9

Contraction Rates: Examples

In this chapter we obtain explicit posterior contraction rates by applying the general theorems from the preceding chapter to common priors and models: log-spline models, Dirichlet process mixtures of normal or Bernstein polynomial kernels for density estimation, nonparametric and semiparametric regression models with priors obtained from bracketing or the Dirichlet process on the link function or error density, spectral density estimation and current status and interval censoring with the Dirichlet process prior.

9.1 Log-Spline Priors

Spline functions $f\colon [0,1] \to \mathbb{R}$ of order q are piecewise polynomials of degree $q-1$ whose pieces connect smoothly of order $q-2$ at the boundaries of their domains. They possess excellent approximation properties for smooth functions, and permit a numerically stable representation as linear combinations of the *B-spline* basis (see Section E.2 for an introduction). Splines can be used for density estimation by exponentiaton and normalizing, as in Section 2.3.1, thus leading to an exponential family with the B-spline basis functions as the sufficient statistics. A prior density is then induced through a prior on the parameter vector of the family.

In the nonparametric setup it is necessary to let the dimension of the family increase, so that it can approximate an arbitrary continuous true density. In this section we show that a prior with of the order $n^{1/(2\alpha+1)}$ basis functions yields a posterior distribution that contracts at the optimal rate for a true density in the Hölder space $\mathfrak{C}^\alpha[0,1]$. (In Chapter 10 the construction will be extended with a prior on α or the dimension, in order to achieve this rate simultaneously for every $\alpha > 0$.)

Fix some *order* $q \in \mathbb{N}$, and for given $K \in \mathbb{N}$ partition $[0,1)$ into the K equal-length subintervals $\big[(k-1)/K, k/K\big)$, for $k = 1, \ldots, K$. The B-spline functions $B_{J,1}, \ldots, B_{J,J}$, described in Section E.2, generate the $J = q + K - 1$-dimensional linear space of all *splines of order* q relative to this partition. Consider probability densities relative to Lebesgue measure on $[0,1]$ given by the exponential family

$$p_{J,\theta}(x) = e^{\theta^\top B_J(x) - c(\theta)}, \quad \theta^\top B_J := \sum_{j=1}^J \theta_j B_{J,j}, \quad c(\theta) = \log \int_0^1 e^{\theta^\top B_J(x)}\,dx, \quad \theta \in \mathbb{R}^J.$$

Because the B-splines add up to unity, the family is actually of dimension $J-1$ and we can restrict the parameter to $\{\theta \in \mathbb{R}^J : \theta^\top 1 = 0\}$. We place a prior on the latter set, and this next induces a prior on probability densities through the map $\theta \mapsto p_{J,\theta}$. In the special case

$q = 1$, the splines are piecewise constant functions, and the prior sits on histograms with cell boundaries k/K, for $k = 0, 1, \ldots, K$.

The true density p_0 need not be of the form $p_{J,\theta}$ for any θ and/or J, and therefore needs to be approximated by some $p_{J,\theta}$ for the posterior distribution to be consistent. To ensure this for a large class of p_0, the dimension $J - 1$ of the log-spline model must tend to infinity with n. The minimal rate at which $J = J_n$ must grow is determined by the approximation properties of the spline functions. For $\log p_0$ belonging to the Hölder space $\mathfrak{C}^\alpha[0, 1]$ and $q \geq \alpha$, Lemma E.5 asserts, for a constant C depending only on q and α,

$$\inf_{\theta \in \mathbb{R}^J} \|\theta^\top B_J - \log p_0\|_\infty \leq C J^{-\alpha} \|\log p_0\|_{\mathfrak{C}^\alpha}.$$

This bound cannot be improved for a general $p_0 \in \mathfrak{C}^\alpha[0, 1]$: it is sharp for p_0 that are "exactly of smoothness α." Consequently, as it certainly cannot be smaller than the distance of the model to the true density, the posterior contraction rate ϵ_n always satisfies $\epsilon_n \gtrsim J^{-\alpha}$ for some p_0. This favors high-dimensional models (large J), and for consistency it is needed that $J = J_n \to \infty$. On the other hand, the local entropy of the spline model turns out to be a multiple of its dimension J, whence the entropy condition (ii) of Theorem 8.11 takes the form $J \lesssim n\epsilon_n^2$, or $\epsilon_n \gtrsim \sqrt{J/n}$. The maximum of the two lower bounds $J^{-\alpha}$ and $\sqrt{J/n}$ is minimized by $J \sim n^{1/(2\alpha+1)}$, which leads to $\epsilon_n \sim n^{-\alpha/(2\alpha+1)}$, the minimax rate of estimation for $\mathfrak{C}^\alpha[0, 1]$. The following theorem shows that the posterior distribution corresponding to flat priors on the coefficients attains this rate.

We restrict to true densities that are bounded away from zero, and use a compactly supported prior for the coefficients θ.

Theorem 9.1 (Contraction rate, log-splines) *Let $\theta \sim \pi_n$ for π_n a Lebesgue density on $\{\theta \in [-M, M]^{J_n} \colon \theta^\top 1 = 0\}$ such that $\underline{c}^{J_n} \leq \pi_n(\theta) \leq \bar{c}^{J_n}$, for some $0 < \underline{c} < \bar{c} < \infty$ and all θ, where $M \geq 1$ is fixed and $J_n \sim n^{1/(2\alpha+1)}$. If $p_0 \in \mathfrak{C}^\alpha[0, 1]$, with $q \geq \alpha \geq 1/2$ and $\|\log p_0\|_\infty \leq d_0 M/2$, where d_0 is the universal constant in Lemma 9.2, then the posterior distribution relative to the prior $p_{J_n,\theta}$ contracts at p_0 at the rate $n^{-\alpha/(2\alpha+1)}$ with respect to the Hellinger distance.*

Proof The proof is based on Theorem 8.11 in combination with a sequence of lemmas listed below.

In Lemma 9.5 the local entropy relative to the Hellinger distance of the full support of the prior is estimated to be bounded above by a multiple of J_n, and hence the entropy equation in (ii) of Theorem 8.11 is satisfied if $J_n \lesssim n\epsilon_n^2$, which is the case for the proposed dimension and ϵ_n a multiple of the claimed contraction rate.

By Lemma 9.4(iii), a Kullback-Leibler ball around p_0 contains a Hellinger ball around p_0 of a multiple of the radius. By Lemma 9.6(ii) and (i) there exists θ_J with $\|\theta_J\|_\infty \leq M$ such that $d_H(p_0, p_{J,\theta_J}) \lesssim J^{-\alpha}$, which is $n^{-\alpha/(2\alpha+1)}$ for $J = J_n$. Next by Lemma 9.4(ii) a Hellinger ball of radius a multiple of ϵ_n contains a set of the form $\Theta(J_n, \epsilon_n)$, for $\Theta(J, \epsilon) = \{\theta \in [-M, M]^J \colon \|\theta - \theta_J\|_2 \leq \sqrt{J}\epsilon\}$.

By Lemma 9.4(i) Hellinger balls around p_0 of radius a multiple of $2j\epsilon_n$ are contained in a set $\Theta(J_n, Cj\epsilon_n)$, for some constant C depending on M only. We conclude that, for ϵ_n a multiple of $n^{-a(2\alpha+1)}$ and c and C constants that depend on M only,

$$\begin{aligned}\frac{\Pi_n(p_{J_n,\theta}\colon d_H(p_{J_n,\theta},p_0)\le 2j\epsilon_n)}{\Pi_n(B_2(p_0,\epsilon_n))} &\le \frac{\Pi_n(\Theta(J_n,Cj\epsilon_n))}{\Pi_n(\Theta(J_n,c\epsilon_n))}\\ &\le \frac{\sup_\theta \pi_n(\theta)\,(\sqrt{J_n}Cj\epsilon_n)^{J_n-1}\mathrm{vol}\{x\in\mathbb{R}^{J_n}\colon \|x\|\le 1\}}{\inf_\theta \pi_n(\theta)\,(\sqrt{J_n}c\epsilon_n)^{J_n-1}\mathrm{vol}\{x\in\mathbb{R}^{J_n}\colon \|x\|\le 1\}} \lesssim \Big(\frac{\bar{c}Cj}{\underline{c}}\Big)^{J_n-1}.\end{aligned}$$

This easily satisfies the bound in (i) of Theorem 8.11. □

The following lemmas compare the statistical distances on the log-spline densities to distances on their parameters. They were used in the preceding proof, and will be useful later on as well. Define

$$\begin{aligned}\Theta_{J,M} &= \{\theta\in[-M,M]^J\colon \theta^\top 1 = 0\},\\ B_{J,M}(p_0,\epsilon) &= \{\theta\in\Theta_{J,M}\colon K(p_0;p_{J,\theta})<\epsilon^2,\, V_2(p_0;p_{J,\theta})<\epsilon^2\},\\ C_{J,M}(p_0,\epsilon) &= \{\theta\in\Theta_{J,M}\colon d_H(p_0,p_{J,\theta})<\epsilon\}.\end{aligned} \tag{9.1}$$

Abusing notation, we shall use the notations $B_{J,M}(p_0,\epsilon)$ and $C_{J,M}(p_0,\epsilon)$ also for the set of densities $p_{J,\theta}$ when θ ranges over the sets as in the display.

Lemma 9.2 $d_0\|\theta\|_\infty \le \|\log p_{J,\theta}\|_\infty \le 2\|\theta\|_\infty$, *for any* $\theta\in\mathbb{R}^J$ *with* $\theta^\top 1 = 0$, *and a universal constant* d_0.

Lemma 9.3 $c_0e^{-M}(\|\theta_1-\theta_2\|\wedge\sqrt{J}) \le \sqrt{J}\,d_H(p_{J,\theta_1},p_{J,\theta_2}) \le C_0e^M\|\theta_1-\theta_2\|$, *for every* $\theta_1,\theta_2\in\Theta_{J,M}$, *and universal constants* $0<c_0<C_0<\infty$.

Proofs Since $\|\theta^\top B_J\|_\infty \le \|\theta\|_\infty =: M$, by the second inequality in Lemma E.6, the number $e^{c(\theta)} = \int_0^1 e^{\theta^\top B_J(x)}dx$ is contained in the interval $[e^{-M},e^M]$. Hence $|c(\theta)|\le M$, and next $\|\log p_{J,\theta}\|_\infty = \|\theta^\top B_J - c(\theta)\|_\infty \le 2M$, by the triangle inequality, which is the upper bound in the first lemma. For the lower bound, we use Lemma E.6 to see that $\|\theta\|_\infty \lesssim \|\theta^\top B_J\|_\infty$, which is bounded above by $\|\log p_{J,\theta}\|_\infty + |c(\theta)|$, by the triangle inequality. Since $\theta^\top 1 = 0$, the second term can be rewritten as $|(\theta-c(\theta)1)^\top 1|/J \le \|\theta-c(\theta)1\|_\infty$. Finally we use Lemma E.6 again to bound this by $\|(\theta-c(\theta)1)^\top B_J\|_\infty = \|\log p_{J,\theta}\|_\infty$.

For the proof of the second lemma we derive, by direct calculation and Taylor's theorem,

$$\begin{aligned}d_H^2(p_{J,\theta_1},p_{J,\theta_2}) &= 2\Big(1-\exp\big[c(\tfrac{1}{2}\theta_1+\tfrac{1}{2}\theta_2)-\tfrac{1}{2}c(\theta_1)-\tfrac{1}{2}c(\theta_2)\big]\Big)\\ &= 2\Big(1-\exp\big[-\tfrac{1}{16}(\theta_1-\theta_2)^\top(\ddot{c}(\tilde\theta)+\ddot{c}(\bar\theta))(\theta_1-\theta_2)\big]\Big),\end{aligned} \tag{9.2}$$

for $\tilde\theta,\bar\theta$ on the line segment connecting θ_1 and θ_2, and $\ddot{c}(\theta)$ the Hessian of c. By well-known properties of exponential families $t^\top\ddot{c}(\theta)t = \mathrm{var}_\theta(t^\top B_J)$. Because $1^\top B_J$ is degenerate, this is equal to the minimum over μ of $\int_0^1\big[(t-\mu1)^\top B(x)\big]^2 p_{J,\theta}(x)\,dx$, which is bounded below and above by $\|(t-\mu1)^\top B_J\|_2^2$ times the infimum and supremum of $p_{J,\theta}(x)$ over $x\in[0,1]$ and θ. By Lemma 9.2 the latter infimum and supremum can be bounded below and above by multiples of e^{-2M} and e^{2M}, respectively, while the square norm is comparable to $\|t-\mu1\|_2^2/J$, by Lemma E.6. We minimize the latter with respect to μ to reduce it to $\|t\|^2/J$ when $1^\top t = 0$.

To finish the proof, we insert the last bound in (9.2), and apply the elementary inequalities $(x\wedge 1)/2 \le 1-e^{-x}\le x$ for $x\ge 0$, and $(cx)\wedge 1 \ge c(x\wedge 1)$ for $x\ge 0$ and $c\le 1$. □

Lemma 9.4 *If $\theta_{J,M}$ minimizes $\theta \mapsto d_H(p_{J,\theta}, p_0)$ over $\Theta_{J,M}$, then, for universal constants c_0, C_0, d_0, D_0,*

(i) $C_{J,M}(p_0, \epsilon) \subset \{\theta \in \Theta_{J,M}: \|\theta - \theta_{J,M}\|_2 \leq 2e^M c_0^{-1}\sqrt{J}\epsilon\}$, *for* $2\epsilon < c_0 e^{-M}$.
(ii) $C_{J,M}(p_0, 2\epsilon) \supset \{\theta \in \Theta_{J,M}: \|\theta - \theta_{J,M}\|_2 \leq e^{-M} C_0^{-1}\sqrt{J}\epsilon\}$, *for* $\epsilon \geq d_H(p_0, p_{J,\theta_{J,M}})$.
(iii) $C_{J,M}(p_0, \epsilon) \subset B_{J,M}(p_0, D_0 M\epsilon)$, *if* $\|\log p_0\|_\infty \leq M$.

Proof The set $C_{J,M}(p_0, \epsilon)$ is void if $\epsilon < d_H(p_0, p_{J,\theta_{J,M}})$, so for the proof of (i) we may assume the opposite. Then the triangle inequality gives that $d_H(p_{J,\theta}, p_{J,\theta_{J,M}}) \leq 2\epsilon$, for any $p_{J,\theta} \in C_{J,M}(p_0, \epsilon)$. If also $2\epsilon < c_0 e^{-M}$, then the lower bound of Lemma 9.3 shows that $c_0 e^{-M}\|\theta - \theta_{J,M}\|_2 \leq 2\sqrt{J}\epsilon$. Statement (ii) follows similarly, and more easily, from the upper bound in Lemma 9.3 combined with the triangle inequality. Inclusion (iii) is a consequence of the bound $2M$ on $\|\log p_{J,\theta}\|_\infty$ for $\theta \in \Theta_{J,M}$, by Lemma 9.2, and the equivalence of Hellinger distance and Kullback-Leibler divergence and variation, for densities with bounded likelihood ratios, as quantified in Lemma B.2. □

Lemma 9.5 $\log N(\epsilon/5, C_{J,M}(p_0, \epsilon), d_H) \leq (2M + \log(30C_0/c_0))J$, *for every* $\epsilon > 0$.

Proof If $2\epsilon < c_0 e^{-M}$, then by Lemma 9.4(i) it suffices to bound the entropy of the set of functions $p_{J,\theta}$ with $\theta \in \{\theta \in \Theta_{J,M}: \|\theta - \theta_{J,M}\|_2 \leq 2\sqrt{J}\epsilon\, e^M/c_0\}$. By Lemma 9.3 on this set the Hellinger distance is bounded by $C_0 e^M/\sqrt{J}$ times the Euclidean norm. Therefore, the $\epsilon/5$-entropy in the lemma is bounded by the $\sqrt{J}e^{-M}\epsilon/(5C_0)$-entropy of a Euclidean ball of radius $2\sqrt{J}\epsilon\, e^M/c_0$, for the Euclidean metric, which follows from Proposition C.2.

The entropy at a value of $\epsilon/5$ with $2\epsilon \geq c_0 e^{-M}$ is bounded by the entropy at the values below this threshold, as just obtained. □

Lemma 9.6 *If* $\log p_0 \in \mathfrak{C}^\alpha[0, 1]$, *then for every J there exists $\theta_J \in \mathbb{R}^J$ with, for universal constants d_0, d_1,*

(i) $d_0\|\theta_J\|_\infty \leq \|\log p_0\|_\infty + d_1\|\log p_0\|_{\mathfrak{C}^\alpha} J^{-\alpha}$.
(ii) $d_H(p_{J,\theta_J}, p_0) \leq d_1\|\log p_0\|_{\mathfrak{C}^\alpha} J^{-\alpha} e^{d_1\|\log p_0\|_{\mathfrak{C}^\alpha} J^{-\alpha}}$.
(iii) $K(p_0; p_{J,\theta_J}) \leq d_1\|\log p_0\|_{\mathfrak{C}^\alpha} J^{-2\alpha} e^{d_1\|\log p_0\|_{\mathfrak{C}^\alpha} J^{-\alpha}}$.
(iv) $V_2(p_0; p_{J,\theta_J}) \leq d_1\|\log p_0\|_{\mathfrak{C}^\alpha} J^{-2\alpha} e^{d_1\|\log p_0\|_{\mathfrak{C}^\alpha} J^{-\alpha}}$.

Proof By Lemma E.5 there exists θ_J such that $\|\theta_J^\top B_J - \log p_0\|_\infty \leq DJ^{-\alpha}$, for D a universal multiple of $\|\log p_0\|_{\mathfrak{C}^\alpha}$. Inequalities (ii)–(iv) are next immediate from Lemma 2.5, with $f = \theta_J^\top B_J$ and $g = \log p_0$, where we increase the universal constant in D, if necessary. For (i) we apply the triangle inequality to obtain that $\|\theta_J^\top B_J\|_\infty \leq \|\log p_0\|_\infty + DJ^{-\alpha}$, and next combine this with Lemma E.6. □

9.2 Priors Based on Dirichlet Processes

In this section, we obtain posterior contraction rates for some priors based on Dirichlet processes.

Example 9.7 (Current status censoring) Consider the problem of estimating a cumulative distribution function F on $[0, \infty)$ when only indirect data is available: for a random sample $Y_1, \ldots, Y_n$ from F and an independent sample of "observation times" $C_1, \ldots, C_n$ we only observe the observation times C_i and the information $\Delta_i = \mathbb{1}\{Y_i \le C_i\}$ whether Y_i had realized or not. Thus our observations are a random sample $X_1, \ldots, X_n$ of pairs $X_i = (\Delta_i, C_i)$. Censored data models will be treated in more details in Chapter 13.

If the times C_i have distribution G with density g relative to a measure μ, then the density p_F of the X_i with respect to the product ν of counting measure on $\{0, 1\}$ and μ is given by

$$p_F(\delta, c) = F(c)^\delta (1 - F(c))^{1-\delta} g(c), \qquad \delta \in \{0, 1\}, c > 0. \tag{9.3}$$

Since this factorizes into expressions depending on F and g only, for a product prior on the pair (F, g) the factors involving g will cancel from the expression for the posterior distribution of F. Consequently, as we are interested in F, we may treat g as known, and need not specify a prior for it.

We assume that g is supported on a compact interval $[a, b]$, and that the true distribution F_0 is continuous and contains $[a, b]$ in the interior of its support (equivalently, $F_0(a-) > 0$ and $F_0(b) < 1$). We assign a Dirichlet prior $F \sim \mathrm{DP}(\alpha)$ with base measure α that has a positive, continuous density on a compact interval containing $[a, b]$. We shall show that the conditions of Theorem 8.9 are satisfied for ϵ_n a large multiple of $n^{-1/3}(\log n)^{1/3}$, and d the $\mathbb{L}_2$-metric on F. Let $\mathcal{F}$ stand for the set of all distribution functions on $(0, \infty)$.

Since $\|p_F - p_{F_0}\|_{2,\nu} = 2\|F - F_0\|_{2,G}$, we have,

$$N(\epsilon, \{p_F : F \in \mathcal{F}\}, \|\cdot\|_{2,\mu}) \le N(\epsilon/2, \mathcal{F}, \|\cdot\|_{2,G}).$$

The corresponding entropy is bounded above by a multiple of ϵ^{-1}, by Proposition C.8. Therefore the entropy condition (ii) of Theorem 8.9 with the $\mathbb{L}_2$-metric is verified for $\epsilon_n \asymp n^{-1/3}$.

Under our conditions F_0 is bounded away from zero and one on the interval $[a, b]$ that contains all observation times $C_1, \ldots, C_n$. Consequently, the quotients p_{F_0}/p_F are uniformly bounded away from zero and infinity, uniformly in F that are uniformly close to F_0 on the interval $[a, b]$. For such F,

$$\begin{aligned} d_H^2(p_F, p_{F_0}) &\le \int |F^{1/2} - F_0^{1/2}|^2 \, dG + \int |(1-F)^{1/2} - (1-F_0)^{1/2}|^2 \, dG \\ &\le C \sup_{c \in [a,b]} |F(c) - F_0(c)|^2, \end{aligned}$$

for a constant C that depends on F_0 only. Thus by (8.8) a neighborhood $B_2(p_0, \epsilon')$ as in the prior mass condition (i) of Theorem 8.9 contains a Kolmogorov-Smirnov neighborhood $\{F : d_{KS}(F, F_0) \le \epsilon\}$, for ϵ a multiple of ϵ'.

Given $\epsilon > 0$, partition the positive half line in intervals $E_1, \ldots, E_N$ such that $F_0(E_i) < \epsilon$ and $A\epsilon \le \alpha(E_i) < 1$ for every i and such that $N \le B\epsilon^{-1}$, for some fixed $A, B > 0$. If $\sum_{i=1}^N |F(E_i) - F_0(E_i)| < \epsilon$, then $|F(c) - F_0(c)| < \epsilon$ at every end point c of an interval E_i, and then $d_{KS}(F, F_0) < 2\epsilon$ by monotonicity and the fact that by construction F_0 varies by at most ϵ on each interval. Because $A\epsilon \le \alpha(E_i) < 1$ for every i, Lemma G.13 gives that $\Pi(F : \sum_{i=1}^N |F(E_i) - F_0(E_i)| < \epsilon) \ge \exp(-c\epsilon^{-1} \log \epsilon^{-1})$, for some $c > 0$. We conclude that the prior mass condition (i) is satisfied for $\epsilon_n \asymp n^{-1/3}(\log n)^{1/3}$.

The rate $n^{-1/3}(\log n)^{1/3}$ is close to the optimal rate of contraction $n^{-1/3}$ in this model. The small discrepancy may be an artifact of the prior mass estimation rather than a deficit of the Dirichlet prior. In Section 8.2.2 the optimal rate $n^{-1/3}$ was obtained for a prior based on bracketing approximation.

9.3 Bernstein Polynomials

Mixtures of beta densities are natural priors for the density p of observations $X_1, X_2, \ldots$ in the unit interval [0, 1]. In this section we consider the two types of mixtures:

$$\text{type I} \qquad b(x; k, w) = \sum_{j=1}^{k} w_{j,k}\,\text{be}(x; j, k-j+1),$$

$$\text{type II} \qquad \tilde{b}(x; k^2, w) = \sum_{i=1}^{k} w_{i,k}\,\frac{1}{k} \sum_{j=(i-1)k+1}^{ik} \text{be}(x; j, k^2-j+1).$$

In both types k is equal to the number of parameters $w_{j,k}$; the degrees of the beta polynomials are k and k^2, respectively. We can form priors for p by equipping k and the coefficients $w_{j,k}$ with priors. For instance,

$$w_k := (w_{1,k}, \ldots, w_{k,k})|\, k \sim \text{Dir}(k; \alpha_{1,k}, \ldots, \alpha_{k,k}), \qquad k \sim \rho.$$

In particular, in Example 5.10 it was suggested to link the Dirichlet priors across k, by letting $F \sim \text{DP}(\alpha)$, and setting $w_k = (F(0, 1/k], F(1/k, 2/k], \ldots, F((k-1)/k, 1])$, giving *Bernstein-Dirichlet processes*. For a given F the first type of mixture is then the derivative of the Bernstein polynomial corresponding to F. The second type of mixture is of interest, as it can repair the suboptimal approximation properties of these polynomials. The k^2 beta densities in these mixtures are grouped in k groups and the random coefficients $w_{j,k}$ are assigned to the averages of these groups rather than to the individual beta densities. See Lemma E.4 for further discussion.

The Bernstein polynomial prior is a role model for mixture priors for densities on bounded intervals with a discrete smoothing parameter (see Problem 9.20 for the example of a triangular density kernel). The prior on the smoothing parameter k is crucial for the concentration rate. Its natural discreteness makes that a point mass may be assigned to each value of the smoothing parameter.

Theorem 9.8 *Let the true density p_0 be strictly positive and be contained in $\mathfrak{C}^\alpha[0, 1]$, for $\alpha \in (0, 2]$. Assume that $A_1 k^{-b} \le \alpha_{j,k} \le A_2$ and that $B_1 e^{-\beta_1 k} \le \rho(k) \le B_2 e^{-\beta_2 k}$, for all j and k, where $A_1, A_2, b, B_1, B_2, \beta_1, \beta_2$ are fixed positive constants.*

(i) *If $\alpha \in (0, 2]$, then the posterior distribution relative to the type I mixtures contracts at the rate $n^{-\alpha/(2+2\alpha)}(\log n)^{(1+2\alpha)/(2+2\alpha)}$ with respect to the Hellinger distance.*

(ii) *If $\alpha \in (0, 1]$, then the posterior distribution relative to the type II mixtures contracts at the rate $n^{-\alpha/(1+2\alpha)}(\log n)^{(1+4\alpha)/(2+4\alpha)}$ with respect to the Hellinger distance.*

(iii) *If $p_0 = b(\cdot; k, w^0)$, for some $w^0 \in \mathbb{S}_k$, then the posterior distribution relative to the type I mixtures contracts at the rate $n^{-1/2}\log n$ with respect to the Hellinger distance.*

Proof We only give the details of the proof of (ii). The proof of the other parts is similar, and slightly simpler. The slower rate in (i) is due to the weaker approximation rate $k^{-\alpha/2}$ of order k type I Bernstein polynomials, as given in Lemma E.3, as opposed to the rate for type II mixtures, given in Lemma E.4.

For given $w_k \in \mathbb{S}_k$ we have $\tilde{b}(\cdot; k^2, w_k) = b(\cdot; k^2, \tilde{w}_k)$, for $\tilde{w}_k$ the vector in $\mathbb{S}_{k^2}$ given by $(w_{k,1}, \ldots, w_{k,1}, w_{k,2}, \ldots, w_{k,2}, \ldots, w_{k,k})/k$, obtained by repeating each coordinate $w_{k,j}$ of w_k exactly k times and renormalizing. Therefore, Lemma E.2 implies that, for any $w_k, w_k' \in \mathbb{S}_k$,

$$\|\tilde{b}(\cdot; k^2, w_k) - \tilde{b}(\cdot; k^2, w_k')\|_\infty \leq k^2 \|w_k - w_k'\|_1. \tag{9.4}$$

We now verify the conditions of Theorem 8.9.

For $w_k^0 \in \mathbb{S}_k$ the vector with coordinates $w_{k,j}^0 = P_0((j-1)/k, j/k]$, for $j = 1, \ldots, k$, Lemma E.4 gives that $\|p_0 - \tilde{b}(\cdot; k^2, w_k^0)\|_\infty \lesssim k^{-\alpha}$. Thus, by the triangle inequality, for any $w_k \in \mathbb{S}_k$,

$$\|p_0 - \tilde{b}(\cdot; k^2, w_k)\|_\infty \lesssim k^{-\alpha} + k^2 \|w_k - w_k^0\|_1.$$

If we choose $k = k_n$ the integer part of a big multiple of $\epsilon_n^{-1/\alpha}$, then the right side will be smaller than a multiple of ϵ_n, for any w_k with $\|w_k - w_k^0\|_1 \leq \epsilon_n^{1+2/\alpha}$. In particular, the function $\tilde{b}(\cdot; k^2, w_k)$ will be bounded away from zero, for sufficiently large n, and we can apply Lemma B.1 (v) and (vii) and the last two assertions of Lemma B.2 to obtain that $\tilde{b}(\cdot; k^2, w_k) \in B_2(p_0, C_1\epsilon_n)$, for B_2 the Kullback-Leibler neighborhood defined in (8.3). It follows that the left side of the prior mass condition (8.4) is bounded below by

$$\rho(k_n)\, \Pi(w_{k_n}\colon \|w_{k_n} - w_{k_n}^0\|_1 \leq \epsilon_n{}^{1+2/\alpha}) \gtrsim e^{-\beta_1 k_n} e^{-c_1 k_n \log_- \epsilon_n},$$

for some $c_1 > 0$, in view of Lemma G.13. (The condition $\epsilon \leq 1/(Mk)$ in the lemma is easily satisfied, by virtue of the relation $\epsilon_n^{1+2/\alpha} \lesssim k_n^{-1}$.) For $\epsilon_n = n^{-\alpha/(1+2\alpha)}(\log n)^{\alpha/(1+2\alpha)}$ the right side is bounded below by a constant multiple of $e^{-c_2 n\epsilon_n^2}$, for some constant $c_2 > 0$, whence (8.4) is satisfied.

To verify (8.5) we use the sieve $\mathcal{P}_{n,1} = \cup_{j=1}^{s_n} \mathcal{C}_j$, where $\mathcal{C}_k = \{\tilde{b}(\cdot; k^2, w_k)\colon w_k \in \mathbb{S}_k\}$, and s_n is the integer part of a multiple of $n^{1/(1+2\alpha)}(\log n)^{2\alpha/(1+2\alpha)}$. In view of (9.4) we have

$$D(\epsilon, \mathcal{C}_k, \|\cdot\|_1) \leq D(\epsilon/k^2, \mathbb{S}_k, \|\cdot\|_1) \leq \left(\frac{5k^2}{\epsilon}\right)^k,$$

for $\epsilon/k^2 < 1$, by Proposition C.1. Because the Hellinger distance is bounded above by the root of the $\mathbb{L}_1$-distance, it follows that $D(\epsilon, \mathcal{P}_{n,1}, d_H) \leq \sum_{k=1}^{s_n} (5k^2/\epsilon^2)^k \leq s_n (5s_n^2/\epsilon^2)^{s_n}$, for $\epsilon < 1$. Thus condition (8.5) is satisfied for $\bar{\epsilon}_n = n^{-\alpha/(1+2\alpha)}(\log n)^{(1+4\alpha)/(2+4\alpha)}$.

Finally, since $\Pi(\mathcal{P}_{n,1}^c) = \sum_{k>s_n} \rho(k) \lesssim e^{-\beta_2 s_n}$, condition (8.6) is verified by the choice of s_n. □

Part (ii) of the theorem shows that the type II mixtures give the minimax rate up to a logarithmic factor, while part (i) suggests a suboptimal rate. Part (ii) of the theorem is restricted to $\alpha \leq 1$, but for $\alpha = 1$ it gives a better contraction rate than part (i) for $1 \leq \alpha < 2$. Thus part (ii) gives a better result than part (i) even for $\alpha \in [1, 2)$.

9.4 Dirichlet Process Mixtures of Normal Kernel

Mixtures of normal densities are popular for estimating densities on a full Euclidean space. Finite mixtures are appropriate for a population that consists of finitely many subclasses. Kernel density estimators are often normal mixtures centered at the observations. In the Bayesian setup nonparametric normal mixtures with a Dirichlet process prior on the mixing distribution lead to elegant computational algorithms, as seen in Chapter 5, and turn out to be remarkably flexible.

In this section we study the contraction rate of the posterior distribution of the density p of an i.i.d. sample $X_1, \ldots, X_n$ in $\mathbb{R}^d$ for a Dirichlet mixture of normal prior, in two scenarios for the true density: either the true density itself is a mixture of normal densities, or it belongs to a Hölder class of functions. In the first situation, the "supersmooth case," a nearly parametric rate of contraction is obtained, while in the second situation, the "ordinary smooth case," a nonparametric rate prevails. Remarkly the same Dirichlet mixture of normal prior yields an almost minimax contraction rate in either situation, and in the ordinary smooth case also for any smoothness level $\beta > 0$. This is in striking contrast with classical kernel density estimation, where the normal kernel, which is a second order kernel, gives the minimax rate only for smoothness up to order 2, and better rates require higher-order kernels that are not probability densities. The Bayesian procedure achieves this by spreading the mass of the posterior mixing distribution in a clever way, rather than putting equal masses at the observations. A key lemma in the derivations below is that a β-smooth density can be approximated closely by a normal mixture of the form $f_\sigma * \phi_{0,\sigma^2 I}$, where f_σ belongs to the support of prior and is different from the usual choice p_0, for which the approximation rate is only σ^2 (see Problem 9.10).

Let $\phi(\cdot; \Sigma)$ be the density of the $\mathrm{Nor}_d(0, \Sigma)$-distribution, where Σ is a positive-definite $(d \times d)$-matrix, and denote the normal mixture with mixing distribution F on $\mathbb{R}^d$ by

$$p_{F,\Sigma}(x) = \int \phi(x - z; \Sigma)\, dF(z).$$

We induce a prior on densities on $\mathbb{R}^d$ by equipping F with a $\mathrm{DP}(\alpha)$ prior and independently Σ with a prior G. We assume that the base measure α and the measure G have continuous and positive densities in the interior of their supports, and satisfy, for positive constants $a_1, a_2, a_3, a_4, a_5, b_1, b_2, b_3, b_4$, and C_1, C_2, C_3, κ,

$$1 - \bar\alpha([-z, z]^d) \leq b_1 e^{-C_1 z^{a_1}}, \qquad z > 0, \tag{9.5}$$

$$G(\Sigma\colon \mathrm{eig}_d(\Sigma^{-1}) \geq s) \leq b_2 e^{-C_2 s^{a_2}}, \qquad s > 0, \tag{9.6}$$

$$G(\Sigma\colon \mathrm{eig}_1(\Sigma^{-1}) \leq s) \quad \leq b_3 s^{a_3}, \qquad s > 0, \tag{9.7}$$

$$G\Bigl(\Sigma\colon \bigcap_{1 \leq j \leq d} \{s_j < \mathrm{eig}_j(\Sigma^{-1}) < s_j(1+t)\}\Bigr) \geq b_4 s_1^{a_4} t^{a_5} e^{-C_3 s_d^{\kappa/2}}, 0 < s_1 \leq \cdots \leq s_d,$$
$$0 < t < 1. \tag{9.8}$$

Here $\mathrm{eig}_1(\Sigma) \leq \cdots \leq \mathrm{eig}_d(\Sigma)$ denote the ordered eigenvalues of a matrix Σ.

If Σ is a diagonal matrix $\mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ with $\sigma_j \overset{\mathrm{iid}}{\sim} G_0$, then (9.6) and (9.7) hold if G_0 has a polynomial tail at 0 and an exponential tail at infinity. Furthermore, (9.8) is satisfied with $\kappa = 2$ if G_0 is an inverse-gamma distribution, and with $\kappa = 1$ if G_0 is the distribution of the square of an inverse-gamma random variable. In Lemma 9.16 inequalities (9.6)–(9.8) are shown to hold with $\kappa = 2$ if Σ^{-1} is distributed according to a Wishart distribution with positive-definite scale matrix.

For a multi-index $k = (k_1, \ldots, k_d)$ of nonnegative integers k_i, define $k. = \sum_{j=1}^d k_j$, and let $D^k = \partial^{k.}/\partial x_1^{k_1} \cdots \partial x_d^{k_d}$ denote the mixed partial derivative operator. We assume that the true density p_0 of the observations satisfies one of the two conditions:

- *Supersmooth case*: $p_0(x) = \int \phi(x - z; \Sigma_0)\, dF_0(z)$, for some positive-definite matrix Σ_0, and a probability measure F_0 on $\mathbb{R}^d$ satisfying $1 - F_0([z, z]^d) \lesssim e^{-c_0 z^{r_0}}$, for all $z > 0$ and some fixed $c_0 > 0$ and $r_0 \geq 2$. (The case that F_0 is compactly supported can be recovered by formally substituting $r_0 = \infty$.)
- *β-smooth case*: p_0 has mixed partial derivatives $D^k p_0$ of order up to $k. \leq \underline{\beta} := \lceil \beta - 1 \rceil$, satisfying for a function $L\colon \mathbb{R}^d \to [0, \infty)$,

$$|(D^k p_0)(x + y) - (D^k p_0)(x)| \leq L(x)\, e^{c_0 \|y\|^2} \|y\|^{\beta - \underline{\beta}}, \qquad k. = \underline{\beta},\ x, y \in \mathbb{R}^d, \tag{9.9}$$

$$P_0\Big[\Big(\frac{L}{p_0}\Big)^2 + \Big(\frac{|D^k p_0|}{p_0}\Big)^{2\beta/k.}\Big] < \infty, \qquad k. \leq \underline{\beta}. \tag{9.10}$$

Furthermore, $p_0(x) \leq c e^{-b\|x\|^\tau}$, for every $\|x\| > a$, for positive constants a, b, c, τ.

Theorem 9.9 *If the Dirichlet mixture of normal prior satisfies conditions* (9.5)–(9.8) *and the true density p_0 is either supersmooth with $r_0 \geq 2$ such that $a_1 > r_0(d + 1) + 1$, or β-smooth, then the posterior distribution contracts at the rate $\epsilon_n = n^{-1/2}(\log n)^{(d+1+1/r_0)/2}$ in the supersmooth case and at the rate $\epsilon_n = n^{-\beta/(2\beta+d^*)}(\log n)^t$ in the β-smooth case, where $d^* = d \vee \kappa$ and $t > (\beta d^* + \beta d^*/\tau + d^* + \beta)/(2\beta + d^*)$, relative to the Hellinger distance.*

In the ordinary smooth case the minimax rate $n^{-\beta/(2\beta+d)}$ is attained up to a logarithmic factor provided that $\kappa \leq d$. This is true for the inverse-Wishart prior on Σ if $d \geq 2$, but not for $d = 1$. The prior based on diagonal Σ with each entry distributed as the square of an inverse-gamma variable (interestingly, not the inverse-gamma, which is conjugate) gives the optimal rate for all dimensions, including $d = 1$.

An essentially identical posterior contraction rate is obtained with a prior on the mixing distribution supported on a random number of random points with random weights, instead of the Dirichlet process prior (see Problem 9.16). Such a prior is intuitively appealing, but posterior computation may need to involve a reversible jump MCMC procedure. The advantage of the generalized Pólya urn scheme is available only for the Dirichlet process and similarly structured priors.

The theorem can be extended to anisotropic smoothness classes, where the smoothness level is different for different coordinates. A naive application of Theorem 9.9 would only give a rate based on the minimal smoothness level of the coordinates, and hence would not take advantage of higher smoothness in other components. A generalization of the theorem, presented in Problem 9.17, gives a rate corresponding to the harmonic mean of the smoothness levels for the different coordinates. This rate coincides with the minimax rate for an anisotropic class (up to a logarithmic factor).

The remainder of this section is devoted to the proof of the theorem, which encompasses approximation results, entropy and prior mass bounds, and an application of the basic contraction theorem, Theorem 8.9 with suitable sieves $\mathcal{P}_{n,1}$.

9.4.1 Approximation

In this section we prove first that any smooth density can be closely approximated by a normal mixture with appropriate bandwidth, and second that general normal mixtures can be closely approximated by finite normal mixtures with few components.

For a nonnegative multi-integer $k = (k_1, \ldots, k_d)$, let $m_k = \int \prod_{j=1}^d x_j^{k_j}\, \phi(x; I)\, dx$ denote the kth mixed moment of the standard normal distribution on $\mathbb{R}^d$; in particular $m_k = 0$ if one or more of the coordinates of k are odd. Recursively define sequences c_n and d_n by setting $c_n = d_n = 0$ for $n_. = 1$ and, for $n_. \geq 2$ and $k! = \prod_{j=1}^d k_j!$,

$$c_n = -\sum_{\substack{l+k=n \\ l_. \geq 1, k_. \geq 1}} \frac{(-1)^{k_.}}{k!} m_k d_l, \qquad d_n = \frac{(-1)^{n_.} m_n}{n!} + c_n. \tag{9.11}$$

For positive constants β, c and a function $L\colon \mathbb{R}^d \to [0, \infty)$, let $\mathfrak{C}^{\beta,L,c}(\mathbb{R}^d)$ be the set of functions $f\colon \mathbb{R}^d \to \mathbb{R}$ whose mixed partial derivatives of order $k_. = \underline{\beta}$ exist and satisfy

$$|(D^k f)(x+y) - (D^k f)(x)| \leq L(x)\, e^{c\|y\|^2} \|y\|^{\beta - \underline{\beta}}, \qquad x, y \in \mathbb{R}^d.$$

For f belonging to this set and $\sigma > 0$, define, with the sum ranging over nonnegative multi-integers $k = (k_1, \ldots, k_d)$,

$$T_{\beta,\sigma} f = f - \sum_{2 \leq k_. \leq \underline{\beta}} d_k \sigma^{k_.} D^k f.$$

Lemma 9.10 *For any $\beta, c > 0$ there exists a positive constant M such that $|\phi(\cdot; \sigma^2 I) * T_{\beta,\sigma} f(x) - f(x)| \leq M L(x) \sigma^\beta$, for all $x \in \mathbb{R}^d$ and all $\sigma \in (0, (3c)^{-1/2})$, and any $f \in \mathfrak{C}^{\beta,L,c}(\mathbb{R}^d)$.*

Proof Abbreviate $\phi(\cdot; \sigma^2 I)$ to ϕ_σ. A multivariate Taylor expansion allows to decompose $f(x-y) - f(x) = \sum_{1 \leq k_. \leq \underline{\beta}} (-y)^{k_.}/k!(D^k f)(x) + R(x,y)$, where the remainder satisfies $|R(x,y)| \lesssim L(x) \exp(c\|y\|^2)\|y\|^\beta$, for $x, y \in \mathbb{R}^d$. Consequently, the difference $\phi_\sigma * T_{\beta,\sigma} f(x) - f(x)$ can be written in the form

$$\int \phi_\sigma(y)(f(x-y) - f(x))\, dy - \sum_{2 \leq k_. \leq \underline{\beta}} d_k \sigma^{k_.} (\phi_\sigma * D^k f)(x)$$

$$= \int \phi_\sigma(y) R(x,y)\, dy + \sum_{2 \leq k_. \leq \underline{\beta}} \Big[\frac{(-\sigma)^{k_.} m_k}{k!} (D^k f)(x) - d_k \sigma^{k_.} (\phi_\sigma * D^k f)(x) \Big].$$

For $\sigma \in (0, (3c)^{-1/2})$ the first term is bounded by a multiple of $L(x)\sigma^{\beta}$. For $\underline{\beta} < 2$, the second sum is empty, and the result follows. For $\underline{\beta} \geq 2$, we decompose the second term as

$$\sum_{2\leq k.\leq\underline{\beta}} \frac{(-\sigma)^{k.} m_k}{k!}\big[D^k f - \phi_\sigma * T_{\beta-k.,\sigma} D^k f\big]$$
$$+ \sum_{2\leq k.\leq\underline{\beta}} \phi_\sigma * \Big[\frac{(-\sigma)^{k.} m_k}{k!} T_{\beta-k.,\sigma} D^k f \quad - d_k \sigma^{k.} D^k f\Big].$$

The function $D^k f - \phi_\sigma * T_{\beta-k.,\sigma} D^k f$ in the first sum is of the same form as in the lemma, but with the function $D^k f$ instead of f, which is approximated with the help of $T_{\beta-k.,\sigma}$. Since $D^k f \in \mathfrak{C}^{\beta-k.,L,c}(\mathbb{R}^d)$, an induction argument on $\underline{\beta}$ can show that it is bounded above by a multiple of $L(x)\sigma^{\beta-k.}$, for every $x \in \mathbb{R}^d$. The second sum is actually identically zero. To see this we substitute the definitions of $T_{\beta-k.,\sigma}$ and d_k, and see that the sum of the functions within square brackets, without the convolution, is identical to

$$\sum_{2\leq k.\leq\underline{\beta}} \Big[\frac{-(-\sigma)^{k.} m_k}{k!} \sum_{1\leq j.\leq\underline{\beta}-k.} d_j \sigma^{j.} D^{j+k} f - c_k \sigma^{k.} D^k f\Big]$$
$$= \sum_{3\leq n.\leq\underline{\beta}} \Big[\sum_{\substack{j+k=n\\ j.\geq 1, k.\geq 2}} \frac{(-1)^{k.+1}}{k!} m_k d_j - c_n\Big]\sigma^{n.} D^n f.$$

The right side vanishes by the definition (9.11) of c_n. □

When Lemma 9.10 is applied to a probability density p_0, then the resulting approximation $T_{\beta,\sigma} p_0$ is not necessarily a probability density. This is remedied in the following result, which also gives approximation in the Hellinger distance.

Lemma 9.11 (Smooth approximation) *For any probability density p_0 satisfying* (9.9)–(9.10) *there exist constants a_0, τ, K, s_0 and for any $0 < \sigma < s_0$ a probability density f_σ supported within the interval $[-a_0(\log_- \sigma)^{1/\tau}, a_0(\log_- \sigma)^{1/\tau}]$ such that $d_H(p_0, \phi(\cdot;\sigma^2 I) * f_\sigma) \leq K\sigma^{\beta}$.*

Proof Let λ stand for the Lebesgue measure. Write T_σ for $T_{\beta,\sigma}$ and set

$$f_\sigma = \frac{|T_\sigma p_0|\mathbb{1}\{E_\sigma\}}{\int_{E_\sigma} |T_\sigma p_0|\, d\lambda}, \qquad E_\sigma = \big[-a_0(\log_- \sigma)^{1/\tau}, a_0(\log_- \sigma)^{1/\tau}\big].$$

This first makes the approximation given in Lemma 9.10 nonnegative, and next truncates it to the interval E_σ. We shall show that the effects of these operations are small.

Because $\int D^k p_0\, d\lambda = 0$ for $k \neq 0$, the definition of T_σ gives that $\int T_\sigma p_0\, d\lambda = 1$. We shall first prove that $\int |T_\sigma p_0|\, d\lambda = 1 + O(\sigma^{2\beta})$. By the definition of T_σ,

$$|T_\sigma p_0 - p_0| \leq \sum_{k.\leq\underline{\beta}} |d_k| \sigma^{k.} |D^k p_0|.$$

Thus $|T_\sigma p_0 - p_0| \lesssim \eta p_0$ on the set $A_\sigma = \{\sigma^{k.}|D^k p_0| \le \eta p_0, \forall k. \le \underline{\beta}\}$, whence $T_\sigma p_0 \ge 0$, for sufficiently small $\eta > 0$. Since also $|T_\sigma p_0| - T_\sigma p_0 \le 2|T_\sigma p_0 - p_0|$ by the nonnegativity of p_0,

$$\int (|T_\sigma p_0| - T_\sigma p_0)\, d\lambda = \int_{A_\sigma^c} (|T_\sigma p_0| - T_\sigma p_0)\, d\lambda \le 2\int_{A_\sigma^c} |T_\sigma p_0 - p_0|\, d\lambda,$$
$$\le 2\sum_{k.\le\underline{\beta}} |d_k|\sigma^{k.}\, \|D^k p_0/p_0\|_{2\beta/k.,p_0} P_0(A_\sigma^c)^{1-k./(2\beta)},$$

by the preceding display and Hölder's inequality. The definition of A_σ^c, the moment condition (9.10), and Markov's inequality give that $P_0(A_\sigma^c) \lesssim \sigma^{2\beta}$. Thus the preceding display is of this same order, completing the proof that $\int |T_\sigma p_0|\, d\lambda = 1 + O(\sigma^{2\beta})$.

Next

$$\begin{aligned} d_H^2(p_0, \phi_\sigma * |T_\sigma p_0|) &\le \int \frac{(p_0 - \phi_\sigma * |T_\sigma p_0|)^2}{p_0 + \phi_\sigma * |T_\sigma p_0|}\, d\lambda \\ &\le 2\int \frac{(p_0 - \phi_\sigma * T_\sigma p_0)^2}{p_0}\, d\lambda + 2\int \frac{(\phi_\sigma * T_\sigma p_0 - \phi_\sigma * |T_\sigma p_0|)^2}{\phi_\sigma * |T_\sigma p_0|}\, d\lambda \\ &\lesssim \int \frac{L^2}{p_0}\sigma^{2\beta}\, d\lambda + \int \phi_\sigma * (|T_\sigma p_0| - T_\sigma p_0)\, d\lambda, \end{aligned}$$

where we use Lemma 9.10 for the first term, and the fact that $|\phi_\sigma * (g - |g|)| \le \phi_\sigma * (|g| - g) \le 2\phi_\sigma * |g|$, for any measurable function g, in the second term. Since convolution retains total integral, the second term further evaluates to $\int (|T_\sigma p_0| - T_\sigma p_0)\, d\lambda$, which was seen to be also of the order $O(\sigma^{2\beta})$.

By convexity of the function $(u, v) \mapsto (\sqrt{u} - \sqrt{v})^2$, we have that $|(\phi_\sigma * f)^{1/2} - (\phi_\sigma * g)^{1/2}|^2 \le \phi_\sigma * (\sqrt{f} - \sqrt{g})^2$, for any nonnegative measurable functions f and g. Therefore

$$d_H^2(\phi_\sigma * |T_\sigma p_0|, \phi_\sigma * f_\sigma) \le d_H^2(|T_\sigma p_0|, f_\sigma) = \int |T_\sigma p_0|\, d\lambda + 1 - 2\Big(\int_{E_\sigma} |T_\sigma p_0|\, d\lambda\Big)^{1/2}.$$

The first term on the right side was seen to be $1 + O(\sigma^{2\beta})$. By the tail condition on p_0, we have $P_0(E_\sigma^c) = O(\sigma^{2\beta})$ if a_0 is chosen sufficiently large (easily, with a bigger power of σ if a_0 is chosen larger). By the same argument as used for A_σ^c we obtain that $\int_{E_\sigma^c} |T_\sigma p_0 - p_0|\, d\lambda = O(\sigma^{2\beta})$ and hence $\int_{E_\sigma} |T_\sigma p_0|\, d\lambda = 1 + O(\sigma^{2\beta})$. It follows that the right side of the display is $O(\sigma^{2\beta})$. □

General normal mixtures can be approximated by discrete normal mixtures with a small number of support points.

Lemma 9.12 (Finite approximation) *Given a probability measure F_0 on $[-a, a]^d$, a positive-definite matrix Σ, and $\epsilon > 0$, there exists a discrete probability measure F^* on $[-a, a]^d$ with no more than $D(a/\underline{\sigma} \vee 1)^d(\log_- \epsilon)^d$ support points such that $\|p_{F_0,\Sigma} - p_{F^*,\Sigma}\|_\infty \lesssim \epsilon/\underline{\sigma}^d$ and $\|p_{F_0,\Sigma} - p_{F^*,\Sigma}\|_1 \lesssim \epsilon(\log_- \epsilon)^{d/2}$, where D is a constant that depends on d and $\underline{\sigma}$ is the smallest eigenvalue of $\Sigma^{1/2}$. Without loss of generality the support points can be chosen in the set $\underline{\sigma}\epsilon\, \mathbb{Z}^d \cap [-a, a]^d$.*

Proof The identity $p_{F,\Sigma}(x) = \det\Sigma^{-1/2} p_{F\circ\Sigma^{-1},I}(\Sigma^{-1/2}x)$ gives that the distances in the lemma are bounded above by $\|p_{F_0\circ\Sigma^{-1/2},I} - p_{F^*\circ\Sigma^{-1/2},I}\|_1$ and $\underline{\sigma}^{-d}\|p_{F_0\circ\Sigma^{-1/2},I} - p_{F^*\circ\Sigma^{-1/2},I}\|_\infty$, respectively. The measures $F\circ\Sigma^{-1/2}$ concentrate on the set $\Sigma^{-1/2}[-a,a]^d \subset [-\sqrt{d}\,a/\underline{\sigma}, \sqrt{d}\,a/\underline{\sigma}]^d$, since $\|\Sigma^{-1/2}x\|_\infty \le \|\Sigma^{-1/2}x\|_2 \le \|x\|_2/\underline{\sigma} \le \sqrt{d}\,\|x\|_\infty/\underline{\sigma}$. Thus the problem can be reduced to mixtures of the standard normal kernel relative to mixing distributions on $[-\sqrt{d}\,a/\underline{\sigma}, \sqrt{d}\,a/\underline{\sigma}]^d$.

We can partition the latter cube into fewer than $D_1(a/\underline{\sigma}\vee 1)^d$ rectangles $I_1,\ldots,I_k$ with sides of length at most 1. Decomposing a probability measure on the cube as $F = \sum_{i=1}^k F(I_i)F_i$, where each F_i is a probability measure on I_i, we have $p_{F,I} = \sum_{i=1}^k F(I_i)p_{F_i,I}$. We shall show that for each $i = 1,\ldots,k$ there exists a discrete distribution F_i^* on I_i with at most $D_2(\log_-\epsilon)^d$ many support points such that the $\mathbb{L}_\infty$- and $\mathbb{L}_1$-norms between $p_{F_i^*,I}$ and $p_{F_{0,i},I}$ are bounded above by ϵ and $\epsilon(\log_-\epsilon)^{d/2}$. Then $F^* = \sum_{i=1}^k F_0(I_i)F_i^*$ will be the appropriate approximation of F_0. Because we can shift the rectangles I_i to the origin and the two distances are invariant under shifting, it is no loss of generality to construct the approximation only for I_i equal to the unit cube. For simplicity of notation, consider a probability measure F_0 on $[0,1]^d$ and the mixture $p_{F_0,I}$.

For $\|x\|_\infty \ge \sqrt{8\log_-\epsilon}$ and sufficiently small ϵ, we have that $\phi(x-z)\le\epsilon$ for all $z\in[0,1]^d$, so that $p_{F,I}(x)\le\epsilon$ for every F concentrated on $[0,1]^d$. For $\|x\|_\infty\le\sqrt{8\log_-\epsilon}$ Taylor's expansion of the exponential function, gives

$$\phi(x-z) = \frac{1}{(2\pi)^{d/2}}\prod_{j=1}^d\left[\sum_{r=0}^{k-1}\frac{\big[-(x_j-z_j)^2/2\big]^r}{r!} + R(x_j-z_j)\right],$$

where the remainder satisfies $|R(x)|\le(x^2/2)^k/k!$. If $|x_j|\le\sqrt{8\log_-\epsilon}$ and $|z_j|\le 1$, then $|R(x_j-z_j)|\le(8e\log_-\epsilon/k)^k$, in view of the inequality $k!\ge k^k e^{-k}$, which will be bounded by 1 for sufficiently large k. Since the univariate standard normal density is uniformly bounded, it follows that the sums, which are the differences of the density and the remainder terms, are also uniformly bounded. Hence the remainder can be pulled out of the product, and for $\|x\|_\infty\le\sqrt{8\log_-\epsilon}$ we can write $\phi(x-z)$ as

$$\frac{1}{(2\pi)^{d/2}}\prod_{j=1}^d\left[\sum_{r=0}^{k-1}\frac{(-1)^r}{2^r r!}\sum_{l=0}^{2r}\binom{2r}{l}x_j^{2r-l}z_j^l\right] + \bar{R}(x-z), \qquad |\bar{R}|\lesssim\left|\frac{(8e\log_-\epsilon)}{k}\right|^k.$$

Let F^* be a probability measure on $[0,1]^d$ such that, for integers l_j,

$$\int z_1^{l_1}\cdots z_d^{l_d}\,dF^*(z) = \int z_1^{l_1}\cdots z_d^{l_d}\,dF_0(z), \qquad 0\le l_1,\ldots,l_d\le 2k-2,$$

Then integrating the second last display with respect to F^*-F_0, giving $p_{F^*,I}-p_{F_0,I}$, leaves only the integral over $\bar{R}$. For k the smallest integer exceeding $(1+c^2)\log_-\epsilon$, the latter integral is bounded by a multiple of ϵ. The preceding display requires matching $(2k-1)^d$ expectations, and hence F^* can be chosen a discrete distribution with at most $(2k-1)^d+1$ support points, by Lemma L.1.

This concludes the proof for the $\mathbb{L}_\infty$-norm. For the $\mathbb{L}_1$-norm we note that $p_{F,I}(x)\le e^{-x_j^{2/8}}$ if $|x_j|\ge 2$, for any probability measure F on $[0,1]^d$, whence $\int_{\|x\|_\infty\ge T} p_{F,I}(x)\,dx \le$

$2^d e^{-T^2/8}$, for sufficiently large T. Letting F^* as for the $\mathbb{L}_\infty$-bound, we have that $|p_{F^*,I} - p_{F_0,I}| \lesssim \epsilon$ on $[-T, T]^d$. Integrating this and combining with the first bound we see that $\|p_{F^*,I} - p_{F_0,I}\|_1 \lesssim e^{-T^2/8} + T^d\epsilon$. We choose T a suitable multiple of $(\log_- \epsilon)^{1/2}$ to establish the second assertion of the lemma.

The final assertion of the lemma follows from the fact that moving the support points of F^* to a closest point in the given lattice increases the $\mathbb{L}_1$-error by at most a multiple of ϵ, as $z \mapsto \phi(x - z; \Sigma)$ is Lipschitz continuous with constant $1/\underline{\sigma}$ relative to the $\mathbb{L}_1$-norm; and increases the $\mathbb{L}_\infty$-error by at most $\epsilon/\underline{\sigma}^d$, as $\|\phi'(\cdot; \Sigma)\|_\infty \lesssim 1/\underline{\sigma}^{d+1}$. □

9.4.2 Prior Concentration

In this section we estimate the prior mass of a ball $B_2(p_0, \epsilon)$, as in (8.3), around a true density from below.

Lemma 9.13 *For a measurable partition $\mathbb{R}^d = \cup_{j=0}^N U_j$ and points $z_j \in U_j$, for $j = 1, \ldots, N$, let $F^* = \sum_{j=1}^N w_j \delta_{z_j}$ be a probability measure. Then, for any probability measure F on $\mathbb{R}^d$ and any positive-definite matrix Σ with eigenvalues bounded below by $\underline{\sigma}^2$,*

$$\|p_{F,\Sigma} - p_{F^*,\Sigma}\|_\infty \lesssim \frac{1}{\underline{\sigma}^{d+1}} \max_{1\le j\le N} \operatorname{diam}(U_j) + \frac{1}{\underline{\sigma}^d} \sum_{j=1}^N |F(U_j) - w_j|,$$

$$\|p_{F,\Sigma} - p_{F^*,\Sigma}\|_1 \lesssim \frac{1}{\underline{\sigma}} \max_{1\le j\le N} \operatorname{diam}(U_j) + \sum_{j=1}^N |F(U_j) - w_j|.$$

Proof We can decompose $p_{F,\Sigma}(x) - p_{F^*,\Sigma}(x)$ as

$$\int_{U_0} \phi(x - z; \Sigma)\, dF(z) + \sum_{j=1}^N \int_{U_j} \big[\phi(x - z; \Sigma) - \phi(x - z_j; \Sigma)\big]\, dF(z)$$
$$+ \sum_{j=1}^N \phi(x - z_j; \Sigma)\big[F(U_j) - w_j\big].$$

Here the mass of the set $U_0 = \mathbb{R}^d \setminus \cup_{j\ge1} U_j$ is bounded as $F(U_0) \le \sum_{j=1}^N |F(U_j) - w_j|$, since $(w_1, \ldots, w_N)$ is assumed to be a probability vector. The inequality for the $\mathbb{L}_\infty$-norm next follows from the estimates $\|\phi(\cdot; \Sigma)\|_\infty \lesssim \underline{\sigma}^{-d}$ and $\|\phi'(\cdot; \Sigma)\|_\infty \lesssim \underline{\sigma}^{-(d+1)}$, while for the inequality for the $\mathbb{L}_1$-norm we employ the estimates $\|\phi(\cdot - z; \Sigma) - \phi(\cdot - z'; \Sigma)\|_1 \lesssim \underline{\sigma}^{-1}\|z - z'\|_\infty$ and $\|\phi(\cdot; \Sigma)\|_1 = 1$. □

Proposition 9.14 (Prior mass) *If α has a continuous, positive density and satisfies* (9.5), *then there exist constants $A, C > 0$ such that* $(\mathrm{DP}_\alpha \times G)((F, \Sigma)\colon p_{F,\Sigma} \in B_2(p_0, A\epsilon_n)) \ge e^{-Cn\epsilon_n^2}$, *where ϵ_n is given by*

(i) $n^{-1/2}(\log n)^{(d+1+1/r_0)/2}$ *if $p_0 = p_{F_0,\Sigma_0}$ is supersmooth with Σ_0 an interior point of the support of G,*

(ii) $n^{-\beta/(2\beta+d^*)}(\log n)^{t_0}$ *if* p_0 *is* β*-smooth and* G *satisfies* (9.6)–(9.8)*, where* $d^* = d \vee \kappa$ *and* $t_0 = (\beta d^* + \beta d^*/\tau + d^* + \beta)/(2\beta + d^*)$.

Proof (i). The assumption on the tail of F_0 gives that $1 - F_0([-a,a]^d) \le \epsilon^2$, for $a = a_0(\log_- \epsilon)^{1/r_0}$, sufficiently small $\epsilon > 0$ and large a_0, whence the renormalized restriction $\tilde F_0$ of F_0 to $[-a,a]^d$ satisfies $\|p_{F_0,\Sigma_0} - p_{\tilde F_0,\Sigma_0}\|_1 \le \epsilon^2$, by Lemma K.10. Next Lemma 9.12 gives a discrete distribution $F^* = \sum_{j=1}^N w_j \delta_{z_j}$ on $[-a,a]^d$ with at most $N \lesssim a(\log_- \epsilon)^d \lesssim (\log_- \epsilon)^{d+1/r_0}$ support points such that $\|p_{\tilde F_0,\Sigma_0} - p_{F^*,\Sigma_0}\|_1 \lesssim \epsilon^2$. The points z_j can be chosen ϵ^2-separated without loss of generality, whence they possess disjoint neighborhoods $U_j \subset [-a,a]^d$ of diameter of the order ϵ^2. If F is a probability measure with $\sum_{j=1}^N |F(U_j) - w_j| \le \epsilon^2$, then $\|p_{F^*,\Sigma_0} - p_{F,\Sigma_0}\|_1 \lesssim \epsilon^2$, by Lemma 9.13.

Combining the preceding with the triangle inequality gives $\|p_{F_0,\Sigma_0} - p_{F,\Sigma_0}\|_1 \lesssim \epsilon^2$ and hence $d_H(p_{F_0,\Sigma_0}, p_{F,\Sigma_0}) \lesssim \epsilon$.

For any probability measure F and positive-definite matrix Σ,

$$d_H^2(p_{F,\Sigma}, p_{F,\Sigma_0}) \le d_H^2(\phi(\cdot;\Sigma), \phi(\cdot;\Sigma_0)) = 2\big[1 - \det\big[2(\Sigma+\Sigma_0)^{-1}\Sigma^{1/2}\Sigma_0^{1/2}\big]^{1/2}\big]. \tag{9.12}$$

If $\|\Sigma - \Sigma_0\| \le \epsilon$, then the right side is bounded above by a multiple of ϵ^2.

When combined the preceding paragraphs give that

$$\Big\{(F,\Sigma)\colon d_H(p_{F,\Sigma}, p_{F_0,\Sigma_0}) \lesssim \epsilon\Big\} \supset B{:=}\Big\{(F,\Sigma)\colon \sum_{j=1}^N |F(U_j) - w_j| \le \epsilon^2,\ \|\Sigma - \Sigma_0\| \le \epsilon\Big\}.$$

By Lemma G.13, applied to the Dirichlet vector $(F(U_0), F(U_1), \dots, F(U_N))$ where $U_0 = \mathbb{R}^d \setminus \cup_j U_j$, and the assumption that the prior density for Σ is bounded away from zero, the prior mass of the set on the right is bounded below by $e^{-cN\log(1/\epsilon)}\epsilon^q$, for $q > 0$ depending on the dimension of the support of Σ.

For any F and Σ and $\underline{\sigma}^2$ the minimal eigenvalue of Σ,

$$p_{F,\Sigma}(x) \ge \frac{1}{\underline{\sigma}^d}\int_{\|z\|\le a} e^{-\|x-z\|^2/2\underline{\sigma}^2}\,dF(z) \gtrsim \begin{cases} \frac{1}{\underline{\sigma}^d} e^{-2da^2/\underline{\sigma}^2} F[-a,a]^d, & \|x\|_\infty \le a, \\ \frac{1}{\underline{\sigma}^d} e^{-2d\|x\|^2/\underline{\sigma}^2} F[-a,a]^d, & \|x\|_\infty > a. \end{cases} \tag{9.13}$$

If (F,Σ) is in the set on the right side of the second last display, then the probability $F[-a,a]^d$ and the smallest eigenvalue $\underline{\sigma}^2$ are bounded away from zero. Since p_{F_0,Σ_0} is uniformly bounded with sub-Gaussian tails, it then follows that $P_0(p_{F_0,\Sigma_0}/p_{F,\Sigma})^\delta$ is uniformly bounded, for sufficiently small $\delta > 0$. Consequently $K(p_{F_0,\Sigma_0}; p_{F,\Sigma})$ and $V_2(p_{F_0,\Sigma_0}; p_{F,\Sigma})$ are bounded above by a multiple of $d_H^2(p_{F,\Sigma}, p_{F_0,\Sigma_0})\log_- d_H(p_{F,\Sigma}, p_{F_0,\Sigma_0})$, by Lemma B.2. The lower bound on the prior mass found in the preceding paragraph, $e^{-cN\log(1/\epsilon)}\epsilon q$, translates into a lower bound on the prior mass of the $B_2(p_{F_0,\Sigma_0}, \eta)$-neighborhood, with $\eta \sim \epsilon(\log_- \epsilon)$. Replacing ϵ by $\epsilon/(\log_- \epsilon)$ in the exponent of the lower bound, does not affect its form, and assertion (i) follows upon equating this exponent to $n\epsilon_n^2$.

(ii). Given small $\sigma > 0$ and large a_0, set $a_\sigma = a_0(\log_- \sigma)^{1/\tau}$. By Lemma 9.11 there exists a probability distribution F_σ on $[-a_\sigma, a_\sigma]^d$ such that $d_H(p_0, p_{F_\sigma,\sigma^2 I}) \lesssim \sigma^\beta$. By Lemma 9.12, applied with $\epsilon^2/(\log_- \epsilon)^d$ instead of ϵ, there exists a discrete probability measure $F_\sigma^* = \sum_{j=1}^N w_j \delta_{z_j}$ with at most $N \lesssim (a_\sigma/\sigma)^d (\log_- \epsilon)^d$ support points inside $[-a_\sigma, a_\sigma]^d$ such that $\|p_{F_\sigma,\sigma^2 I} - p_{F_\sigma^*,\sigma^2 I}\|_1 \lesssim \epsilon^2$ and hence $d_H(p_{F_\sigma,\sigma^2 I}, p_{F_\sigma^*,\sigma^2 I}) \lesssim \epsilon$.

The support points of F_σ^* can be chosen on a lattice of mesh width $\sigma\epsilon^2$ without loss of generality. Then there exist disjoint neighborhoods $U_1, \ldots, U_N$ of $z_1, \ldots, z_N$ of diameters of the order $\sigma\epsilon^2$, which can be extended into a partition $\{U_1, \ldots, U_{N'}\}$ of $[-a_\sigma, a_\sigma]^d$ in at most $N' \lesssim N$ sets of diameter at most σ, and next into a partition $\{U_1, \ldots, U_M\}$ of $\mathbb{R}^d$ in $M \lesssim N$ sets, all with the property that $(\sigma\epsilon^2)^d \lesssim \alpha(U_j) \lesssim 1$, for all $j = 1, \ldots, M$. By Lemma 9.13 applied with $U_0 = \cup_{j>N} U_j$ and $w_{N+1} = \cdots = w_M = 0$, if F is a probability measure with $\sum_{j=1}^M |F(U_j) - w_j| \leq \epsilon^2$ then $\|p_{F_\sigma^*,\sigma^2 I} - p_{F,\sigma^2 I}\|_1 \lesssim \epsilon^2$, and hence $d_H(p_{F_\sigma^*,\sigma^2 I}, p_{F,\sigma^2 I}) \lesssim \epsilon$.

By (9.12), for any probability measure F,

$$d_H^2(p_{F,\Sigma}, p_{F,\sigma^2 I}) \leq 2 - 2\prod_{j=1}^d \left(1 - \frac{(\mathrm{eig}_j(\Sigma)^{1/2} - \sigma)^2}{\mathrm{eig}_j(\Sigma) + \sigma^2}\right)^{1/2}.$$

If all eigenvalues of Σ are contained in the interval $[\sigma^2/(1+\sigma^\beta), \sigma^2]$, then the right side is bounded by a multiple of $\sigma^{2\beta}$.

When combined the preceding paragraphs give that

$$\begin{aligned}&\Bigl\{(F, \Sigma)\colon d_H(p_0, p_{F,\Sigma}) \lesssim \sigma^\beta + \epsilon\Bigr\} \qquad (9.14)\\ &\supset B{:=} \Bigl\{(F, \Sigma)\colon \sum_{j=1}^M |F(U_j) - w_j| \leq \epsilon^2,\ \min_{1\leq j\leq M} F(U_j) \geq \epsilon^4,\ 1 \leq \sigma^2 \mathrm{eig}(\Sigma^{-1}) \leq 1 + \sigma^\beta\Bigr\}.\end{aligned}$$

By Lemma G.13 and (9.8), the prior mass of the set B on the right side is bounded below by a multiple of $e^{-c_1 M \log_- \epsilon}\,\sigma^{-2a_4}\sigma^{\beta a_5} e^{-C_3 \sigma^{-\kappa}}$. For M as chosen previously and small σ and ϵ this is bounded below by $e^{-c' E(\sigma,\epsilon)}$, for $E(\sigma, \epsilon) = (\log_- \sigma)^{d/\tau}\sigma^{-d}(\log_- \epsilon)^{d+1} + \sigma^{-\kappa}$.

Because $\{U_1, \ldots, U_{N'}\}$ forms a partition of $[-a_\sigma, a_\sigma]^d$ in sets of diameter smaller than σ, for every $x \in [-a_\sigma, a_\sigma]^d$ there exists a set $U_{j(x)}$ that is contained in the ball of radius σ around x. Hence, for any (F, Σ) in the set B of (9.14) and $\|x\|_\infty \leq a_\sigma$,

$$p_{F,\Sigma}(x) \geq \frac{1}{\sigma^d}\int_{\|x-z\|\leq\sigma} e^{-\|x-z\|^2/\sigma^2}\,dF(z) \geq \frac{e^{-1}}{\sigma^d} F(U_{j(x)}) \geq \frac{e^{-1}\epsilon^4}{\sigma^d}.$$

On the other hand, for $\|x\|_\infty > a_\sigma$, and $(F, \Sigma) \in B$, the second estimate in (9.13), with $a = a_\sigma$, gives that $p_{F,\Sigma}(x) \gtrsim \sigma^{-d} e^{-2d\|x\|^2/\sigma^2} F[-a_\sigma, a_\sigma]^d$, where $F[-a_\sigma, a_\sigma]^d$ is bounded away from zero. This shows that $\log p_{F,\Sigma}(x) \lesssim \log_- \sigma + \|x\|^2/\sigma^2$, whence

$$P_0\Bigl[\Bigl(\log\frac{p_0}{p_{F,\Sigma}}\Bigr)^2 \mathbb{1}\Bigl\{\frac{p_{F,\Sigma}}{p_0} < \frac{e^{-1}\epsilon^d}{\|p_0\|_\infty \sigma^d}\Bigr\}\Bigr] \lesssim \int_{\|x\|>a_\sigma}\Bigl[(\log_- \sigma)^2 + \frac{\|x\|^4}{\sigma^4}\Bigr]\,p_0(x)\,dx.$$

By the definition of a_σ and the tail condition on p_0 the right side is smaller than any given power of σ, if a_0 is sufficiently large. By Lemma B.2 we see that both $K(p_0; p_{F,\Sigma})$ and $V_2(p_0; p_{F,\Sigma})$ are bounded above by a multiple of $d_H^2(p_0; p_{F,\Sigma})(\log_-(\epsilon^4/\sigma^d))^2 + o(\sigma^{2\beta})$, provided that $e^{-1}\epsilon^4/\sigma^d \le 0.44\|p_0\|_\infty$. This implies that the prior mass of the neighborhood $B_2(p_0, \epsilon_n)$ is bounded below by the prior mass in the set B of (9.14) if ϵ and σ are chosen so that $(\sigma^\beta + \epsilon)\log_-(\epsilon^4/\sigma^d) \lesssim \epsilon_n$ and ϵ^4/σ^d is sufficiently small. The prior mass is bounded below by $e^{-Cn\epsilon_n^2}$ for some C if

$$(\sigma^{2\beta}+\epsilon^2)\big[\log_-(\epsilon^4/\sigma^d)\big]^2 \lesssim \epsilon_n^2, \quad \epsilon^4/\sigma^d \ll 1, \quad (\log_-\sigma)^{d/\tau}\sigma^{-d}(\log_-\epsilon)^{d+1}+\sigma^{-\kappa} \le n\epsilon_n^2.$$

We choose $\epsilon^4 \asymp \sigma^d \wedge \sigma^{2\beta}$ in order to satisfy the second requirement and reduce the first requirement to $\sigma^{2\beta}(\log_-\sigma)^2 \lesssim \epsilon_n^2$. If $\kappa \le d$, then we further choose $\sigma^{2\beta+d} \sim n^{-1}(\log n)^{d/\tau+d-1}$ to satisfy the remaining requirements with the rate ϵ_n as given in the theorem. If $\kappa > d$, then we choose $\sigma^{2\beta+\kappa} \sim n^{-1}(\log n)^{-2}$, which leads to a slightly smaller ϵ_n than given in the theorem. □

9.4.3 Entropy Estimate and Controlling Complexity

The following lemma bounds the metric entropy of a suitable sieve of discrete normal mixtures and, the prior probability of its complement. For positive constants ϵ, a, σ and integers M, N, define

$$\begin{aligned}\mathcal{F} &= \Big\{\sum_{j=1}^{\infty} w_j\delta_{z_j}: \sum_{j=N+1}^{\infty} w_j < \epsilon^2, \quad z_1,\ldots,z_N \in [-a,a]^d\Big\},\\ \mathcal{S} &= \Big\{\Sigma: \sigma^2 \le \operatorname{eig}_1(\Sigma) \le \operatorname{eig}_d(\Sigma) < \sigma^2(1+\epsilon^2)^M\Big\}.\end{aligned} \tag{9.15}$$

Lemma 9.15 (Entropy) *For $a \ge \sigma\epsilon$ and $M \ge d$, the sets $\mathcal{F}$ and $\mathcal{S}$ satisfy, for some $A > 0$,*

(i) $N(A\epsilon, \{p_{F,\Sigma}: F \in \mathcal{F}, \Sigma \in \mathcal{S}\}, d_H) \lesssim (5/\epsilon^2)^N (3a/(\sigma\epsilon^2))^{dN}(5/\epsilon^2)^{d^2}(1+\epsilon^2)^{Md^2}M^d$.
(ii) $\mathrm{DP}_\alpha(\mathcal{F}^c) \le (2e|\alpha|\log_-\epsilon/N)^N + N(1-\bar\alpha([-a,a]^d))$.

Proof Let $\mathcal{F}_\epsilon$ be the set of all mixtures $F = \sum_{j=1}^N w_j\delta_{z_j}$, with weight vector $(w_1,\ldots,w_N)$ belonging to a fixed maximal ϵ^2-net in the N-dimensional unit simplex $\mathbb{S}_N$, and support points z_j belonging to a fixed maximal $\sigma\epsilon^2$-net in $[-a,a]^d$ (relative to the $\mathbb{L}_1$- and maximum norms). For $\eta = \epsilon^2/(1+\epsilon^2)^M$, let $\mathcal{S}_\epsilon$ be the set of all matrices $O\Lambda O^\top$ for O belonging to a fixed maximal η-net over the set of all orthogonal matrices (relative to the Frobenius norm), and Λ a diagonal matrix with diagonal entries belonging to the set of M points $\sigma^2(1+\epsilon^2)^m$, for $m = 0, 1, \ldots, M-1$.

The cardinality of the set $\mathcal{F}_\epsilon \times \mathcal{S}_\epsilon$ is bounded above by the upper bound in (i) (see Propositions C.1 and C.2). We shall show that the corresponding set of $p_{F,\Sigma}$ forms an $A\epsilon$-net for d_H over the set of all such densities with $(F,\Sigma) \in \mathcal{F}\times\mathcal{S}$, for a sufficiently large constant A.

Given probability measures $F = \sum_{j=1}^\infty w_j\delta_{z_j}$ and $F' = \sum_{j=1}^N w'_j\delta_{z'_j}$ arguments similar as in the proof of Lemma 9.13 show that $\|p_{F,\Sigma} - p_{F',\Sigma}\|_1 \le \sigma^{-1}\|z-z'\|_\infty + \sum_{j=1}^\infty |w_j - w'_j|$, for $\Sigma \in \mathcal{S}$ (so that its eigenvalues are bounded below by σ^2). For z and w corresponding to some $F \in \mathcal{F}$, the right side can be reduced to $(\sigma^{-1})\sigma\epsilon^2 + 3\epsilon^2$ by choice of z' and w' from the nets. Thus for every $F \in \mathcal{F}$ and $\Sigma \in \mathcal{S}$, there exists $F' \in \mathcal{F}_\epsilon$ such that $d_H(p_{F,\Sigma}, p_{F',\Sigma}) \le 2\epsilon$.

For an orthogonal matrix O and diagonal matrices $\Lambda = \mathrm{diag}(\lambda_j)$ and $\Lambda' = \mathrm{diag}(\lambda'_j)$, the matrices $O\Lambda O^\top$ and $O\Lambda' O^\top$ possess the same eigenvectors. Therefore, by (9.12), for any probability measure F,

$$d_H^2(p_{F,O\Lambda O^\top}, p_{F,O\Lambda' O^\top}) \le 2 - 2\prod_{i=1}^{d}\Big(1 - \frac{(\sqrt{\lambda_i} - \sqrt{\lambda'_i})^2}{\lambda_i + \lambda'_i}\Big)^{1/2}.$$

For $1 \le \lambda'_i/\lambda_i \le 1 + \epsilon^2$ this is of the order ϵ^4. For any diagonal matrix Λ and orthogonal matrices O and O', by the first inequality in (9.12) followed by Lemma B.2(iv) and a calculation of the Kullback-Leibler divergence, since $\det(O\Lambda O^\top) = \det(O'\Lambda(O')^\top)$,

$$\begin{aligned} d_H^2(p_{F,O\Lambda O^\top}, p_{F,O'\Lambda(O')^\top}) &\le K(\phi(\cdot; O\Lambda O^\top), \phi(\cdot; O'\Lambda(O')^\top)) \\ &= \mathrm{tr}([O'\Lambda(O')^\top]^{-1} O\Lambda O^\top - I). \end{aligned}$$

The right side can be written $\mathrm{tr}(Q\Lambda^{-1}Q^\top\Lambda - I)$, for $Q = O^\top O'$, which can be further rewritten as $2\mathrm{tr}(Q - I) + \mathrm{tr}((Q-I)\Lambda^{-1}(Q^\top - I)\Lambda)$. Since $|\mathrm{tr}(A)| \lesssim \|A\|$, this is bounded above by $2\|Q - I\| + \|Q - I\|^2\|\,\|\Lambda\|\,\|\Lambda^{-1}\|$. Since $\|Q - I\| = \|Q - O^\top O\| \le \|O' - O\|$, it follows that there exists O' in the net such that the right side is bounded above by $2\eta + \eta^2\|\Lambda\|\,\|\Lambda^{-1}\| \le 2\eta + \eta^2(1+\epsilon^2)^M$, if $\Lambda \in \mathcal{S}$, which is bounded by a multiple of ϵ^2 by the definition of η.

Combined, the preceding inequalities show that for any $(F, \Sigma) \in \mathcal{F} \times \mathcal{S}$ there exists $(F', \Sigma') \in \mathcal{F}_\epsilon \times \mathcal{S}_\epsilon$ such that $d_H(p_{F,\sigma}, p_{F',\sigma'}) \lesssim \epsilon$.

For the proof of (ii) we use the stick-breaking representation $F = \sum_{j=1}^\infty W_j\delta_{Z_j}$ of a Dirichlet process, given in Theorem 4.12. Then $W_j = V_j\prod_{l=1}^{j-1}(1 - V_l)$ for $V_j \overset{\text{iid}}{\sim} \mathrm{Be}(1, |\alpha|)$, and $R := -\log\sum_{j>N} W_j = -\sum_{j=1}^N \log(1 - V_j)$ is distributed as a $\mathrm{Ga}(N, |\alpha|)$-variable. Hence

$$\begin{aligned} \mathrm{DP}_\alpha(\mathcal{F}^c) &\le \mathrm{P}(R < -\log\epsilon^2) + \sum_{j=1}^N \mathrm{P}(Z_j \notin [-a,a]^d) \\ &\le \frac{(2|\alpha|\log_- \epsilon)^N}{N!} + N(1 - \bar\alpha([-a,a]^d)). \end{aligned}$$

The bound in (ii) follows upon using that $N^N/N! \le e^N$. □

9.4.4 Proof of Theorem 9.9

We apply Theorem 8.9 with $\bar\epsilon_n$ the rate found in Proposition 9.14 on the prior mass, different in the supersmooth and β-smooth cases, and the sieve $\mathcal{P}_{n,1}$ consisting of all densities $p_{F,\Sigma}$ with F in $\mathcal{F}$ and S in $\mathcal{S}$ as given in (9.15), with the choices, for a large constant C,

$$N = \frac{Cn\bar\epsilon_n^2}{\log(n\bar\epsilon_n^2)}, \qquad n\epsilon^2 = C\,N\log n, \qquad M = n, \qquad a^{a_1} = n\epsilon^2, \qquad \sigma^{-2a_2} = n\epsilon^2.$$

This gives that $\epsilon_n^2 := \epsilon^2 = C^2\bar\epsilon_n^2\log n/\log(n\bar\epsilon_n^2)$, which implies that $\epsilon_n \gg \bar\epsilon_n$ in the supersmooth case and $\epsilon_n \ge \sqrt{C}\bar\epsilon_n$ in the β-smooth case.

The prior mass condition (8.4) of Theorem 8.9 is satisfied by choice of $\bar{\epsilon}_n$ and Proposition 9.14. By Lemma 9.15,

$$\log N(A\epsilon_n, \mathcal{P}_{n,1}, d_H) \lesssim N\Big[\log \frac{3a}{\sigma\epsilon_n^2} + \log \frac{5}{\epsilon_n}\Big] + M\log(1+\epsilon_n^2) + \log M.$$

From the second to fifth definitions the right side can be seen to be bounded by a multiple of $n\epsilon_n^2$, as required by (8.5) of Theorem 8.9. By (ii) of Lemma 9.15 and the conditions (9.5)–(9.7), the prior probability of $\mathcal{P}_{n,1}^c$ is bounded by a multiple of

$$\Big(\frac{2e|\alpha|\log_- \epsilon_n}{N}\Big)^N + Ne^{-C_1 a^{a_1}} + e^{-C_2\sigma^{-2a_2}} + \sigma^{-2a_3}(1+\epsilon_n^2)^{-a_3 M}.$$

It is immediate from the definitions that the second and third terms are bounded above by $e^{-C_i n\epsilon_n^2} \leq e^{-C_i C^2 n\bar{\epsilon}_n^2}$. In the fourth term $(1+\epsilon_n^2)^{-M} \leq e^{-n\epsilon_n^2/2}$ is similarly small, and dominates σ^{-2a_3}. Finally in the first term the quotient $U_n = \log_- \epsilon_n/N = C\log_- \epsilon_n \log n/(n\epsilon_n^2)$ is bounded above by $C(\log n)^2/n\epsilon_n^2$. In the supersmooth case $n\epsilon_n^2 \geq n\bar{\epsilon}_n^2 = (\log n)^{d+1+1/r_0}$ and hence $U_n \leq C(\log n)^{-\gamma}$, for some $\gamma > 0$, whence $U_n^N \leq e^{-N[\gamma \log\log n - \log C]} \leq e^{-Cn\bar{\epsilon}_n^2/2}$. In the ordinary smooth case $(\log n)^2/n\epsilon_n^2$ is bounded above by a power of $1/n$ and hence $U_n^N \leq e^{-N\delta\log n} \leq e^{-C\delta_1 n\bar{\epsilon}_n^2}$. In both cases it follows that (8.6) of Theorem 8.9 is satisfied as well.

9.4.5 Wishart Prior

Conditions (9.6)–(9.8) hold for the example of a conjugate inverse-Wishart prior.

Lemma 9.16 *If* $\Sigma^{-1} \sim \mathrm{Wis}(\nu, \Psi)$, *with positive-definite scale matrix* Ψ, *then* (9.6)–(9.8) *hold with* $\kappa = 2$.

Proof First consider the case Ψ is the identity matrix. Then $\mathrm{tr}(\Sigma^{-1}) \sim \chi^2_{\nu d}$, and hence $\mathrm{P}(\mathrm{tr}(\Sigma^{-1}) > s) \leq e^{-s/4}$, for all sufficiently large s, whence (9.6) holds, as $\mathrm{eig}_d(\Sigma^{-1}) \leq \mathrm{tr}(\Sigma^{-1})$. The joint probability density of $\mathrm{eig}_1(\Sigma^{-1}), \ldots, \mathrm{eig}_d(\Sigma^{-1})$ satisfies (cf. Muirhead 1982, page 106)

$$f(s_1, \ldots, s_d) \propto e^{-\sum_j s_j/2} \prod_{j=1}^{d} s_j^{(\nu+1-d)/2} \prod_{1\leq j<k\leq d} (s_k - s_j), \qquad 0 < s_1 < \cdots < s_d. \quad (9.16)$$

Since $\prod_{j<k}(s_k - s_j) \leq \prod_{j<k} s_k = \prod_{k=2}^{d} s_k^{k-1}$, the marginal density of $\mathrm{eig}_1(\Sigma^{-1})$ is bounded by a multiple of

$$s_1^{(\nu+1-d)/2} e^{-s_1/2} \prod_{k=2}^{d} \int_0^\infty s_k^{(\nu+1-d)/2+k-1} e^{-s_k/2}\, ds_k \propto s_1^{(\nu+1-d)/2} e^{-s_1/2}, \qquad s_1 > 0.$$

This leads to relation (9.7). For (9.8) it suffices to lower bound the probability of the event $\cap_{j=1}^{d}\{\mathrm{eig}_j(\Sigma^{-1}) \in I_j\}$, where $I_j = (s_j(1+(j-1/2)t/d), s_j(1+jt/d))$, for $j = 1, \ldots, d$. If

$s_j \in I_j$ and $s_k \in I_k$, for $j < k$, then $s_k - s_j > s_k(1+(k-1/2)t/d) - s_j(1+jt/d) \geq s_1 t/(2d)$, and hence, for

$$\begin{aligned}
\mathrm{P}(s_j < \mathrm{eig}_j(\Sigma^{-1}) < s_j(1+t), \quad j = 1, \ldots, d) &\geq \int_{I_1} \cdots \int_{I_1} f(s_1, \ldots, s_d)\, ds_1 \cdots ds_d \\
&\geq c_{d,\nu} e^{-d s_d} s_1^{d\nu/2} (t/(2d))^{d(d-1)/2} \int_{I_d} \cdots \int_{I_1} dx_1 \cdots dx_d \\
&= c_{d,\nu} (2d)^{-d(d+1)/2} e^{-d s_d} s_1^{d(\nu+2)/2} t^{d(d+1)/2},
\end{aligned}$$

where $c_{d,\nu}$ is the norming constant in (9.16). This gives (9.8) for positive constants a_4, a_5, b_4, C_3.

If $\Psi \neq I$, then the random matrix $\Omega = \Psi^{1/2} \Sigma \Psi^{1/2}$ follows the inverse-Wishart distribution with identity scale matrix, and hence satisfies (9.6) and (9.7) by the special case. Since $\mathrm{eig}_d(\Sigma^{-1}) = \|\Sigma^{-1}\| = \|\Psi^{1/2} \Omega^{-1} \Psi^{1/2}\| \leq \|\Psi\| \, \|\Omega^{-1}\|$, it follows that (9.6) holds for Σ as well, with a different constant C_2. The smallest eigenvalue satisfies $\mathrm{eig}_1(\Sigma^{-1}) = \|\Sigma\|^{-1} \geq \|\Psi^{-1}\|^{-1} \|\Omega\|^{-1} = \|\Psi^{-1}\|^{-1} \mathrm{eig}_1(\Omega^{-1})$. Hence condition (9.7) follows again from the special case, with a different choice of b_3. Finally, to check (9.8), it suffices to show that the joint density of the eigenvalues admits a lower bound of the form (9.16). The joint density is proportional to (cf. Muirhead (1982), page 106) by

$$\prod_{j=1}^d s_j^{(\nu+1-d)/2} \prod_{j<k} (s_k - s_j) \int_{\mathcal{O}(d)} e^{-\nu \mathrm{tr}(\Psi^{-1} H \Delta(s) H^\top /2)}\, dH,$$

where $\Delta(s) = \mathrm{diag}(s_1, \ldots, s_d)$ and $\mathcal{O}(d)$ is the group of $d \times d$ orthogonal matrices. Here $\mathrm{tr}(\Psi^{-1/2} H \Delta(s) H^\top \Psi^{-1/2})$ increases if we replace $\Delta(s)$ by $s_d I$, whence this expression is bounded above by $s_d \mathrm{tr}(\Psi^{-1/2} H H^\top \Psi^{-1/2}) = s_d \mathrm{tr}(\Psi^{-1})$, in view of the orthogonality of H. □

9.5 Non-i.i.d. Models

In this section we present a number of applications of Theorem 8.19 in the non-i.i.d. setup. We consider various combinations of models and prior distributions.

9.5.1 Finite Sieves

In Theorem 8.24 we constructed a sequence of prior distributions using bounds for bracketing numbers such that the posterior converges at the optimal rate. We illustrate the construction for two concrete models.

Example 9.17 (Nonparametric Poisson regression) Suppose we observe $X_1, \ldots, X_n$, where $X_i \overset{\mathrm{ind}}{\sim} \mathrm{Poi}(\psi(z_i))$ for an an unknown increasing function $\psi: \mathbb{R} \to (0, \infty)$ and real, deterministic covariates $z_1, z_2, \ldots$. We assume that $L \leq \psi \leq U$ for some constants $0 < L < U < \infty$.

If $l \leq \psi \leq u$, then $e^{-\psi(z)} \psi(z)^x / x! \leq e^{-l(z)} u(z)^x / x!$, for all z, x. Thus for $q_{l,u,z}(x)$ defined as the right side of this inequality, the vector $(q_{l,u,z_1}, \ldots, q_{l,u,z_n})$ is a componentwise upper bracket for the density of an observation with parameter ψ. For any constants $L < \lambda_1, \lambda_2, \mu_1, \mu_2 < U$,

$$\begin{aligned}\sum_{x=0}^{\infty}\Big[\Big(e^{-\lambda_1}\frac{\mu_1^x}{x!}\Big)^{1/2} - \Big(e^{-\lambda_2}\frac{\mu_2^x}{x!}\Big)^{1/2}\Big]^2 &= \Big[e^{-(\lambda_1+\mu_1)/2} - e^{-(\lambda_2+\mu_2)/2}\Big]^2 \\ &\quad + 2e^{-(\lambda_1+\lambda_2)/2}\Big[e^{(\mu_1+\mu_2)/2} - e^{\sqrt{\mu_1\mu_2}}\Big] \\ &\leq \Big(\frac{1}{2} + \frac{1}{4L}\Big)e^{U-L}\left(|\lambda_1-\lambda_2|^2 + |\mu_1-\mu_2|^2\right).\end{aligned}$$

Therefore, for given pairs of functions $l_1 \leq u_1$ and $l_2 \leq u_2$ taking their values in the interval $[L, U]$,

$$\frac{1}{n}\sum_{i=1}^{n} d_H^2(q_{l_1,u_1,z_i}, q_{l_2,u_2,z_i}) \lesssim \int (|l_1 - l_2|^2 + |u_1 - u_2|^2)\, d\mathbb{P}_n^z,$$

for $\mathbb{P}_n^z = n^{-1}\sum_{i=1}^n \delta_{z_i}$ the empirical distribution of the covariates. Hence an ϵ-bracketing for the class of functions ψ in the $\mathbb{L}_2(\mathbb{P}_n^z)$-metric yields a componentwise Hellinger upper bracketing for the model of size a multiple of ϵ. The ϵ-bracketing entropy number in $\mathbb{L}_2(\mathbb{P}_n^z)$ of the class of all monotone, uniformly bounded functions is bounded by a multiple of ϵ^{-1}, in view of Proposition C.8. Equating this to $n\epsilon^2$, we obtain the rate $\epsilon_n = n^{-1/3}$ relative to the root mean square Hellinger distance, for the posterior distribution based on the discrete prior on a minimal set of renormalized upper brackets.

The normalization of the upper bracket $q_{l,u,z}$ is the density of the Poisson distribution with mean $u(z)$. Hence in this example the bracketing prior charges the original model, which makes its interpretation, and that of the posterior contraction rate, more transparent. As the parameter space is fixed, proceeding as in Theorem 8.15, a prior not depending on n can be constructed such that the posterior converges at the same $n^{-1/3}$ rate.

Example 9.18 (Semiparametric Poisson regression) Suppose we observe $X_1, \ldots, X_n$ for $X_i \overset{\text{ind}}{\sim} \text{Poi}(\psi(\beta^\top z_i))$, where $z_1, z_2, \ldots$ are deterministic covariates in $\mathbb{R}^d$, β is an unknown vector of the same dimension and $\psi\colon \mathbb{R} \to (0, \infty)$ is an unknown function. We assume that the function ψ is smooth, but it need not be monotone. The parameters ψ and β are not jointly identifiable, but this will not concern us as we measure distances based on the distributions of the observations rather than the parameters. Assume that the possible values of the regressors $z_1, z_2, \ldots$ and the regression coefficient β lie in a known compact set. Let ψ lie in the Hölder class $\mathfrak{C}^\alpha(\mathbb{K})$, defined in Definition C.4, with norm $\|\psi\|_{\mathfrak{C}^\alpha}$ bounded by M, for given $\alpha \geq 1$ and $M > 0$, and $\mathbb{K}$ a compact interval containing the range of all possible linear combinations $\beta^\top z_i$.

We consider the discrete prior on a minimal set of renormalized componentwise upper brackets. To construct these we bracket the set of functions $z \mapsto \psi(\beta^\top z)$ and can next use the Poisson brackets as in the preceding example. For B a bound for the norm of the regressors find an $\epsilon/(3MB)$ net $\beta_1, \beta_2, \ldots, \beta_{N_1}$ for β and an $\epsilon/3$-bracketing $(l_1, u_1), \ldots, (l_{N_2}, u_{N_2})$ for ψ and consider the collection $(l_k(\beta_j^\top z) - \epsilon/3, u_k(\beta_j^\top z) + \epsilon/3)$, $j = 1, 2, \ldots, N_1, k = 1, 2, \ldots, N_2$. It is straightforward to check that this is an ϵ-bracketing for the functions $z \mapsto \psi(\beta^\top z)$. As we can choose $\log N_1 \lesssim \log_- \epsilon$ and $\log N_2 \lesssim \epsilon^{-1/\alpha}$, the ϵ-bracketing entropy numbers for the desired class is bounded by a multiple of $\epsilon^{-1/\alpha}$. Therefore the posterior contracts at the rate $n^{-\alpha/(1+2\alpha)}$ with respect to the root mean square Hellinger distance.

9.5.2 Whittle Estimation of a Spectral Density

Consider estimating the spectral density f of a stationary Gaussian time series $(X_t: t \in \mathbb{Z})$ using the Whittle likelihood, as considered in Section 7.3.3. Thus we act as if the periodogram values $U_l = I_n(\omega_l)$, for $l = 1, \ldots, \nu$, are independent exponential variables with means $f(\omega_l)$, for $\omega_l = 2\pi l/n$ the natural frequencies, and form a posterior distribution based on the corresponding (pseudo) likelihood. By Theorem L.8, the sequence of actual distributions of the periodogram vectors $(U_1, \ldots, U_\nu)$ and their exponential approximation are contiguous. Thus a rate of contraction for this (pseudo) posterior distribution is also valid relative to the original distribution of the time series.

Because the observations $U_1, \ldots, U_n$ are independent exponential variables, rates of contraction can be obtained with the help of Theorem 8.23. The Hellinger distance and Kullback-Leibler discrepancies were computed in Section 7.3.3. If the spectral densities are bounded away from zero and infinity, then the root average square Hellinger distance d_n of the joint densities can be bounded by the distance $\bar{d}_n$ on spectral densities defined by $\bar{d}_n^2(f_1, f_2) = \nu^{-1} \sum_{l=1}^{\nu} (f_1(\omega_l) - f_2(\omega_l))^2$. Indeed if $m \leq f_1, f_2 \leq M$, then for $U_l \overset{\text{ind}}{\sim} P_{f,l}$, $l = 1, \ldots, \nu$,

$$\frac{1}{4M^2}\bar{d}_n^2(f_1, f_2) \leq d_{n,H}^2(p_{f_1}, p_{f_2}) \leq \frac{1}{4m^2}\bar{d}_n^2(f_1, f_2) \leq \frac{1}{4m^2}\|f_1 - f_2\|_\infty^2,$$

$$\frac{1}{\nu}\sum_{l=1}^{\nu} K(P_{f_0,l}; P_{f,l}) \vee V_{2,0}(P_{f_0,l}; P_{f,l}) \lesssim \bar{d}_n^2(f_0, f) \lesssim \|f - f_0\|_\infty^2.$$

This shows that the entropy and prior mass conditions may be verified for the metric $\bar{d}_n$, or even the uniform distance. If, furthermore, the spectral densities are Lipschitz, then we may also use the $\mathbb{L}_2$-distance $\|\cdot\|_2$ relative to the Lebesgue measure on $[0, \pi]$. This is because $\bar{d}_n(f_1, f_2)$ is equal to $\sqrt{2\pi\nu/n}$ times the $\mathbb{L}_2$-distance $\|f_{1,n} - f_{2,n}\|_2$ between the discretizations $f_{i,n} = \sum_{l=1}^{\nu} \mathbb{1}_{(\omega_{l-1},\omega_l]} f_i(\omega_l)$ of the functions f_i, while $\|f - f_n\|_2 \lesssim \|f\|_{\text{Lip}}/n$. If all spectral densities in the sieve are Lipschitz with constant L_n and $\epsilon_n \gg L_n/n$, then we may replace d_n also by the $\mathbb{L}_2$-norm $\|\cdot\|_2$ when verifying (8.27). Similarly we may estimate the prior mass of an $\mathbb{L}_2$-ball of Lipschitz functions.

Example 9.19 (Bernstein polynomial prior for spectral density) Consider the prior on $f = \tau\bar{f}$ induced by equipping the scalar τ with a nonsingular prior density on $(0, \infty)$ and the probability density $\bar{f}$ with the Dirichlet-Bernstein prior of Example 5.10 (rescaled to the domain $[0, \pi]$), and then restricting the prior of $\tau\bar{f}$ to the set $\mathcal{F} = \{f: m < f < M\}$, for given $0 < m < M < \infty$. Let α be the base measure of the Dirichlet process, and let the order of the Bernstein polynomial possess prior $k \sim \rho$ such that $e^{-\beta_1 k \log k} \lesssim \rho(k) \lesssim e^{-\beta_2 k}$. Let Π stand for the resulting prior. Assume that $f_0 \in \mathfrak{C}^\alpha[0, \pi]$, for some $\alpha \in (0, 2]$.

If $f_0 \in \mathcal{F}$, then restricting the prior to $\mathcal{F}$ can only increase the prior probability of the set $\{f: \|f - f_0\|_\infty < \epsilon\}$, whence the restriction can be disregarded for lower bounding the prior concentration. For any k such that $\|f_0 - b(\cdot;\, f_0, k)\|_\infty < \epsilon/2$, a lower bound for the prior mass of the set $\{f\colon \|f - f_0\|_\infty < \epsilon\}$ is given by $\rho(k)\mathrm{DP}_\alpha(\|b(\cdot;\, f_0, k) - b(\cdot;\, F, k)\|_\infty < \epsilon/2)$. As in Section 9.3, this can be bounded below by $\rho(k)e^{-Ck\log_- \epsilon}$ for some constant $C > 0$. If $f_0 \in \mathfrak{C}^\alpha[0, \pi]$, then the uniform distance $\|f_0 - b(\cdot;\, f_0, k)\|_\infty$ is bounded by a multiple of $k^{-\alpha/2}$, by Lemma E.3. Optimizing over all k such that $\|f_0 - b(\cdot;\, f_0, k)\|_\infty < \epsilon/2$, which forces $k \gtrsim \epsilon^{-2/\alpha}$, gives a lower bound of the form $e^{-c\epsilon^{-2/\alpha}\log_- \epsilon}$, for some $c > 0$. Therefore (8.26) is satisfied for ϵ_n of the order $(n/\log n)^{-\alpha/(2+2\alpha)}$.

Let $\mathcal{F}_n \subset \mathcal{F}$ be the set of consisting of only Bernstein polynomials of order k_n or less. By similar calculations as in Section 9.3, based on the entropy of the unit simplex of dimension k_n, it can be shown that $\log D(\epsilon, \mathcal{F}_n, \|\cdot\|_\infty) \lesssim k_n \log k_n + k_n \log_- \epsilon$. Thus the entropy condition (8.27) is satisfied for $\epsilon_n \gtrsim \sqrt{k_n \log k_n}/\sqrt{n}$.

Finally the prior of the complement of the sieve satisfies $\Pi(\mathcal{F}_n^c) = \rho(k > k_n) \lesssim e^{-\beta_2 k_n}$. This is $o(e^{-Cn\epsilon_n^2})$ for an (arbitrarily) large constant C if k_n is a sufficiently large multiple of $n^{1/(1+\alpha)}(\log n)^{\alpha/(1+\alpha)}$. The posterior probability of $\mathcal{F}_n^c$ then goes to 0 by Lemma 8.20. For this choice of k_n the contraction rate on $\mathcal{F}_n$, and hence on $\mathcal{F}$, is $n^{-\alpha/(2+2\alpha)}(\log n)^{(1+2\alpha)/(2+2\alpha)}$.

This is the contraction rate relative to root average square Hellinger distance, or equivalently the discrete $\mathbb{L}_2$-distance $\bar{d}_n$. Because the functions in $\mathcal{F}_n$ are Lipschitz with Lipschitz constant at most k_n^2, this distance differs at most of the order k_n^2/n from the ordinary $\mathbb{L}_2$-distance, uniformly in $\mathcal{F}_n$. It follows that the discrete $\mathbb{L}_2$-contraction rate carries over into a rate $k_n^2/n \vee n^{-\alpha/(2+2\alpha)}(\log n)^{(1+2\alpha)/(2+2\alpha)}$ for the ordinary $\mathbb{L}_2$-distance. For k_n as chosen previously, the term k_n^2/n dominates for all $\alpha \leq 2$, and tends to zero only for $\alpha > 1$, thus giving a slower rate for the ordinary $\mathbb{L}_2$-distance in these cases. (For $\alpha = 2$ the difference is only in the logarithmic term.)

For $f_0 \in \mathfrak{C}^\alpha[0, \pi]$, for $\alpha \in (0, 1]$, a coarsened Bernstein polynomial prior as used in Section 9.3 will lead to the posterior contraction rate $n^{-\alpha/(1+2\alpha)}(\log n)^{(1+4\alpha)/(2+4\alpha)}$, which is nearly minimax. Other priors such as those based on finite-dimensional approximation lead to a nearly optimal rate for any α; see Section 10.4.6.

9.5.3 Nonlinear Autoregression

Consider a nonlinear autoregressive stationary time series $\{X_t\colon t \in \mathbb{Z}\}$ given by

$$X_i = f(X_{i-1}) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{9.17}$$

where f is an unknown function and $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} \mathrm{Nor}(0, 1)$. Then X_n is a Markov chain with transition density $p_f(y|x) = \phi(y - f(x))$, where ϕ is the standard normal density. Assume that $f \in \mathcal{F}$, a class of functions such that $|f(x)| \leq M$ and $|f(x) - f(y)| \leq L|x - y|$ for all x, y and $f \in \mathcal{F}$.

Set $r(y) = (\phi(y - M) + \phi(y + M))/2$. Then $r(y) \lesssim p_f(y|x) \lesssim r(y)$ for all $x, y \in \mathbb{R}$ and $f \in \mathcal{F}$. Further, $\sup\{\int |p(y|x_1) - p(y|x_2)|\, dy\colon x_1, x_2 \in \mathbb{R}\} < 2$. Hence the chain is α-mixing with exponentially decaying mixing coefficients by the discussion following Theorem 8.29, has a unique stationary distribution Q_f whose density q_f satisfies $r \lesssim q_f \lesssim r$. Let $\|f\|_s = (\int |f|^s dr)^{1/s}$.

Because $d_H^2(\mathrm{Nor}(\mu_1,1),\mathrm{Nor}(\mu_2,1)) = 2[1-e^{-|\mu_1-\mu_2|^2/8}]$ (see Problem B.1), it easily follows that for $f_1, f_2 \in \mathcal{F}$, d defined in (8.35) and $d\nu = r\,d\lambda$ that $\|f_1-f_2\|_2 \lesssim d(f_1,f_2) \lesssim \|f_1-f_2\|_2$. Thus we may verify (8.37) relative to the $\mathbb{L}_2(r)$-metric. It can also be computed that

$$P_{f_0}\log\frac{p_{f_0}(X_2|X_1)}{p_f(X_2|X_1)} = \frac{1}{2}\int (f_0-f)^2 q_{f_0}\,d\lambda \lesssim \|f-f_0\|_2^2.$$

$$P_{f_0}\Big|\log\frac{p_{f_0}(X_2|X_1)}{p_f(X_2|X_1)}\Big|^s \lesssim \int |f_0-f|^s q_{f_0}\,d\lambda \lesssim \|f-f_0\|_s^s.$$

Thus $B^*(f_0,\epsilon;s) \supset \{f: \|f-f_0\|_s \le c\epsilon\}$ for some constant $c>0$, where $B^*(f_0,\epsilon;s)$ is as in Theorem 8.29. Thus it suffices to verify (8.36) with $s>2$.

Example 9.20 (Random histograms) As a prior on the regression functions f, consider a random histogram as follows. For a given number $K \in \mathbb{N}$, partition a given compact interval in $\mathbb{R}$ into K subintervals $I_1,\ldots,I_K$ and let $I_0 = (\cup_k I_k)^c$. Let the prior Π_n on f be induced by the map $\alpha \mapsto f_\alpha$ given by $f_\alpha = \sum_{k=1}^K \alpha_k \mathbb{1}_{I_k}$, where $\alpha = (\alpha_1,\ldots,\alpha_K) \in \mathbb{R}^K$, and a priori $\alpha_j \overset{\text{iid}}{\sim} \mathrm{Unif}[-M,M]$, $j=1,\ldots,K$, and $K=K_n$ is to be chosen later. Let $r(I_k) = \int_{I_k} r\,d\lambda$.

The support of Π_n consists of all functions with values in $[-M,M]$ that are piecewise constant on each interval I_k for $k=1,\ldots,K$, and vanish on I_0. For any pair f_α and f_β of such functions we have, for any $s \in [2,\infty]$, $\|f_\alpha - f_\beta\|_s = \|\alpha-\beta\|_s$, where $\|\alpha\|_s$ is the r-weighted ℓ_s-norm of α given by $\|\alpha\|_s^s = \sum_k |\alpha_k|^s r(I_k)$. The dual use of $\|\cdot\|_s$ should not lead to any confusion, as it will be clear from the context whether $\|\cdot\|_s$ is a norm on functions or that on vectors. The $\mathbb{L}_2(r)$-projection of f_0 onto this support is the function f_{α_0} for $\alpha_{0,k} = \int_{I_k} f_0 r\,d\lambda/r(I_k)$, whence, by Pythagoras' theorem, $\|f_\alpha - f_0\|_2^2 = \|f_\alpha - f_{\alpha_0}\|_2^2 + \|f_{\alpha_0}-f_0\|_2^2$, for any $\alpha \in [-M,M]^K$. In particular, $\|f_\alpha - f_0\|_2 \ge c\|\alpha-\alpha_0\|_2$ for some constant c, and hence, with $\mathcal{F}_n$ denoting the support of Π_n,

$$\begin{aligned} &N(\epsilon, \{f\in\mathcal{F}_n: \|f-f_0\|_2 \le 16\epsilon\}, \|\cdot\|_2)\\ &\qquad \le N(\epsilon, \{\alpha\in\mathbb{R}^K: \|\alpha-\alpha_0\|_2 \le 16c\epsilon\}, \|\cdot\|_2) \le (80c)^K \end{aligned}$$

by Proposition C.1. Thus (8.37) holds if $n\epsilon_n^2 \gtrsim K$.

To verify (8.36), note that, for $\lambda = (\lambda(I_1),\ldots,\lambda(I_K))$

$$\|f_{\alpha_0} - f_0\|_s^s = \int_{I_0} |f_0|^s r\,d\lambda + \sum_{k=1}^K \int_{I_k} |\alpha_{0,k} - f_0|^s r\,d\lambda \le M^s r(I_0) + L^s \|\lambda\|_s^s.$$

Hence as $f_0 \in \mathcal{F}$, for every $\alpha \in [-M,M]^K$,

$$\|f_\alpha - f_0\|_s \lesssim \|\alpha-\alpha_0\|_s + r(I_0)^{1/s} + \|\lambda\|_s \le \|\alpha-\alpha_0\|_\infty + r(I_0)^{1/s} + \|\lambda\|_s,$$

where $\|\cdot\|_\infty$ is the ordinary maximum norm on $\mathbb{R}^K$. For $r(I_0)^{1/s} + \|\lambda\|_s \le \epsilon/2$, we have that $\{f: \|f-f_0\|_s \le \epsilon\} \supset \{f_\alpha: \|\alpha-\alpha_0\|_\infty \le \epsilon/2\}$. Using $\|\alpha-\alpha_0\|_2 \le c\|f_\alpha - f_0\|_2$, for any $\epsilon>0$ such that $r(I_0)^{1/s} + \|\lambda\|_s \le \epsilon/2$,

$$\frac{\Pi_n(f: \|f-f_0\|_2 \le j\epsilon)}{\Pi_n(f: \|f-f_0\|_s \le \epsilon)} \le \frac{\Pi_n(\alpha: \|\alpha-\alpha_0\|_2 \le j\epsilon)}{\Pi_n(\alpha: \|\alpha-\alpha_0\|_\infty \le \epsilon c/2)}.$$

We show that the right-hand side is bounded by $e^{Cn\epsilon^2/8}$ for some C.

For $\{I_1, \ldots, I_K\}$ a regular partition of an interval $[-A, A]$, we have that $\|\lambda\|_s = 2A/K$ and, since $r(I_k) \geq \lambda(I_k) \inf\{r(x): x \in I_k\}$, $k \geq 1$, the norm $\|\cdot\|_2$ is bounded below by the Euclidean norm multiplied by $\sqrt{2A\phi(A)/K} \gtrsim \sqrt{\phi(A)/K}$. In this case, the preceding display is bounded above by

$$\frac{(Cj\epsilon\sqrt{K/\phi(A)}/(2M))^K \operatorname{vol}\{x \in \mathbb{R}^K: \|x\| \leq 1\}}{(\epsilon c/(4M))^K} \sim \Big(\frac{2Cj\sqrt{2\pi e}}{c\sqrt{\phi(A)}}\Big)^K \frac{1}{\sqrt{\pi K}},$$

by Lemma K.13. The probability $r(I_0)$ is bounded above by $1 - 2\Phi(A) \lesssim \phi(A)$. Hence (8.36) will hold if $K \log_- \phi(A) \lesssim n\epsilon_n^2$, $\phi(A) \lesssim \epsilon_n^s$, and $A/K \lesssim \epsilon_n$. With $K \asymp \epsilon_n^{-1}(\log_- \epsilon_n)^{1/2}$ and $A \asymp (\log_- \epsilon_n)^{1/2}$, all conditions are met for ϵ_n a sufficiently large multiple of $n^{-1/3}(\log n)^{1/2}$. This is only marginally weaker than the minimax rate which is $n^{-1/3}$ for this problem provided that the autoregressive functions are assumed to be only Lipschitz continuous.

The logarithmic factor in the contraction rate appears to be a consequence of the fact that the regression functions are defined on the full real line. The present prior is a special case of a spline-based prior to be discussed in Subsection 9.5.5. If f has smoothness beyond Lipschitz continuity, then the use of higher order splines should yield a faster contraction rate.

9.5.4 White Noise with Conjugate Priors

The observation in the white noise model is an infinite-dimensional vector $X^{(n)} = (X_{n,1}, X_{n,2}, \ldots)$, where $X_{n,i} \overset{\text{ind}}{\sim} \mathrm{Nor}(\theta_i, n^{-1})$, and the prior Π_n is given by $\theta_i \overset{\text{ind}}{\sim} \mathrm{Nor}(0, \sigma_{i,k}^2)$, $i = 1, \ldots, k$, and $\theta_{k+1} = \theta_{k+2} = \cdots = 0$, where $k = k_n = \lfloor n^{1/(2\alpha+1)} \rfloor$ for some $\alpha > 0$. The posterior distribution and rates of contraction can be computed simply and explicitly, as shown Example 8.6, but as illustration we derive the rate in this section from the general theorem in Section 8.3.4.

Assume that

$$\min\{\sigma_{i,k}^2 i^{2\alpha}: 1 \leq i \leq k\} \sim k^{-1}. \tag{9.18}$$

This is valid if $\sigma_{i,k}^2 = k^{-1}$, for $i = 1, \ldots, k$, but also if $\sigma_{i,k}^2 = i^{-(2\alpha+1)}$, for $i = 1, \ldots, k$. Even though quite different both priors give the optimal contraction rate if the true parameter θ_0 is "α-regular."

Theorem 9.21 *If* (9.18) *holds, then the posterior contracts at the rate $\epsilon_n = n^{-\alpha/(2\alpha+1)}$ for any θ_0 such that $\sum_{i=1}^\infty \theta_{0,i}^2 i^{2\alpha} < \infty$.*

Proof The support Θ_n of the prior is the set of all $\theta \in \ell_2$ with $\theta_i = 0$ for $i > k$, and can be identified with $\mathbb{R}^k$. Moreover, the ℓ_2-norm $\|\cdot\|$ on the support can be identified with the Euclidean norm on $\mathbb{R}^k$, temporarily to be denoted by $\|\cdot\|_k$ to make the dependence on k explicit. Let $B_k(x, \epsilon)$ denote the k-dimensional Euclidean balls of radius ϵ and $x \in \mathbb{R}^k$. For any true parameter $\theta_0 \in \ell_2$ we have $\|\theta - \theta_0\| \geq \|\operatorname{Proj}\theta - \operatorname{Proj}\theta_0\|_k$, where Proj is the projection operator on Θ_n and hence

$$N(\epsilon/8, \{\theta \in \Theta_n : \|\theta - \theta_0\| \le \epsilon\}, \|\cdot\|) \le N(\epsilon/8, B_k(\operatorname{Proj}\theta_0, \epsilon), \|\cdot\|_k) \le (40)^k$$

in view of Proposition C.1. It follows that (8.41) is satisfied if $n\epsilon_n^2 \gtrsim k$, equivalently, if $\epsilon_n \gtrsim n^{-\alpha/(2\alpha+1)}$.

By Pythagoras's theorem we have that $\|\theta - \theta_0\|^2 = \|\operatorname{Proj}\theta - \operatorname{Proj}\theta_0\|^2 + \sum_{i>k}\theta_{0,i}^2$ for any $\theta \in \operatorname{supp}(\Pi_n)$. Hence for $\sum_{i>k}\theta_{0,i}^2 \le \epsilon_n^2/2$ we have that

$$\Pi_n(\theta \in \Theta_n : \|\theta - \theta_0\| \le \epsilon_n) \ge \Pi_n(\theta \in \mathbb{R}^k : \|\theta - \operatorname{Proj}\theta_0\|_k \le \epsilon_n/2).$$

By the definition of the prior, the right-hand side involves a quadratic form in Gaussian variables. Set $\Sigma = \operatorname{diag}(\sigma_{i,k}^2 : i = 1, \ldots, k)$ and let Φ_k refer to the probability content of a k-dimensional normal distribution. The quotient on the left-hand side of (8.40) can be bounded as

$$\frac{\Pi_n(\theta \in \Theta_n : \|\theta - \theta_0\| \le j\epsilon_n)}{\Pi_n(\theta \in \Theta_n : \|\theta - \theta_0\| \le \epsilon_n)} \le \frac{\Phi_k(-\operatorname{Proj}\theta_0, \Sigma)(B(0, j\epsilon_n))}{\Phi_k(-\operatorname{Proj}\theta_0, \Sigma)(B(0, \epsilon_n/2))}.$$

The probability in the numerator increases if we center the normal distribution at 0 rather than at $-\operatorname{Proj}\theta_0$, by Anderson's lemma (Lemma K.12). Furthermore, for any $\mu \in \mathbb{R}^k$, the normal densities satisfy

$$\frac{\phi_{\mu,\Sigma}(\theta)}{\phi_{0,\Sigma/2}(\theta)} = \frac{e^{-\sum_{i=1}^k (\theta_i - \mu_i)^2/(2\sigma_{i,k}^2)}}{\sqrt{2}^k e^{-\sum_{i=1}^k \theta_i^2/\sigma_{i,k}^2}} \ge 2^{-k/2} \exp\Big(-\sum_{i=1}^k \frac{\mu_i^2}{\sigma_{i,k}^2}\Big).$$

Therefore, we may recenter the denominator at 0 at the cost of adding the factor on the right, with $\mu = \theta_0$, and dividing the covariance matrix by 2. We obtain that the left-hand side of (8.40) is bounded above by

$$\begin{aligned} &2^{k/2} e^{\sum_{i=1}^k \theta_{0,i}^2/\sigma_{i,k}^2} \frac{\Phi_k(0, \Sigma)(B(0, j\epsilon_n))}{\Phi_k(0, \Sigma/2)(B(0, \epsilon_n/2))} \\ &\qquad \le 2^{k/2} \exp\Big(\sum_{i=1}^k \frac{\theta_{0,i}^2}{\sigma_{i,k}^2}\Big)\Big(\frac{\bar\sigma_k}{\underline{\sigma}_k}\Big)^k \frac{\Phi_k(0, \bar\sigma_k^2 I)(B(0, j\epsilon_n))}{\Phi_k(0, \underline{\sigma}_k^2 I/2)(B(0, \epsilon_n/2))}, \end{aligned}$$

where $\bar\sigma_k$ and $\underline{\sigma}_k$ denote the maximum and the minimum of $\sigma_{i,k}$ for $i = 1, 2, \ldots, k$. The probabilities on the right are left-tail probabilities of chi-square distributions with k degrees of freedom, and can be expressed as integrals. The preceding display is bounded above by

$$2^{k/2} \exp\Big\{\sum_{i=1}^k \theta_{0,i}^2/\sigma_{i,k}^2\Big\} \Big(\frac{\bar\sigma_k}{\underline{\sigma}_k}\Big)^k \frac{\int_0^{j^2\epsilon_n^2/\bar\sigma_k^2} x^{k/2-1} e^{-x/2}\,dx}{\int_0^{\epsilon_n^2/(2\underline{\sigma}_k^2)} x^{k/2-1} e^{-x/2}\,dx}. \tag{9.19}$$

The exponential in the integral in the numerator is bounded by 1 and hence this integral is bounded above by $j^k \epsilon_n^k/(k\bar\sigma_k^k)$.

Now consider two separate cases.

(i) If $\epsilon_n^2/\underline{\sigma}_k^2$ remains bounded, then the exponential in the integral in the denominator of the expression in (9.19) is bounded below by a constant, and we have that the expression in (9.19) is bounded above by a multiple of $4^k j^k \exp(\sum_{i=1}^k \theta_{0,i}^2/\sigma_{i,k}^2)$.

(ii) If $\epsilon_n^2/\underline{\sigma}_k^2 \to \infty$, then we bound the integral in the denominator of the expression in (9.19) by $(\eta/2)^{k/2-1}\int_{\eta/2}^{\eta} e^{-x/2}\,dx$ for $\eta = \epsilon_n^2/(2\underline{\sigma}_k^2)$. This leads to the upper bound a multiple of $8^k j^k \exp\big(\sum_{i=1}^k \theta_{0,i}^2 \sigma_{i,k}^{-2}\big)\, \epsilon_n^2 \underline{\sigma}_k^{-2} \exp(\epsilon_n^2 \underline{\sigma}_k^{-2}/8)$.

By assumption (9.18) we have that $\underline{\sigma}_k^2 \gtrsim k^{-(2\alpha+1)} \asymp n^{-1}$. We also have that $k \asymp n\epsilon_n^2$. It follows that $\epsilon_n^2/\underline{\sigma}_k^2 \lesssim n\epsilon_n^2$, and $\underline{\sigma}_k^{-2}$ is bounded by a polynomial in k. Thus (8.40) holds if ϵ_n satisfies $\sum_{i=1}^k \theta_{0,i}^2/\sigma_{i,k}^2 \lesssim n\epsilon_n^2$, and $\sum_{i>k}\theta_{0,i}^2 \le \epsilon_n^2/2$. Since $\sum_{i=1}^\infty \theta_{0,i}^2 i^{2\alpha} < \infty$, then for ϵ_n a sufficiently large multiple of $n^{-\alpha/(2\alpha+1)}$, all required conditions hold, proving the theorem. □

9.5.5 Nonparametric Regression Using Splines

Consider the nonparametric regression model, where we observe independent random variables $X_1, \ldots, X_n$ distributed as $X_i = f(z_i) + \varepsilon_i$ for an unknown regression function $f\colon [0,1] \to \mathbb{R}$, deterministic covariates $z_1, \ldots, z_n$ in $[0,1]$, and $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{Nor}(0, \sigma^2)$, for $i = 1, \ldots, n$. For simplicity, we assume that the error variance σ^2 is known. Let $\mathbb{P}_n^z = n^{-1}\sum_{i=1}^n \delta_{z_i}$ be the empirical measure of the covariates, and let $\|\cdot\|_{2,n}$ denote the norm of $\mathbb{L}_2(\mathbb{P}_n^z)$.

Assume that the true regression function f_0 belongs to the unit ball of a Hölder class $\mathfrak{C}^\alpha[0,1]$ defined in Definition C.4 for some $\alpha \ge \frac{1}{2}$, and without loss of generality by rescaling. We shall construct priors based on a spline series representation like in Section E.2, namely, $f_\beta(z) = \beta^\top B(z)$ and induce a prior on f from a prior on $\beta = (\beta_1, \ldots, \beta_J)$, for instance by $\beta_j \overset{\text{iid}}{\sim} \mathrm{Nor}(0,1)$, $j = 1, \ldots, J$.[1]

We need the regressors $z_1, z_2, \ldots, z_n$ to be sufficiently regularly distributed in the interval $[0,1]$. In view of the spatial separation property of the B-spline functions, the precise condition can be expressed in the covariance matrix $\Sigma_n = (\!(\int B_i B_j \, d\mathbb{P}_n^z)\!)$, namely

$$J^{-1}\|\beta\|^2 \lesssim \beta^\top \Sigma_n \beta \lesssim J^{-1}\|\beta\|^2, \tag{9.20}$$

where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^J$.

Under condition (9.20) we have that, for all $\beta_1, \beta_2 \in \mathbb{R}^J$,

$$C\|\beta_1 - \beta_2\| \le \sqrt{J}\|f_{\beta_1} - f_{\beta_2}\|_{2,n} \le C'\|\beta_1 - \beta_2\|, \tag{9.21}$$

for some constants C and C'. This enables us to perform all calculations in terms of the Euclidean norms on the spline coefficients.

Theorem 9.22 *If $J = J_n \asymp n^{1/(1+2\alpha)}$, then the posterior contracts at the minimax rate $n^{-\alpha/(1+2\alpha)}$ relative to $\|\cdot\|_{2,n}$.*

Proof We verify the conditions of Theorem 8.23. Let f_{β_n} be the $\mathbb{L}_2(\mathbb{P}_n^z)$-projection of f_0 onto the J-dimensional space of splines $f_\beta = \beta^\top B$. Then $\|f_{\beta_n} - f_\beta\|_{2,n} \le \|f_0 - f_\beta\|_{2,n}$ for every $\beta \in \mathbb{R}^J$ and hence, by (9.21), for every $\epsilon > 0$, we have $\{\beta\colon \|f_\beta - f_0\|_{2,n} \le \epsilon\} \subset \{\beta\colon \|\beta - \beta_n\| \le C'\sqrt{J}\epsilon\}$. It follows that the ϵ-covering numbers of the set $\{f_\beta\colon \|f_\beta -$

[1] Unlike the case for densities, in the present situation a restriction to a compact interval is unnecessary.

$f_0\|_{2,n} \le \epsilon\}$ for $\|\cdot\|_{2,n}$ are bounded by the $C\sqrt{J}\epsilon$-covering numbers of a Euclidean ball of radius $C'\sqrt{J}\epsilon$, which are of the order D^J for some constant D. Thus the entropy condition (8.27) is satisfied provided that $J \lesssim n\epsilon_n^2$.

By the projection property, with β_∞ as in Lemma E.5,

$$\|f_{\beta_n} - f_0\|_{2,n} \le \|f_{\beta_\infty} - f_0\|_{2,n} \le \|f_{\beta_\infty} - f_0\|_\infty \lesssim J^{-\alpha}. \tag{9.22}$$

Combination with (9.21) shows that there exists a constant C'' such that for every $\epsilon \gtrsim 2J^{-\alpha}$, $\{\beta\colon \|f_\beta - f_0\|_n \le \epsilon\} \supset \{\beta\colon \|\beta - \beta_n\| \le C''\sqrt{J}\epsilon\}$. Thus

$$\begin{aligned}\frac{\Pi_n(f\colon \|f - f_0\|_n \le j\epsilon)}{\Pi_n(f\colon \|f - f_0\|_n \le \epsilon)} &\le \frac{\Phi_J(0, I)(\beta\colon \|\beta - \beta_n\| \le C' j\sqrt{J}\epsilon)}{\Phi_J(0, I)(\beta\colon \|\beta - \beta_n\| \le C''\sqrt{J}\epsilon)} \\ &\le \frac{\Phi_J(0, I)(\beta\colon \|\beta\| \le C' j\sqrt{J}\epsilon)}{2^{-J/2}e^{-\|\beta_n\|^2}\Phi_J(0, I)(\beta\colon \|\beta\| \le C''\sqrt{J}\epsilon/\sqrt{2})},\end{aligned}$$

where Φ_J refers to the normal probability content in $\mathbb{R}^J$. This follows since in the last step, the numerator increases if we replace the centering β_n by the origin in view of Lemma K.12, whereas the normal densities satisfy

$$\frac{\phi_{\beta_n, I}(\beta)}{\phi_{0,I/2}(\beta)} = \frac{e^{-\|\beta-\beta_n\|^2/2}}{2^{J/2}e^{-\|\beta\|^2}} \ge 2^{-J/2}e^{-\|\beta_n\|^2}.$$

Here, by the triangle inequality, (9.21) and (9.22), we have that $\|\beta_n\| \lesssim \sqrt{J}\|f_{\beta_n}\|_{2,n} \lesssim \sqrt{J}(J^{-\alpha} + \|f_0\|_\infty) \lesssim \sqrt{J}$. Furthermore, the two Gaussian probabilities are left tail probabilities of the chi-square distribution with J degrees of freedom. The quotient can be evaluated as

$$2^{J/2}e^{\|\beta_n\|^2}\frac{\int_0^{(C')^2 j^2 J\epsilon^2} x^{J/2-1}e^{-x/2}\,dx}{\int_0^{(C'')^2 J\epsilon^2/2} x^{J/2-1}e^{-x/2}\,dx}.$$

This is bounded above by $(Cj)^J$ for some constant C, by rescaling the domain of integration in the numerator to match with the denominator and using the fact that $\|\beta_n\| \le \sqrt{J}\|\beta\|_\infty = O(\sqrt{J})$. Hence to satisfy (8.26) it suffices again that $n\epsilon_n^2 \gtrsim J$.

We conclude the proof by choosing $J = J_n \asymp n^{1/(1+2\alpha)}$. □

Nonparametric Regression Using Orthonormal Series

The arguments in the preceding subsection use the special nature of the B-spline basis only through the approximation result Lemma E.5 and the comparison of norms (9.21). Theorem 9.22 thus may be extended to many other possible bases. For instance, we can use a sequence of orthonormal bases with good approximation properties for a given class of regression functions f_0. Then (9.20) should be replaced by

$$\|\beta_1 - \beta_2\| \lesssim \|f_{\beta_1} - f_{\beta_2}\|_{2,n} \lesssim \|\beta_1 - \beta_2\|. \tag{9.23}$$

This is trivially true if the bases are orthonormal in $\mathbb{L}_2(\mathbb{P}_n^z)$, such as the discrete wavelet bases relative to the design points. (Then the basis functions must necessarily change with the design points $z_1, \ldots, z_n$.)

9.5.6 Binary Nonparametric Regression with a Dirichlet Process Prior

We revisit the example in Section 7.4.3 and obtain the contraction rate. The observations are independent Bernoulli variables $Y_1, \dots, Y_n$ with $Y_i \overset{\text{ind}}{\sim} \mathrm{Bin}(1, H(x_i))$ for deterministic, known real-valued covariates $x_1, \dots, x_n$ and an unknown, monotone function $H\colon \mathbb{R} \to (0, 1)$. The prior is constructed as $H(x) = F(\alpha + \beta x)$, where $F \sim \mathrm{DP}(\gamma)$ is independent of (α, β), which receives a prior on $\mathbb{R} \times (0, \infty)$. Thus H possesses a mixture of Dirichlet process prior, and given (α, β), follows the Dirichlet process prior with base measure $\gamma((\cdot - \alpha)/\beta)$.

Assume that the true function H_0 is continuous and that $x_1, \dots, x_n$ lie in an interval $[a, b]$ strictly within its support. Furthermore assume that γ is absolutely continuous with a positive, continuous density on its support.

Theorem 9.23 *If there exists a compact set $\mathbb{K}$ inside the support of the prior for (α, β) such that the support of the base measure $\gamma((\cdot - \alpha)/\beta)$ strictly contains the interval $[a, b]$ for every $(\alpha, \beta) \in \mathbb{K}$, then the posterior distribution of H contracts at the rate $n^{-1/3}(\log n)^{1/3}$ with respect to the root mean square Hellinger distance d_n.*

Proof Let $\mathbb{P}_n$ be the empirical distribution of $x_1, \dots, x_n$. In view of Problem B.2,

$$d^2_{n,H}(H_1, H_2) \le \int |H_1^{1/2} - H_2^{1/2}|^2 \, d\mathbb{P}_n + \int |(1 - H_1)^{1/2} - (1 - H_2)^{1/2}|^2 \, d\mathbb{P}_n.$$

Both the class of all functions $H^{1/2}$ and the class of all functions $(1 - H)^{1/2}$ when H ranges over all cumulative distribution functions, possess ϵ-entropy bounded by a multiple of $1/\epsilon$ by Proposition C.1. Thus any $\epsilon_n \gtrsim n^{-1/3}$ satisfies (8.27).

By Problem B.2 again, for K_i and $V_{2;i}$ the Kullback-Leibler divergence and variation for the distributions of Y_i,

$$K_i(H_0; H) = H_0(x_i) \log \frac{H_0(x_i)}{H(x_i)} + (1 - H_0(x_i)) \log \frac{1 - H_0(x_i)}{1 - H(x_i)},$$
$$V_i(H_0; H) = H_0(x_i) \log^2 \frac{H_0(x_i)}{H(x_i)} + (1 - H_0(x_i)) \log^2 \frac{1 - H_0(x_i)}{1 - H(x_i)}.$$

By assumption, the numbers $H_0(x_i)$ are bounded away from 0 and 1. Therefore, by a Taylor expansion it can be seen that, with $\|H - H_0\|_\infty = \sup\{|H(z) - H_0(z)|\colon z \in [a, b]\}$,

$$\frac{1}{n} \sum_{i=1}^{n} K_i(H_0; H) \vee V_{2;i}(H_0; H) \lesssim \|H - H_0\|_\infty^2.$$

Hence, in order to verify (8.26), it suffices to lower bound the prior probability of the set $\{H\colon \|H - H_0\|_\infty \le \epsilon\}$.

For a given $\epsilon > 0$, partition the line into $N \lesssim \epsilon^{-1}$ intervals $E_1, \dots, E_N$ such that $H_0(E_j) \le \epsilon$ and such that $A\epsilon \le \gamma((E_j - \alpha)/\beta) \le 1$ for all $j = 1, \dots, N$ and $(\alpha, \beta) \in \mathbb{K}$, for some $A > 0$. Existence of such a partition follows from the continuity of H_0 and absolute continuity of γ. It can be seen that $\|H - H_0\|_\infty \lesssim \epsilon$ for every H such that $\sum_{j=1}^N |H(E_j) - H_0(E_j)| \le \epsilon$. To compute the prior probability of $\sum_{j=1}^N |H(E_j) - H_0(E_j)| \le \epsilon$, first condition on a fixed value of (α, β) in $\mathbb{K}$. By

Lemma G.13, the prior probability is at least $\exp(-c\epsilon^{-1}\log_-\epsilon)$ for some constant c. A uniform estimate works for all $(\alpha,\beta)\in\mathbb{K}$. Hence (8.26) holds for ϵ_n the solution of $n\epsilon^2=\epsilon^{-1}\log_-\epsilon$, or $\epsilon_n=n^{-1/3}(\log n)^{1/3}$. □

9.5.7 Interval Censoring Using a Dirichlet Process Prior

Let $T_1,T_2,\ldots,T_n$ be an i.i.d. sample from a life distribution F on $(0,\infty)$. However, the observations are subject to interval censoring by deterministic intervals $(l_1,u_1),\ldots,(l_n,u_n)$. Along with the values of these interval end points, we observe $(\delta_1,\eta_1),\ldots,(\delta_n,\eta_n)$, where $\delta_i=\mathbb{1}\{T_i\le l_i\}$ and $\eta_i=\mathbb{1}\{l_i<T_i<u_i\}$, $i=1,2,\ldots,n$. The observations (δ_i,η_i), $i=1,2,\ldots,n$ are therefore i.n.i.d. multinomial with probability mass function

$$(F(l_i))^{\delta_i}(F(u_i)-F(l_i))^{\eta_i}(1-F(u_i))^{1-\delta_i-\eta_i}. \tag{9.24}$$

Theorem 9.24 *Consider a prior $F\sim\mathrm{DP}(\alpha)$. Let the true distribution function F_0 be continuous. Assume that $(l_1,u_1),\ldots,(l_n,u_n)$ lie in an interval $[a,b]$ where $F_0(a-)>0$ and $F_0(b)<1$ and the density of α is positive and continuous on $[a,b]$. Then the posterior contracts at the rate $n^{-1/3}(\log n)^{1/3}$ with respect to the root average square Hellinger distance d_n.*[2]

Proof If F_1 and F_2 are two life distributions, then the squared Hellinger distance between the distributions of (δ_i,η_i) under F_1 and F_2 is given by

$$\begin{aligned}(\sqrt{F_1(l_i)}-\sqrt{F_2(l_i)})^2+(\sqrt{F_1(u_i)-F_1(l_i)}-\sqrt{F_2(u_i)-F_2(l_i)})^2&\\+(\sqrt{1-F_1(u_i)}-\sqrt{1-F_2(u_i)})^2,&\end{aligned}$$

and hence $d_n^2(F_1,F_2)$ is given by

$$\begin{aligned}\int(\sqrt{F_1(l)}-\sqrt{F_2(l)})^2\,d\mathbb{P}_n(u,l)+\int(\sqrt{F_1(u)-F_1(l)}-\sqrt{F_2(u)-F_2(l)})^2\,d\mathbb{P}_n(u,l)&\\+\int(\sqrt{1-F_1(u)}-\sqrt{1-F_2(u)})^2\,d\mathbb{P}_n(u,l),&\end{aligned} \tag{9.25}$$

where $\mathbb{P}_n$ stands for the empirical distribution of $(l_1,u_1),\ldots,(l_n,u_n)$. Therefore, to bound the ϵ-entropy with respect to d_n, it suffices to bound the $\mathbb{L}_2(\mathbb{P}_n)$ entropies of the classes $(l,u)\mapsto(F(l))^{1/2}$, $(l,u)\mapsto(F(u)-F(l))^{1/2}$ and $(l,u)\mapsto(1-F(u))^{1/2}$. As in Section 9.5.6, it follows from Proposition C.1 that the first and the third are of the order ϵ^{-1}. It therefore suffices to bound the second.

Let $\mathbb{H}_n=(2n)^{-1}\sum_{i=1}^n(\delta_{l_i}+\delta_{u_i})$. Let $(\underline{H}_1,\overline{H}_1),\ldots,(\underline{H}_N,\overline{H}_N)$ be an ϵ-bracketing for the class of life distributions with respect to $\mathbb{L}_1(\mathbb{H}_n)$. By Proposition C.1 again, we may choose $\log N\lesssim\epsilon^{-1}$. We shall show that for any bracket $(\underline{H},\overline{H})$ any distributions F_1,F_2 enclosed by the bracket, the second term of (9.25) is at most 2ϵ. Then it will follow that the

[2] If the censoring intervals are stochastic and are enclosed inside $[a,b]$ with $0<F_0(a-)<F_0(b)<1$, then by similar arguments it also follows that the posterior contracts at the rate $n^{-1/3}(\log n)^{1/3}$ for the Hellinger distance.

ϵ-entropy of the class of all life distributions for the metric d_n is bounded by a multiple of ϵ^{-1}. Hence (8.27) will hold for $\epsilon_n = n^{-1/3}$.

Without loss of generality, we may assume that $\underline{H}$ and $\overline{H}$ are monotone increasing and $0 \leq \underline{H} \leq \overline{H} \leq 1$. Take any two distributions F_1, F_2 enclosed by $(\underline{H}, \overline{H})$. Then for $j = 1, 2$, $l \leq u$, we have $(\underline{H}(u) - \overline{H}(l))_+ \leq F_j(u) - F_j(l) \leq \overline{H}(u) - \underline{H}(l)$, and hence using $(a-b)^2 \leq a^2 - b^2$ for $a \geq b \geq 0$, we obtain

$$\left(\sqrt{F_1(u) - F_1(l)} - \sqrt{F_2(u) - F_2(l)}\right)^2 \leq (\overline{H}(u) - \underline{H}(l)) - (\underline{H}(u) - \overline{H}(l))_+. \qquad (9.26)$$

If $\underline{H}(u) \geq \overline{H}(l)$, then the right-hand side of (9.26) is equal to $(\overline{H}(u) - \underline{H}(u)) + (\overline{H}(l) - \underline{H}(l))$. If $\underline{H}(u) < \overline{H}(l)$, then the right-hand side of (9.26) is equal to

$$\begin{aligned}\overline{H}(u) - \underline{H}(l) &= (\overline{H}(u) - \underline{H}(u)) + (\underline{H}(u) - \overline{H}(l)) + (\overline{H}(l) - \underline{H}(l))\\ &\leq (\overline{H}(u) - \underline{H}(u)) + (\overline{H}(l) - \underline{H}(l)).\end{aligned}$$

Therefore, the second terms in (9.25) can be bounded by $2\int(\overline{H}(x) - \underline{H}(x))\, d\mathbb{H}_n(x) < 2\epsilon$.

To estimate the prior probability in (8.26), we proceed as in Subsection 9.5.6. By similar arguments, the required probability is bounded below by the probability of $\{F\colon \|F - F_0\|_\infty \leq c\epsilon\}$ under the Dirichlet process, which is at least a multiple of $\exp[-\beta\epsilon^{-1}\log_- \epsilon]$. □

9.6 Historical Notes

Log-spline models were introduced for maximum likelihood density estimation by Stone (1990), and generalized to the multivariate case by Stone (1994). Ghosal et al. (2000) used the log-spline model to construct a prior distribution on smooth densities, as in Theorem 9.1. Example 9.7 is also from this paper. The rate of contraction for Bernstein polynomial mixtures was obtained by Ghosal (2001) for supersmooth Bernstein polynomials and for twice continuously differentiable densities. The improved contraction rate using coarsened Bernstein polynomials was derived by Kruijer and van der Vaart (2008) for smoothness level $0 < \alpha < 2$; they established adaptation to smoothness and a nearly minimax rate in the range $0 < \alpha \leq 1$. The minimax rate of estimation for analytic functions was obtained by Ibragimov and Has'minskiĭ (1982) with respect to $\mathbb{L}_p$-distances. Ghosal and van der Vaart (2001) obtained the posterior contraction rate with respect to the Hellinger distance of Dirichlet process mixtures of the location and location-scale family of univariate normals when the true density is a normal mixture. The rate matches the minimax rate up to a logarithmic factor. They introduced the moment-matching technique for entropy calculation. However the prior on σ was supported in a bounded interval, ruling out the common conjugate inverse-gamma prior. Posterior contraction rates for ordinary smooth normal mixtures were first obtained Ghosal and van der Vaart (2007b) for univariate twice continuously differentiable densities and a sequence of priors on the scale. Their result was subsequently improved in steps. Kruijer et al. (2010), obtained the (nearly) optimal rate for any smoothness level of the underlying density (i.e., adaptive posterior contraction rate) using discrete mixtures of normal densities with a random number of points, but their results did not cover the case of Dirichlet process mixtures. The primary technique used in their result was an adaptive construction of a KL-approximation in the model for the true density that improves with the level of smoothness of the true density. A construction of this type was first used by Rousseau (2010) in the context of beta mixtures. De Jonge and van Zanten (2010) constructed a similar approximation in the multivariate case for global Hölder class of functions. Finally the (adaptive) posterior contraction rate for ordinary smooth normal mixtures using Dirichlet process mixtures of normals with a general prior on scale was obtained by Shen et al. (2013). They also treated anisotropic smooth classes of multivariate densities (see Problem 9.17 below). Their construction of the adaptive KL-approximation uses only local Hölder classes, and can be applied to the log-density. They also introduced the entropy calculation based on the stick-breaking representation of the Dirichlet process. The unification of the the smooth and supersmooth cases (which in particular contains multivariate supersmooth mixtures) presented in the present chapter appears to be new. Posterior rates of contraction for the examples in Section 9.5 were obtained by Ghosal and van der Vaart (2007a).

Problems

9.1 (Ghosal and van der Vaart 2001) Let F^* be any probability measure with sub-Gaussian tails and σ^* be fixed. Show that the density $p_{F^*,\sigma^*}(x) = \int \phi_{\sigma^*}(x-z)\, dF^*(z)$ also has sub-Gaussian tails.

9.2 (Ghosal and van der Vaart 2001) Let $X_i \overset{\text{iid}}{\sim} p$, $i = 1, 2, \ldots$, where p is a density in $\mathbb{R}$ that can be written as (i) $p(x) = \int \phi_{\sigma_0}(x - z)\, dF(z)$, (ii) $p(x) = \int\int \phi_\sigma(x - z)\, dF(z)\, dG(\sigma)$, or (iii) $p(x) = \int \phi_\sigma(x - z)\, dH(z, \sigma)$. In model (i), let F be any probability measure supported on $[-a, a]$ where $a \lesssim (\log n)^\gamma$ and σ take any arbitrary value on the interval $[\underline{\sigma}, \overline{\sigma}]$. For model (ii), we assume that F is as above and the distribution G of σ is supported on $[\underline{\sigma}, \overline{\sigma}]$. In model (iii), the distribution H for (z, σ) is supported on $[-a, a] \times [\underline{\sigma}, \overline{\sigma}]$, where a is as above. Assume that the true F_0 has compact support in case of model (i) and (ii); for model (iii), H_0 is assumed to have compact support. Show that for a sufficiently large constant M, the MLE $\hat{p}_n$ satisfies $P_0(d(\hat{p}_n, p_0) > M\epsilon_n) \lesssim e^{-c\log^2 n}$ where $\epsilon_n = (\log n)^{\max(\gamma,\frac{1}{2})+\frac{1}{2}}/\sqrt{n}$ for models (i) and (ii) while $\epsilon_n = (\log n)^{2\max(\gamma,\frac{1}{2})+\frac{1}{2}}/\sqrt{n}$ for model (iii), and c is a constant. In particular, $\hat{p}_n$ converges to p_0 in Hellinger distance at the rate ϵ_n in P_0-probability, and a.s. $[P_0]$. (It is known that the MLE exists; see Theorem 18 of Lindsay 1995.)

9.3 (Ghosal and van der Vaart 2001) Consider the setting of Problem 9.2. For model (i), let $\hat{p}_{n,k}$ be the maximizer of the likelihood on $\{p_{F,\sigma}: F = \sum_{j=1}^k w_j\delta_{z_j}, w_j \geq 0, \sum_{j=1}^k w_j = 1, z_j \in [-a, a]\}$. Define $\hat{p}_{n,k}$ in model (ii) similarly by restricting the mixing distributions F and G to have at most k support points, while in model (iii), let H to be supported on k^2 points. Show that if $k \geq C\log n$ for some sufficiently large C, then $\hat{p}_{n,k}$ converges at the rate ϵ_n given in Problem 9.2.

9.4 (Ghosal and van der Vaart 2001) Let $X_i \overset{\text{iid}}{\sim} p$, $i = 1, 2, \ldots$, where p belongs to one of the models (i), (ii) or (iii). Consider the sieve $\mathcal{P}_n = \{g_1, \ldots, g_N\}$, where $N = N_{[\,]}(\epsilon_n, \mathcal{P}, d)$, ϵ_n is the solution of the entropy equation $\log N_{[\,]}(\epsilon, \mathcal{P}, d) \leq n\epsilon^2$, $g_j = u_j/\int u_j$, $j = 1, \ldots, N$, and $[l_1, u_1], \ldots, [l_N, u_N]$ is a Hellinger bracketing for $\mathcal{P}$ of size ϵ_n. If we choose $a \lesssim \sqrt{\log n}$, then $\epsilon_n \lesssim n^{-1/2}\log n$ for models (i) and (ii), $\epsilon_n \lesssim n^{-1/2}(\log n)^{3/2}$ for model (iii) and the sieve MLE $\hat{p}_n$ satisfies $P_0(d(\hat{p}_n, p_0) > M\epsilon_n) \lesssim e^{-c\log^2 n}$.

9.5 (Ghosal and van der Vaart 2001) The sieve in Problem 9.4 can also be used to construct a prior for which the posterior converges at rate ϵ_n. Put the uniform distribution Π_j on $\mathcal{P}_j$ and consider the prior $\Pi = \sum_{j=1}^{\infty} \lambda_j \Pi_j$, where $\lambda_j > 0$, $\sum_{j=1}^{\infty} \lambda_j = 1$ and $\log \lambda_j^{-1} = O(\log j)$ as $j \to \infty$. Alternatively, for a sample of size n, simply consider the prior Π_n. Using Theorem 8.15, show that the posterior contracts at the intended rate ϵ_n.

9.6 (Ghosal and van der Vaart 2001) Assume the setup and conditions of Theorem 9.9, except that $m = \alpha(\mathbb{R})$ varies with the sample size. If $1 \lesssim m \lesssim \log n$ and p_0 is supersmooth, then for a sufficiently large constant M, show that the posterior contracts at the same rate as given in the theorem. Formulate a similar result for the β-smooth case.

9.7 (Ghosal and van der Vaart 2001) Assume the setup of Theorem 9.9, but the base measure α of the Dirichlet process prior for F depends on a parameter θ, where θ is given an arbitrary prior and α_θ are base measures satisfying the following conditions:

(i) $\alpha_\theta(\mathbb{R})$ is bounded above and below in θ,

(ii) There exist constants B, b and $\delta > 0$ such that $\alpha_\theta(z\colon |z| > t) \leq Be^{-bt^\delta}$ for all $t > 0$ and θ,

(iii) Every α_θ has a density α'_θ such that for some $\epsilon > 0$, $\alpha'_\theta(x) \geq \epsilon$ for all θ and $x \in [-k_0, k_0]$.

(These conditions usually hold if the parametric family α_θ is "well behaved" and θ has a compact range. However, conditions (ii) and (iii) are not expected to hold if the range of θ is unbounded.) If p_0 is supersmooth, show that the posterior contracts at the same rate given in the theorem. Formulate a similar result for the β-smooth case.

9.8 (Ghosal and van der Vaart 2001) In certain situations, the conclusion of Problem 9.7 may still hold even if the hyperparameters are not compactly supported. Let $d = 1$ and the base measure α be $\mathrm{Nor}(\mu, \tau)$, $\mu \sim \mathrm{Nor}(\mu_0, A)$ and τ is either given, or has compactly supported prior distribution. Then the posterior contracts at the same rate.

9.9 (Walker et al. 2007, Xing 2010) Consider a model consisting of univariate supersmooth normal location mixtures with a known σ and the let true density be $\phi_\sigma * F_0$ for some compactly supportd F_0. For all $n \in \mathbb{N}$, define $a_{nj} \uparrow \infty$ and consider a covering for the space of densities given by $\mathcal{G}_{\underline{\sigma},a_{n1},\delta_n^2} = \{\phi_\sigma * F\colon F([-a_{n1}, a_{n1}]) \geq 1 - \delta_n^2\}$, $\mathcal{G}_{\underline{\sigma},a_{nj},\delta_n^2} = \{\phi_\sigma * P\colon F([-a_{nj}, a_{nj}]) \geq 1 - \delta_n^2, F([-a_{n,j-1}, a_{n,j-1}]) \geq 1 - \delta_n^2\}$, $j \geq 2$. Show that for $\delta_n = n^{-1/2} \log n$, we have $e^{-n\delta_n^2} \sum_{j=1}^{\infty} \sqrt{\Pi(\mathcal{G}_{\underline{\sigma},a_{nj},\delta_n^2})} \to 0$. Hence using Problem 8.11, conclude that the posterior contracts at the rate $n^{-1/2} \log n$ a.s. Reach the same conclusion using Problem 8.13.

9.10 (Ghosal and van der Vaart 2007b) Let p_0 be a twice continuously differentiable probability density.

(i) If $\int (p_0''/p_0)^2\, p_0 \, d\lambda < \infty$ and $\int (p_0'/p_0)^4\, p_0 \, d\lambda < \infty$, then $d_H(p_0, p_0 * \phi_\sigma) \lesssim \sigma^2$.

(ii) If p_0 is a bounded with $\int |p_0''| \, d\lambda < \infty$, then $\|p_0 - p_0 * \phi_\sigma\|_1 \lesssim \sigma^2$.

In both cases the constants in "$\lesssim$" depend on the given integrals only.

9.11 (Ghosal and van der Vaart 2007b, Wu and Ghosal 2010) Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} P$, where P has density p, and let the true value of p be p_0, which is bounded. Let p be given a Dirichlet mixture prior Π: $p = F * \phi_\sigma$, $F \sim \mathrm{DP}(\alpha)$, $\sigma/\sigma_n \sim G$. Let α have positive and continuous density on $[-a, a]$, where a is fixed. Then for any $\epsilon > 0$ and $0 < b < a\sigma_n^{-1}$, there exists K not depending on n such that

$$P_0[\Pi_n(F[-2a, 2a]^c > \epsilon | X_1, \dots, X_n)\mathbb{1}\{\max_{1\le i\le n} |X_i| \le a\}]$$
$$\lesssim P_0\Pi_n(\sigma > b\sigma_n | X_1, \dots, X_n) + \frac{\alpha[-2a, 2a]^c}{\epsilon(\alpha(\mathbb{R}) + n)} + Kn\epsilon^{-1}e^{-a^2/4b^2\sigma_n^2}.$$

Moreover, if P_0 is compactly supported, α has positive and continuous density on an interval containing $\mathrm{supp}(P_0)$, $b_n \to \infty$ such that $b_n\sigma_n \to 0$, $n\epsilon_n^{-2}e^{-a^2/4b_n^2\sigma_n^2} \to 0$ and $\mathrm{P}(\sigma > b_n\sigma_n) = o(e^{-n\epsilon_n^2})$ for some sequence ϵ_n satisfying Condition (8.4) on prior concentration rate, then

$$P_0^n\Pi_n(F\colon F[-2a, 2a]^c > \epsilon_n^2 | X_1, \dots, X_n) \to 0. \tag{9.27}$$

Generalize the result when $a = a_n \to \infty$.

9.12 (Ghosal and van der Vaart 2007b) Let $X_i \overset{\text{iid}}{\sim} p$, $i = 1, 2, \dots$, $p \in \mathcal{P}$. Consider the sieve $\mathcal{P}_n = \{p_{F,\sigma}\colon F[-a_n, a_n] = 1, b_1\sigma_n \le \sigma \le b_2\sigma_n\}$, where $a_n \ge e$ and $\sigma_n \to 0$ are positive sequences such that $\log n \lesssim \log(a_n/\sigma_n) \lesssim \log n$. Let $\hat{p}_n = \arg\max\{\prod_{i=1}^n p(X_i)\colon p \in \mathcal{P}_n\}$. Assume that P_0 has compact support and $[-a_n, a_n] \supset \mathrm{supp}(P_0)$ for all sufficiently large n. Using Theorem F.4, show that the sieve-MLE $\hat{p}_n$ converges at the rate $\epsilon_n = \max\{(n\sigma_n)^{-1/2}a_n \log n, \sigma_n^2\}$.

9.13 (Ghosal and van der Vaart 2007b) Consider the setting of Problem 9.12 except that P_0 is not compactly supported. Consider the sieve $\mathcal{P}_n = \{p_{F,\sigma}\colon F[-a, a]^c \le A(a)$ for all $a > 0, b_1\sigma_n \le \sigma \le b_2\sigma_n\}$, where $A(a) = e^{-da^{1/\delta}}$, $d, \delta > 0$ constants and $\log n \lesssim \log\sigma_n^{-1} \lesssim \log n$. Assume that $P_0[-a, a]^c \le A(a)$ for every $a > 0$ and $p_0/(p_0 * \phi_{\sigma_n})$ are uniformly bounded. Show that the sieve-MLE $\hat{p}_n$ converges at the rate $\epsilon_n = \max\{(n\sigma_n)^{-1/2}(\log n)^{1+(1\vee 2\delta)/4}, \sigma_n^2\}$.

9.14 (Ghosal and van der Vaart 2007b) If p_0 is increasing on $(-\infty, a]$, bounded below on $[a, b]$ and decreasing on $[b, \infty)$ for some $a < b$, show that $p_0/(p_0 * \phi_{\sigma_n})$ are uniformly bounded in Problem 9.13.

9.15 (Ghosal and van der Vaart 2007b) The sieves $\mathcal{P}_n$ in Problem 9.12 and Problem 9.13 can also be used to construct a prior for which the posterior contracts at the same rate ϵ_n as the corresponding sieve-MLEs. Consider a minimal collection of Hellinger ϵ_n-brackets that cover $\mathcal{P}_n$ and impose the uniform prior Π_n on the renormalized upper brackets. Using Theorem 8.15, show that under the same assumptions on p_0, the resulting posterior converges at the rate ϵ_n.

9.16 (Shen et al. 2013) For a finite mixture prior specification Π, where the density function f is represented by $f(x) = \sum_{j=1}^N \omega_j\phi_\Sigma(x - \mu_j)$ and priors are assigned on N, Σ, $\omega = (\omega_1, \dots, \omega_N)$ and $\mu_1, \dots, \mu_N$. We assume $\Sigma \sim G$, which satisfies (9.6), (9.7) and (9.8), and that there exist positive constants $a_4, b_4, b_5, b_6, b_7, C_4, C_5, C_6, C_7$ such that for sufficiently large $x > 0$,

$$b_4\exp\{-C_4t(\log t)^{\tau_1}\} \le \Pi(N \ge t) \le b_5\exp\{-C_5t(\log t)^{\tau_1}\}$$

while for every fixed value of N,

$$\Pi(\mu_i \notin [-x, x]^d) \le b_6 \exp(-C_6 x^{a_4}), \quad \text{for sufficiently large } x > 0, \quad i = 1, \dots, N,$$
$$\Pi(\|\omega - \omega_0\| \le \epsilon) \ge b_7 \exp\{-C_7 h \log_- \epsilon\}, \quad \text{for all } 0 < \epsilon < 1/N \text{ and all } \omega_0 \in \mathbb{S}_N.$$

Show that the conclusion of Theorem 9.9 also holds for this prior.

9.17 (Shen et al. 2013) For any $a = (a_1, \dots, a_d)$ and $b = (b_1, \dots, b_d)$, let $\langle a, b\rangle$ denote $\sum_{j=1}^d a_j b_j$ and for $y = (y_1, \dots, y_d)$, let $\|y\|_1$ denote the ℓ_1-norm $\sum_{j=1}^d |y_j|$. For a $\beta > 0$, an $\alpha = (\alpha_1, \dots, \alpha_d) \in (0, \infty)^d$ with sum $\alpha = d$ and an $L\colon \mathbb{R}^d \to (0, \infty)$ satisfying $L(x + y) \le L(x) \exp(c_0 \|y\|_1^2)$ for all $x, y \in \mathbb{R}^d$ and some $c_0 > 0$, the α-anisotropic β-Hölder class with envelope L is defined as the set of all functions $f\colon \mathbb{R}^d \to \mathbb{R}$ that have continuous mixed partial derivatives $D^k f$ of all orders $k \in \mathbb{N}_0^d$, $\beta - \alpha_{\max} \le \langle k, \alpha\rangle < \beta$, with

$$|D^k f(x + y) - D^k f(x)| \le L(x) e^{c_0 \|y\|_1^2} \sum_{j=1}^d |y_j|^{\min(\beta/\alpha_j - k_j, 1)}, \quad x, y \in \mathbb{R}^d,$$

where $\alpha_{\max} = \max(\alpha_1, \dots, \alpha_d)$. We denote this set of functions by $\mathfrak{C}^{\alpha,\beta,L,c_0}(\mathbb{R}^d)$; here β refers to the mean smoothness and α refers to the anisotropy index. An $f \in \mathfrak{C}^{\alpha,\beta,L,c_0}$ has partial derivatives of all orders up to $\underline{\beta}_j$ along axis j where $\beta_j = \beta/\alpha_j$, and β is the harmonic mean $d/(\sum_{j=1}^d \beta_j^{-1})$ of these axial smoothness coefficients. (In the special case of $\alpha = (1, \dots, 1)$, the anisotropic set $\mathfrak{C}^{\alpha,\beta,L,c_0}(\mathbb{R}^d)$ equals the isotropic set $\mathfrak{C}^{\beta,L,c_0}(\mathbb{R}^d)$.)

Suppose that $p_0 \in \mathfrak{C}^{\alpha,\beta,L,c_0}(\mathbb{R}^d)$ is a probability density function satisfying

$$P_0 \left(|D^k p_0|/p_0\right)^{(2\beta+\epsilon)/\langle k,\alpha\rangle} < \infty, \quad k \in \mathbb{N}_0^d, \langle k, \alpha\rangle < \beta, \qquad P_0 \left(L/p_0\right)^{(2\beta+\epsilon)/\beta} < \infty \tag{9.28}$$

for some $\epsilon > 0$ and that $p_0(x) \le c e^{-b\|x\|^\tau}$, $\|x\| > b$, holds for some constants $a, b, c, \tau > 0$. If Π is as in Section 9.4, then the posterior contraction rate at p_0 in the Hellinger or the $\mathbb{L}_1$-metric is $\epsilon_n = n^{-\beta/(2\beta+d^*)}(\log n)^t$, where $t > \{d^*(1 + \tau^{-1} + \beta^{-1}) + 1\}/(2 + d^*/\beta)$, and $d^* = \max(d, \kappa\alpha_{\max})$.

9.18 (Kleijn and van der Vaart 2006) In a supersmooth normal location mixture model (assume standard deviation σ is known to be 1), consider a Dirichlet mixture prior with compact base measure. If the true density p_0 lies outside the model, then using the results of Section 8.5, show that the posterior contracts at $p_{F^*} = \phi * F^*$ at the rate $n^{-1/2} \log n$ provided that p_0/p_{F^*} is bounded, where $F^* = \operatorname{argmin} K(p_0; \phi * F)$.

9.19 (Petrone and Wasserman 2002, Ghosal 2001) Let $X_i \overset{\text{iid}}{\sim} p, i = 1, 2, \dots, p \in \mathcal{P}$, where p is a density on the unit interval. An extended Bernstein polynomial prior, which is supported on extended Bernstein polynomial densities $\sum_{r=1}^k \sum_{j=1}^r w_{j,r} \beta(x; j, r - j + 1)$. Show that if the true density is of the extended Bernstein type, then posterior contracts at the rate $n^{-1/2} \log n$.

9.20 (McVinish et al. 2009) Let $X_i \overset{\text{iid}}{\sim} p, i = 1, 2, \dots, p \in \mathcal{P}$, where p is a density on the unit interval. Consider a prior described by the triangular kernel given below:
Let $0 = t_0 < t_1 < \cdots < t_{k-1} < t_k = 1$ and let $\Delta_j(x)$ be the triangular density on $[t_{j-1}, t_{j+1}]$ with mode of height $2/(t_{j+1} - t_{j-1})$ at $t = t_j$, $j = 1, \dots, k - 1$. Also

define $\Delta_0(x) = 2(t_1 - x)/(t_1 - t_0)$, $t_0 \le x \le t_1$, and $\Delta_k(x) = 2(x - t_{k-1})/(t_k - t_{k-1})$, $t_{k-1} \le x \le t_k$. Let $p(x) = \sum_{j=0}^{k} w_j \Delta_j(x)$. Induce a prior on p in one of the following two ways:

Type I mixture: $t_j = j/k$, $(w_0, \ldots, w_k)|k \sim \mathrm{Dir}(k{+}1\colon \alpha_{0,k}, \ldots, \alpha_{k+1,k})$, $k \sim \rho(\cdot)$;
Type II mixture: $w_j = 1/(k+1)$, $(t_1, \ldots, t_{k-1})|k \sim h_k(\cdot)$, a positive density h_k on the set $\{(t_1, \ldots, t_{k-1})\colon t_1 < t_2 < \cdots < t_{k-1}\}$, $k \sim \rho(\cdot)$.

For the type I mixture prior, assume that $\{\alpha_{j,k}\}$ is uniformly bounded, $e^{-c_1 k \log k} \lesssim \rho(k) \lesssim e^{-c_2 k}$, and the true density $p_0 \in \mathfrak{C}^\beta[0, 1]$, $\beta \le 2$, $p_0(x) \ge x(1 - x)$. Then the posterior contracts at the rate $n^{-\beta/(2\beta+1)}(\log n)^{(4\beta+1)/4\beta}$ a.s.

For the type II mixture prior, assume that $e^{-c_1 k \log k} \lesssim \sum_{j=k+1}^{\infty} \rho(j) \lesssim e^{-c_2 k \log k}$, $\mathrm{P}(\max_j |t_j - t_{j-1}| < e^{-c_3 n^\gamma}|k) \le \exp(-c_4 n^{1/(2\beta+1)} \log n)$, $k \le k_0 n^{1/(2\beta+1)}$, and $h_k \gtrsim (\Gamma(k))^{-r}$ for some $r > 0$, and suppose that the true density $p_0 \in \mathfrak{C}^\beta(0, 1)$, $\beta \le 2$, p_0 bounded away from 0. Then the posterior contracts at the rate $n^{-\beta/(2\beta+1)} \log n$ a.s.

9.21 (Scricciolo 2006) Consider the estimation of a periodic density p on $[0, 1]$, $p \in \mathfrak{W}^\beta(L)$, the Sobolev space of smoothness index β with Sobolev norm $\|p\|_{2,2,\beta}$ bounded by L (see Section E.3) and a prior on p induced by the infinite-dimensional exponential family of trigonometric series

$$p(x) = \frac{\exp[\sqrt{2}\sum_{k=1}^{\infty}\{a_k \sin(2\pi k x) + b_k \cos(2\pi k x)\}]}{\int_0^1 \exp[\sqrt{2}\sum_{k=1}^{\infty}\{a_k \sin(2\pi k y) + b_k \cos(2\pi k y)\}]dy}, \qquad a_k, b_k \overset{\mathrm{ind}}{\sim} \mathrm{Nor}(0, k^{-2q})$$

conditioned on $\sum_{k=1}^{\infty} k^{2p}(a_k^2 + b_k^2) < \pi^{-2p} L^2$. Then the prior is supported within $\mathfrak{W}^\beta(L)$ and the posterior for p contracts at the rate $n^{-p/(2p+1)}$ at p_0 with respect to d_H.

9.22 (Ghosal and van der Vaart 2007a) For the Gaussian white noise model with conjugate prior on it, show that Theorem 9.21 can be modified to give the same rate of contraction for a fixed prior of the mixture type $\sum_n \lambda_n \Pi_n$ under suitable conditions on the weights λ_n.

9.23 (Jiang 2007) The posterior contraction rate theorems apply also to finite-dimensional models where the dimension grows with the sample size. Consider a generalized linear model with K_n-dimensional predictor $X \in [-1, 1]^{K_n}$, where $K_n \to \infty$, and $Y|\, X = x$ has density $p(y|\, x) = \exp\{a(x^\top \beta)y + b(x^\top \beta) + c(y)\}$ with respect to a dominating σ-finite measure ν. Suppose that $X_i \overset{\mathrm{iid}}{\sim} G$, the true value of β is β_0, the true value of p is p_0, and assume that the ℓ_1-sparsity condition $\limsup_{n\to\infty} \sum_{j=1}^{K_n} |\beta_{0j}| < \infty$ holds. Define a metric for p by $d^2(p_1, p_2) = \int\int (\sqrt{p}_1(y|\, x) - \sqrt{p}_2(y|\, x)^2\, d\nu(y)\, dG(x)$.

Let a prior be specified through the following variable selection scheme. Fix sequences $r_n, \bar{r}_n \to \infty$ with $r_n/K_n \to 0$ and $1 \le r_n \le \bar{r}_n \le K_n$. Let $\gamma_1, \ldots \gamma_{K_n}$ be indicators of inclusion of the corresponding predictors in the model, and set $|\gamma| = \sum_{j=1}^{K_n} \gamma_j$. Let $\gamma_1, \ldots \gamma_{K_n}$ be a priori i.i.d. $\mathrm{Bin}(1, r_n/K_n)$ conditioned on $|\gamma| \le \bar{r}_n$. Given $\gamma_1, \ldots \gamma_{K_n}$ let $\beta_j = 0$ if $\gamma_j = 0$ and let $(\beta_j\colon \gamma_j = 1) \sim \mathrm{Nor}(0, V_\gamma)$.

Define $\Delta(r_n) = \inf\{\sum_{j\notin\gamma} |\beta_{0j}|\colon |\gamma| = r_n\}$, $\underline{B}_n = \sup\{\mathrm{eig}(V_\gamma^{-1})\colon |\gamma| \le \bar{r}_n\}$, $\bar{B}_n = \sup\{\mathrm{eig}(V_\gamma)\colon |\gamma| \le \bar{r}_n\}$, and $D(R) = 1 + R \sup\{|a'(h)|\colon |h| \le R\} \sup\{|b'(h)/a'(h)|\colon |h| \le R\}$. Suppose that ϵ_n satisfy $0 < \epsilon_n < 1$, $n\epsilon_n^2 \gg 0$ and

(i) $\bar{r}_n \log_- \epsilon_n = o(n\epsilon_n^2)$;
(ii) $\bar{r}_n \log K_n = o(n\epsilon_n^2)$;
(iii) $\bar{r}_n \log D(\bar{r}_n\sqrt{n\bar{B}_n}\epsilon_n)$;
(iv) $\Delta(r_n) = o(\epsilon_n^2)$;
(v) $\underline{B}_n = o(n\epsilon_n^2)$;
(vi) $r_n \log \bar{B}_n = o(n\epsilon_n^2)$.

Then the posterior for p contracts at p_0 with respect to d_H at the rate ϵ_n at p_0 a.s. In particular, if $\max\{\underline{B}_n, \bar{B}_n\} \lesssim \bar{r}_n^v$ for some $v \geq 0$, $\bar{r}_n = o(n^{1/(4+v)})$, $\bar{r}_n/K_n \to 0$ and $\bar{r}_n \log K_n = o(n)$, then the posterior for p is consistent at p_0 with respect to d.

Under appropriate conditions $\log K_n$ can be as large as n^ξ, for $\xi < 1$, with ϵ_n of the order $n^{-(1-\xi)/2}(\log n)^{k/2}$ if $\bar{r}_n = o(\log^k n)$. The rate is close to the parametric rate $n^{-1/2}$ if ξ can be taken arbitrarily close to 0, for instance when K_n has polynomial growth.

9.24 (Jiang 2007) Under the setting of Problem 9.23, consider specific regression models to obtain explicit conditions. Let $\max\{\underline{B}_n, \bar{B}_n\} \lesssim r_n^v$ for some $v \geq 0$.

(a) *Poisson regression:* $Y|X \sim \mathrm{Poi}(e^{X^\top\beta})$. Assume (ii), (iv) and $\bar{r}_n^{4+v} = o(n\epsilon_n^2)$. Then the posterior contracts at the rate ϵ_n.
(b) *Normal regression:* $Y|X \sim \mathrm{Nor}(X^\top\beta, 1)$. Assume (ii), (iv) and $\bar{r}_n^v = o(n\epsilon_n^2)$. Then the posterior contracts at the rate ϵ_n.
(c) *Exponential regression:* $Y|X \sim \mathrm{Exp}(e^{X^\top\beta})$. Assume (ii), (iv) and $\bar{r}_n^{4+v} = o(n\epsilon_n^2)$. Then the posterior contracts at the rate ϵ_n.
(d) *Logistic binary regression:* $Y|X \sim \mathrm{Bin}(1, (1+e^{-X^\top\beta})^{-1})$. Assume (ii), (iv) and $\bar{r}_n^v = o(n\epsilon_n^2)$. Then the posterior contracts at the rate ϵ_n.
(e) *Probit binary regression:* $Y|X \sim \mathrm{Bin}(1, \Phi(X^\top\beta))$. Assume (ii), (iv) and $\bar{r}_n^v = o(n\epsilon_n^2)$. Then the posterior contracts at the rate ϵ_n.

9.25 (Jiang 2007) This example shows that in high dimensional problems, even consistency for the posterior distribution of p may not hold without the variable selection step in the prior. Adopt the notations in Problem 9.23. Let Z_i be i.i.d. with $\mathrm{P}(Z = j/K) = K^{-1}$, $j = 1, \ldots, K$. Let $X_{ij} = \mathbb{1}\{Z_i = j/K\}$, $K > n$. Let $Y_i|X_i \sim \mathrm{Nor}(X_i^\top\beta, 1)$. Consider a prior $\beta \sim \mathrm{Nor}_K(0, I_K)$. Let $\beta_0 = 0$. Show that $\Pi(d_H(p, p_0) \geq \sqrt{\eta}|X_1, Y_1, \ldots, X_n, Y_n) \geq 1 - \eta^{-2}n^{-1}$ a.s. for $\eta = \frac{1}{2} - \frac{1}{\sqrt{5}}$.