

ADAPTIVE NONPARAMETRIC BAYESIAN INFERENCE USING LOCATION-SCALE MIXTURE PRIORS

BY R. DE JONGE AND J.H. VAN ZANTEN

Eindhoven University of Technology

We study location-scale mixture priors for nonparametric statistical problems, including multivariate regression, density estimation and classification. We show that a rate-adaptive procedure can be obtained if the prior is properly constructed. In particular, we show that adaptation is achieved if a kernel mixture prior on a regression function is constructed using a Gaussian kernel, an inverse gamma bandwidth, and Gaussian mixing weights.

Revised version, December 22, 2009

1. Introduction. In Bayesian nonparametrics, the use of location-scale mixtures of kernels for the construction of priors on probability densities is well established. The methodology is used in a variety of practical settings, and in recent years there has been substantial progress on the mathematical, asymptotic theory for kernel mixture priors as well, cf. [3], [23], [6], [5], [29], [15]. At the present time we have a well-developed understanding of important aspects including consistency, convergence rates, rate-optimality, and adaptation properties. A similar, parallel development has taken place in the area of beta mixture priors, cf. [20], [4], [14], [21].

A discrete location-scale mixture of a fixed probability density p on \mathbb{R}^d can be expressed as

$$(1.1) \quad x \mapsto \sum_{j=1}^m w_j \frac{1}{\sigma^d} p\left(\frac{x - x_j}{\sigma}\right),$$

where $m \in \mathbb{N}$, $x_1, \dots, x_m \in \mathbb{R}^d$, $w_1, \dots, w_m \geq 0$ and $\sum w_j = 1$, and $\sigma > 0$. A prior on densities is obtained by putting prior distributions on m , the locations x_j , the scale σ and the weights w_j . When p satisfies some regularity conditions, a wide class of probability densities can be well approximated by mixtures of the form (1.1). This indicates that if the priors on the coefficients

*Research supported by the Netherlands Organization for Scientific Research NWO

AMS 2000 subject classifications: Primary 62G08, 62C10; secondary 62G20

Keywords and phrases: Rate of convergence, posterior distribution, adaptation, Bayesian inference, nonparametric regression, kernel mixture priors

are suitably chosen, the resulting prior and posterior on probability densities can be expected to have good asymptotic properties. The cited papers give precise conditions under which this is indeed the case.

Obviously, a much wider class of functions is well approximated by mixtures of the form (1.1) if we lift the restriction that the weights w_j should be nonnegative and sum up to 1. This suggests that location-scale mixtures might be attractive priors not just in the setting of density estimation, but for instance also in nonparametric regression. Although this idea has been proposed in the applied literature, cf. e.g. [22], [11], it does not seem to have attracted a great deal of attention. The few examples do show however that the approach can yield quite satisfactory results.

In the paper [22], location-scale mixture priors are used in an astrophysical setting for the analysis of data from galactic radio sources. The statistical problem essentially boils down to a bivariate, nonparametric, fixed design regression problem. The use of a mixture prior is natural in that particular application because it reflects the idea that the function of interest, which describes the strength of the magnetic field caused by our planet and its “neighborhood” in space, is in fact an aggregate of contributions from a large number of locations, with different weights, which can be positive or negative.

Another reason for using a location-scale mixture prior in multivariate regression, instead of for instance the popular Gaussian squared exponential or Matérn priors, are computational advantages. Conditional on the gridsize m the prior only involves finitely many terms, so no artificial truncation or approximation is necessary for computation. As argued also in [22], the mixture prior allows to avoid the inversion or decomposition of non-trivial and often ill-behaved $n \times n$ matrices (with n the sample size), which can become cumbersome already for moderate sample sizes (cf. also the discussion in [1]). In the astrophysical application of [22], the sample size is of the order 1500 and it is shown that samples of this order can be dealt with effectively using kernel mixture priors.

On the theoretical side, little or nothing seems to be known for kernel mixture priors in a regression setting. In the present paper we therefore take up the study of asymptotic properties, in order to assess the fundamental potential of the methodology and to provide a theoretical underpinning of its use in practice. We will show that if the kernel and the priors on locations and scales are appropriately chosen, kernel mixture priors yield posteriors with very good asymptotic properties. It is well known that for the estimation of an α -regular function of d variables, the best possible rate of convergence is of the order $n^{-\alpha/(d+2\alpha)}$, where n is the number of observations available. We

will prove that up to a logarithmic factor, this optimal rate can be attained with location-scale mixture priors. More importantly, the near optimal rate can be achieved by a prior that does *not* depend on the unknown smoothness level α of the regression function. In other words, we can obtain a fully adaptive procedure.

The bounds for the convergence rates that we will obtain depend crucially on the smoothness of the kernel p that is used. For kernels with only a finite degree of regularity, we get sub-optimal rates. We only obtain the optimal minimax rate (up to a logarithmic factor) for kernels that are infinitely smooth, in the sense that they admit an analytic extension to a strip in complex space. The standard normal kernel is an example of an optimal choice in this respect. We also have to put (mild) conditions on the priors on the grid size m and the scale σ . In particular, the popular inverse gamma choice for the scale is included in our setup.

Perhaps surprising is the fact that although we use a probability density p to construct the mixtures, we can still achieve adaptation to all smoothness levels. Intuition from kernel estimation might suggest that when p is a centered probability density, we have good approximation behaviour for regression functions with regularity at most 2, and that for more regular functions we should use higher order kernels. This turns out not to be the case however. To prove this fact we adapt an observation of J. Rousseau, who uses a similar idea to prove that for densities on the unit interval, using appropriate mixtures of beta densities yields adaptation to all smoothness levels, see [21]. The recent preprint [15], which was written at the same time and independently of the present work, employs the same idea to prove adaptation for kernel mixture priors for density estimation. In the present paper we extend the technique to a multivariate setting (see Lemma 3.4 ahead).

The literature on Bayesian adaptation is still relatively young. Earlier papers include [2], [12], [10], [9], [17], [21], and [26]. Priors that yield adaptation across a continuum of regularities in nonparametric regression have been exhibited in [12], where priors based on spline expansions are considered, and [26], who use randomly rescaled Gaussian processes as priors.

The location-scale priors we consider in this paper are conditionally Gaussian, since we will put Gaussian priors on the mixing weights. This allows us to use the machinery for Gaussian process priors developed in [27] and [28] in our proofs. Other technical ingredients include metric entropy results for spaces of analytic functions, as can be found for instance in [13], and the connection between metric entropy and small deviations results for Gaussian process (cf. [16], [18]). We will obtain a general result for a conditionally

Gaussian kernel mixture process, which can in fact be used in a variety of statistical settings. To illustrate this we present rate of contraction results not just for nonparametric regression, which is our main motivation, but also for density estimation and classification settings.

In the next section we present the main results of the paper. In Section 2.1 we state a general result for a conditionally Gaussian location-scale mixture process whose law will be used to define the kernel mixture prior in the various statistical settings. Rate of contraction results for nonparametric regression, density estimation and classification are given in Section 2.2. The proof of the general theorem can be found in Section 3.

1.1. Notation.

- $\Im z, \Re z$: imaginary and real part of a complex number z .
- $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$.
- For $k \in \mathbb{N}_0^d$: $k. = k_1 + \dots + k_d$, $k! = k_1! \dots k_d!$.
- $f * g$: convolution of f and g .
- $a \vee b = \max\{a, b\}$, $a \wedge b = \min\{a, b\}$, $a_+ = a \vee 0$.
- $C(X)$: continuous functions on X .
- $C^\alpha(X)$ for $\alpha > 0$ and $X \subseteq \mathbb{R}^d$: functions on X with bounded partial derivatives up to the order β , which is the largest integer strictly smaller than α , and such that the partial derivatives of order β are Hölder continuous of order $\alpha - \beta$. For $f \in C^\alpha(X)$ we denote by $\|f\|_\alpha$ the associated Hölder norm of f , cf. [25], Section 2.7.1. The Hölder ball of radius $R > 0$ is defined as $C_R^\alpha(X) = \{f \in C^\alpha(X) : \|f\|_\alpha \leq R\}$.

2. Main results.

2.1. *General result for Gaussian location-scale mixtures.* On a common probability space, let M be an \mathbb{N} -valued random variable, Σ a $(0, \infty)$ -valued random variable and $(Z_k : k \in \mathbb{N}^d)$ standard Gaussian random variables, all independent. The stochastic process W indexed by $[0, 1]^d$ is defined by

$$(2.1) \quad W(x) = \sum_{k \in \{1, \dots, M\}^d} Z_k \frac{1}{M^{d/2}} \frac{1}{\Sigma^d} p\left(\frac{x - k/M}{\Sigma}\right)$$

for $x \in [0, 1]^d$, where $p : \mathbb{R}^d \rightarrow \mathbb{R}$ is a function that belongs to the class \mathcal{P}_γ of γ -regular kernels defined as follows:

DEFINITION 2.1. For $\gamma \in (d/2, \infty]$, an integrable function p on \mathbb{R}^d belongs to \mathcal{P}_γ if $\int_{\mathbb{R}^d} p(x) dx = 1$, it is uniformly Lipschitz on \mathbb{R}^d , it has finite moments of every order, and it satisfies one of the following conditions, depending on whether $\gamma < \infty$ or $\gamma = \infty$:

- For $\gamma < \infty$: p belongs to $C^\gamma(\mathbb{R}^d)$.
- For $\gamma = \infty$: p is the restriction to \mathbb{R}^d of a function that is defined on the set $S = \{(z_1, \dots, z_d) \in \mathbb{C}^d : |\Im z_j| \leq 1 \text{ for } j = 1, \dots, d\}$, and that is bounded and analytic on S .

Examples of kernels belonging to \mathcal{P}_γ for $\gamma < \infty$ are abundant. Using Fourier inversion it is not difficult to see that an integrable function p belongs to \mathcal{P}_∞ if it has a characteristic function

$$\psi(\lambda) = \int_{\mathbb{R}^d} e^{i(\lambda, x)} p(x) dx$$

which is infinitely often differentiable at 0, which satisfies $\psi(0) = 1$, and which satisfies the exponential moment condition

$$\int_{\mathbb{R}^d} e^{\|\lambda\|} |\psi(\lambda)| d\lambda < \infty.$$

The prime example is the standard normal density on \mathbb{R}^d , which is easily seen to belong to \mathcal{P}_∞ . Note that we do not require that $p \geq 0$ in Definition 2.1. So in fact, higher order kernels are allowed as well.

The index γ of the class of kernels quantifies the regularity of the kernel that is employed. We will see that this regularity influences the rate of convergence that we can obtain for the corresponding location-scale mixture prior. The restriction $\gamma > d/2$ is connected to the fact that in order to obtain bounds for the process W independent of M , we want the process in (2.1) to be well defined if the sum is taken over all k in \mathbb{N}^d .

For $\varepsilon > 0$, the metric entropy of a set B in a metric space with metric d is defined as $\log N(\varepsilon, B, d)$, where $N(\varepsilon, B, d)$ is the minimum number of balls of radius ε needed to cover B . Fix $0 < a < b < 1$ and define $\mathcal{X} = [a, b]^d$. Let $d_\gamma = 2d(d + \gamma)/(2\gamma - d)$ and $\delta_\gamma = d/(2\gamma - d)$.

THEOREM 2.2. *Suppose that $p \in \mathcal{P}_\gamma$ for $\gamma \in (d/2, \infty]$, that $\mathbb{P}(M = m) \geq Cm^{-s}$ for some $C > 0$, $s > 1$, and that Σ has a Lebesgue density g that, for some $D_1, D_2, D_3, D_4 > 0$ and $q, r \geq 0$, satisfies*

$$(2.2) \quad D_1 \sigma^{-q} e^{-D_2 (\frac{1}{\sigma})^{d_\gamma} (\log \frac{1}{\sigma})^r} \leq g(\sigma) \leq D_3 \sigma^{-q} e^{-D_4 (\frac{1}{\sigma})^{d_\gamma} (\log \frac{1}{\sigma})^r}$$

for all σ in a neighborhood of 0.

Then if $w_0 \in C^\alpha(\mathcal{X})$ for $\alpha > 0$, there exist for every constant $C > 1$ measurable subsets B_n of $C([0, 1]^d)$ and a constant $D > 0$ such that, for n

large enough,

$$(2.3) \quad \log N(\bar{\varepsilon}_n, B_n, \|\cdot\|_\infty) \leq Dn\bar{\varepsilon}_n^2,$$

$$(2.4) \quad \mathbb{P}(W \notin B_n) \leq e^{-Cn\varepsilon_n^2},$$

$$(2.5) \quad \mathbb{P}\left(\sup_{x \in \mathcal{X}} |W(x) - w_0(x)| \leq \varepsilon_n\right) \geq e^{-n\varepsilon_n^2}.$$

Here if $\gamma < \infty$,

$$\varepsilon_n = n^{-\frac{\alpha}{d_\gamma + 2\alpha(1+\delta_\gamma)}}, \quad \bar{\varepsilon}_n = n^{-\frac{\alpha(1-(d\delta_\gamma)/(2\gamma))}{(d_\gamma + 2\alpha(1+\delta_\gamma))(1+d/(2\gamma))}},$$

and if $\gamma = \infty$,

$$\varepsilon_n = n^{-\frac{\alpha}{d+2\alpha}} \log^{\frac{r\sqrt{(1+d)}}{2+d/\alpha}} n, \quad \bar{\varepsilon}_n = n^{-\frac{\alpha}{d+2\alpha}} \log^{\frac{r\sqrt{(1+d)}}{2+d/\alpha} + (\frac{1+d-r}{2})_+} n.$$

A few remarks about the result are in order. First of all, the process W is indexed by the unit cube, but the supremum in (2.5) is over the strictly smaller set \mathcal{X} . This is due to the fact that to obtain good enough approximations of the given function w_0 defined on \mathcal{X} by location-scale mixtures of the kernel p , we also need kernels centered at points just outside the set \mathcal{X} . A result like (2.5) with the supremum over the entire unit cube is only possible under additional assumptions on the boundary behaviour of the function w_0 .

Theorem 2.2 connects to existing results for nonparametric Bayes procedures, which give sufficient conditions of the form (2.3)–(2.5) for having a certain rate of posterior contraction, cf. e.g. [8], [7], [24]. In the next subsection we will single out the most important particular cases. In all cases the statistical results will state that the posterior will asymptotically concentrate on balls of radius of the order $\bar{\varepsilon}_n$ around the true parameter (relative to a natural statistical metric depending on the specific setting). Note that in the case $\gamma < \infty$, this means we only obtain a rate if $(d\delta_\gamma)/(2\gamma) < 1$, which is true if and only if $\gamma > (1/4)(1 + \sqrt{5})d \approx (0.81)d$. In particular, the choice $\gamma \geq d$ suffices to have consistency. As the smoothness γ of the kernel p that is employed is increased, the rate of contraction improves. Since $d_\gamma \rightarrow d$ and $\delta_\gamma \rightarrow 0$ as $\gamma \rightarrow \infty$, the power of n^{-1} in the expression for the rate $\bar{\varepsilon}_n$ tends to $\alpha/(d + 2\alpha)$ as $\gamma \rightarrow \infty$, which corresponds to the optimal minimax rate of convergence for estimating an α -regular function of d variables. If an analytic kernel $p \in \mathcal{P}_\infty$ is used the minimax rate $n^{-\alpha/(d+2\alpha)}$ itself is attained, up to a logarithmic factor.

The proof of the theorem is deferred to Section 3. In the next subsection we give the precise rate of contraction result for nonparametric regression, density estimation, and classification settings. The first case, which was the original motivation for this study, is worked out in some detail. The analogous results for the second and third settings are presented more briefly, to avoid unnecessary duplications.

2.2. Rate of contraction results for specific statistical settings.

2.2.1. *Regression with Gaussian errors.* Consider a multivariate regression problem where we have known design points $x_1, x_2, \dots \in \mathcal{X} = [a, b]^d$ for some $a < b$ and $d \in \mathbb{N}$, and we observe real-valued variables Y_1, \dots, Y_n satisfying the regression relation

$$Y_i = \theta(x_i) + \varepsilon_i,$$

for $\theta : \mathcal{X} \rightarrow \mathbb{R}$ an unknown regression function and error variables ε_i that are independent and Gaussian, with mean 0 and variance τ^2 . We assume that $0 < a < b < 1$, so that the design space \mathcal{X} is strictly contained in the interior of the unit cube in \mathbb{R}^d .

As prior on the regression function we employ the law Π_Θ that the stochastic process W defined by (2.1) generates on the space $C(\mathcal{X})$ of continuous functions on \mathcal{X} . The total prior Π on the pair (θ, τ) is then defined by $\Pi(d\theta, d\tau) = \Pi_\Theta(d\theta) \times \Pi_T(d\tau)$, for Π_T a prior on a compact interval that is assumed to contain the true value τ_0 , with a Lebesgue density that is bounded away from 0.

The posterior distribution for (θ, τ) given the data Y_1, \dots, Y_n is denoted by $\Pi(\cdot | Y_1, \dots, Y_n)$. By Bayes' formula, it is given by the expression

$$\Pi(B | Y_1, \dots, Y_n) = \frac{\int_B L(\theta, \tau; Y_1, \dots, Y_n) \Pi(d\theta, d\tau)}{\int L(\theta, \tau; Y_1, \dots, Y_n) \Pi(d\theta, d\tau)},$$

where

$$L(\theta, \tau; Y_1, \dots, Y_n) = \frac{1}{(2\pi\tau^2)^{n/2}} \exp\left(-\frac{1}{2\tau^2} \sum_{i=1}^n (Y_i - \theta(x_i))^2\right)$$

is the likelihood. For a given sequence of positive numbers $\varepsilon_n \downarrow 0$, the posterior is said to contract around the true parameter (θ_0, τ_0) at the rate ε_n if for $L > 0$ sufficiently large,

$$\Pi\left((\theta, \tau) : \frac{1}{n} \sum_{j=1}^n (\theta(x_j) - \theta_0(x_j))^2 + |\tau - \tau_0|^2 > L^2 \varepsilon_n^2 \mid Y_1, \dots, Y_n\right) \xrightarrow{P^{(\theta_0, \tau_0)}} 0$$

as $n \rightarrow \infty$, where the convergence is in probability under the true distribution governed by (θ_0, τ_0) . This means in particular that asymptotically, the marginal posterior for θ is concentrated on balls with radius of the order ε_n around the true regression function θ_0 , where we use the natural L^2 -norm associated to the empirical measure of the design points to measure distance.

The following theorem follows from Theorem 2.2, in combination with the results in [7] (slightly adapted like Theorem 2.1 of [6] in the density estimation case, cf. also the discussion following Theorem 3.1 of [26]).

THEOREM 2.3. *Suppose that the conditions of Theorem 2.2 are fulfilled. Then if $\theta_0 \in C^\alpha(\mathcal{X})$ for $\alpha > 0$, the posterior contracts at the rate*

$$n^{-\frac{\alpha(1-(d\delta\gamma)/(2\gamma))}{(d\gamma+2\alpha(1+\delta\gamma))(1+d/(2\gamma))}}$$

if $\gamma < \infty$, or at the rate

$$n^{-\frac{\alpha}{d+2\alpha} \log \frac{r\sqrt{(1+d)} + (\frac{1+d-r}{2})_+}{2+d/\alpha}} n$$

if $\gamma = \infty$.

As discussed above already the choice $p \in \mathcal{P}_\infty$ yields the best rate of contraction, namely the optimal minimax rate, up to a logarithmic factor. Also note that the prior does not depend on the unknown regularity α of the true regression function, so the procedure is rate-adaptive. Observe that for $p \in \mathcal{P}_\infty$ and $r = 1 + d$ we obtain the rate $(n/\log^{1+d} n)^{-\alpha/(d+2\alpha)}$. If r is strictly larger or smaller than $1 + d$ we get a slightly worse rate, in the sense that the power of the logarithm in our upper bound for the rate increases.

In the following corollary we single out the important special case of a standard Gaussian kernel and an inverse gamma prior (or a power of it in the multivariate case) on the scale.

COROLLARY 2.4. *Suppose that p is the standard Gaussian density on \mathbb{R}^d , Σ^d is inverse gamma, and M is such that $\mathbb{P}(M = m) \geq Cm^{-s}$ for some $C > 0$ and $s > 1$. Then if $\theta_0 \in C^\alpha(\mathcal{X})$ for $\alpha > 0$, the posterior contracts at the rate*

$$n^{-\frac{\alpha}{d+2\alpha} \log \frac{4\alpha+4\alpha d+d+d^2}{4\alpha+2d}} n.$$

PROOF. Simply note that the standard normal kernel belongs to \mathcal{P}_∞ and that if Σ^d has an inverse gamma law, then (2.2) is satisfied with $r = 0$. \square

2.2.2. *Density estimation.* Let X_1, \dots, X_n be a sample from a positive density f_0 on the set $\mathcal{X} = [a, b]^d$, for $0 < a < b < 1$. The aim is to estimate the unknown density.

We consider the prior Π on densities defined as the law that is generated on the function space $C(\mathcal{X})$ by the random function

$$(2.6) \quad x \mapsto \frac{e^{W(x)}}{\int_{\mathcal{X}} e^{W(y)} dy},$$

for W the process defined by (2.1). In this case we say that the posterior $\Pi(\cdot | X_1, \dots, X_n)$ contracts around the true density f_0 at the rate ε_n if for all $L > 0$ large enough,

$$\Pi(f : h(f, f_0) > L\varepsilon_n | X_1, \dots, X_n) \xrightarrow{P_{f_0}} 0$$

as $n \rightarrow \infty$, where h is the Hellinger distance.

Theorem 2.2, the general rate of contraction results for Bayesian density estimation (cf. [8], [6]) and the relations between the uniform norm on the paths of W and the relevant statistical metrics on the densities (2.6) (cf. [27]) yield the following result.

THEOREM 2.5. *In this setting the assertions of Theorem 2.3 and Corollary 2.4 are true for $\theta_0 = \log f_0$.*

2.2.3. *Classification.* Consider i.i.d. observations $(X_1, Y_1), \dots, (X_n, Y_n)$, where the X_i take values in the set $\mathcal{X} = [a, b]^d$, $0 < a < b < 1$ and the Y_i take values in $\{0, 1\}$. The aim is to estimate the regression function $r_0(x) = \mathbb{P}(Y_1 = 1 | X_1 = x)$.

As prior on r_0 we use the law Π of the process $\Psi(W)$, where W is as in (2.1) and the link function $\Psi : \mathbb{R} \rightarrow (0, 1)$ is the logistic or normal distribution function. Let $\Pi(\cdot | (X_1, Y_1), \dots, (X_n, Y_n))$ denote the corresponding posterior and let G be the distribution of the covariate X_1 . With $\|\cdot\|_{2,G}$ the associated L^2 -norm, we say that the posterior contracts around the truth r_0 at the rate ε_n if for all large enough $L > 0$,

$$\Pi(r : \|r - r_0\|_{2,G} > L\varepsilon_n | (X_1, Y_1), \dots, (X_n, Y_n)) \xrightarrow{P_{r_0}} 0$$

as $n \rightarrow \infty$.

Theorem 2.2, the general rate of contraction results (cf. [8]) and the relations between the relevant norms (cf. [27]) yield the following result.

THEOREM 2.6. *In this setting the assertions of Theorem 2.3 and Corollary 2.4 are true for $\theta_0 = \Psi^{-1}(r_0)$.*

3. Proof of Theorem 2.2. We will find the appropriate sieves B_n and derive the inequalities (2.3)–(2.5) by using the fact that conditionally on the grid size M and the scale Σ , the process W is Gaussian. For fixed $m \in \mathbb{N}$ and $\sigma > 0$, we define the stochastic process $(W^{m,\sigma}(x) : x \in [0, 1]^d)$ by setting

$$W^{m,\sigma}(x) = \sum_{k \in \{1, \dots, m\}^d} Z_k \frac{1}{m^{d/2}} \frac{1}{\sigma^d} p\left(\frac{x - k/m}{\sigma}\right).$$

In the following subsection we first study some properties of the Gaussian process $W^{m,\sigma}$ that we will need to establish (2.3)–(2.5).

3.1. *Properties of $W^{m,\sigma}$.* Recall that in general, the reproducing kernel Hilbert space (RKHS) \mathbb{H} attached to a zero-mean Gaussian process X is defined as the completion of the linear space of functions $t \mapsto \mathbb{E}X(t)H$ relative to the inner product

$$\langle \mathbb{E}X(\cdot)H_1, \mathbb{E}X(\cdot)H_2 \rangle_{\mathbb{H}} = \mathbb{E}H_1H_2,$$

where H, H_1 and H_2 are finite linear combinations of the form $\sum_i a_i X(s_i)$ with $a_i \in \mathbb{R}$ and s_i in the index set of X . The following lemma describes the RKHS of the process $W^{m,\sigma}$. It is a direct consequence of a general result describing the RKHS of a Gaussian process admitting a series expansion, cf. Theorem 4.2 of [28] and the discussion following it.

LEMMA 3.1. *The reproducing kernel Hilbert space $\mathbb{H}^{m,\sigma}$ of $W^{m,\sigma}$ consists of all functions of the form*

$$(3.1) \quad h(x) = \sum_{k \in \{1, \dots, m\}^d} w_k \frac{1}{\sigma^d} p\left(\frac{x - k/m}{\sigma}\right), \quad x \in [0, 1]^d,$$

where the weights w_k range over the entire set of real numbers. The RKHS-norm is given by

$$(3.2) \quad \|h\|_{\mathbb{H}^{m,\sigma}}^2 = m^d \min_w \sum_{k \in \{1, \dots, m\}^d} w_k^2,$$

where the minimum is over all weights w_k for which the representation (3.1) holds true.

We remark that if the functions $x \mapsto p((x - k/m)/\sigma)$ on $[0, 1]^d$ are linearly independent, then the representation (3.1) of an element of the RKHS is necessarily unique and hence the minimum in (3.2) can be removed. For our

purpose it is however not important that these functions are independent for every fixed σ and m .

Next we consider the so-called centered small ball probabilities of the process $W^{m,\sigma}$, which are determined by its reproducing kernel Hilbert space. We use well-known results by Kuelbs and Li [16] and Li and Linde [18] that relate the metric entropy of the unit ball in the RKHS to the centered small ball probabilities of the process. The unit ball $\mathbb{H}_1^{m,\sigma}$ in the reproducing kernel Hilbert space $\mathbb{H}^{m,\sigma}$ is the set of all elements $h \in \mathbb{H}^{m,\sigma}$ such that $\|h\|_{\mathbb{H}^{m,\sigma}} \leq 1$.

To find an upper bound for the metric entropy of the unit ball, we embed it in appropriate space of functions for which an upper bound for the entropy is known, depending on the value of γ . First we consider the case $\gamma < \infty$. Let h be an element of $\mathbb{H}^{m,\sigma}$. By Lemma 3.1, it admits a representation (3.1), with the weights w_k such that $\|h\|_{\mathbb{H}^{m,\sigma}}^2 = m^d \sum w_k^2$. If $p \in \mathcal{P}_\gamma$ with $\gamma < \infty$, we get that $h \in C^\gamma([0,1]^d)$ and $\|h\|_\gamma \leq \sigma^{-(d+\gamma)} \|p\|_\gamma \|h\|_{\mathbb{H}^{m,\sigma}}$. Hence, we have $\mathbb{H}_1^{m,\sigma} \subset C_R^\gamma([0,1]^d)$ in this case, where $R = \sigma^{-(d+\gamma)} \|p\|_\gamma$. For $\gamma = \infty$ and h as before, it follows from the assumptions on p that the function h is in fact well defined on $S_\sigma = \{z \in \mathbb{C}^d : \forall j | \Im z_j | \leq \sigma\}$, is analytic on this set, and takes real values on \mathbb{R}^d . By the Cauchy-Schwarz inequality, it follows that

$$|h(z)|^2 \leq \frac{1}{\sigma^{2d}} \left(\sum_{k \in \{1, \dots, m\}^d} w_k^2 \right) \left(\sum_{k \in \{1, \dots, m\}^d} \left| p\left(\frac{z - k/m}{\sigma}\right) \right|^2 \right).$$

The last factor in the right-hand side is bounded from above by a multiple of m^d on the set S_σ . Hence, we obtain

$$(3.3) \quad |h(z)| \leq K \sigma^{-d} \|h\|_{\mathbb{H}^{m,\sigma}}$$

for every $z \in S_\sigma$, where the constant K only depends on the density p . Let \mathcal{G}_σ the set of all analytic functions on S_σ , uniformly bounded by $K \sigma^{-d}$ on that set, with K the same constant as in (3.3). The preceding shows that for the RKHS unit ball we have $\mathbb{H}_1^{m,\sigma} \subset \mathcal{G}_\sigma$ if $\gamma = \infty$.

We see that in all cases we can embed the RKHS unit ball $\mathbb{H}_1^{m,\sigma}$ in a function space independent of m , for which the metric entropy relative to the supremum norm on $[0,1]^d$ is essentially known. We have the following result.

LEMMA 3.2.

- If $\gamma < \infty$, then

$$\log N(\varepsilon, C_{\sigma^{-(d+\gamma)} \|p\|_\gamma}^\gamma([0,1]^d), \|\cdot\|_\infty) \leq K_0 \left(\frac{1}{\varepsilon \sigma^{d+\gamma}} \right)^{\frac{d}{\gamma}}$$

for all $\sigma, \varepsilon > 0$, with K_0 a constant independent of ε, m and σ .

- There exist $\varepsilon_0, \sigma_0 > 0$ such that

$$\log N(\varepsilon, \mathcal{G}_\sigma, \|\cdot\|_\infty) \leq K_1 \frac{1}{\sigma^d} \left(\log \frac{K_2}{\varepsilon \sigma^d} \right)^{1+d}$$

for $\varepsilon \in (0, \varepsilon_0)$ and $\sigma \in (0, \sigma_0)$, with constants $K_1, K_2 > 0$ that do not depend on ε or σ . For $\sigma > \sigma_0$, it holds that

$$\log N(\varepsilon, \mathcal{G}_\sigma, \|\cdot\|_\infty) \leq K_3 \left(\log \frac{1}{\varepsilon} \right)^{1+d}$$

for all $\varepsilon \in (0, \varepsilon_0)$, with $K_3 > 0$ a constant independent of ε and σ .

PROOF. The first statement is well known, see for instance Theorem 2.7.1 of [25]. The second statement is similar to the classical result given by Theorem 23 of [13], which gives the entropy for the class of analytic functions bounded by a constant on a strip in complex space. However, the proof of the present statement requires extra care to identify the role of σ , because it should not be considered as an irrelevant constant in our framework. We omit the details, since the proof of Lemma 4.5 of [26] is very similar. \square

In view of the observations preceding Lemma 3.2 we now have entropy bounds for the unit ball of the RKHS in all cases. Using the results from [16] and [18], these translate into results on the centered small ball probability of $W^{m,\sigma}$. The first statement of the following lemma follows from the preceding lemma in combination with the results of [18]. The second statement is derived from Lemma 3.2 by arguing as in the proof of lemma 4.6 in [26].

LEMMA 3.3.

- If $d/2 < \gamma < \infty$,

$$-\log \mathbb{P}(\|W^{m,\sigma}\|_\infty < \varepsilon) \leq K_0 \left(\frac{1}{\varepsilon \sigma^{d+\gamma}} \right)^{\frac{2d}{2\gamma-d}}$$

for all $\varepsilon, \sigma > 0$, with K_0 a constant independent of ε and σ .

- If $\gamma = \infty$, there exist $\varepsilon_0, \sigma_0, K_4 > 0$, not depending on ε and σ , such that

$$-\log \mathbb{P}(\|W^{m,\sigma}\|_\infty < \varepsilon) \leq K_4 \frac{1}{\sigma^d} \left(\log \frac{1}{\varepsilon \sigma^{1+d}} \right)^{1+d}$$

for all $\varepsilon \in (0, \varepsilon_0)$ and $\sigma \in (0, \sigma_0)$. For $\sigma \geq \sigma_0$ we have

$$-\log \mathbb{P}(\|W^{m,\sigma}\|_\infty < \varepsilon) \leq K_5 \left(\log \frac{1}{\varepsilon} \right)^{1+d}$$

for all $\varepsilon \in (0, \varepsilon_0)$, where $K_5 > 0$ is independent of ε and σ .

With condition (2.5) in mind, we now consider the non-centered small ball probabilities of the process $W^{m,\sigma}$. According to Lemma 5.3 of [28] we have for $w_0 \in C([0, 1]^d)$ the inequality

$$(3.4) \quad -\log \mathbb{P}\left(\|W^{m,\sigma} - w_0\|_\infty < 2\varepsilon\right) \leq \varphi_{w_0}^{m,\sigma}(\varepsilon),$$

with $\varphi_{w_0}^{m,\sigma}$ the so-called concentration function, defined as follows:

$$(3.5) \quad \varphi_{w_0}^{m,\sigma}(\varepsilon) = \inf_{h \in \mathbb{H}^{m,\sigma}: \|h - \theta_0\|_\infty \leq \varepsilon} \|h\|_{\mathbb{H}^{m,\sigma}}^2 - \log \mathbb{P}(\|W^{m,\sigma}\|_\infty < \varepsilon).$$

(Our function w_0 is actually defined only on \mathcal{X} , but we will extend it to all of $[0, 1]^d$ in an appropriate way later). That is to say, the exponent of the non-centered small ball probability involves the exponent of the centered small ball probability that we considered above and an approximation term that quantifies how well w_0 can be approximated by elements of the RKHS.

To obtain a suitable approximation we need an auxiliary result concerning the approximation of a smooth function f by convolutions. Define $m_k = \int y^k p(y) dy$ for $k \in \mathbb{N}_0^d$. Next, for $n \in \mathbb{N}_0^d$ we recursively define two collections of numbers c_n and d_n as follows. If $n = 1$ we put $c_n = 0$ and $d_n = -m_n/n!$. For $n \geq 2$, we define

$$(3.6) \quad c_n = - \sum_{\substack{n=l+k \\ l \geq 1, k \geq 1}} \frac{(-1)^k}{k!} m_k d_l, \quad d_n = \frac{(-1)^n m_n}{n!} + c_n.$$

Note that the numbers c_n and d_n are well defined and that they only depend on the moments of p . For a function $f \in C^\alpha(\mathbb{R}^d)$ and $\sigma > 0$ we define the transform $T_{\alpha,\sigma} f$ as follows:

$$(3.7) \quad T_{\alpha,\sigma} f = f - \sum_{j=1}^{\beta} \sum_{k=:j} d_k \sigma^j (D_k^j f).$$

Here β is largest integer strictly smaller than α and for a positive integer j and a multi-index $k \in \mathbb{N}_0^d$ with $k = j$, D_k^j is the j th order differential operator

$$D_k^j = \frac{\partial^j}{\partial x_1^{k_1} \dots \partial x_d^{k_d}}.$$

Let $p_\sigma(x) = \sigma^{-d} p(x/\sigma)$.

LEMMA 3.4. *For $\alpha, \sigma > 0$ and $f \in C^\alpha(\mathbb{R}^d)$ we have*

$$\|p_\sigma * (T_{\alpha,\sigma} f) - f\|_\infty \leq K_6 \sigma^\alpha,$$

where $K_6 > 0$ is a constant independent of σ .

The lemma is an extension of an idea of [21], where a similar method is employed to approximate arbitrary smooth densities by beta mixtures. The proof follows the same lines but is somewhat more involved in the present higher-dimensional case, see Appendix A.

The following lemma deals with the approximation of the function w_0 by elements of the RKHS of the process $W^{m,\sigma}$.

LEMMA 3.5. *For all $\sigma > 0$, $m \geq 1$ and $w_0 \in C^\alpha(\mathcal{X})$ there exists an $h \in \mathbb{H}^{m,\sigma}$ such that $\|h\|_{\mathbb{H}^{m,\sigma}} \leq K_7(1 \vee \sigma)$ and*

$$\sup_{x \in \mathcal{X}} |h(x) - w_0(x)| \leq \frac{K_8(1 \vee \sigma^{\beta+1})}{\sigma^{1+d} m^{\alpha-\beta}} + K_9 \sigma^\alpha,$$

for $K_7, K_8, K_9 > 0$ constants independent of σ and m and β the largest integer strictly smaller than α .

PROOF. Since $\mathcal{X} = [a, b]^d \subset (0, 1)^d$ we can extend w_0 to all of \mathbb{R}^d in such a way that that the resulting function belongs to $C^\alpha(\mathbb{R}^d)$ and has support strictly inside $(0, 1)^d$. Using the operator $T_{\alpha,\sigma}$ introduced above (see (3.7)) we define

$$h(x) = \sum_{k \in \{1, \dots, m\}^d} (T_{\alpha,\sigma} w_0)(k/m) \frac{1}{m^d} \frac{1}{\sigma^d} p\left(\frac{x - k/m}{\sigma}\right)$$

for $x \in [0, 1]^d$. By Lemma 3.1 it holds that $h \in \mathbb{H}^{m,\sigma}$ and

$$\|h\|_{\mathbb{H}^{m,\sigma}}^2 \leq \frac{1}{m^d} \sum_{k \in \{1, \dots, m\}^d} \left((T_{\alpha,\sigma} w_0)(k/m) \right)^2 \leq \|T_{\alpha,\sigma} w_0\|_\infty^2.$$

It follows from the definition of $T_{\alpha,\sigma}$ that this is bounded by a constant times $(1 \vee \sigma^\beta)^2$.

It remains to prove the bound for the approximation error. By the triangle inequality,

$$(3.8) \quad \|h - w_0\|_\infty \leq \|h - p_\sigma * (T_{\alpha,\sigma} w_0)\|_\infty + \|p_\sigma * (T_{\alpha,\sigma} w_0) - w_0\|_\infty.$$

The first term on the right is the difference between the convolution $p_\sigma * T_{\alpha,\sigma} w_0$ and the corresponding Riemann sum. Using again the triangle in-

equality we get

$$\begin{aligned}
& |h(x) - (p_\sigma * T_{\alpha,\sigma} w_0)(x)| \\
& \leq \sup_{\|y-z\|_\infty \leq 1/m} |T_{\alpha,\sigma} w_0(y) p_\sigma(x-y) - T_{\alpha,\sigma} w_0(z) p_\sigma(x-z)| \\
& \leq \|T_{\alpha,\sigma} w_0\|_\infty \sup_{\|y-z\|_\infty \leq 1/m} |p_\sigma(x-y) - p_\sigma(x-z)| \\
& \quad + \|p_\sigma\|_\infty \sup_{\|y-z\|_\infty \leq 1/m} |T_{\alpha,\sigma} w_0(y) - T_{\alpha,\sigma} w_0(z)|.
\end{aligned}$$

Now use the facts that $T_{\alpha,\sigma} w_0$ is bounded by a constant times $1 \vee \sigma^\beta$, p_σ is bounded by σ^{-d} times a constant, p is Lipschitz and the definition of $T_{\alpha,\sigma} w_0$ to see that

$$\|h - p_\sigma * T_{\alpha,\sigma} w_0\|_\infty \leq \frac{C_1(1 \vee \sigma^\beta)}{\sigma^{1+d}m} + \frac{C_2(1 \vee \sigma^\beta)}{\sigma^d m^{\alpha-\beta}} \leq \frac{C_3(1 \vee \sigma^{\beta+1})}{\sigma^{1+d} m^{\alpha-\beta}},$$

which covers the first term on the right of (3.8). Lemma 3.4 implies that the second term is bounded by a constant times σ^α . \square

By combining the preceding lemma with Lemma 3.3 and (3.4) we obtain the following result.

LEMMA 3.6. *Let $w_0 \in C^\alpha(\mathcal{X})$.*

- *If $\gamma < \infty$, there exist constants $\varepsilon_0, \sigma_0, K_1, K_2, K_3, K_4 > 0$, independent of σ and m , such that*

$$-\log \mathbb{P}\left(\sup_{x \in \mathcal{X}} |W^{m,\sigma}(x) - w_0(x)| < 2\varepsilon\right) \leq K_1 + K_2 \left(\frac{1}{\varepsilon \sigma^{d+\gamma}}\right)^{\frac{2d}{2\gamma-d}},$$

provided that

$$\frac{K_3}{\sigma^{1+d} m^{\alpha-\beta}} + K_4 \sigma^\alpha < \varepsilon < \varepsilon_0$$

and $\sigma \in (0, \sigma_0)$.

- *If $\gamma = \infty$, there exist constants $\varepsilon_0, \sigma_0, K_1, K_2, K_3, K_4 > 0$, independent of σ and m , such that*

$$-\log \mathbb{P}\left(\sup_{x \in \mathcal{X}} |W^{m,\sigma}(x) - w_0(x)| < 2\varepsilon\right) \leq K_1 + K_2 \frac{1}{\sigma^d} \left(\log \frac{1}{\varepsilon \sigma^{1+d}}\right)^{1+d},$$

provided that

$$\frac{K_3}{\sigma^{1+d} m^{\alpha-\beta}} + K_4 \sigma^\alpha < \varepsilon < \varepsilon_0$$

and $\sigma \in (0, \sigma_0)$.

3.2. Proof of Theorem 2.2.

3.2.1. Condition (2.5). By definition of the process W and conditioning,

$$\begin{aligned} & \mathbb{P}\left(\sup_{x \in \mathcal{X}} |W(x) - w_0(x)| \leq \varepsilon\right) \\ &= \sum_{m=1}^{\infty} \lambda_m \int_0^{\infty} g(\sigma) \mathbb{P}\left(\sup_{x \in \mathcal{X}} |W^{m,\sigma}(x) - w_0(x)| < \varepsilon\right) d\sigma, \end{aligned}$$

where $\lambda_m = \mathbb{P}(M = m)$. If $\gamma < \infty$, Lemma 3.6 implies that there exist constants $\varepsilon_0, C_1, C_2, C_3, C_4 > 0$, independent of σ and m , such that if $\varepsilon < \varepsilon_0$ and

$$\frac{1}{2} C_1 \varepsilon^{1/\alpha} < \sigma < C_1 \varepsilon^{1/\alpha} \leq 1, \quad m \geq C_2 \varepsilon^{-\frac{1+d+\alpha}{\alpha(\alpha-\beta)}},$$

then

$$-\log \mathbb{P}\left(\sup_{x \in \mathcal{X}} |W^{m,\sigma}(x) - w_0(x)| < \varepsilon\right) \leq C_3 + C_4 \left(\frac{1}{\varepsilon \sigma^{d+\gamma}}\right)^{\frac{2d}{2\gamma-d}}.$$

Hence, the probability of interest is bounded from below, for $\varepsilon < \varepsilon_0$, by

$$\begin{aligned} & e^{-C_3} \sum_{m \geq C_2 \varepsilon^{-\frac{1+d+\alpha}{\alpha(\alpha-\beta)}}} \lambda_m \int_{C_1 \varepsilon^{1/\alpha}/2}^{C_1 \varepsilon^{1/\alpha}} g(\sigma) \exp\left(-C_4 \left(\frac{1}{\varepsilon \sigma^{d+\gamma}}\right)^{\frac{2d}{2\gamma-d}}\right) d\sigma \\ & \geq C_5 \exp\left(-C_6 \varepsilon^{-\frac{\alpha+d+\gamma}{\alpha} \frac{2d}{2\gamma-d}}\right) \end{aligned}$$

for constants $C_5, C_6 > 0$. It follows that condition (2.5) is fulfilled for

$$(3.9) \quad \varepsilon_n = M_1 n^{-\frac{\alpha}{d_\gamma + 2\alpha(1+\delta_\gamma)}}$$

for $M_1 > 0$ an appropriate constant and $d_\gamma = 2d(d+\gamma)/(2\gamma-d)$, $\delta_\gamma = d/(2\gamma-d)$.

If $\gamma = \infty$, the same reasoning implies that there exist constants $C_5, C_6 > 0$ such that, for $\varepsilon > 0$ small enough,

$$\mathbb{P}\left(\sup_{x \in \mathcal{X}} |W(x) - w_0(x)| \leq \varepsilon\right) \geq C_5 e^{-C_6 \varepsilon^{-d/\alpha} \log^{r \vee (1+d)}(1/\varepsilon)}.$$

It follows that in this case, condition (2.5) is fulfilled for

$$(3.10) \quad \varepsilon_n = M_1 n^{-\frac{\alpha}{d+2\alpha}} \log^t n$$

for $M_1 > 0$ an appropriate constant, provided that $t \geq (r \vee (1+d))/(2+d/\alpha)$.

3.2.2. *Construction of the sets B_n and condition (2.4).* First suppose that $\gamma < \infty$ again. For $L, R, \varepsilon > 0$, we define

$$B = LC_{R^{-(d+\gamma)}\|p\|_\gamma}^\gamma([0, 1]^d) + \varepsilon\mathbb{B}_1,$$

where \mathbb{B}_1 is the unit ball of the space $C([0, 1]^d)$. The sieves B_n will be defined by making appropriate choices for the L, R and ε below. Recall that in this case $\mathbb{H}_1^{m,\sigma} \subset C_{\sigma^{-(d+\gamma)}\|p\|_\gamma}^\gamma([0, 1]^d)$. Hence, by Borell-Sudakov (see for instance [19]), with Φ the standard normal distribution function and for $\sigma \geq R$,

$$\begin{aligned} \mathbb{P}(W^{m,\sigma} \notin B) &\leq \mathbb{P}(W^{m,\sigma} \notin L\mathbb{H}_1^{m,\sigma} + \varepsilon\mathbb{B}_1) \\ &\leq 1 - \Phi(\Phi^{-1}(\mathbb{P}(\|W^{m,\sigma}\|_\infty \leq \varepsilon)) + L). \end{aligned}$$

By Lemma 3.3 we have, for $\sigma \geq R$ and $R \leq 1$,

$$\mathbb{P}(\|W^{m,\sigma}\|_\infty \leq \varepsilon) \geq e^{-K_6 R^{-d\gamma} \varepsilon^{-2d/(2\gamma-d)}}$$

for a constant $K_6 > 0$ and $\varepsilon > 0$ small enough. Since $\Phi^{-1}(y) \geq -\sqrt{(5/2)\log(1/y)}$ for $y \in (0, 1/2)$, it follows that

$$\begin{aligned} \mathbb{P}(W^{m,\sigma} \notin B) &\leq 1 - \Phi\left(L - \sqrt{(5/2)K_6 R^{-d\gamma} \varepsilon^{-2d/(2\gamma-d)}}\right) \\ &\leq e^{-\frac{1}{2}(L - \sqrt{(5/2)K_6 R^{-d\gamma} \varepsilon^{-2d/(2\gamma-d)}})^2}, \end{aligned}$$

for $\sigma \geq R$ and $L \geq \sqrt{(5/2)K_6 R^{-d\gamma} \varepsilon^{-2d/(2\gamma-d)}}$. By definition of W and conditioning,

$$\mathbb{P}(W \notin B) \leq \sum_{m=1}^{\infty} \lambda_m \int_R^{\infty} g(\sigma) \mathbb{P}(W^{m,\sigma} \notin B) d\sigma + \mathbb{P}(\Sigma < R).$$

By the preceding, the first term on the right is bounded by

$$e^{-\frac{1}{2}(L - \sqrt{(5/2)K_6 R^{-d\gamma} \varepsilon^{-2d/(2\gamma-d)}})^2}.$$

The assumption on g and a substitution show that the second term is bounded by

$$D_3 \int_{1/R}^{\infty} x^{q-2} e^{-D_4 x^{d\gamma} (\log x)^r} dx.$$

By Lemma 4.9 of [26] this is further bounded by

$$\frac{2D_3}{dD_4} \frac{(1/R)^{q-2-d\gamma+1}}{(\log(1/R))^r} e^{-D_4(1/R)^{d\gamma} (\log(1/R))^r} \leq e^{-\frac{1}{2}D_4(1/R)^{d\gamma} (\log(1/R))^r}$$

for R small enough.

Given $C > 1$, we now define the sieve B_n by

$$B_n = L_n C_{R_n^{-(d+\gamma)} \|p\|_\gamma}^\gamma ([0, 1]^d) + \varepsilon_n \mathbb{B}_1,$$

where ε_n is given by (3.9). To show that (2.4) holds, we have to show we can choose R_n and L_n such that

$$\frac{1}{R_n^{d_\gamma}} \log^r \frac{1}{R_n} \geq Cn\varepsilon_n^2$$

and

$$(L_n - \sqrt{(5/2)K_6 R_n^{-d_\gamma} \varepsilon_n^{-2d/(2\gamma-d)}})^2 \geq Cn\varepsilon_n^2.$$

Observe that if we take

$$\frac{1}{R_n^{d_\gamma}} = Mn^{\frac{d_\gamma + 2\alpha\delta_\gamma}{d_\gamma + 2\alpha(1+\delta_\gamma)}}$$

for a large enough constant M , the first condition is satisfied. The second condition is then fulfilled if we choose

$$L_n^2 = Nn^{\frac{d_\gamma + 4\alpha\delta_\gamma}{d_\gamma + 2\alpha(1+\delta_\gamma)}},$$

for N large enough.

Next we consider the case $\gamma = \infty$. Recall that \mathcal{G}_σ is the set of all analytic functions defined on the strip $S_\sigma = \{z \in \mathbb{C}^d : \forall j | \Im z_j | \leq \sigma\}$ that are bounded by $K\sigma^{-d}$ on S_σ . Arguing as before and now using that $\mathbb{H}_1^{m,\sigma} \subset \mathcal{G}_\sigma$ and $\mathcal{G}_{\sigma_1} \subseteq \mathcal{G}_{\sigma_2}$ if $\sigma_1 \geq \sigma_2$, we get, for $L, R, \varepsilon > 0$ and $B = L\mathcal{G}_R + \varepsilon\mathbb{B}_1$,

$$\mathbb{P}(W^{m,\sigma} \notin B) \leq e^{-\frac{1}{2}(L - \sqrt{(5/2)K_6 R^{-d} (\log(1/(\varepsilon R^{1+d})))^{1+d}})^2},$$

for $\sigma \geq R$ and $L \geq \sqrt{(5/2)K_6 R^{-d} (\log(1/(\varepsilon R^{1+d})))^{1+d}}$. By the same conditioning argument as before it follows that if, given $C > 1$, we define B_n in this case by

$$B_n = L_n \mathcal{G}_{R_n} + \varepsilon_n \mathbb{B}_1,$$

where ε_n is given by (3.10), then condition (2.4) is fulfilled if we choose R_n and L_n such that

$$\frac{1}{R_n^d} \log^r \frac{1}{R_n} \geq Cn\varepsilon_n^2$$

and

$$\left(L_n - \sqrt{(5/2)K_6 R_n^{-d} (\log(1/(\varepsilon_n R_n^{1+d})))^{1+d}} \right)^2 \geq Cn\varepsilon_n^2.$$

Observe that we can take

$$\frac{1}{R_n^d} = Mn^{\frac{d}{d+2\alpha}} \log^v n$$

for a large enough constant M and $v \geq 2t - r$ (with t as in (3.10)), and L_n a large enough power of n .

3.2.3. Entropy condition. Suppose $\gamma < \infty$. For the entropy of the sieve B_n we have in this case, for $\bar{\varepsilon}_n \geq \varepsilon_n$,

$$\begin{aligned} N(2\bar{\varepsilon}_n, B_n, \|\cdot\|_\infty) &\leq N(\bar{\varepsilon}_n, L_n C_{R_n^{-(d+\gamma)} \|p\|_\gamma}^\gamma([0, 1]^d), \|\cdot\|_\infty) \\ &\leq N(\bar{\varepsilon}_n R_n^{d+\gamma} / (L_n \|p\|_\gamma), C_1^\gamma([0, 1]^d), \|\cdot\|_\infty). \end{aligned}$$

Hence, see Lemma 3.2,

$$\log N(2\bar{\varepsilon}_n, B_n, \|\cdot\|_\infty) \leq K_1 \left(\frac{L_n}{\bar{\varepsilon}_n R_n^{d+\gamma}} \right)^{d/\gamma}.$$

This is bounded by a constant times $n\bar{\varepsilon}_n^2$ for

$$\bar{\varepsilon}_n \gtrsim \frac{L_n^{d/(d+2\gamma)}}{n^{\frac{\gamma}{d+2\gamma}} R_n^{\frac{d(d+\gamma)}{d+2\gamma}}}.$$

For L_n and R_n chosen as above this yields

$$\bar{\varepsilon}_n \gtrsim n^{-\frac{\alpha(1-(d\delta\gamma)/(2\gamma))}{d_\gamma+2\alpha(1+\delta\gamma)+(d(d_\gamma+2\alpha(1+\delta\gamma)))/(2\gamma)}}.$$

Note that $\bar{\varepsilon}_n$ is always larger than ε_n , as was required.

Let now $\gamma = \infty$. Arguing as before we have in this case, for $\bar{\varepsilon}_n \geq \varepsilon_n$,

$$N(2\bar{\varepsilon}_n, B_n, \|\cdot\|_\infty) \leq N(\bar{\varepsilon}_n/L_n, \mathcal{G}_{R_n}, \|\cdot\|_\infty) \leq K_1 \frac{1}{R_n^d} \left(\log \frac{L_n}{\bar{\varepsilon}_n R_n^d} \right)^{1+d},$$

by Lemma 3.2. With the choices of R_n and L_n made in this case above and for $\bar{\varepsilon}_n$ bounded from below by a power of n , this is bounded by a constant times $n^{\frac{d}{d+2\alpha}} \log^{1+d+v} n$. This is further bounded by a constant times $n\bar{\varepsilon}_n^2$ for

$$\bar{\varepsilon}_n = n^{-\frac{\alpha}{d+2\alpha}} \log^a n,$$

provided $a \geq (1 + d + v)/2$. The requirement that $\bar{\varepsilon}_n \geq \varepsilon_n$ translates into the condition $a \geq t$.

APPENDIX A

PROOF OF LEMMA 3.4. The proof is by induction on β , which is the largest integer strictly smaller than α . If $\beta = 0$ then $\alpha \in (0, 1]$ and $T_{\alpha, \sigma} f = f$ and the statement of the claim is standard. To prove the induction step, suppose now that $\beta \geq 1$. By definition of $T_{\alpha, \sigma} f$ we have

$$\begin{aligned} & (p_\sigma * T_{\alpha, \sigma} f - f)(x) \\ &= \int p_\sigma(y) \left(f(x-y) - f(x) - \sum_{j=1}^{\beta} \sum_{k=j} d_k \sigma^j (D_k^j f)(x-y) \right) dy. \end{aligned}$$

By Taylor's formula and the fact that $f \in C^\alpha$,

$$f(x-y) - f(x) = \sum_{j=1}^{\beta} \sum_{k=j} \frac{(-y)^k}{k!} (D_k^j f)(x) + R(x, y),$$

where $|R(x, y)| \leq C \|y\|^\alpha$. It follows that

$$\begin{aligned} & (p_\sigma * T_{\alpha, \sigma} f - f)(x) \\ &= \int p_\sigma(y) R(x, y) dy \\ &+ \sum_{j=1}^{\beta} \sum_{k=j} \left(\frac{1}{k!} (-1)^j (D_k^j f)(x) \sigma^j m_k - d_k \sigma^j (p_\sigma * (D_k^j f))(x) \right). \end{aligned}$$

The first term on the right is easily seen to be bounded by a constant times σ^α . To see that this holds for the second term as well we use the induction hypothesis.

By definition of the constants c_k and d_k (see (3.6)), the second term can be written as

$$\sum_{j=1}^{\beta} \sum_{k=j} \left(\frac{(-1)^j}{k!} \sigma^j m_k (D_k^j f - p_\sigma * (D_k^j f))(x) - c_k \sigma^j (p_\sigma * (D_k^j f))(x) \right).$$

Now for $j \leq \beta$ and $k. = j$, consider the decomposition

$$\begin{aligned} & D_k^j f - p_\sigma * (D_k^j f) \\ &= \left(D_k^j f - p_\sigma * (T_{\alpha-j, \sigma} D_k^j f) \right) + \left(p_\sigma * (T_{\alpha-j, \sigma} D_k^j f) - p_\sigma * (D_k^j f) \right). \end{aligned}$$

Since $D_k^j f \in C^{\alpha-j}$, the induction hypothesis implies that the first term on the right is uniformly bounded by a constant times $\sigma^{\alpha-j}$. Combined with

the first display of the paragraph, this shows that it suffices to show that

$$\sum_{j=1}^{\beta} \sum_{k=j} \left(\frac{(-1)^j}{k!} \sigma^j m_k \left(T_{\alpha-j, \sigma} D_k^j f - D_k^j f \right) - c_k \sigma^j \left(D_k^j f \right) \right) = 0$$

identically. Straightforward algebra shows that

$$T_{\alpha-j, \sigma} D_k^j f - D_k^j f = - \sum_{i=1}^{\beta-j} \sum_{l=i} d_l \sigma^i D_{k+l}^{i+j} f.$$

Hence,

$$\begin{aligned} \sum_{j=1}^{\beta} \sum_{k=j} \frac{(-1)^j}{k!} \sigma^j m_k \left(T_{\alpha-j, \sigma} D_k^j f - D_k^j f \right) &= - \sum_{j=1}^{\beta} \sum_{k=j} \sum_{i=1}^{\beta-j} \sum_{l=i} \frac{(-1)^j}{k!} m_k d_l \sigma^{i+j} D_{l+k}^{i+j} f \\ &= - \sum_{s=2}^{\beta} \sum_{n=s} \left(\sum_{\substack{n=l+k \\ l \geq 1, k \geq 1}} \frac{(-1)^k}{k!} m_k d_l \right) \sigma^s D_n^s f. \end{aligned}$$

By definition of the numbers c_n and d_n this equals

$$\sum_{s=1}^{\beta} \sum_{n=s} c_n \sigma^s D_n^s f,$$

and the proof is complete. \square

REFERENCES

- [1] Suddipto Banerjee, Alan E. Gelfand, Andrew O. Finley, and Huiyan Sang. Gaussian predictive process models for large spatial data sets. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(4):825–848, 2008. ISSN 1369-7412.
- [2] Eduard Belitser and Subhashis Ghosal. Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist.*, 31(2):536–559, 2003. ISSN 0090-5364. Dedicated to the memory of Herbert E. Robbins.
- [3] S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.*, 27(1):143–158, 1999. ISSN 0090-5364.
- [4] Subhashis Ghosal. Convergence rates for density estimation with Bernstein polynomials. *Ann. Statist.*, 29(5):1264–1280, 2001. ISSN 0090-5364.

- [5] Subhashis Ghosal and Aad van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.*, 35(2):697–723, 2007. ISSN 0090-5364.
- [6] 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. ISSN 0090-5364.
- [7] Subhashis Ghosal and Aad W. van der Vaart. Convergence rates for posterior distributions for noniid observations. *Ann. Statist.*, 35:697–723, 2007.
- [8] 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.
- [9] 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.
- [10] Subhashis Ghosal, Jüri Lember, and Aad Van Der Vaart. Nonparametric bayesian model selection and averaging. *Electronic Journal of Statistics*, 2:63–89, 2008.
- [11] Dave Higdon. Space and space-time modeling using process convolutions. In *Quantitative methods for current environmental issues*, pages 37–56. Springer, London, 2002.
- [12] Tzee-Ming Huang. Convergence rates for posterior distributions and adaptive estimation. *Ann. Statist.*, 32(4):1556–1593, 2004. ISSN 0090-5364.
- [13] A. N. Kolmogorov and V. M. Tihomirov. ε -entropy and ε -capacity of sets in functional space. *Amer. Math. Soc. Transl. (2)*, 17:277–364, 1961. ISSN 0065-9290.
- [14] Willem Kruijer and Aad van der Vaart. Posterior convergence rates for Dirichlet mixtures of beta densities. *J. Statist. Plann. Inference*, 138(7):1981–1992, 2008. ISSN 0378-3758.
- [15] Willem Kruijer, Judith Rousseau, and A.W. van der Vaart. Adaptive bayesian density estimation with location-scale mixtures. Preprint, 2009.
- [16] James Kuelbs and Wenbo V. Li. Metric entropy and the small ball problem for Gaussian measures. *J. Funct. Anal.*, 116(1):133–157, 1993. ISSN 0022-1236.
- [17] J. Lember and A.W. van der Vaart. On universal bayesian adaptation. *Statistics and Decisions*, 25:127–152, 2007.
- [18] Wenbo V. Li and Werner Linde. Approximation, metric entropy and small ball estimates for gaussian measures. *The Annals of Probability*, 27(3):1556–1578, 1999.
- [19] M. A. Lifshits. *Gaussian random functions*, volume 322 of *Mathematics and its Applications*. Kluwer Academic Publishers, Dordrecht, 1995. ISBN 0-7923-3385-3.
- [20] Sonia Petrone and Larry Wasserman. Consistency of Bernstein polynomial posteriors. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(1):79–100, 2002. ISSN 1369-7412.
- [21] Judith Rousseau. Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. Preprint, 2008.
- [22] Margaret B. Short, David M. Higdon, and Philipp P. Kronberg. Estimation of Faraday rotation measures of the near galactic sky using Gaussian process models. *Bayesian Anal.*, 2(4):665–680, 2007. ISSN 1931-6690.
- [23] Surya T. Tokdar. Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā*, 68(1):90–110, 2006. ISSN 0972-7671.
- [24] 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. ISSN 1350-7265.
- [25] 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.
- [26] A.W. van der Vaart and J.H. van Zanten. Adaptive bayesian estimation using

- a gaussian random field with inverse gamma bandwidth. *Annals of Statistics*, to appear, 2009.
- [27] A.W. Van der Vaart and J.H. Van Zanten. Rates of contraction of posterior distributions based on gaussian process priors. *Annals of Statistics*, 36, 2008.
- [28] 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.
- [29] Yuefeng Wu and Subhashis Ghosal. Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electron. J. Stat.*, 2:298–331, 2008. ISSN 1935-7524.

DEPARTMENT OF MATHEMATICS
P.O. BOX 513
5600 MB EINDHOVEN
THE NETHERLANDS
E-MAIL: r.d.jonge@tue.nl
j.h.v.zanten@tue.nl