-2} \lesssim d_n/(2N)$, as implied by the hypothesis $d_n N \to \infty$. Now use Chebishev inequality to get

$$P_0\Big(\sum_{j=1}^N n_j I_{n,j} \geq d_n\, n/N \,|\, n_{1:N}\Big) \leq \frac{n/N^2}{n^2 d_n^2/(4N^2)} = \frac{1}{4nd_n^2}$$

which goes to zero by the hypothesis $nd_n^2 \to \infty$. The proof is then complete. □

Proof of Proposition 3.3. The proof is an adaptation of Lemma 8.1 in Ghosal et al. (2000) and of the proof of Theorem 2.1 therein. It goes along by first showing that, for any ϵ and for $\tilde\Pi_1, \ldots, \tilde\Pi_N$ independent probability measures concentrated on the sets $B(\epsilon, f_{0,1}), \ldots, B(\epsilon, f_{0,N})$, it is that for every $\delta > 0$,

$$P_0\Big\{\prod_{j=1}^N \int \prod_{i\in\mathcal{I}_j} \frac{f(Y_i)}{f_{0,j}(Y_i)}\, \tilde\Pi_j(\mathrm{d}f) \leq \exp(-(1+\delta)n\epsilon^2)\Big\} \leq \frac{1}{\delta^2 n\epsilon^2}. \tag{28}$$

Recall the notation for $\mathbb{P}_{n,j}\, g = \frac{1}{n_j}\sum_{i\in\mathcal{I}_j} g(Y_i)$ and $P_{0,j}\, g = \int g(y) f_{0,j}(y)\mathrm{d}y$, so that $P_0[\mathbb{P}_{n,j}\, g] = P_0(P_0[\mathbb{P}_{n,j}\, g|n_j]) = P_{0,j}\, g$. By Jensen's inequality applied to the logarithm,

$$\log \int \prod_{i\in\mathcal{I}_j} \frac{f(y_i)}{f_{0,j}(y_i)}\, \tilde\Pi_j(\mathrm{d}f) \geq \sum_{i\in\mathcal{I}_j} \int \log \frac{f(y_i)}{f_{0,j}(y_i)}\, \tilde\Pi_j(\mathrm{d}f) = n_j \mathbb{P}_{n,j} \int \log \frac{f}{f_{0,j}}\, \tilde\Pi_j(\mathrm{d}f)$$

so the left hand side of (28) is

$$\leq P_0\Big\{\sum_{j=1}^N n_j \mathbb{P}_{n,j} \int \log \frac{f}{f_{0,j}}\, \tilde\Pi_j(\mathrm{d}f) \leq -(1+\delta)n\epsilon^2\Big\}$$

$$= P_0\Big\{\sum_{j=1}^N n_j (\mathbb{P}_{n,j} - P_{0,j}) \int \log \frac{f}{f_{0,j}}\, \tilde\Pi_j(\mathrm{d}f)$$

$$\leq -(1+\delta)n\epsilon^2 - \sum_{j=1}^N n_j P_{0,j} \int \log \frac{f}{f_{0,j}}\, \tilde\Pi_j(\mathrm{d}f)\Big\}$$

$$\leq P_0\Big\{\sum_{j=1}^N n_j (\mathbb{P}_{n,j} - P_{0,j}) \int \log \frac{f}{f_{0,j}}\, \tilde\Pi_j(\mathrm{d}f) \leq -\delta n\epsilon^2\Big\}$$

where the last inequality follows by an application of Fubini's theorem together with the assumptions on $\tilde{\Pi}_j$ being supported on $B_{j,\epsilon}$, i.e.

$$P_{0,j}\int \log(f/f_{0,j})\,\tilde{\Pi}_j(\mathrm{d}f) \geq -\epsilon^2.$$

Now write

$$P_0\Big\{\sum_{j=1}^N n_j(\mathbb{P}_{n,j}-P_{0,j})\int \log\frac{f}{f_{0,j}}\,\tilde{\Pi}_j(\mathrm{d}f) \leq -\delta n\epsilon^2\Big\}$$
$$=\sum_{n_{1:N}} P_0\Big\{\sum_{j=1}^N n_j(\mathbb{P}_{n,j}-P_{0,j})\int \log\frac{f}{f_{0,j}}\,\tilde{\Pi}_j(\mathrm{d}f) \leq -\delta n\epsilon^2 \,|\, n_{1:N}\Big\}P_0(n_{1:N})$$

and use Chebichev's inequality to get the following upper bound for the left hand side of (28)

$$\frac{1}{\delta^2 n^2\epsilon^4}\sum_{n_{1:N}}\mathrm{Var}\Big(\sum_{j=1}^N n_j(\mathbb{P}_{n,j}-P_{0,j})\int \log\frac{f}{f_{0,j}}\,\tilde{\Pi}_j(\mathrm{d}f)\,\Big|\,n_{1:N}\Big)P_0(n_{1:N})$$
$$=\frac{1}{\delta^2 n^2\epsilon^4}\sum_{n_{1:N}}\Big[\sum_{j=1}^N \mathrm{Var}\Big(n_j(\mathbb{P}_{n,j}-P_{0,j})\int \log\frac{f}{f_{0,j}}\,\tilde{\Pi}_j(\mathrm{d}f)\,\Big|\,n_{1:N}\Big)\Big]P_0(n_{1:N})$$
$$\leq\frac{1}{\delta^2 n^2\epsilon^4}\sum_{n_{1:N}}\Big[\sum_{j=1}^N n_j P_{0,j}\Big(\int \log\frac{f_{0,j}}{f}\tilde{\Pi}_j(\mathrm{d}f)\Big)^2\Big]P_0(n_{1:N})$$
$$\leq\frac{1}{\delta^2 n^2\epsilon^4}\sum_{n_{1:N}}\Big[\sum_{j=1}^N n_j P_{0,j}\int \Big(\log\frac{f_{0,j}}{f}\Big)^2\tilde{\Pi}_j(\mathrm{d}f)\Big]P_0(n_{1:N})$$

where in the last inequality Jensen's inequality has been used. By Fubini's theorem and the assumptions on $\tilde{\Pi}_j$, $P_{0,j}\int[\log(f_{0,j}/f)]^2\tilde{\Pi}_j(\mathrm{d}f)\leq\epsilon^2$, hence the last display is

$$\leq\frac{1}{\delta^2 n^2\epsilon^4}\sum_{n_{1:N}}\sum_{j=1}^N n_j\epsilon^2 P_0(n_{1:N})=\frac{1}{\delta^2 n^2\epsilon^4}n\epsilon^2=\frac{1}{\delta^2 n\epsilon^2}$$

so that (28) is proved.

Next, for the constant C' in (8),

$$P_0\Big\{\prod_{j=1}^N\int\prod_{i=1}^n\frac{f(y_i)}{f_{0,j}(y_i)}\mathbb{1}_{C_j}(x_i)\,\tilde{\Pi}_j(\mathrm{d}f)\geq\exp\Big(-(1+\delta+C')n\tilde{\epsilon}_{n/N}^2\Big)\Big\}$$

$$\geq P_0\Big\{\prod_{j=1}^N\tilde{\Pi}_j(B(\tilde{\epsilon}_{n/N},f_{0,j}))\int_{B(\tilde{\epsilon}_{n/N},f_{0,j})}\prod_{i=1}^n\frac{f(y_i)}{f_{0,j}(y_i)}\mathbb{1}_{C_j}(x_i)\,\frac{\tilde{\Pi}_j(\mathrm{d}f)}{\tilde{\Pi}_j(B(\tilde{\epsilon}_{n/N},f_{0,j}))}$$
$$\geq\exp\Big(-(1+\delta+C')n\tilde{\epsilon}_{n/N}^2\Big)\Big\}$$

$$\geq P_0\Big\{\mathrm{e}^{-C'n\tilde{\epsilon}_{n/N}^2}\prod_{j=1}^N\int_{B(\tilde{\epsilon}_{n/N},f_{0,j})}\prod_{i=1}^n\frac{f(y_i)}{f_{0,j}(y_i)}\mathbb{1}_{C_j}(x_i)\,\frac{\tilde{\Pi}_j(\mathrm{d}f)}{\tilde{\Pi}_j(B(\tilde{\epsilon}_{n/N},f_{0,j}))}$$

$$\geq\exp\Big(-(1+\delta+C')n\tilde{\epsilon}_{n/N}^2\Big)\Big\}$$

$$=P_0\Big\{\prod_{j=1}^N\int_{B(\tilde{\epsilon}_{n/N},f_{0,j})}\prod_{i=1}^n\frac{f(y_i)}{f_{0,j}(y_i)}\mathbb{1}_{C_j}(x_i)\,\frac{\tilde{\Pi}_j(\mathrm{d}f)}{\tilde{\Pi}_j(B(\tilde{\epsilon}_{n/N},f_{0,j}))}$$
$$\geq\exp\big(-(1+\delta)n\tilde{\epsilon}_{n/N}^2\big)\Big\}$$

where (8) has been used in the third inequality. The integral on the left hand side is with respect to the prior $\tilde{\Pi}_j$ restricted on the set $B(\tilde{\epsilon}_{n/N},f_{0,j})$, so we can use (28) for $\epsilon=\tilde{\epsilon}_{n/N}$. The last probability is then lower bounded by $1-1/(\delta^2n\tilde{\epsilon}_{n/N}^2)\uparrow 1$. The proof is then complete. □

Lemma .1. *Under (5), for all j and N sufficiently large,*

$$\sup_{x\in C_j}\int_{\mathbb{R}}f_0(y|x)\bigg(\log\frac{f_0(y|x)}{f_{0,j}(y)}\bigg)^r\mathrm{d}y\lesssim N^{-2},\quad r=1,2.$$

Proof. We start off dealing with $r=1$ and fix an $x\in C_j$. By the convexity of $-\log(\cdot)$, Jensen's inequality gives

$$\log\frac{f_0(y|x)}{f_{0,j}(y)}\leq\frac{1}{Q(C_j)}\int_{C_j}\log\frac{f_0(y|x)}{f_0(y|x')}q(x')\mathrm{d}x'\tag{29}$$

and hence

$$\begin{aligned}\int_{\mathbb{R}}f_0(y|x)\log\frac{f_0(y|x)}{f_{0,j}(y)}\mathrm{d}y&\leq\int_{\mathbb{R}}f_0(y|x)\frac{1}{Q(C_j)}\int_{C_j}\log\frac{f_0(y|x)}{f_0(y|x')}q(x')\mathrm{d}x'\,\mathrm{d}y\\&=\frac{1}{Q(C_j)}\int_{C_j}\int_{\mathbb{R}}f_0(y|x)\log\frac{f_0(y|x)}{f_0(y|x')}\mathrm{d}y\,q(x')\mathrm{d}x'\\&\leq\sup_{x'\in C_j}\int_{\mathbb{R}}f_0(y|x)\log\frac{f_0(y|x)}{f_0(y|x')}\mathrm{d}y,\end{aligned}$$

where Fubini's theorem has been used to derive the equality. The thesis follows by (5) and (2).

Now we deal with the case $r = 2$. We can not use Jensen's inequality since $[\log(a/b)]^2$ is not convex in b. Hence, we need to split the integral into two parts. As before, fix an $x \in C_j$, and write

$$\int_{\mathbb{R}} f_0(y|x)\left(\log\frac{f_0(y|x)}{f_{0,j}(y)}\right)^2 \mathrm{d}y = \int_A f_0(y|x)\left(\log\frac{f_0(y|x)}{f_{0,j}(y)}\right)^2 \mathrm{d}y + \int_B f_0(y|x)\left(\log\frac{f_0(y|x)}{f_{0,j}(y)}\right)^2 \mathrm{d}y$$

where

$$A = \left\{y : \ \tfrac{f_0(y|x)}{f_{0,j}(y)} > 1\right\} \quad \text{and} \quad B = \left\{y : \ \tfrac{f_0(y|x)}{f_{0,j}(y)} \le 1\right\}.$$

For $y \in A$, we can use (29) to get

$$0 < \left(\log\frac{f_0(y|x)}{f_{0,j}(y)}\right)^2 \le \left(\frac{1}{Q(C_j)}\int_{C_j} \log\frac{f_0(y|x)}{f_0(y|x')} q(x')\mathrm{d}x'\right)^2.$$

A second application of Jensen's inequality yields

$$\left(\log\frac{f_0(y|x)}{f_{0,j}(y)}\right)^2 \le \frac{1}{Q(C_j)}\int_{C_j}\left(\log\frac{f_0(y|x)}{f_0(y|x')}\right)^2 q(x')\mathrm{d}x'$$

so that

$$\begin{aligned}
\int_A f_0(y|x)\left(\log\frac{f_0(y|x)}{f_{0,j}(y)}\right)^2 \mathrm{d}y &\le \int_A f_0(y|x)\frac{1}{Q(C_j)}\int_{C_j}\left(\log\frac{f_0(y|x)}{f_0(y|x')}\right)^2 q(x')\mathrm{d}x'\,\mathrm{d}y \\
&= \frac{1}{Q(C_j)}\int_{C_j}\int_A f_0(y|x)\left(\log\frac{f_0(y|x)}{f_0(y|x')}\right)^2 \mathrm{d}y\, q(x')\mathrm{d}x' \\
&\le \sup_{x'\in C_j}\int_A f_0(y|x)\left(\log\frac{f_0(y|x)}{f_0(y|x')}\right)^2 \mathrm{d}y \\
&\le \sup_{x'\in C_j}\int_{\mathbb{R}} f_0(y|x)\left(\log\frac{f_0(y|x)}{f_0(y|x')}\right)^2 \mathrm{d}y.
\end{aligned}$$

For $y \in B$ we can use the fact that $|\log x| \le 2|x^{1/2} - 1|$ for $x \ge 1$ so that

$$\begin{aligned}
\int_B f_0(y|x)\left(\log\frac{f_0(y|x)}{f_{0,j}(y)}\right)^2 \mathrm{d}y &\le 4\int_B f_0(y|x)\left(1 - \sqrt{f_{0,j}(y)/f_0(y|x)}\right)^2 \mathrm{d}y \\
&\le 4H^2(f_0(\cdot,x), f_{0,j}) \\
&\le 4\frac{1}{Q(C_j)}\int_{C_j} H^2(f_0(\cdot,x), f_0(\cdot|x'))q(x')\mathrm{d}x' \\
&\le 4\sup_{x'\in C_j} H^2(f_0(\cdot,x), f_0(\cdot|x')) \\
&\le 4\sup_{x'\in C_j}\int_{\mathbb{R}} f_0(y|x)\log\frac{f_0(y|x)}{f_0(y|x')}\mathrm{d}y,
\end{aligned}$$

where we have used $H^2(f,g) = \int f\{1-(g/f)^{1/2}\}^2$ in the second inequality, the convexity of the Hellinger distance together with Jensen's inequality in the third inequality, and $H^2(f,g) \leq \int f \log(f/g)$ in the last inequality. We conclude that, for any $x \in C_j$,

$$\int_{\mathbb{R}} f_0(y|x)\left(\log \frac{f_0(y|x)}{f_{0,j}(y)}\right)^2 \mathrm{d}y \leq \sup_{x' \in C_j} \int_{\mathbb{R}} f_0(y|x)\left(\log \frac{f_0(y|x)}{f_0(y|x')}\right)^2 \mathrm{d}y + 4 \sup_{x' \in C_j} \int_{\mathbb{R}} f_0(y|x) \log \frac{f_0(y|x)}{f_0(y|x')} \mathrm{d}y$$

The thesis follows again by (5) and (2). □

Acknowledgements

The authors are grateful to two anonymous referees and the Associate Editor for carefully reading and for providing comments that helped to improve the paper.

References

Castillo, I. (2014). On bayesian supremum norm contraction rates. *Ann. Statist.*, 42(5):2058–2091.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian cart model search. *J. Amer. Statist. Assoc.*, 93:935–948.

De Blasi, P., James, L. F., and Lau, J. W. (2010). Bayesian nonparametric estimation and consistency of mixed multinomial logit choice models. *Bernoulli*, 16:679–704.

Efromovich, S. (2007). Conditional density estimation in a regression setting. *The Annals of Statistics*, 35(6):2504–2535.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531.

Ghosal, S. and Roy, A. (2006). Posterior consistency of gaussian process prior for nonparametric binary regression. *Ann. Statist.*, 34:2413–2429.

Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.*, 29(5):1233–1263.

Ghosal, S. and van der Vaart, A. W. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.*, 35(5):697–723.

Ginè, E. and Nickl, R. (2011). Rates of contraction for posterior distributions in l^r-metrics, $1 \leq r \leq \infty$. *Ann. Statist.*, 39:2883–2911.

Hoffman, M., Rousseau, J., and Schmidt-Hieber, J. (2015). On adaptive posterior contraction rates. *Ann. Statist.*, 43(5):2259–2295.

Kruijer, W., Rousseau, J., and van der Vaart, A. W. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257.

Lijoi, A., Prünster, I., and Walker, S. G. (2005). On consistency of nonparametric normal mixtures for bayesian density estimation. *J. Amer. Statist. Assoc.*, 100(472):1292–1296.

Ma, L. (2012). Recursive partitioning and bayesian inference on conditional distributions. Technical report, Duke Statistical Science Discussion Paper 2012-03.

Norets, A. and Pati, D. (2014). Adaptive bayesian estimation of conditional densities. Technical report, arXiv:1408.5355v2.

Pati, D., Dunson, D. B., and Tokdar, S. T. (2013). Posterior consistency in conditional distribution estimation. *J. Multivariate Anal.*, 116:456–472.

Shen, W. and Ghosal, S. (2016). Adaptive bayesian density regression for high-dimensional data. *Bernoulli*, 22(1):396–420.

Shen, W., Tokdar, S. T., and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100:623–640.

Shively, T. S., Sager, T. W., and Walker, S. G. (2009). A bayesian approach to non-parametric monotone function estimation. *J. R. Stat. Soc. Ser. B*, 71:159–175.

Tang, Y. and Ghosal, S. (2007). Posterior consistency of dirichlet mixture for estimating a transition density. *J. Statist. Plann. Inference*, 137:1711–1726.

Tokdar, S. T., Zhu, Y., and Ghosh, J. K. (2010). Bayesian density regression with logistic gaussian process and subspace projection. *Bayesian Analysis*, 5:1–26.

Walker, S. G. (2004). New approaches to bayesian consistency. *Ann. Statist.*, 32:2028–2043.

Walker, S. G., Lijoi, A., and Prünster, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *Ann. Statist.*, 35:738–746.

Xiang, F. and Walker, S. G. (2013). Bayesian consistency for regression models under a supremum distance. *J. Statist. Plann. Inference*, 143:468–478.

Yoo, W. W. and Ghosal, S. (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *Ann. Statist.*, 44(3):1069–1102.