10

Adaptation and Model Selection

Many nonparametric priors possess a given regularity or complexity, and are appropriate when the true parameter of the data is of similar regularity. In this chapter we investigate general constructions that combine priors of varying regularity into a mixture with the aim of achieving optimal contraction rates for a variety of regularity levels simultaneously. Mixture priors arise naturally from placing a prior on a hyperparameter that expresses regularity or complexity. In this chapter we obtain parallels of the main results of Chapter 8 for mixture priors, for fairly general priors on the hyperparameters. Examples include finite-dimensional models, sieves, the white noise model, log-spline densities and random series with a random number of terms. Adaptation is connected, but certainly does not require consistent selection of the complexity parameter of a model, as is explored in a final section of the chapter. The general message of the chapter is that adaptation is easily achieved, in particular if one is satisfied with optimal rates up to logarithmic factors. To some extent this conclusion also emanates from the adaptivity of special priors, such as Dirichlet process mixtures or Gaussian processes, which is explored by direct methods in the respective chapters.

10.1 Introduction

Many nonparametric procedures involve a *bandwidth* or *regularization* parameter, which adapts the procedure to the "regularity level" or "smoothness" of the function being estimated. This parameter is typically not known, and within the Bayesian framework it is natural to put a prior on it and let the data decide on a correct value through the corresponding posterior distribution. The resulting procedure then combines suitable priors Π_α on statistical models of regularities α, with a prior λ on the regularity α itself, resulting in the "overall" mixture prior $\int \Pi_\alpha \, d\lambda(\alpha)$. The parameter α may concretely refer to a smoothness level (e.g. the number of derivatives of a Gaussian process), but may also be more abstract (e.g. the number of terms in a series expansion, the bandwidth of a kernel, or the scale of a stochastic process prior), as long as the collection of priors Π_α for varying α covers (or can approach) all targeted true densities.

Such a hierarchical Bayesian procedure fits naturally within the framework of *adaptive estimation*, which focuses on constructing estimators that are rate-optimal simultaneously across a scale of models. Informally, given a collection of models, an estimator is said to be *rate-adaptive* if it attains at any given true parameter in the union of all models the rate of contraction that would have been attained had only the best model been used. For instance, the minimax rate for estimating a probability density or regression function on $[0, 1]^d$ that is known to have α derivatives (is "α-smooth"), is $n^{-\alpha/(2\alpha+d)}$, where n is the number of

observations. For smoother functions, faster rates of estimation are possible, with the rate $n^{-\alpha/(2\alpha+d)}$ approaching the "parametric rate" $n^{-1/2}$ as $\alpha \to \infty$. An estimator would be rate-adaptive in this context if its construction is free of α and it attains the rate $n^{-\alpha/(2\alpha+d)}$ whenever the true density is α-smooth, for any $\alpha > 0$. An adaptive estimator possesses an *oracle property* in that it performs as if it were an "oracle" in possession of a priori knowledge on the smoothness of the true function.

In the Bayesian context we may investigate whether the contraction rate of a posterior distribution adapts to the regularity of a true parameter. In an abstract setting we have parameter sets $\Theta_{n,\alpha}$ indexed by a parameter α ranging over a set A, and assume that the true parameter belongs to the union $\Theta_n = \cup_\alpha \Theta_{n,\alpha}$. We are given an observation $X^{(n)}$ in statistical experiments $(\mathfrak{X}^{(n)}, \mathscr{X}^{(n)}, P_\theta^{(n)}: \theta \in \Theta_n)$, and form a posterior distribution $\Pi_n(\cdot \mid X^{(n)})$ based on a prior Π_n on Θ_n. Furthermore, for each model $\Theta_{n,\alpha}$ we have a targeted contraction rate $\epsilon_{n,\alpha}$ relative to some semimetric d_n on Θ_n.

Definition 10.1 (Adaptation) The prior Π_n is *rate-adaptive* (or adaptive, in short) if for every $\alpha \in A$ and every $M_n \to \infty$,

$$\sup_{\theta_0 \in \Theta_{n,\alpha}} P_{\theta_0}^{(n)} \Pi_n(\theta: d_n(\theta, \theta_0) \geq M_n \epsilon_{n,\alpha} \mid X^{(n)}) \to 0.$$

Without specifying the rates $\epsilon_{n,\alpha}$, the definition can merely serve as a template for the desired results. One typical specialization would be to a mixture prior $\Pi_n = \int \Pi_{n,\alpha}\, d\lambda(\alpha)$, with $\Pi_{n,\alpha}$ prior distributions on the models $\Theta_{n,\alpha}$, and $\epsilon_{n,\alpha}$ the contraction rates of the posterior distributions $\Pi_{n,\alpha}(\cdot \mid X^{(n)})$ induced by $\Pi_{n,\alpha}$, at parameters in the model $\Theta_{n,\alpha}$. It is natural to require the rate of contraction to be uniform over true parameters θ_0 in a model, as in the definition, but for simplicity we often formulate pointwise results.

The examples in Chapters 9 and 11 may guide the choice of the priors $\Pi_{n,\alpha}$, given a collection of models $\Theta_{n,\alpha}$. The mixing weights λ are a new element. As the models $\Theta_{n,\alpha}$ may have very different complexities, weighting them equally may not be a good choice. In this chapter we shall see that there may be a wide range of choices for λ. In particular, if we are willing to add a logarithmic factor to the rates $\epsilon_{n,\alpha}$ for a given model, then simple, universal choices may do.

The posterior distribution induced by the mixture $\Pi_n = \int \Pi_{n,\alpha}\, d\lambda(\alpha)$ is itself a mixture, given by

$$\Pi_n(\theta \in B \mid X^{(n)}) = \int \Pi_{n,\alpha}(\theta \in B \mid X^{(n)})\, d\lambda(\alpha \mid X^{(n)}), \tag{10.1}$$

with $\Pi_{n,\alpha}(\cdot \mid X^{(n)})$ the posterior distribution when using the prior $\Pi_{n,a}$ and $\lambda(\cdot \mid X^{(n)})$ the posterior distribution of α. Adaptation in this setting would easily be seen to occur if the latter weights concentrate on a "true" index α. However, the notion of a true smoothness or regularity level is generally misguided. For a given true parameter θ_0, multiple models $\Theta_{n,\alpha}$ may contain parameters that approximate θ_0 closely, and adaptation will occur when the posterior distribution concentrates on these parameters, without the posterior distribution of α settling down on specific values. In Section 10.5 we further investigate the posterior distribution of the model index.

The main interest in the present chapter is to exhibit general choices of the weights λ. Specific priors may come with natural choices of hyperpriors on the bandwidth parameter. In particular, with Dirichlet process mixtures of normal kernel, an inverse-gamma prior on the scale of the normal kernel is natural, and for Gaussian process priors changing the length scale is customary. These adaptation schemes are described elsewhere together with their general constructions (see e.g. Sections 9.4 and 11.6).

The following result shows that a rate-adaptive prior induces rate-adaptive Bayesian point estimators. The theorem is an immediate corollary of Theorem 8.7.

Theorem 10.2 (Point estimator) *If the prior Π_n is rate-adaptive in the sense of Definition 10.1, then the center $\hat{\theta}_n$ of the (nearly) smallest d_n-ball that contains posterior mass at least $1/2$ satisfies $d_n(\hat{\theta}_n, \theta_{n,0}) = O_P(\epsilon_{n,\alpha})$ in $P_{\theta_{n,0}}^{(n)}$-probability, for any $\theta_{n,0} \in \Theta_{n,\alpha}$, for every α.*

10.2 Independent Identically Distributed Observations

In this section we obtain adaptive versions of the posterior contraction results for i.i.d. observations in Section 8.2. We take the parameter equal to the probability density p of the observations $X_1, \ldots, X_n$, and consider a countable collection of models $\mathcal{P}_{n,\alpha}$ for this density, indexed by a parameter α ranging through a set A_n, which may depend on n. The index α may refer to the regularity of the elements of $\mathcal{P}_{n,\alpha}$, but in general may be arbitrary. Thus $\mathcal{P}_{n,\alpha}$ is an arbitrary set of probability densities with respect to a σ-finite measure ν on the sample space $(\mathfrak{X}, \mathscr{X})$ of the observations, for every α in the arbitrary countable set A_n.

Consider a metric d on the set of all ν-probability densities for which tests as in (8.2) exist. For instance, d may be the Hellinger or $\mathbb{L}_1$-distance, or the $\mathbb{L}_2$-distance if the densities are uniformly bounded.

Let $\Pi_{n,\alpha}$ be a probability measure on $\mathcal{P}_{n,\alpha}$, and $\lambda_n = (\lambda_{n,\alpha}: \alpha \in A_n)$ a probability measure on A_n, viewed as prior distributions for p within the model $\mathcal{P}_{n,\alpha}$ and for the index α, respectively, at stage n. Thus the overall prior is a probability measure on the set of probability densities, given by

$$\Pi_n = \sum_{\alpha \in A_n} \lambda_{n,\alpha} \Pi_{n,\alpha}.$$

The corresponding posterior distribution is given by Bayes's rule as

$$\begin{aligned} \Pi_n(B \mid X_1, \ldots, X_n) &= \frac{\int_B \prod_{i=1}^n p(X_i)\, d\Pi_n(p)}{\int \prod_{i=1}^n p(X_i)\, d\Pi_n(p)} \\ &= \frac{\sum_{\alpha \in A_n} \lambda_{n,\alpha} \int_{p \in \mathcal{P}_{n,\alpha}: p \in B} \prod_{i=1}^n p(X_i)\, d\Pi_{n,\alpha}(p)}{\sum_{\alpha \in A_n} \lambda_{n,\alpha} \int_{p \in \mathcal{P}_{n,\alpha}} \prod_{i=1}^n p(X_i)\, d\Pi_{n,\alpha}(p)}. \end{aligned} \tag{10.2}$$

Here we have implicitly assumed the existence of a suitable measurability structure on $\mathcal{P}_{n,\alpha}$, for which the priors $\Pi_{n,\alpha}$ are defined and the maps $(x, p) \mapsto p(x)$ are jointly measurable.

The true density p_0 of the observations need not belong to any of the models $\mathcal{P}_{n,\alpha}$. However, we denote by β_n parameters in A_n that should be thought of as giving "best" models $\mathcal{P}_{n,\beta_n}$. The purpose is to state conditions under which the mixture posterior distribution

(10.2) attains the same contraction rate as obtained with the single prior Π_{n,β_n}, whenever the true density p_0 belongs to $\mathcal{P}_{n,\beta_n}$. Then the hierarchical Bayesian procedure would automatically adapt to the set of models $\mathcal{P}_{n,\alpha}$ in that it performs at par with an oracle that uses the knowledge of the "correct model" $\mathcal{P}_{n,\beta_n}$.

Similarly as in (8.3), define

$$B_{n,\alpha}(p_0,\epsilon) = \{p \in \mathcal{P}_{n,\alpha}: K(p_0; p) \le \epsilon^2, V_2(p_0; p) \le \epsilon^2\}, \tag{10.3}$$

$$C_{n,\alpha}(p_0,\epsilon) = \{p \in \mathcal{P}_{n,\alpha}: d(p_0, p) \le \epsilon\}. \tag{10.4}$$

Let $\epsilon_{n,\alpha}$ be given positive sequences of numbers tending to zero, to be thought of as the rate attached to the model $\mathcal{P}_{n,\alpha}$ if this is (approximately) correct.

For $\beta_n \in A_n$, thought of as the index of a best model for a given p_0, and a fixed constant $H \ge 1$, decompose A_n into

$$A_{n,\succeq\beta_n} = \{\alpha \in A_n: \epsilon_{n,\alpha}^2 \le H\epsilon_{n,\beta_n}^2\},$$
$$A_{n,\prec\beta_n} = \{\alpha \in A_n: \epsilon_{n,\alpha}^2 > H\epsilon_{n,\beta_n}^2\}.$$

These two sets are thought of as indices of the collections of models consisting of densities that are "more regular" or "less regular" than the given true density p_0. Thus, as models they are less complex (or lower dimensional) or more complex, respectively. Even though we do not assume that A_n is ordered, we shall write $\alpha \succeq \beta_n$ and $\alpha \prec \beta_n$ if α belongs to the sets $A_{n,\succeq\beta_n}$ or $A_{n,\prec\beta_n}$, respectively. The set $A_{n,\succeq\beta_n}$ contains β_n and hence is never empty, but the set $A_{n,\prec\beta_n}$ can be empty (if β_n is the "smallest" possible index). In the latter case, conditions involving $\alpha \prec \beta_n$ are understood to be automatically satisfied.

The assumptions of the following theorem are analogous to those in Theorem 8.9 on posterior contraction rate in a single model. They entail a bound on the complexity of the models and a condition on the concentration of the priors. The complexity bound is exactly as in Section 8.2, namely, for some constants E_α,

$$\sup_{\epsilon \ge \epsilon_{n,\alpha}} \log N\Big(\epsilon/3, C_{n,\alpha}(2\epsilon), d\Big) \le E_\alpha n\epsilon_{n,\alpha}^2, \qquad \alpha \in A_n. \tag{10.5}$$

The conditions on the priors involve comparisons of the prior masses of balls of various sizes in various models. These conditions are split in conditions on the models that are smaller or bigger than the best model: for given constants $\mu_{n,\alpha}$, L, H, I,

$$\frac{\lambda_{n,\alpha}}{\lambda_{n,\beta_n}} \frac{\Pi_{n,\alpha}(C_{n,\alpha}(p_0, i\epsilon_{n,\alpha}))}{\Pi_{n,\beta_n}(B_{n,\beta_n}(p_0,\epsilon_{n,\beta_n}))} \le \mu_{n,\alpha} e^{Li^2 n\epsilon_{n,\alpha}^2}, \qquad \alpha \prec \beta_n, \quad i \ge I, \tag{10.6}$$

$$\frac{\lambda_{n,\alpha}}{\lambda_{n,\beta_n}} \frac{\Pi_{n,\alpha}(C_{n,\alpha}(p_0, i\epsilon_{n,\beta_n}))}{\Pi_{n,\beta_n}(B_{n,\beta_n}(p_0,\epsilon_{n,\beta_n}))} \le \mu_{n,\alpha} e^{Li^2 n\epsilon_{n,\beta_n}^2}, \qquad \alpha \succeq \beta_n, \quad i \ge I. \tag{10.7}$$

A final condition requires that the prior mass in a ball of radius of the order $\epsilon_{n,\alpha}$ in a bigger model (i.e., smaller α) is significantly smaller than in a small model: for a constant $M > I$,

$$\sum_{\alpha \prec \beta_n} \frac{\lambda_{n,\alpha}}{\lambda_{n,\beta_n}} \frac{\Pi_{n,\alpha}(C_{n,\alpha}(p_0, M\epsilon_{n,\alpha}))}{\Pi_{n,\beta_n}(B_{n,\beta_n}(p_0,\epsilon_{n,\beta_n}))} = o(e^{-2n\epsilon_{n,\beta_n}^2}). \tag{10.8}$$

Let K be the universal testing constant appearing in (8.2), and assume that there exists a finite constant (with an empty supremum defined as 0) such that

$$E \geq \sup_{\alpha \prec \beta_n} E_\alpha \vee \sup_{\alpha \succeq \beta_n} E_\alpha \frac{\epsilon_{n,\alpha}^2}{\epsilon_{n,\beta_n}^2}.$$

Theorem 10.3 (Adaptive contraction rates) *Assume that* (10.5)–(10.8) *hold for constants* $\mu_{n,\alpha} > 0$, $E_\alpha > 0$, $L > 0$, $H \geq 1$, *and* $M > I > 2$ *such that* $KM^2/I^2 > E + 1$, *and* $M^2 H(K - 2L) > 3$, *and* $\sum_{\alpha \in A_n} \sqrt{\mu_{n,\alpha}} \leq e^{n\epsilon_{n,\beta_n}^2}$. *If* $\beta_n \in A_n$ *for every* n *and satisfies* $n\epsilon_{n,\beta_n}^2 \to \infty$, *then* $P_0^n \Pi_n(p: d(p, p_0) \geq M\sqrt{H}\,\epsilon_{n,\beta_n} | X_1, \cdots, X_n) \to 0$.

Remark 10.4 In view of Theorem 8.20, the entropy condition (10.5) can, as usual, be relaxed to the same condition on submodels $\mathcal{P}'_{n,\alpha} \subset \mathcal{P}_{n,\alpha}$ that carry most of the prior mass, in the sense that

$$\frac{\sum_\alpha \lambda_{n,\alpha} \Pi_{n,\alpha}(\mathcal{P}_{n,\alpha} \setminus \mathcal{P}'_{n,\alpha})}{\lambda_{n,\beta_n} \Pi_{n,\beta_n}(B_{n,\beta_n}(p_0, \epsilon_{n,\beta_n}))} = o(e^{-2n\epsilon_{n,\beta_n}^2}).$$

Proof of Theorem 10.3 Abbreviate $J_{n,\alpha} = n\epsilon_{n,\alpha}^2$, so that $E \geq \sup\{E_\alpha J_{n,\alpha}/J_{n,\beta_n} : \alpha \succeq \beta_n\}$. For $\alpha \succeq \beta_n$ we have $B\epsilon_{n,\beta_n} \geq B/\sqrt{H}\epsilon_{n,\alpha} \geq \epsilon_{n,\alpha}$, whence, in view of the entropy bound (10.5) and Lemma D.3 with $\epsilon = B\epsilon_{n,\beta_n}$ and $\log N(\epsilon) = E_\alpha J_{n,\alpha}$ (constant in ϵ), there exists for every $\alpha \succeq \beta_n$ a test $\phi_{n,\alpha}$ with,

$$\begin{aligned} P_0^n \phi_{n,\alpha} &\leq \sqrt{\mu_{n,\alpha}} \frac{e^{E_\alpha J_{n,\alpha} - KB^2 J_{n,\beta_n}}}{1 - e^{-KB^2 J_{n,\beta_n}}} \lesssim \sqrt{\mu_{n,\alpha}} e^{(E - KB^2) J_{n,\beta_n}}, \\ \sup_{p \in \mathcal{P}_{n,\alpha}: d(p,p_0) \geq iB\epsilon_{n,\beta_n}} P^n(1 - \phi_{n,\alpha}) &\leq \frac{1}{\sqrt{\mu_{n,\alpha}}} e^{-KB^2 i^2 J_{n,\beta_n}}. \end{aligned} \tag{10.9}$$

For $\alpha \prec \beta_n$ we have $\epsilon_{n,\alpha} > \sqrt{H}\epsilon_{n,\beta_n}$ and we cannot similarly test balls of radius proportional to ϵ_{n,β_n} in $\mathcal{P}_{n,\alpha}$. However, Lemma D.3 with $\epsilon = B'\epsilon_{n,\alpha}$ and $B' := B/\sqrt{H} > 1$, still gives tests $\phi_{n,\alpha}$ such that for every $i \in \mathbb{N}$,

$$\begin{aligned} P_0^n \phi_{n,\alpha} &\leq \sqrt{\mu_{n,\alpha}} \frac{e^{(E_\alpha - KB'^2) J_{n,\alpha}}}{1 - e^{-KB'^2 J_{n,\alpha}}} \lesssim \sqrt{\mu_{n,\alpha}} e^{(E - KB'^2) J_{n,\alpha}}, \\ \sup_{p \in \mathcal{P}_{n,\alpha}: d(p,p_0) > iB'\epsilon_{n,\alpha}} P^n(1 - \phi_{n,\alpha}) &\leq \frac{1}{\sqrt{\mu_{n,\alpha}}} e^{-KB'^2 i^2 J_{n,\alpha}}. \end{aligned} \tag{10.10}$$

Let $\phi_n = \sup\{\phi_{n,\alpha} : \alpha \in A_n\}$ be the supremum of all tests so constructed.

The test ϕ_n is more powerful than all the tests $\phi_{n,\alpha}$, and has error of the first kind $P_0^n \phi_n$ bounded by $P_0^n \sum_{\alpha \in A_n} \phi_{n,\alpha} \lesssim \sum_{\alpha \in A_n} \sqrt{\mu_{n,\alpha}} e^{-cJ_{n,\beta_n}}$, for c equal to the minimum of $KB^2 - E$ and $KB^2 - EH$. This tends to zero, because the sum is bounded by $e^{(1-c)J_{n,\beta_n}}$, by assumption, and $c > 1$ and $J_{n,\beta_n} \to \infty$, by assumption. Consequently, for any M we have that $P_0^n\big[\Pi_n(p: d(p, p_0) > M\sqrt{H}\epsilon_{n,\beta_n} | X_1, \ldots, X_n)\phi_n\big] \leq P_0^n \phi_n \to 0$. We shall complement this with an analysis of the posterior multiplied by $1 - \phi_n$.

For $i \in \mathbb{N}$, define

$$\mathcal{S}_{n,\alpha,i} = \{p \in \mathcal{P}_{n,\alpha}: iB'\epsilon_{n,\alpha} < d(p, p_0) \le (i+1)B'\epsilon_{n,\alpha}\}, \qquad \alpha \prec \beta_n,$$
$$\mathcal{S}_{n,\alpha,i} = \{p \in \mathcal{P}_{n,\alpha}: iB\epsilon_{n,\beta_n} < d(p, p_0) \le (i+1)B\epsilon_{n,\beta_n}\}, \qquad \alpha \succeq \beta_n.$$

Then the set $\{p: d(p, p_0) > IB\epsilon_{n,\beta_n}\}$ is contained in the union of the set $\cup_\alpha \cup_{i\ge I} \mathcal{S}_{n,\alpha,i}$ and the set $\cup_{\alpha\prec\beta_n} C_{n,\alpha}(p_0, IB'\epsilon_{n,\alpha})$. We shall estimate the posterior mass in the sets of the first union with the help of the tests, and handle the sets in the second union using their prior mass only.

We estimate numerator and denominator of the posterior distribution separately with the help of the inequalities, for any set C and suitable events A_n,

$$P_0^n \int_C \prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)}(1-\phi_n)\, d\Pi_{n,\alpha}(p) \le \sup_{p\in C} P^n(1-\phi_n)\, \Pi_{n,\alpha}(C),$$
$$\int \prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)} d\Pi_n(p) \mathbb{1}_{A_n} \ge e^{-2J_{n,\beta_n}} \lambda_{n,\beta_n} \Pi_{n,\beta_n}(B_{n,\beta_n}(p_0, \epsilon_{n,\beta_n})).$$

The first inequality follows by Fubini's theorem. Because $\Pi_n \ge \lambda_{n,\beta_n}\Pi_{n,\beta_n}$, the second is true on events A_n such that $P_0^n(A_n) \ge 1 - (n\epsilon_{n,\beta_n}^2)^{-1} \to 1$, by Lemma 8.10. Combining these two inequalities, and also the first inequality but with $\phi_n \equiv 0$, with (10.9) and (10.10), we see that

$$\begin{aligned} P_0^n \Big[\Pi_n\Big(d(p, p_0) > IB\epsilon_{n,\beta_n} \,|\, X_1, \ldots, X_n\Big)(1-\phi_n)\mathbb{1}_{A_n}\Big] \\ \le \sum_{\alpha\in A_n:\alpha\succeq\beta_n} \sum_{i\ge I} \frac{\lambda_{n,\alpha}}{\lambda_{n,\beta_n}} \frac{e^{-KB^2i^2J_{n,\beta_n}} \Pi_{n,\alpha}(\mathcal{S}_{n,\alpha,i})}{e^{-2J_{n,\beta_n}} \Pi_{n,\beta_n}(B_{n,\beta_n}(p_0,\epsilon_{n,\beta_n}))} \frac{1}{\sqrt{\mu_{n,\alpha}}} \\ + \sum_{\alpha\in A_n:\alpha\prec\beta_n} \sum_{i\ge I} \frac{\lambda_{n,\alpha}}{\lambda_{n,\beta_n}} \frac{e^{-KB'^2i^2J_{n,\alpha}} \Pi_{n,\alpha}(\mathcal{S}_{n,\alpha,i})}{e^{-2J_{n,\beta_n}} \Pi_{n,\beta_n}(B_{n,\beta_n}(p_0,\epsilon_{n,\beta_n}))} \frac{1}{\sqrt{\mu_{n,\alpha}}} \\ + \sum_{\alpha\in A_n:\alpha\prec\beta_n} \frac{\lambda_{n,\alpha}}{\lambda_{n,\beta_n}} \frac{\Pi_{n,\alpha}(C_{n,\alpha}(p_0, IB'\epsilon_{n,\alpha}))}{e^{-2J_{n,\beta_n}} \Pi_{n,\beta_n}(B_{n,\beta_n}(p_0,\epsilon_{n,\beta_n}))}. \end{aligned} \tag{10.11}$$

The third term on the right-hand side tends to zero by assumption (10.8) if M is defined as $M = IB' = IB/\sqrt{H}$. Because for $\alpha \succeq \beta_n$ and for $i \ge I \ge 3$, we have $\mathcal{S}_{n,\alpha,i} \subset C_{n,\alpha}(p_0, \sqrt{2}iB\epsilon_{n,\beta_n})$, the assumption (10.7) shows that the first term is bounded by

$$\sum_{\alpha\in A_n:\alpha\succeq\beta_n} \sum_{i\ge I} \frac{\lambda_{n,\alpha}}{\lambda_{n,\beta_n}} \frac{e^{-KB^2i^2J_{n,\beta_n}} \Pi_{n,\alpha}(C_{n,\alpha}(p_0, \sqrt{2}iB\epsilon_{n,\beta_n}))}{e^{-2J_{n,\beta_n}} \Pi_{n,\beta_n}(B_{n,\beta_n}(p_0,\epsilon_{n,\beta_n}))} \frac{1}{\sqrt{\mu_{n,\alpha}}}$$
$$\le \sum_{\alpha\in A_n:\alpha\succeq\beta_n} \sqrt{\mu_{n,\alpha}} e^{2J_{n,\beta_n}} \sum_{i\ge I} e^{(2L-K)B^2J_{n,\beta_n}i^2} \le \sum_{\alpha\in A_n:\alpha\succeq\beta_n} \frac{e^{(2L-K)B^2J_{n,\beta_n}I^2}}{1 - e^{(2L-K)B^2J_{n,\beta_n}}}.$$

This tends to zero if $(K-2L)I^2B^2 > 3$, which is ensured by the assumption that $M^2H(K-2L) > 3$. Similarly, for $\alpha \prec \beta_n$ the second term is bounded by, in view of (10.6),

$$\sum_{\alpha\in A_n:\alpha\prec\beta_n}\sum_{i\geq I}\frac{\lambda_{n,\alpha}}{\lambda_{n,\beta_n}}\frac{e^{-KB'^2i^2J_{n,\alpha}}\Pi_{n,\alpha}(C_{n,\alpha}(p_0,\sqrt{2}iB'\epsilon_{n,\alpha}))}{e^{-2J_{n,\beta_n}}\Pi_{n,\beta_n}(B_{n,\beta_n}(p_0,\epsilon_{n,\beta_n}))}\frac{1}{\sqrt{\mu_{n,\alpha}}}$$

$$\leq \sum_{\alpha\in A_n:\alpha\prec\beta_n}\sqrt{\mu_{n,\alpha}}e^{2J_{n,\beta_n}}\sum_{i\geq I}e^{(2L-K)B'^2i^2J_{n,\alpha}} \leq \sum_{\alpha\in A_n:\alpha\prec\beta_n}\frac{e^{(2L-K)B'^2J_{n,\alpha}I^2}}{1-e^{(2L-K)B'^2J_{n,\alpha}}}.$$

Here $J_{n,\alpha} > HJ_{n,\beta_n}$ for every $\alpha \prec \beta_n$, and hence this tends to zero, again because $(K-2L)B^2I^2 > 3$. □

The conditions of Theorem 10.3 look complicated, but the theorem should be interpreted as saying that adaptation is easy. It appears that even the model weights $\lambda_{n,\alpha}$ need not be very specific. To see this, we simplify the theorem along the same lines as the posterior contraction theorems in Section 8.2.

In analogy to the "crude" prior mass condition (8.4), consider the lower bound

$$\Pi_{n,\beta_n}(B_{n,\beta_n}(p_0,\epsilon_{n,\beta_n})) \geq e^{-Fn\epsilon_{n,\beta_n}^2}. \tag{10.12}$$

Combined with the trivial bound $\Pi_{n,\alpha}(C) \leq 1$, for any set C, we see that (10.6) and (10.7) hold (for sufficiently large I) if, for all $\alpha \in A_n$,

$$\frac{\lambda_{n,\alpha}}{\lambda_{n,\beta_n}} \leq \mu_{n,\alpha}e^{n(H^{-1}\epsilon_{n,\alpha}^2\vee\epsilon_{n,\beta_n}^2)}. \tag{10.13}$$

As the right side tends to infinity fast, this does not seem restrictive. Similarly a sufficient condition for (10.8) is that

$$\sum_{\alpha\in A_n:\alpha\prec\beta_n}\frac{\lambda_{n,\alpha}}{\lambda_{n,\beta_n}}\Pi_{n,\alpha}(C_{n,\alpha}(p_0,M\epsilon_{n,\alpha})) = o(e^{-(F+2)n\epsilon_{n,\beta_n}^2}). \tag{10.14}$$

This condition is more involved, but ought also to be true for fairly general model weights, as the prior probabilities $\Pi_{n,\alpha}(C_{n,\alpha}(p_0,M\epsilon_{n,\alpha}))$ ought to be small. In particular, if a reverse of the bound (10.12) holds, for α instead of β_n, then these prior probabilities would be of the order $e^{-F_0n\epsilon_{n,\alpha}^2}$. Since $\epsilon_{n,\alpha}^2 \geq H\epsilon_{n,\beta_n}^2$ for $\alpha \prec \beta_n$, this would easily give the condition, for sufficiently large H. These observations are summarized in the following corollary.

Corollary 10.5 (Adaptive contraction rates) *Assume that* (10.5) *and* (10.12)–(10.14) *hold for constants* $\mu_{n,\alpha} > 0$, $E_\alpha > 0$, $F > 0$, *and* $H \geq 1$ *such that* $K^2M^2 > 3(1+F)(K+E+1)$, *and* $KM^2H > 9$, *and* $\sum_{\alpha\in A_n}\sqrt{\mu_{n,\alpha}} \leq e^{n\epsilon_{n,\beta_n}^2}$. *If* $\beta_n \in A_n$ *for every* n *and satisfies* $n\epsilon_{n,\beta_n}^2 \to \infty$, *then* $P_0^n\Pi_n(p\colon d(p,p_0) \geq M\sqrt{H}\,\epsilon_{n,\beta_n}\,|\,X_1,\cdots,X_n) \to 0$.

Proof Choose $L = K/3$, so that $K-2L = K/3$. Then conditions (10.6)–(10.7) are satisfied for $i^2 \geq I^2 := 3(1+F)$, in view of (10.12) and (10.13). The conditions on the constants in Theorem 10.3 now translate in the conditions as given. □

10.2.1 Universal Weights

Even though the prior probabilities in (10.14) will typically be small, the condition may also be forced by the model weights $\lambda_{n,\alpha}$, for *any* priors Π_α. For given nonnegative numbers (λ_α), consider the "universal weights"

$$\lambda_{n,\alpha} = \frac{\lambda_\alpha e^{-Cn\epsilon_{n,\alpha}^2}}{\sum_{\gamma\in A_n} \lambda_\gamma e^{-Cn\epsilon_{n,\gamma}^2}}. \tag{10.15}$$

These weights put more weight on less complex models (with small $\epsilon_{n,\alpha}$), and thus down-weight the bigger models (with bigger $\epsilon_{n,\alpha}$).

Corollary 10.6 (Adaptive contraction rates, universal weights) *Assume that* (10.5) *and* (10.12) *hold for positive constants E_α and F such that and $\sum_\alpha (\lambda_\alpha/\lambda_\beta) e^{-Cn\epsilon_{n,\alpha}^2/4} = O(1)$ and $\sup_\alpha E_\alpha < \infty$. If $\beta_n \in A_n$ eventually and satisfies $n\epsilon_{n,\beta_n}^2 \to \infty$, then the posterior distribution based on the weights* (10.15) *attains contraction rate ϵ_{n,β_n} at p_0.*

Proof We have $\lambda_{n,\alpha}/\lambda_{n,\beta} = (\lambda_\alpha/\lambda_\beta) e^{-C(n\epsilon_{n,\alpha}^2 - n\epsilon_{n,\beta}^2)}$ by the definition (10.15) of the weights. Combining this with (10.12) we see that

$$\begin{aligned}\frac{\lambda_{n,\alpha}}{\lambda_{n,\beta_n}} \frac{1}{\Pi_{n,\beta_n}(B_{n,\beta_n}(p_0,\epsilon_{n,\beta_n}))} &\le \frac{\lambda_\alpha}{\lambda_{\beta_n}} e^{-Cn\epsilon_{n,\alpha}^2 + (F+C)n\epsilon_{n,\beta_n}^2} \\ &\le \frac{\lambda_\alpha}{\lambda_\beta} e^{-Cn\epsilon_{n,\alpha}^2/2} \begin{cases} e^{(-C/2+(F+C)/H)n\epsilon_{n,\alpha}^2}, & \text{if } \alpha \prec \beta_n, \\ e^{(-CH/2+F+C)n\epsilon_{n,\beta_n}^2}, & \text{if } \alpha \succeq \beta_n. \end{cases}\end{aligned}$$

This is a bound on the left sides of both (10.6) and (10.7), and on the terms of the sum in (10.8). The first two inequalities are valid, with $\mu_{n,\alpha} = (\lambda_\alpha/\lambda_\beta) e^{-Cn\epsilon_{n,\alpha}^2}$, for $i^2 \ge F + C$. Condition (10.8) is seen to be satisfied if H is chosen sufficiently large that $-C/2 + (F + C + 2)/H < 0$. □

10.2.2 Parametric Rate

Theorem 10.3 excludes the case that ϵ_{n,β_n} is equal to the "parametric rate" $n^{-1/2}$. To cover this case the statement of the theorem must be slightly modified.

Theorem 10.7 (Adaptive contraction, parametric rate) *Assume there exist constants $B > 0$, $E_\alpha > 0$, $0 < L < K/2$, $H \ge 1$ and $I > 0$ such that* (10.5)–(10.8) *hold for every sufficiently large I. Furthermore, assume that $\sum_{\alpha\in A_n} \sqrt{\mu_{n,\alpha}} = O(1)$. If $\beta_n \in A_n$ for every n and $\epsilon_{n,\beta_n} = n^{-1/2}$, then for every $I_n \to \infty$, $P_0^n \Pi_n(p\colon d(p, p_0) \ge I_n \epsilon_{n,\beta_n} \mid X_1, \cdots, X_n) \to 0$.*

Proof We follow the line of argument of the proof of Theorem 10.3, the main difference being that presently $J_{n,\beta_n} = 1$ and hence does not tend to infinity. To make sure that $P_0^n \phi_n$ is small, we choose the constant B sufficiently large, and to make $P_0^n(A_n)$ sufficiently large we apply Lemma 8.10 with C a large constant instead of $C = 1$. This gives a factor $e^{-(1+C)J_{n,\beta_n}}$

instead of $e^{-2J_{n,\beta_n}}$ in the denominators of (10.11), but this is fixed for fixed C. The arguments then show that for an event A_n with probability arbitrarily close to 1 the expectation $P_0^n[\Pi_n(d(p, p_0) > IB\epsilon_{n,\beta_n} | X_1, \ldots, X_n)\mathbb{1}_{A_n}]$ can be made arbitrarily small by choosing sufficiently large I and B. □

10.2.3 Two Models

It is instructive to apply the theorems to the situation of two models, say $\mathcal{P}_{n,1}$ and $\mathcal{P}_{n,2}$ with rates $\epsilon_{n,1} > \epsilon_{n,2}$. For simplicity we shall also assume (10.12) and use universal constants.

Corollary 10.8 *Assume that* (10.5) *holds for* $\alpha \in A_n = \{1, 2\}$ *and sequences* $\epsilon_{n,1} > \epsilon_{n,2}$ *such that* $n\epsilon_{n,2}^2 \to \infty$.

(i) *If* $\Pi_{n,1}(B_{n,1}(p_0, \epsilon_{n,1})) \geq e^{-n\epsilon_{n,1}^2}$ *and* $\lambda_{n,2}/\lambda_{n,1} \leq e^{n\epsilon_{n,1}^2}$, *then the posterior contracts at the rate* $\epsilon_{n,1}$.

(ii) *If* $\Pi_{n,1}(B_{n,2}(p_0, \epsilon_{n,2})) \geq e^{-n\epsilon_{n,2}^2}$ *and* $\lambda_{n,2}/\lambda_{n,1} \geq e^{-n\epsilon_{n,1}^2}$, *and for some* M *with* $M^2 \geq 3 \vee (9/K) \vee ((6/K^2 \vee (3/K)(E + 1))$ *we have that* $\Pi_{n,1}(C_{n,1}(p_0, M\epsilon_{n,1})) \leq (\lambda_{n,2}/\lambda_{n,1})o(e^{-3n\epsilon_{n,2}^2})$, *then the posterior contracts at the rate* $\epsilon_{n,2}$.

Proof We apply the preceding theorems with $\beta_n = 1$, $A_{n,\prec\beta_n} = \varnothing$ and $A_{n,\succeq\beta_n} = \{1, 2\}$ in case (i) and $\beta_n = 2$, $A_{n,\prec\beta_n} = \{1\}$ and $A_{n,\succeq\beta_n} = \{2\}$ in case (ii), both times with $H = 1$ and $\mu_{n,1} = \mu_{n,2} = 1$ and $L = K/3$. □

Statement (i) of the corollary gives the slower $\epsilon_{n,1}$ of the two rates under the assumption that the bigger model satisfies the prior mass condition (10.12) and a condition on the weights $\lambda_{n,i}$ that ensures that the smaller model is not overly down-weighted. The latter condition is very mild, as it allows the weights of the two models to be very different. Apart from this, Statement (i) is not surprising, and could also be obtained from Theorem 8.9.

Statement (ii) gives the faster rate $\epsilon_{n,2}$ under the same two conditions, but with the two models swapped, and an additional condition on the prior weight $\Pi_{n,1}(C_{n,1}(p_0, M\epsilon_{n,1}))$ that the bigger model attaches to a neighborhood of the true distribution. If this were of the expected order $e^{-2n\epsilon_{n,1}^2}$ and $\epsilon_{n,1} \gg \epsilon_{n,2}$, then the union of the conditions on the weights $\lambda_{n,1}$ and $\lambda_{n,2}$ in (i) and (ii) could be summarized as

$$e^{-n\epsilon_{n,1}^2} \leq \frac{\lambda_{n,2}}{\lambda_{n,1}} \leq e^{n\epsilon_{n,1}^2}.$$

This is a remarkably big range of weights, showing that Bayesian methods are very robust to the prior specification of model weights.

10.3 Examples

10.3.1 Priors Based on Finite Approximating Sets

Priors based on finite approximating sets were shown in Section 8.2.2 to yield posterior distributions that contract at the rate determined by the (bracketing) entropy approximations.

In this section we extend the construction to multiple models, and combine the resulting priors using the universal weights described in Section 10.2.1.

For each $\alpha \in A_n$ let $\mathcal{Q}_{n,\alpha}$ be a set of nonnegative, integrable functions on the sample space with finite upper bracketing numbers relative to the Hellinger distance d_H (not necessarily probability densities). Let target rates $\epsilon_{n,\alpha}$ satisfy, for every $\alpha \in A_n$,

$$\log N_{]}(\epsilon_{n,\alpha}, \mathcal{Q}_{n,\alpha}, d_H) \lesssim n\epsilon_{n,\alpha}^2.$$

For each α choose a set $\mathcal{U}_{n,\alpha} = \{u_{1,n,\alpha}, \ldots, u_{N_n,n,\alpha}\}$ of $\epsilon_{n,\alpha}$-upper brackets over $\mathcal{Q}_{n,\alpha}$, and let $\Pi_{n,\alpha}$ be the uniform discrete probability measure on the set of re-normalized functions $u/\int u\,d\nu$, for $u \in \mathcal{U}_{n,\alpha}$. The collection $\mathcal{U}_{n,\alpha}$ may be a minimal set of $\epsilon_{n,\alpha}$-upper brackets over $\mathcal{Q}_{n,\alpha}$, but the following theorem only assumes a bound on its cardinality, in agreement with its entropy, and that every bracket intersects the set $\mathcal{Q}_{n,\alpha}$.

We combine the priors $\Pi_{n,\alpha}$ with the universal model weights (10.15). The following theorem shows that the resulting mixture prior is appropriate whenever the true density is contained in the union $\cup_{M>0}(M\mathcal{Q}_{n,\alpha})$, for some α.

Theorem 10.9 (Priors on nets) *Construct the priors $\Pi_{n,\alpha}$ as described, using at most $\exp(En\epsilon_{n,\alpha}^2)$ upper brackets, for some constant E. If there exist $\beta_n \in A_n$ and M_0 such that $p_0 \in M_0\mathcal{Q}_{n,\beta_n}$ eventually and such that $\sum_\alpha (\lambda_\alpha/\lambda_{\beta_n})e^{-Cn\epsilon_{n,\alpha}^2/4} = O(1)$ and $n\epsilon_{n,\beta_n}^2 \to \infty$, then the posterior distributions relative to the model weights* (10.15) *contract to p_0 at the rate ϵ_{n,β_n} relative to the Hellinger distance.*

Proof Because the support $\mathcal{P}_{n,\alpha}$ of $\Pi_{n,\alpha}$ has cardinality at most $\exp(En\epsilon_{n,\alpha}^2)$, by construction, the entropy condition (10.5) with d equal to the Hellinger distance is trivially satisfied. If we can also show that (10.12) holds, then the theorem follows from Corollary 10.6.

Because $p_0/M_0 \in \mathcal{Q}_{n,\beta_n}$, there exists $u_n \in \mathcal{U}_{n,\beta_n}$ such that $p_0/M_0 \le u_n$ and $\|\sqrt{p_0} - \sqrt{M_0u_n}\|_2 \le \sqrt{M_0}\epsilon_{n,\beta_n}$. It follows that $\|\sqrt{M_0u_n}\|_2 \ge \|\sqrt{p_0}\|_2 = 1$, and by the triangle inequality also that $\|\sqrt{M_0u_n}\|_2 \le \sqrt{M_0}\epsilon_{n,\beta_n} + 1$. By construction the function $p_n = u_n/\int u_n d\nu$ belongs to $\mathcal{P}_{n,\beta_n}$. Furthermore, by the triangle inequality,

$$\begin{aligned} d_H(p_0, p_n) &\le d_H(p_0, M_0u_n) + d_H(M_0u_n, p_n) \\ &= d_H(p_0, M_0u_n) + |\|\sqrt{M_0u_n}\|_2 - 1| \le 2\sqrt{M_0}\epsilon_{n,\beta_n}. \end{aligned}$$

The inequality in the second line follows from the fact that $\|r - r/\|r\|\| = |1 - \|r\||$ for every norm and function r, applied with $r = \sqrt{M_0u_n}$. We also have $p_0/p_n \le M_0\int u_n\,d\nu = \|\sqrt{M_0u_n}\|_2^2$, which is bounded by the square of $\sqrt{M_0}\epsilon_{n,\beta_n}$. Therefore, in view of the comparison of Hellinger and Kuillback-Leibler discrepancies given in Lemma B.2, it follows that $p_n \in B_{n,\beta_n}(p_0, D\sqrt{M_0}\epsilon_{n,\beta_n})$ for a sufficiently large constant D, whence

$$\Pi_{n,\beta_n}(B_{n,\beta_n}(p_0, D\sqrt{M_0}\epsilon_{n,\beta_n})) \ge \Pi_{n,\beta_n}(\{p_n\}) \ge \frac{1}{\#\mathcal{P}_{n,\beta_n}} \ge e^{-En\epsilon_{n,\beta_n}^2}.$$

It follows that the prior mass condition (10.12) holds, but for a multiple of ϵ_{n,β_n} instead of ϵ_{n,β_n}. By redefining the rates $\epsilon_{n,\alpha}$, we can still draw the desired conclusion from Corollary 10.6. □

If the base collection $\mathcal{Q}_{n,\alpha}$ is the unit ball in a Banach space, then $\cup_{M>0}(M\mathcal{Q}_{n,\alpha})$ is the full space, and the preceding theorem simply assumes that p_0 is contained in the Banach space. For instance, if the Banach space is defined by a regularity norm, then the assumption is that the true density is "α-regular," for some α, without having to satisfy quantitative regularity bounds.

Example 10.10 (Hölder spaces) Consider for each $\alpha > 0$ a Banach space $\mathbb{B}^\alpha(\mathfrak{X})$ of measurable functions $f: \mathfrak{X} \to \mathbb{R}$ whose unit ball $\mathbb{B}_1^\alpha(\mathfrak{X})$ processes finite upper bracketing numbers relative to the $\mathbb{L}_2(\nu)$-norm. Let the constants E_α and functions $H_\alpha: (0,\infty) \to (0,\infty)$ satisfy

$$\log N_{]}(\epsilon, \mathbb{B}_1^\alpha(\mathfrak{X}), \|\cdot\|_2) \le E_\alpha H_\alpha(\epsilon). \tag{10.16}$$

For $\alpha \in (0,\infty)$, define $\mathcal{Q}_\alpha$ as the set of all nonnegative functions $p: \mathfrak{X} \to \mathbb{R}$ such that the root $\sqrt{p}$ is contained in the unit ball $\mathbb{B}_1^\alpha(\mathfrak{X})$. Because the Hellinger distance on $\mathcal{Q}_\alpha$ corresponds to the $\mathbb{L}_2(\nu)$-distance on the roots of the elements of $\mathcal{Q}_\alpha$, the inequality (10.16) implies that there exists a set of $\epsilon_{n,\alpha}$-brackets of cardinality as in the preceding theorem, for the rates $\epsilon_{n,\alpha}$ satisfying

$$H_\alpha(\epsilon_{n,\alpha}) = n\epsilon_{n,\alpha}^2.$$

The root $\sqrt{p_0}$ of the true density belongs to the Banach space $\mathbb{B}^\beta(\mathfrak{X})$, if and only if $p_0 \in \cup_{M>0}(M\mathcal{Q}_\beta)$. Therefore, the priors chosen as in Theorem 10.9 yield the rate of contraction ϵ_{n,β_n} for any β_n such that $\sqrt{p_0} \in \mathbb{B}^{\beta_n}(\mathfrak{X})$ and $\beta_n \in A_n$ eventually. The prior construction does not use information about the norm of $\sqrt{p_0}$ in $\mathbb{B}^{\beta_n}(\mathfrak{X})$; it suffices that the square root of p_0 be contained in $\mathbb{B}^{\beta_n}(\mathfrak{X})$.

A typical concrete example is a Hölder space $\mathfrak{C}^\alpha([0,1]^d)$ of α-smooth functions $f: [0,1]^d \to \mathbb{R}$. Let By Proposition C.5 the ϵ-entropy of the unit ball of these spaces relative to the uniform norm are of the order $\epsilon^{-d/\alpha}$, whence (10.16) is satisfied with $H_\alpha(\epsilon) = \epsilon^{-d/\alpha}$. This yields posterior contraction at the rate $n^{-\alpha/(2\alpha+d)}$ whenever $\sqrt{p_0} \in \mathfrak{C}^\alpha([0,1]^d)$.

There are many ways of constructing an ϵ-net for the uniform norm over $C_1^\alpha[0,1]^d$, some of which are only of theoretical interest, but others of a constructive nature. Splines of an appropriate degree and dimension are one example.

10.3.2 White Noise Model

In the white noise model, considered previously in Example 8.6 and Section 9.5.4, the observation is (equivalent with) an infinite-dimensional random vector $X^{(n)} = (X_{n,1}, X_{n,2}, \ldots)$, whose coordinates are independent variables $X_{n,i} \overset{\text{ind}}{\sim} \text{Nor}(\theta_i, n^{-1})$, for a parameter vector $\theta = (\theta_1, \theta_2, \ldots)$ belonging to ℓ_2. This fits into the i.i.d. setup by writing the observation as the average of n independent variables distributed as $X^{(1)}$. In the latter model, the average is sufficient and hence the resulting posterior distribution is the same as the posterior distribution given $X^{(n)}$.

Consider the prior Π_α that models the coordinates $\theta_1, \theta_2, \ldots$ as independent variables with $\theta_i \sim \text{Nor}(0, i^{-2\alpha-1})$, for $i = 1, 2, \ldots$ and $\alpha > 0$. In Example 8.6 and Section 9.5.4 the posterior contraction rate for a given α was seen to be the minimax rate $n^{-\alpha/(2\alpha+1)}$ if θ_0 belongs to the Sobolev space $\mathfrak{W}^\alpha$ of sequences θ with $\sum_i i^{2\alpha}\theta_i^2 < \infty$, but to be a possibly

slower rate if θ is contained in a Sobolev space of a different order β. By combining the priors into an average $\sum_\alpha \lambda_\alpha \Pi_\alpha$ we may achieve the optimal rate for a Sobolev space of any order.

Theorem 10.3 is limited to a countable set A_n of indices α, but this is not a real restriction, as two rate sequences $n^{-\alpha/(2\alpha+1)}$ and $n^{-\alpha_n/(2\alpha_n+1)}$ are equivalent up to constants whenever $|\alpha - \alpha_n| \lesssim 1/\log n$. Thus a collection of α in a grid of mesh width of the order $1/\log n$ suffices to construct a prior that is rate-adaptive for any $\alpha > 0$.

Theorem 10.11 (White noise model) *If there exists $\beta_n \in A_n$ with $|\beta_n - \beta| \lesssim 1/\log n$ and $\lambda_{\beta_n} e^{n^{1/(2\beta+1)}} \gtrsim 1$, then $\Pi_n(\|\theta - \theta_0\|_2 \le M n^{-\beta/(2\beta+1)} | X^{(n)}) \to 1$, for any $\theta_0 \in \mathfrak{W}^\beta$, for a constant M that depends on β and $\sum_i i^{2\beta}\theta_{i,0}^2$ only.*

Proof We use estimates on prior masses that may be obtained by direct methods, or derived from general results on Gaussian measures, as discussed in Section 11.4.5. First, by Lemmas 11.47 and 11.48 there exist positive constants d_1 and F_1 depending only on β and $\sum_i i^{2\beta}\theta_{i,0}^2$, if this is finite, so that, for $0 < \epsilon < d_1$,

$$\Pi_{\beta_n}(\theta\colon \|\theta - \theta_0\|_2 < \epsilon) \ge e^{-F_1 \epsilon^{-1/\beta_n}}. \tag{10.17}$$

Second, there exist universal constants $d > 0$ and $0 < c < C < \infty$ such that, for $0 < \epsilon < d^\alpha$ and every $\alpha > 0$, and B_1 the unit ball of ℓ_2,

$$e^{-C(\epsilon/\sqrt{2})^{-1/\alpha}} \le \Pi_\alpha(\theta\colon \|\theta\|_2 < \epsilon) \le e^{-c(\epsilon\sqrt{2})^{-1/\alpha}}, \tag{10.18}$$

$$\Pi_\alpha(M\mathfrak{W}_1^{\alpha+1/2} + \epsilon B_1) \ge \Phi\Big(-\sqrt{2C(\epsilon/2^{1/2})^{-1/\alpha}} + M\Big). \tag{10.19}$$

The two "small ball probability" inequalities are given in Lemma 11.47. Inequality (10.19) is a consequence of Borell's inequality, the lower bound on the small ball probability given by (10.18), and Lemma K.6(ii).

We derive the theorem as a corollary of Theorem 10.3, with d the ℓ_2-norm, making the choices, for a constant $D > 0$ to be determined:

$$\begin{aligned}
\epsilon_{n,\alpha} &= \begin{cases} 6\sqrt{2}\, n^{-\alpha/(2\alpha+1)}, & \text{if } \alpha < \beta, \\ 6\sqrt{2}\, n^{-\beta/(2\beta+1)}, & \text{if } \alpha \ge \beta, \end{cases} \\
M_\alpha^2 &= D n^{1/(2\alpha+1)} = D n \epsilon_{n,\alpha}^2/72, \\
\mathcal{P}_{n,\alpha} &= \begin{cases} M_\alpha \mathfrak{W}_1^{\alpha+1/2} + \frac{1}{6}\epsilon_{n,\alpha} B_1, & \text{if } \alpha < \beta, \\ M_\beta \mathfrak{W}_1^{\beta+1/2} + \frac{1}{6}\epsilon_{n,\beta} B_1, & \text{if } \alpha \ge \beta. \end{cases}
\end{aligned}$$

Condition (10.12) is satisfied, by (10.17) and the assumption on λ_{β_n}, with $F = F_1 + 1$. We verify the remaining conditions of Corollary 10.5.

By Proposition C.10, for $\alpha < \beta$,

$$\log N\Big(\frac{\epsilon_{n,\alpha}}{3}, \mathcal{P}_{n,\alpha}, \|\cdot\|_2\Big) \le \log N\Big(\frac{\epsilon_{n,\alpha}}{6}, M_\alpha \mathfrak{W}_1^{\alpha+1/2}, \|\cdot\|_2\Big) \lesssim \beta\Big(\frac{18 M_\alpha}{\epsilon_{n,\alpha}}\Big)^{1/(\alpha+1/2)}.$$

The right side evaluates as $E_\alpha n \epsilon_{n,\alpha}^2$, for the constants $E_\alpha = \beta(3\sqrt{D/2})^{1/\alpha+1/2}/72$, which are bounded in $\alpha \le \beta$. For $\alpha \ge \beta$, the statement reduces to the single statement for $\alpha =$

β, by the definitions, and hence is also true. This verifies (10.5) for the models $\mathcal{P}_{n,\alpha}$. As these models do not fully support the priors Π_α, we supplement this with showing that the prior mass in their complements is negligible, following Remark 10.4. For $\alpha < \beta$ we have $(\epsilon_{n,\alpha}/6\sqrt{2})^{-1/\alpha} = n^{1/(2\alpha+1)} = M_\alpha^2/D$ and hence, by (10.19), for $D > 8C$,

$$\Pi_\alpha(\mathcal{P}_{n,\alpha}^c) \le 1 - \Phi\Big(-\sqrt{2CM_\alpha^2/D} + M_\alpha\Big) \le e^{-M_\alpha^2/8} \le e^{-Dn\epsilon_{n,\beta}^2/576}.$$

For $\alpha > \beta$ we have $(\epsilon_{n,\beta}/6\sqrt{2})^{-1/\alpha} \le (\epsilon_{n,\beta}/6\sqrt{2})^{-1/\beta} = M_\beta^2/D$. Since the Sobolev unit balls $\mathfrak{W}_1^\alpha$ are decreasing in α, it follows that $\mathcal{P}_{n,\alpha} \supset M_\beta \mathfrak{W}_1^{\alpha+1/2} + \epsilon_{n,\beta} B_1/6$ and hence, by (10.19), for $D > 8C$,

$$\Pi_\alpha(\mathcal{P}_{n,\alpha}^c) \le 1 - \Phi\Big(-\sqrt{2CM_\beta^2/D} + M_\beta\Big) \le e^{-Dn\epsilon_{n,\beta}^2/576}.$$

Combining the preceding two displays we see that $\sum_\alpha \lambda_\alpha \Pi_\alpha(\mathcal{P}_{n,\alpha}^c) = o(e^{-(F+2)n\epsilon_{n,\beta}^2})$, if D is chosen to satisfy $D > 576(F+2)$.

We are left to verify conditions (10.13) and (10.14). The first is easily satisfied, with $\lambda_{n,\alpha} = \lambda_\alpha = \mu_{n,\alpha}$, by the assumption on λ_{β_n}. By the upper inequality in (10.18) the left side of the second is bounded above by

$$\sum_{\alpha \prec \beta_n} \frac{\lambda_\alpha}{\lambda_{\beta_n}} e^{-c(M\epsilon_{n,\alpha}\sqrt{2})^{-1/\alpha}} = \sum_{\alpha \prec \beta_n} \frac{\lambda_\alpha}{\lambda_{\beta_n}} e^{-c(12M)^{-1/\alpha} n\epsilon_{n,\alpha}^2/72}.$$

For $\alpha \prec \beta_n$ minus the exponent is bigger than $c(12M)^{-1/\alpha} H n\epsilon_{n,\beta}^2/72$, which can be made bigger than $(3+F)n\epsilon_{n,\beta}^2$, for every α and fixed M, by choosing a large H. Condition (10.14) follows. □

10.3.3 Finite-Dimensional Approximations

Consider models $\mathcal{P}_{n,J,M}$ indexed by a parameter $\alpha = (J, M)$ consisting of a "dimension parameter" $J \in \mathbb{N}$ and a second parameter $M \in \mathfrak{M}$, such that, for every (J, M) and constants A_M, for every $\epsilon > 0$,

$$\log N\Big(\frac{\epsilon}{5}, C_{n,J,M}(p_0, 2\epsilon), d\Big) \le A_M J. \tag{10.20}$$

Thus the models $\mathcal{P}_{n,J,M}$ are J-dimensional in the sense of the Le Cam entropy. Such finite-dimensional models may arise from approximation of a collection of target densities through a set of basis functions (e.g. trigonometric functions, splines or wavelets), where a model of dimension J is generated by a selection of J basis functions. The index M may refer to a restriction on the coefficients used with this selection, but also to multiple selections of dimension J.

In this context an abstract definition of "regularity" of order β of a true density p_0, given the list of models $\mathcal{P}_{n,J,M}$, could be that, for some $M_0 \in \mathfrak{M}$,

$$d(p_0, \mathcal{P}_{n,J,M_0}) \lesssim J^{-\beta}.$$

If p_0 is β-regular in this sense, then one might hope that a suitable estimation scheme using the model $\mathcal{P}_{n,J,M_0}$ would lead to a bias of order $J^{-\beta}$, and to a variance term of order J/n.

The best dimension J would balance the square bias and the variance, leading to an optimal dimension J satisfying $J^{-2\beta} \asymp J/n$. This is solved by $J \asymp n^{1/(2\beta+1)}$ and would lead to an "optimal" rate of contraction $n^{-\beta/(2\beta+1)}$.

For super-regular densities satisfying $d(p_0, \mathcal{P}_{n,J,M_0}) \lesssim \exp(-J^\beta)$, or even $p_0 \in \mathcal{P}_{n,J_0,M_0}$ for some J_0 and M_0, a similar argument would lead to rates closer to $1/\sqrt{n}$.

The theorem in this section shows that an adaptive Bayesian scheme, using fairly simple priors, can yield these optimal rates up to a logarithmic factor. As finite-dimensional models are widely applicable, this illustrates the ease by which adaptation is achievable if one is willing to accept a logarithmic factor in the rate. This factor can be avoided by using other schemes (e.g. based on a discretization of the coefficient space as in Section 10.3.1, a smooth prior on restricted coefficient space, or more complicated model weights as in Section 10.3.4).

Le Cam's definition of dimension is combinatorial rather than geometric. A "geometrically J-dimensional" model can be described smoothly by a J-dimensional parameter $\theta \in \mathbb{R}^J$. In that case it is natural to construct a prior on $\mathcal{P}_{n,J,M}$ by putting a prior on the parameter θ. If this prior is chosen to be smooth on $\mathbb{R}^J$, and a ball of d-radius ϵ in $\mathcal{P}_{n,J,M}$ corresponds to a ball of radius $\bar{B}_J \bar{C}_M \epsilon$ on the coefficients $\theta \in \mathbb{R}^J$ (for some constants $\bar{B}_J \bar{C}_M$), then we may expect that, for some constant D_M,

$$\Pi_{n,J,M}(B_{n,J,M}(p_0, \epsilon)) \geq (B_J C_M \epsilon)^J, \qquad \text{if } \epsilon > D_M\, d(p_0, \mathcal{P}_{n,J,M}). \tag{10.21}$$

Here the constants B_J and C_M incorporate the constants $\bar{B}_J$ and $\bar{C}_M$, the prior density on $\mathbb{R}^J$, and the volume of a J-dimensional ball. A restriction of the type $\epsilon \gtrsim d(p_0, \mathcal{P}_{n,J,M})$ is necessary, because by their definition the sets $B_{n,J,M}(p_0, \epsilon)$ are centered around p_0, and this may be at a positive distance to $\mathcal{P}_{n,J,M}$. If $\epsilon > 2d(p_0, \mathcal{P}_{n,J,M})$, then a ball of radius $\epsilon/2$ around a projection of p_0 into $\mathcal{P}_{n,J,M}$ is contained in $C_{n,J,M}(p_0, \epsilon)$. The general constant D_M in (10.21), instead of the universal constant 2, is meant to make up for the difference between the neighborhoods $B_{n,J,M}(p_0, \epsilon)$ and $C_{n,J,M}(p_0, \epsilon)$.

For a large constant A, an arbitrary positive constant C and finite sets $\mathcal{J}_n \subset \mathbb{N}$ and $\mathcal{M}_n \subset \mathcal{M}$, define

$$\epsilon_{n,J,M} = \sqrt{\frac{J \log n}{n} A_M A},$$
$$\lambda_{n,J,M} = \frac{\exp[-Cn\epsilon_{n,J,M}^2]}{\sum_{(J,M)\in\mathcal{J}_n\times\mathcal{M}_n} \exp[-Cn\epsilon_{n,J,M}^2]}\, \mathbb{1}_{\mathcal{J}_n\times\mathcal{M}_n}(J, M).$$

The $\lambda_{n,J,M}$ are the "universal weights" (10.15), with $\alpha = (J, M)$ and $\lambda_{J,M} = 1$, restricted to the set $A_n = \mathcal{J}_n \times \mathcal{M}_n$.

Theorem 10.12 (Finite-dimensional approximation) *Suppose that* (10.20)–(10.21) *hold for every J and M, where $A_M A \geq 1$, $C_M^2 A_M A \geq e$ and $B_J \sqrt{J} \geq e$. Let $\mathcal{J}_n$ and $\mathcal{M}_n$ be such that $\sum_{M\in\mathcal{M}_n} e^{-LA_M} = O(1)$ for some $L > 0$. Then for every sequences $J_n \in \mathcal{M}_n$ and $M_n \in \mathcal{M}_n$ with $D_{M_n} d(p_0, \mathcal{P}_{n,J_n,M_n}) \leq \epsilon_{n,J_n,M_n}$, the posterior distribution relative to the model weights $\lambda_{n,J,M}$ attains rate of contraction ϵ_{n,J_n,M_n} at p_0 relative to d.*

Proof We apply Corollary 10.6 with α equal to the pair (J, M) and $\beta_n = (J_n, M_n)$. Condition (10.5) is satisfied (easily with an extra logarithmic factor) by (10.20) in virtue of the definition of the numbers $\epsilon_{n,J,M}$, with $E_\alpha = 1$, for which $n\epsilon_{n,J,M}^2 = J(\log n)A_M A$.

Because $D_{M_n} d(p_0, \mathcal{P}_{n,J_n,M_n}) \le \epsilon_{n,J_n,M_n}$ by assumption, condition (10.21) implies that the prior mass in the left side of (10.12) can be bounded below by

$$\big(B_{J_n} C_{M_n} \epsilon_{n,J_n,M_n}\big)^{J_n} = \exp\{J_n \log(B_{J_n}\sqrt{J_n}) + \frac{1}{2} J_n \log(C_{M_n}^2 A_{M_n} A)\} \exp\{-\frac{1}{2} J_n \log(n/\log n)\}.$$

The first factor on the right is bounded below by 1 in view of the assumptions on the constants. Because $n\epsilon_{n,J,M}^2 = J(\log n)A_M A$ and $A_M A \ge 1$, it follows that (10.12) is satisfied with $F = 1$.

Finally, we verify that $\sum_\alpha (\lambda_\alpha/\lambda_\beta) e^{-Cn\epsilon_{n,\alpha}^2/4} = O(1)$. Because presently $\lambda_\alpha = 1$, the left side of this equation takes the form

$$\sum_{J \in \mathcal{J}_n} \sum_{M \in \mathcal{M}_n} e^{-CJ(\log n)A_M A/4} \le \sum_{M \in \mathcal{M}_n} e^{-LA_M},$$

for any constant L and n sufficiently large. The right side is bounded for some L, by assumption. □

Example 10.13 (Supersmooth true density) If $p_0 \in \mathcal{P}_{n,J_0,M_0}$ for some pair of constants (J_0, M_0), then we can apply the preceding theorem with $(J_n, M_n) = (J_0, M_0)$, yielding a rate $(\log n)^{1/2}/\sqrt{n}$.

If $d(p_0, \mathcal{P}_{n,J,M_0}) \lesssim e^{-J^\beta}$ for every J, then we can apply the preceding theorem with J_n a multiple of $(\log n)^{1/\beta}$, yielding a rate $(\log n)^{1/(2\beta)+1/2}/\sqrt{n}$.

Example 10.14 (Regular true density) If $d(p_0, \mathcal{P}_{n,J,M_0}) \lesssim J^{-\beta}$ for every J and some M_0, then we can apply the preceding theorem with J_n a multiple of $(n/\log n)^{1/(2\beta+1)}$, yielding a rate $(n/\log n)^{-\beta/(2\beta+1)}$.

As the two examples illustrate, the logarithmic factor present in the definition of $\epsilon_{n,J,M}$ typically works through in the contraction rate, which tends to be suboptimal by a logarithmic factor. This appears to be inherent in the construction, using smooth priors on the coefficients of the model. The discrete priors of Section 10.3.1 use similar model weights, and do not have this deficit. This point is explored further in the concrete situation of spline approximations in Section 10.3.4.

10.3.4 Log-Spline Models

In Section 9.1 log-spline models of dimension $J_{n,\alpha} \sim n^{1/(2\alpha+1)}$ were seen to yield the minimax posterior contraction rate for a true density whose logarithm belongs to $\mathfrak{C}^\alpha[0, 1]$. The prior, which depends on the target smoothness α through the dimension of the model, can be made independent of α with a hyperprior on the dimension. As long as this gives the right amount of weight to (neighborhoods) of the dimension $J_{n,\alpha}$, the posterior distribution should adapt to α.

In the following four subsections we discuss several strategies, which differ in their choices of prior for the coefficients θ and prior for the dimension. The first strategy simply follows the general finite-dimensional construction of Section 10.3.3. It combines non-informative priors on the coefficients, with the universal model weights (10.15), which down-weight higher-dimensional models. The second construction uses similar flat priors on the coefficients, but fixed model weights. Notwithstanding the big difference, both constructions yield adaptation up to a logarithmic factor in the contraction rate. The third and fourth constructions are focused on removing this factor. The third is limited to finitely many smoothness levels, and somewhat surprisingly, manages to remove the logarithmic factor by down-weighting lower-dimensional models. The fourth construction uses again the universal model weights on the dimension, but changes the prior on the coefficients to a discrete prior, in the spirit of Section 10.3.1. One may conclude from these examples that a variety of adaptation schemes may work, in particular if an unnecessary logarithmic factor is taken for granted.

Spline functions only have good approximation properties for α-smooth functions if their order is at least α. For adaptation to a finite range of α, the order can, of course, be chosen an upper bound to this range. Furthermore, if the hyperprior on dimension is placed indirectly and induced by a prior on α and the map $\alpha \mapsto J_{n,\alpha}$, then the splines for that particular model can be chosen of order at least α. For a construction with a hyperparameter directly on the dimension, independent of the sample size, it is in general not possible to link the order of the splines to a target smoothness. Adaptation may be limited to smoothness smaller than the order of the splines, unless an additional layer of models consisting of spline bases of different orders is introduced.

Throughout this section we assume that the true density p_0 is contained in the Hölder space $\mathfrak{C}^\beta[0, 1]$, for some $\beta > 0$, and strictly positive. In the first construction the posterior adapts to the minimal and maximal values of the true density, but in the second to fourth constructions the prior depends on known lower and upper bounds on the true density. An additional layer of adaptation might remove this dependency, but is not explored here.

We adopt the notation of Section 9.1. In particular, a log-spline density is denoted by $p_{J,\theta}$, where J is the dimension, and $\theta \in \mathbb{R}^J$ the parameter of the exponential family. We also adopt the notations $B_{J,M}(p_0, \epsilon)$ and $C_{J,M}(p_0, \epsilon)$ for a Kullback-Leibler and Hellinger ball of radius ϵ around p_0 within the restricted log-spline model with parameter set $\Theta_{J,M} = \{\theta \in [-M, M]^J : \theta^\top 1 = 0\}$, as given in (9.1).

Throughout we denote by $\Pi_{n,J,M}$ the prior distribution both for the parameter $\theta \in \Theta_{J,M}$ and for the induced density $p_{J,\theta}$.

Smooth Coefficients, Universal Model Weights

We combine the uniform probability measures $\Pi_{n,J,M}$ on the sets $\Theta_{J,M}$, for $(J, M) \in \mathbb{N}^2$, with the model weights $\lambda_{n,J,M}$ proportional to $e^{-CJ\log(nM)}$. This corresponds to the general construction of Section 10.3.3, with the models $\mathcal{P}_{J,M}$ consisting of the spline densities $p_{J,\theta}$ with $\theta \in \Theta_{J,M}$. We thus obtain the following result.

Corollary 10.15 *If* $\log p_0 \in \mathfrak{C}^\beta[0, 1]$, *for some* $\beta > 0$, *then the rate of posterior contraction for the Hellinger distance is* $(n/\log n)^{-\beta/(2\beta+1)}$.

Proof From Lemmas 9.4–9.6 it can be seen that conditions (10.20)–(10.21) of Theorem 10.12 are satisfied with

$$A_M = 2M + \log(30C_0/c_0), \qquad B_J = \sqrt{J}(\mathrm{vol}\{x \in \mathbb{R}^J \colon \|x\| \le 1\})^{1/J},$$
$$C_M = M^{-2}e^M/(8D_0C_0), \qquad D_M = 2D_0M.$$

The constants B_J tend to a fixed value as $J \to \infty$, by Lemma K.13. If we choose A_M to satisfy $A_M = C_M^{-2}$, then the conditions of Theorem 10.12 are satisfied. Since $d_H(p_0, \mathcal{P}_{J,M}) \lesssim J^{-\beta}$, for every $M \ge 2\|\log p_0\|_\infty/d_0$, by Lemma 9.6(ii) and (i), the theorem applies with $J_n \asymp n^{1/(2\beta+1)}$ and fixed large $M = M_n$. □

Flat Priors, Fixed Model Weights

We combine the uniform probability measures $\Pi_{n,\alpha}$ on the sets $\Theta_{J_{n,\alpha},M}$, for $\alpha \in A := \{\alpha \in \mathbb{Q}^+ \colon \alpha \ge \underline{\alpha}\}$ for some constant $\underline{\alpha} > 0$ and a fixed $M > 0$, with fixed model weights $\lambda_{n,\alpha} = \lambda_\alpha > 0$ concentrated on A.

Corollary 10.16 *If $p_0 \in \mathfrak{C}^\beta[0,1]$ for some $\beta \in \mathbb{Q}^+ \cap [\underline{\alpha}, \infty)$ and $\|\log p_0\|_\infty < d_0M/2$ and $\sum_\alpha \sqrt{\lambda_\alpha} < \infty$, then then the rate of posterior contraction for the Hellinger distance is $n^{-\beta/(2\beta+1)}\sqrt{\log n}$.*

Proof We shall show that the conditions of Theorem 10.3 hold for $\epsilon_{n,\alpha} := n^{-\alpha/(2\alpha+1)}\sqrt{\log n}$. These numbers satisfy $n\epsilon_{n,\alpha}^2 \asymp J_{n,\alpha}\log n$ and $\epsilon_{n,\alpha} \gtrsim J_{n,\alpha}^{-\alpha}$, for $J_{n,\alpha} \asymp n^{1/(2\alpha+1)}$, and all $\alpha > 0$.

By Lemma 9.5 relation (10.5) holds whenever $n\epsilon_{n,\alpha}^2 \gtrsim J_{n,\alpha}$, with constants E_α that are independent of α.

Because $\|\log p_0\|_\infty < d_0M/2$ by assumption, $d_H(p_0, \mathcal{P}_{J,M}) \lesssim J^{-\beta}$, by Lemma 9.6. Since $\epsilon_{n,\beta} \gtrsim J_{n,\beta}^{-\beta}$, Lemma 9.4 implies that for some positive constants c and C, and sufficiently large n,

$$\Pi_{n,\alpha}(C_{n,\alpha}(p_0,\epsilon)) \le \Pi_{n,\alpha}(\theta \in \Theta_{J_{n,\alpha},M} \colon \|\theta - \theta_{J_{n,\alpha},M}\| \le c\sqrt{J_{n,\alpha}}\epsilon),$$
$$\Pi_{n,\beta}(B_{n,\beta}(p_0,\epsilon_{n,\beta})) \ge \Pi_{n,\beta}(\theta \in \Theta_{J_{n,\beta},M} \colon \|\theta - \theta_{J_{n,\beta}}\| \le 2CA\sqrt{J_{n,\alpha}}\epsilon_{n,\beta}).$$

For $\theta_{J,M} \in \Theta_{J,M}$ defined in Lemma 9.6, $\mathrm{vol}(\theta \in \Theta_{J,M} \colon \|\theta - \theta_{J,M}\| \le \epsilon) \ge 2^{-J}\mathrm{vol}(\theta \colon \|\theta - \theta_J\|_2 \le \epsilon)$. Thus by Lemma K.13, for any $\alpha, \beta \in A$ and every small ϵ,

$$\frac{\Pi_{n,\alpha}(C_{n,\alpha}(p_0, i\epsilon))}{\Pi_{n,\beta}(B_{n,\beta}(p_0,\epsilon_{n,\beta}))} \le \frac{(cDi\epsilon\sqrt{J_{n,\alpha}})^{J_{n,\alpha}}\mathrm{vol}\{x \in \mathbb{R}^{J_{n,\alpha}} \colon \|x\| \le 1\}}{(Cd\epsilon_{n,\beta}\sqrt{J_{n,\beta}})^{J_{n,\beta}}\mathrm{vol}\{x \in \mathbb{R}^{J_{n,\beta}} \colon \|x\| \le 1\}} \lesssim \frac{(Ai\epsilon)^{J_{n,\alpha}}}{(a\epsilon_{n,\beta})^{J_{n,\beta}}}, \tag{10.22}$$

for suitable constants a and A.

If $\alpha \prec \beta$, then with $\epsilon = \epsilon_{n,\alpha}$, inequality (10.22) yields

$$\frac{\Pi_{n,\alpha}(C_{n,\alpha}(p_0, i\epsilon_{n,\alpha}))}{\Pi_{n,\beta}(B_{n,\beta}(p_0,\epsilon_{n,\beta}))} \lesssim \exp\Big[J_{n,\alpha}\Big(\log(Ai\epsilon_{n,\alpha}) - \frac{J_{n,\beta}}{J_{n,\alpha}}\log(a\epsilon_{n,\beta})\Big)\Big]$$
$$\le \exp\Big[J_{n,\alpha}\Big(\log(Ai) + H^{-1}|\log a| + \log\epsilon_{n,\alpha} - H^{-1}\log\epsilon_{n,\beta}\Big)\Big],$$

because $J_{n,\alpha} > H J_{n,\beta}$ for $\alpha < \beta$ for sufficiently large n, no matter which fixed constant H is chosen. Also for some sufficiently large H, simultaneously for all $\alpha \geq \underline{\alpha}$,

$$\log \epsilon_{n,\alpha} - \frac{1}{H} \log \epsilon_{n,\beta} = \Big(\frac{1}{H} \frac{\beta}{2\beta+1} - \frac{\alpha}{2\alpha+1} \Big) \log n + \tfrac{1}{2} \Big(1 - \frac{1}{H} \Big) \log \log n$$

is bounded by a negative multiple of $\log n$, so (10.6) holds with $\mu_{n,\alpha} = \mu_\alpha / \mu_\beta$ and arbitrarily small $L > 0$.

If $\alpha \prec \beta$, then with $\epsilon = IB\epsilon_{n,\alpha}$, inequality (10.22) and similar calculations yield

$$\begin{aligned} & e^{2n\epsilon_{n,\beta}^2} \frac{\Pi_{n,\alpha}(C_{n,\alpha}(p_0, IB\epsilon_{n,\alpha}))}{\Pi_{n,\beta}(B_{n,\beta}(p_0, \epsilon_{n,\beta}))} \\ & \lesssim \exp\Big[J_{n,\alpha} \Big(\log(AIB) + \log \epsilon_{n,\alpha} - \frac{J_{n,\beta}}{J_{n,\alpha}} \log(a\epsilon_{n,\beta}) + 2 \frac{J_{n,\beta}}{J_{n,\alpha}} \log n \Big) \Big] \\ & \leq \exp\Big[J_{n,\alpha} \Big(\log(AIB) + \frac{1}{H} |\log a| + \log \epsilon_{n,\alpha} - \frac{1}{H} \log \epsilon_{n,\beta} + \frac{2}{H} \log n \Big) \Big]. \end{aligned}$$

As before, for sufficiently large H, the exponent is smaller than $-J_{n,\alpha} c \log n$ for a positive constant c, simultaneously for all $\alpha \geq \underline{\alpha}$, eventually. This implies that Condition (10.8) is satisfied.

With $\epsilon = \epsilon_{n,\beta}$, inequality (10.22) yields

$$\frac{\Pi_{n,\alpha}(C_{n,\alpha}(p_0, i\epsilon_{n,\beta}))}{\Pi_{n,\beta}(B_{n,\beta}(p_0, \epsilon_{n,\beta}))} \lesssim \exp\Big[J_{n,\beta} \Big(\frac{J_{n,\alpha}}{J_{n,\beta}} (\log(Ai) + \log \epsilon_{n,\beta}) - \log(a\epsilon_{n,\beta}) \Big) \Big].$$

If $\alpha \succeq \beta$ the right-hand side is bounded above by

$$\exp\big[J_{n,\beta} (H |\log(Ai)| - \log(a\epsilon_{n,\beta})) \big] \leq e^{J_{n,\beta} L i^2 \log n},$$

for sufficiently large i, for any arbitrarily small constant L. Thus condition (10.7) holds. □

Flat Priors, Decreasing Model Weights

We combine the uniform probability measures $\Pi_{n,\alpha}$ on the sets $\Theta_{J_{n,\alpha},M}$, for α in a finite set $\{\alpha_1, \alpha_2, \ldots, \alpha_N\} \subset (0, \infty)$ with the weights

$$\lambda_{n,\alpha} \propto \prod_{\gamma \in A : \gamma < \alpha} (C\epsilon_{n,\gamma})^{J_{n,\gamma}}.$$

These weights vary with n and are decreasing in α, unlike the universal weights (10.15).

Corollary 10.17 *If $p_0 \in \mathfrak{C}^\beta[0,1]$ for some $\beta \in \{\alpha_1, \alpha_2, \ldots, \alpha_N\}$ and $\|\log p_0\|_\infty < d_0 M/2$, then the rate of posterior contraction for the Hellinger distance is $n^{-\beta/(2\beta+1)}$.*

Proof Let $\epsilon_{n,\alpha} = n^{-\alpha/(2\alpha+1)}$, so that $J_{n,\alpha} \asymp n\epsilon_{n,\alpha}^2$, and $J_{n,\alpha'}/J_{n,\alpha} \ll n^{-c}$ for some $c > 0$ whenever $\alpha' > \alpha$. Assume without loss of generality that $\alpha_1 < \cdots < \alpha_N$.

If $r < s$, and hence $\alpha_r < \alpha_s$, then, by inequality (10.22),

$$\frac{\lambda_{n,\alpha_r}\Pi_{n,\alpha_r}(C_{n,\alpha_r}(p_0, i\epsilon_{n,\alpha_r}))}{\lambda_{n,\alpha_s}\Pi_{n,\alpha_s}(B_{n,\alpha_s}(p_0,\epsilon_{n,\alpha_s}))}$$
$$\lesssim \exp\Big[J_{n,\alpha_r}\Big(\log(Ai\epsilon_{n,\alpha_r}) - \frac{J_{n,\alpha_s}}{J_{n,\alpha_r}}\log(a\epsilon_{n,\alpha_s}) - \sum_{k=r}^{s-1}\frac{J_{n,\alpha_k}}{J_{n,\alpha_r}}\log(C\epsilon_{n,\alpha_k})\Big)\Big]$$
$$= \exp\Big[J_{n,\alpha_r}\Big(\log\Big(\frac{Ai}{C}\Big) - \frac{J_{n,\alpha_s}}{J_{n,\alpha_r}}\log(a\epsilon_{n,\alpha_s}) - \sum_{k=r+1}^{s-1}\frac{J_{n,\alpha_k}}{J_{n,\alpha_r}}\log(C\epsilon_{n,\alpha_k})\Big)\Big].$$

The exponent takes the form $J_{n,\alpha_r}(\log(Ai/C) + o(1))$. Applying this with $\alpha_r = \alpha < \beta = \alpha_s$, it follows that, for any chosen C, (10.6) holds with $\mu_{n,\alpha} = 1$ and any arbitrarily small $L > 0$, for sufficiently large i and n.

Similarly, again with $\alpha_r = \alpha < \beta = \alpha_s$,

$$e^{2n\epsilon_{n,\alpha_s}^2}\frac{\lambda_{n,\alpha_r}\Pi_{n,\alpha_r}(C_{n,\alpha_r}(p_0, IB\epsilon_{n,\alpha_r}))}{\lambda_{n,\alpha_s}\Pi_{n,\alpha_s}(B_{n,\alpha_s}(p_0,\epsilon_{n,\alpha_s}))}$$
$$= \exp\Big[J_{n,\alpha_r}\Big(\log\Big(\frac{AIB}{C}\Big) - \frac{J_{n,\alpha_s}}{J_{n,\alpha_r}}\log(a\epsilon_{n,\alpha_s}) - \sum_{k=r+1}^{s-1}\frac{J_{n,\alpha_k}}{J_{n,\alpha_r}}\log(C\epsilon_{n,\alpha_k}) + \frac{2J_{n,\alpha_s}}{J_{n,\alpha_r}}\Big)\Big].$$

This tends to 0 if $C > AIB$. Hence, for C big enough, condition (10.8) is fulfilled as well.

Finally, choose $\alpha_r = \beta < \alpha = \alpha_s$ and note that

$$\frac{\lambda_{n,\alpha_s}\Pi_{n,\alpha_s}(C_{n,\alpha_s}(p_0, i\epsilon_{n,\alpha_r}))}{\lambda_{n,\alpha_r}\Pi_{n,\alpha_r}(B_{n,\alpha_r}(p_0,\epsilon_{n,\alpha_r}))}$$
$$\lesssim \exp\Big[J_{n,\alpha_r}\Big(\frac{J_{n,\alpha_s}}{J_{n,\alpha_r}}(\log(Ai\epsilon_{n,\alpha_r}) - \log(a\epsilon_{n,\alpha_r}) + \sum_{k=r}^{s-1}\frac{J_{n,\alpha_k}}{J_{n,\alpha_r}}\log(C\epsilon_{n,\alpha_k})\Big)\Big]$$
$$= \exp\Big[J_{n,\alpha_r}\Big(\frac{J_{n,\alpha_s}}{J_{n,\alpha_r}}\log(Ai\epsilon_{n,\alpha_r}) + \log\Big(\frac{C}{a}\Big) + \sum_{k=r+1}^{s-1}\frac{J_{n,\alpha_k}}{J_{n,\alpha_r}}\log(C\epsilon_{n,\alpha_k})\Big)\Big].$$

Here the exponent is of the order $J_{n,\alpha_r}(\log(C/a) + o(1)\log i + o(1))$. We conclude that the condition (10.7) holds.

Now the assertion follows from Theorem 10.3, with $A_n = A$. □

Discrete Priors, Universal Model Weights

We combine discrete priors $\Pi_{n,\alpha}$ on the sets $\Theta_{J_{n,\alpha},M}$, for $\alpha \in \mathbb{Q}^+$, with model weights, for fixed λ_α and $C > 0$,

$$\lambda_{n,\alpha} \propto \lambda_\alpha e^{-Cn^{1/(2\alpha+1)}}.$$

The support of the prior $\Pi_{n,\alpha}$ are finitely many spline functions, constructed as follows. By Proposition C.5 there exists an $\epsilon_{n,\alpha}$-net $f_1, \ldots, f_{N_{n,\alpha}}$ in the collection of $f \in \mathfrak{C}^\alpha[0,1]$ with $\|f\|_{\mathfrak{C}^\alpha} \le M$ of cardinality $N_{n,\alpha} \lesssim (M/\epsilon_{n,\alpha})^{1/\alpha}$. By Lemma E.5 there exist $\theta_i \in \mathbb{R}^{J_{n,\alpha}}$ such that $\|\theta_i^\top B_{J_{n,\alpha}} - f_i\|_\infty \lesssim MJ_{n,\alpha}^{-\alpha}$, for $i = 1, 2, \ldots, N_{n,\alpha}$. The points $p_{J_{n,\alpha},\theta_i}$ form the support of $\Pi_{n,\alpha}$.

This construction follows the general construction using discrete priors, described in Section 10.3.1, with model weights of the form (10.15). The following is a corollary of Theorem 10.9.

Corollary 10.18 *If $p_0 \in \mathfrak{C}^\beta[0,1]$ for some $\beta \in \mathbb{Q}^+$, $\|\log p_0\|_{\mathfrak{C}^\beta} < M$ and $\|\log p_0\|_\infty < d_0 M/2$, then the rate of posterior contraction for the Hellinger distance is $n^{-\beta/(2\beta+1)}$.*

10.4 Finite Random Series

In this section we consider combining priors defined by finite random, linear combinations of given basis functions into an overall prior by equipping the number J of terms with a hyperprior. Adaptation then occurs provided some suitable linear combination of the basis functions approximates the "true" function well enough and the prior on J is chosen correctly. Given the great variety of possible bases, this provides a flexible method of prior construction, which we illustrate with several examples. The priors are a concrete implementation of the finite-dimensional approximations of Section 10.3.3. The resulting contraction rates likewise suffer from logarithmic factors that can only be avoided with more carefully constructed priors.

For given $J \in \mathbb{N}$, fix basis functions $\psi_{J,1}, \ldots, \psi_{J,J}: \mathfrak{X} \to \mathbb{R}$ on a domain $\mathfrak{X}$, which is typically a bounded convex set in $\mathbb{R}^d$. Construct a prior on functions $w: \mathfrak{X} \to \mathbb{R}$ by choosing a random dimension J and random coefficients $\theta_J = (\theta_{J,1}, \ldots, \theta_{J,J}) \in \mathbb{R}^J$, and next forming the function

$$\theta_J^\top \psi_J := \sum_{j=1}^{J} \theta_{J,j} \psi_{J,j}.$$

As indicated in the notation, the basis functions, and the priors on coefficients and dimension, may depend on J, and in the asymptotic results they may also further change from one stage to the next. However, the constants in the following conditions should be universal. With some abuse of notation we denote by $\theta^\top \psi$ a generic function, for an unspecified dimension J, and denote by Π both the prior on the pair (θ, J) and on the functions $\theta^\top \psi$.

Throughout the section we impose the following conditions on these priors:

(A1) $A(j) \le \Pi(J = j) \le \Pi(J \ge j) \le B(j)$, for sufficiently large j, for nonnegative, strictly decreasing functions $A, B: \mathbb{N} \to \mathbb{R}$ with limit $A(\infty) = B(\infty) = 0$ at ∞.

(A2) $\Pi(\|\theta_j - \theta_{0,j}\| \le \epsilon) \ge e^{-c_2 j \log \frac{1}{\epsilon}}$, for every $\theta_{0,j} \in \mathbb{R}^j$ with $\|\theta_{0,j}\|_\infty \le H$, for given positive constants c_2 and H, and every sufficiently small $\epsilon > 0$.

(A3) $\Pi(\theta_j \notin [-M, M]^j) \le j e^{-R(M)}$, for every j and M, for an increasing function $R: [0, \infty) \to [0, \infty)$ with limit $R(\infty) = \infty$ at ∞.

Geometric and negative binomial distributions on J satisfy (A1) with $A(j) \asymp B(j) \asymp e^{-cj}$, while Poisson distributions satisfy (A1) with $A(j) \asymp B(j) \asymp e^{-cj \log j}$, for some $c > 0$. Priors that satisfy (A2) include the gamma and exponential distributions assigned independently to every coordinate of θ_j, and multivariate normal and Dirichlet distributions if the parameters lie in a fixed compact set (see Lemma G.13).

Let d be a metric such that, for a positive increasing function $\rho: \mathbb{N} \to \mathbb{R}$, and every $j \in \mathbb{N}$,

$$d(\theta_1^\top \psi_j, \theta_2^\top \psi_j) \le \rho(j) \, \|\theta_1 - \theta_2\|, \qquad \text{every } \theta_1, \theta_2 \in \mathbb{R}^j. \tag{10.23}$$

Example 10.19 (Metrics) For any uniformly bounded basis functions on a domain of finite dominating measure μ the conformity (10.23) of the metric d on the functions $\theta^\top \psi$ with

the Euclidean metric on the coefficients is valid for d equal to any $\mathbb{L}_r(\mu)$-metric (where $1 \le r \le \infty$) and $\rho(j) \asymp \sqrt{j}$. This follows from the estimate $\|(\theta_1 - \theta_2)^\top \psi\|_r \le \|\theta_1 - \theta_2\|_1 \max_{1\le j\le J} \|\psi_j\|_r$, and the relation $\|\cdot\|_1 \le \sqrt{j}\|\cdot\|_2$ between the ℓ_1- and ℓ_2-norms on $\mathbb{R}^j$. This observation applies for instance to the classical Fourier base, B-splines and many wavelet bases. For B-splines and d equal to the uniform norm, this estimate is sharp, while for d equal to the $\mathbb{L}_2$-norm, inequality (10.23) is even valid with $\rho(j) \asymp j^{-1/2}$ (see Lemma E.6).

For orthonormal basis functions and d equal to the $\mathbb{L}_2$-norm, the correspondence is of course valid with $\rho(j) = 1$.

We shall see below that in the case of a polynomial function ρ its degree is essentially irrelevant for the contraction rate. In particular $\rho(j) = \sqrt{j}$ is not worse than a constant function ρ.

The following lemma reveals the roles of the preceding assumptions in a proof of a contraction rate, with sieves of the form $\mathcal{W}_{J,M} = \{\theta^\top \psi_j : \theta \in \mathbb{R}^j, j \le J, \|\theta\|_\infty \le M\}$.

Lemma 10.20 *If (A1), (A2) and (A3) and* (10.23) *are satisfied, and for a given function $w_0 : \mathfrak{X} \to \mathbb{R}$ there exist integers $\overline{J}_n \in \mathbb{N}$ and vectors $\theta_0 \in [-H, H]^{\overline{J}_n}$ with $d(w_0, \theta_0^\top \psi_{\overline{J}_n}) \le \bar{\epsilon}_n$, then, for $\bar{\epsilon}_n/\rho(\bar{J}_n)$ sufficiently small and $\epsilon_n \le J_n \rho(J_n) M_n$,*

$$\Pi(\theta^\top \psi : d(w_0, \theta^\top \psi) \le 2\bar{\epsilon}_n) \ge A(\overline{J}_n)(\bar{\epsilon}_n/\rho(\overline{J}_n))^{c_2 \overline{J}_n},$$
$$\log D(\epsilon_n, \mathcal{W}_{J_n,M_n}, d) \lesssim J_n \log(3 J_n \rho(J_n) M_n/\epsilon_n),$$
$$\Pi(\mathcal{W}^c_{J_n,M_n}) \le B(J_n) + J_n e^{-R(M_n)}.$$

Proof To prove the first inequality we use the triangle inequality and the assumption on $\overline{J}_n$ to see that the set in its left side contains the set of all functions $\theta^\top \psi_{\overline{J}_n}$ with $d(\theta^\top \psi_{\overline{J}_n}, \theta_0^\top \psi_{\overline{J}_n}) \le \bar{\epsilon}_n$. By (10.23) the latter inequality can be translated into the coefficients θ, giving that the prior probability of this set of functions is bounded below by $\Pi(J = \overline{J}_n)\Pi(\|\theta_{\overline{J}_n} - \theta_0\| \le \bar{\epsilon}_n/\rho(\overline{J}_n))$. The inequality then follows by (A2) and (A3).

Using that the packing number of a union of sets is smaller than the sum of the packing numbers of the individual sets and again assumption (10.23), we obtain that $D(\epsilon, \mathcal{W}_{J_n,M_n}, d) \le \sum_{j=1}^{J_n} D(\epsilon/\rho(j), [-M_n, M_n]^j, \|\cdot\|)$, for any $\epsilon > 0$. Since ρ is increasing by assumption, the right side is bounded above by $J_n D(\epsilon/\rho(J_n), [-M_n, M_n]^{J_n}, \|\cdot\|)$, which is further bounded by $J_n D(\epsilon/(\sqrt{J_n}\rho(J_n)), [-M_n, M_n]^{J_n}, \|\cdot\|_\infty)$. The second assertion next follows from Proposition C.2.

The third Inequality is immediate from (A1) and (A3), upon noting that $\Pi(\mathcal{W}^c_{J_n,M_n}) \le \Pi(J > J_n) + \sum_{j=1}^{J_n} \Pi(\theta \notin [-M_n, M_n]^j)$. □

In the intended application, $\bar{\epsilon}_n$ will be (close to) the targeted rate of contraction as in Theorem 8.9 or Theorem 8.19. The dimension $\overline{J}_n$ will be required to be sufficiently large in order to have sufficiently good approximation to w_0 (i.e. control of the bias of the model), but a too large value will verify the prior mass condition only for a slow rate. The dimension J_n will typically have to be larger than $\overline{J}_n$ to control the remaining mass $\Pi(\mathcal{W}^c_{J_n,M_n})$, but small enough to make the sieve not too complex. In the upper bounds given by the lemma,

the proportionality function ρ from (10.23) influences the log-prior probability and entropy only through a logarithmic factor. This means that as long as ρ is polynomial, its exact degree hardly matters for the final result, the contraction rate.

The following theorem combines the lemma with Theorem 8.19 to obtain contraction rates. Following the setup of Section 8.3, we consider a general observation $X^{(n)}$. We assume that its distribution $P^{(n)}_{w,\eta}$ is parameterized by a pair of a function $w\colon \mathfrak{X} \to \mathbb{R}$ and possibly an additional nuisance parameter $\eta \in \mathbb{R}^{d_0}$ (which may be vacuous). For simplicity we assume that the two metrics d_n and e_n on the parameters (w, η), which should satisfy the testing assumption (8.17), are identical, and write it as d_n. This notation should be discriminated from the metric d on the set of functions w introduced in (10.23). For instance, in the case that $X^{(n)}$ is a vector of n independent observations, the metric d_n can be taken equal to the root average squared Hellinger distance $d_{n,H}$, defined in (8.25), as shown in Theorem 8.23.

We equip the function w with the finite random series prior as considered previously, and independently equip η with an additional prior.

Theorem 10.21 (Adaptation, finite random series) *Let (A1)–(A3) hold with* $\log A(j) \asymp -j(\log j)^{t_1}$ *and* $\log B(j) \asymp -j(\log j)^{t_2}$, *for* $t_1 \geq t_2 \geq 0$, *and let* (10.23) *hold with* $\rho(j) = j^k$, *for some* $k \geq 0$. *Let* $\epsilon_n \geq \bar{\epsilon}_n$ *be sequences of positive numbers with* $\epsilon_n \to 0$ *and* $n\bar{\epsilon}_n^2 \to \infty$, *and let* J_n, $\overline{J}_n \geq 3$, $M_n > 0$ *be such that, for some sequence* $b_n \to \infty$,

$$\inf_{\theta \in [-H,H]^{\overline{J}_n}} d(w_0, \theta^\top \psi_{\overline{J}_n}) \leq \bar{\epsilon}_n, \tag{10.24}$$

$$\overline{J}_n (\log \overline{J}_n)^{t_1 \vee 1} + \overline{J}_n \log_- \bar{\epsilon}_n \lesssim n\bar{\epsilon}_n^2, \tag{10.25}$$

$$J_n \log \frac{J_n M_n n}{\epsilon_n} \lesssim n\epsilon_n^2, \tag{10.26}$$

$$J_n (\log J_n)^{t_2} \geq b_n n\bar{\epsilon}_n^2, \tag{10.27}$$

$$\log J_n + b_n n\bar{\epsilon}_n^2 \leq R(M_n). \tag{10.28}$$

Suppose that the prior on η *satisfies, for some* $\mathcal{H}_n \subset \mathbb{R}^{d_0}$,

$$\Pi(\eta\colon \|\eta - \eta_0\| < \bar{\epsilon}_n) \geq e^{-n\bar{\epsilon}_n^2}, \tag{10.29}$$

$$\log_+ \operatorname{diam}(\mathcal{H}_n) \leq n\epsilon_n^2, \tag{10.30}$$

$$\Pi(\eta \notin \mathcal{H}_n) \leq e^{-b_n n\bar{\epsilon}_n^2}. \tag{10.31}$$

Furthermore, assume that $d_n = e_n$ *is a metric for which tests as in* (8.17) *exist, and such that, for some* $a, c > 0$ *and every* $w_1, w_2 \in \mathcal{W}_{J_n,M_n}$ *and* $\eta_1, \eta_2 \in \mathcal{H}_n$,

$$n^{-c}\, d_n((w_1, \eta_1), (w_2, \eta_2)) \lesssim d(w_1, w_2)^a + \|\eta_1 - \eta_2\|^a, \tag{10.32}$$

and for all w, η *such that* $d(w_0, w)$ *and* $\|\eta - \eta_0\|$ *are sufficiently small,*

$$n^{-1} K(P^{(n)}_{w_0,\eta_0}; P^{(n)}_{w,\eta}) \lesssim d^2(w_0, w) + \|\eta - \eta_0\|^2. \tag{10.33}$$

Then $P^{(n)}_{w_0,\eta_0} \Pi_n(d_n((w, \eta), (w_0, \eta_0)) > K_n\epsilon_n) \to 0$, *for every* $K_n \to \infty$.

Proof We verify the conditions of Theorems 8.19 and 8.20, with $\Theta_{n,1} = \mathcal{W}_{J_n,M_n} \times \mathcal{H}_n$ and $\Theta_{n,2}$ its complement, where $\mathcal{W}_{J_n,M_n}$ is defined in Lemma 10.20.

The latter lemma combined with (10.24) gives that $\Pi(\theta^\top\psi: d(w_0, \theta^\top\psi) \leq 2\bar\epsilon_n) \geq e^{-c_3 n\bar\epsilon_n^2}$, in view of (10.25), for some constant $c_3 > 0$. The prior of η satisfies a similar lower bound by assumption (10.29). The prior independence of w and η and (10.33) show that the simplified prior mass condition (8.22) is satisfied with ϵ_n in the latter condition taken equal to a multiple of the sequence $\bar\epsilon_n$. This implies condition (i) of Theorem 8.19.

Lemma 10.20 and condition (10.26) give the bound $\log D(n^{-c/a}\epsilon_n^{1/a}, \mathcal{W}_{J_n,M_n}, d) \lesssim n\epsilon_n^2$. Condition (10.30) and again (10.26) give a similar bound on $\log D(n^{-c/a}\epsilon_n^{1/a}, \mathcal{H}_n, \|\cdot\|)$. Combined with (10.32) this verifies the global entropy condition (8.23), with $e_n = d_n$ and ϵ_n replaced by a multiple, and hence the local entropy condition (ii) of Theorem 8.19.

Lemma 10.20 and conditions (10.27) and (10.28) show that $\Pi(\mathcal{W}_{J_n,M_n}^c) \leq 2e^{-b_n n\bar\epsilon_n^2}$. A similar bound on the remaining mass for the prior on η holds by (10.31). Together with the simplified prior mass condition (8.22) found previously, this verifies condition (iii) as given in Theorem 8.20. □

Conditions (10.32) and (10.33) relate the statistical problem to the metric structure of the parameter space. They are verified for several examples below. It is notable that condition (10.33) requires a correspondence between the Kullback-Leibler divergence (a square statistical discrepancy) and the square distance d^2, whereas the condition (10.32) allows any polynomial relation between the statistical distance d_n and the distance d, and even an extra factor n^{-c}. A technical explanation for this difference is that the entropy of the present finite-dimensional models is mostly driven by their dimensions, in the form of estimates $J_n \log(1/\epsilon)$, with J_n the dimension. Relation (10.32) is employed to bound the d_n-entropy of the statistical models in terms of the d-entropy of the parameter space, but the nature of the relation between the metrics (if at least polynomial) will affect the end result only through the logarithmic factor $\log(1/\epsilon)$. (Sharp contraction results, without logarithmic factors, would require a more precise analysis, and possibly different priors, as discussed in Sections 10.3.4 and Section 9.1.)

It is helpful that the more stringent of the two conditions is needed only at the true parameter, where the Kullback-Leibler divergence is often better behaved.

Conditions (10.29)–(10.31) concern the prior on the parameter η, and do not usually play a determining role for the contraction rate.

In many situations a best approximation of a true function w_0 by a linear combination of J basis functions satisfies, for some $\alpha > 0$ and every $J \in \mathbb{N}$,

$$\inf_{\theta\in\mathbb{R}^J} d(w_0, \theta^\top\psi_J) \lesssim J^{-\alpha}. \tag{10.34}$$

The parameter α depends on w_0 and can be considered the regularity level of this function, relative to the basis functions. If the coordinates of the minimizing vectors θ can be limited to the fixed interval $[-H, H]$ of condition (A2), then (10.24) will be satisfied for $\bar\epsilon_n \geq \overline{J}_n^{-\alpha}$. In many examples M_n and $R(M_n)$ can be taken (large) powers of n. It may be verified that inequalities (10.24)–(10.28) are then satisfied for the choices, for $\bar t_1 = t_1 \vee 1$,

$$\overline{J}_n \asymp (n/\log^{\bar t_1} n)^{1/(2\alpha+1)}, \qquad J_n \gg n^{1/(2\alpha+1)}(\log n)^{2\alpha/(2\alpha+1)\bar t_1 - t_2}$$
$$\bar\epsilon_n \asymp (n/\log^{\bar t_1} n)^{-\alpha/(2\alpha+1)}, \qquad \epsilon_n \gg n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)\bar t_1 - t_2/2 + 1/1}.$$

The maximum of $\bar\epsilon_n$ and ϵ_n is the posterior contraction rate under the preceding theorem; it misses the usual rate $n^{-\alpha/(2\alpha+1)}$ by only a logarithmic factor.

The preceding applies to functions on univariate and multivariate domains alike, albeit that when α is taken to be the usual smoothness of a function of d variables (e.g. the number of bounded derivatives), then the index α in (10.34) should be replaced by α/d. In terms of the usual smoothness index, the resulting rate is then $n^{-\alpha/(2\alpha+d)}$ times a logarithmic factor. A function $w_0 \colon [0,1]^d \to \mathbb{R}$ may possess different regularity levels in its d coordinates. For such an *anisotropic* function, an approximation exploiting different dimensions for the different coordinates is appropriate. If $\alpha_1, \ldots, \alpha_d$ are the regularity levels, with $\alpha^* = d/(\sum_{k=1}^d \alpha_k^{-1})$ their *harmonic mean*, and $J_n(k)$ is the dimension used for the kth coordinate, a typical approximation rate takes the form

$$\inf_{\theta \in \mathbb{R}^{J_n(1)} \times \cdots \times \mathbb{R}^{J_n(d)}} d(w_0, \theta^\top \psi_J) \lesssim \sum_{k=1}^d J_n(k)^{-\alpha_k}.$$

For $J_n = \prod_{k=1}^d J_n(k)$ the overall dimension, optimal choices satisfying inequalities (10.26)–(10.25) are then

$$J_n(k) \asymp (n/\log n)^{(\alpha^*/\alpha_k)/(2\alpha^*+d)}, \quad \overline{J}_n = \prod_{k=1}^d J_n(k) \asymp \bar{\epsilon}_n^{-d/\alpha^*},$$
$$\epsilon_n \asymp (n/\log n)^{-\alpha^*/(2\alpha^*+d)} (\log n)^{(1-t_1)/2}.$$

Example 10.22 (Classical regularity bases) Relation (10.34) and the conclusion of the preceding paragraph applies to most of the classical regularity spaces: Hölder functions $w_0 \in \mathfrak{C}^\alpha[0,1]$ in combination with classical Fourier series (with d the $\mathbb{L}_2$- or the uniform distance; see Jackson 1912), polynomials (with d the uniform norm; see Hesthaven et al. 2007)), B-splines (with d the $\mathbb{L}_2$ or supremum metric; see Section E.2), and wavelets (see Cohen et al. 1993 and Section E.3).

In all examples tensor products of the basis functions can be used to approximate functions on a multi-dimensional domain. If $w_0 \in \mathfrak{C}^\alpha([0,1]^d)$, then relation (10.34) is again valid, but with α replaced by α/d (e.g. see Lemma E.7 for multivariate B-splines).

The approximation rate of such tensor products for a function w_0 in an anisotropic Hölder class of functions on $[0,1]^d$, with regularity levels $\alpha_1, \ldots, \alpha_d$ (see (E.15)), is $\sum_{k=1}^d J(k)^{-\alpha_k}$, when the tensors are formed by using the first $J(k)$ basis functions for the kth coordinate, for $k = 1, \ldots, d$ (see Section E.2 for a precise statement for tensor B-splines). This leads to the rates described previously, with the harmonic mean of the regularity levels replacing α.

Example 10.23 (Bernstein polynomials) Bernstein polynomials, as in Example 5.10, possess the slower approximation rate of $J^{-\alpha/2}$, for $\alpha \leq 2$ (see Lemma E.3). This leads to the same rates as in the preceding discussion, but with α replaced by $\alpha/2$ throughout. In particular, the contraction rate, even though dependent on α, will be suboptimal for all α. For $\alpha \leq 1$, this can be repaired using the coarsened Bernstein polynomials, as in Lemma E.4. This is similar to adaptation by the Dirichlet priors considered in Section 9.3.

10.4.1 Density Estimation

A linear combination of basis functions is not naturally nonnegative or integrates to one. However, a finite random series prior $\theta^\top \psi$ can be transformed into a prior for a probability density by applying a link function $\Psi : \mathbb{R} \to (0, \infty)$ followed by renormalization, giving the prior

$$\frac{\Psi(\theta^\top \psi)}{\int \Psi(\theta^\top \psi)\, d\mu}.$$

The exponential link function $\Psi(x) = e^x$ is particularly attractive, as the statistical distances and discrepancies of the transformed densities relate to the uniform norm on the functions $\theta^\top \psi$, as shown in Lemma 2.5, but other link functions may do as well. (Link functions such that $\log \Psi$ is Lipschitz behave exactly as the exponential link, as shown in Problem 2.4; for properties of other link functions see Problem 2.5.) The special case of the B-spline basis with the exponential link function is discussed in detail in Section 10.3.4.

For nonnegative basis functions employing a link function is unnecessary if the linear combinations are restricted to vectors θ with positive coordinates, and the prior can be constructed through renormalization only, giving $\theta^\top \psi / \theta^\top c$, for $c = \int \psi\, d\mu$. It seems that in general the approximation (10.34) might suffer from a restriction to nonnegative coefficients, as basis functions may need to partially cancel each other for good approximation, but for localized bases, such as wavelets or B-splines, such a problem should not occur. For the B-spline basis, this is verified in Lemma E.5, part (d), under the assumption that the true function is bounded away from zero. The B-spline basis, normalized to integrate to one, has the further elegant property that $\int \theta^\top B^*_{J,j}(x)\, dx = \sum_j \theta_j$, so that renormalization becomes unnecessary if the parameter vector θ is further restricted to the unit simplex $\mathbb{S}_J$. A Dirichlet prior on θ has the further attraction of allowing an analytic expression for posterior moments (see Problem 10.11).

The preceding observations are valid for functions on a univariate and on multivariate domains alike. The domains would typically have to be compact to ensure the approximation property (10.34).

10.4.2 Nonparametric Normal Regression

Consider independent observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ following the regression model $Y_i = w(X_i) + \varepsilon_i$, where $\varepsilon_i \overset{\text{iid}}{\sim} \text{Nor}(0, \sigma^2)$, and w and σ are unknown parameters. The covariates $X_1, \ldots, X_n$ can be either fixed or random; in the latter case they are assumed i.i.d. and independent of the errors.

We model the regression function w by a finite random series prior $\theta^\top \psi$, and put an independent prior distribution on σ. We assume that for any $\sigma_0 > 0$, there exist positive numbers t_4, t_5 and t_6 such that for sufficiently small $\sigma_1, \sigma_2, \sigma_3 > 0$,

$$\Pi(|\sigma - \sigma_0| < \sigma_3) \gtrsim \sigma_3^{t_4},$$
$$\Pi(\sigma < \sigma_1) + \Pi(\sigma > \sigma_2^{-1}) \lesssim \exp(-\sigma_1^{-t_5}) + \sigma_2^{t_6}.$$

For instance, these conditions hold if some positive power of σ is inverse-gamma distributed. The first condition (easily) ensures (10.29) as soon as $n\bar{\epsilon}_n^2$ is some power of n, while the second makes (10.31) valid when the sieve for σ is taken equal to $\mathcal{H}_n = (n^{-1/t_5}, e^{n\bar{\epsilon}_n^2/t_6})$.

The random design model is an i.i.d. model, and hence the Hellinger distance on the law of the observations is a possible choice for the metric d_n. For known σ this distance is bounded above by the $\mathbb{L}_2(G)$-distance on the regression functions, for G the marginal distribution of the design points, while the Kullback-Leibler discrepancies are equal to the square of this distance, by Lemma 2.7. Relations (10.32) and (10.33) are then verified for d the $\mathbb{L}_2(G)$-distance, or a stronger distance, such as the uniform distance. For unknown σ the situation is more complicated, but Lemma 2.7(i) shows that for σ restricted to the sieve $\mathcal{H}_n$, so that $\sigma \geq n^{-1/t_5}$, the Hellinger distance is bounded above by a multiple of n^{1/t_5} times the sum of the $\mathbb{L}_2(G)$ distance and the Euclidean distance on σ. Thus relation (10.32) is still valid in the weaker form with $c = 1/t_5$. In the neighborhood of the true parameter (f_0, σ_0) the Kullback-Leibler divergence still translates by Lemma 2.7(ii) into the sum of the $\mathbb{L}_2(G)$- and the Euclidean metric, whence (10.33) remains valid. Thus under the conditions of Theorem 10.21 for the finite series prior we obtain a rate of contraction relative to the Hellinger distance on the densities of the observations. If the prior would be restricted to uniformly bounded regression functions, or the posterior could be shown to concentrate on such functions, then this translates into a contraction rate relative to the $\mathbb{L}_2(G)$-distance on the regression functions, but in general this requires additional arguments and possibly conditions.

For the fixed design model we can reach the same conclusion, but with metric d_n taken equal to the root average square Hellinger distance $d_{n,H}$ and with the empirical distribution G_n of the design points replacing the marginal distribution G. (It may then be attractive to bound this by the uniform distance, which does not change with n.) The rate of contraction will then also be relative to $d_{n,H}$, which entails convergence to zero of

$$\frac{1}{n}\sum_{i=1}^{n} d_H^2(\mathrm{Nor}(f_0(x_i), \sigma_0^2), \mathrm{Nor}(f(x_i), \sigma^2))$$

$$= 2\Big[1 - \sqrt{1 - \frac{(\sigma_0 - \sigma)^2}{\sigma_0^2 + \sigma^2}} \int e^{-(f_0 - f)^2/(4\sigma_0^2 + 4\sigma^2)}\, dG_n\Big].$$

This readily gives consistency and a contraction rate for σ, but as in the random design case translates into a rate relative to the more natural $\mathbb{L}_2(G_n)$-metric only under additional conditions. For the fixed design case a better alternative it to not employ $d_{n,H}$. For known σ, it is shown in Section 8.3.2 that the general contraction theorem is valid when $d_n = e_n$ is taken equal to the $\mathbb{L}_2(G_n)$-norm of the regression functions. Lemma 8.27 in the same section shows that this extends to the case of unknown σ if this is restricted to a bounded interval. Since we already concluded that the posterior distribution of σ is consistent for σ_0, the latter restriction can be made without loss of generality. Thus in the fixed design case we obtain under the conditions of Theorem 10.21 for the finite series prior (relative to the uniform norm on the series) a rate of contraction relative to the $\mathbb{L}_2(G_n)$-distance on the regression functions.

10.4.3 Nonparametric Binary Regression

Consider independent observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ in the binary regression model discussed in Section 2.5. Thus Y_i takes is values in $\{0, 1\}$, and given X_i is Bernoulli distributed with success probability $p(X_i)$, for a given function p. The covariates $X_1, \ldots, X_n$ can be either i.i.d. or fixed values in a set $\mathcal{X}$.

For a given link function $H: \mathbb{R} \to (0, 1)$ and vector of basis functions ψ, we model the success probability p through a prior of the form $H(\theta^\top \psi)$.

The logistic link function is convenient as it allows a direct relationship between the Hellinger distance, Kullback-Leibler divergence, and the $\mathbb{L}_2$-norm of the functions $\theta^\top \psi$: by Lemma 2.8(ii)–(iv), relations (10.32) and (10.33) hold (in its most natural form with $c = 0$ and $a = 2$) for d the $\mathbb{L}_2(G)$-distance, where G is the marginal distribution of a covariate X_i in the case of a random design, and the empirical distribution of the design points given fixed design. Theorem 10.21 therefore gives adaptive contraction rates relative to the Hellinger distance on the densities of the observations as defined in (2.6). In view of Problem B.2(i)+(iii), this also implies the same contraction rates relative to the $\mathbb{L}_1$ and Hellinger distances on the response probability function p relative to G.

Other link functions are possible as well. For instance, by Lemma 2.8(ii) the Hellinger distance is bounded above by the $\mathbb{L}_2(G)$-distance for any link function with finite Fisher information for location, while by part (i) of the lemma and Lemma B.1(ii) the Hellinger distance is bounded above by the root of the $\mathbb{L}_1(G)$-distance for the slightly larger class of all Lipschitz link functions. This readily gives relation (10.32) for d one of these distances, with $a = 1/2$ in the second case. A fortiori the relation is true for d the uniform distance. For the latter choice we can also retain (10.33), at least when the true regression function is bounded away from zero and one. Indeed a function in a uniform neighborhood of this true function is then also bounded and for bounded functions the Kullback-Leibler divergence of the observations are upper bounded by the uniform distance, by the last assertion of Lemma 2.8.

For $\mathcal{X} = (0, 1)$ the B-spline basis may be used with the identity link function, and the coordinates of the parameter vector θ restricted to the interval $(0, 1)$. In view of part (c) of Lemma E.5 this restriction does not affect the approximation of the splines. Problem 10.11 suggests that independent beta priors on the coefficients allow an explicit expression of posterior moments.

10.4.4 Nonparametric Poisson Regression

Consider independent observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ in the Poisson regression model discussed in Section 2.6. Thus $Y_i \mid X_i \overset{\text{ind}}{\sim} \text{Poi}(H(w(X_i))$, for a fixed link function $H: \mathbb{R} \to (0, \infty)$ and an unknown function $w: \mathcal{X} \to \mathbb{R}$. The covariates $X_1, \ldots, X_n$ can be either i.i.d. or fixed values in the set $\mathcal{X}$; denote by G and G_n the marginal density or the empirical distribution of $X_1, \ldots, X_n$ in these two cases.

We model the conditional mean value function $H(w)$ through a prior of the form $H(\theta^\top \psi)$.

For a link function such that the function $H'/\sqrt{H}$ is bounded, the Hellinger distance on the densities p_w of an observation (X_i, Y_i) in the random design model is bounded above by the $\mathbb{L}_2(G)$-distance on the functions w, by Lemma 2.9(i). Furthermore, if the true mean function $H(w_0)$ is bounded, then part (ii) of the lemma shows that the Kullback-Leibler

divergence is bounded above by a multiple of the square of the $\mathbb{L}_2(G)$-distance. This verifies relations (10.32) (in its most natural form with $c = 0$ and $a = 2$ and vacuous η) and (10.33) for d the $\mathbb{L}_2(G)$-distance. Theorem 10.21 therefore gives adaptive contraction rates relative to the Hellinger distance on the densities of the observations. If the design is fixed rather than random, the same reasoning, but using the root average square Hellinger distance, leads to the same conclusion, with G_n taking the place of G.

Link functions such that the function $H'/\sqrt{H}$ is only locally bounded, such as the exponential link function, are not covered by the preceding paragraph. We may still reach the same conclusion if $|\sqrt{H}(\theta_1^\top\psi) - \sqrt{H}(\theta_2^\top\psi)| \leq n^c\|(\theta_1 - \theta_2)^\top\psi\|_\infty$, for some constant $c \geq 0$ for every $\|\theta_1\|, \|\theta_2\| \leq M_n$, for M_n a positive sequence satisfying $R(M_n) \leq bn$ for some $b > 0$ (and R the function in (A2)). This inequality will then replace Lemma 2.9(i) and ensures (10.32) with d the uniform distance on the functions $\theta^\top\psi$ (and $a = 1$ and vacuous η), while part (ii) of the lemma still verifies (10.33) with this distance d as long as the true mean function $H(w_0)$ is bounded.

Contraction relative to the Hellinger distance on the densities p_w does not generally imply contraction relative to (other) natural distances on $H(w)$. However, the same contraction rate will be implied for the Hellinger distance $\|\sqrt{H(w)} - \sqrt{H(w_0)}\|_{2,G}$ (or G_n) on the mean functions if it can be shown that the posterior distribution concentrates asymptotically all its mass on a uniformly bounded set of functions (i.e. $\Pi_n(\|H(w) - H(w_0)\|_\infty > B|\, X_1, Y_1, \ldots, X_n, Y_n)$ tends to zero in probability for some B) . This follows because the Hellinger distance between two Poisson measures is bounded below by a constant times the distance between the roots of their means, where the constant is uniform in bounded means. There are at least two ways to meet the extra condition — showing that the posterior is consistent for the uniform distance on $H(w)$, or ensuring that the prior on $H(w)$ charges only functions that are bounded away from zero and infinity.

If $\mathfrak{X}$ is the unit interval and the B-spline basis is used for the prior $\theta^\top\psi$, then we can use the identity function as the link function and restrict the coefficients θ to the positive half line, in view of in view of part (b) of Lemma E.5. Problem 10.11 suggests an explicit expression of the Bayes estimator if these coefficients are given independent gamma priors. This remark extends to multivariate covariates and tensor products of B-splines.

10.4.5 Functional Regression

In the (linear) *functional regression* model the covariates X and the parameter w are functions $X\colon \mathfrak{T} \to \mathbb{R}$ and $w\colon \mathfrak{T} \to \mathbb{R}$ on a set $\mathfrak{T} \subset \mathbb{R}^d$, and the usual linear regression is replaced by the integral $\int_{\mathfrak{T}} X(t)w(t)\,dt$, which we observe with a random error. Given Gaussian errors $\varepsilon_i \overset{\text{iid}}{\sim} \text{Nor}(0, \sigma^2)$, for $i = 1, \ldots, n$, the observations are an i.i.d. sample of pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ following the model

$$Y_i = \int_{\mathfrak{T}} X_i w\,d\lambda + \varepsilon_i. \tag{10.35}$$

Induce a prior on w through an expansion $w = \theta^\top\psi$. For simplicity we consider σ as fixed; alternatively it can be treated as in Section 10.4.2. Assume that $\text{E}\int_{\mathfrak{T}} X^2\,d\lambda < \infty$ and write $\|\cdot\|_2$ for the $\mathbb{L}_2$-norm relative to the Lebesgue measure λ on $\mathfrak{T}$.

The square Hellinger distance and Kullback-Leibler divergence on the laws P_w of a single observation (X_i, Y_i) satisfy, by the Cauchy-Schwartz inequality,

$$d_H^2(P_v, P_w) \le \frac{1}{2\sigma^2}\mathrm{E}_X\Big(\int_{\mathfrak{T}} X\,(v-w)\,d\lambda\Big)^2 \lesssim \|v-w\|_2^2,$$
$$K(P_v; P_w) = \frac{1}{2\sigma^2}\mathrm{E}_X\Big(\int_{\mathfrak{T}} X(v-w)\,d\lambda\Big)^2 \lesssim \|v-w\|_2^2.$$

Therefore (10.32) and (10.33) are verified for the $\mathbb{L}_2$-distance, and under the conditions of Theorem 10.21 we obtain an adaptive rate of contraction relative to the Hellinger distance on P_w.

Instead of the preceding model we can also consider a longitudinal version, in which instead of Y_i we observe $Y_i(T_i) = X_i(T_i)w(T_i) + \varepsilon_i$, for $T_1, \ldots, T_n$ i.i.d. random points in $\mathfrak{T}$, independent of $X_1, \ldots, X_n, \varepsilon_1, \ldots, \varepsilon_n$. The statistical distances for this model are identical to the preceding model and hence we can draw the same conclusions.

10.4.6 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.

As a prior model for the spectral density we set $f = H(\theta^\top \psi)$ for a finite series prior $\theta^\top \psi$ and a given monotone link function $H : \mathbb{R} \to (0, \infty)$, which should have the property that $\log H$ is Lipschitz on bounded intervals.

As the observations are independent, the root average square Hellinger distance $d_{n,H}$ is the natural candidate for the metric d_n in the application of Theorem 10.21. In Section 7.3.3 this is seen to be equivalent to the root of

$$e_n(f_1, f_2)^2 = \frac{1}{\nu}\sum_{i=1}^{\nu} \frac{(f_1(\omega_l) - f_2(\omega_l))^2}{(f_1(\omega_l) + f_2(\omega_l))^2}.$$

Because $\log u - \log v = \int_u^v s^{-1}\,ds \ge \frac{1}{2}(v-u)/(v+u)$, for any $0 < u < v$, we see that

$$d_n(\theta_1^\top \psi, \theta_2^\top \psi) := e_n(H(\theta_1^\top \psi), H(\theta_2^\top \psi)) \le \|\log H(\theta_1^\top \psi) - \log H(\theta_2^\top \psi)\|_\infty.$$

Assume that the right side is bounded above by $n^c \|(\theta_1 - \theta_2)^\top \psi\|_\infty$, for some constant $c \ge 0$ and every $\|\theta\|_\infty \le M_n$, for some sequence $M_n > 0$ satisfying $R(M_n) \le bn$ for some $b > 0$ and R as in (A2). Then (10.32) is verified with d taken equal to the uniform distance on the functions $\theta^\top \psi$.

In Section 7.3.3 the Kullback-Leibler divergence between exponential densities is bounded above by a multiple of the square of the difference of their means, if the means

are bounded away from zero. If we assume that the true spectral density $f_0 = H(w_0)$ is bounded away from zero, then this gives, for every w such that $\|w_0 - w\|_\infty$ is small enough,

$$\frac{1}{\nu}\sum_{l=1}^{\nu} K(P_{w_0,l};\, P_{w,l}) \lesssim \|\log H(w_0) - \log H(w)\|_\infty^2 \lesssim \|w - w_0\|_\infty^2,$$

in view of the assumed Lipschitz continuity of $\log H$. Hence (10.33) is verified for the uniform distance on w.

The root average square Hellinger distance is perhaps not the most natural distance on the spectral densities. On the set of spectral densities that are bounded away from zero, we have that $e_n(f_1, f_2) \gtrsim \nu^{-1}\sum_{l=1}^{\nu} \big|\log f_1(\omega_l) - \log f_2(\omega_l)\big|$. (This follows, because the inequality involving the Hellinger distance between exponential distributions used in the preceding can be reversed if the means are bounded away from zero.) Thus if the posterior probability of the event $\{\|\log f_0 - \log f\|_\infty > B\}$ tends to zero for some B, then the contraction rate with respect to e_n implies the same contraction rate for the average $\mathbb{L}_1$-distance on the log spectral densities. The additional condition certainly holds if the posterior is consistent for the uniform distance on $\log f$, or if the prior on f charges only functions that are bounded away from zero and infinity.

The B-spline basis may be used with the reciprocal function as the link and the coefficients restricted to the positive half line. Problem 10.11 suggests explicit expressions of posterior moments when using independent gamma priors on the coefficients.

10.5 Model Selection Consistency

When multiple models are under consideration one may ask whether the posterior distribution can identify the "correct model" from the given class of models. Within our asymptotic framework this is equivalent to asymptotic contraction of the posterior distribution of the model index to the "true index." However, a useful precision of this idea is not so straightforward to formulate. This is true in particular for nonparametric model, where a "true model" is often defined only by approximation. It may not be unique, and the true density may not belong to any of the models under consideration.

The issue of consistency of model selection is connected to adaptation, in that consistency for model selection will force rate adaptation. This is clear from representation (10.1) of the posterior distribution. Model selection consistency would entail that the posterior $\lambda(\cdot|\, X^{(n)})$ on the model index contract to the "true index" β, whence the full posterior distribution will be asymptotic to the posterior $\Pi_{n,\beta}(\cdot|\, X^{(n)})$ if only the true model had been used. However, this argument cannot be reversed: adaptation may well occur without consistency of model selection. The main reason is that a "true model" may not exist, as noted previously. Under the (reasonable) conditions of the main theorem of this chapter, the model index prior will choose models of low complexity given equal approximation properties, but may choose a model of too low complexity if this approximates the true parameter well enough.

In the situation that there are precisely two models $\mathcal{P}_0$ and $\mathcal{P}_1$ with equal prior weights, model selection consistency can be conveniently described by the behavior of the *Bayes factor* of the two models, the quotient of the marginal likelihoods of the two models. The posterior distribution is consistent for model selection if and only if the Bayes factor of $\mathcal{P}_0$

versus $\mathcal{P}_1$ has a dichotomous behavior: it tends to infinity if the true density p_0 belongs to $\mathcal{P}_0$, and it converges to zero otherwise, so-called *Bayes factor consistency*. This consistency property is especially relevant for Bayesian goodness of fit testing against a nonparametric alternative. A computationally convenient method of such goodness of fit test using a mixture of Pólya tree prior on the nonparametric alternative is described in Problem 10.16.

10.5.1 Testing a Point Null

Model selection with a singleton model is essentially a problem of testing goodness of fit. It is natural to formulate it as testing a null hypothesis $H_0\colon p = p_*$ against the alternative $H_1\colon p \neq p_*$. If $\lambda \in (0, 1)$ is the prior weight of the null model and Π_1 is the prior distribution on p under the alternative model, then the overall prior is the mixture $\Pi = (1-\lambda)\delta_{p_*} + \lambda\Pi_1$. The posterior probability of the null model and the Bayes factor are given by

$$\Pi_n(p = p_* | X_1, \ldots, X_n) = \frac{(1-\lambda)\prod_{i=1}^n p_*(X_i)}{(1-\lambda)\prod_{i=1}^n p_*(X_i) + \lambda \int \prod_{i=1}^n p(X_i)\, d\Pi_1(p)}$$

and

$$\mathbb{B}_n = \frac{\prod_{i=1}^n p_*(X_i)}{\int \prod_{i=1}^n p(X_i)\, d\Pi_1(p)}.$$

Theorem 10.24 *As as* $n \to \infty$,

(i) $\mathbb{B}_n \to \infty$ *a.s.* $[P_*^\infty]$.
(ii) $\mathbb{B}_n \to 0$ *a.s.* $[P^\infty]$ *for any* $p \neq p_*$ *with* $p \in \mathrm{KL}(\Pi_1)$.

Proof The Bayes factor tends to 0 or ∞ if and only if the posterior probability $\Pi_n(p = p_* | X_1, \ldots, X_n)$ of the null hypothesis tends to 0 or 1.

Since the prior mass of $\{p_*\}$ is positive, Doob's theorem, Theorem 6.9, applied with the function $f(p) := \mathbb{1}_{\{p_0\}}(p)$, gives that $\Pi_n(p = p_* | X_1, \ldots, X_n) \to 1$ a.s. under the null hypothesis. This implies (i).

If $p \in \mathrm{KL}(\Pi_1)$, then also $p \in \mathrm{KL}(\Pi)$ and hence $\Pi_n(\mathcal{W} | X_1, \ldots, X_n) \to 1$ a.s. $[P^\infty]$ for any weak neighborhood $\mathcal{W}$ of p, by Schwartz's theorem and Example 6.20. If $p \neq p_*$, then the neighborhood can be chosen to exclude p_* and hence the posterior mass of the null hypothesis tends to zero. This proves (ii). □

The role of Doob's theorem in giving a painless proof for the relatively complicated issue of Bayes factor consistency is intriguing. It is not difficult to see that the result goes through when the null hypothesis is a countable closed set in which every point receives positive prior mass. (Also Schwartz's theorem is not needed if the alternative hypothesis has a similar form.) Unfortunately, the method of proof does not seem to generalize any further.

10.5.2 General Case

Consider the setting of Section 10.2, where we consider a collection of models $\mathcal{P}_{n,\alpha}$, for $\alpha \in A_n$, for the density of i.i.d. observations $X_1, X_2, \ldots$ We retain the notations $A_{n,\prec\beta_n}$ and

$A_{n,\succeq\beta_n}$, and, as before, consider a distance d on the set of densities for which tests as in (8.2) exist.

Somewhat abusing notation, we denote the posterior distribution of the model index by the same symbol as the posterior distribution: for any set $B \subset A_n$ of indices,

$$\Pi_n(\alpha \in B | X_1, \ldots, X_n) = \frac{\sum_{\alpha\in B} \lambda_{n,\alpha} \int \prod_{i=1}^n p(X_i)\, d\Pi_{n,\alpha}(p)}{\sum_{\alpha\in A_n} \lambda_{n,\alpha} \int \prod_{i=1}^n p(X_i)\, d\Pi_{n,\alpha}(p)}.$$

Theorem 10.25 *Under the conditions of Theorem 10.3,*

$$P_0^n \Pi_n(\alpha \in A_{n,\prec\beta_n} | X_1, \cdots, X_n) \to 0, \tag{10.36}$$

$$P_0^n \Pi_n(\alpha \in A_{n,\succeq\beta_n} : d(p_0, \mathcal{P}_{n,\alpha}) > M\sqrt{H}\,\epsilon_{n,\beta_n} | X_1, \cdots, X_n) \to 0. \tag{10.37}$$

Under the conditions of Theorem 10.7 assertion (10.36) *is satisfied and* (10.37) *holds with* $M\sqrt{H}$ *replaced by* I_n*, for any* $I_n \to \infty$.

Proof Theorems 10.3 and 10.7 show that the posterior concentrates all its mass on balls of radius $M\sqrt{H}\epsilon_{n,\beta_n}$ or $I_n\epsilon_{n,\beta_n}$ around p_0, respectively. Consequently the posterior cannot charge any model that does not intersect these balls and (10.37), or its adapted version under the conditions of Theorem 10.7, follows.

The first assertion can be proved using exactly the proof of Theorems 10.3 and 10.7, except that the references to $\alpha \succeq \beta_n$ can be omitted. In the notation of the proof of Theorem 10.3, we have that

$$\bigcup_{\alpha\prec\beta_n} \mathcal{P}_{n,\alpha} \subset \Big(\bigcup_{\alpha\prec\beta_n} \bigcup_{i\geq I} \mathcal{S}_{n,\alpha,i} \Big) \bigcup \Big(\bigcup_{\alpha\prec\beta_n} C_{n,\alpha}(p_0, IB'\epsilon_{n,\alpha}) \Big).$$

It follows that $P_0[\Pi_n(A_{n,\prec\beta_n} | X_1, \ldots, X_n)(1-\phi_n)\mathbb{1}_{A_n}]$ can be bounded by the sum of the second and third terms on the right-hand side of (10.11), which tend to zero under the conditions of Theorem 10.3, or can be made arbitrarily small under the conditions of Theorem 10.7 by choosing B and I sufficiently large. □

The first assertion of the theorem is pleasing. It can be interpreted as saying that the models that are bigger than the model $\mathcal{P}_{n,\beta_n}$ that contains the true distribution eventually receive negligible posterior weight. The second assertion makes a similar claim about the smaller models, but it is restricted to the smaller models that keep a certain distance to the true distribution. Such a restriction appears not unnatural, as a small model that can represent the true distribution well ought to be favored by the posterior, which looks at the data through the likelihood and hence will judge a model by its approximation properties rather than its parameterization. That big models with similarly good approximation properties are not favored is caused by the fact that (under our conditions) the prior mass on the big models is more spread out, yielding relatively less prior mass near good approximants within the big models.

It is insightful to specialize the theorem to the case of two models, and simplify the prior mass conditions to (10.12). The behavior of the posterior of the model index can then be described through the posterior odds ratio $\mathbb{D}_n = (\lambda_{n,2}/\lambda_{n,1})\mathbb{B}_n$, which is the prior odds

times the Bayes factor

$$\mathbb{B}_n = \frac{\int \prod_{i=1}^n p(X_i)\,\Pi_{n,2}(p)}{\int \prod_{i=1}^n p(X_i)\,\Pi_{n,1}(p)}.$$

Corollary 10.26 *Adopt the notation* (10.3)–(10.4). *Assume that* (10.5) *holds for* $\alpha \in A_n = \{1, 2\}$ *and sequences* $\epsilon_{n,1} > \epsilon_{n,2}$ *such that* $n\epsilon_{n,2}^2 \to \infty$.

(i) *If* $\Pi_{n,1}(B_{n,1}(p_0, \epsilon_{n,1})) \geq e^{-n\epsilon_{n,1}^2}$ *and* $\lambda_{n,2}/\lambda_{n,1} \leq e^{n\epsilon_{n,1}^2}$ *and* $d(p_0, \mathcal{P}_{n,2}) \geq I_n\epsilon_{n,1}$ *for every* n *and some* $I_n \to \infty$, *then* $\mathbb{D}_n \to 0$ *in* P_0^n*-probability.*

(ii) *If* $\Pi_{n,2}(B_{n,2}(p_0, \epsilon_{n,2})) \geq e^{-n\epsilon_{n,2}^2}$ *and* $\lambda_{n,2}/\lambda_{n,1} \geq e^{-n\epsilon_{n,1}^2}$, *and also* $\Pi_{n,1}(C_{n,1}(p_0, M\epsilon_{n,1})) \leq (\lambda_{n,2}/\lambda_{n,1})o(e^{-3n\epsilon_{n,2}^2})$ *for a sufficiently large constant* M, *then* $\mathbb{D}_n \to \infty$ *in* P_0^n*-probability.*

Proof The posterior odds ratio $\mathbb{D}_n$ converges respectively to 0 or ∞ if the posterior probability of model $\mathcal{P}_{n,2}$ or $\mathcal{P}_{n,1}$ tends to zero, respectively. Therefore, we can apply Theorem 10.25 with the same choices as in the proof of Corollary 10.8. □

10.5.3 Testing Parametric versus Nonparametric Models

Suppose that there are two models, with the bigger model $\mathcal{P}_{n,1}$ infinite dimensional, and the alternative model a fixed parametric model $\mathcal{P}_{n,2} = \mathcal{P}_2 = \{p_\theta : \theta \in \Theta\}$, for $\Theta \subset \mathbb{R}^d$, equipped with a fixed prior $\Pi_{n,2} = \Pi_2$. We shall show that the posterior odds ratio $\mathbb{D}_n$ is typically consistent in this situation: $\mathbb{D}_n \to \infty$ if $p_0 \in \mathcal{P}_2$, and $\mathbb{D}_n \to 0$ if $p_0 \notin \mathcal{P}_2$.

If the prior Π_2 is smooth in the parameter and the parameterization $\theta \mapsto p_\theta$ is regular, then, for any $\theta_0 \in \Theta$ where the prior density is positive and continuous, as $\epsilon \to 0$,

$$\Pi_2\Big(\theta : K(p_{\theta_0}; p_\theta) \leq \epsilon^2, V_2(p_{\theta_0}; p_\theta) \leq \epsilon^2\Big) \sim C(\theta_0)\epsilon^d$$

for some positive constant $C(\theta_0)$ depending on θ_0. Therefore, if the true density p_0 is contained in $\mathcal{P}_2$, then $\Pi_{n,2}(B_{n,2}(p_0, \epsilon_{n,2})) \gtrsim \epsilon_{n,2}^d > e^{-n\epsilon_{n,2}^2}$ for $\epsilon_{n,2}^2 = Dn^{-1}\log n$, for $D \geq d/2$ and sufficiently large n.[1]

For this choice of $\epsilon_{n,2}$ we have $e^{n\epsilon_{n,2}^2} = n^D$. Therefore, it follows from (ii) of Corollary 10.26, that if $p_0 \in \mathcal{P}_2$, then the posterior odds ratio $\mathbb{D}_n$ tends to ∞, as $n \to \infty$, as soon as there exists $\epsilon_{n,1} > \epsilon_{n,2}$ such that, for every I,

$$\Pi_{n,1}(p : d(p, p_0) \leq I\epsilon_{n1}) = o(n^{-3D}). \tag{10.38}$$

For an infinite-dimensional model $\mathcal{P}_{n,1}$, relation (10.38) is typically true. In fact, for $p_0 \in \mathcal{P}_{n,1}$ the left-hand side is typically of the order $e^{-Fn\epsilon_{n,1}^2}$, for $\epsilon_{n,1}$ the rate attached to the model $\mathcal{P}_{n,1}$. For a true infinite-dimensional model this rate is not faster than n^{-a} for some $a < 1/2$, leading to an upper bound of the type e^{-n^b} for some $b > 0$, which is certainly $o(n^{-3D})$. If $p_0 \notin \mathcal{P}_{n,1}$, then the prior mass in (10.38) ought to be even smaller.

[1] The logarithmic factor enters because we use the crude prior mass condition (10.12) instead of the comparisons of prior mass in the main theorems, but it does not matter for the purpose of this subsection.

If p_0 is not contained in the parametric model, then typically $d(p_0, \mathcal{P}_2) > 0$, and hence $d(p_0, \mathcal{P}_2) > I_n \epsilon_{n,1}$ for any $\epsilon_{n,1} \to 0$ and sufficiently slowly increasing I_n, as required in (i) of Corollary 10.26. To ensure that $\mathbb{D}_n \to 0$, it suffices that for some $\epsilon_{n,1} > \epsilon_{n,2}$,

$$\Pi_{n,1}\Big(p\colon K(p_0; p) \le \epsilon_{n,1}^2,\, V_2(p_0; p) \le \epsilon_{n,1}^2\Big) \ge e^{-n\epsilon_{n,1}^2}. \tag{10.39}$$

This is the usual prior mass condition (8.4) for obtaining the rate of contraction $\epsilon_{n,1}$ using the prior $\Pi_{n,1}$ on the model $\mathcal{P}_{n,1}$.

We illustrate the preceding in three examples. In each, we assume that the prior model weights are fixed, positive numbers.

Example 10.27 (Bernstein polynomial prior for density estimation) Let $\mathcal{P}_{n,1}$ be the set of densities realized by the Bernstein polynomial prior in Section 9.3, with a geometric or Poisson prior on the index parameter k. The rate of contraction for this model is given by $\epsilon_{n,1} = n^{-1/3}(\log n)^{1/3}$. As the prior spreads its mass over an infinite-dimensional set that can approximate any smooth function, condition (10.38) will be satisfied for most true densities p_0. In particular, if k_n is the minimal degree of a polynomial that is within Hellinger distance $n^{-1/3} \log n$ of p_0, then the left-hand side of (10.38) is bounded by the prior mass of all Bernstein-Dirichlet polynomials of degree at least k_n, which is e^{-ck_n} for some constant c, by construction. Thus the condition is certainly satisfied if $k_n \gg \log n$. Consequently, the Bayes factor is consistent for true densities that are not well approximable by polynomials of low degree.

Example 10.28 (Log-spline prior for density estimation) Let $\mathcal{P}_{n,1}$ be the set of log-spline densities described of dimension $J \asymp n^{1/(2\alpha+1)}$, in Section 10.3.4, equipped with the prior obtained by equipping the coefficients with the uniform distribution on $[-M, M]^J$. The corresponding rate can then be taken to be $\epsilon_{n,1} = n^{-\alpha/(2\alpha+1)}\sqrt{\log n}$, by the calculations in Section 10.3.4. Condition (10.38) can be verified easily by computations on the uniform prior, after translating the distances on the spline densities into the Euclidean distance on the coefficients as in Lemma 9.4.

Example 10.29 (White noise model) Let $\mathcal{P}_{n,1}$ be the set of $\mathrm{Nor}_\infty(\theta, I)$-distributions with $\theta \in \mathfrak{W}^\alpha$, as in Section 10.3.2, with prior given by $\theta_i \overset{\text{ind}}{\sim} \mathrm{Nor}(0, i^{-(2\alpha+1)})$. Relations (10.38)–(10.39) follow from the estimates (10.17) and (10.18). Thus the Bayes factor is consistent for testing $\mathcal{P}_{n,1}$ against finite-dimensional alternatives.

10.6 Historical Notes

Adaptive estimation in a minimax sense was first studied by Efroĭmovich and Pinsker (1984) in the context of the white noise model. Earlier Pinsker (1980) had derived the the minimax risk for squared error loss over ellipsoids in this model. Efroĭmovich and Pinsker (1984) constructed an estimator by adaptively determining optimal damping coefficients in an orthogonal series. There has been considerable work on adaptive estimation in this and many other models since then. The white noise model with the conjugate prior $\theta_i \overset{\text{ind}}{\sim} \mathrm{Nor}(0, \tau_i^2)$,

for $i = 1, 2, \ldots$, was studied by Cox (1993) and Freedman (1999), with particular interest for the posterior distribution of the nonlinear functional $\|\theta - \hat{\theta}\|^2$, where $\hat{\theta}$ is the Bayes estimator. Zhao (2000) noted the minimaxity for ellipsoids and the prior with $\tau_i^2 = i^{-(2\alpha+1)}$, for $i = 1, 2, \ldots$, and also pointed out that this prior gives probability zero to any ellipsoid of order α. She then also studied mixtures of finite-dimensional normal priors to rectify the problem and retain minimaxity. Such priors are also considered by Shen and Wasserman (2001) and Ghosal and van der Vaart (2007a). Bayesian adaptation in this framework was studied by Belitser and Ghosal (2003), with direct methods, for a discrete set of α. Bayesian adaptation for i.i.d. models was studied by Ghosal et al. (2003) in the setting of a finite collection of smoothness indices and log-spline models. They put fixed but arbitrary weights on each model and obtained minimax rate up to a logarithmic factor. Huang (2004) used both log-spline and wavelet series priors and avoided logarithmic factors by carefully constructing the priors on the coefficient spaces, combined with appropriate model weights. Scricciolo (2006) proved adaptation over Sobolev balls using exponential family models, without a logarithmic factor, but with an additional condition on the true parameters. Theorem 10.3 on adaptation in i.i.d. models was obtained by Ghosal et al. (2008), together with the case study on log-spline models presented in Section 10.3.4. The universal weights discussed in Section 10.2.1 were described by Lember and van der Vaart (2007). The applications to priors based on nets and finite-dimensional models given in Sections 10.3.1 and 10.3.1 are taken from their paper. The example of priors on nets, which do not lead to additional logarithmic factors in the rates, suggests that these factors are due to an interaction between the way mass is spread over a model or divided between the models. Section 10.5 is adapted from Shen and Ghosal (2015). Other references on adaptation include Scricciolo (2007, 2014) and van der Vaart (2010). An alternative approach developed in Xing (2008), using the concept of Hausdorff entropy (see Problem 8.13), gives contraction results in the a.s. sense, based on a stronger condition on concentration in terms of a modification of the Hellinger distance. Some of these results are presented in Problems 10.8–10.10 and 10.13. Certain kernel mixture priors are able to adapt automatically to the smoothness level of the true density if augmented by an appropriate prior on a bandwidth parameter. In order for this to happen the mixture class must approximate the true density with increasingly higher accuracy for smoother true densities. The mixture with the true distribution as mixing distribution may not be a close projection into the class of mixtures. Rousseau (2010) devised a technique of constructing better approximations to smoother true densities in the context of a certain type of nonparametric mixtures of beta kernel (see Problem 10.15). In Section 9.4 a similar idea is applied to Dirichlet mixtures of normal distributions. Bayesian model selection in the parametric context is the subject of a rich literature, of which Schwarz (1978) is notable for introducing the Bayesian Information Criterion (BIC) to approximate Bayes factors. The first paper dealing with model selection consistency in the nonparametric context is Dass and Lee (2004), who looked at point null hypotheses (Subsection 10.5.1). Verdinelli and Wasserman (1998) considered testing a uniform density against an infinite-dimensional exponential family, where they showed that the prior satisfies the KL property, and showed by direct calculations that the Bayes factor is consistent. The general case, which requires a different proof, was studied by Ghosal et al. (2008). They also noted the (dis)connection between Bayesian adaptation and model selection consistency. Section 10.5.2 and the examples of Section 10.5.3 are due to Ghosal

et al. (2008). Berger and Guglielmi (2001) developed a method of testing a parametric model using a mixture of Pólya tree prior on the nonparametric alternative (see Problem 10.16).

Problems

10.1 (Belitser and Ghosal 2003) For the white noise model with a prior λ_α on a discrete set of α without limit points in $(0, \infty)$, show that, for any l,

$$\mathrm{E}_{\theta_0}\Pi(\alpha = q_m \mid X^{(n)}) \le \frac{\lambda_m}{\lambda_l} \exp\left[\frac{1}{2}\sum_{i=1}^{\infty} \frac{(i^{-(2q_l+1)} - i^{-(2q_m+1)})(i^{-(2q_l+1)} - \theta_{i0}^2)}{i^{-2(q_l+q_m+1)} + 2n^{-1}i^{-(2q_l+1)} + n^{-2}}\right].$$

10.2 (Belitser and Ghosal 2003) Consider the infinite-dimensional normal model of Section 10.3.2. Suppose that $q_0 \in \mathcal{Q}$, where $\mathcal{Q}$ is an arbitrary subset of $(0, \infty)$. Choose a countable dense subset $\mathcal{Q}^*$ of $\mathcal{Q}$, and put a prior λ on it such that $\lambda(q = s) > 0$ for each $s \in \mathcal{Q}^*$. For any sequence $M_n \to \infty$ and $\delta > 0$, show that $\Pi\{\theta: n^{q_0/(2q_0+1)-\delta}\|\theta - \theta_0\| > M_n \mid X\} \to 0$ in P_{θ_0}-probability as $n \to \infty$.

10.3 (Belitser and Ghosal 2003) Consider the infinite-dimensional normal model of Section 10.3.2. For any $\theta^* \in \Theta_{q_0}$, show that there exists $\epsilon > 0$ such that

$$\sup_{\theta_0 \in \mathcal{E}_{q_0}(\theta^*, \epsilon)} \mathrm{E}_{\theta_0}\Pi\{\theta: n^{q_0/(2q_0+1)}\|\theta - \theta_0\| > M_n \mid X\} \to 0\,,$$

where $\mathcal{E}_{q_0}(\theta^*, \epsilon) = \{\theta_0: \sum_{i=1}^{\infty} i^{2q_0}(\theta_{i0} - \theta_i^*)^2 < \epsilon\}$.

10.4 (Belitser and Ghosal 2003) In the infinite-dimensional normal model of Section 10.3.2, let $\mathcal{Q} = \{q_0, q_1\}$ and $\theta_{i0} = i^{-p}$. Show that if $p \ge q_1 + 1 - \frac{2q_1+1}{2(2q_0+1)}$, then $\Pi(q = q_1|X) \to 1$ in probability even if $\theta_0 \in \Theta_{q_0} \setminus \Theta_{q_1}$.

Show that the prior Π_{q_1} leads to a slower contraction rate than Π_{q_0} only when $p < q_1 + 1 - (2q_1 + 1)/(2(2q_0 + 1))$.

10.5 (Huang 2004) Consider $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p$, p an unknown density on $[0, 1]$, and let the true density $p_0 \in \mathfrak{W}^\alpha[0, 1]$ be such that $\|\log p_0\|_\infty \le M_0$. Let $J = \{(l, L): l \in \mathbb{N} \cup \{0\}, L \in \mathbb{N}\}$. For $j \in J$, let $m_j = 2^{l+1}$ and $B_{j,i}$, $i = 1, \ldots, m_j - 1$ stand for the Haar wavelets ψ_{j_1,k_1}, $0 \le j_1 \le l$, $0 \le k_1 \le 2^{j_1} - 1$, respectively. Define a prior on p by the relation $p(x) \propto \exp[\sum_{i=1}^{m_j-1} \theta_i B_{j,i}(x)]$, $\theta := (\theta_1, \ldots, \theta_{m_j-1}) \in \Theta_j$, where $\Theta_j = \{\theta \in \mathbb{R}^{m_j-1}: \|\sum_{i=1}^{m_j-1} \theta_j B_{j,i}\|_\infty \le L\}$. Given j, let θ be uniformly distributed over Θ_j. Choose $A_j = 19.28 \cdot 2^{(l+1)/2}(2L + 1)e^L + 0.06$, $C_j = m_j + L$ and assign a prior on various models by $a_j = \alpha \exp\{-(1+0.5\sigma^{-2}+0.0056\sigma^{-1})\eta_j\}$, where $\eta_j = 4m_j \log(1072.5A_j)c^{-1}(1 - 4\gamma)^{-1} + C_j \max(1, 8c^{-1}(1 - 4\gamma)^{-1})$, $c = 0.5 \min\{M^{-2}(1 - e^{-M^2/(2\sigma^2)}, \sigma^{-2})$, $0.13\gamma(1 - 4\gamma)^{-1/2} = 0.0056$. Show that the posterior distribution for p contracts at p_0 at the rate $n^{-\alpha/(1+2\alpha)}\sqrt{\log n}$.

10.6 (Huang 2004) Consider the regression problem $Y_i = f(X_i) + \varepsilon_i$, $X_i \overset{\text{iid}}{\sim} G$, $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{Nor}(0, \sigma^2)$, σ known, p_f the density of (X, Y). Consider a sieve-prior on f given by $f = f_{\theta,j}$, $\theta \in \Theta_j$, a subset of a Euclidean space, $\theta|j \sim \pi_j$, $j \in J$, a countable set, $j \sim a(\cdot)$. Let d_j the metric on Θ_j, $d_{j,\infty}(\theta, \eta) = \|f_{\theta,j} - f_{\eta,j}\|_\infty$ for $\theta, \eta \in \Theta_j$, $B_{2,j}(\eta, r) = \{\theta \in \Theta_j: \|f_{\theta,j} - f_{\eta,j}\|_{2,G} \le r\}$ and $\|f_{\theta,j}\|_\infty \le M$ for all $j, \theta \in \Theta_j$. Let f_0 be the true regression function. Assume that the following conditions hold:

(a) There are $0 < A_j < 0.0056$, $m_j \geq 1$ such that $N(B_{2,j}(\theta, r), \delta, d_{j,\infty}) \leq (A_j r/\delta)^{m_j}$ for all $r > 0$ and $\delta < 0.00056r$.
(b) Let $a_j = \alpha \exp\{-(1 + 0.5\sigma^{-2} + 0.0056\sigma^{-1})\eta_j\}$, $c = 0.5 \min\{M^{-2}(1 - e^{-M^2/(2\sigma^2)}, \sigma^{-2}), 0.13\gamma(1-4\gamma)^{-1/2} = 0.0056$, $\eta_j = 4m_j \log(1072.5A_j)c^{-1}(1-4\gamma)^{-1} + C_j \max(1, 8c^{-1}(1-4\gamma)^{-1})$, $\sum_{j=1}^{\infty} a_j = 1$, $\sum_{j=1}^{\infty} e^{-C_j} \leq 1$. There exist j_n and $\epsilon_n > 0$, $\epsilon_n \to 0$, $n\epsilon_n^2 \to \infty$, $\beta_n \in \Theta_{j_n}$ such that $\max\{K(p_{f_0}, p_{\beta_n, j_n}), V_2(p_{f_0}, p_{\beta_n, j_n})\} + n^{-1}\eta_{j_n}^2 \leq \epsilon_n^2$.
(c) $\|f_{\theta,j_n} - f_{\eta,j_n}\|_{2,G} \lesssim d_{j_n}(\theta, \eta)$ for all $\theta, \eta \in \Theta_{j_n}$.
(d) $\log N(\Theta_{j_n}, \epsilon_n, d_n) \leq m_{j_n}(K + b \log A_{j_n})$ for some constants K and b.

Show that the posterior for f contracts at the rate ϵ_n with respect to $\|f_{\theta,j} - f_{\eta,j}\|_{2,G}$.

10.7 (Huang 2004) In the setting of the nonparametric regression in Problem 10.6, let the true regression function $f_0 \in \mathfrak{W}^s[0,1]$, $\|f_0\|_\infty \leq M$ with known M, G be Unif[0, 1] and $J = \{(k, q, L): k \in \mathbb{N} \cup \{0\}, q \in \mathbb{N}, L \in \mathbb{N}\}$. For $j = (k, q, L)$, let $m_j = k + q$, $B_{j,i}$, $i = 1, \ldots, k$, be the normalized B-splines of order q and k regularly spaced knots, $f_{\theta,j} = \sum_{i=1}^{m_j} \theta_i B_{j,i}$, $(\theta_1, \ldots, \theta_{m_j}) \in \mathbb{R}^{m_j}$. Choose $A_j = 9.64\sqrt{q}(2q+1)9^{q-1} + 0.06$, $C_j = m_j + L$ and define η_j and a_j as in Problem 10.6. Show that the posterior distribution for f contracts at f_0 at the rate $n^{-s/(1+2s)}$ with respect to $\mathbb{L}_2$-distance.

Since the prior does not depend on the true smoothness level s, the posterior is rate adaptive.

10.8 (Xing 2008) Consider the setting and notations of Section 10.2. Suppose that there exist positive constants $H \geq 1$, E_α, $\mu_{n,\alpha}$, G, J, L, C and $0 < \gamma < 1$ such that $1 - \gamma > 18\gamma L$, $n\epsilon_{n,\beta_n}^2 \geq (1 + C^{-1}) \log n$, $E_\alpha \epsilon_{n,\alpha}^2 \leq G\epsilon_{n,\beta_n}^2$ for all $\alpha \in A_{n,\geq\beta_n}$, $E_\alpha \leq G$ for all $\alpha \in A_{n,>\beta_n}$ and $\sum_{\alpha\in A_n} \mu_{n,\alpha}^\gamma = O(e^{Jn\epsilon_{n,\beta_n}^2})$. Let $M \geq \sqrt{H} + 1 + 18(C + J + G + 3\gamma + 2\gamma C)/(1 - \alpha - 18\alpha L)$ such that

(a) $\log N(\epsilon/3, C_{n,\alpha}(p_0, 2\epsilon), d) \leq E_\alpha n\epsilon_{n,\alpha}^2$ for all $\alpha \in A_n$, $\epsilon \geq \epsilon_n$;
(b) $\dfrac{\lambda_{n,\alpha}\Pi_{n,\alpha}(C_{n,\alpha}(p_0, j\epsilon_{n,\alpha}))}{\lambda_{n,\beta_n}\Pi_{n,\beta_n}(\mathcal{H}_{n,\beta_n}(\epsilon_{n,\beta_n}))} \leq \mu_{n,\alpha} e^{Lj^2 n\epsilon_{n,\alpha}^2}$ for all $\alpha \in A_{n,>\beta_n}$ and $j \geq M$, where $\mathcal{H}_{n,\beta_n}(\epsilon) = \{p \in \mathcal{P}_{n,\beta_n}: \|(\sqrt{p} - \sqrt{p_0})(2\sqrt{p_0/p} + 1)/\sqrt{3}\|_2 < \epsilon\}$;
(c) $\dfrac{\lambda_{n,\alpha}\Pi_{n,\alpha}(C_{n,\alpha}(p_0, j\epsilon_{n,\alpha}))}{\lambda_{n,\beta_n}\Pi_{n,\beta_n}(\mathcal{H}_{n,\beta_n}(\epsilon_{n,\beta_n}))} \leq \mu_{n,\alpha} e^{Lj^2 n\epsilon_{n,\beta_n}^2}$ for all $\alpha \in A_{n,\geq\beta_n}$ and $j \geq M$;
(d) $\displaystyle\sum_{n=1}^{\infty} \sum_{\alpha\in A_{n,>\beta_n}} \frac{\lambda_{n,\alpha}\Pi_{n,\alpha}(C_{n,\alpha}(p_0, j\epsilon_{n,\alpha}))}{\lambda_{n,\beta_n}\Pi_{n,\beta_n}(\mathcal{H}_{n,\beta_n}(\epsilon_{n,\beta_n}))} e^{(1+2C)n\epsilon_{n,\beta_n}^2} < \infty.$

Show that $\Pi_n(p: d(p_0, p) \geq M\epsilon_{n,\beta_n} | X_1, \ldots, X_n) \to 0$ a.s. $[P_0^\infty]$ as $n \to \infty$.

10.9 (Xing 2008) In Problem 10.8, replace Condition (iii) by the following condition: There exists $K \geq 1$ independent of n, α and j such that $\dfrac{\lambda_{n,\alpha}\Pi_{n,\alpha}(C_{n,\alpha}(p_0, j\epsilon_{n,\alpha}))}{\lambda_{n,\beta_n}\Pi_{n,\beta_n}(\mathcal{H}_{n,\beta_n}(K\epsilon_{n,\beta_n}))} \leq \mu_{n,\alpha} e^{Lj^2 n\epsilon_{n,\beta_n}^2}$ for all $\alpha \in A_{n,\geq\beta_n}$ and $j \geq M$. Assume that $M \geq \sqrt{H} + 1 + 18\frac{C+J+G+3\gamma K+2\gamma CK}{1-\alpha-18\alpha L}$. Show that $\Pi_n(p: d(p_0, p) \geq M\epsilon_{n,\beta_n} | X_1, \ldots, X_n) \to 0$ a.s. $[P_0^\infty]$ as $n \to \infty$.

10.10 (Xing 2008) Applying Problem 10.9 to the log-spline prior with a finite index set A_n, show that the posterior adapts to the correct rate $n^{-\beta/(1+2\beta)}$ a.s., whenever $f \in \mathfrak{C}^\beta[0,1]$ for the $\mathbb{L}_2$-distance.

10.11 (Shen and Ghosal 2015) Show that in each of the following occasions, the posterior mean (and other moments) are analytically expressible (but may have a large number of terms):

- White noise model, identity link function, random series prior using multivariate normal distribution on the coefficients of B-spline functions — $f(x) = \sum_{j=1}^J \theta_j B_j(x)$, $(\theta_1, \ldots, \theta_J) \sim \mathrm{Nor}_J(\mu_J, \Sigma_{J\times J})$, $J \sim \pi$;
- Density estimation, identity link function, random series prior using Dirichlet distribution on the coefficients of normalized B-spline functions — $p(x) = \sum_{j=1}^J \theta_j B_j^*(x)$, $(\theta_1, \ldots, \theta_J) \sim \mathrm{Dir}(J; \alpha_1, \ldots, \alpha_J)$, $J \sim \pi$;
- Nonparametric normal regression, identity link function, random series prior using multivariate normal distribution on the coefficients of B-spline functions — $f(x) = \sum_{j=1}^J \theta_j B_j(x)$, $(\theta_1, \ldots, \theta_J) \sim \mathrm{Nor}_J(\mu_J, \Sigma_{J\times J})$, $J \sim \pi$;
- Nonparametric Poisson regression, identity link function, random series prior using independent gamma distribution on the coefficients of B-spline functions — $f(x) = \sum_{j=1}^J \theta_j B_j(x)$, $\theta_j \overset{\text{ind}}{\sim} \mathrm{Ga}(\alpha_j, \beta_j)$, $J \sim \pi$;
- Functional regression, identity link function, random series prior using multivariate normal distribution on the coefficients of B-spline functions — $\beta(x) = \sum_{j=1}^J \theta_j B_j(x)$, $(\theta_1, \ldots, \theta_J) \sim \mathrm{Nor}_J(\mu_J, \Sigma_{J\times J})$, $J \sim \pi$;
- Spectral density estimation, reciprocal link function, random series prior using independent gamma distribution on the coefficients of B-spline functions — $1/f(x) = \sum_{j=1}^J \theta_j B_j(x)$, $\theta_j \overset{\text{ind}}{\sim} \mathrm{Ga}(\alpha_j, \beta_j)$, $J \sim \pi$.

10.12 (Walker et al. 2005) Consider two models $\mathcal{P}_1$ and $\mathcal{P}_2$ for density estimation. Suppose that for a prior distribution Π, $\mathrm{KL}(\Pi) \cap \mathcal{P}_2 = \varnothing$. Show that the Bayes factor of model $\mathcal{P}_1$ to model $\mathcal{P}_2$ goes to infinity a.s. $[P_0^\infty]$ for every $p_0 \in \mathcal{P}_1 \cap \mathrm{KL}(\Pi)$.

10.13 (Xing 2008) Under the setting of Problem 10.8, show that $\Pi_n(A_{n,\prec\beta_n} | X_1, \ldots, X_n) \to 0$ a.s. $[P_0^\infty]$. If further

$$\sum_{n=1}^\infty \sum_{\alpha\in A_{n,\succeq\beta_n}} \frac{\lambda_{n,\alpha}\Pi_{n,\alpha}(C_{n,\alpha}(p_0, M\epsilon_{n,\beta_n}))e^{(3+2C)n\epsilon_{n,\beta_n}^2}}{\lambda_{n,\beta_n}\Pi_{n,\beta_n}(\mathcal{H}_{n,\beta_n}(\epsilon_{n,\beta_n}))} < \infty,$$

then show that $\Pi_n(\alpha \in A_n : H^{-1/2}\epsilon_{n,\beta_n}\epsilon_{n,\alpha} \le \sqrt{H}\epsilon_{n,\beta_n} | X_1, \ldots, X_n) \to 1$ a.s. $[P_0^\infty]$.

10.14 (Xing 2008) Assume the setting of Corollary 10.26 of two nested models. Let the assumption of Problem 10.8 hold and further assume that $\epsilon_{n,1} > \epsilon_{n,2} \ge n^{-1/2}\sqrt{(1+C^{-1})\log n}$ for some $C > 0$ and $M > 700(2C + G + 2)$. Show that

(a) if $\Pi_{n,2}(\mathcal{H}_{n,2}(\epsilon_{n,2}) \ge e^{-n\epsilon_{n,2}^2}$, $\frac{\lambda_{n,1}}{\lambda_{n,2}}\Pi_{n,1}(C_{n,1}(p_0, M\epsilon_{n,1})) = O(e^{-(4+3C)n\epsilon_{n,2}^2})$ and $\frac{\lambda_{n,1}}{\lambda_{n,2}} \le e^{n\epsilon_{n,1}^2}$, then $\mathbb{B}_n \to \infty$ a.s. $[P_0^\infty]$;

(b) if $\Pi_{n,1}(\mathcal{H}_{n,1}(\epsilon_{n,1}) \geq e^{-n\epsilon_{n,1}^2}$, $\frac{\lambda_{n,1}}{\lambda_{n,2}}\Pi_{n,2}(C_{n,2}(p_0, M\epsilon_{n,2})) = O(e^{-(4+3C)n\epsilon_{n,2}^2})$ and $\frac{\lambda_{n,2}}{\lambda_{n,1}} \leq e^{n\epsilon_{n,1}^2}$, then $\mathbb{B}_n \to 0$ a.s. $[P_0^\infty]$.

10.15 (Rousseau 2010) A key step in the derivation of a posterior contraction rate at a smooth density p_0 using a kernel mixture prior $\int \psi(x;\theta,\sigma)\,dF(\theta)$ is to construct a density p_1 of mixture type such that $\max\{K(p_0; p_1), V_2(p_0; p_1)\} \leq \epsilon_n^2$. Typically p_1 is chosen as $\int \psi(x;\theta,\sigma)p_0(\theta)\,d\theta$, or $\int \psi(x;\theta,\sigma)p_0^*(\theta)\,d\theta$, where p_0^* is some suitably truncated version of p_0, and the order of approximation is then $\max(\sigma^\beta, \sigma^2)$, where β is the smoothness level of p_0. In particular, the order does not improve if $\beta > 2$, leading to no improvement in contraction rate. In some situations it is possible to consider a different KL approximation $\int \psi(x;\theta,\sigma)p_1(\theta)\,d\theta$ which will lead to smaller approximation error $O(\sigma^\beta)$ for $\beta \geq 2$. The density p_1 used in the approximation is related to p_0, but the relation depends on the smoothness level of p_0. The following describes a situation where the higher order approximation error is achievable.

Consider estimating a density on $[0, 1]$. Let $\psi(x;\theta,\sigma) = x^{a-1}(1-x)^{b-1}/B(a,b)$, where $0 < x < 1$, $a^{-1} = \sigma(1-\theta)$, $b^{-1} = \sigma\theta$, $0 < \theta < 1$, $\sigma > 0$. Assume that $p_0 \in \mathfrak{C}^\beta[0,1]$ with $p^{(m_1}(0) > 0$, $p^{(m_2)}(1) > 0$ for some $m_1, m_2 \in \mathbb{N}$. Show that there exists functions $r_1(x), r_2(x), \ldots$ such that $p_0(x)[1+\sum_{j=2}^{\lfloor\beta\rfloor-1}\sigma^{j/2}r_j(x)]$ is proportional to a probability density $f_1(x)$ on $(0,1)$ and $\int p_0(x)|\log(p_0/p_1)|^m \lesssim \sigma^\beta$, where $p_1(x) = \int \psi(x;\theta,\sigma)f_1(x)\,dx$.

10.16 (Berger and Guglielmi 2001) Let $X_1, X_2, \ldots \overset{\text{iid}}{\sim} p$. Consider the problem of testing a parametric model $p \in \{p_\theta : \theta \in \Theta\}$ versus the nonparametric alternative that p is an arbitrary density. Let π be a prior density on θ in the parametric model. For a prior on the nonparametric alternative, consider a mixture of Pólya trees of the second kind introduced in Subsection 3.7.2 as follows: Let F_θ be the c.d.f. of p_θ, P_θ the corresponding probability measure,

$$\begin{aligned}\mathcal{T}_m &= \{B_{\varepsilon_1\cdots\varepsilon_m} : (\varepsilon_1\cdots\varepsilon_m) \in \{0,1\}^m\}\\ &= \{(F^{-1}((k-1)/2^m), F^{-1}(k/2^m)] : k = 1, \ldots, 2^m\},\end{aligned}$$

for some fixed distribution with a density (for instance, p_{θ^*}, where θ^* is a prior guess for θ in the parametric family). Let

$$\begin{aligned}c_{\varepsilon_1\cdots\varepsilon_{m-1}}(\theta) &= P_\theta(B_{\varepsilon_1\cdots\varepsilon_{m-1}1})/P_\theta(B_{\varepsilon_1\cdots\varepsilon_{m-1}0}),\\ \alpha_{\varepsilon_1\cdots\varepsilon_{m-1}0}(\theta) &= h^{-1}a_m/\sqrt{c_{\varepsilon_1\cdots\varepsilon_{m-1}}(\theta)},\\ \alpha_{\varepsilon_1\cdots\varepsilon_{m-1}1}(\theta) &= h^{-1}a_m\sqrt{c_{\varepsilon_1\cdots\varepsilon_{m-1}}(\theta)},\end{aligned}$$

a_m a sequence with $\sum_{m=1}^\infty a_m^{-1} < \infty$ (like $a_m = m^2$) and h a hyperparameter for scaling. Let θ have density π in the alternative as well.

Let $m^*(X_i) = \min\{m\colon X_{i'} \notin B_{\varepsilon_1 \cdots \varepsilon_m(X_i)},\, i' < i\}$, where $\varepsilon_1 \cdots \varepsilon_m(x)$ denotes the first m digits in the binary expansion for $F(x)$. Let

$$\psi(\theta \mid X_1, \ldots, X_n) = \prod_{i=2}^{n} \prod_{m=1}^{m^*(X_i)} \frac{\alpha'_{\varepsilon_1 \cdots \varepsilon_m(X_i)}(\theta)[\alpha_{\varepsilon_1 \cdots \varepsilon_{m-1}(X_i)0}(\theta) + \alpha_{\varepsilon_1 \cdots \varepsilon_{m-1}(X_i)1}(\theta)]}{\alpha_{\varepsilon_1 \cdots \varepsilon_m(X_i)}(\theta)[\alpha'_{\varepsilon_1 \cdots \varepsilon_{m-1}(X_i)0}(\theta) + \alpha'_{\varepsilon_1 \cdots \varepsilon_{m-1}(X_i)1}(\theta)]}.$$

Show that the Bayes factor for testing the parametric model versus the nonparametric alternative is given by

$$\mathbb{B}_n(X_1, \ldots, X_n) = \frac{\int \prod_{i=1}^{n} p_\theta(X_i)\pi(\theta)\, d\theta}{\int m_1(X_1, \ldots, X_n \mid \theta)\pi(\theta)\, d\theta},$$

where $m_1(X_1, \ldots, X_n \mid \theta) = \prod_{i=1}^{n} p_\theta(X_i)\psi(\theta \mid X_1, \ldots, X_n)$.

This allows the Bayes factor to be computed numerically by simulating $\theta_1, \ldots, \theta_N$ from a proposal density $q(\cdot)$ and computing the importance sampling Monte Carlo approximation

$$\frac{\sum_{j=1}^{N} \prod_{i=1}^{n} p_{\theta_j}(X_i)\pi(\theta_j)/q(\theta_j)}{\sum_{j=1}^{N} \psi(\theta_j \mid X_1, \ldots, X_n)\pi(\theta_j)/q(\theta_j)}.$$