Submitted to the Annals of Statistics

ADAPTIVE DIMENSION REDUCTION WITH A GAUSSIAN PROCESS PRIOR

By Anirban Bhattacharya*, Debdeep Pati* and David Dunson†

Department of Statistical Science, Duke University‡

In nonparametric regression problems involving multiple predictors, there is typically interest in estimating the multivariate regression surface in the important predictors while discarding the unimportant ones. Our focus is on defining a Bayesian procedure that leads to the minimax optimal rate of posterior contraction (up to a log factor) adapting to the unknown dimension and anisotropic smoothness of the true surface. We propose such an approach based on a Gaussian process prior with dimension-specific scalings, which are assigned carefully-chosen hyperpriors. The prior also leads to consistent Bayesian variable selection. We additionally show that using a homogenous Gaussian process with a single bandwidth leads to a sub-optimal rate in anisotropic cases.

1. Intorduction. Non-parametric function estimation methods have been immensely popular due to their ability to adapt to a wide variety of function classes with unknown regularities. In Bayesian nonparametrics, Gaussian processes (Rasmussen, 2004; van der Vaart and van Zanten, 2008b) are widely used as priors on functions due to tractable posterior computation and attractive theoretical properties. The law of a mean zero Gaussian process W_t is entirely characterized by its covariance kernel $C(s,t) = E(W_s W_t)$, and given any positive definite function $C(s,t)$, one can construct a valid Gaussian process with C as its covariance kernel. A squared exponential covariance kernel given by $C(s,t) = \exp(-a\,\|s-t\|^2)$ is commonly used in the literature.

It is well established (Stone, 1982) that given n independent observations, the optimal rate of estimation of a d-variable function that is only known to be α-smooth is $n^{-\alpha/(2\alpha+d)}$. The quality of estimation thus improves with increasing smoothness of the "true" function. In practice, the smoothness level α is typically unknown and thus one would like to have a unified estimation procedure that automatically adapts to all possible smoothness levels of

*PhD. Student, Department of Statistical Science

†Professor, Department of Statistical Science

AMS 2000 subject classifications: Primary 62G07, 62G20; secondary 60K35

Keywords and phrases: Adaptive, Anisotropic, Bayesian nonparametrics, Function estimation, Gaussian process, Rate of convergence, Variable selection

the true function. Accordingly, a lot of effort has been employed to develop adaptive estimation methods that are rate-optimal for every regularity level of the unknown function.

The literature on adaptive estimation in a minimax setting was initiated by Lepski in a series of papers (Lepski, 1990, 1991, 1992); see also Birgé (2001) for a discussion on this topic. We also refer the reader to Hoffmann and Lepski (2002), which contains an extensive list of developments in the frequentist literature on adaptive estimation. A key idea in adaptive estimation is to narrow down the search for an "optimal" estimator within a class of estimators indexed by a smoothness or bandwidth parameter, and make a data-driven choice to select the proper bandwidth.

In a Bayesian context, one would place a prior on the bandwidth parameter and model-average across different values of the bandwidth through the posterior distribution. The parameter a in the squared-exponential covariance kernel C plays the role of a scaling or inverse bandwidth. van der Vaart and van Zanten (2009) showed that with a gamma prior on a^d, one obtains the minimax rate of posterior contraction $n^{-\alpha/(2\alpha+d)}$ up to a logarithmic factor for α-smooth functions adaptively over all $\alpha > 0$.

In multivariate problems involving even moderate number of dimensions, the assumption of the true function being in an isotropic smoothness class characterized by a single smoothness parameter seems restrictive. Practitioners often use a non-homogeneous variant of the squared exponential covariance kernel given by $C(s,t) = \exp(-\sum_{j=1}^{d} a_j|s_j - t_j|^2)$. A separate scaling variable a_j for the different dimensions incorporates dimension specific effects in the covariance kernel, intuitively enabling better approximation of functions in anisotropic smoothness classes. In particular, one can let a subset of the covariates drop out of the covariance kernel by setting some of the scales a_j to zero. Such a model was recently studied in Savitsky, Vannucci and Sha (2011), who used a point mass mixture prior on $\rho_j = -\log a_j \in [0,1]$. Zou et al. (2010) also used a similar model for high-dimensional non-parametric variable selection. Although this is an attractive scheme for anisotropic modeling and dimension reduction in non-parametric regression problems with encouraging empirical performance, there hasn't been any theoretical study on rates of convergence in related models in a Bayesian framework to our knowledge.

In the frequentist literature, minimax rates of convergence in anisotropic Sobolev, Besov and Hölder spaces have been studied in Birgé (1986); Ibragimov and Khasminski (1981); Nussbaum (1985), with adaptive estimation procedures developed in Barron, Birgé and Massart (1999); Hoffmann and Lepski (2002); Kerkyacharian, Lepski and Picard (2001); Klutchnikoff (2005)

among others. The traditional way of dealing with anisotropy is to employ a separate bandwidth or scaling parameter for the different dimensions, and choose an optimal combination of scales in a data-driven way. However, the multidimensional nature of the problem makes the optimal bandwidth selection considerably difficult compared to the isotropic case, as there is no natural ordering among the estimators with multiple bandwidths.

It is known (Hoffmann and Lepski, 2002) that the minimax rate of convergence for a function with smoothness α_i along the ith dimension is given by $n^{-\alpha_0/(2\alpha_0+1)}$, where $\alpha_0^{-1} = \sum_{i=1}^{d} \alpha_i^{-1}$ is an *exponent of global smoothness* (Birgé, 1986). When $\alpha_i = \alpha$ for all $i = 1, \ldots, d$, one reduces back to the optimal rate for isotropic classes. On the contrary, if the true function belongs to an anisotropic class, the assumption of isotropy would lead to loss of efficiency which would be more and more accentuated in higher dimensions. In addition, if the true function depends on a subset of coordinates $I = \{i_1, \ldots, i_{d_0}\} \subset \{1, \ldots, d\}$ for some $1 \leq d_0 \leq d$, the minimax rate would further improve to $n^{-\alpha_{0I}/(2\alpha_{0I}+1)}$, with $\alpha_{0I}^{-1} = \sum_{j \in I} \alpha_j^{-1}$.

The objective of this article is to study whether one can fully adapt to this larger class of functions in a Bayesian framework using dimension specific rescalings of a homogenous Gaussian process, referred to as a multi-bandwidth Gaussian process (GP) from now on. We answer the question in the affirmative and exhibit a class of priors which lead to the optimal rate $n^{-\alpha_{0I}/(2\alpha_{0I}+1)}$ of posterior contraction (up to a log term) for any α and I without prior knowledge of either of them.

The general sufficient conditions for obtaining posterior rates of convergence (Ghosal, Ghosh and van der Vaart, 2000) involve finding a sequence of compact and increasing subsets of the parameter space, usually referred to as sieves, which are "not to large" in the sense of metric entropy and yet capture most of the prior mass. van der Vaart and van Zanten (2008a) developed a general technique for constructing such sieves with Gaussian process priors, which involved subtle manipulations with the reproducing kernel Hilbert space (RKHS) of a Gaussian process (van der Vaart and van Zanten, 2008b). A key technical advancement in van der Vaart and van Zanten (2009) was to extend the above theoretical framework to the setting of conditionally Gaussian random fields. In particular, they crucially exploited a containment relation among the unit RKHS balls with different bandwidths to construct the sieves B_n in their framework. Their construction can be conceptually related to the general framework for adaptive estimation developed in Lepski (1990, 1991, 1992), where a natural ordering among kernel estimators with different scalar bandwidths is utilized to compare different estimators and balance the bias-variance trade-off. However, it gets

significantly more complicated in situations involving multiple bandwidths to compare kernel estimators with different vectors of bandwidths. In multi-bandwidth Gaussian processes, a similar problem arises in comparing unit RKHS balls of Gaussian processes with different vectors of bandwidths, and the techniques of van der Vaart and van Zanten (2009) cannot be immediately extended to this case.

Our main contribution is to address the above issue by a novel prior specification on the vector of bandwidths and a careful construction of the sieves B_n, which can be used to establish rate adaptiveness of the posterior distribution in variety of settings involving a multi-bandwidth Gaussian process. For simplicity of exposition, we initially study the problem in two parts: (i) adaptive estimation over anisotropic Hölder functions of d arguments, and (ii) adaptive estimation over functions that can possibly depend on fewer coordinates and have isotropic Hölder smoothness over the remaining coordinates. In each of these cases, we propose a joint prior on the bandwidths induced through a hierarchical Bayesian framework. To avoid the problem of comparing between different vector of scales, we aggregate over a collection of bandwidth vectors to construct the sets B_n. New results are developed to bound the metric entropy of such collections of unit RKHS balls. Combining these results, we balance the metric entropy of the sieve and the prior probability of its complement. The prior specifications for the two cases above are easy to interpret intuitively and can be easily connected to combine the cases (i) and (ii).

Although our prior specification involving dimension-specific bandwidth parameters leads to adaptivity, a stronger result is required to conclude that a single bandwidth would be inadequate for the above classes of functions. We prove that the optimal prior choice in the isotropic case leads to a sub-optimal convergence rate if the true function depends on fewer coordinates by obtaining a lower bound on the posterior contraction rate. The general sufficient conditions for rates of posterior contraction provide an upper bound on the rate of convergence implying the posterior contracts at least as fast as the rate obtained. Castillo (2008) studied lower bounds for posterior contraction rate with a class of Gaussian process priors. e extend the results of Castillo (2008) to the setting of rescaled Gaussian process priors. We develop a novel technique for deriving a sharp lower bound to the concentration function of a rescaled Gaussian process. This technique can be used for comparing the posterior convergence rates obtained for different prior distributions on the bandwidth parameter.

The remaining paper is organized as follows. In Section 2, we introduce relevant notations. Section 3 discusses the main developments with appli-

cations to anisotropic Gaussian process mean regression and Logistic Gaussian process density estimation described in subsection 3.3. In Section 4, we study various properties of multi-bandwidth Gaussian processes which are crucially used in the proofs of the main theorems in Section 5 and should also be of independent interest. Section 6 establishes the necessity of the multi-bandwidth GP by showing that a single rescaling can lead to sub-optimal rates when the true function is lower-dimensional.

2. Notations. We shall make frequent use of the following multi-index notations. For vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, let $\mathbf{a}. = \sum_{j=1}^d a_j, \mathbf{a}^* = \prod_{j=1}^d a_j, \mathbf{a}! = \prod_{j=1}^d a_j!, \bar{\mathbf{a}} = \max_j a_j, \underline{\mathbf{a}} = \min_j a_j, \mathbf{a}./\mathbf{b} = (a_1/b_1, \ldots, a_d/b_d)^{\mathrm{T}}, \mathbf{a} \cdot \mathbf{b} = (a_1 b_1, \ldots, a_d b_d)^{\mathrm{T}}, \mathbf{a}^{\mathbf{b}} = \prod_{j=1}^d a_j^{b_j}$. Denote $\mathbf{a} \leq \mathbf{b}$ if $a_j \leq b_j$ for all $j = 1, \ldots, d$. For $\boldsymbol{n} = (n_1, \ldots, n_d)$, let $D^{\boldsymbol{n}} f$ denote the mixed partial derivatives of order $(n_1, \ldots, n_d)$ of f.

Let $C[0,1]^d$ and $C^\beta[0,1]^d$ denote the space of all continuous functions and the Hölder space of β-smooth functions $f : [0,1]^d \to \mathbb{R}$ respectively, endowed with the supremum norm $\|f\|_\infty = \sup_{t \in [0,1]^d} |f(t)|$. For $\beta > 0$, the Hölder space $C^\beta[0,1]^d$ consists of functions $f \in C[0,1]^d$ that have bounded mixed partial derivatives up to order $\lfloor \beta \rfloor$, with the partial derivatives of order $\lfloor \beta \rfloor$ being Lipschitz continuous of order $\beta - \lfloor \beta \rfloor$.

Next, we define an anisotropic Hölder class of functions previously used in Barron, Birgé and Massart (1999); Klutchnikoff (2005). For a function $f \in C[0,1]^d$, $x \in [0,1]^d$, and $1 \leq i \leq d$, let $f_i(\cdot \mid x)$ denote the univariate function $y \mapsto f(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_d)$. For a vector of positive numbers $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$, the anisotropic Hölder space $C^{\boldsymbol{\alpha}}[0,1]^d$ consists of functions f which satisfy, for some $L > 0$,

$$\max_{1 \leq i \leq n} \sup_{x \in [0,1]^d} \sum_{j=0}^{\lfloor \alpha_i \rfloor} \left\| D^j f_i(\cdot \mid x) \right\|_\infty \leq L, \tag{2.1}$$

and, for any $y \in [0,1]$, h small such that $y + h \in [0,1]$ and for all $1 \leq i \leq d$,

$$\sup_{x \in [0,1]^d} \left\| D^{\lfloor \alpha_i \rfloor} f_i(y + h \mid x) - D^{\lfloor \alpha_i \rfloor} f_i(y \mid x) \right\|_\infty \leq L \, |h|^{\alpha_i - \lfloor \alpha_i \rfloor}. \tag{2.2}$$

For $t \in \mathbb{R}^d$ and a subset $I \subset \{1, \ldots, d\}$ of size $|I| = \tilde{d}$ with $1 \leq \tilde{d} \leq d$, let t_I denote the vector of size $\tilde{d}$ consisting of the coordinates $(t_j : j \in I)$. Let $C[0,1]^I$ (resp. $C^{\boldsymbol{\alpha}}[0,1]^I$) denote the subset of $C[0,1]^d$ (resp. $C^{\boldsymbol{\alpha}}[0,1]^d$) consisting of functions f such that $f(t) = g(t_I)$ for some function $g \in C[0,1]^{\tilde{d}}$ (resp. $C^{\boldsymbol{\alpha}_I}[0,1]^{\tilde{d}}$).

The ϵ-covering number $N(\epsilon, S, d)$ of a semi-metric space S relative to the semi-metric d is the minimal number of balls of radius ϵ needed to cover S. The logarithm of the covering number is referred to as the entropy.

We write "$\precsim$" for inequality up to a constant multiple. Let $\phi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$ denote the standard normal density, and let $\phi_\sigma(x) = (1/\sigma)\phi(x/\sigma)$. Let an asterisk denote a convolution e.g., $(\phi_\sigma * f)(y) = \int \phi_\sigma(y-x)f(x)dx$. Let $\hat{f}$ denote the Fourier transform of a function f whenever it is defined. Denote by $\mathcal{S}_{d-1}$ the $d-1$-dimensional simplex consisting of points $\{x \in \mathbb{R}^d : x_i \geq 0, 1 \leq i \leq d, \sum_{i=1}^d x_i = 1\}$.

2.1. *RKHS of Gaussian processes.* We briefly recall the definition of the RKHS of a Gaussian process prior next; a detailed review of the facts relevant to the present application can be found in van der Vaart and van Zanten (2008b). A Borel measurable random element W with values in a separable Banach space $(\mathbb{B}, \|\cdot\|)$ (e.g., $C[0,1]$) is called Gaussian if the random variable b^*W is normally distributed for any element $b^* \in \mathbb{B}^*$, the dual space of $\mathbb{B}$. The reproducing kernel Hilbert space (RKHS) $\mathbb{H}$ attached to a zero-mean Gaussian process W is defined as the completion of the linear space of functions $t \mapsto EW(t)H$ relative to the inner product

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

where H, H_1 and H_2 are finite linear combinations of the form $\sum_i a_i W(s_i)$ with $a_i \in \mathbb{R}$ and s_i in the index set of W. The RKHS of a Gaussian process plays an important role in determining the support and concentration properties of the process.

A key feature of a Gaussian process controlling the rate of posterior contraction is the concentration function $\phi_{w_0}(\epsilon)$ (van der Vaart and van Zanten, 2008b, 2009) given by,

$$\phi_{w_0}(\epsilon) = \inf_{h:\mathbb{H}:\|h-w_0\|_\infty \leq \epsilon} \|h\|_{\mathbb{H}}^2 - \log P(\|W\|_\infty \leq \epsilon). \tag{2.3}$$

3. Main results. Let $W = \{W_t : t \in [0,1]^d\}$ be a centered homogeneous Gaussian process with covariance function $\mathrm{E}(W_sW_t) = c(s-t)$. By Bochner's theorem, there exists a finite positive measure ν on $\mathbb{R}^d$, called the spectral measure of W, such that

$$c(t) = \int e^{-i\langle\lambda,t\rangle}\nu(d\lambda).$$

As in van der Vaart and van Zanten (2009), we shall restrict ourselves to processes with spectral measure ν having sub-exponential tails, i.e., for some

$\delta > 0$,

$$\int e^{\delta\|\lambda\|}\nu(d\lambda) < \infty. \tag{3.1}$$

The spectral measure of a squared exponential covariance kernel with $c(t) = \exp(-\|t\|^2)$ has a density w.r.t. the Lebesgue measure given by $f(\lambda) = 1/(2^d\pi^{d/2})\exp(-\|\lambda\|^2/4)$ which clearly satisfies (3.1).

Rates of posterior contraction with Gaussian process priors were first studied by van der Vaart and van Zanten (2008a), who gave sufficient conditions in terms of the concentration function of a Gaussian random element for optimal rate of convergence in a variety of statistical problems including density estimation using the Logistic Gaussian process (Lenk, 1988, 1991), Gaussian process mean regression, latent Gaussian process regression (e.g., in logit, probit models), binary classification, etc. As indicated in the introduction, one needs to build appropriate sieves in the space of continuous functions to get a handle on the posterior rates of convergence in such models. van der Vaart and van Zanten (2008a) constructed the sieves as a collection of continuous functions within a small (sup-norm) neighborhood of a norm-bounded subset of the RKHS. Sharp bounds on the complement probability of such sets can be obtained using Borell's inequality (**?**), and the metric entropy can also be appropriately controlled exploiting the fact that the RKHS consists of smooth functions if the covariance kernel is smooth. It is important to mention here that a similar strategy involving a subset of continuous functions bounded in sup-norm doesn't work beyond the uni-dimensional case (Tokdar and Ghosh, 2007).

Given a homogeneous smooth Gaussian random field W, van der Vaart and van Zanten (2009) considered rescaled Gaussian processes $W^A = \{W_{At} : t \in [0,1]^d\}$ for a real positive random variable A stochastically independent of W, extending the framework of van der Vaart and van Zanten (2008a) to the setting of conditionally Gaussian random elements (see also De Jonge and van Zanten (2010) for a different class of conditionally Gaussian processes). Since W has infinitely smooth sample paths, it is not suitable for modeling less smooth functions. Rescaling the sample paths of an infinitely smooth Gaussian process is a powerful technique to improve the approximation of α-Hölder functions from the RKHS of the scaled process. Intuitively, for large values of A, the scaled process traverses the sample path of an unscaled process on the larger interval $[0, A]^d$, thereby incorporating more "roughness".

The key result of van der Vaart and van Zanten (2009) was to construct the sieves $B_n \subset C[0,1]^d$ so that given $\alpha > 0$, a function $w_0 \in C^\alpha[0,1]^d$

and a constant $C > 1$, there exists a constant $D > 0$ such that, for every sufficiently large n,

$$\log N(\bar{\epsilon}_n, B_n, \|\cdot\|_\infty) \leq Dn\bar{\epsilon}_n^2, \tag{3.2}$$

$$\mathrm{P}(W^A \notin B_n) \leq exp(-Cn\epsilon_n^2), \tag{3.3}$$

$$\mathrm{P}(\left\|W^A - w_0\right\|_\infty \leq \epsilon_n) \geq \exp(-n\epsilon_n^2), \tag{3.4}$$

with $\epsilon_n = n^{-\alpha/(2\alpha+1)}(\log n)^{\kappa_1}, \bar{\epsilon}_n = n^{-\alpha/(2\alpha+1)}(\log n)^{\kappa_2}$ for constants $\kappa_1, \kappa_2 > 0$.

There is a deep connection between the above measure theoretic result involving the concentration probability and complexity of the support of the conditional Gaussian process W^A and rates of posterior contraction with Gaussian process priors. van der Vaart and van Zanten (2008a) mention that the conditions (3.2) - (3.4) have a one-to-one correspondence with the general sufficient conditions for rates of posterior contraction (Theorem 2.1 of Ghosal, Ghosh and van der Vaart (2000)). In a specific statistical setting involving Gaussian process priors on some function, sieves in the parameter space of interest can be easily obtained by restricting the unknown function to such sets B_n. It only remains to appropriately relate the norm of discrepancy specific to the problem (e.g., Hellinger norm for density estimation) to the Banach space norm (sup-norm in this case) of the Gaussian random element to conclude that $\max\{\epsilon_n, \bar{\epsilon}_n\}$ is the rate of posterior contraction; refer to the discussion following Theorem 3.1 in van der Vaart and van Zanten (2009).

Based on (3.2) - (3.4), van der Vaart and van Zanten (2009) showed that using a rescaled Gaussian random field with a Gamma prior on the (transformed) scaling parameter A^d, one obtains the minimax-optimal rate of convergence $n^{-\alpha/(2\alpha+d)}$ (up to a logarithmic factor) for α-smooth functions. Since their prior specification does not involve the unknown smoothness α, the procedure is fully adaptive.

In this article, we shall consider two function classes defined in Section 2, (i) Hölder class of functions $C^\alpha[0,1]^I$ with isotropic smoothness that can possibly depend on fewer dimensions ($\alpha > 0$ and $I \subset \{1, \ldots, d\}$), and (ii) Hölder class of functions $C^{\boldsymbol{\alpha}}[0,1]^d$ with anisotropic smoothness ($\boldsymbol{\alpha} \in \mathbb{R}_+^d$).

For a continuous function in the support of a Gaussian process, the probability assigned to an ϵ sup-norm neighborhood of the function is controlled by the centered small ball probability and how well the function can be approximated from the RKHS of the process (Section 5 of van der Vaart and van Zanten (2008b)). With the target class of functions as in (i) or (ii), a single scaling seems inadequate and it is intuitively appealing to introduce multiple bandwidth parameters to enlarge the RKHS and facilitate

improved approximation from the RKHS. We aim to find sets B_n so that (3.2) - (3.4) are satisfied with w_0 in the above function classes and ϵ_n being the optimal rate of convergence for the same. Once we establish existence of such sets B_n, one can utilize that result in various statistical settings such as density estimation, mean regression, etc to obtain adaptive estimation. To that end, we first discuss rescaling schemes specific to cases (i) and (ii) in Section 3.1 and 3.2 respectively. Connections between these two schemes are discussed and a unified framework is prescribed for the function class $\left\{C^{\boldsymbol{\alpha}}[0,1]^I : \boldsymbol{\alpha} \in \mathbb{R}^d_+, I \subset \{1,\ldots,d\}\right\}$ combining (i) and (ii). Applications to specific statistical problems are described in Section 3.3.

3.1. *Anisotropic function estimation.* Let $\mathbf{A} = (A_1, \ldots A_d)^{\mathrm{T}}$ be a random vector in $\mathbb{R}^d$ with each A_j a non-negative random variable stochastically independent of W. We can then define a scaled process $W^{\mathbf{A}} = \{W_{\mathbf{A}\cdot t} : t \in [0,1]^d\}$, to be interpreted as a Borel measurable map in $C[0,1]^d$ equipped with the sup-norm $\|\cdot\|_\infty$. The basic idea here is to stretch or shrink the different dimensions by different amounts so that the resulting process becomes suitable for approximating functions having differential smoothness along the different coordinate axes.

We shall define a joint distribution on $\mathbf{A}$ induced through the following hierarchical specification. Let $\Theta = (\Theta_1, \ldots, \Theta_d)$ denote a random vector with a density supported on the simplex $\mathcal{S}_{d-1}$. Given $\Theta = \theta$, we let the elements of $\mathbf{A}$ be conditionally independent, with $A_j^{1/\theta_j} \sim g$, where g is a density on the positive real line satisfying,

$$C_1 x^p \exp(-D_1 x \log^q x) \leq g(x) \leq C_2 x^p \exp(-D_2 x \log^q x),$$

for positive constants C_1, C_2, D_1, D_2 and every sufficiently large $x > 0$.

In particular, the conditions in the above display are satisfied with $q = 0$ if A_j^{1/θ_j} follows a gamma distribution. For notational simplicity, we shall assume g to be a gamma density from now on, noting that the main results would all hold for the general form of g above.

Let $\pi_{\mathbf{A}}$ denote the induced joint prior on $\mathbf{A}$, so that $\pi_{\mathbf{A}}(\mathbf{a}) = \int \prod_{j=1}^d \pi(a_j \mid \theta_j) d\pi(\theta)$. We now state our main theorem for the rescaled process $W^{\mathbf{A}}$.

THEOREM 3.1. *Let W be centered homogeneous Gaussian random field on $\mathbb{R}^d$ with spectral measure ν that satisfies (3.1) and has a Lebesgue density f such that for any $\mathbf{a} \leq \mathbf{b}$, $f(\lambda./\mathbf{a}) \leq f(\lambda./\mathbf{b})$ for all λ. Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$ be a vector of positive numbers and $\alpha_0 = (\sum_{i=1}^d \alpha_i^{-1})^{-1}$. Suppose w_0 belongs to the anisotropic Hölder space $C^{\boldsymbol{\alpha}}[0,1]^d$. Then for every constant $C > 1$, there exist Borel measurable subsets B_n of $C[0,1]^d$ and a constant $D > 0$*

such that, for every sufficiently large n, the conditions (3.2) - (3.4) are satisfied with $\epsilon_n = n^{-\alpha_0/(2\alpha_0+1)}(\log n)^{\kappa_1}, \bar{\epsilon}_n = n^{-\alpha_0/(2\alpha_0+1)}(\log n)^{\kappa_2}$ for constants $\kappa_1, \kappa_2 > 0$.

3.2. *Adaptive dimension reduction.* Consider a joint prior on $\mathbf{A}$ induced through the following hierarchical scheme: (i) draw $\tilde{d}$ according to some prior distribution (with full support) on $\{1, \ldots, d\}$, (ii) given $\tilde{d}$, draw a subset S of size $\tilde{d}$ from $\{1, \ldots, d\}$ following some prior distribution assigning positive prior probability to all $\binom{d}{\tilde{d}}$ subsets of size $\tilde{d}$, (iii) generate a pair of random variables (a, b) with $a^{\tilde{d}} \sim$ gamma and b drawn from a compactly supported density, and finally, (iv) let $A_j = a$ for $j \in S$ and $A_j = b$ for $j \notin S$.

THEOREM 3.2. *Let W be centered homogeneous Gaussian random field on $\mathbb{R}^d$ with spectral measure ν that satisfies (3.1) and has a Lebesgue density f such that for any $\mathbf{a} \leq \mathbf{b}$, $f(\lambda./\mathbf{a}) \leq f(\lambda./\mathbf{b})$ for all λ. Suppose w_0 belongs to the Hölder space $C^{\alpha}[0,1]^I$ for some subset I of $\{1, \ldots, d\}$. Then for every constant $C > 1$, there exist Borel measurable subsets B_n of $C[0,1]^d$ and a constant $D > 0$ such that, for every sufficiently large n, the conditions (3.2) - (3.4) are satisfied with $\epsilon_n = n^{-\alpha/(2\alpha+d_0)}(\log n)^{\kappa_1}, \bar{\epsilon}_n = n^{-\alpha/(2\alpha+d_0)}(\log n)^{\kappa_2}$ for constants $\kappa_1, \kappa_2 > 0$ and $d_0 = |I|$.*

The joint distributions on $\mathbf{A}$ specified in Section 3.1 and 3.2 are closely connected. To begin with, note that if we set $A_j = A$ and $\theta_j = 1/d$ for all j, one obtains a gamma prior on A^d which was previously suggested by van der Vaart and van Zanten (2009). In the general anisotropic case, the joint distribution can be motivated as follows. Recall that the purpose of rescaling is to traverse the sample paths of an infinite smooth stochastic process on a larger domain to make it more suitable for less smooth functions. If the true function has anisotropic smoothness, then we would like to stretch those directions more where the function is less smooth. Now note that for smaller values of θ_j, the marginal distribution of a_j has lighter tails compared to larger values of θ_j. We would thus like θ_j to assume smaller values for the directions j where the function is more smooth and larger values corresponding to the less smooth directions. Moreover, θ_j assuming larger or smaller values only depends on the relative smoothness for a particular dimension. For example, if we scaled the d smoothness levels by a constant factor, intuitively the same set of θ_j's should work. This motivates us to constrain θ to the simplex which serves as a weak identifiability condition.

In the limit as $\theta_j \to 0$, the distribution of a_j converges to a point mass at zero. Accordingly, if the true function doesn't depend on a set of $(d - d^*)$

dimensions, we would set $\theta_j = 0$ for those dimensions and choose the remaining θ_j's from a $d^* - 1$ dimensional simplex. In particular, if the function has isotropic smoothness in the remaining d^* coordinates, one can simple choose $\theta_j = 1/d^*$ for those dimensions. This explains our choice of letting a^{d^*} follow a gamma distribution in Section .

Based on the above discussion, we combine the results in Section 3.1 and 3.2 to prescribe a unified framework for adaptively estimating functions which possibly depend on fewer coordinates and have anisotropic smoothness in the remaining ones, i.e., functions in $C^{\boldsymbol{\alpha}}[0,1]^I$ for $\boldsymbol{\alpha} \in \mathbb{R}_+^d$ and $I \subset \{1, \ldots, d\}$.

3.3. *Rates of convergence in specific settings.* The above two theorems are in the same spirit as Theorem 3.1 of van der Vaart and van Zanten (2009) and Theorem 2.2 of De Jonge and van Zanten (2010) and can be used to derive rates of posterior contraction in a variety of statistical problems involving Gaussian random fields. We shall consider a couple of specific problems with the message that similar results can be obtained for a large class of problems involving rescaled Gaussian random fields.

We first consider a regression problem where given independent response variables y_i and covariates $x_i \in [0,1]^d$, the response is modeled as random perturbations around a smooth regression surface, i.e., $y_i = \mu(x_i) + \epsilon_i$. We assume $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ with a prior on σ supported on some interval $[a, b] \subset [0, \infty)$.

As motivated before, the regression surface might depend only on a subset of variables in $[0,1]^d$ and have anisotropic smoothness in the remaining variables. It is thus appealing to place a Gaussian process prior with dimension specific rescalings on μ as follows. Let W denote a Gaussian process with squared exponential covariance kernel $c(t) = \exp(-\|t\|^2)$ and $\mathbf{A} = (A_1, \ldots, A_d)^{\mathrm{T}}$ be a vector of positive random variables stochastically independent of W. We use the conditionally Gaussian process $W^{\mathbf{A}} = \{W_{\mathbf{A} \cdot t} : t \in [0,1]^d\}$ as a prior for μ, with a joint prior on $\mathbf{A}$ induced through the following hierarchical specification: (i) draw $\tilde{d}$ uniformly on $\{1, \ldots, d\}$, (ii) given $\tilde{d}$, draw a subset $S = \{i_1, \ldots, i_{\tilde{d}}\}$ of size $\tilde{d}$ uniformly from $\{1, \ldots, d\}$, (iii) draw $\theta = (\theta_1, \ldots, \theta_{\tilde{d}})$ from the $\tilde{d} - 1$-dimensional simplex $\mathcal{S}_{\tilde{d}-1}$, (iv) let $A_j^{1/\theta_j} \sim$ gamma for $j \in S$, and set the remaining A_j's to zero.

We denote the posterior distribution by $\Pi(\cdot \mid y_1, \ldots, y_n)$. Let $\|\mu\|_n^2 = n^{-1} \sum_{i=1}^n \mu^2(x_i)$ denote the L_2 norm corresponding to the empirical distribution of the design points. Let the true value σ_0 of σ be contained in the interval $[a, b]$. The posterior is said to contract a rate ϵ_n, if for every

sufficiently large M,

$$\mathrm{E}_{\mu_0,\sigma_0}\Pi\big[(\mu,\sigma) : \|\mu-\mu_0\|_n + |\sigma-\sigma_0| > M\epsilon_n \mid y_1,\ldots,y_n\big] \to 0.$$

THEOREM 3.3. *Let $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_d)$ be a vector of positive numbers and I be a subset of $\{1,\ldots,d\}$. If $w_0 \in C^{\boldsymbol{\alpha}}[0,1]^I$, then the posterior contracts at the rate $\epsilon_n = n^{-\alpha_{0I}/(2\alpha_{0I}+1)}\log^{\kappa} n$, where $\alpha_{0I}^{-1} = \sum_{j\in I}\alpha_j^{-1}$.*

Thus, one obtains the minimax optimal rate up to a log factor adapting to the unknown dimensionally and anisotropic smoothness.

A similar result holds for density estimation using the Logistic Gaussian process.

4. Properties of the rescaled process. We now summarize some properties of the RKHS of the scaled process W^a for a fixed vector of scales a, which shall be crucially used to prove our main theorems.The first five lemmas generalize the results in section 4 of van der Vaart and van Zanten (2009) from a single scaling to a vector of scales. A key idea in van der Vaart and van Zanten (2009) to construct the sets B_n was to exploit a containment relation among the unit balls of the RKHS with different amounts of scaling. Such a result sufficed in the single rescaling framework as the set of positive reals in totally ordered. However, the result can only be generalized with respect to the partial order on $\mathbb{R}_+^d$ which is not sufficient for our purpose. We develop a technique to circumvent this curse of dimensionality by precisely calculating the metric entropy of a collection of unit RKHS balls.

Assume that the spectral measure ν of W has a spectral density f. Then, for $\mathbf{a} \in \mathbb{R}_+^d$, the rescaled proces $W^{\mathbf{a}}$ has a spectral measure $\nu_{\mathbf{a}}$ given by $\nu_{\mathbf{a}}(B) = \nu(B./\mathbf{a})$. Further, $\nu_{\mathbf{a}}$ admits a spectral density $f_{\mathbf{a}}$, with $f_{\mathbf{a}}(\lambda) = \mathbf{a}^{-1}f(\lambda./\mathbf{a})$. For $w_0 \in C[0,1]^d$, define $\phi_{w_0}^{\mathbf{a}}(\epsilon)$ to be the concentration function of the rescaled Gaussian process $W^{\mathbf{a}}$. As an easy consequence of Lemma 4.1 and 4.2 in van der Vaart and van Zanten (2009), it turns out that the RKHS of the process $W^{\mathbf{a}}$ can be characterized as below.

LEMMA 4.1. *The RKHS $\mathbb{H}^{\mathbf{a}}$ of the process $\{W_t^{\mathbf{a}} : t \in [0,1]^d\}$ consists of real parts of the functions*

$$t \to \int e^{i(\lambda,t)}g(\lambda)\nu_{\mathbf{a}}(d\lambda),$$

where g runs over the complex Hilbert space $L_2(\nu_{\mathbf{a}})$. Further, the RKHS norm of the element in the above display is given by $\|g\|_{L_2(\nu_{\mathbf{a}})}$.

Lemma 4.3 of van der Vaart and van Zanten (2009) shows that for any isotropic Hölder smooth function w, convolutions with an appropriately chosen class of higher order kernels indexed by the scaling parameter a belong to the RKHS. This suggests that driving the bandwidth $1/a$ to zero, one can obtain improved approximations to any Hölder smooth function. The following Lemma 4.2 illustrates the usefulness of using separate bandwidths for each dimensions for approximating anisotropic Hölder functions from the RKHS.

LEMMA 4.2. *Assume ν has a density with respect to the Lebesgue measure which is bounded away from zero on a neighborhood of the origin. Let $\boldsymbol{\alpha} \in \mathbb{R}_+^d$ be given. Then, for any subset I of $\{1, \ldots, d\}$ and $w \in C^{\boldsymbol{\alpha}}[0,1]^I$, there exists constants C and D depending only on ν and w such that, for $\mathbf{a}$ large enough,*

$$\inf\{\|h\|_{\mathbb{H}\mathbf{a}}^2 : \|h - w\|_\infty \leq C \sum_{i \in I} a_i^{-\alpha_i}\} \leq D\mathbf{a}^*$$

PROOF. We shall prove the result for $w \in C^{\boldsymbol{\alpha}}[0,1]^d$ and sketch an argument for extending the proof to any $w \in C^{\boldsymbol{\alpha}}[0,1]^I$.

Let $\psi_j, j = 1, \ldots, d$, be a set of higher order kernel as in the proof of Lemma 4.3 of van der Vaart and van Zanten (2009), which satisfy $\int \psi_j(t_j)dt_j = 1$, $\int t_j^k \psi_j(t_j)dt_j = 0$ for any positive integer k and $\int |t_j|^{\alpha_j}|\psi_j(t_j)|dt_j \leq 1$. Define $\psi : \mathbb{R}^d \to \mathbb{C}$ by $\psi(t) = \psi_1(t_1)\ldots\psi_d(t_d)$ so that one has $\int \psi(t)dt = 1$, $\int t^k \psi(t)dt = 0$ for any non-zero multi-index $k = (k_1, \ldots, k_d)$, and the functions $\left|\hat{\psi}\right|/f$ and $\left|\hat{\psi}\right|^2/f$ are uniformly bounded, where $\hat{\psi}$ denotes the Fourier transform of ψ.

For a vector of positive numbers $\mathbf{a} = (a_1, \ldots, a_d)$, let $\psi_{\mathbf{a}}(t) = \mathbf{a}^* \psi(\mathbf{a} \cdot t)$, where $\mathbf{a}^* = \prod_{j=1}^d a_j$. By Whitney's theorem, w can be extended to a function $w : \mathbb{R}^d \to \mathbb{R}$ with compact support and $\|w\|_{\boldsymbol{\alpha}} < \infty$. Working with this extension, we shall first show that the convolution $\psi_{\mathbf{a}} * w$ is contained in the RKHS $\mathbb{H}^{\mathbf{a}}$. To that end, note that,

$$\frac{1}{(2\pi)^d}(\psi_{\mathbf{a}} * w)(t) = \int e^{-i(t,\lambda)} \hat{w}(\lambda)\hat{\psi}_{\mathbf{a}}(\lambda)d\lambda = \int e^{i(t,\lambda)} \frac{\hat{w}(-\lambda)\hat{\psi}_{\mathbf{a}}(\lambda)}{f_{\mathbf{a}}(\lambda)} \nu_{\mathbf{a}}(d\lambda).$$

Thus, following Lemma 4.1, we need to show that $\hat{w}(-\lambda)\hat{\psi}_{\mathbf{a}}(\lambda)/f_{\mathbf{a}}(\lambda) \in L_2(\nu_{\mathbf{a}})$ to conclude that $\psi_{\mathbf{a}} * w$ belongs to $\mathbb{H}^{\mathbf{a}}$. Since $\hat{\psi}_{\mathbf{a}}(\lambda) = \hat{\psi}(\lambda./\mathbf{a})$, one has

$$\int \left| \frac{\hat{w}(-\lambda)\hat{\psi}_{\mathbf{a}}(\lambda)}{f_{\mathbf{a}}(\lambda)} \right|^2 \nu_{\mathbf{a}}(d\lambda) \leq \mathbf{a}^* \left\| \frac{|\hat{\psi}|^2}{f} \right\|_\infty \int |\hat{w}(\lambda)|^2 \, d\lambda.$$

The above assertion is thus proved by noting that $|\hat{\psi}|^2/f$ is uniformly bounded by construction and $(2\pi)^d \int |\hat{w}^2(\lambda)d\lambda = \int |w(t)|^2 dt < \infty$. Also, the squared RKHS norm of $\psi_{\mathbf{a}} * w$ is bounded by $D\mathbf{a}^*$, with D depending only on ν and w. Thus, the proof of Lemma 4.2 would be completed if we can show that $\|\psi_{\mathbf{a}} * w - w\|_\infty \leq C \sum_{j=1}^d a_j^{-\alpha_j}$.

We have, for any $t \in \mathbb{R}^d$,

$$\psi_{\mathbf{a}} * w(t) - w(t) = \int \psi(s)\{w(t - s./\mathbf{a}) - w(t)\}ds.$$

For $1 \leq j \leq d-1$, let $u^{(j)}$ denote the vector in $\mathbb{R}^d$ with $u_i^{(j)} = 0$ for $i = 1, \ldots, j$ and $u_i^{(j)} = 1$ for $i = j+1, \ldots, d$. For any two vectors $x, y \in \mathbb{R}^d$, we can navigate from x to y in a piecewise linear fashion traveling parallel to one of the coordinate axes at a time. The vertices of the path will be given by $x^{(0} = x$, $x^{(j)} = u^{(j)} \cdot x + (1 - u^{(j)}) \cdot y$ for $j = 1, \ldots, d-1$ and $x^{(d)} = y$.

A multivariate Taylor expansion of $w(t - s./a)$ around $w(t)$ cannot take advantage of the anisotropic smoothness of w across different coordinate axes. Letting $x = t, y = t - s./a$ and $x^{(j)}, j = 0, 1, \ldots, d$ as above, let us write $w(y) - w(x)$ in the following telescoping form,

$$w(y) - w(x) = \sum_{j=1}^d w(x^{(j)}) - w(x^{(j-1)}) = \sum_{j=1}^d w_j(t_j - s_j/a_j \mid x^{(j)}) - w_j(t_j \mid x^{(j)}),$$

where the functions w_j are as defined in Section 2, with $w_j(t \mid x) = w(x_1, \ldots, x_{j-1}, t, x_{j+1}, \ldots, x_d)$ for any $t \in \mathbb{R}$ and $x \in \mathbb{R}^d$.

Thus,

$$w(t - s./a) - w(t) = \sum_{j=1}^d \left[\sum_{i=1}^{\lfloor \alpha_j \rfloor} D^i w_j(t_j \mid x^{(j)}) \frac{(-s_j/a_j)^i}{i!} + S_j(t_j, -s_j/a_j) \right],$$

where $|S_j(t_j, -s_j/a_j)| \leq K s_j^{\alpha_j} a_j^{-\alpha_j}$ by (2.2), for a constant K depending on ν and w but not on t and s. Combining the above, we have

$$\left| \int \psi(s)\{w(t - s./\mathbf{a}) - w(t)\} \right| = \left| \sum_{j=1}^d \int S_j(t_j, -s_j/a_j) dt_j \right| \leq C \sum_{j=1}^d a_j^{-\alpha_j}.$$

If, $w \in C^{\boldsymbol{\alpha}}[0,1]^I$ for some subset I of $\{1, \ldots, d\}$, then we only need to scale the co-ordinates in I.

□

We next study the metric entropy of the unit ball of the RKHS and the centered small ball probability of the resealed process. Let $\mathbb{H}_1^{\mathbf{a}}$ denote the unit ball in the RKHS of $W^{\mathbf{a}}$.

LEMMA 4.3. *There exists a constant K, depending only on ν and d, such that, for $\epsilon < 1/2$,*

$$\log N(\epsilon, \mathbb{H}_1^{\mathbf{a}}, \|\cdot\|_\infty) \leq K\mathbf{a}^* \left(\log \frac{1}{\epsilon}\right)^{d+1}.$$

PROOF. By Lemma 4.1, an element of $\mathbb{H}_1^{\mathbf{a}}$ can be written as the real part of the function $h : [0,1]^d \to \mathbb{C}$ given by

$$h(t) = \int e^{i(\lambda,t)} g(\lambda) \nu_{\mathbf{a}}(d\lambda) \tag{4.1}$$

for $g : \mathbb{R}^d \to \mathbb{C}$ a function with $\int |g|^2 \nu_{\mathbf{a}}(d\lambda) \leq 1$.

Viewing h as a function of it, we would like to exploit the sub-exponential tails of ν as in (3.1) to extend h analytically over a larger domain in $\mathbb{C}^d$. For $z \in \mathbf{C}^d$, we shall continue to denote the function $z \mapsto \int e^{(\lambda,z)} \psi(\lambda) \nu_a(d\lambda)$ by h. Using the Cauchy-Schwartz inequality and the change of variable theorem,

$$|h(z)|^2 \leq \int e^{(\lambda, 2\mathbf{a}.\mathrm{Re}(z))} \nu(d\lambda). \tag{4.2}$$

Thus, using the dominated convergence theorem, h (as a function of it) can be extended to an analytic function $z \mapsto \int e^{(\lambda,z)} \psi(\lambda) \nu_a(d\lambda)$ on the strip $\Omega = \{z \in \mathbb{C}^d : |\mathrm{Re} z_j| \leq R_j, j = 1, \ldots, d\}$ with $R_j = \delta/(6a_j\sqrt{d})$. Also, for every $z \in \Omega$, h satisfies the uniform bound $|h(z)|^2 \leq \int e^{\delta\|\lambda\|} \nu(d\lambda) = C^2$.

The analytic extension of h to a strip containing the product of the imaginary axes allows us to precisely estimate the error term of a k-order Taylor expansion of $h(t)$. For $t \in [0,1]^d$, Let $C_1, \ldots, C_d$ denote circles of radius $R_1, \ldots, R_d$ in the complex plane around the coordinates $it_1, \ldots, it_d$ of it respectively. Using the Cauchy integral formula,

$$\left|\frac{D^n h(t)}{\boldsymbol{n}!}\right| = \left|\frac{1}{(2\pi i)^d} \oint_{C_1} \cdots \oint_{C_d} \frac{h(z)}{(z-t)^{\boldsymbol{n}+1}} dz_1 \cdots dz_d\right| \leq \frac{C}{R^{\boldsymbol{n}.}},$$

where $D^{\boldsymbol{n}}$ denotes the partial derivative of order $\boldsymbol{n} = (n_1, \ldots, n_d)$. This suggests using a net of piecewise polynomials for approximating the elements of $\mathbb{H}_1^{\mathbf{a}}$. One can discretize the coefficients and centers of the piecewise polynomials to obtain a finite set of functions that approximate the leading terms

of a Taylor expansion of a function in $\mathbb{H}_1^{\mathbf{a}}$ and the remainder terms can be controlled using the bound in the above display.

To elaborate, let $R = (R_1, \ldots, R_d)^{\mathrm{T}}$. Partition $T = [0,1]^d$ into rectangles $\Gamma_1, \ldots, \Gamma_m$ with centers $\{t_1, t_2, \ldots, t_m\}$ such that given any $z \in T$, there exists Γ_j with center $t_j = (t_{j1}, \ldots, t_{jd})^{\mathrm{T}}$ with $|z_i - t_{ji}| \leq R_i/4, i = 1, \ldots, d$. Consider the piecewise polynomials $P = \sum_{j=1}^m P_{j,\gamma_j} 1_{\Gamma_j}$ with

$$P_{j,\gamma_j}(t) = \sum_{\boldsymbol{n}.\leq k} \gamma_{j,\boldsymbol{n}} (t - t_j)^{\boldsymbol{n}}.$$

We obtain a finite set of functions $\mathcal{P}_{\mathbf{a}}$ by discretizing the coefficients $\gamma_{j,\boldsymbol{n}}$ for each j and $\boldsymbol{n}$ over a grid of mesh width $\epsilon/R^{\boldsymbol{n}}$ in the interval $[-C/R^{\boldsymbol{n}}, C/R^{\boldsymbol{n}}]$, with C defined as above. As in van der Vaart and van Zanten (2009), the log cardinality of the set is bounded above by

$$\log \left(\prod_j \prod_{\boldsymbol{n}:\boldsymbol{n}.\leq k} \#\gamma_{j,n} \right) \leq m k^d \log \left(\frac{2C}{\epsilon} \right). \tag{4.3}$$

We can choose $m \precsim 1/R^*$. The proof is complete if we how that the resulting set of functions is a $K\epsilon$-net for constants C and K depending on ν and $k \precsim \log(1/\epsilon)$. The rest of the proof follows exactly as in the proof for Lemma 4.5 in van der Vaart and van Zanten (2009) by showing that

$$\left| \sum_{n.>k} \frac{D^n h_\psi(t_i)}{n!} (z - t_i)^n \right| \leq \sum_{n.>k} \frac{C}{R^n} (R/2)^n \leq KC \left(\frac{2}{3} \right)^k. \tag{4.4}$$

and

$$\left| \sum_{n.\leq k} \frac{D^n h_\psi(t_i)}{n!} (z - t_i)^n - P_{i,\gamma_i}(z) \right| \leq K\epsilon. \tag{4.5}$$

The proof is completed by choosing k large enough such that $(2/3)^k \leq K\epsilon$. □

LEMMA 4.4. *For any a_0 positive, there exists constants C and $\epsilon_0 > 0$ such that for $a \geq a_0$ and $\epsilon < \epsilon_0$,*

$$-\log P\big(\|W^a\|_\infty \leq \epsilon \big) \leq C a^* \left(\log \frac{\bar{a}}{\epsilon} \right)^{d+1}.$$

PROOF. This follows from Theorem 2 in Kuelbs and Li (1993) and Lemma 4.6 in van der Vaart and van Zanten (2009). Proceeding as in Lemma 4.6 in van der Vaart and van Zanten (2009) and Lemma 4.3, we obtain

$$\phi^a(\epsilon) + \log 0.5 \leq K_1 a^* \left(\log \frac{\phi_0^a(\epsilon)}{\epsilon} \right)^{1+d}. \tag{4.6}$$

for some constant $K_1 > 0$. Note that with $L = [0, a_1] \times \cdots \times [0, a_d]$,

$$\phi_0^a(\epsilon) = -\log P(\|W^a\|_\infty \leq \epsilon) = -\log P(\sup_{t \in L} |W| \leq \epsilon) \tag{4.7}$$

$$\leq -\log P(\sup_{t \in [0, \bar{a}]} |W| \leq \epsilon) \tag{4.8}$$

$$\leq K_2 \left(\frac{\bar{a}}{\epsilon} \right)^\tau \tag{4.9}$$

for some constant K_2 and $\tau > 0$, where the last inequality follows from the proof of Lemma 4.6 in van der Vaart and van Zanten (2009). Inserting this bound in (4.6), we obtain

$$-\log \mathrm{P}\big(\|W^a\|_\infty \leq \epsilon \big) \leq C a^* \left(\log \frac{\bar{a}}{\epsilon} \right)^{d+1}. \tag{4.10}$$

for some constant $C > 0$. □

We next state a nesting property of the unit ball $\mathbb{H}_1^{\mathbf{a}}$ of the RKHS of $W^{\mathbf{a}}$ for different values of $\mathbf{a}$, generalizing Lemma 4.7 of van der Vaart and van Zanten (2009).

LEMMA 4.5. *Assume the spectral measure ν satisfies (3.1) and has a density f with respect to the Lebesgue measure on $\mathbb{R}^d$ which satisfies $f(t./\mathbf{a}) \leq f(t./\mathbf{b})$ for any $\mathbf{a} \leq \mathbf{b}$. Then,*

$$\sqrt{a_1 \ldots a_d}\ \mathbb{H}_1^{\mathbf{a}} \subset \sqrt{b_1 \ldots b_d}\ \mathbb{H}_1^{\mathbf{b}}.$$

PROOF. Let $h \in \mathbb{H}_1^{\mathbf{a}}$. Following Lemma 4.1, $h(t) = \int e^{i\langle \lambda, t \rangle} \psi(\lambda) \nu_{\mathbf{a}}(d\lambda)$. Since $\|h\|_{\mathbb{H}^{\mathbf{a}}}^2 = \|\psi\|_{L_2(\nu_{\mathbf{a}})}^2$, it follows that $\int |\psi(\lambda)|^2 f_a(\lambda) d\lambda \leq 1$. Now, $h(t) = \int e^{i\langle \lambda, t \rangle} \{\psi(\lambda) f_{\mathbf{a}}(\lambda)/f_{\mathbf{b}}(\lambda)\} \nu_{\mathbf{b}}(d\lambda)$. The conclusion follows since,

$$\|h\|_{\mathbb{H}^{\mathbf{b}}}^2 = \int |\psi(\lambda)|^2 \left\{ \frac{f_{\mathbf{a}}(\lambda)}{f_{\mathbf{b}}(\lambda)} \right\}^2 \nu_{\mathbf{b}}(d\lambda) \leq \left\| \frac{f_{\mathbf{a}}(\lambda)}{f_{\mathbf{b}}(\lambda)} \right\|_\infty \int |\psi(\lambda)|^2 \nu_{\mathbf{a}}(d\lambda) \leq \frac{\mathbf{a}^*}{\mathbf{b}^*},$$

using the fact that $f_{\mathbf{a}}(\lambda)/f_{\mathbf{b}}(\lambda) = (\mathbf{b}^*/\mathbf{a}^*) f(\lambda./\mathbf{a})/f(\lambda./\mathbf{b}) \leq (\mathbf{b}^*/\mathbf{a}^*)$ by assumption. □

The next two Lemmas are crucially used in our treatment of the multi-bandwidth case.

LEMMA 4.6. *For a positive number $r > 0$ and $\theta \in \mathcal{S}_{d-1}$, let $\mathbb{H}_1^{r,\theta}$ denote the unit ball of the RKHS of $W^{\mathbf{a}}$ with $a_j = r^{\theta_j}$ for $1 \leq j \leq d$. Then, there exists a constant K, depending only on ν and d, such that, for $\epsilon < 1/2$,*

$$\log N\left(\epsilon, \cup_{\theta \in \mathcal{S}_{d-1}} \mathbb{H}_1^{r,\theta}, \|\cdot\|_\infty\right) \leq Kr\left(\log \frac{1}{\epsilon}\right)^{d+1}.$$

PROOF. Let $Q = \{\mathbf{a} \in \mathbb{R}_+^d : \ a_j \geq 1 \forall j = 1, \ldots, d, \ \mathbf{a}^* = r\}$. For $\mathbf{a}, \mathbf{b} \in Q$, the idea of the proof is to show that the piecewise polynomials $\mathcal{P}_{\mathbf{a}}$ that form a $K\epsilon$-net for $\mathbb{H}_1^{\mathbf{a}}$ in the proof of Lemma 4.3 are also a $K\epsilon$-net for $\mathbb{H}_1^{\mathbf{b}}$ if $\mathbf{b}$ is "close enough" to $\mathbf{a}$.

Fix $\mathbf{a} \in Q$. Let $\Omega = \{z \in \mathbb{C}^d : |\text{Re} z_j| \leq R_j, j = 1, \ldots, d\}$ with $R_j = \delta/(6a_j\sqrt{d})$ denote the strip in $\mathbb{C}^d$. From the proof of Lemma 4.3, any function $h \in \mathbb{H}_1^{\mathbf{a}}$ has an analytic extension to Ω. Let $\mathbf{b} \in Q$ satisfy $\max_j |a_j - b_j| \leq 1$. We shall show that any $h \in \mathbb{H}_1^{\mathbf{b}}$ can also be extended analytically to the same strip Ω.

From (4.2), it follows that $h \in \mathbb{H}_1^{\mathbf{b}}$ can be analytically extended to $\{z \in \mathbb{C}^d : |(\lambda, 2\mathbf{b} \cdot \text{Re}(z))| \leq \delta \|\lambda\|\}$. Now,

$$\begin{aligned} |(\lambda, 2\mathbf{b} \cdot \text{Re}(z))| &\leq |(\lambda, 2\mathbf{a} \cdot \text{Re}(z))| + |(\lambda, 2(\mathbf{b} - \mathbf{a}) \cdot \text{Re}(z))| \\ &\leq 2\,|(\lambda, 2\mathbf{a} \cdot \text{Re}(z))| \leq 2\delta \|\lambda\| /3, \end{aligned}$$

where the last inequality follows from Cauchy-Schwartz and the one before uses $|b_j - a_j| \leq 1 \leq a_j$ for all $j = 1, \ldots, d$.

Clearly, the same tail estimate as in (4.4) works for any $h \in \mathbb{H}_1^{\mathbf{b}}$. From (4.5), it thus follows that the set of functions $\mathcal{P}_{\mathbf{a}}$ form a $K\epsilon$-net for $\mathbb{H}_1^{\mathbf{b}}$. Let $\mathcal{A}$ be a set of points in Q such that for any $\mathbf{b} \in Q$, there exists $\mathbf{a} \in \mathcal{A}$ such that $\max_j |a_j - b_j| \leq 1$. One can clearly find an $\mathcal{A}$ with $|\mathcal{A}| \leq r^d$. The proof is completed by observing that $\cup_{\mathbf{a} \in \mathcal{A}} \mathcal{P}_{\mathbf{a}}$ form a $K\epsilon$ net for $\cup_{\theta \in \mathcal{S}_{d-1}} \mathbb{H}_1^{r,\theta}$. □

Another lemma on entropy of collection of unit RKHS balls.

LEMMA 4.7. *For $r \in \mathbb{R}_+^d$, let $\mathcal{C}_r$ denote the subset of $\mathbb{R}_+^d$ consisting of all vectors $\mathbf{a} \leq r$, i.e., $a_j \leq r_j$ for all $j = 1, \ldots, d$. Then,*

$$\log N\left(\epsilon, \cup_{\mathbf{a} \in \mathcal{C}_r} \mathbb{H}_1^{\mathbf{a}}, \|\cdot\|_\infty\right) \leq Kr^*\left(\log \frac{1}{\epsilon}\right)^{d+1}.$$

5. Proof of main results.

5.1. *Proof of Theorem 3.1.* Let us begin by observing that,

$$\begin{aligned}\mathrm{P}(\left\|W^{\mathbf{A}}-w_0\right\|_\infty \le 2\epsilon) &= \int \mathrm{P}(\left\|W^{\mathbf{a}}-w_0\right\|_\infty \le 2\epsilon)\pi_{\mathbf{A}}(d\mathbf{a}) \\ &= \int \left\{ \int \mathrm{P}(\left\|W^{\mathbf{a}}-w_0\right\|_\infty \le 2\epsilon)\pi(\mathbf{a}\mid\theta)d\mathbf{a}\right\}\pi(\theta)d\theta.\end{aligned}$$

As in van der Vaart and van Zanten (2009), we first derive bounds on the non-centered small ball probability for a fixed rescaling $\mathbf{a}$, and then integrate over the distribution of $\mathbf{a}$ to derive the same for $W^{\mathbf{A}}$.

Given $\mathbf{a} \in \mathbb{R}_+^d$, recall the definition of the centered and non-centered concentration functions of the process $W^{\mathbf{a}}$,

$$\phi_0^{\mathbf{a}}(\epsilon) = -\log \mathrm{P}(\left\|W^{\mathbf{a}}\right\|_\infty \le \epsilon), \tag{5.1}$$

$$\phi_{w_0}^{\mathbf{a}}(\epsilon) = \inf_{h\in\mathbb{H}_1^{\mathbf{a}}:\left\|h-w_0\right\|_\infty\le\epsilon} \|h\|_{\mathbb{H}^{\mathbf{a}}}^2 - \log \mathrm{P}(\left\|W^{\mathbf{a}}\right\|_\infty \le \epsilon). \tag{5.2}$$

For a fixed $\mathbf{a}$, the non-centered small ball probability of $W^{\mathbf{a}}$ can be bound in terms of the concentration function as follows (van der Vaart and van Zanten, 2008b),

$$\mathrm{P}(\left\|W^{\mathbf{a}}-w_0\right\|_\infty \le 2\epsilon) \ge e^{-\phi_{w_0}^{\mathbf{a}}(\epsilon)}.$$

Now, suppose that $w_0 \in C^{\boldsymbol{\alpha}}[0,1]^d$ for some $\boldsymbol{\alpha} \in \mathbb{R}_+^d$. From Lemma 4.2 and 4.4, it follows that for every $a_0 > 0$, there exist positive constants $\epsilon_0 < 1/2$, C, D and E that depend on w_0 and ν only such that, for $\mathbf{a} > a_0$, $\epsilon < \epsilon_0$ and $C\sum_{i=1}^d a_i^{-\alpha_i} < \epsilon$,

$$\phi_{w_0}^{\mathbf{a}}(\epsilon) \le D\mathbf{a}^* + E\mathbf{a}^*\left(\log\frac{\bar{\mathbf{a}}}{\epsilon}\right)^{1+d} \le K_1\mathbf{a}^*\left(\log\frac{\bar{\mathbf{a}}}{\epsilon}\right)^{1+d},$$

with K_1 depending only on a_0, ν and d. Thus, for $\epsilon < \min\{\epsilon_0, C_1 a_0^{-\bar{\alpha}}\}$, by (5.1),

$$\begin{aligned}\mathrm{P}(\left\|W^{\mathbf{A}}-w_0\right\|_\infty \le 2\epsilon) &\ge \int \left\{ \int e^{-\phi_{w_0}^{\mathbf{a}}(\epsilon)}\pi(\mathbf{a}\mid\theta)d\mathbf{a}\right\}\pi(\theta)d\theta \\ &\ge \int \left\{ \int_{a_1=(C_1/\epsilon)^{1/\alpha_1}}^{2(C_1/\epsilon)^{1/\alpha_1}} \cdots \int_{a_d=(C_1/\epsilon)^{1/\alpha_d}}^{2(C_1/\epsilon)^{1/\alpha_d}} e^{-K_1\mathbf{a}^*\log^{1+d}(\bar{\mathbf{a}}/\epsilon)}\pi(\mathbf{a}\mid\theta)d\mathbf{a}\right\}\pi(\theta)d\theta \\ &\ge C_2 e^{-K_2(1/\epsilon)^{1/\alpha_0}\log^{1+d}(1/\epsilon)} \int \left\{ \int_{a_1=(C_1/\epsilon)^{1/\alpha_1}}^{2(C_1/\epsilon)^{1/\alpha_1}} \cdots \int_{a_d=(C_1/\epsilon)^{1/\alpha_d}}^{2(C_1/\epsilon)^{1/\alpha_d}} \pi(\mathbf{a}\mid\theta)d\mathbf{a}\right\}\pi(\theta)d\theta.\end{aligned}$$

Now,

$$\begin{aligned}
&\int \left\{ \int_{a_1=(C_1/\epsilon)^{1/\alpha_1}}^{2(C_1/\epsilon)^{1/\alpha_1}} \cdots \int_{a_d=(C_1/\epsilon)^{1/\alpha_d}}^{2(C_1/\epsilon)^{1/\alpha_d}} \pi(\mathbf{a} \mid \theta) d\mathbf{a} \right\} \pi(\theta) d\theta \\
&\geq \int \left\{ \int_{a_1=(C_1/\epsilon)^{1/\alpha_1}}^{2(C_1/\epsilon)^{1/\alpha_1}} \cdots \int_{a_d=(C_1/\epsilon)^{1/\alpha_d}}^{2(C_1/\epsilon)^{1/\alpha_d}} e^{-\sum_{j=1}^d a_j^{1/\theta_j}} d\mathbf{a} \right\} \pi(\theta) d\theta \\
&\geq \int e^{-K_3 \sum_{j=1}^d (1/\epsilon)^{1/\alpha_j \theta_j}} \pi(\theta) d\theta.
\end{aligned}$$

Let Γ denote the region in the simplex $\mathcal{S}_{d-1}$ given by $\Gamma = \{\theta \in \mathcal{S}_{d-1} : \tau < \theta_j - \frac{\alpha_0}{\alpha_1} < 2\tau, j = 1, \ldots, d-1\}$. Since $\sum_{j=1}^d \alpha_0/\alpha_j = 1$, we can choose $\tau > 0$ small enough to guarantee that any θ satisfying the set of inequalities lies inside the simplex. Moreover, with $\theta_d = 1 - \sum_{j=1}^{d-1} \theta_j$, one has $(d-1)\tau < \theta_d < 2(d-1)\tau$. Choosing $\tau = C_5/\log(1/\epsilon)$, one can show that $\sum_{j=1}^d (1/\epsilon)^{1/(\alpha_j \theta_j)} \leq C_6 (1/\epsilon)^{1/\alpha_0}$ for any $\theta \in \Gamma$. Hence, the probability in the above display can be bounded below by $e^{-n\epsilon_n^2}$ for $\epsilon_n = n^{-\alpha_0/(2\alpha_0+1)}$.

Let $\mathbb{B}_1$ denote the unit sup-norm ball of $C[0,1]^d$. For a vector $\theta \in \mathcal{S}_{d-1}$ and positive constants M, r, ξ, ϵ, let $B^\theta = B^\theta(M, r, \xi, \epsilon)$ denote the set,

$$B^\theta = \cup_{\mathbf{a} \leq r^\theta} (M \mathbb{H}_1^{r,\theta}) + \epsilon \mathbb{B}_1.$$

Let us first calculate the probability $\mathrm{P}(W^{\mathbf{A}} \notin B^\theta \mid \theta)$. Note that,

$$\begin{aligned}
\mathrm{P}(W^{\mathbf{a}} \notin B^\theta \mid \theta) &= \int \mathrm{P}(W^\theta \notin B^\theta) \pi(\mathbf{a} \mid \theta) d\mathbf{a} \\
&\leq \int_{\mathbf{a} \leq r^\theta} \mathrm{P}(W^{\mathbf{a}} \notin B^\theta) \pi(\mathbf{a} \mid \theta) d\mathbf{a} + \mathrm{P}(\mathbf{A} \not\leq r^\theta \mid \theta),
\end{aligned}$$

where $\mathrm{P}(W^{\mathbf{A}} \not\leq r \mid \theta)$ is a shorthand notation for $\mathrm{P}(\text{at least one } A_j > r^{\theta_j} \mid \theta)$. Since A_j^{1/θ_j} follows a gamma distribution given θ_j, the above probability can be bound above by e^{-r} following Lemma 4.9 of van der Vaart and van Zanten (2009). To tackle the first term in the last display, note that B^θ contains the set $M\mathbb{H}_1^{\mathbf{a}} + \epsilon \mathbb{B}_1$ for any $\mathbf{a} \leq r^\theta$ by definition. Hence, for any $\mathbf{a} \leq r^\theta$, by Borell's inequality,

$$\begin{aligned}
\mathrm{P}(W^{\mathbf{a}} \notin B^\theta) &\leq \mathrm{P}(W^{\mathbf{a}} \notin M\mathbb{H}_1^{\mathbf{a}} + \epsilon \mathbb{B}_1) \\
&\leq 1 - \Phi\left\{ M + \Phi^{-1}\left(e^{-\phi_0^{\mathbf{a}}(\epsilon)} \right) \right\} \\
&\leq 1 - \Phi\left\{ M + \Phi^{-1}\left(e^{-\phi_0^{r^\theta}(\epsilon)} \right) \right\} \leq e^{-\phi_0^{r^\theta}(\epsilon)},
\end{aligned}$$

if $M \geq -2\Phi^{-1}\big(e^{-\phi_0^{r^\theta}(\epsilon)}\big)$, where the penultimate inequality follows from the fact that, with $T = [0,1]^d$,

$$e^{-\phi_0^{\mathbf{a}}(\epsilon)} = \mathrm{P}(\sup_{t\in\mathbf{a}\cdot T}|W_t| \leq \epsilon) \leq \mathrm{P}(\sup_{t\in r^\theta\cdot T}|W_t| \leq \epsilon) = e^{-\phi_0^{r^\theta}(\epsilon)}.$$

By Lemma 4.10 of van der Vaart and van Zanten (2009), $\Phi^{-1}(u) \geq -\{2\log(1/u)\}^{1/2}$ for $u \in (0,1)$. Hence, the last inequality in the above display remains valid if we choose

$$M \geq 4\sqrt{\phi_0^{r^\theta}(\epsilon)}.$$

The proof is completed by letting $B = \cup_{\theta\in\mathcal{S}_{d-1}} B^\theta$. Derivation to be added later.

6. Lower bounds on posterior contraction rates. In this section, we will demonstrate that when the true density is dependent on a smaller number of variables, a Gaussian process prior with a single bandwidth leads to a sub-optimal rate of convergence. To illustrate this, we will focus on the example of density estimation using the Logistic Gaussian process prior. We will show that the posterior contraction rate using a a single bandwidth Logistic Gaussian process with respect to the sup-norm topology is bounded below by $n^{-\alpha/(2\alpha+d)}$ when the true density is

$$f_0(x_1,\ldots,x_d) = Ce^{|x_1-0.5|^{1.5}}, (x_1,\ldots,x_d) \in [0,1]^d. \tag{6.1}$$

This shows the necessity of using an inhomogeneous Gaussian process in high-dimensional density estimation when the true density is actually lower dimensional. Although lower bounds on the posterior contraction rates in Gaussian process settings have been previously addressed by Castillo (2008), the literature is restricted to series expansion priors and the Riemann-Liouville process priors. In this section, we have extended the results to Gaussian process with exponential covariance kernel having a single bandwidth. In particular, we have derived a lower bound to the concentration function around $w_0(x_1,\ldots,x_d) = |x_1 - 0.5|^{1.5}$ using a single inverse-Gamma bandwidth.

For $d^* \in \{1,\ldots,d\}$, let $A^{d^*} \sim g$, where g is the gamma density. Let the prior distribution for $\mathbf{A}$ is specified by assuming the same prior for all the bandwidths across the different dimensions, i.e., $\mathbf{A} \overset{d}{=} (A, A, \ldots, A)$. Recall that the Logistic Gaussian process prior for a density f on $[0,1]^d$ is given by

$$f(x) = \frac{\exp\{W^{\mathbf{A}}x\}}{\int_{[0,1]^d}\exp\{W^{\mathbf{A}}t\}dt}, x \in [0,1]^d. \tag{6.2}$$

We will start with a few auxiliary lemmas which enable us to provide an upper bound to the concentration function of the Gaussian process $W^{\mathbf{A}}$. Denote by $\mathbb{H}^a$ the reproductive kernel Hilbert space of the Gaussian process W^a.

In the following, we define a Gaussian based higher order kernel as in Wand and Schucany (1990). For $r \geq 1$, let Q_{2r-2} be the polynomial given by $Q_{2r-2}(x) = \sum_{i=0}^{r-1} c_{2i} x^{2i}$ where

$$c_{2i} = \frac{(-1)^i 2^{i-2r+1}(2r)!}{r!(2i+1)!(r-i-1)!}. \tag{6.3}$$

Then Q_{2r-2} is the unique polynomial of degree $\leq 2r-2$ for which $G_{2r} \equiv Q_{2r-2}\phi$ is a $2r$ order kernel. It is easy to see (also refer to Wand and Schucany (1990)) that $r = 1$ corresponds to the standard Gaussian kernel and for $r > 1$ and $1 \leq j \leq r-1$,

$$\int x^{2j} G_{2r}(x) = 0. \tag{6.4}$$

Define a higher order kernel with scaling $(a_1, \ldots, a_d)$ in d dimensions using G_{2r} as

$$\psi_a^{2r}(x) = G_{2r}(ax_1) \cdots G_{2r}(ax_d). \tag{6.5}$$

In the following Lemma 6.1, we calculate the Fourier transform of $\psi^{2r}(t)$.

LEMMA 6.1. $\hat{\psi}^{2r}(t) = e^{-\|t\|^2/2} \prod_{j=1}^d \sum_{s=0}^{r-1} \frac{t_j^{2s}}{2^s s!}$

PROOF.

$$\begin{aligned}
\hat{\psi}^{2r}(t) &= \int e^{i\langle \lambda, t\rangle} \psi^{2r}(\lambda) d\lambda \\
&= \int e^{i\langle \lambda, t\rangle} G_{2r}(\lambda_1) \cdots G_{2r}(\lambda_d) d\lambda \\
&= \prod_{j=1}^d \int e^{i(\lambda_j, t_j)} G_{2r}(\lambda_j) d\lambda_j \\
&= \prod_{j=1}^d e^{-t_j^2/2} \sum_{s=0}^{r-1} \frac{t_j^{2s}}{2^s s!} \\
&= e^{-\|t\|^2/2} \prod_{j=1}^d \sum_{s=0}^{r-1} \frac{t_j^{2s}}{2^s s!}
\end{aligned}$$

where the penultimate identity follows from Wand and Schucany (1990). □

Lemma 4.1 of van der Vaart and van Zanten (2009) gives a nice characterization of $\mathbb{H}^a$ in view of the isometry with the space $L_2(\nu_a)$. In the following Lemma 6.2, we express each element of $\mathbb{H}^a$ as a convolution of ψ_a^{2r} with a function in $C(\mathbb{R}^d)$ for any given $r \geq 1$. In other words, every element of $\mathbb{H}^a$ arises as a convolution of a higher order kernel with a function in $C(\mathbb{R}^d)$ showing that the search for the best approximator of a $C^\alpha[0,1]$ function in the space $\mathbb{H}^a$ can be restricted to only convolutions of continuous functions with a higher order kernel.

LEMMA 6.2. *Given any $h \in \mathbb{H}^a$ and $r \geq 1$, there exists $w \in C(\mathbb{R}^d)$ such that $h = \psi_{2a}^{2r} * w$.*

PROOF. By Lemma 4.1 of van der Vaart and van Zanten (2009), we obtain that any $h \in \mathbb{H}^a$ can be written as

$$t \to \int e^{i(\lambda,t)} g(\lambda) f_a(\lambda) d\lambda, \tag{6.6}$$

where $\int g(\lambda)^2 f_a(\lambda) d\lambda < \infty$.

By change of variable,

$$h(t) = \int e^{-i(\lambda,t)} g(-\lambda) f_a(\lambda) d\lambda \tag{6.7}$$

for $\int g(-\lambda)^2 f_a(\lambda) d\lambda < \infty$. Then $\hat{h}(\lambda) = (2\pi)^d g(-\lambda) f_a(\lambda)$. Now observe that $\hat{\psi}^{2r}(t)$ is real and positive for all values of t and $\hat{\psi}^{2r}(t) > e^{-\|t\|^2/2}$. Also note that $\hat{\psi}_{2a}^{2r}(t) = \hat{\psi}^{2r}(t/2a)$. Hence setting $\hat{w}(\lambda) = \frac{\hat{h}(\lambda)}{\hat{\psi}_{2a}^{2r}(\lambda)}$, we obtain

$$\hat{w}(\lambda) = \frac{g(-\lambda)\pi^{d/2} \exp\{-\|\lambda\|^2/4a^2\}}{\exp\{-\|\lambda\|^2/8a^2\} \prod_{j=1}^d \sum_{s=0}^{r-1} \frac{\lambda_j^{2s}}{(2a)^{2s} 2^s s!}}.$$

Thus $|\hat{w}(\lambda)| \leq \exp\{-\|\lambda\|^2/8a^2\} |g(-\lambda)|$ and

$$\begin{aligned} \left\{ \int |\hat{w}(\lambda)| \, d\lambda \right\}^2 &\leq \left\{ \int \exp\{-\|\lambda\|^2/8a^2\} |g(-\lambda)| \, d\lambda \right\}^2 \\ &\leq \int \exp\{-\|\lambda\|^2/4a^2\} |g(-\lambda)|^2 \, d\lambda \\ &< \infty \end{aligned}$$

As $\hat{w} \in L_1(\lambda)$, there exists a w for which $\hat{w}$ is a Fourier transform of w. Now, since $\hat{h} = \hat{\psi^{2r}}_{2a} \hat{w}$ and there exists a w for which $\hat{w}$ is a Fourier transform

of w, we immediately have $h = \psi_{2a}^{2r} * w$. In the following, we will find an expression for w and show that $w \in C(\mathbb{R}^d)$.

$$\begin{aligned} w(\lambda) &= \frac{1}{(2\pi)^d} \int e^{-i(\lambda,t)} \hat{w}(t) dt \\ &= \frac{1}{(2\pi)^d} \int e^{-i(\lambda,t)} \frac{g(-t)\pi^{d/2} \exp\{-\|t\|^2/4a^2\}}{\exp\{-\|t\|^2/8a^2\} \prod_{j=1}^d \sum_{s=0}^{r-1} \frac{t_j^{2s}}{(2a)^{2s} 2^s s!}} dt \end{aligned}$$

Thus $w \in C(\mathbb{R}^d)$ and its smoothness is controlled by g. □

The following Lemma 6.3 says that $\psi_a^{2r} * w_0$ can better approximate $w_0 \in C(\mathbb{R}^d)$ compared to $\psi_a^{2r} * w$ for any $w \neq w_0$. Lemma 6.3 further restricts the search for the best approximator of a $C(\mathbb{R}^d)$ function to only convolutions of the higher order kernel ψ_a^{2r} with the function w_0 itself.

LEMMA 6.3. *Given any $w_0 \in C(\mathbb{R}^d)$ compactly supported and $r \geq 1$,*

$$\left\| w_0 - \psi_a^{2r} * w_0 \right\|_\infty \leq \left\| w_0 - \psi_a^{2r} * w \right\|_\infty$$

for sufficiently large $a > 0$ and for any $w \in C(\mathbb{R}^d)$ compactly supported with $\|w - w_0\| > \delta$ for some $\delta > 0$.

PROOF. Note that

$$\left\| \psi_a^{2r} * w - w_0 \right\|_\infty \geq \|w - w_0\|_\infty - \left\| \phi_a^{2r} * w - w \right\|_\infty .$$

Since w is compactly supported, there exists $a_0 > 0$ such that for $a > a_0$, $\left\| \phi_a^{2r} * w - w \right\|_\infty < \delta/2$. The conlusion of the lemma follows by observing that for $a > a_0$, $\left\| \psi_a^{2r} * w - w_0 \right\|_\infty > \delta/2$. □

The following Lemma 6.3 provides a lower bound to the approximation error for $w_0(x_1, \ldots, x_d) = |x_1 - 0.5|^{1.5}, (x_1, \ldots, x_d) \in [0,1]^d$ with $\psi_a^2 * w_0$.

LEMMA 6.4. *For $w_0(x_1, \ldots, x_d) = |x_1 - 0.5|^{1.5}$,*

$$\left\| w_0 - \psi_{2a}^2 * w_0 \right\|_\infty \geq C_0 a^{-1.5} \tag{6.8}$$

for some global constant $C_0 > 0$.

PROOF. Since $w_0 \in C^{1.5}[0,1]^d$, by Whitney's theorem we can extend it to $\mathbb{R}^d$ so that w_0 has a compact support with $\|w_0\|_{1.5} < \infty$. Without loss of

generality, assume w_0 is non-negative and the support of w_0 is $[-L, L]^d$ for some large L. Observe that

$$\psi_{2a}^2 * w_0(1/2) - w_0(1/2) \quad = \quad \int \psi^2(s) w_0(1/2 - s/(2a)) ds$$

Now since $w_0(1/2 - s/(2a)) = 0$ if $|1/2 - s/(2a)| > L$, so for $a > 1/2$, $\{s : |1/2 - s/(2a)| \leq L\} \supset [-2L+1, 2L+1]^d$. Thus

$$\begin{aligned} \int \psi^2(s) w_0(1/2 - s/(2a)) ds &\geq \int_{[-2L+1,2L+1]^d} \psi^2(s) w_0(1/2 - s/(2a)) ds \\ &= 1/(2a)^{1.5} \int_{[-2L+1,2L+1]^d} \psi^2(s) \left|s_1\right|^{1.5} ds. \end{aligned}$$

This shows that $\left\|w_0 - \psi_{2a}^2 * w_0\right\|_\infty \geq C_0 a^{-1.5}$ where

$$C_0 = \frac{1}{2^{1.5}} \int_{[-2L+1,2L+1]^d} \psi^2(s) \left|s_1\right|^{1.5} ds.$$

Also it follows from the last part of Lemma 4.3 of van der Vaart and van Zanten (2009) that $\psi_{2a}^2 * w_0 \in \mathbb{H}^a$ since $(\hat{\psi}_{2a}^2)^2(\lambda) = f_a(\lambda)$. □

Note that the lower bound is same as the upper bound to the approximation error of any $C^{1.5}[0,1]$ function using $\psi_{2a}^2 * w$ upto constants.

The following Lemma 6.5 is crucial to the derivation of a lower bound to the concentration function $\phi_a(w_0)$. Lemma 6.5 complements Lemma 4.6 of van der Vaart and van Zanten (2009) and is an application of Theorem 2 of Kuelbs and Li (1993).

LEMMA 6.5. *There exists $\epsilon_0 > 0$, possibly depending on a, such that for all $\epsilon < \epsilon_0$,*

$$-\log P(\|W^a\|_\infty < \epsilon) \gtrsim a^d \log\left(\frac{|\log \epsilon|^{1/2}}{\epsilon}\right)^{d+1}. \tag{6.9}$$

PROOF. Obtaining the lower bound is a simple application of Lemma 4.5 of van der Vaart and van Zanten (2009) and Theorem 2 of Kuelbs and Li (1993). The proof of Lemma 4.3 of van der Vaart and van Zanten (2009) shows that

$$\log \mathrm{N}(\epsilon, \mathbb{H}_1^a, \|\cdot\|_\infty) \approx a^d \left(\log \frac{1}{\epsilon}\right)^{d+1}.$$

If we define $g_a(x) = a^d\big(\log \frac{1}{x}\big)^{d+1}$, it is easy to observe that g is a slowly varying function. Then by Theorem 2 of Kuelbs and Li (1993), we obtain

$$\phi_0^a(\epsilon) \geq C_1 g_a\left(\frac{\epsilon}{\sqrt{\phi_0^a(\epsilon)}}\right) = a^d\left(\log \frac{\sqrt{\phi_0^a(\epsilon)}}{\epsilon}\right)^{d+1}. \tag{6.10}$$

Below we show that we only need to find a crude lower bound to $\phi_0^a(\epsilon)$ to obtain the required bound. Observe that

$$\phi_0^a(\epsilon) = -\log P(\|W^a\|_\infty \leq \epsilon) \geq -\log P(\big|W^0\big| \leq \epsilon). \tag{6.11}$$

Note that $W^0 \sim \mathrm{N}(0,1)$ and hence $P(\big|W^0\big| \leq \epsilon) = \{2\Phi(\epsilon) - 1\} \approx 1 + |\log \epsilon|$ as $\epsilon \to 0$. Hence we obtain for sufficiently small ϵ,

$$\phi_0^a(\epsilon) \gtrsim |\log \epsilon|\,. \tag{6.12}$$

Plugging in the bound (6.12) in (6.10), we obtain

$$\phi_0^a(\epsilon) \gtrsim a^d \log\left(\frac{|\log \epsilon|^{1/2}}{\epsilon}\right)^{d+1}. \tag{6.13}$$

□

Note that the lower bound in Lemma 6.5 differs from the upper bound in Lemma 4.6 of van der Vaart and van Zanten (2009) only by a logarithmic factor suggesting that the lower bound obtained in reasonably tight.

Finally, we calculate the tail probability of the supremum of the Gaussian process $W^{\mathbf{A}}$ which will be crucially used to derive the lower bound to the posterior concentration rate. Although this is an application of Borel's Inequality van der Vaart and van Zanten (2008b), we will provide an independent proof to carefully identify the role of the prior for the bandwidth.

LEMMA 6.6. *For $r > 1$,*

$$P\big(\big\|W^{\mathbf{A}}\big\|_\infty > M\big) \leq$$
$$P(A > r) + 2(aM)^d \exp\left[-\frac{1}{2}M^2 + C\{(\log r)^{1/2} + (\log M)^{1/2}\}\right]$$

for some constant $C > 0$.

PROOF. From Theorem 5.2 of Adler (1990) it follows that if X is a centered Gaussian process on a compact set $T \subset \mathbb{R}^d$ and σ_T^2 is the maximum variance attained by the Gaussian process on T, then for large M,

$$P(\|X\|_\infty > M) \le 2N(1/M, T, \|\cdot\|) \exp\left[-\frac{1}{2\sigma_T^2}\{M - \nu(M)\}^2 \right],$$

where $\nu(M) = C_1 \int_0^{1/M} \{\log N(1/M, T, \|\cdot\|)\}^{1/2} d(1/M)$ for some constant $C_1 > 0$. Observe that W^a is rescaled to $T = [0, a]^d$ and the maximum variance attained by W^a is 1. Note that $N(1/M, T, \|\cdot\|) = (aM)^d$. Now

$$\begin{aligned} \nu(M) &\le C_2 \int_0^{1/M} \{d \log(aM)\}^{1/2} d(1/M) \\ &\le C_3 \int_0^{1/M} \{(\log a)^{1/2} + (\log M)^{1/2}\} d(1/M) \\ &\le C_3 \frac{1}{M} \{(\log a)^{1/2} + (\log M)^{1/2}\} \end{aligned}$$

for some constants $C_2, C_3 > 0$. Using W^a in place of X, we obtain,

$$P(\|W^a\|_\infty > M) \le 2(aM)^d \exp\left[-\frac{1}{2}M^2 + C_3\{(\log a)^{1/2} + (\log M)^{1/2}\} \right]$$

The conclusion of the lemma follows immediately. □

7. Main result. Below we state the main theorem on the lower bound of the posterior concentration rate using a Logistic Gaussian process prior when the true density is given by (6.1).

THEOREM 7.1. *If f_0 is given by (6.1) and the prior for a density f on $[0, 1]^d$ is given as in (6.2) for any $d^* \in \{1, \ldots, d\}$, then*

$$P(\|f - f_0\|_\infty \le n^{-3/(6+2d)} \log^{t_0} n \mid Y_1, \ldots, Y_n) \to 0 \tag{7.1}$$

a.s. as $n \to \infty$ for some constant $t_0 > 0$.

PROOF. To obtain the lower bound, we will verify the conditions of Lemma 1 in Castillo (2008) with $B_n = \{f : \|f - f_0\|_\infty \le \xi_n\}$ for $\xi_n = n^{-3/(6+2d)} \log^{t_0} n$ for some constant t_0 chosen appropriately in the subsequent analysis. From the proof of Lemma 5 in Castillo (2008) it follows that for $c_k = kd\xi_n, k = -N, \ldots, N$ and N the smallest integer larger than $C\sqrt{n}$,

$$\begin{aligned} &P\big(\|f - f_0\|_\infty \le \xi_n \big) \\ (7.2) \le &\sum_{k=-N}^{N} P\Big(\Big\| W^{\mathbf{A}} - w_0 - c_k \Big\|_\infty \le 2d\xi_n \Big) + P\Big(\Big\| W^{\mathbf{A}} \Big\|_\infty > C\sqrt{n}\xi_n \Big) \end{aligned}$$

An application of Lemma 6.6 with $M_n^2, r_n^{d^*} = O(n\xi_n^2)$ yields

$$\begin{aligned} P\big(\left\|W^{\mathbf{A}}\right\|_\infty > C\sqrt{n}\xi_n\big) &\leq P(A > r_n) + \exp\{-K_1 M_n^2\} \\ &\leq \exp\{-r_n^{d^*}\} + \exp\{-K_1 M_n^2\} \\ &\leq \exp\{-K_2 n\xi_n^2\} \end{aligned} \tag{7.3}$$

for some constants $K_1, K_2 > 0$.

Lemma 6.2, 6.3, 6.4 and the observation that $w_0 \notin C^{1.5+\delta}[0,1]^d$ for any $\delta > 0$ together imply that given any $\epsilon > 0$ there does not exist any element in $\mathbb{H}^a$ such that for each $k = -N, \ldots, N$

$$\|w_0 - h - c_k\|_\infty < \epsilon \tag{7.4}$$

if $a < C_0\epsilon^{-1/\alpha}$, where w_0 is given by $w_0(x_1, \ldots, x_d) = |x_1 - 0.5|^{1.5}$. From Lemma 6.5, we obtain if $a > C_0\epsilon^{-1/\alpha}$,

$$\phi^a_{w_0+c_k}(\epsilon) \geq \inf_{h\in\mathbb{H}^a:\|h-w_0\|_\infty<\epsilon} \frac{1}{2}\|h\|^2_{\mathbb{H}} + a^d \log\left(\frac{|\log\epsilon|^{1/2}}{\epsilon}\right)^{d+1} \tag{7.5}$$

$$\geq a^d \log\left(\frac{|\log\epsilon|^{1/2}}{\epsilon}\right)^{d+1}. \tag{7.6}$$

Hence for $k = -N, \ldots, N$,

$$P\left(\left\|W^{\mathbf{A}} - w_0 - c_k\right\|_\infty < \epsilon\right) \leq \int_{a=C_0\epsilon^{-1/\alpha}}^\infty \exp\left\{-a^d \log\left(\frac{|\log\epsilon|^{1/2}}{\epsilon}\right)^{d+1}\right\} da.$$

Using the inequality

$$\int_v^\infty \exp\{-t^r\}dt \leq 2r^{-1}v^{1-r}\exp\{-v^r\}, \tag{7.7}$$

we obtain that

$$P\left(\left\|W^{\mathbf{A}} - w_0 - c_k\right\|_\infty < \epsilon\right) \leq C_1 \exp\{-C_2\epsilon^{-d/\alpha}|\log\epsilon|^{d+1}\}.$$

for some constants $C_1, C_2 > 0$. Thus we obtain from (7.3) and (7.2),

$$P\big(\|f - f_0\|_\infty \leq \xi_n\big) \leq C_3 N \exp\{-C_4\xi_n^{-d/\alpha}\}. \tag{7.8}$$

for some constant $C_3 > 0$. From van der Vaart and van Zanten (2009) it also follows that

$$P(B_{KL}(f_0, \xi_n)) \geq e^{-C_5 n\xi_n^2} \tag{7.9}$$

for some constant $t_0 > 0$ and $C_5 > 0$ where

$$(7.10)\quad B_{KL}(f_0, \epsilon) = \left\{ f : \int f_0 \log \frac{f_0}{f} < \epsilon^2, \int f_0 \left(\log \frac{f_0}{f} \right)^2 < \epsilon^2 \right\}.$$

By adjusting t_0, C_4 and C_5, we have from (7.8) and (7.9)

$$\frac{P\big(\|f - f_0\|_\infty \le \xi_n \big)}{P(B_{KL}(f_0, \xi_n))} \le \exp\{-2n\xi_n^2\},$$

which proves the assertion of the theorem by Lemma 1 of Castillo (2008). □

Note that the lower bound $n^{-3/(6+2d)} \log^{t_0} n$ for $d > 1$ is only a sub-optimal rate for estimating w_0, the optimal rate being given by $n^{-3/8}$.

References.

Adler, R. J. (1990). *An introduction to continuity, extrema, and related topics for general Gaussian processes* **12**. Institute of Mathematical Statistics.

Barron, A., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probability theory and related fields* **113** 301–413.

Birgé, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probability theory and related fields* **71** 271–291.

Birgé, L. (2001). An alternative point of view on Lepski's method. *Lecture Notes-Monograph Series* 113–133.

Castillo, I. (2008). Lower bounds for posterior rates with Gaussian process priors. *Electronic Journal of Statistics* **2** 1281–1299.

De Jonge, R. and van Zanten, J. (2010). Adaptive nonparametric bayesian inference using location-scale mixture priors. *The Annals of Statistics* **38** 3300–3320.

Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Annals of Statistics* **28** 500–531.

Hoffmann, M. and Lepski, O. (2002). Random rates in anisotropic regression. *Annals of statistics* 325–358.

Ibragimov, I. A. and Khasminski, R. Z. (1981). *Statistical estimation–asymptotic theory* **16**. Springer.

Kerkyacharian, G., Lepski, O. and Picard, D. (2001). Nonlinear estimation in anisotropic multi-index denoising. *Probability theory and related fields* **121** 137–170.

Klutchnikoff, N. (2005). On the adaptive estimation of anisotropic functions PhD thesis, Ph. D. thesis, Univ. Aix–Marseille I.

Kuelbs, J. and Li, W. V. (1993). Metric entropy and the small ball problem for Gaussian measures. *J. Funct. Anal* **116** 133–157.

Lenk, P. J. (1988). The logistic normal distribution for Bayesian, nonparametric, predictive densities. *Journal of the American Statistical Association* **83** 509–516.

Lenk, P. J. (1991). Towards a practicable Bayesian nonparametric density estimator. *Biometrika* **78** 531.

Lepski, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Teoriya Veroyatnostei i ee Primeneniya* **35** 459–470.

Lepski, O. (1991). Asymptotic minimax adaptive estimation. —. Upper bounds. *Theory Probab. Appl* **36** 645-659.

Lepski, O. (1992). Asymptotic minimax adaptive estimation. 2.— Statistical model without optimal adaptation. Adaptive estimators. *Theory Probab. Appl* **37** 468–481.

Nussbaum, M. (1985). Spline smoothing in regression models and asymptotic efficiency in L2. *The Annals of Statistics* 984–997.

Rasmussen, C. E. (2004). Gaussian processes in machine learning. *Advanced Lectures on Machine Learning* 63–71.

Savitsky, T., Vannucci, M. and Sha, N. (2011). Variable selection for nonparametric Gaussian process priors: Models and computational strategies. *Statistical Science* **26** 130–149.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics* 1040–1053.

Tokdar, S. T. and Ghosh, J. K. (2007). Posterior consistency of logistic Gaussian process priors in density estimation. *Journal of Statistical Planning and Inference* **137** 34–42.

van der Vaart, A. and van Zanten, J. (2008a). Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics* **36** 1435–1463.

van der Vaart, A. and van Zanten, J. (2008b). Reproducing kernel Hilbert spaces of Gaussian priors. *IMS Collections* **3** 200–222.

van der Vaart, A. W. and van Zanten, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *The Annals of Statistics* **37** 2655–2675.

Wand, M. P. and Schucany, W. R. (1990). Gaussian-based kernels. *Canadian Journal of Statistics* **18** 197–204.

Zou, F., Huang, H., Lee, S. and Hoeschele, I. (2010). Nonparametric Bayesian Variable Selection With Applications to Multiple Quantitative Trait Loci Mapping With Epistasis and Gene–Environment Interaction. *Genetics* **186** 385.

BOX 90251, Old Chemsitry Building
Durham, NC 27708
E-mail: dp55@stat.duke.edu
dunson@stat.duke.edu; ab179@stat.duke.edu