Electronic Journal of Statistics
Vol. 6 (2012) 1984–2001
ISSN: 1935-7524
DOI: 10.1214/12-EJS735

Adaptive estimation of multivariate functions using conditionally Gaussian tensor-product spline priors

R. de Jonge

Department of Mathematics
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven
The Netherlands
e-mail: r.d.jonge@tue.nl

and

J.H. van Zanten

Korteweg-de Vries Institute for Mathematics
University of Amsterdam
P.O. Box 94248, 1090 GE Amsterdam
The Netherlands
e-mail: j.h.vanzanten@uva.nl

Abstract: We investigate posterior contraction rates for priors on multivariate functions that are constructed using tensor-product B-spline expansions. We prove that using a hierarchical prior with an appropriate prior distribution on the partition size and Gaussian prior weights on the B-spline coefficients, procedures can be obtained that adapt to the degree of smoothness of the unknown function up to the order of the splines that are used. We take a unified approach including important nonparametric statistical settings like density estimation, regression, and classification.

AMS 2000 subject classifications: Primary 62C10; secondary 62G20.
Keywords and phrases: Nonparametric Bayes procedure, tensor-product splines, posterior contraction rate, adaptive estimation.

Received February 2012.

1. Introduction

In recent years the use of Bayesian methods in nonparametric function estimation problems has increased considerably. It is by now well known however that the construction of sensible priors on functional, infinite-dimensional parameters is a delicate matter. Intuition is often not enough to guarantee important properties like consistency and optimal convergence rates of nonparametric Bayes procedures. Over the last decade a mathematical framework has been developed to study these frequentist concepts for Bayesian methods, starting with

[*]Research supported by the Netherlands Organization for Scientific Research NWO.

the papers [6] and [19]. Concrete families of nonparametric priors for which consistency, contraction rates and related matters like adaptation have been investigated include priors based on Dirichlet processes, Bernstein polynomials, kernel mixture priors, beta mixtures, Gaussian process priors, wavelet based priors, etc. See for instance the papers [8, 23, 24, 16, 11, 2, 13, 5], to mention but a few.

In this work we consider prior distributions on functions of one or more variables that are constructed using so-called splines. A spline function is a piecewise polynomial function on either an interval in the real line or some multi-dimensional Euclidean space. Spline functions provide good approximations for Hölder smooth functions, see for instance De Boor [1] or Schumaker [17]. Therefore, splines can be a useful tool for constructing prior distributions on smooth functions.

There are several papers in the literature that obtain rates of estimation for smooth functions using splines, in particular ones using log-spline models in a density estimation setting. It was shown for instance by Stone [20] that a smooth probability density can be estimated at the minimax rate in a log-spline model of growing dimension using a maximum likelihood estimator (MLE). This was extended in [21] to the multivariate case. A Bayesian version of the former result was obtained by Ghosal, Ghosh and Van der Vaart [6]. They consider priors on densities that are constructed by postulating the same log-spline model for the density as in Stone and putting an appropriate prior distribution on the coefficients in the B-spline expansion of the log-density.

Ghosal, Ghosh and Van der Vaart [6] show that it is possible to attain the minimax rate as the posterior contraction rate if the log-density is bounded (by a known constant) and satisfies a smoothness condition. Specifically, the results state that in the univariate case that a sample from an unknown density f on an interval is observed, then if $\log f$ is uniformly bounded by a known constant and is r times continuously differentiable, a rate of convergence relative to the Hellinger metric (for the MLE in the case of Stone [20] and for the posterior in the case of Ghosal, Ghosh and Van der Vaart [6]) of the optimal order $n^{-r/(1+2r)}$ can be attained. Stone [21] also obtains the optimal rate $n^{-r/(d+2r)}$ in the case that f is a d-variate density. The procedures in the cited papers are non-adaptive, in the sense that they rely on knowledge of the smoothness level r of the unknown density.

Rate-adaptive results for spline priors have been obtained by Huang [10] and by Ghosal, Lember and Van der Vaart [7]. The paper [7] deals with univariate density estimation again. Instead of letting the dimension J of the log-spline expansion tend to infinity with sample size in a deterministic manner, the "model index" J is viewed as a hyper-parameter and is endowed with an additional prior. Put differently, the density estimation problem is viewed as a model selection problem: a sequence of finite-dimensional log-spline models for the density is considered, each with there own (finite-dimensional) prior. Then appropriate prior weights are assigned to each of the models to obtain an overall prior for f. The resulting hierarchical prior does not depend on the regularity r of the density f, but it still yields a posterior contraction rate of the order

$n^{-r/(1+2r)}$ if $\log f$ is r times continuously differentiable. Huang [10] presents a very similar adaption result, but with more complicated prior weights on the finite-dimensional models. This is accompanied by a similar result in a univariate nonparametric regression context. The two settings in [10] are not treated in a unified approach however. Priors weights for the models are chosen separately for each case.

A joint feature of the approaches of [10] and [7] is that both the order and the knots of the splines (see the next section for definitions of these notions) are changing between models. In view of the approximation properties of splines (see Section 2), allowing the orders of the splines to become arbitrarily large is indeed necessary when adaption to arbitrarily large smoothness levels is desired. On the other hand, it makes the priors rather involved and might be less attractive from the computational perspective. Implementation schemes that have been proposed in the literature typically take the order of the splines fixed, cf. e.g. [3, 4].

Our approach and the results we derive complement and extend the existing literature in a number of directions. First of all, we do not study specific settings like density estimation and regression separately. Instead, we present general theorems about random spline processes (Theorems 4.1 and 4.2) that, in combination with existing general rate of contraction results for specific statistical settings such as given for instance in [6, 9, 23, 22], or [12], lead to concrete results for, for instance, density estimation, regression, classification, or drift estimation for diffusions.

Secondly, we consider multivariate function estimation problems. Similar to what Stone [21] did for the frequentist approach, we show that sensible priors on multivariate functions can be constructed using tensor-product splines. We prove that adaptive, rate-optimal procedures for multivariate function estimation problems can be obtained in this way.

Another difference concerns the fact that the existing approaches in [6, 10] and [7] assume known uniform bounds on the log-density or the regression function that is being estimated, allowing the use of bounded priors on the B-spline coefficients. As is indicated in [7] this restriction could be removed by adding another hierarchical layer, treating the bound as an additional hyper parameter. In our approach this is not necessary however and we do not need to assume any uniform bounds. This is a consequence of the fact that we use unbounded, namely Gaussian prior weights on the B-spline coefficients. In our rates we get additional logarithmic factors, which might in part be due to this issue.

Finally, we keep the order of the splines that we use fixed in the construction of the prior. Only the number of knots is viewed as a hyper-parameter, which we either send to infinity with sample size or endow with a prior. As a result our priors are simpler and conceivably also computationally more attractive. On the down side, with this approach we can not obtain adaption up to arbitrary high smoothness levels, but only up to the order of the splines that are used. Since we can freely choose this order however, we feel this is not a serious restriction.

As mentioned already, we build our spline priors from random splines with independent, Gaussian B-spline coefficients. We keep the order of the splines fixed and treat the number of knots as a hyper-parameter. The latter will be either deterministic, or endowed with a second, independent prior. As a result, the priors we construct will be (transformations of) Gaussian or conditionally Gaussian process priors. This allows us to use powerful tools from the general theory of Gaussian process priors (cf. [23, 25]) for their analysis.

A different approach is taken in the recent papers [15] and [18]. Both papers deal with series priors as well. The former considers univariate density estimation and considers series priors with Fourier or wavelet basis functions. The latter studies univariate estimation problems in a general framework, allowing also different kinds of basis functions. The papers show that desirable results can also be obtained with non-Gaussian prior weights. It is conceivable that this is true for tensor product spline priors as well, but proving this would require a different technical approach than the one we take in this paper.

The remainder of the paper is organized as follows. In Section 2 we review the notions of spline functions and B-splines, and formulate a result that gives a bound on the uniform distance between splines and a given smooth function. In Section 3 we define our spline process with Gaussian coefficients and derive small ball probability bounds that will be key ingredients for the rate of contraction results. In Section 4.1 we show that optimal posterior rates (up to logarithmic factors) can be achieved using Gaussian spline priors by letting the number of knots tend to infinity with sample size in an appropriate way. In Section 4.2 we present a hierarchical procedure by choosing a prior distribution on the partition size hyper-parameter. We show that this hierarchical procedure also achieves a near-optimal rate of posterior contraction and adapts to the smoothness of the truth.

2. Preliminaries on splines

In this section we recall some necessary elements of the theory of splines. We follow the definitions given by Schumaker [17] and refer to that book for an exhaustive treatment of the subject.

2.1. Spline functions on intervals

A *spline function* or *spline* is a piecewise polynomial function that satisfies a global smoothness condition. Let us first consider spline functions defined on an interval. The domain of such a spline function can be partitioned into disjoint subintervals in such a way that the function coincides with a polynomial on every subinterval. A spline function is said to be of *order* q if all polynomials in its definition are of degree at most $q-1$. Without any further requirements, this set of piecewise polynomials is a linear space of dimension qm, where m is the number of partitioning intervals. Linear subspaces of lower dimension can be obtained by further imposing that adjacent polynomials are tied together smoothly at the knots of the partition.

In this paper we use splines of order q that satisfy such a smoothness condition. We consider a space S_m of splines of order q on the unit interval that is partitioned into m subintervals of equal length. We first define

$$P_q = \Big\{x \mapsto \sum_{k=0}^{q-1} c_k x^k, c_0, \ldots, c_{q-1} \in \mathbb{R}\Big\}$$

the space of polynomials of degree at most $q-1$. Let $y_j = j/m$ and denote the corresponding subintervals of $[0,1]$ by $I_j = [y_{j-1}, y_j)$ for $j = 1, \ldots, m-1$, and $I_m = [y_{m-1}, 1]$. A function $s : [0,1] \to \mathbb{R}$ is then defined to be in S_m if there exist polynomials $p_1, \ldots, p_m$ in P_q such that $s(x) = p_j(x)$ for $x \in I_j$ and, moreover, s is $q-2$ times continuously differentiable[1]. According to the terminology of [17], S_m is the space of *polynomial splines of order q with simple knots at the points* $1/m, 2/m, \ldots, (m-1)/m$. We will always take $q \geq 2$, so that all the splines in S_m are continuous functions.

The space S_m has dimension $q + m - 1$, cf. Theorem 4.4 of [17]. A convenient basis of the space is given by the so-called *B-splines*. The exact definition of these functions (see Theorem 4.9 of Schumaker [17]) is of no importance to us. Important properties of B-splines are that they are nonnegative and supported on relative small parts of the domain and that the sum of all B-splines at any given location equals one, i.e. they form a partition of unity: if we denote the B-splines by $B_1^m, \ldots, B_{q+m-1}^m$, then

$$\sum_{j=1}^{q+m-1} B_j^m(x) = 1$$

for all $x \in [0,1]$.

2.2. Tensor-product splines

Spline functions can also be defined on multi-dimensional domains using multivariate polynomials. One can construct linear spaces of such multivariate splines by taking tensor-products of the spline spaces mentioned above. This just means that we associate a direction with every linear space in the tensor product, that we introduce a different variable for each direction, and that we then multiply polynomials of a single variable defined on intervals to obtain multivariate polynomials defined on rectangles.

The space of tensor-product splines is spanned by the *tensor-product B-splines*, which are just products of the B-splines associated with the different directions. The dimension of the tensor-product space is thus found by multiplying the dimensions of the spline spaces from which it was constructed. The properties of univariate B-splines carry over to similar properties for their tensor-product analogues.

In the following we consider tensor-product splines from the d-fold tensor product space $\mathcal{S}_m = S_m \otimes \cdots \otimes S_m$ (d times), with S_m the space of univariate

[1] Here -1 times continuous differentiability is an empty condition and 0 times just means continuity.

splines defined above. The tensor-product splines are thus defined on the unit cube $[0,1]^d$ in the Euclidean space of dimension d and this unit cube is partitioned into m^d equal cubes $I_{k_1} \times \cdots \times I_{k_d}$. On every such set the splines coincide with a polynomial of the form

$$\sum_{k_1=0}^{q-1} \cdots \sum_{k_d=0}^{q-1} c_{k_1,\dots,k_d} x_1^{k_1} \cdots x_d^{k_d}. \tag{2.1}$$

The space $\mathcal{S}_m$ has dimension $(q+m-1)^d$ and a basis is given by the tensor-product B-splines

$$B_j^m(x_1,\dots,x_d) = B_{j_1}^m(x_1) \cdots B_{j_d}^m(x_d), \quad 1 \le j_i \le q+m-1.$$

From now on these multivariate B-splines are denoted by $B_1^m,\dots,B_J^m$ for $J = (q+m-1)^d$. It is easy to see that we again have the partition of unity property

$$\sum_{j=1}^{J} B_j^m(x) = 1 \tag{2.2}$$

for all $x \in [0,1]^d$.

The total degree of a polynomial of the form (2.1) is the maximum of $|k| = k_1 + \cdots + k_d$ over all k for which the coefficient c_k is nonzero. The total degree of these polynomials is thus at most $d(q-1)$, but not any polynomial of total degree at most $d(q-1)$ is an element of $\mathcal{S}_m$. This is only true if the degree in each single variable $x_1,\dots,x_d$ is at most $q-1$. In particular the polynomials of total *order* q are in $\mathcal{S}_m$, i.e. the polynomials of the form (2.1) with $c_k = 0$ if $|k| > q-1$. The approximating properties of such polynomials determine the approximating capabilities of the tensor-product splines in $\mathcal{S}_m$, see Lemma 2.1 ahead.

This approximation result is proved using a dual basis of the tensor-product space. Given a set of linear functionals $\lambda_j : \mathcal{S}_m \to \mathbb{R}$, we say that $\lambda_1,\dots,\lambda_J$ is a dual basis of $\mathcal{S}_m$ if

$$\lambda_i B_j^m = \delta_{i,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

for any $i,j = 1,\dots,J$. For the spline $s \in \mathcal{S}_m$ given by

$$s = \sum_{j=1}^{J} a_j B_j^m \tag{2.3}$$

we have that $\lambda_j(s) = a_j$. Thus λ_j finds the coefficient belonging to the B-spline B_j^m.

2.3. Approximation properties

The following result describes how well splines in the space $\mathcal{S}_m$ can approximate functions with a smoothness level r that does not exceed the order q of the splines. We first explain what the appropriate notion of smoothness is in this situation.

Let $C([0,1]^d)$ be the space of continuous functions $f : [0,1]^d \to \mathbb{R}$ and denote the supremum norm of f over $[0,1]^d$ by $\|f\|_\infty$. For a multi-index $\alpha = (\alpha_1, \dots, \alpha_d)$, we define $|\alpha| = \alpha_1 + \cdots + \alpha_d$ and the partial derivative

$$D^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}.$$

For $r \in \mathbb{N}$, we define the Hölder space $C^r([0,1]^d)$ of all functions $f \in C([0,1]^d)$ with partial derivatives $D^\alpha f \in C([0,1]^d)$ for any $|\alpha| \le r$, and we equip it with the norm

$$\|f\|_{C^r} = \|f\|_\infty + \sum_{\alpha:|\alpha|=r} \|D^\alpha f\|_\infty.$$

The lemma below gives an upper bound on the uniform distance of a function $f \in C^r([0,1]^d)$ and some spline in $\mathcal{S}_m$. The distance can be controlled by choosing the partition size m sufficiently large. The proof of the lemma is similar to the proof of Theorem 12.7 in Schumaker [17]. We only need to apply the multidimensional Taylor expansion in Theorem 13.18 of Schumaker with a total Taylor expansion (13.33) in [17] instead of a tensor Taylor expansion (13.44), so that this expansion produces a polynomial of total order r.

Lemma 2.1. *For any $m, d, q \in \mathbb{N}$, $r \le q$, and $f \in C^r([0,1]^d)$ there exists a spline $s \in \mathcal{S}_m$ and a constant $C > 0$ that only depends on d, q and r such that*

$$\|f - s\|_\infty \le Cm^{-r} \sum_{\alpha:|\alpha|=r} \|D^\alpha f\|_\infty Cm^{-r} \|f\|_{C^r}.$$

Proof. Let Q be the bounded linear operator in (12.29) of [17] that maps $C^r([0,1]^d)$ onto $\mathcal{S}_m$. It is given by $Qf(x) = \sum_{j=1}^J (\lambda_j f) B_j(x)$, for λ_j the (extensions of the) elements of the dual space of $\mathcal{S}_m$ given in Theorem 12.5 of [17]. Let H be a hypercube in the partition of $[0,1]^d$ and let $\|\cdot\|$ be the supremum over H. We will bound $\|f - Qf\|$ from above. It is obvious that $\|f - Qf\|_\infty$ is then bounded from above by the maximum of these bounds for the various cubes in the partition.

We have $\|Qf\| \le C\|f\|$ for any $f \in C^r([0,1]^d)$ according to (12.31) of [17]. The constant C does not depend on the cube H as can be seen from (12.25) of [17], but it does depend on q. According to Theorem 13.18 of [17] there exists a polynomial $p = p_j$ of total order r such that $\|f - p\| \le Dm^{-r} \sum_{\alpha:|\alpha|=r} \|D^\alpha f\|$ for some constant D that only depends on d, r and thus not on H. We have $Qp = p$ (see (12.30) in [17]) and hence $\|f - Qf\| \le \|f - p\| + \|Q(f - p)\| \le (C+1)\|f - p\|$. □

2.4. The size of a spline and its coefficients

In Section 3 we will use the fact that a smooth function can be approximated by a spline in $\mathcal{S}_m$ in the sense of Lemma 2.1. For our purposes, we do not need to know the approximating spline or its coefficients in full detail, but rather an expression that quantifies its size. We will use the following lemma, which states that the uniform norm of a spline is equivalent to the maximal norm of the vector of its B-spline coefficients.

Recall that the B-spline coefficients of a spline can be obtained from a dual basis of $\mathcal{S}_m$. We now assume that $\lambda_1, \ldots, \lambda_J$ is the dual basis given in Theorem 12.5 of [17]. Let $\|\lambda_j\|$ be the norm of the bounded linear functional λ_j. That is to say, $\|\lambda_j\|$ is the smallest constant K for which $|\lambda_j(s)| \leq K\|s\|_\infty$ holds for any $s \in \mathcal{S}_m$ Although $\max_{1\leq j\leq J} \|\lambda_j\|$ depends on m, it can actually be replaced by a constant that does not depend on m, cf. Theorem 12.5 in [17].

Lemma 2.2. *Let $s \in \mathcal{S}_m$ be given by (2.3). Then*

$$\|s\|_\infty \leq \max_{1\leq j\leq J} |a_j| \leq \big(\max_{1\leq j\leq J} \|\lambda_j\|\big)\|s\|_\infty \leq C\|s\|_\infty,$$

where $C > 0$ is a constant independent of m.

Proof. Because the B-splines are nonnegative, $|s(x)| \leq \sum_{j=1}^J |a_j|B_j(x)$. Take the maximum of the absolute values $|a_j|$ outside the sum. The first inequality now follows from the partition of unity property (2.2). For the second inequality, use that $|a_j| \leq \|\lambda_j\|\|s\|_\infty$, by definition. The third inequality follows from Theorem 12.5 in [17]. □

3. Gaussian random splines

In this section we introduce and study a class of Gaussian processes that we will use to construct prior distributions for various statistical settings. The corresponding posterior contraction rates will be determined in Section 4.1. We use the tensor-product splines from the preceding section to define the stochastic process via its sample paths.

We have seen that the space $\mathcal{S}_m$ of tensor-product splines depends on two parameters q and m. The parameter q is the order of the splines and m quantifies the partition size. We fix some natural number $q \geq 2$ and from now on it will be understood that all splines are of order q. The remaining parameter m will be referred to as the partition size parameter.

As before, let $B_1^m, \ldots, B_J^m$ be the tensor-product B-spline basis of $\mathcal{S}_m$, with $J = (m + q - 1)^d$. For any $m \in \mathbb{N}$ we now define the Gaussian random element W^m in $\mathcal{S}_m$ as follows. Let $Z_1, \ldots, Z_J$ be independent, standard Gaussian random variables, and let W^m be the random process on $[0, 1]^d$ defined by

$$W^m(x) = \sum_{j=1}^J Z_j B_j^m(x), \qquad x \in [0, 1]^d. \tag{3.1}$$

It follows from Theorem 4.2 in [25] that the reproducing kernel Hilbert space (RKHS) $\mathbb{H}^m$ associated with W^m consists of all splines of order q with respect to the given partition. Moreover, since they are linearly independent, the B-splines

B_j^m form an orthonormal basis of the RKHS. It follows immediately that the RKHS-norm of a spline in the RKHS is equal to the Euclidean norm of the vector of its B-spline coefficients. In other words, the RKHS of W^m is equal to the set $\mathcal{S}_m$ equipped with the norm $\|\cdot\|_{\mathbb{H}^m}$ given by

$$\Big\|\sum_{j=1}^{J} a_j B_j^m\Big\|_{\mathbb{H}^m}^2 = \sum_{j=1}^{J} a_j^2. \tag{3.2}$$

It is known (cf. [23]) that in general, the contraction rate of a posterior corresponding to a Gaussian process prior is determined by its concentration function, i.e. its non-centered small ball probabilities around the truth. The concentration function can be determined from the centered small ball probabilities of the process in addition to a term that quantifies the size of an approximation of the truth in the reproducing kernel Hilbert space of the process. We study these two quantities in the next two subsections.

3.1. Centered small ball probabilities

The following lemma is a straightforward consequence of the definition of the process W^m and the basic properties of the B-splines.

Lemma 3.1. *For all $q, m \in \mathbb{N}$ such that $m \geq q-1$ we have*

$$\mathbb{P}(\|W^m\|_\infty \leq \varepsilon) \geq (\varepsilon/2)^{2^d m^d}$$

for all $\varepsilon \in (0, 1/2)$.

Proof. By Lemma 2.2 and the fact that the random variables Z_j are independent and identically distributed we have

$$\mathbb{P}(\|W^m\|_\infty \leq \varepsilon) \geq \mathbb{P}(\max |Z_j| \leq \varepsilon) = (\mathbb{P}(|Z_1| \leq \varepsilon))^J.$$

The probability $\mathbb{P}(|Z_1| \leq \varepsilon)$ is bounded from below $2\varepsilon\varphi(\varepsilon_0)$ for for any $\varepsilon \in (0, \varepsilon_0)$, with φ the standard normal density. Since $J = (q+m-1)^d \leq (2m)^d$ for $m \geq q-1$, it follows that for $\varepsilon \in (0, \varepsilon_0)$ and any $q \geq 1$ and $m \geq q-1$,

$$(\mathbb{P}(|Z_1| \leq \varepsilon))^J \geq (2\varphi(\varepsilon_0)\varepsilon)^{2^d m^d}$$

This proves the assertion, since $2\varphi(1/2) \geq 1/2$. □

3.2. Non-centered small ball probabilities

Consider $w_0 \in C^r([0,1]^d)$. The non-centered ball probability $\mathbb{P}(\|W^m - w_0\|_\infty \leq 2\varepsilon)$ is the probability that a realization of W^m ends up in a uniform ball of radius 2ε around w_0. To obtain contraction rates for priors based on the process W^m we need a lower bound for this quantity.

Lemma 3.2. *Let $w_0 \in C^r([0,1]^d)$, for $r \leq q$. There exist constants $C, D > 0$, independent of m, such that for any $\varepsilon \in (0, 1/2)$ and any $m \geq q-1$ such that $Dm^{-r} \leq \varepsilon$, we have*

$$\mathbb{P}(\|W^m - w_0\|_\infty \leq 2\varepsilon) \geq \exp(-Cm^d \log 1/\varepsilon).$$

Let $\varphi^m_{w_0}$ be the concentration function of W^m around w_0 as defined in [23]:

$$\varphi^m_{w_0}(\varepsilon) = \inf_{h \in \mathbb{H}^m : \|h - w_0\|_\infty \leq \varepsilon} \|h\|^2_{\mathbb{H}^m} - \log \mathbb{P}(\|W^m\|_\infty \leq \varepsilon). \tag{3.3}$$

Then by Lemma 5.3 of [25],

$$\mathbb{P}(\|W^m - w_0\|_\infty \leq 2\varepsilon) \geq \exp(-\varphi^m_{w_0}(\varepsilon))$$

and a similar inequality holds for the upper bound. Now Lemma 3.2 is a consequence of the following result.

Lemma 3.3. *Let $w_0 \in C^r([0,1]^d)$, for $r \leq q$. There exist constant $C, D > 0$, independent of m, such that for any $\varepsilon \in (0, 1/2)$ and any $m \geq q-1$ such that $Dm^{-r} \leq \varepsilon$, we have*

$$\varphi^m_{w_0}(\varepsilon) \leq Cm^d \log 1/\varepsilon. \tag{3.4}$$

Proof. The second term in the concentration function (3.3) can be bounded from above using Lemma 3.1. For $\varepsilon \in (0, 1/2)$ we have

$$-\log \mathbb{P}(\|W^m\|_\infty \leq \varepsilon) \leq 2^d m^d \log\left(\frac{2}{\varepsilon}\right). \tag{3.5}$$

As for the infimum part in (3.3), Lemma 2.1 shows that for every $m \in \mathbb{N}$ there exists a spline $s \in \mathcal{S}_m = \mathbb{H}^m$ such that $\|s - w_0\|_\infty \leq Dm^{-r}$, for $D > 0$ a constant that only depends on d, q, r and w_0. Now fix $\varepsilon \in (0, 1/2)$ and $m \in \mathbb{N}$ such that $Dm^{-r} < \varepsilon$. Then with s the spline above,

$$\inf_{h \in \mathbb{H}^m : \|w_0 - h\|_\infty \leq \varepsilon} \|h\|^2_{\mathbb{H}^m} \leq \|s\|^2_{\mathbb{H}^m}.$$

Suppose that the spline $s \in \mathcal{S}_m$ is given by $s = \sum_{j=1}^J a_j B^m_j$. Then the squared RKHS-norm of s is given by (3.2) and satisfies

$$\|s\|^2_{\mathbb{H}^m} = \sum_{j=1}^J a_j^2 \leq J\bigl(\max_{1 \leq j \leq J} |a_j|\bigr)^2.$$

We have seen in Lemma 2.2 that the absolute maximum $\max_{1 \leq j \leq J} |a_j|$ of the coefficients can be bounded from above by $C'\|s\|_\infty$ for some $C' > 0$ that does not depend on m. Note that by the triangle inequality and the fact that $Dm^{-r} < \varepsilon$, we have that $\|s\|_\infty \leq \|w_0\|_\infty + \varepsilon$. Since $J \leq (2m)^d$, we obtain upper bound for $\|s\|^2_{\mathbb{H}^m}$ that can be written as a multiple of m^d. This concludes the proof. □

4. Posterior contraction results

4.1. Gaussian spline priors

The Gaussian spline processes W^m can be used to construct priors in various nonparametric statistical settings. In order for the priors to have large enough support to ensure for instance consistency, one has to either let the partition size parameter m tend to infinity with sample size, or view it as a hyper parameter that itself is estimated from the data. In this section we consider the former construction, leading to sequences of Gaussian process priors. We give bounds on the contraction rates of the corresponding posteriors. In the next section we investigate the possibility of endowing m with a prior distribution.

Let $m_n \to \infty$ be a sequence of natural numbers, fix an order $q \geq 2$ for the splines and consider the corresponding sequence W^{m_n} of Gaussian spline processes on $[0,1]^d$ defined by (3.1). For a natural number $r \leq q$ and $w_0 \in C^r([0,1]^d)$, let $\varphi_{w_0}^{m_n}$ be the sequence of concentration functions defined by (3.3), with $\mathbb{H}^m$ the RKHS of the process W^m. The general theory of Gaussian process priors says that posterior contraction rates are obtained by solving the inequality

$$\varphi_{w_0}^{m_n}(\varepsilon_n) \leq n\varepsilon_n^2, \tag{4.1}$$

see Van der Vaart and Van Zanten [23]. By Lemma 3.3 this inequality holds if

$$\begin{aligned} Cm_n^d \log m_n &\leq n\varepsilon_n^2, \\ Dm_n^{-r} &\leq \varepsilon_n, \end{aligned}$$

with $C, D > 0$ the constants from the statement of the lemma. The optimal solution of these inequalities is easily found and given in the following theorem.

Theorem 4.1. *In the setting described above, let $m_n \sim (n/\log n)^{1/(d+2r)}$. Then inequality (4.1) holds with $\varepsilon_n \sim (n/\log n)^{-r/(d+2r)}$.*

In combination with the results given in [23] this theorem immediately yields rate of contraction results for a number of important nonparametric statistical problems. In a nonparametric fixed design regression problem for instance, where we have observations $Y_1, \dots, Y_n$ satisfying

$$Y_i = w_0(x_i) + \xi_i$$

for known $x_i \in [0,1]^d$ and independent $N(0,\sigma^2)$-distributed ξ_i, the law Π_n of W^{m_n} can be used directly as a prior on the regression function w_0. Combining Theorem 4.1 and Theorem 3.3 of [23] then shows that if $w_0 \in C^r([0,1]^d)$, the posterior distribution $\Pi_n(\cdot \,|\, Y_1, \dots, Y_n)$ of the regression function concentrates around the truth at the rate $\varepsilon_n = (n/\log n)^{-r/(d+2r)}$ in the sense that we have, for all $L > 0$ large enough,

$$\mathbb{E}_0 \Pi_n\Big(w : \frac{1}{n}\sum_{i \leq n}(w(x_i) - w_0(x_i))^2 > L^2\varepsilon_n^2 \,|\, Y_1, \dots, Y_n\Big) \to 0$$

as $n \to \infty$. Here $\mathbb{E}_0$ is the expectation corresponding to the true regression function w_0.

After exponentiation and renormalization a Gaussian process can be used as a prior model for probability densities as well. Theorem 4.1 and Theorem 3.1 of [23] imply that if the true log-density is in $C^r([0,1]^d)$, a contraction rate $(n/\log n)^{-r/(d+2r)}$ relative to the Hellinger distance is attained. Similar results can be obtained for instance for classification settings (by combining Theorem 4.1 and Theorem 3.2 of [23]) and nonparametric drift estimation for diffusions (using results of [22] or [12]).

Generally speaking, the results show that if the law of the Gaussian spline process W^{m_n} is used as a prior on an r-regular function of d variables (possibly after a suitable transformation), then with the choice $m_n \sim (n/\log n)^{1/(d+2r)}$ this leads to a posterior contraction rate of the order $n^{-r/(d+2r)}$, up to a logarithmic factor. This is typically the optimal rate for estimating an r-regular function of d variables (up to a logarithmic factor), for instance in a minimax sense. Note however that through the partition size parameter m_n, the prior depends on the unknown smoothness level of the function of interest. Hence, the procedure is not rate-adaptive. In Section 4.2 we construct a hierarchical, conditionally Gaussian prior that does lead to adaption.

4.2. Adaptation using conditionally Gaussian priors

In the previous section we saw, for several statistical settings, that under a certain smoothness condition on the truth w_0, posterior contraction can be achieved at an optimal rate for an appropriate sequence of our Gaussian spline priors. We assumed that w_0 is contained in $C^r([0,1]^d)$ for a given $r \leq q$ and used the knowledge of the degree of regularity r to define a sequence of Gaussian priors via the partition size parameter m_n.

In practice however, the exact degree of smoothness is typically not known a-priori. Therefore, we will in this section only assume that for $q \geq 2$ fixed in advance, w_0 is contained in $C^r([0,1]^d)$ for r some unknown smoothness level such that $r \leq q$. In other words, we only assume a known upper bound on the smoothness. The aim now is to construct a prior independent of r such that the posterior achieves the same optimal rate as in the preceding section (perhaps up to a logarithmic factor) for every possible value of $r \leq q$. Such a procedure is said to adapt to the regularity of the truth up to the level q.

As before we take the Gaussian spline process W^m as the starting point for the definition of our priors. However, we now take a different approach to choosing m. In the Bayesian paradigm it is quite common to view unknown tuning parameters of this type as so-called hyper parameters and to endow them with a separate prior, leading to hierarchical priors. We adopt this approach and show that if the prior on m is chosen carefully, we can achieve our goal of constructing a rate-adaptive procedure in this way.

Concretely, we define a new, conditionally Gaussian spline process W by setting $W = W^M$, for W^m the Gaussian process defined in (3.1) and M an independent $\mathbb{N}$-valued random variable. This construction is hierarchical in the sense that a sample path of W is generated in two steps: first draw a realization

m of the random variable M, then given m, draw a sample path of the Gaussian process W^m.

The hierarchical spline process can be used to construct priors for various statistical settings again. The following general theorem about the process W will lead to the desired adaptive rate of contraction results.

Theorem 4.2. *Suppose that for every $m \geq 1$,*

$$C_1 \exp(-D_1 m^d \log^t m) \leq \mathbb{P}(M = m) \leq C_2 \exp(-D_2 m^d \log^t m) \tag{4.2}$$

for some constants $C_1, C_2, D_1, D_2, t \geq 0$. If $w_0 \in C^r([0,1]^d)$ for some integer $r \leq q$, then there exists for every constant $C > 0$, a constant $D > 0$ and measurable subsets U_n of $C([0,1]^d)$ such that

$$\mathbb{P}(\|W - w_0\|_\infty \leq 2\varepsilon_n) \geq \exp(-n\varepsilon_n^2), \tag{4.3}$$

$$\mathbb{P}(W \notin U_n) \leq \exp(-Cn\varepsilon_n^2), \tag{4.4}$$

$$\log N(2\bar{\varepsilon}_n, U_n, \|\cdot\|_\infty) \leq Dn\bar{\varepsilon}_n^2, \tag{4.5}$$

are satisfied for sufficiently large n, and for ε_n and $\bar{\varepsilon}_n$ given by

$$\varepsilon_n = c(n/\log^{1\vee t} n)^{-\frac{r}{d+2r}} \qquad \bar{\varepsilon}_n = n^{-\frac{r}{d+2r}} (\log n)^{\frac{(1\vee t)r}{d+2r} + (\frac{1-t}{2})_+}, \tag{4.6}$$

for $c > 0$ a large enough constant.

Combined with existing results from [6, 9] and [23], which give posterior contraction rates for various statistical settings under conditions of the form (4.3)–(4.5), this general theorem will lead to results that state that in the various settings, using the law of W^M as a prior will lead to posterior contraction at the rate $\varepsilon_n \vee \bar{\varepsilon}_n$, provided that the true function has smoothness degree $r \leq q$. Hence, up to a logarithmic factor, the posteriors attain optimal convergence rates in this case. Moreover, since the prior does not depend on the unknown smoothness level r, we indeed obtain rate-adaptive procedures.

Note that condition (4.2) holds in particular, for $t = 0$, if M^d has a geometric distribution. The best rate $\varepsilon_n \vee \bar{\varepsilon}_n$ is obtained when t is chosen equal to 1. The resulting rate is $(n/\log n)^{-r/(d+2r)}$ in that case, which coincides with the rate obtained in Theorem 4.1 for the non-adaptive sequence of spline priors.

In our approach the order q of the splines remains fixed, contrary to for instance in [10] or [7]. This keeps the priors simple and easy to deal with, but of course in practice q has to be chosen. From the theoretical perspective q can be chosen as large as one would like, although it might be chosen not too large for computational reasons. In practice, cubic splines ($q = 4$ in our notation) are a popular choice.

4.3. Proof of the general Theorem 4.2

4.3.1. Prior mass condition (4.3)

Let $\varepsilon_n \to 0$ be given. Note that the inequality

$$\mathbb{P}(\|W - w_0\|_\infty \leq 2\varepsilon_n) \geq \mathbb{P}(M = m)\mathbb{P}(\|W^m - w_0\| \leq 2\varepsilon_n)$$

holds for any $m \geq 1$ by construction of W. According to Lemma 3.2 the second factor on the right is bound from below by $\exp(-Cm_n^d \log m_n)$ for sufficiently large n and m_n such that $\varepsilon_n \geq Dm_n^{-r}$. The probability $\mathbb{P}(M = m_n)$ is bounded from below by $C_1 \exp(-D_1 m_n^d \log^t m_n)$ by assumption (4.2). We conclude that

$$\mathbb{P}(\|W - w_0\|_\infty \leq 2\varepsilon_n) \geq C_1 \exp(-C_2 m_n^d \log^{1\vee t} m_n)$$

for some constants $C_1, C_2 > 0$. The inequalities

$$\begin{aligned} m_n^d \log^{1\vee t} m_n &\lesssim n\varepsilon_n^2, \\ m_n^{-r} &\lesssim \varepsilon_n, \end{aligned}$$

are solved by $m_n \sim (n/\log^{1\vee t} n)^{1/(d+2r)}$ and ε_n as in (4.6). Condition (4.3) thus holds if the constant c in (4.6) is sufficiently large.

4.3.2. Construction of sieves U_n

Recall that $\mathbb{H}_1^m$ is the unit ball of the RKHS $\mathbb{H}^m$ of the Gaussian spline process W^m and $\mathbb{B}_1$ is the unit ball in the Banach space $C([0,1]^d)$. For $m \in \mathbb{N}$, let $U_n^m = L_n \mathbb{H}_1^m + \varepsilon_n \mathbb{B}_1$ for some k_n and L_n specified below, and $U_n = \bigcup_{m=1}^{k_n} U_n^m$.

In the next two subsections we show that conditions (4.4) and (4.5) are fulfilled if L_n and k_n satisfy certain inequalities. In Subsection 4.3.5 we show that these inequalities can be solved.

4.3.3. Remaining mass condition (4.4)

First note that the inequality

$$\mathbb{P}(W \notin U_n) \leq \sum_{m=1}^{k} \mathbb{P}(M = m)\mathbb{P}(W^m \notin U_n) + \mathbb{P}(M \geq k + 1). \qquad (4.7)$$

holds for any k by construction of W. Now take k equal to k_n as defined in the preceding subsection. By assumption (4.2) the tail probability $\mathbb{P}(M \geq k_n + 1)$ is bounded from above by a constant times the geometric series

$$\sum_{m \geq k_n+1} (\exp(-k_n^{d-1} \log^t k_n))^m \leq \exp(-k_n^d \log^t k_n).$$

So the tail probability is bounded by $\exp(-Cn\varepsilon_n^2)/2$ for large n if k_n is chosen such that $k_n^d \log^t k_n > Cn\varepsilon_n^2$, for C as in the assertion of the theorem.

We now show that

$$\mathbb{P}(W^m \notin U_n) \leq \exp(-Cn\varepsilon_n^2)/2$$

for any $m \leq k_n$, so that the first term on the right of (4.7) is also bounded by $\exp(-Cn\varepsilon_n^2)/2$. It follows from the construction of the sieve B_n that

$$\mathbb{P}(W^m \notin U_n) \leq \mathbb{P}(W^m \notin U_n^m)$$

for any $m \leq k_n$. By Borell's inequality (see [25], Theorem 5.1)

$$\mathbb{P}(W^m \notin U_n^m) \leq 1 - \Phi(\Phi^{-1}(\mathbb{P}(\|W^m\|_\infty \leq \varepsilon_n)) + L_n).$$

A lower bound for the centered small ball probability $\mathbb{P}(\|W^m\|_\infty \leq \varepsilon_n)$ was given in Lemma 3.1. The lower bound provided by this lemma is a decreasing function of m. For every $m \leq k_n$ we thus have

$$\mathbb{P}(\|W^m\|_\infty \leq \varepsilon_n) \geq (\varepsilon/2)^{2^d k_n^d}.$$

For $y \in (0, 1/2)$ one has $\Phi^{-1}(y) \geq -\sqrt{(5/2)\log(1/y)}$. Apply this inequality with y equal to $(\varepsilon/2)^{2^d k_n^d}$ to find that

$$\mathbb{P}(W^m \notin U_n^m) \leq \Phi\Big(\sqrt{(5/2)2^d k_n^d \log(2/\varepsilon_n)} - L_n\Big)$$

for every $m \leq k_n$. Using the bound $\Phi(y) \leq \exp(-y^2/2)$ we obtain

$$\mathbb{P}(W^m \notin U_n) \leq e^{-\frac{1}{2}\left(L_n - \sqrt{(5/2)2^d k_n^d \log(2/\varepsilon_n)}\right)^2} \tag{4.8}$$

for every $m \leq k_n$. Hence if L_n and k_n are chosen such that

$$\frac{1}{2}\left(L_n - \sqrt{(5/2)2^d k_n^d \log(2/\varepsilon_n)}\right)^2 > Cn\varepsilon_n^2,$$

then the first term on the right of (4.7) is bounded by $\exp(-Cn\varepsilon_n^2)/2$ as well.

4.3.4. *Proof of entropy condition (4.5)*

Let $\bar{\varepsilon}_n$ be given by (4.6). Because U_n is a union of the sets U_n^m for $m = 1, \ldots, k_n$, its $2\bar{\varepsilon}_n$-covering number satisfies

$$N(2\bar{\varepsilon}_n, U_n, \|\cdot\|_\infty) \leq \sum_{m=1}^{k_n} N(2\bar{\varepsilon}_n, U_n^m, \|\cdot\|_\infty).$$

If $A_1, \ldots, A_N$ is a minimal covering of $\mathbb{H}_1^m$ using balls of radius $\bar{\varepsilon}_n/L_n$, then the sets $L_n A_i + \varepsilon_n \mathbb{B}_1$ are balls of radius $\bar{\varepsilon}_n + \varepsilon_n \leq 2\bar{\varepsilon}_n$ which cover U_n^m. This shows that

$$N(2\bar{\varepsilon}_n, U_n^m, \|\cdot\|_\infty) \leq N(\bar{\varepsilon}_n/L_n, \mathbb{H}_1^m, \|\cdot\|_\infty). \tag{4.9}$$

We now identify splines in $\mathbb{H}^m$ with points in $\mathbb{R}^J$ via the B-spline coefficients. Then $\mathbb{H}_1^m$ corresponds to the unit ball in $\mathbb{R}^J$ (see (3.2)). Moreover, for a spline $s = \sum a_j B_j^m$ in $\mathbb{H}^m$ we have that the uniform norm $\|s\|_\infty$ is bounded by the Euclidean norm $\|a\|$ of the vector of B-spline coefficients, by Cauchy-Schwarz and the basic properties of the B-splines. It follows that the covering number on the right of (4.9) is bounded by the $\bar{\varepsilon}_n/L_n$-covering number of the unit ball in $\mathbb{R}^J$ relative to the Euclidean distance. The latter is bounded from above by $(6L_n/\bar{\varepsilon}_n)^J$ according to e.g. Lemma 4.1 of Pollard [14].

We thus find

$$N(2\bar{\varepsilon}_n, U_n, \|\cdot\|_\infty) \leq k_n (6L_n/\bar{\varepsilon}_n)^{2^d k_n^d}$$

and consequently, if $L_n = O(n^p)$ for some $p > 0$, we have

$$\log N(2\bar{\varepsilon}_n, U_n, \|\cdot\|_\infty) \leq D k_n^d \log n$$

for some positive constant D. So if k_n is taken such that $k_n^d \log n$ is bounded by a multiple of $n\bar{\varepsilon}_n^2$, then condition (4.5) holds.

4.3.5. End of the proof of Theorem 4.2

The preceding subsections show that the proof of Theorem 4.2 is complete once we show that there exist sequences L_n and k_n such that

$$k_n^d \log^t k_n > C n \varepsilon_n^2 \tag{4.10}$$

$$\bar{\varepsilon}_n \geq \varepsilon_n \tag{4.11}$$

$$k_n^d \log n \leq C' n \bar{\varepsilon}_n^2 \tag{4.12}$$

$$L_n - \sqrt{(5/2) 2^d k_n^d \log(2/\varepsilon_n)} > \sqrt{2Cn\varepsilon_n^2} \tag{4.13}$$

$$L_n = O(n^p), \tag{4.14}$$

where C is a given positive constant and p and C' may be chosen arbitrarily.

We have $n\varepsilon_n^2 = c^2 n^{d/(d+2r)} (\log n)^{(2r(1\vee t))/(d+2r)}$, hence (4.10) is fulfilled if

$$k_n^d = A n^{\frac{d}{d+2r}} (\log n)^{\frac{2r(1\vee t)}{d+2r} - t},$$

with A a large enough positive constant. Conditions (4.11) and (4.12) are then fulfilled as well if C' is chosen large enough, by definition of the sequence $\bar{\varepsilon}_n$. Finally, conditions (4.13) and (4.14) are then easily taken care of by taking L_n to be a large enough power of n.

References

[1] Carl de Boor. *A practical guide to splines*, volume 27 of *Applied Mathematical Sciences*. Springer-Verlag, New York, revised edition, 2001. MR1900298

[2] R. de Jonge and J.H. van Zanten. Adaptive nonparametric Bayesian inference using location-scale mixture priors. *Ann. Statist.*, 38(6):3300–3320, 2010. MR2766853

[3] D.G.T. Denison, B.K. Mallick, and A.F.M. Smith. Automatic Bayesian curve fitting. *J. R. Statist. Soc. B*, 60(60):333–350, 1998. MR1616029

[4] I. DiMatteo, C.R. Genovese, and R.E. Kass. Bayesian curve-fitting with free-knot splines. *Biometrika*, 88(4):1055–1071, 2001. MR1872219

[5] Subhashis Ghosal. Convergence rates for density estimation with Bernstein polynomials. *Ann. Statist.*, 29(5):1264–1280, 2001. MR1873330

[6] Subhashis Ghosal, Jayanta K. Ghosh, and Aad W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 2000. MR1790007

[7] Subhashis Ghosal, Jüri Lember, and Aad Van Der Vaart. Nonparametric Bayesian model selection and averaging. *Electronic Journal of Statistics*, 2:63–89, 2008. MR2386086

[8] Subhashis Ghosal and Aad W. van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.*, 29(5):1233–1263, 2001. MR1873329

[9] Subhashis Ghosal and Aad W. van der Vaart. Convergence rates for posterior distributions for noniid observations. *Ann. Statist.*, 35:697–723, 2007. MR2336864

[10] Tzee-Ming Huang. Convergence rates for posterior distributions and adaptive estimation. *Ann. Statist.*, 32(4):1556–1593, 2004. MR2089134

[11] Willem Kruijer, Judith Rousseau, and Aad van der Vaart. Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.*, 4:1225–1257, 2010. MR2735885

[12] L. Panzar and J.H. van Zanten. Nonparametric Bayesian inference for ergodic diffusions. *J. Statist. Plann. Inference*, 139:4204–4210, 2009. MR2558361

[13] Sonia Petrone and Larry Wasserman. Consistency of Bernstein polynomial posteriors. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(1):79–100, 2002. MR1881846

[14] David Pollard. *Empirical processes: theory and applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, 2. Institute of Mathematical Statistics, Hayward, CA, 1990. MR1089429

[15] V. Rivoirard and J. Rousseau. Posterior concentration rates for infinite dimensional exponential families. *Bayesian Anal.*, to appear, 2012.

[16] Judith Rousseau. Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *Ann. Statist.*, 38(1):146–180, 2010. MR2589319

[17] Larry L. Schumaker. *Spline functions: Basic theory*. John Wiley & Sons Inc., New York, 1981. Pure and Applied Mathematics, A Wiley-Interscience Publication. MR0606200

[18] W. Shen and S. Ghosal. MCMC-free adaptive Bayesian procedures using random series prior. Preprint, 2012.

[19] Xiaotong Shen and Larry Wasserman. Rates of convergence of posterior distributions. *Ann. Statist.*, 29(3):687–714, 2001. MR1865337

[20] Charles J. Stone. Large-sample inference for log-spline models. *Ann. Statist.*, 18(2):717–741, 1990. MR1056333

[21] Charles J. Stone. The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.*, 22(1):118–184, 1994. With discussion by Andreas Buja and Trevor Hastie and a rejoinder by the author. MR1272079

[22] F.H. van der Meulen, A.W. van der Vaart, and J.H. van Zanten. Convergence rates of posterior distributions for Brownian semimartingale models. *Bernoulli*, 12(5):863–888, 2006. MR2265666

[23] A.W. van der Vaart and J.H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.*, 36(3):1435–1463, 2008. MR2418663

[24] A.W. van der Vaart and J.H. van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.*, 37(5B):2655–2675, 2009. MR2541442

[25] A.W. Van der Vaart and J.H. Van Zanten. *Reproducing Kernel Hilbert Spaces of Gaussian priors*, volume 3 of *IMS Collections*, pages 200–222. Institute of Mathematical Statistics, 2008. MR2459226