Statistics & Decisions , 1– (2007)

On Universal Bayesian Adaptation

Jüri Lember, Aad van der Vaart

Received: Month-1 99, 2003; Accepted: Month-2 99, 2004

Summary: We consider estimating a probability density p based on a random sample from this density by a Bayesian approach. The prior is constructed in two steps, by first constructing priors on a collection of models each expressing a qualitative prior guess on the true density, and next combining these priors in an overall prior by attaching prior weights to the models. The purpose is to show that the posterior distribution contracts to the true distribution at a rate that is (nearly) equal to the rate that would have been obtained had only the model that is most suitable for the true density been used. We study special model weights that yield this adaptation property in some generality. Examples include minimal discrete priors and finite-dimensional models, with special attention to scales of Banach spaces, such as Hölder spaces, spline models, and classes of densities that are not uniformly bounded away from zero or infinity.

1 Introduction

Consider the problem of estimating a probability density p based on a random sample $X_1, \ldots, X_n$ from this density. If p is a density on d-dimensional Euclidean space and is a-priori known to possess α derivatives, then it is well known that p can be estimated at the rate $\varepsilon_{n,\alpha} = n^{-\alpha/(2\alpha+d)}$, relative to, for instance, the Hellinger distance on the set of probability densities (under some restrictions). A variety of methods achieve the rate $\varepsilon_{n,\alpha}$, and this is known to be optimal in a minimax sense if nothing more is known concerning p. Furthermore, it is well known that the rate $\varepsilon_{n,\alpha}$ can be achieved even if the value of α is not known a-priori. So-called rate-adaptive estimators achieve the rate $\varepsilon_{n,\alpha}$ if p is α-smooth for all values of α simultaneously. In this paper we investigate such rate-adaptation within a fully Bayesian set-up, where a suitable order α is selected through Bayesian model averaging.

More generally, we study the posterior distribution relative to a prior that is constructed by combining prior probability measures $\Pi_{n,\alpha}$ on the members $\mathcal{P}_{n,\alpha}$ of a list of models indexed by a parameter α with a prior probability measure λ_n on the set of indices α. The index α may be a regularity parameter, but in our general set-up it may

combine into the overall prior

$$\Pi_n = \int_A \Pi_{n,\alpha}\, \lambda_n(d\alpha). \tag{1.1}$$

Given the prior distribution (1.1), the corresponding posterior distribution is the random measure given by

$$\begin{aligned}\Pi_n(B|X_1,\ldots,X_n) &= \frac{\int_B \prod_{i=1}^n p(X_i)\, d\Pi_n(p)}{\int \prod_{i=1}^n p(X_i)\, d\Pi_n(p)} \\ &= \frac{\int_A \int_{p\in\mathcal{P}_{n,\alpha}:p\in B} \prod_{i=1}^n p(X_i)\, d\Pi_{n,\alpha}(p)\, \lambda_n(d\alpha)}{\int_A \int_{p\in\mathcal{P}_{n,\alpha}} \prod_{i=1}^n p(X_i)\, d\Pi_{n,\alpha}(p)\, \lambda_n(d\alpha)}.\end{aligned} \tag{1.2}$$

Of course, we make appropriate (measurability) conditions to ensure that this expression is well defined.

We say that the posterior distributions have *rate of convergence at least* ε_n if, for every sufficiently large constant M, as $n \to \infty$, in probability,

$$\Pi_n\big(d(p,p_0) > M\varepsilon_n|X_1,\ldots,X_n\big) \to 0.$$

Here the distribution of the random measure (1.2) is evaluated under the assumption that $X_1,\ldots,X_n$ are an i.i.d. sample from P_0, and d is the Hellinger, L_1- or L_2-metric on the set of densities.

This set-up falls within the set-up considered in earlier work on posterior distributions, for instance Ghosal et al. (2000). Application of their result would yield the statement that if the true density p_0 is well approximated by the overall model $\mathcal{P}_n = \cup_\alpha \mathcal{P}_{n,\alpha}$, then the posterior will concentrate in balls around p_0 of a radius ε_n that converges to zero at a speed depending on the complexity of the model $\mathcal{P}_n$ and its approximation properties. Because this rate ε_n refers to the union model $\mathcal{P}_n = \cup_\alpha \mathcal{P}_{n,\alpha}$, it will in most cases be determined by the worst case in the list, e.g. the index α for which the densities in $\mathcal{P}_{n,\alpha}$ are least regular. For instance, in the case that α is a smoothness parameter and A possesses a smallest element $\underline{\alpha}$, we would (at best) find the rate $n^{-\underline{\alpha}/(2\underline{\alpha}+d)}$ for the posterior. This is an interesting result, but it is pessimistic and not the result we would like to prove. In the present paper, we want to refine the result to the statement that if the true density p_0 is contained in $\mathcal{P}_{n,\beta}$ for some $\beta \in A$, or is well approximated by $\mathcal{P}_{n,\beta}$, then the posterior will concentrate in a ball of radius $\varepsilon_{n,\beta}$ around p_0, where $\varepsilon_{n,\beta}$ is the "correct rate" if only the model $\mathcal{P}_{n,\beta}$ were used. We prove that this is true (possibly up to logarithmic factors) for general priors $\Pi_{n,\alpha}$ on the models if combined with certain special model weights λ_n, defined in terms of the rates $\varepsilon_{n,\alpha}$ attached to the models (see e.g. (2.3)). Under these "universal" weights "small" models (the ones with fast rate of convergence $\varepsilon_{n,\alpha}$ attached) receive more prior mass than large models.

finite-dimensional models and treats several examples also covered in the present paper, with different priors. Huang's methods of proof are based on the methods developed by Barron and Cover (1991), who studied adaptation using penalized maximum likelihood estimators. In this paper we use the testing method originating in Schwartz (1964) and LeCam (1973), and further developed in Ghosal et al. (2000). The paper Ghosal et al. (2003) considers adaptation to finitely many log spline density models and uses similar ideas as the present paper.

The paper is organized as follows. In Section 2 we derive two main results and several corollaries. In Sections 3 and 4 we apply these results to two substantial methods of constructing priors: finite discrete priors on ε-nets over given Banach spaces and smooth finite-dimensional priors. In both cases these are made concrete for the scale of Hölder spaces, where the finite-dimensional priors are based on log spline models. The last section contains auxiliary and technical results.

1.1 Notation

Throughout the paper the data are a random sample $X_1, \ldots, X_n$ from a probability measure P_0 on a measurable space $(\mathcal{X}, \mathcal{A})$ with density p_0 relative to a given reference measure μ. In general we write p and P for a density and the corresponding probability measure. In our main results we use the Hellinger distance, total variation distance, or L_2-distance on the set of densities, denoted by d. For two probability densities p and q relative to μ these are defined as, respectively,

$$h(p,q) = \sqrt{\int |\sqrt{p} - \sqrt{q}|^2 \, d\mu},$$
$$\|p - q\|_1 = \int |p - q| \, d\mu,$$
$$\|p - q\|_2 = \sqrt{\int |p - q|^2 \, d\mu}.$$

If the L_2-distance is used, then it is implicitly assumed that all densities involved are bounded by a common upper bound. The ε-covering numbers of a metric space $(\mathcal{P}, d)$, denoted by $N(\varepsilon, \mathcal{P}, d)$, are defined as the minimal numbers of balls of radius ε needed to cover $\mathcal{P}$.

The index set is an arbitrary measurable space $(A, \mathcal{A}_0)$, and for every $n \in \mathbb{N}$ and every $\alpha \in A$ the set $\mathcal{P}_{n,\alpha}$ is a set of μ-probability densities on $(\mathcal{X}, \mathcal{A})$ equipped with a σ-field such that the maps $(x, p) \mapsto p(x)$ are measurable. Furthermore, $\Pi_{n,\alpha}$ denotes a probability measure on $\mathcal{P}_{n,\alpha}$ such that $(\alpha, B) \mapsto \Pi_{n,\alpha}(B)$ is a Markov kernel, and λ_n is a probability measure on A. This allows to form the prior $\Pi_n = \int_A \Pi_{n,\alpha} \, \lambda_n(d\alpha)$ on $\mathcal{P}_n = \cup_{\alpha \in A} \mathcal{P}_{n,\alpha}$ and corresponding posterior distirbution (1.2). We define

Throughout the paper $\varepsilon_{n,\alpha}$ are given positive numbers with $\varepsilon_{n,\alpha} \to 0$ as $n \to \infty$. These should be thought of as the rate attached to the model $\mathcal{P}_{n,\alpha}$ if this is (approximately) correct.

The notation $a \lesssim b$ means that $a \leq Cb$ for a constant C that is universal or fixed in the proof. For sequences a_n and b_n we write $a_n \ll b_n$ if $a_n/b_n \to 0$ and $a_n \gg 0$ if $a_n > 0$ for every n and $\liminf a_n > 0$. For a measure P and a measurable function f we write Pf for the integral of f relative to P.

2 Main results

The main results of the paper are adaptive versions of Theorem 2.1 of Ghosal et al. (2000). These authors consider the case of a prior $\Pi_{n,\beta}$ on a single model $\mathcal{P}_{n,\beta}$ such that, for some positive constants $\underline{E}, E, \underline{F}, F$,

$$\log N(\underline{E}\varepsilon_{n,\beta}, \mathcal{P}_{n,\beta}, d) \leq En\varepsilon_{n,\beta}^2, \tag{2.1}$$

$$\Pi_{n,\beta}\big(B_{n,\beta}(\underline{F}\varepsilon_{n,\beta})\big) \geq \exp[-Fn\varepsilon_{n,\beta}^2]. \tag{2.2}$$

The first condition measures the complexity of the model $\mathcal{P}_{n,\beta}$, whereas the second condition lower bounds the amount of prior mass close to the true density p_0. Theorem 2.1 of Ghosal et al. (2000) shows that under conditions (2.1)-(2.2) the posterior rate is at least $\varepsilon_{n,\beta}$. In many (but not all) examples this upper bound on the rate appears to be (almost) sharp. The theorems in this section show that this rate of convergence remains valid if the model $\mathcal{P}_{n,\beta}$ makes part of a family of models $\{\mathcal{P}_{n,\alpha}\colon \alpha \in A\}$ and the prior mass is distributed appropriately over the elements of the list. The theorems impose the same two conditions on the models $\mathcal{P}_{n,\alpha}$, or very similar ones, and in addition assume that the weights λ_n have the form: for a constant $C > 0$, subsets $A_n \subset A$, and a fixed measure λ on A

$$\lambda_n(d\alpha) = \frac{\exp[-Cn\varepsilon_{n,\alpha}^2]\, 1_{A_n}(\alpha)\, \lambda(d\alpha)}{\int_{A_n} \exp[-Cn\varepsilon_{n,\alpha}^2]\, \lambda(d\alpha)}, \tag{2.3}$$

The measure λ may be infinite (e.g. counting measure on a countable set), as long as the denominator of (2.3) is finite. The function λ_n gives larger weight to models with fast rate of convergence (which we think of as the "smaller" models).

For $\beta \in A$ and $1 \leq H < \infty$ define

$$A_{n,\approx\beta,H} := \big\{\alpha \in A_n\colon \varepsilon_{n,\beta}^2 \leq \varepsilon_{n,\alpha}^2 \leq H\varepsilon_{n,\beta}^2\big\}, \tag{2.4}$$

$$A_{n,\gtrsim\beta,H} := \big\{\alpha \in A_n\colon \varepsilon_{n,\alpha}^2 \leq H\varepsilon_{n,\beta}^2\big\}. \tag{2.5}$$

These are the sets of all model indices that give "approximately the same" rate, or "not really bigger" rate as the index β. The notations $\approx \beta$ and $\gtrsim \beta$ are used even though we do not require that the index set A is partially ordered. The spirit of this notation is that

Thus $U_n(\beta, H, \underline{F}, F)$ is a "neighbourhood" around β of model indices that satisfy the prior mass condition (2.2) with $\alpha = \beta$. Because $B_{n,\alpha}(\varepsilon_{n,\alpha})$ is a neighbourhood of the true density p_0 within $\mathcal{P}_{n,a}$, we interpret inclusion of an index α in $U_n(\beta, H, \underline{F}, F)$ as meaning that the model $\mathcal{P}_{n,\alpha}$ is appropriate for p_0, in the sense that if just this model were used the rate would be $\varepsilon_{n,\alpha}$ if also $\mathcal{P}_{n,\alpha}$ were of the correct complexity (i.e. satisfy both (2.1) and (2.2).

Rather than a fixed index of a "true model" we shall employ a sequence of indices $\beta_n \in A_n$ and we establish the rate ε_{n,β_n} attached to this sequence. This will allow to approximate a "regularity index" that is not in the support of λ_n, and will also provide sufficient flexibility to handle applications where the index α does not refer to a fixed regularity measure, but for instance to the dimension of the model $\mathcal{P}_{n,\alpha}$.

Theorem 2.1 *Assume that there exist positive constants $C, \underline{E}, E, \underline{F}, F, H, T$ and a sequence of indices $\beta_n \in A_n$ such that $HC \geq 2(2\underline{F}^2 + CT + FT)$ and*

$$\log N\Big(\underline{E}\varepsilon_{n,\beta_n}, \cup_{\alpha \in A_{n,\gtrsim \beta_n, H}} \mathcal{P}_{n,\alpha}, d\Big) \leq En\varepsilon_{n,\beta_n}^2. \tag{2.7}$$

If $n\varepsilon_{n,\beta_n}^2 \to \infty$ and

$$\frac{\int_{A_n} \exp[-Cn\varepsilon_{n,\alpha}^2/4]\, \lambda(d\alpha)}{\lambda\big(U(\beta_n, T, \underline{F}, F)\big)} = O(1), \tag{2.8}$$

then the posterior distributions (1.2) relative to the weight functions λ_n given by (2.3) satisfy $P_0^n \Pi_n\big(p\colon d(p, p_0) \geq M\varepsilon_{n,\beta_n} | X_1, \cdots, X_n\big) \to 0$.

The proof of the theorem can be found in Section 5.

The conditions (2.7)-(2.8) play a similar role to the conditions (2.1)-(2.2). The model with index β_n should be thought of as a "true" model. Condition (2.7) requires that the complexity of the union of all models that are "smaller" than the true model is "not larger" than the complexity of the true model. Condition (2.8) is essentially a lower bound on the prior mass in a neighbourhood of the true density p_0 within the models that have the "same" rate as model β_n. Before interpreting these conditions further, we derive a corollary for the situation that A is a countable set.

Consider priors of the form (2.3) with λ a measure on a countable set A_n with atoms written as $\lambda\{\alpha\} = \lambda_\alpha$. Thus for a positive constant C and arbitrary positive numbers λ_α,

$$\lambda_n\{\alpha\} = \frac{\lambda_\alpha \exp[-Cn\varepsilon_{n,\alpha}^2]}{\sum_{\alpha \in A_n} \lambda_\alpha \exp[-Cn\varepsilon_{n,\alpha}^2]} 1_{A_n}(\alpha). \tag{2.9}$$

For $\beta_n \in A_n$ assume that

Corollary 2.2 *Assume that there exist positive constants $C, \underline{E}, E, \underline{F}, F, H$ and a sequence of indices $\beta_n \in A_n$ such that $HC \geq 2(2\underline{F}^2 + C + F)$ and such that (2.7), (2.10) and (2.11) hold. If $n\varepsilon_{n,\beta_n}^2 \to \infty$, then the posterior distributions (1.2) relative to the weight functions λ_n given by (2.9) have rate of convergence at least ε_{n,β_n}.*

Proof: We set $T = 1$ and note that $\beta_n \in U_n(\beta_n, 1, \underline{F}, F)$ by (2.11). Consequently, $\lambda\big(U_n(\beta_n, 1, \underline{F}, F)\big) \geq \lambda_{\beta_n}$ and (2.8) is implied by (2.10). □

Condition (2.11) is a direct generalization of the prior mass condition (2.2). Thus, the rate ε_{n,β_n} in this condition corresponds to the rate we would obtain if we would employ only the model $\mathcal{P}_{n,\beta_n}$ with its associated prior Π_{n,β_n}.

Condition (2.10) is trivially satisfied if λ is a finite measure with $\lambda_{\beta_n} \gg 0$. It is also satisfied if $\lambda_\alpha = 1$ for every α and the set of models is not too large, as the terms $\exp[-Cn\varepsilon_{n,\alpha}^2/4]$ will typically be very small. Thus this condition seems to be mostly of a technical character.

The entropy condition (2.7) appears not unnatural. In particular, if the index set A is ordered and the models $\mathcal{P}_{n,\alpha}$ are nested with the rates $\varepsilon_{n,\alpha}$ "decreasing" in α, then the condition is implied by the bound (2.1) with β taken equal to the smallest index $\alpha \in A_{n,\gtrsim\beta_n,H}$. (Because necessarily $H > 1$ the latter index may be "smaller" than β_n, but the difference is typically small.)

In the remainder of this section we shall investigate other situations in which condition (2.7) can be reduced to bounds of type (2.1) on single models. We shall assume that (2.1) holds for every α, where the constant E may depend on α, i.e.

$$\log N(\varepsilon_{n,\alpha}, \mathcal{P}_{n,\alpha}, d) \leq E_\alpha n\varepsilon_{n,\alpha}^2, \qquad \alpha \in A. \tag{2.12}$$

A first observation is that we always have the crude bounds

$$N\Big(\sqrt{H}\varepsilon_{n,\beta_n}, \bigcup_{\alpha\gtrsim\beta_n} \mathcal{P}_{n,\alpha}, d\Big) \leq \sum_{\alpha\gtrsim\beta_n} N(\sqrt{H}\varepsilon_{n,\beta_n}, \mathcal{P}_{n,\alpha}, d) \leq \sum_{\alpha\gtrsim\beta_n} N(\varepsilon_{n,\alpha}, \mathcal{P}_{n,\alpha}, d).$$

Both inequalities in this display are pessimistic. The first inequality ignores the fact that the models may overlap, and is particularly crude in the nested case. The second inequality ignores that the rates $\varepsilon_{n,\alpha}$ for $\alpha \gtrsim \beta_n$ may be much smaller than ε_{n,β_n}, so that the entropy at $\varepsilon_{n,\alpha}$ may be much bigger than the entropy at ε_{n,β_n}. Nevertheless, the bounds can be useful.

In particular, if A is a finite set, then we can bound the sum on the right by its cardinality $\#A$ times its maximum term. Then from (2.1) we obtain that the logarithm of each of the terms in the sum is of the order $n\varepsilon_{n,\alpha}^2 \leq Hn\varepsilon_{n,\beta_n}^2$, and (2.7) holds. This yields the following corollary.

An infinite set of models cannot be treated as easily, but the argument can be extended to infinite (possibly uncountable) A by employing priors λ_n that are truncated to a finite set A_n. Assume that for a suitable constant H,

$$\max_{\alpha\in A_n:\varepsilon_{n,\alpha}^2\leq H\varepsilon_{n,\beta_n}^2} E_\alpha \frac{\varepsilon_{n,\alpha}^2}{\varepsilon_{n,\beta_n}^2} = O(1), \tag{2.13}$$

$$\log(\#A_n) \lesssim n\varepsilon_{n,\beta_n}^2. \tag{2.14}$$

Theorem 2.4 *Assume that there exist positive constants $E_\alpha, \underline{F}, C, H$ and a sequence β_n of indices such that $HC \geq 2(2\underline{F}^2 + C + F)$ and such that (2.10), (2.11), (2.12), (2.13) and (2.14) hold. Let $A_n \subset A$ be such that $\beta_n \in A_n$ for every sufficiently large n. If $n\varepsilon_{n,\beta_n}^2 \to \infty$, then the posterior distributions (1.2) relative to the weight functions λ_n given by (2.9) have rate of convergence at least ε_{n,β_n}.*

Proof: Let $\mathcal{P}_{n,A_n,\gtrsim\beta_n}$ be the union of the models $\mathcal{P}_{n,\alpha}$ with $\alpha \in A_n$ and $\alpha \gtrsim \beta_n$, where the latter inequality $\alpha \gtrsim \beta_n$ refers to the indices with $\varepsilon_{n,\alpha}^2 \leq H\varepsilon_{n,\beta_n}^2$. Then, by the argument given previously,

$$N\big(\sqrt{H}\varepsilon_{n,\beta_n}, \mathcal{P}_{n,A_n,\gtrsim\beta_n}, d\big) \leq \#A_n \max_{\alpha\in A_n:\alpha\gtrsim\beta_n} N(\varepsilon_{n,\alpha}, \mathcal{P}_{n,\alpha}, d).$$

Hence by (2.12) the logarithm of the left side is bounded above by $\log(\#A_n)+E_n n\varepsilon_{n,\beta_n}^2$, for the numbers E_n defined as the left side of (2.13). By the assumptions (2.13) and (2.14) this is bounded above by $En\varepsilon_{n,\beta_n}^2$ for some constant E. The theorem is a corollary of Corollary 2.2. □

Remark 2.5 As can be seen from the proofs, the global entropy conditions (2.7) and (2.12) can be replaced by the weaker *local entropy* conditions

$$\sup_{\varepsilon\geq\underline{E}\varepsilon_{n,\beta_n}} \log N\Big(\frac{\varepsilon}{3}, \cup_{\alpha\gtrsim\beta_n} C_{n,\alpha}(2\varepsilon), d\Big) \leq En\varepsilon_{n,\beta_n}^2,$$

$$\sup_{\varepsilon\geq\varepsilon_{n,\alpha}} \log N\Big(\frac{\varepsilon}{5}, C_{n,\alpha}(2\varepsilon), d\Big) \leq E_\alpha n\varepsilon_{n,\alpha}^2. \tag{2.15}$$

These strengthenings are especially of interest when the models are of finite dimension, as in Section 4.

Suppose now that there is a fixed index $\beta \in A$ that gives the "true" model", and the rate $\varepsilon_{n,\beta}$ is to be achieved. Theorem 2.4 can be used with $\beta_n = \beta$ provided $\beta \in A_n$ eventually. If A is countable, then the latter is guaranteed as soon as $A_n \nearrow A$. For

Corollary 2.6 *Assume (2.12). Let $\beta \in A$ be such that $n\varepsilon_{n,\beta}^2 \to \infty$. Let $C, \underline{F}, F, H, T$ be positive constants with $HTC \geq 2(2\underline{F}^2 + C + F)$. Let $A_n \subset A$ be finite subsets such that $\log(\#A_n) \lesssim n\varepsilon_{n,\beta}^2$, $U_n(\beta, T, \underline{F}, F) \neq \emptyset$ for all sufficiently large n, and*

$$\sum_{\alpha \in A_n} \max_{\alpha' \in A_{n,\approx\beta,T}} \Big(\frac{\lambda_\alpha}{\lambda_{\alpha'}}\Big) \exp[-Cn\varepsilon_{n,\alpha}^2/4] = O(1). \tag{2.16}$$

If (2.13) with β substituted for β_n holds, then the posterior distributions (1.2) relative to the weight functions λ_n given by (2.9) have rate of convergence at least $\varepsilon_{n,\beta}$.

Proof: By assumption, there exists $\beta_n \in U_n(\beta, T, \underline{F}, F)$, which implies that $\beta_n \in A_n$ and that (2.11) holds, and also that ε_{n,β_n} and $\varepsilon_{n,\beta}$ are of the same order. The latter and the assumptions show that $n\varepsilon_{n,\beta_n}^2 \to \infty$ and $\log(\#A_n) \lesssim n\varepsilon_{n,\beta_n}^2$ for n big enough. The condition (2.16) implies (2.10). Finally, if (2.13) holds with respect to β for a given constant H, then (2.13) holds with respect to β_n for the constant HT. Hence all conditions of Theorem 2.4 are fulfilled, so that this theorem gives the rate of convergence $\varepsilon_{n,\beta_n} \leq \sqrt{T}\varepsilon_{n,\beta}$. □

Theorem 2.4 and the preceding corollary add to the natural assumptions (2.11) and (2.12) on the single models, the conditions (2.10), (2.13) and (2.14). These conditions are not strong. Moreover, in the remainder of this section we show that they can often be arranged by choosing the sets A_n small.

Of course, the sets A_n must also be chosen rich enough so that every possible index α is (asymptotically) represented by an index $\beta_n \in A_n$ that satisfies (2.11). For a countable set A of target values α we may choose $A_n \uparrow A$ at a slow rate. More generally, we might construct the sets A_n by a "discretization" of the set A. This discretization must be rich enough to contain (eventually) for every $\beta \in A$ an index β_n with rate ε_{n,β_n} that is of the same order as the "target rate" $\varepsilon_{n,\beta}$ (and for which (2.11) holds). Because $\varepsilon_{n,\alpha} \leq \sqrt{H}\,\varepsilon_{n,\alpha'}$ if and only if $\log \varepsilon_{n,\alpha} \leq \log \varepsilon_{n,\alpha'} + \frac{1}{2}\log H$, this is the case as soon as the set of logarithmic rates $\{\log \varepsilon_{n,\alpha} \colon \alpha \in A_n\}$ forms a grid of fixed meshwidth over (parts of) A (eventually). The following three examples give (theoretical) constructions of such grids that satisfy (2.10), (2.13) and (2.14).

Example 2.7 Let $A \subset (0, \infty)$ and suppose that $n\varepsilon_{n,\alpha}^2 = n^{g(\alpha)}c_n$ for some strictly decreasing $g \colon [0, \infty) \to [0, \infty)$. If the range of g is contained in a finite interval $[\underline{g}, \overline{g}]$, then the set $\{\log \varepsilon_{n,\alpha}, \alpha \in A\}$ of all possible log rates is the interval $\frac{1}{2}\log(c_n/n) + \frac{1}{2}\log n[\underline{g}, \overline{g}]$, which has length $\frac{1}{2}\log n(\overline{g} - \underline{g})$. Therefore, a grid A_n with of the order $\log n$ points can suitably represent A. In terms of the indices α, for a diffeomorphism g it suffices that the grid includes an index in each interval of length of the order $1/\log n$.

This applies for instance to rates of the form $\varepsilon_{n,\alpha} = n^{-\alpha/(2\alpha+d)}(\log n)^k$ with α

bounded for $\alpha \in A_\beta$, and $\varepsilon_{n,\alpha}/\varepsilon_{n,\beta} \to 0$ as $n \to \infty$, for every $\alpha \in A_{>\beta}$. Then we can always construct sets A_n with $A_n \uparrow A$ that satisfy (2.13) for every fixed sequence $\beta_n = \beta$, as follows.

We order the set A in a sequence and define $\eta_{n,m}$ as the maximum of the numbers $E_\alpha \varepsilon_{n,\alpha}^2/\varepsilon_{n,\beta}^2$ for α and β ranging over the first m elements of A and with $\alpha > \beta$. Because $\varepsilon_{n,\alpha}/\varepsilon_{n,\beta} \to 0$ for $\alpha > \beta$, the sequence $\eta_{n,m}$ converges to zero as $n \to \infty$ for every fixed m, and hence there exists $m_n \to \infty$ with $\eta_{n,m_n} \to 0$ also. We may now set A_n equal to the first m_n elements of A.

Example 2.9 Let $A \subset (0,\infty)$ and suppose that $n\varepsilon_{n,\alpha}^2 = n^{g(\alpha)}$ for some strictly decreasing, continuous function $g\colon [0,\infty) \to [0,\infty)$. Suppose that, for all $0 < a < b < \infty$,

$$\sup_{\alpha\in[a,b]} E_\alpha =: E(a,b) < \infty \tag{2.17}$$

Then there exist subsets $A_n \subset A$ such that, with $\lambda_\alpha = 1$, (2.10), (2.13) and (2.14) are satisfied and the set $A_{n,\approx\beta,H}$ is nonempty, provided n is big enough, for any fixed $\beta > 0$.

One construction to prove this claim is as follows. For natural numbers $m \geq 2$ define

$$\eta_{n,m} := \sup_{\alpha\in[m,m+1]} E_\alpha n^{g(\alpha)-g(m-1)}.$$

Since g is strictly decreasing and (2.17) holds, $\eta_{n,m} \to 0$ as $n \to \infty$, for every fixed m, and hence also $\max_{2\leq m\leq k} \eta_{n,m} \to 0$ as $n \to \infty$, for every fixed k. This implies that there exists $m_n \nearrow \infty$ such that $\max_{2\leq m\leq m_n} \eta_{n,m} \leq 1$, for every sufficiently large n. Consequently, if $2 \leq m < m_n$ is a natural number, then

$$\sup_{\alpha\in[m,m_n]} E_\alpha n^{g(\alpha)-g(m-1)} \leq \eta_{n,m} \vee \eta_{n,m+1} \vee \cdots \vee \eta_{n,m_n-1} \leq 1. \tag{2.18}$$

Choose $\overline{\alpha}_n \nearrow \infty$ slowly enough that $n^{g(\overline{\alpha}_n)} \gg \log\log n$, and define $\tilde{\alpha}_n := \overline{\alpha}_n \wedge m_n$. Let $A_n \subset (0,\tilde{\alpha}_n]$ be such that $\{\log \varepsilon_{n,\alpha}\colon \alpha \in A_n\}$ is a $\frac{1}{2}\log H$-net in the interval $\frac{1}{2}\log(1/n) + \frac{1}{2}\log n\big[g(\tilde{\alpha}_n), g(0)\big]$. Since $\tilde{\alpha}_n \nearrow \infty$, this interval contains $\log \varepsilon_{n,\beta} = \frac{1}{2}\log(1/n) + \frac{1}{2}g(\beta)\log n$, eventually. Therefore, for sufficiently large n, the set $A_{n,\approx\beta,H}$ is nonempty. Furthermore, (2.10) and (2.14) are satisfied, because $\#A_n \lesssim \log n$, while $\min_{\alpha\in A_n} n\varepsilon_{n,\alpha}^2 \geq n^{g(\overline{\alpha}_n)} \gg \log\log n$.

Define $\underline{\beta}_n$ by $\varepsilon_{n,\underline{\beta}_n}^2 = H\varepsilon_{n,\beta}^2$, or equivalently by $g(\underline{\beta}_n) = g(\beta) + \log H/\log n$. The continuity of g implies that $\underline{\beta}_n \nearrow \beta$. Let $m(\beta) = \min\{m \in \mathbb{N}\colon m - 1 \geq \beta\}$. Since $\beta > 0$, $m(\beta) \geq 2$. Because $A_n \subset (0, m_n]$ and $\varepsilon_{n,\alpha} \leq \sqrt{H}\varepsilon_{n,\beta}$ if and only if $\alpha \geq \underline{\beta}_n$,

$$\max_{\alpha\in A_n:\varepsilon_{n,\alpha}\leq\sqrt{H}\varepsilon_{n,\beta}} E_\alpha \frac{\varepsilon_{n,\alpha}^2}{\varepsilon_{n,\beta}^2} \leq \sup_{\alpha\in[\underline{\beta}_n,m_n]} E_\alpha \frac{\varepsilon_{n,\alpha}^2}{\varepsilon_{n,\beta}^2}$$

bounded by $E\big(\beta, m(\beta)\big)$. Finally, because of (2.18), the third supremum can be bounded by

$$\sup_{\alpha\in[m(\beta),m_n]} E_\alpha n^{g(\alpha)-g(\beta)} \le \sup_{\alpha\in[m(\beta),m_n]} E_\alpha n^{g(\alpha)-g(m(\beta)-1)} \le 1.$$

3 Priors based on nets

Given a metric d and $\varepsilon > 0$ say that a set of functions $u_1, \ldots, u_N \colon \mathcal{X} \to \mathbb{R}$ on a measurable space $(\mathcal{X}, \mathcal{A})$ is a set of *ε-upper brackets* for a given set $\mathcal{P}$ of densities if for every $p \in \mathcal{P}$ there exist a function u_i with both $p \le u_i$ and $d(u_i, p) < \varepsilon$. The *ε-upper bracketing number* $N_](\varepsilon, \mathcal{P}, d)$ is defined as the minimal number of functions in such a set of ε-brackets. These upper bracketing numbers are smaller than the more usual bracketing numbers employed in empirical process theory (e.g. van der Vaart and Wellner (1996), Definition 2.1.6), but still bigger than the covering numbers $N(\varepsilon/2, \mathcal{P}, d)$. However, in many situations the three types of complexity measures are of the same order in ε as $\varepsilon \searrow 0$.

The optimal rate of convergence ε_n for a model $\mathcal{P}$ relative to the Hellinger distance h can typically be related to its entropy, through the equation

$$\log N_](\varepsilon_n, \mathcal{P}, h) \asymp n\varepsilon_n^2.$$

See Birgé (1986). In Ghosal et al. (2000) posterior distributions relative to priors constructed on minimal brackets were shown to contract at this rate. Here we extend these results to adaptation to multiple models.

For each $\alpha \in A$ let $\mathcal{Q}_{n,\alpha}$ be a set of nonnegative, integrable functions with finite upper bracketing numbers relative to the Hellinger distance h (not necessarily probability densities). In agreement with the preceding display, let target rates $\varepsilon_{n,\alpha}$ satisfy, for every $\alpha \in A$,

$$\log N_](\varepsilon_{n,\alpha}, \mathcal{Q}_{n,\alpha}, h) \lesssim n\varepsilon_{n,\alpha}^2. \tag{3.1}$$

Next for each α choose a set $\mathcal{U}_{n,\alpha} = \{u_1, \ldots, u_N\}$ of $\varepsilon_{n,\alpha}$-upper brackets over $\mathcal{Q}_{n,\alpha}$ and let $\mathcal{P}_{n,\alpha}$ be the set of re-normalized functions

$$\Big\{ \frac{u_j}{\int u_j \, d\mu} \colon j = 1, \ldots, N \Big\} = \Big\{ \frac{u}{\int u \, d\mu} \colon u \in \mathcal{U}_{n,\alpha} \Big\}. \tag{3.2}$$

Let the prior $\Pi_{n,\alpha}$ be the uniform probability measure on $\mathcal{P}_{n,\alpha}$.

In particular, we may use a minimal set of $\varepsilon_{n,\alpha}$-upper brackets over $\mathcal{Q}_{n,\alpha}$. In this section we show that the resulting prior $\Pi_{n,\alpha}$ is then appropriate if the true density is contained in the union $\cup_{M>0}(M\mathcal{Q}_{n,\alpha})$ of the sets $\mathcal{Q}_{n,\alpha}$. The base collection $\mathcal{Q}_{n,\alpha}$ could for instance be the unit ball in a regularity space, and then it suffices that p_0 is contained in this space. The set of brackets need not be minimal, but we assume that its cardinality

$p_0 \in M_0\mathcal{Q}_{n,\beta_n}$ for every sufficiently large n. Let $A_n \subset A$ be such that $\beta_n \in A_n$ eventually, and such that (2.10), (2.13) and (2.14) hold, for every $H > 0$ and some $C > 0$. Let $n\varepsilon_{n,\beta_n}^2 \to \infty$. Then the posterior distributions (1.2) relative to the weight functions λ_n given by (2.9) have rate of convergence at least ε_{n,β_n} relative to the Hellinger distance.

Proof: By construction $\#\mathcal{P}_{n,\alpha} \leq \exp(E_\alpha n\varepsilon_{n,\alpha}^2)$, implying that (2.12) with d equal to the Hellinger distance is trivially satisfied. Since (2.13) holds for every H, it also holds for some H with $HC \geq 2(2\underline{F}^2 + C + F)$, for any constants $\underline{F}, F, C$. If we can also show that (2.11) holds, then the theorem follows from Theorem 2.4.

By assumption, there exist constants $M_0 > 0$ such that $p_0/M_0 \in \mathcal{Q}_{n,\beta_n}$ for every sufficiently large n. So, there exists $u_n \in \mathcal{U}_{n,\beta_n}$ such that $p_0/M_0 \leq u_n$ and $\big\|\sqrt{p_0/M_0} - \sqrt{u_n}\big\|_2 = h(p_0/M_0, u_n) \leq \varepsilon_{n,\beta_n}$. It follows that

$$1 = \|\sqrt{p_0}\|_2 \leq \|\sqrt{M_0u_n}\|_2 \leq \|\sqrt{M_0u_n} - \sqrt{p_0}\|_2 + \|\sqrt{p_0}\|_2 \leq \sqrt{M_0}\varepsilon_{n,\beta_n} + 1.$$

By construction the function $p_n = u_n/\int u_n d\mu$ belongs to $\mathcal{P}_{n,\beta_n}$. Furthermore, by the triangle inequality,

$$\begin{aligned} h(p_0, p_n) &\leq h(p_0, M_0u_n) + h(M_0u_n, p_n) \\ &= h(p_0, M_0u_n) + \big|\|\sqrt{M_0u_n}\|_2 - 1\big| \leq 2\sqrt{M_0}\varepsilon_{n,\beta_n}. \end{aligned}$$

The inequality in the second line follows from the fact that $\|r - r/\|r\|\| = \big|1 - \|r\|\big|$ for every norm and function r, applied with $r = \sqrt{M_0u_n}$. We also have

$$\frac{p_0}{p_n} \leq M_0 \int u_n \, d\mu = \|\sqrt{M_0u_n}\|_2^2 \lesssim 1 + M_0\varepsilon_{n,\beta_n}^2,$$

which is uniformly bounded by assumption. In view of Lemma 5.3, it follows that $p_n \in B_{n,\beta_n}(D\sqrt{M_0}\varepsilon_{n,\beta_n})$ for a sufficiently large constant D, whence

$$\Pi_{n,\beta_n}\big(B_{n,\beta_n}(D\sqrt{M_0}\varepsilon_{n,\beta_n})\big) \geq \Pi_{n,\beta_n}(\{p_n\}) \geq (1/\#\mathcal{P}_{n,\beta_n}) \geq \exp[-E_{\beta_n}n\varepsilon_{n,\beta_n}^2].$$

The assumption (2.13) implies that E_{β_n} is bounded above. It follows that the prior probability $\Pi_{n,\beta_n}\big(B_{n,\beta_n}(\underline{F}\varepsilon_{n,\beta_n})\big)$ is bounded below by $\exp[-Fn\varepsilon_{n,\beta_n}^2]$, for constants, $\underline{F}, F$. This completes the verification of (2.11). □

Because bounds on bracketing numbers have been established for many situations, and typically give sharp rates of convergence, the preceding theorem can be seen as confirmation that in many situations there exist priors that give Bayesian adaptation across a scale of models of interest. That the base collections $\mathcal{Q}_{n,\alpha}$ need not be collections of probability densities is helpful, because in this form the theorem applies to any true

A disadvantage of this otherwise attractive construction is that the resulting models $\mathcal{P}_{n,\alpha}$ need not be nested, even if the base collections $\mathcal{Q}_{n,\alpha}$ may be. For instance, scales of regularity spaces are nested, and it is natural to build this into the priors. Technically, this would permit to use condition (2.7) of Theorem 2.1 instead of (2.13) and (2.14), which result from truncating the weight function λ_n. Assume that A is totally ordered and let $\underline{\beta}_n$ be the minimum of the set $A_{n,\gtrsim\beta_n,H} = \{\alpha \in A_n : \varepsilon_{n,\alpha}^2 \leq H\varepsilon_{n,\beta_n}^2\}$, which we assume to exist.

Lemma 3.2 *If A is a totally ordered set and $\mathcal{Q}_{n,\alpha}$ are sets of probability densities with $\mathcal{Q}_{n,\alpha} \subset \mathcal{Q}_{n,\beta}$ for $\alpha \geq \beta$, then conditions (2.13) and (2.14) in Theorem 3.1 may be replaced by the condition that $E_{\underline{\beta}_n} = O(1)$.*

Proof: We employ Corollary 2.2 rather than Theorem 2.4. It suffices to verify (2.7), the other part of the proof being the same as before. (Note that $p_0 \in M_0\mathcal{Q}_{n,\beta_n} \subset M_0\mathcal{Q}_{n,\underline{\beta}_n}$.)

If $p \in \mathcal{P}_{n,\alpha}$ for some $\alpha \geq \underline{\beta}_n$, then p is a renormalized upper bracket $u \in \mathcal{U}_{n,\alpha}$. By construction of $\mathcal{U}_{n,\alpha}$ there exists $q \in \mathcal{Q}_{n,\alpha}$ with $h(u,q) \leq \varepsilon_{n,\alpha}$. Arguing as in the proof of Theorem 3.1, using the assumption that q is a probability density, we can see that $\int u\,d\mu$ differs from 1 by at most $\varepsilon_{n,\alpha}$, and consequently $h(p,u) \leq \varepsilon_{n,\alpha}$. Because the $\mathcal{Q}_{n,\alpha}$ are nested, the function q is also contained in $\mathcal{Q}_{n,\underline{\beta}_n}$, so there exists an upper bracket $v \in \mathcal{U}_{n,\underline{\beta}_n}$ such that $h(q,v) \leq \varepsilon_{n,\underline{\beta}_n}$ and $h(q,v') \leq \varepsilon_{n,\underline{\beta}_n}$, where $v' \in \mathcal{P}_{n,\underline{\beta}_n}$ is a the renormalized v. Combination shows that $h(p,v') \leq h(p,u) + h(u,q) + h(q,v) + h(v,v') \leq 4\varepsilon_{n,\underline{\beta}_n} \leq 4\sqrt{H}\varepsilon_{n,\beta_n}$, implying that $h(p,\mathcal{P}_{n,\underline{\beta}_n}) \lesssim \varepsilon_{n,\beta_n}$.

We conclude that the union of the sets $\mathcal{P}_{n,\alpha}$ for $\alpha \geq \underline{\beta}_n$ is at a distance of a multiple of ε_{n,β_n} from $\mathcal{P}_{n,\underline{\beta}_n}$. Condition (2.7) therefore follows from (3.1) and the assumption that $E_{\underline{\beta}_n} = O(1)$. □

Remark 3.3 The set $\mathcal{Q}_{n,\alpha}$ in the preceding lemma can be the set of all probability densities in a multiple of the unit ball in a regularity space, but not all densities (because of the restriction to probability densities) and not all probability densities in the whole regularity space (because (3.1) would fail). It would be of interest to extend the theorem to models $\mathcal{Q}_{n,\alpha,M}$ indexed by a pair (α, M), where α can refer to regularity and give a nested scale of models and M can refer to a "multiple". For instance $\mathcal{Q}_{n,\alpha,M} = M\mathcal{Q}_{n,\alpha}$. This requires a result intermediate between Theorems 2.1 and 2.4. We omit a discussion.

3.1 Banach spaces

Consider for each $\alpha > 0$ a Banach space $\mathbb{B}^\alpha(\mathcal{X})$ of measurable functions $f : \mathcal{X} \to \mathbb{R}$ whose unit ball $\mathbb{B}_1^\alpha(\mathcal{X})$ processes finite upper bracketing numbers relative to the $L_2(\mu)$-norm, denoted by $\|\cdot\|_2$. Let the constants E_α and functions $H_\alpha : (0,\infty) \to (0,\infty)$

corresponds to the $L_2(\mu)$-distance on the roots of the elements of $\mathcal{Q}_\alpha$, the inequality (3.3) implies (3.1) for the rates $\varepsilon_{n,\alpha}$ satisfying

$$H_\alpha(\varepsilon_{n,\alpha}) = n\varepsilon_{n,\alpha}^2. \tag{3.4}$$

The root $\sqrt{p_0}$ of the true density belongs to the Banach space $\mathbb{B}^\beta(\mathcal{X})$, for a $\beta \in A$, if and only if $p_0 \in \cup_{M>0}(M\mathcal{Q}_\beta)$. Therefore, the priors chosen as in Theorem 3.1 yield the rate of convergence ε_{n,β_n} for any β_n such that $\sqrt{p_0} \in \mathbb{B}^{\beta_n}(\mathcal{X})$ and $\beta_n \in A_n$ eventually, under growth conditions on A_n (e.g. (2.10), (2.13) and (2.14)). Note that the prior construction does not use any information about the norm of $\sqrt{p_0}$ in $\mathbb{B}^{\beta_n}(\mathcal{X})$; it suffices that the square root of p_0 be contained in $\mathbb{B}^{\beta_n}(\mathcal{X})$.

Assuming that the root $\sqrt{p_0}$ of the true density is regular, rather than p_0 itself, is convenient in the preceding construction, because it allows to relate the (complicated) Hellinger distance to the $L_2(\mu)$-norm. However, it appears to be not merely a trick. For instance, if the scale of Banach spaces $\mathbb{B}^\alpha(\mathcal{X})$ corresponds to smoothness, then α-regularity of p_0 and its square root $\sqrt{p_0}$ are equivalent if p_0 is bounded away from zero, but not if p_0 can approach zero. It is intuitively clear that it is hard to estimate p_0 in regions where it is small, as few observations will land in such a region. This indeed can be shown rigorously, and removing the root from the regularity assumption, i.e. assuming that $p_0 \in \mathbb{B}^\alpha(\mathcal{X})$ instead of $\sqrt{p_0} \in \mathbb{B}^\alpha(\mathcal{X})$ will decrease the rate of the convergence. (Cf. Birgé (1986).) In the following we shall consider, more generally, the situation that $p_0^{1/s} \in \mathbb{B}^\alpha(\mathcal{X})$ for some $s \in [1,2]$, and consider adaptation to both s and α.

For every α and $s \in [1,2]$ let $\mathcal{Q}_{\alpha,s}$ be the set of all nonnegative functions $p\colon \mathcal{X} \to \mathbb{R}$ such that $p^{1/s} \in \mathbb{B}_1^\alpha(\mathcal{X})$. The optimal rate of convergence in this case changes to the solution $\varepsilon_{n,\alpha,s}$ of the equation

$$H_\alpha(\varepsilon_{n,\alpha,s}^{2/s}) = n\varepsilon_{n,\alpha,s}^2. \tag{3.5}$$

The case that $\sqrt{p_0}$ is regular corresponds to $s = 2$, and in that case this equation reduces to (3.4), with $\varepsilon_{n,\alpha,2} = \varepsilon_{n,\alpha}$. The claim follows from the following lemma.

Lemma 3.4 *A set $\{v_1, \ldots, v_N\}$ of upper $\varepsilon^{2/s}$-brackets over $\mathbb{B}_1^\alpha(\mathcal{X})$ for the $\|\cdot\|_2$-norm yields a set of upper ε-brackets $\{v_1^s, \ldots, v_N^s\}$ over $\mathcal{Q}_{\alpha,s}$ for the Hellinger distance. Consequently*

$$\log N_]\bigl(\varepsilon, \mathcal{Q}_{\alpha,s}, h\bigr) \le \log N_]\bigl(\varepsilon^{2/s}, \mathbb{B}_1^\alpha(\mathcal{X}), \|\cdot\|_2\bigr). \tag{3.6}$$

Proof: The function v is an upper $\varepsilon^{2/s}$-bracket for $p^{1/s}$ with respect to the $\|\cdot\|_2$-norm if $p^{1/s} \le v$ and $\|v - p^{1/s}\|_2 \le \varepsilon$. Then $p \le v^s$ and, by the inequality $|a^t - b^t| \le |a-b|^t$ for $0 \le t \le 1$,

It follows that the rate $\varepsilon_{n,\alpha,s}$ for the model with index (α, s), as determined by inequality (3.1), is the solution of the equation (3.5). Furthermore, given a set $\mathcal{V}_{n,\alpha,s}$ of upper $\varepsilon_{n,\alpha,s}^{2/s}$-brackets over $\mathbb{B}_1^\alpha(\mathcal{X})$ with respect to the $\|\cdot\|_2$-distance, the functions v^s for $v \in \mathcal{V}_{n,\alpha,s}$ are $\varepsilon_{n,\alpha,s}$-upper brackets over $\mathcal{Q}_{\alpha,s}$ and the corresponding priors $\Pi_{n,\alpha,s}$ of Theorem 3.1 are the uniform distributions on the sets of functions

$$\mathcal{P}_{n,\alpha,s} := \Big\{ \frac{v^s}{\int v^s \, d\mu} : v \in \mathcal{V}_{n,\alpha,s} \Big\}. \tag{3.7}$$

We may consider adaptation to the two indices α and s separately, for a fixed value of the other index, or to the pair of indices jointly.

If $\alpha \in (0, \infty)$ specifies a regularity level, then the unit balls of the Banach spaces are typically nested: $\mathbb{B}_1^\beta(\mathcal{X}) \subset \mathbb{B}_1^\alpha(\mathcal{X})$ for $\alpha \le \beta$. We shall assume this in the remainder of this section. For the power parameter s this may not be the case. However, the sets $Q_{\alpha,s}$ for different values of s are also very related. The following lemma establishes a bound on the bracketing entropy of a union of such spaces.

Lemma 3.5 *Assume that the uniform norm is bounded by the norm of $\mathbb{B}_1^\alpha(\mathcal{X})$. Then, for any α, and $1 \le s' < s'' \le 2$,*

$$\log N_{]}\big(\varepsilon + 2\sqrt{s''-s'}, \cup_{s' \le s < s''} \mathcal{Q}_{\alpha,s}, h\big) \le \log N_{]}\big(\varepsilon, \mathcal{Q}_{\alpha,s'}, h\big).$$

More precisely, a set of ε-brackets over $\mathcal{Q}_{\alpha,s'}$ is a set of $\varepsilon + 2\sqrt{s''-s'}$-brackets over $\cup_{s' \le s < s''} \mathcal{Q}_{\alpha,s}$.

Proof: If $p \in \mathcal{Q}_{\alpha,s}$ for $s' \le s < s''$, then $p^{1/s} \in \mathbb{B}_1^\alpha(\mathcal{X})$ and $p^{s'/s} \in \mathcal{Q}_{\alpha,s'}$. In view of the assumption on the norms, the first implies that $\|p^{1/s}\|_\infty \le 1$, and hence it is also true that $\|p\|_\infty \le 1$. Furthermore,

$$\frac{p}{p^{s'/s}} = (p^{1/s})^{s-s'} \le \|p^{1/s}\|_\infty^{s-s'} \le 1,$$

$$h^2(p, p^{s'/s}) \le \|p - p^{s'/s}\|_\infty \le \max_{u \in [0,1]} (u^{s'/s} - u) \le 4(s - s').$$

In the second last inequality we use that p takes its values in the interval $[0, 1]$; the last equality follows by an explicit calculation, where we use because $s' \le s \in [1, 2]$.

Suppose that $u_1, \ldots, u_N$ are ε-upper brackets over $\mathcal{Q}_{\alpha,s'}$. If $p \in \mathcal{Q}_{\alpha,s}$, then by the preceding paragraph $q = p^{s'/s}$ satisfies $p \le q$, $h(p, q) \le 2\sqrt{s - s'}$ and $q \in Q_{\alpha,s'}$, whence there exists an upper bracket u_i with $q \le u_i$ and $h(q, u_i) \le \varepsilon$. Together this implies that $u_1, \ldots, u_N$ are $\varepsilon + 2\sqrt{s - s'}$-upper brackets over $\mathcal{Q}_{\alpha,s}$. □

Suppose that $p_0^{1/t} \in \mathbb{B}^\beta(\mathcal{X})$ for some unknown $\beta \in A \subset (0, \infty)$ and $t \in [1, 2]$.

This is the weight function (2.9) with (α, s) playing the role of α and $\lambda_{\alpha,s} = 1$. For every $(\alpha, s) \in A_n \times S_n$ we employ the uniform prior $\Pi_{n,\alpha,s}$ on $\mathcal{P}_{n,\alpha,s}$ given by (3.7).

Assume without loss of generality that A_n and S_n are ordered, let α_0^n and $\alpha_{K_n+1}^n$ decrease to the infimum and supremum of A, $s_0^n = 1$, $s_{L_n+1}^n = 2$, and assume for every $j = 0, \ldots, K_n$ and $i = 0, \ldots, L_n$,

$$s_{i+1}^n - s_i^n \leq \min_{\alpha \in A_n} \varepsilon_{n,\alpha,s_i^n}^2, \tag{3.9}$$

$$\varepsilon_{n,\alpha_j^n,s_i^n} \lesssim \varepsilon_{n,\alpha_{j+1}^n,s_{i+1}^n}, \tag{3.10}$$

$$\#A_n \times \#S_n \lesssim \min_{\alpha \in A_n, s \in S_n} \exp[Cn\varepsilon_{n,\alpha,s}^2/36], \tag{3.11}$$

$$\max_{\alpha \in A_n, s \in S_n : \varepsilon_{n,\alpha,s}^2 \leq H\varepsilon_{n,\beta_n,t_n}^2} E_\alpha \frac{\varepsilon_{n,\alpha,s}^2}{\varepsilon_{n,\beta_n,t_n}^2} = O(1) \quad \forall H > 0. \tag{3.12}$$

In Example 2.9 grids satisfying the last three conditions are shown to exist in the case that the rates of convergence are of the form $n\varepsilon_{n,\alpha,s}^2 = n^{g(\alpha,s)}$ for a function that is strictly decreasing in its arguments, as is the case for the classical regularity spaces, and the constants E_α are bounded for α in bounded intervals. By extending this construction we can also ensure the first condition. In general, the first two conditions require that the grid $A_n \times S_n$ is rich enough, whereas the third is a (mild) upper bound on the size of $A_n \times S_n$.

Corollary 3.6 *Assume that the unit balls $\mathbb{B}_1^\alpha(\mathcal{X})$ satisfy (3.3), are nested and consist of functions that are uniformly bounded by 1. Suppose that $p_0^{1/t} \in \mathbb{B}^\beta(\mathcal{X})$ for some $\beta \in A \subset (0, \infty)$ and $t \in [1, 2]$. Let $\Pi_{n,\alpha,s}$ be as indicated. Let $A_n \subset A$ and $S_n \subset [1, 2]$ satisfy (3.9), (3.10), (3.11) and (3.12). Then the posterior distributions (1.2) relative to the weight functions λ_n given by (3.8) have rate of convergence at least $\varepsilon_{n,\beta,t}$ relative to the Hellinger distance.*

Proof: We apply Theorem 3.1 with $(\alpha, s) \in A_n \times S_n$ in the place of $\alpha \in A_n$, and with $\mathcal{Q}_{n,\alpha}$ of that theorem taken equal to

$$\overline{\mathcal{Q}}_{n,\alpha,s_i^n} = \cup_{s_i^n \leq s < s_{i+1}^n} \mathcal{Q}_{\alpha,s}.$$

By construction the prior Π_{n,α,s_i^n} is uniform on the rescaled $\varepsilon_{n,\alpha,s_i^n}$-upper brackets over $\mathcal{Q}_{\alpha,s_i^n}$, which are $3\varepsilon_{n,\alpha,s_i^n}$-upper brackets over $\overline{\mathcal{Q}}_{n,\alpha,s_i^n}$ by Lemma 3.5 and (3.9). Therefore, the priors are as in Theorem 3.1, with $\varepsilon_{n,\alpha}$ taken equal to three times the present $\varepsilon_{n,\alpha_j^n,s_i^n}$. (To be precise, by multiplication of the rate by a constant has the effect of divided the constant C in the prior weights (2.9) by a the square of the constant.)

Remark 3.7 In the preceding corollary we obtain the rate $\varepsilon_{n,\beta,t}$ under the condition that $p_0^{1/t} \in \mathbb{B}_1^\beta(\mathcal{X})$. Because this rate is faster for bigger β, we might want to apply the result for the "true" regularity level of a given density p_0 given by

$$\beta = \sup\bigl\{\alpha \in A\colon {p_0}^{1/t} \in \mathbb{B}^\alpha(\mathcal{X})\bigr\}.$$

If this supremum is attained, then indeed we can apply the corollary with β and obtain the rate $\varepsilon_{n,\beta,t}$. If the supremum is not attained, then we can apply the corollary or Theorem 3.1 with an approximating sequence β_n. For $\beta_n < \beta$ we are quaranteed that $p_0 \in M_n\mathcal{Q}_{\beta_n,t}$ for $M_n = \|p_0^{1/t}\|_{\beta_n}$ the norm of $p_0^{1/t}$ in $\mathbb{B}^{\beta_n}(\mathcal{X})$. If $\|{p_0}^{1/t}\|_\alpha = O(1)$ as $\alpha \uparrow \beta$, then M_n remains bounded as $\beta_n \uparrow \beta$ and hence the rate is $\varepsilon_{n,\beta_n,t}$ by Theorem 3.1, which will be of the same order as $\varepsilon_{n,\beta,t}$ if $\beta_n \uparrow \beta$ fast enough, even if $p_0^{1/t} \notin \mathbb{B}^\beta(\mathcal{X})$. However, it may be that $M_n = \|p_0^{1/t}\|_{\beta_n}$ increases indefinitely. By closer inspection the the rate can then be seen to be at least $M_n^{t/2}\varepsilon_{n,\beta_n,t}$, short of the rate $\varepsilon_{n,\beta,t}$.

Remark 3.8 Nestedness $\mathbb{B}_1^\beta(\mathcal{X}) \subset \mathbb{B}_1^\alpha(\mathcal{X})$ for $\alpha \le \beta$ of the unit balls can also be used within the context of Lemma 3.2, by defining the set $\mathcal{Q}_{\alpha,t}$ as the set of all probability densities p such that $p^{1/t}$ is contained in a fixed multiple of the unit ball $\mathbb{B}^\alpha(\mathcal{X})$, say $M\mathbb{B}_1^\alpha(\mathcal{X})$. Then if $p_0^{1/t} \in M\mathbb{B}_1^{\beta_n}(\mathcal{X})$, for some fixed t, we obtain the rate of convergence $\varepsilon_{n,\beta_n,t}$ as soon as $A_n \subset (0,\infty)$ is chosen to satisfy (2.10) and the constants E_α in (3.1) are bounded in bounded intervals. For instance, we may choose λ_α to be a finite measure on $A_n = \mathbb{Q}^+$, in which case $\varepsilon_{n,\beta_n,t}$ can typically represent any rate $\varepsilon_{n,\beta,t}$ for $\beta > 0$ with arbitrary precision.

3.1.1 Hölder spaces

A typical example of a scale of Banach spaces are the Hölder spaces $C^\alpha[0,1]^d$ of α-smooth functions $f\colon [0,1]^d \to \mathbb{R}$. Let $\mathcal{X} = [0,1]^d$ equipped with the Lebesgue measure μ. For $\alpha > 0$ let $\underline{\alpha}$ be the largest integer strictly smaller than α. The unit ball $C_1^\alpha[0,1]^d$ of the space $C^\alpha[0,1]^d$ consists of the functions $f \in C^\alpha[0,1]^d$ with partial derivatives of orders $0, 1, \ldots, \underline{\alpha}$ bounded by 1 and the partial derivatives of order $\underline{\alpha}$ Lipschitz of order $\alpha - \underline{\alpha}$ with Lipschitz constant 1.

From Kolmogorov and Tihomirov (1961) it is known that the entropy of this unit ball relative to the uniform norm satisfies

$$\log N\bigl(\varepsilon, C_1^\alpha[0,1]^d, \|\cdot\|_\infty\bigr) \le E_\alpha \varepsilon^{-d/\alpha}. \tag{3.13}$$

A ball of radius ε around a function f for the uniform norm yields a bracket $[f-\varepsilon, f+\varepsilon]$ of size 2ε for the uniform norm, and a-fortiori for the $L_2(\mu)$-norm. Hence

$$\log N_{[]}\bigl(\varepsilon, C_1^\alpha[0,1]^d, \|\cdot\|_2\bigr) \le N\bigl(\varepsilon/2, C_1^\alpha[0,1]^d, \|\cdot\|_\infty\bigr) \le E_\alpha(\varepsilon/2)^{-d/\alpha}.$$

If $\sqrt{p_0} \in C^\alpha[0,1]^d$ (corresponding to $s = 2$), the rate is $n^{-\alpha/(2\alpha+d)}$, which is well known to be optimal in the minimax sense. If it is only known that $p_0 \in C^\alpha[0,1]^d$ (corresponding to $s = 1$), then the rate is the slower $n^{-\alpha/(2\alpha+2d)}$, which again is optimal, as shown by Birgé (1986). The other cases give intermediate rates.

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 being constructive. Splines of an appropriate degree and dimension are one example. Given an $\varepsilon_{n,\alpha,s}^{2/s}$-net $\mathcal{V}_{n,\alpha,s}$, we let $\Pi_{n,\alpha,s}$ be the uniform prior on the functions (3.7), for any (α, s). Condition (3.1) is satisfied for $Q_{n,\alpha,s}$ the set of all densities p such that $p^{1/s} \in C^\alpha[0,1]^d$. To adapt to α or s it suffices to construct suitable weight functions λ_n, i.e. sets A_n and S_n of regularity indices and powers, and a measure λ.

Example 3.9 Within the context of Corollary 3.6 construct the sets A_n and S_n as uniform grids of sizes K_n and L_n over intervals $[\underline{\alpha}_n, \overline{\alpha}_n]$ and $[1,2]$. The fastest rate $\varepsilon_{n,\alpha,s}$ for $(\alpha, s) \in A_n \times S_n$ is $\varepsilon_{n,\overline{\alpha}_n,2} = n^{-\overline{\alpha}_n/(2\overline{\alpha}_n+d)}$. Therefore, condition (3.9) is satisfied as soon as $L_n \geq n^{\overline{\alpha}_n/(2\overline{\alpha}_n+d)}$. The numbers $\varepsilon_{n,\alpha,s}$ given by (3.14) satisfy

$$\log \frac{\varepsilon_{n,\alpha_j^n,s_i^n}}{\varepsilon_{n,\alpha_{j+1}^n,s_{i+1}^n}} = \Big[\frac{ds_i^n(\alpha_{j+1}^n - \alpha_j^n)/2}{(s_i^n\alpha_{j+1}^n + d)(s_i^n\alpha_j^n + d)} + \frac{d\alpha_{j+1}^n(s_{i+1}^n - s_i^n)/2}{(s_i^n\alpha_{j+1}^n + 1)(s_{i+1}^n\alpha_{j+1}^n + 1)}\Big] \log n.$$

Therefore, condition (3.10) is satisfied as soon as the meshwidths of the two grids are of order $O(1/\log n)$, i.e. as soon as $K_n \vee L_n \gtrsim \log n$. These two restrictions on the sizes K_n and L_n of the number of grid points are easily compatible with (3.11). The constants E_α can be shown to be bounded for α ranging over compact intervals, by inspection of the proofs in Kolmogorov and Tihomirov (1961) or van der Vaart and Wellner (1996). Because the present constants do not depend on s, condition (2.17) is satisfied, and hence the construction of Example 2.9 can be used to also satisfy (3.12).

4 Finite-dimensional models

LeCam (1973), Le Cam (1986) calls a model *finite-dimensional* if its local entropy function is bounded. In this section we consider a list of models $\mathcal{P}_{n,J,M}$ indexed by a dimension parameter $J \in \mathbb{N}$ and a second parameter $M \in \mathcal{M}$, such that, for every (J, M) and constants A_M, for every $\varepsilon > 0$,

$$\log N\Big(\frac{\varepsilon}{5}, C_{n,J,M}(2\varepsilon), d\Big) \leq A_M J. \tag{4.1}$$

Here the sets $C_{n,(J,M)}(\varepsilon)$ are the ones given in (1.3), for $\alpha = (J, M)$. Thus the models $\mathcal{P}_{n,J,M}$ are J-dimensional in the sense of Le Cam. Such finite-dimensional models may

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 $M_0 \in \mathcal{M}$,

$$d(p_0, \mathcal{P}_{n,J,M_0}) \lesssim \Big(\frac{1}{J}\Big)^{\beta}.$$

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 $J^{-2\beta} \sim J/n$. This is solved by $J \sim n^{1/(2\beta+1)}$ and would lead to an "optimal" rate of convergence $n^{-\beta/(2\beta+1)}$.

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 $1/\sqrt{n}$.

We shall show in this section that an adaptive Bayesian scheme, using fairly simple priors, can yield these optimal rates up to a logarithmic factor. This logarithmic factor can be avoided by using other schemes (e.g. based on a discretization of the coefficient space as in the preceding section, or a smooth prior on restricted coefficient space as in Huang (2004)), but we believe it cannot be removed from the simple construction used in this section. The advantage of the present priors is that they give adaptation across a wide range of regularity scales, and are easier to implement.

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 $\bar{B}_J\bar{C}_M\varepsilon$ 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,

$$\Pi_{n,J,M}\big(B_{n,J,M}(\varepsilon)\big) \geq \big(B_J C_M \varepsilon\big)^J, \quad \text{if } \varepsilon > D_M\, d(p_0, \mathcal{P}_{n,J,M}). \tag{4.2}$$

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 $\varepsilon \gtrsim d(p_0, \mathcal{P}_{n,J,M})$ is necessary, because by their definition the sets $B_{n,J,M}(\varepsilon)$ are centered around p_0, and this may be at a positive distance to $\mathcal{P}_{n,J,M}$. If $\varepsilon > 2d(p_0, \mathcal{P}_{n,J,M})$, then a ball of radius $\varepsilon/2$ around a projection of p_0 into $\mathcal{P}_{n,J,M}$ is contained in $C_{n,J,M}(\varepsilon)$. The general constant D_M in (4.2) instead of the universal constant 2 is meant to make up for the difference between the neighbourhoods $B_{n,J,M}(\varepsilon)$ and $C_{n,J,M}(\varepsilon)$.

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

$$\varepsilon_{n,J,M} = \sqrt{\frac{J \log n}{n} A_M A},$$

Theorem 4.1 *Suppose that (4.1)-(4.2) 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 $\log \#(\mathcal{J}_n \times \mathcal{M}_n) \lesssim n$ and $\sum_{M\in\mathcal{M}_n} e^{-LA_M} = O(1)$ for some $L > 0$. Then for every sequences $J_n \in \mathcal{M}_n$ and $M_n \in \mathcal{M}_n$ with $D_{M_n} d(p_0, \mathcal{P}_{n,J_n,M_n}) \leq \varepsilon_{n,J_n,M_n}$, there exists a K such that the posterior distribution relative to the weights $\lambda_{n,J,M}$ satisfies that $P_0^n \Pi_n(p\colon d(p,p_0) \geq K\varepsilon_{n,J_n,M_n}) \to 0$.*

Proof: We apply Theorem 2.4 with α equal to the pair (J, M) and $\beta_n = (J_n, M_n)$. Here we replace the global entropy condition by a local condition, as indicated in Remark 2.5.

Condition (2.15) is (easily) satisfied in virtue of the definition of the numbers $\varepsilon_{n,J,M}$, with $E_\alpha = 1$. The choices $E_\alpha = 1$ immediately give that (2.13) is satisfied, for every constant $H > 0$.

Because $D_{M_n} d(p_0, \mathcal{P}_{n,J_n,M_n}) \leq \varepsilon_{n,J_n,M_n}$ by assumption, condition (4.2) implies that the prior mass in the left side of (2.11) can be bounded below by

$$\left(B_{J_n} C_{M_n} \varepsilon_{n,J_n,M_n}\right)^{J_n} = e^{J_n \log(B_{J_n}\sqrt{J_n}) + \frac{1}{2} J_n \log(C_{M_n}^2 A_{M_n} A)} e^{-\frac{1}{2} J_n \log(n/\log n)}.$$

The first factor on the right is bounded below by 1 in view of the assumptions on the constants. Because $n\varepsilon_{n,J,M}^2 = J(\log n) A_M A$ and $A_M A \geq 1$, it follows that (2.11) is satisfied with $F = 1$.

Finally, we verify (2.10). Because presently $\lambda_\alpha = 1$, the left side of (2.10) takes the form

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

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

Example 4.2 (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 of convergence $\sqrt{(\log n)/n}$.

Example 4.3 (Regular true density.) If there exists a constant M_0 such that $d(p_0, \mathcal{P}_{n,J,M_0}) \lesssim J^{-\beta}$ 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 of convergence $(n/\log n)^{-\beta/(2\beta+1)}$.

Example 4.4 (Rough true density.) If there exists a constant M_0 such that $d(p_0, \mathcal{P}_{n,J,M_0}) \lesssim e^{-J^\beta}$ for every J, then we can apply the preceding theorem with J_n a multiple of $\log(n/\log n)^{1/\beta}$, yielding a rate of convergence $(\log n)^{1/\beta+1/2}/\sqrt{n}$.

4.1 Log Spline models

space of splines of order q relative to this partition is the set of all continuous functions $f\colon [0,1] \to \mathbb{R}$ that are $q-2$ times differentiable on $[0,1)$ and whose restriction to every of the partitioning intervals $\big[(k-1)/K, k/K\big)$ is a polynomial of degree strictly less than q. The set of these splines is a $J = q + K - 1$-dimensional vector space. A convenient basis is the set of B-splines $B_{J,1}, \ldots, B_{J,J}$, defined e.g. in de Boor (2001).

For $\theta \in \mathbb{R}^J$ let $\theta^T B_J = \sum_j \theta_j B_{J,j}$ and define

$$p_{J,\theta}(x) = e^{\theta^T B_J(x) - c_J(\theta)}, \qquad e^{c_J(\theta)} = \int_0^1 e^{\theta^T B_J(x)}\, dx.$$

Thus $p_{J,\theta}$ is a probability density that belongs to a J-dimensional exponential family with sufficient statistics the B-spline functions. Since the B-splines add up to unity, the family is actually of dimension $J-1$ and we can restrict θ to the subset of $\theta \in \mathbb{R}^J$ such that $\theta^T 1 = 0$.

We now consider models $\mathcal{P}_{J,M}$ indexed by pairs $(J,M) \in \mathbb{N}^2$ consisting of the spline densities $p_{J,\theta}$ with $\theta \in \Theta_{J,M} = \big\{\theta \in [-M,M]^J\colon \|\theta\|_\infty \le M, \theta^T 1 = 0\big\}$. Let the priors $\Pi_{n,J,M}$ be the distribution of $p_{J,\Theta}$ for Θ a random vector with an absolutely continuous distribution with a density of which the quotient of supremum and infimum on $\Theta_{J,M}$ is bounded by a fixed constant, for instance the uniform distribution.

Lemma 4.5 *Conditions (4.1) and (4.2) hold with the constants $A_M \ge M e^{K_1 M}$, $B_J = \sqrt{J} v_J^{1/J}$, $C_M = (K_3 M)^{-1} e^{-K_1 M^2}$ and $D_M = K_2 M$, where v_J is the volume of the J-dimensional unit ball and K_1, K_2 and K_3 are universal positive constants.*

Proof: Write $\|\theta\|_\infty$ and $\|\theta\|_2$ for the maximum norm and the Euclidean norm of $\theta \in \mathbb{R}^J$. Let $B_{J,M}(\varepsilon)$ and $C_{J,M}(\varepsilon)$ be the sets defined in (1.3), with $\alpha = (J,M)$ and the redundant n suppressed. >From the inequalities (and their derivations) on log spline densities given in Ghosal et al. (2003) (or alternatively Stone (1986) and Ghosal et al. (2000)), it can be obtained that there exist constants K_1 and K_2 such that

$$B_{J,M}(K_2 M\varepsilon) \supset C_{J,M}(\varepsilon),$$
$$\log N\big(\varepsilon/5, C_{J,M}(\varepsilon), h\big) \le M e^{K_1 M} J,$$

Furthermore, if $e^{K_1 M}\varepsilon \ge h(p_0, \mathcal{P}_{J,M})$ also there exists $\theta_{J,M} \in [-M,M]^J$ with

$$C_{J,M}(e^{K_1 M}\varepsilon) \supset \big\{p_{J,\theta}\colon \|\theta\|_\infty \le M, \|\theta - \theta_{J,M}\|_2 \le \sqrt{J}\varepsilon\big\}.$$

Condition (4.1) is immediate from this. By the assumption on the prior the prior probability of the set of θ in the right set of the last display is at least $(cM)^{-J}$ times its Euclidean volume. Claim (4.2) therefore follows from combination of the preceding inequalities.

If $\log p_0 \in C^\beta[0,1]$ and the order q of the splines is larger than β, then the minimizer $\bar\theta_J$ of $\theta \mapsto \|\log p_{J,\theta} - \log p_0\|_\infty$ over $\theta \in \mathbb{R}^J$ with $\theta^T 1 = 0$ satisfies

$$h(p_{J,\bar\theta_J}, p_0) \lesssim \|\log p_{J,\bar\theta_J} - \log p_0\|_\infty \lesssim J^{-\beta}. \tag{4.3}$$

(See Lemmas 5 and 7 in Ghosal et al. (2003).) Because $\|\log p_{J,\theta}\|_\infty \asymp \|\theta\|_\infty$ the vector $\bar\theta_J$ automatically has max-norm $\|\bar\theta_J\|_\infty$ bounded by a multiple of $\|\log p_0\|_\infty$. This implies that a positive density $p_0 \in C^\beta[0,1]$ can be approximated with an error of order $(1/J)^\beta$ by a log spline density in $\mathcal{P}_{J,M_0}$ if M_0 is sufficiently large. The preceding theorem and example then give a rate of contraction of $(n/\log n)^{-\beta/(2\beta+1)}$ for the posterior. This rate is the optimal one in the minimax sense up to the logarithmic factor. Although we have proved only an upper bounded, the rate of contraction of the present posterior appears to contain an additional logarithmic factor indeed. This is due to spreading the prior mass smoothly over the coefficient space.

This complements the result of Ghosal et al. (2003), who considered adaptation to a finite set of regularity levels assuming a fixed and known upper bound M on the absolute values of the log densities.

5 Auxiliary lemmas and proofs

The following lemmas are taken from Ghosal et al. (2000), and are used in the proofs of the main results. The first lemma gives a sufficient condition for the existence of certain tests in terms of the local entropy of a statistical model. The lemma is proved in Ghosal et al. (2000), following work by LeCam (1973) and Birgé (1983). The numbers $D(\varepsilon)$ in the condition of the following lemma are related to the measures of dimension used by these authors. Up to constants Le Cam (1986) calls the numbers $\sup_{\varepsilon>\varepsilon_n} N(\varepsilon)$ the *dimension of $\mathcal{P}$ for the pair (d, ε_n)*.

Lemma 5.1 *Suppose that for some nonincreasing function $N(\varepsilon)$ and some $\varepsilon_n \geq 0$,*

$$\sup_{\varepsilon>\varepsilon_n} N\Big(\frac{\varepsilon}{3}, \big\{p \in \mathcal{P}\colon \varepsilon \leq d(p,p_0) \leq 2\varepsilon\big\}, d\Big) \leq N(\varepsilon), \qquad \varepsilon > \varepsilon_n.$$

Then for every $\varepsilon > \varepsilon_n$ there exist tests ϕ_n (depending on p_0 and ε but not on i) such that, for a universal constant K and every $i \in \mathbb{N}$,

$$P_0^n \phi_n \leq N(\varepsilon) e^{-Kn\varepsilon^2} \frac{1}{1 - e^{-Kn\varepsilon^2}},$$

$$\sup_{p \in \mathcal{P}: d(p,p_0) > i\varepsilon} P^n(1-\phi_n) \leq e^{-Kn\varepsilon^2 i^2},$$

Lemma 5.2 *For every $\varepsilon > 0$ and probability measure Π on the set*

$$\Big\{p \in \mathcal{P}\colon P_0 \log \frac{p_0}{p} \leq \varepsilon^2, P_0\Big(\log \frac{p_0}{p}\Big)^2 \leq \varepsilon^2\Big\},$$

Lemma 5.3 *For any pair of probability measures P and P_0,*

$$h^2(p,p_0) \le P_0 \log \frac{p_0}{p} \le 2h^2(p,p_0)\Big[1+\log\Big\|\frac{p_0}{p}\Big\|_\infty\Big] \le 2h^2(p,p_0)\Big\|\frac{p_0}{p}\Big\|_\infty,$$
$$P_0\Big(\log\frac{p_0}{p}\Big)^2 \lesssim h^2(p,p_0)\Big[1+\log\Big\|\frac{p_0}{p}\Big\|_\infty\Big]^2.$$

Lemma 5.4 *For every $b>0$ there exists a constant $\varepsilon_b>0$ such that for every pair of probability measures P and P_0 with $0<h^2(p,p_0)<\varepsilon_b P_0(p_0/p)^b$,*

$$P_0 \log\frac{p_0}{p} \lesssim h^2(p,p_0)\Big(1+\frac{1}{b}\log_+\frac{1}{h(p,p_0)}+\frac{1}{b}\log_+ P_0\Big(\frac{p_0}{p}\Big)^b\Big),$$
$$P_0\Big(\log\frac{p_0}{p}\Big)^2 \lesssim h^2(p,p_0)\Big(1+\frac{1}{b}\log_+\frac{1}{h(p,p_0)}+\frac{1}{b}\log_+ P_0\Big(\frac{p_0}{p}\Big)^b\Big)^2.$$

Lemma 5.5 *If v_J is the volume of the J-dimensional unit ball, then $J\mapsto \sqrt{J}^J v_J$ is increasing, and, as $J\to\infty$,*

$$\sqrt{J}^J v_J = \frac{\sqrt{J}^J\sqrt{\pi}^J}{\Gamma(J/2+1)} = \frac{\sqrt{2\pi e}^J}{\sqrt{\pi J}}(1+o(1)).$$

Proof:of Theorem 2.1 Abbreviate $\mathcal{P}_{n,\gtrsim\beta} = \cup_{\alpha\in A_{n,\gtrsim\beta_n,H}}\mathcal{P}_{n,\alpha}$ and let $\mathcal{P}_{n,<\beta}$ refer to the union of the $\mathcal{P}_{n,\alpha}$ for α in the complementary part A_n. Furthermore, set $J_{n,\alpha}=n\varepsilon_{n,\alpha}^2$. In view of assumption (2.7), we have, for every $\varepsilon\ge 3\underline{E}\varepsilon_{n,\beta_n}$,

$$N\Big(\frac{\varepsilon}{3},\{p\in\mathcal{P}_{n,\gtrsim\beta_n}\colon \varepsilon<d(p,p_0)<2\varepsilon\},d\Big) \le N\big(\underline{E}\,\varepsilon_{n,\beta_n},\mathcal{P}_{n,\gtrsim\beta_n},d\big) \le e^{EJ_{n,\beta_n}}.$$

Therefore, by Lemma 5.1 with $\varepsilon = M\varepsilon_{n,\beta_n}$ and $N(\varepsilon)=\exp(EJ_{n,\beta_n})$ and sufficiently large M, there exists for each n a test ϕ_n such that for a universal constant K and any $i\in\mathbb{N}$,

$$\begin{aligned} P_0^n\phi_n &\le 3e^{(E-KM^2)J_{n,\beta_n}},\\ \sup_{p\in\mathcal{P}_{n,\gtrsim\beta_n}:d(p,p_0)\ge iM\varepsilon_{n,\beta_n}} P^n(1-\phi_n) &\le e^{-KM^2i^2J_{n,\beta_n}}. \end{aligned} \tag{5.1}$$

We choose M sufficiently large, so that the right side of the first equation tends to zero. Then $P_0^n\phi_n\Pi_n\big(p\colon d(p,p_0)\ge M\varepsilon_{n,\beta_n}|X_1,\ldots,X_n\big)\le P_0^n\phi_n\to 0$.

For every α in $U_n\colon= U_n(\beta_n,T,\underline{F},F)$ as given in (2.6), we have the inequality $\Pi_{n,\alpha}\big(B_{n,\alpha}(\underline{F}\varepsilon_{n,\alpha})\big)\ge\exp[-FTJ_{n,\beta_n}]$ and, therefore, in view of (2.3),

where $\Lambda_n := \int \exp[-Cn\varepsilon_{n,\alpha}^2]\,\lambda_n(d\alpha)$. By Lemma 5.2 with $C = 1$ and $\varepsilon = \underline{F}\varepsilon_{n,\beta_n}$, and because $n\varepsilon_{n,\beta_n}^2 \to \infty$, there exist events E_n with $P_0^n(\mathrm{E}_n) \to 1$ and on E_n,

$$\begin{aligned}\int \prod_{i=1}^n \frac{p}{p_0}(X_i)\,d\Pi_n(p) &\geq e^{-2\underline{F}^2 J_{n,\beta_n}}\,\Pi_n\Big(\bigcup_\alpha B_{n,\alpha}(\underline{F}\varepsilon_{n,\beta_n})\Big)\\ &\geq e^{-2\underline{F}^2 J_{n,\beta_n}}\,\Pi_n\big(\bigcup_{\alpha\in U_n} B_{n,\alpha}(\varepsilon_{n,\alpha})\big)\\ &\geq \exp[-(2\underline{F}^2 + FT + CT)J_{\beta_n,n}]\frac{\lambda(U_n)}{\Lambda_n},\end{aligned}$$

where the last inequality follows from inequality (5.2). Using this lower bound for the denominator of the posterior measure (1.2), Fubini's theorem and next the inequality $P_0^n\prod_{i=1}^n(p/p_0)(X_i)(1-\phi_n) \leq P^n(1-\phi_n)$ and (5.1), we see that

$$\begin{aligned}&P_0^n\big[(1-\phi_n)1_{\mathrm{E}_n}\Pi_n\big(p\in\mathcal{P}_{n,\gtrsim\beta_n}\colon d(p,p_0)\geq M\varepsilon_{n,\beta_n}|X_1,\ldots,X_n\big)\big]\\ &\quad\leq \frac{\Lambda_n}{\lambda(U_n)}e^{(2\underline{F}^2+CT+FT)J_{n,\beta_n}}\int_{p\in\mathcal{P}_{n,\gtrsim\beta_n}\colon d(p,p_0)\geq M\varepsilon_{n,\beta_n}} P_0^n\prod_{i=1}^n\frac{p}{p_0}(X_i)(1-\phi_n)\,d\Pi_n(p)\\ &\quad\leq \frac{\Lambda_n}{\lambda(U_n)}e^{(2\underline{F}^2+CT+FT)J_{n,\beta_n}}e^{-KM^2J_{n,\beta_n}}\Pi_n\big(\mathcal{P}_{n,\gtrsim\beta_n}\big)\\ &\quad\leq e^{(2\underline{F}^2+CT+FT-KM^2)J_{n,\beta_n}}\frac{\int_{\alpha\gtrsim\beta_n}e^{-CJ_{n,\alpha}}\lambda(d\alpha)}{\lambda(U_n)}.\end{aligned}$$

For sufficiently large M the leading exponential term converges to zero. Furthermore, the second term is bounded by assumption (2.8) (even with an additional factor $1/4$ in the exponent).

Using that $P_0^n\prod_{i=1}^n(p/p_0)(X_i) \leq 1$ and again Fubini's theorem, we also see that

$$\begin{aligned}&P_0^n(1-\phi_n)1_{\mathrm{E}_n}\Pi_n\big(p\in\mathcal{P}_{n,<\beta_n}\colon d(p,p_0)\geq M\varepsilon_{n,\beta_n}|X_1,\ldots,X_n\big)\\ &\quad\leq \frac{\Lambda_n}{\lambda(U_n)}e^{(2\underline{F}^2+CT+FT)J_{n,\beta_n}}\Pi_n(\mathcal{P}_{n,<\beta_n})\\ &\quad\leq e^{(2\underline{F}^2+CT+FT)J_{n,\beta_n}}\frac{\int_{\alpha<\beta_n}e^{-CJ_{n,\alpha}}\lambda(d\alpha)}{\lambda(U_n)}\\ &\quad\leq e^{-(2\underline{F}^2+CT+FT)J_{n,\beta_n}/2}\frac{\int_{\alpha<\beta_n}e^{-CJ_{n,\alpha}/4}\lambda(d\alpha)}{\lambda(U_n)},\end{aligned}$$

where the last inequality follows from the bound $CJ_{n,\alpha} \geq CHJ_{n,\alpha} \geq 2(2F^2 + CT +$

6 Acknowledgements

The greatest part of this research was carried out at Eurandom in Eindhoven in 2002-2003.
J. Lember is supported by Estonian Science Foundation Grant 5694.

References

Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999. ISSN 0178-8051.

Andrew R. Barron and Thomas M. Cover. Minimum complexity density estimation. *IEEE Trans. Inform. Theory*, 37(4):1034–1054, 1991. ISSN 0018-9448.

Lucien Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete*, 65(2):181–237, 1983. ISSN 0044-3719.

Lucien Birgé. On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Relat. Fields*, 71(2):271–291, 1986. ISSN 0178-8051.

Carl de Boor. *A practical guide to splines*, volume 27 of *Applied Mathematical Sciences*. Springer-Verlag, New York, revised edition, 2001. ISBN 0-387-95366-3.

David L. Donoho, Iain M. Johnstone, Gérard Kerkyacharian, and Dominique Picard. Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B*, 57(2):301–369, 1995. ISSN 0035-9246. With discussion and a reply by the authors.

David L. Donoho, Iain M. Johnstone, Gérard Kerkyacharian, and Dominique Picard. Density estimation by wavelet thresholding. *Ann. Statist.*, 24(2):508–539, 1996. ISSN 0090-5364.

S. YU. Efromovich and M.S. Pinsker. Learning algorithm for nonparametric filtering. *Autom. Remote Control*, 11:1434–1440, 1984.

Subhashis Ghosal, Jayanta K. Ghosh, and Aad W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 2000. ISSN 0090-5364.

Subhashis Ghosal, Jüri Lember, and Aad Van Der Vaart. On Bayesian adaptation. In *Proceedings of the Eighth Vilnius Conference on Probability Theory and Mathematical Statistics, Part II (2002)*, volume 79, pages 165–175, 2003.

G. K. Golubev. Adaptive asymptotically minimax estimates for smooth signals. *Problemy Peredachi Informatsii*, 23(1):57–67, 1987. ISSN 0555-2923.

Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. Springer-Verlag, New York, 1986. ISBN 0-387-96307-3.

L. LeCam. Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1: 38–53, 1973. ISSN 0090-5364.

O. V. Lepskiĭ. A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.*, 35(3):459–470, 1990. ISSN 0040-361X.

O. V. Lepskiĭ. Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.*, 36(4):645–659, 1991. ISSN 0040-361X.

O. V. Lepskiĭ. Asymptotically minimax adaptive estimation. II. Schemes without optimal adaptation. Adaptive estimates. *Teor. Veroyatnost. i Primenen.*, 37(3):468–481, 1992. ISSN 0040-361X.

Michael Nussbaum. Spline smoothing in regression models and asymptotic efficiency in L_2. *Ann. Statist.*, 13(3):984–997, 1985. ISSN 0090-5364.

Loraine Schwartz. On consistency of Bayes procedures. *Proc. Nat. Acad. Sci. U.S.A.*, 52:46–49, 1964.

Charles J. Stone. An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.*, 12(4):1285–1297, 1984. ISSN 0090-5364.

Charles J. Stone. The dimensionality reduction principle for generalized additive models. *Ann. Statist.*, 14(2):590–606, 1986. ISSN 0090-5364.

Charles J. Stone. Large-sample inference for log-spline models. *Ann. Statist.*, 18(2): 717–741, 1990. ISSN 0090-5364.

Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. ISBN 0-387-94640-3. With applications to statistics.

Jüri Lember
University of Tartu
Institute of Mathematical Statistics
J. Liivi 2
50409 Tartu
ESTONIA
juri.lember@ut.ee

Aad van der Vaart
Department of Mathematics
Vrije Universiteit Amsterdam
De Boelelaan 1081
1081 HV Amsterdam
NETHERLANDS
aad@cs.vu.nl