11

Gaussian Process Priors

Gaussian processes are random functions and can serve as priors on function spaces. The variety of Gaussian processes, given by their covariance kernels, provides many modelling choices, which can be further modified by including hyperparameters in their covariance kernels. The associated geometrical structure of Gaussian processes, encoded in their reproducing kernel Hilbert spaces, makes for an elegant theory of posterior contraction. We begin the chapter with a review and examples of Gaussian processes. Next we present general results on posterior contraction rates, and apply them in several inference problems: density estimation, binary and normal regression, white noise. We proceed with methods of rescaling and adaptation using mixtures of Gaussian processes. The chapter closes with a review of computational techniques.

11.1 Definition and Examples

Definition 11.1 (Gaussian process) A *Gaussian process* is a stochastic process $W = (W_t : t \in T)$ indexed by an arbitrary set T such that the vector $(W_{t_1}, \ldots, W_{t_k})$ possesses a multivariate normal distribution, for every $t_1, \ldots, t_k \in T$ and $k \in \mathbb{N}$. A Gaussian process W indexed by $\mathbb{R}^d$ is called *self-similar* of index α if $(W_{\sigma t} : t \in \mathbb{R}^d)$ is distributed like $(\sigma^\alpha W_t : t \in \mathbb{R}^d)$, for every $\sigma > 0$, and *stationary* if $(W_{t+h} : t \in \mathbb{R}^d)$ has the same distribution as $(W_t : t \in \mathbb{R}^d)$, for every $h \in \mathbb{R}^d$.

The vectors $(W_{t_1}, \ldots, W_{t_k})$ in the definition are called marginals, and their distributions *marginal distributions* or *finite-dimensional distributions*. Since a multivariate normal distribution is determined by its mean vector and covariance matrix, the finite-dimensional distributions of a Gaussian process are determined by the *mean function* and *covariance kernel*, defined by

$$\mu(t) = \mathrm{E}(W_t), \qquad K(s,t) = \mathrm{cov}(W_s, W_t), \qquad s, t \in T.$$

The mean function is an arbitrary function $\mu : T \to \mathbb{R}$, which for prior modeling is often taken equal to zero: a shift to a nonzero mean can be incorporated in the statistical model. The covariance kernel is a bilinear, symmetric nonnegative-definite function $K : T \times T \to \mathbb{R}$, and determines the properties of the process.[1] By Kolmogorov's extension theorem there exists a Gaussian process for any mean function and covariance kernel.

[1] Symmetric nonnegative-definite means that every matrix $((K(t_i, t_j)))_{i,j=1,\ldots,k}$ possesses the property with the same name, for every $t_1, \ldots, t_k$.

Definition 11.1 defines a Gaussian process as a collection of random variables W_t restricted only to be marginally normally distributed. It does not refer to the *sample paths* $t \mapsto W_t$, which define the process as a *random function*. The properties of the sample paths are often of importance for interpreting a prior, and are not fully determined by the mean and covariance functions, as the marginal distributions do not change by changing the variables W_t on null sets, whereas the sample paths do. Often this discrepancy is resolved by considering the version of the process with continuous sample paths, if that exists.[2] Here a process $\tilde{W}$ is a *version* of W if $\tilde{W}_t = W_t$, almost surely for every $t \in T$. Consideration of the sample paths makes it also possible to think of the process as a map $W: \Omega \to \mathbb{B}$ from the underlying probability space to a function space $\mathbb{B}$, such as the space of continuous functions on T or a space of differentiable functions. This connects to the following abstract definition.

Let $\mathbb{B}^*$ denote the *dual space* of a Banach space $\mathbb{B}$: the collection of continuous, linear maps $b^*: \mathbb{B} \to \mathbb{R}$.

Definition 11.2 (Gaussian random element) A *Gaussian random element* is a Borel measurable map W into a Banach space $(\mathbb{B}, \|\cdot\|)$ such that the random variable $b^*(W)$ is normally distributed, for every $b^* \in \mathbb{B}^*$.

The two definitions can be connected in two ways. First a Gaussian random element W always induces the Gaussian stochastic process $(b^*(W): b^* \in T)$, for any subset T of the dual space. Second if the sample paths $t \mapsto W_t$ of the stochastic process $W = (W_t: t \in T)$ belong to a Banach space $\mathbb{B}$ of functions $z: T \to \mathbb{R}$, then under reasonable conditions the process will be a Gaussian random element. The following lemma is sufficient for most purposes. For instance, it applies to the space $\mathfrak{C}(T)$ of continuous functions on a compact metric space T whenever the Gaussian process has continuous sample paths. For a proof and additional remarks, see Lemmas I.5 and I.6 in the appendix.

Lemma 11.3 *If the sample paths $t \mapsto W_t$ of the stochastic process $W = (W_t: t \in T)$ belong to a separable subset of a Banach space and the norm $\|W - w\|$ is a random variable for every w in the subset, then W is a Borel measurable map in this space. Furthermore, if W is a Gaussian process and the Banach space is $\mathfrak{L}_\infty(T)$ equipped with the supremum norm, then W is a Gaussian random element in this space.*

Example 11.4 (Random series) If $Z_1, \ldots, Z_m \overset{\text{iid}}{\sim} \text{Nor}(0, 1)$ variables and $a_1, \ldots, a_m$ are arbitrary functions, then $W_t = \sum_{i=1}^m a_i(t) Z_i$ defines a mean zero Gaussian process with covariance kernel $K(s, t) = \sum_{i=1}^m a_i(s) a_i(t)$.

Special cases are the *polynomial process* given by the functions $a_i(t) = t^i$, and the *trigonometric process* given by the trigonometric functions.

Another case of interest is given by functions $a_i(t) = \sigma^{-d} \psi((t - t_i)/\sigma)$, constructed by shifting and dilating a kernel function ψ, for instance a probability density ψ on $T = [0, 1]^d$ and a regular grid $t_1, \ldots, t_m$ over T.

[2] A weaker restriction is that the process is *separable*, which means that there is a countable subset $T_0 \subset T$ such that the suprema over open sets are equal to suprema over the intersections of these sets with T_0.

If $a_1, a_2, \ldots$ are functions such that $\sum_{i=1}^{\infty} a_i^2(t) < \infty$ for all t, then the construction extends to the case $m = \infty$: the series $W_t = \sum_{i=1}^{\infty} a_i(t) Z_i$ converges a.s. Theorem I.25 shows that all Gaussian processes can be expressed as an infinite series.

Example 11.5 (Brownian motion) The standard *Brownian motion* or *Wiener process* on $[0, 1]$ (or $[0, \infty)$) is the mean zero Gaussian process with continuous sample paths and covariance function $K(s, t) = \min(s, t)$. The process has stationary, independent increments: $B_t - B_s \sim \text{Nor}(0, t - s)$ and is independent of $(B_u : u \le s)$, for any $s < t$. Hence it is a Lévy process, i.e. a process with stationary and independent increments. Brownian motion is self-similar of index $1/2$.

The sample paths of Brownian motion are Lipschitz continuous of any order $\alpha < 1/2$, and hence W can be viewed as a map in the Hölder space $\mathfrak{C}^{\alpha}[0, 1]$, for any $0 \le \alpha < 1/2$. That it is also a *Gaussian random element* in the sense of Definition 11.2 is not immediate, but can be shown along the lines of Lemma 11.3 (see Lemma I.7).

The sample paths also belong to the Besov space $\mathfrak{B}^{1/2}_{1,\infty}[0, 1]$, and hence the process can also be viewed to be of exact regularity $1/2$, relative to a weaker norm.

The process $B_t - t B_1$ is called the *Brownian bridge*. It has covariance function $K(s, t) = \min(s, t) - st$, and can also be derived through conditioning Brownian motion B on the event $B_1 = 0$.

Example 11.6 (Integrated Brownian motion, Riemann-Liouville) In view of its rough appearance Brownian motion is often considered a poor prior model for a function. A simple method to smooth it is to take consecutive primitives. If $I_{0+} f$ denotes the primitive function $I_{0+} f(t) = \int_0^t f(s)\, ds$ of a function f, and $I_{0+}^k f = I_{0+}^{k-1} I_{0+} f$ its k-fold primitive, and B is Brownian motion, then $W = I_{0+}^k B$ is a Gaussian process with sample paths that are k times differentiable with a kth derivative that is Lipschitz of order almost $1/2$. Thus W is nearly $k + 1/2$-smooth.

This operation creates processes of smoothness levels $1/2, 3/2, \ldots$. It is possible to interpolate between these values by *fractional integration*. By partial integration it can be seen that, for $k \in \mathbb{N}$ and a continuous function f,

$$I_{0+}^k f(t) = \frac{1}{\Gamma(k)} \int_0^t (t - s)^{k-1} f(s)\, ds. \tag{11.1}$$

The formula on the right makes good sense also for noninteger values of k, and then is called the fractional integral of f. The process $I_{0+}^{\alpha} B$ resulting from integrating Brownian motion B is called the Riemann-Liouville process with Hurst parameter $\alpha + 1/2$, for $\alpha \ge 0$. The sample paths of this process are nearly $\alpha + 1/2$-smooth. To cover also the smoothness levels in $[0, 1/2]$, the *Riemann-Liouville process* with *Hurst parameter* $\alpha > 0$ is defined more generally as the stochastic integral

$$R_t^{\alpha} = \frac{1}{\Gamma(\alpha + 1/2)} \int_0^t (t - s)^{\alpha - 1/2}\, dB_s, \qquad t \ge 0. \tag{11.2}$$

The process R^{α} can be viewed as the the $(\alpha + 1/2)$-fractional integral of the "derivative dB of Brownian motion." It is Gaussian with zero mean and has nearly α-smooth sample paths.

These primitive processes are "tied at zero," in that their function value as well as their derivatives at zero vanish. This is undesirable for prior modeling. The easiest method to "release" the processes at zero is to add a polynomial $t \mapsto \sum_{j=0}^{k} Z_j t^j$, for independent centered Gaussian random variables $Z_0, \ldots, Z_k$, independent of B. The resulting process is then still Gaussian with mean zero.

The self-similarity property can be seen to "integrate": k times integrated Brownian motion $I_{0+}^k B$ is self-similar of index $k + 1/2$, and the Riemann-Liouville process is self-similar of index α.

Example 11.7 (Ornstein-Uhlenbeck) The standard *Ornstein-Uhlenbeck process* with parameter θ is the (only) mean zero, stationary, Markovian Gaussian process with time set $T = [0, \infty)$ and continuous sample paths. It can be constructed from a Brownian motion B through the relation $W_t = (2\theta)^{-1/2} e^{-\theta t} B_{e^{2\theta t}}$, for $t \geq 0$, and its covariance kernel is given by $K(s,t) = (2\theta)^{-1} e^{-\theta|s-t|}$. It can also be characterized as the solution to the stochastic differential equation $dW_t = -\theta\, W_t\, dt + dB_t$, which exhibits the process as a continuous time autogressive process.

Example 11.8 (Stationary process, square exponential, Matérn) Stationary Gaussian processes with index set $T \subset \mathbb{R}^d$ are characterized by covariance kernels of the form $K(s,t) = K(s-t)$, where $K : \mathbb{R}^d \to \mathbb{R}$ is a positive-definite function, and we abuse notation by giving this the same name as the kernel. Provided the index set is rich enough *Bochner's theorem* gives an alternative characterization in terms of the *spectral measure* of the process. This is a symmetric, finite measure μ on $\mathbb{R}$ such that

$$K(s-t) = \int e^{-i\langle s-t, \lambda\rangle}\, d\mu(\lambda). \tag{11.3}$$

Popular examples are the *square exponential process* and the family of *Matérn processes*, given by the spectral measures

$$d\mu(\lambda) = 2^{-d} \pi^{-d/2} e^{-\|\lambda\|^2/4}\, d\lambda, \tag{11.4}$$

$$d\mu(\lambda) = (1 + \|\lambda\|^2)^{-\alpha - d/2}\, d\lambda, \qquad \alpha > 0. \tag{11.5}$$

The covariance function of the square exponential process takes the simple explicit form $\mathrm{E}[W_s W_t] = e^{-\|s-t\|^2}$. For the Matérn process it can be represented in terms of special functions (see e.g. Rasmussen and Williams 2006, page 84).

As the notation suggests, the sample paths of the Matérn process are α-smooth. The sample paths of the square exponential process are infinitely often differentiable, and even analytic. These claims may be inferred from Proposition I.4.

Example 11.9 (Fractional Brownian motion) The *fractional Brownian motion* (fBm) with *Hurst parameter* $\alpha \in (0,1)$ is the mean zero Gaussian process $W = (W_t : t \in [0,1])$ with continuous sample paths and covariance function

$$\mathrm{E}(W_s W_t) = \tfrac{1}{2}(s^{2\alpha} + t^{2\alpha} - |t-s|^{2\alpha}). \tag{11.6}$$

The choice $\alpha = 1/2$ yields the ordinary Brownian motion. To obtain a process of a given smoothness $\alpha > 1$, we can take an ordinary integral of fractional Brownian motion of order $\alpha - \langle\alpha\rangle$.

Example 11.10 (Kriging) For a given Gaussian process $(W_t : t \in T)$ and fixed, distinct points $t_1, \ldots, t_m \in T$ the conditional expectations $W_t^* = \mathrm{E}(W_t \,|\, W_{t_1}, \ldots, W_{t_m})$ define another Gaussian process. It can be written as $W_t^* = \sum_{i=1}^m a_i(t) W_{t_i}$, where, for K the covariance function of W,

$$\begin{pmatrix} a_1(t) \\ \vdots \\ a_m(t) \end{pmatrix} = \Sigma^{-1}\sigma(t), \qquad \Sigma = (\!(K(t_i, t_j))\!)_{i,j=1,\ldots,m}, \qquad \sigma(t) = \begin{pmatrix} K(t, t_1) \\ \vdots \\ K(t, t_m) \end{pmatrix}.$$

The process W^* takes its randomness from the finitely many random variables $W_{t_1}, \ldots, W_{t_m}$, and coincides with these variables at the points $t_1, \ldots, t_m$. In spatial statistics this interpolation operation is known as *kriging*. The process W^* is referred to as the *interpolating Gaussian process*, the *Gaussian sieve process*, or the *predictive process*.

The covariance kernel of the process W^* is given by

$$K^*(s, t) = \sum_{i=1}^m \sum_{j=1}^m a_i(s) a_j(t) K(t_i, t_j) = \sigma(s)^\top \Sigma \sigma(t). \tag{11.7}$$

If W has continuous sample paths, then so does W^*. In that case the process W^* converges to W when $m \to \infty$ and the interpolating points $t_1, \ldots, t_m$ grow dense in T.

Example 11.11 (Scaling) If $W = (W_t : t \in \mathbb{R}^d)$ is a Gaussian process with covariance kernel K, then the process $(W_{at} : t \in \mathbb{R}^d)$ is another Gaussian process, with covariance kernel $K(as, at)$, for any $a > 0$. A scaling factor $a < 1$ stretches the sample paths, whereas a factor $a > 1$ shrinks them. Intuitively, if restricted to a compact domain stretching decreases the variability of the sample paths, whereas shrinking increases the variability. This is illustrated in Figure 11.1. Even though the smoothness of the sample paths does not change in a qualitative analytic sense, these operations may completely change the properties of the process as a prior, as we shall see in Section 11.5.

11.2 Reproducing Kernel Hilbert Space

Every Gaussian process comes with an intrinsic Hilbert space, determined by its covariance kernel. This space determines the support and shape of the process, and therefore is crucial for the properties of the Gaussian process as a prior. The definition is different for Gaussian processes (as in Definition 11.1) and Gaussian random elements (as in Definition 11.2).

For a Gaussian process $W = (W_t : t \in T)$, let $\overline{\mathrm{lin}}(W)$ be the closure of the set of all linear combinations $\sum_i \alpha_i W_{t_i}$ in the $\mathbb{L}_2$-space of square-integrable variables. The space $\overline{\mathrm{lin}}(W)$ is a Hilbert space, called the *first order chaos* of W.

Definition 11.12 (Stochastic process RKHS) The *reproducing kernel Hilbert space* (*RKHS*) of the mean zero, Gaussian process $W = (W_t : t \in T)$ is the set of all functions

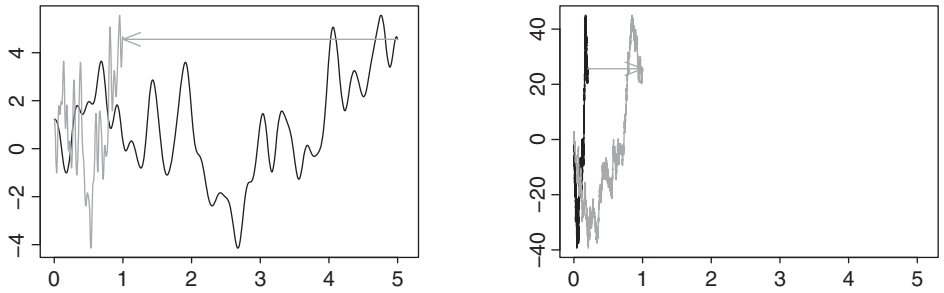

Figure 11.1 Shrinking the time scale of the square exponential process (left) yields a process of more variability on the desired domain [0, 1], whereas stretching a Brownian motion (right) creates a smoother prior.

$z_H : T \to \mathbb{R}$ defined by $z_H(t) = \mathrm{E}(W_t H)$, for H ranging over $\overline{\mathrm{lin}}(W)$. The corresponding inner product is

$$\langle z_{H_1}, z_{H_2} \rangle_{\mathbb{H}} = \mathrm{E}(H_1 H_2).$$

By the definition of the inner product, the correspondence $z_H \leftrightarrow H$ is an isometry between the RKHS $\mathbb{H}$ and the first-order chaos $\overline{\mathrm{lin}}(W)$. This shows that the definition is well posed (the correspondence is one-to-one), and also that $\mathbb{H}$ is indeed a Hilbert space.

The function corresponding to $H = \sum_i \alpha_i W_{s_i}$ is

$$z_H(t) = \mathrm{E}(W_t H) = \sum_{i=1}^{k} \alpha_i K(s_i, t).$$

In particular, the function $K(s, \cdot)$ is contained in the RKHS, for every fixed $s \in T$, and corresponds to the variable W_s. For a general function $z_H \in \mathbb{H}$ we can compute $\langle z_H, K(s, \cdot) \rangle_{\mathbb{H}} = \mathrm{E}(H W_s) = z_H(s)$. In other words, for every $h \in \mathbb{H}$,

$$h(t) = \langle h, K(t, \cdot) \rangle_{\mathbb{H}}. \tag{11.8}$$

This is the *reproducing formula* that gives the RKHS its name. In general, a reproducing kernel Hilbert space is defined to be a Hilbert space of functions for which there exists a "kernel" K that makes the reproducing formula valid.

A Gaussian random element in a separable Banach space (in the sense of Definition 11.2) also comes with a reproducing kernel Hilbert space. This has a more abstract definition, but will be fundamental in the sequel. First we define the map $S : \mathbb{B}^* \to \mathbb{B}$ by

$$Sb^* = \mathrm{E}\left[b^*(W) W\right].$$

The right side of this equation is the *Pettis integral* of the $\mathbb{B}$-valued random variable $b^*(W)W$. By definition this is the unique element $Sb^* \in \mathbb{B}$ such that $b_2^*(Sb^*) = \mathrm{E}\left[b_2^*(W) b^*(W)\right]$, for every $b_2^* \in \mathbb{B}^*$[3].

Definition 11.13 (RKHS) The *reproducing kernel Hilbert space* (RKHS) of the Gaussian random element W is the completion of the range $S\mathbb{B}^*$ of the map $S : \mathbb{B}^* \to \mathbb{B}$ for the inner product

[3] This exists by Lemma I.9, as $\|W b^*(W)\| \le \|b^*\| \, \|W\|^2$, and $\|W\|$ has finite moments.

$$\langle Sb_1^*, Sb_2^* \rangle_{\mathbb{H}} = \mathrm{E}\left[b_1^*(W) b_2^*(W)\right].$$

Even though there is nothing "Gaussian" about the RKHS of a Gaussian process, as this depends on the covariance function only, it may be shown that this completely characterizes the Gaussian measure: every RKHS corresponds to exactly one centered Gaussian measure on a Banach space. Furthermore, to a certain extent the RKHS does not depend on the Banach space in which the process is embedded, but is intrinsic to the process (see Lemma I.17).

In Lemma 11.3 Gaussian processes are related to Gaussian random elements. There is a similar correspondence between the two types of RKHSs. If the sample paths $t \mapsto W_t$ of a mean zero stochastic process $W = (W_t : t \in T)$ belong to a Banach space $\mathbb{B}$ of functions $b : T \to \mathbb{R}$, and the coordinate projections $\pi_t : b \mapsto b(t)$ are elements of $\mathbb{B}^*$, then

$$K(s,t) = \mathrm{E}[W_s W_t] = \mathrm{E}\left[\pi_s(W)\pi_t(W)\right] = \langle S\pi_s, S\pi_t \rangle_{\mathbb{H}}. \tag{11.9}$$

The three equalities are immediate from the definitions of K, π_t and S, respectively. By the reproducing formula (11.8), the left side is equal to $\langle K(s,\cdot), K(t,\cdot)\rangle_{\mathbb{H}}$, and hence we infer the correspondence $K(t,\cdot) \leftrightarrow S\pi_t$ between the stochastic process and Banach space RKHSs. The following lemma makes this precise for the case of most interest.

Lemma 11.14 *If W is a mean zero Gaussian random element in a separable subspace of $\mathfrak{L}_\infty(T)$ equipped with the supremum norm, then the stochastic process RKHS of Definition 11.12 and the Banach space RKHS of Definition 11.13 coincide, with the correspondence given by $K(t,\cdot) = S\pi_t$.*

Proof As argued in the proof of Lemma I.6, the assumptions imply that W can be viewed as a Gaussian random element in the subspace $\mathfrak{UC}(T,\rho)$ of $\mathfrak{L}_\infty(T)$ for some semimetric ρ on T under which T is totally bounded. Furthermore, it is shown in this proof that every element of the dual of the space is the pointwise limit of a sequence of linear combinations of coordinate projections. In other words, the linear span $\mathbb{B}_0^*$ of the coordinate projections is weak-* dense in $\mathfrak{UC}(T,\rho)^*$, and hence the RKHS is the completion of $S\mathbb{B}_0^*$, by Lemma I.11 and the identification given in equation (11.9). □

Example 11.15 (Euclidean space) To gain insight in the RKHS it helps to consider a Gaussian random vector $W \sim \mathrm{Nor}_k(0,\Sigma)$ in $\mathbb{R}^k$. This can be identified with the stochastic process $W = (W_i : i = 1,\ldots,k)$ on the time set $T = \{1,2,\ldots,k\}$ and is of course also a random element in the Banach space $\mathbb{R}^k$. The covariance kernel is $K(i,j) = \Sigma_{i,j}$ and the RKHS is the space of functions $z_\alpha : \{1,\ldots,k\} \to \mathbb{R}$ given by $z_\alpha(i) = \mathrm{E}\left[W_i(\alpha^T W)\right] = (\Sigma\alpha)_i$ indexed by the (coefficients of the) linear combinations $\alpha^\top W \in \mathrm{lin}(W_1,\ldots,W_k)$, with inner product $\langle z_\alpha, z_\beta\rangle_{\mathbb{H}} = \mathrm{E}\left[(\alpha^\top W)(\beta^\top W)\right] = \alpha^\top\Sigma\beta$. We can identify z_α with the vector $\Sigma\alpha$, and the inner product then satisfies $\langle \Sigma\alpha, \Sigma\beta\rangle_{\mathbb{H}} = \alpha^\top\Sigma\beta$. The map S in the definition of the Banach space RKHS is equal to Σ.

In other words, the RKHS is the range of the covariance matrix, with inner product given through the (generalized) inverse of the covariance matrix. If the covariance matrix is non-singular, then the RKHS is $\mathbb{R}^k$, but equipped with the inner product generated by the inverse covariance matrix Σ^{-1}.

The RKHS reflects the familiar ellipsoid contours of the density of the multivariate normal distribution.

Example 11.16 (Series) Any mean zero Gaussian random element in a separable Banach space can be represented as an infinite series

$$W = \sum_{i=1}^{\infty} Z_i h_i,$$

for i.i.d. standard normal variables $Z_1, Z_2, \ldots$ and $h_1, h_2, \ldots$ (deterministic) elements of the Banach space $\mathbb{B}$. The series converges almost surely in the norm of $\mathbb{B}$, and the elements h_i have norm one in the RKHS (if they are chosen linearly independent). For a Banach function space and $h_1, h_2, \ldots$ equal to multiples of the eigen functions of the covariance kernel, this is known as the *Karhunen-Loève expansion*, but in fact any orthonormal basis $h_1, h_2, \ldots$ of the RKHS will do. (See Section I.6 for discussion of infinite Gaussian series, and Theorem I.25 for the existence of series representations.)

The series representation invites to identify W with the infinite vector $(Z_1, Z_2, \ldots)$, whose random coordinates Z_i give the spread of W in the deterministic, pre-chosen directions h_i. One should note here that the RKHS norm is stronger than the original Banach space norm, and necessarily $\|h_i\| \to 0$ as $i \to \infty$, even if $\|h_i\|_{\mathbb{H}} = 1$, for every i. If the directions h_i are placed in the order of decreasing norms, then the spread decreases as $i \to \infty$, in some sense to zero, even though Z_i is unbounded. This reveals the Gaussian distribution as an ellipsoid with smaller and smaller axes.

Assume that the functions $\{h_i : i \in \mathbb{N}\}$ are chosen linearly independent in $\mathbb{B}$ in the sense that $w = 0$ is the only element $w \in \ell_2$ such that $\sum_{i=1}^{\infty} w_i h_i = 0$. Then the RKHS of W can also be described in terms of infinite sums: it consists of all linear combinations $\sum_{i=1}^{\infty} w_i h_i$ for $w = (w_1, w_2, \ldots)$ ranging over ℓ_2, and the correspondence $\sum_{i=1}^{\infty} w_i h_i \leftrightarrow w$ is an isometry between $\mathbb{H}$ and ℓ_2, i.e.

$$\left\langle \sum_{i=1}^{\infty} v_i h_i, \sum_{i=1}^{\infty} w_i h_i \right\rangle_{\mathbb{H}} = \sum_{i=1}^{\infty} v_i w_i.$$

Thus the RKHS gives the ellipsoid shape of the distribution, as for the multivariate-normal distribution.

Three properties of Gaussian variables relating to their RKHS are important for the analysis of Gaussian priors. (For proofs and references for the following, see Appendices I.5 and I.7.) The first relates to the shape of the Gaussian distribution, and is described in *Borell's inequality*. Let $\mathbb{B}_1$ and $\mathbb{H}_1$ stand for the unit balls of the spaces $\mathbb{B}$ and $\mathbb{H}$, respectively, under their norms, and let Φ be the cumulative distribution function of the standard normal distribution. Furthermore, let $\varphi_0(\epsilon)$ be the *small ball exponent*, defined through, for $\epsilon > 0$,

$$\mathrm{P}(\|W\| < \epsilon) = e^{-\varphi_0(\epsilon)}. \tag{11.10}$$

Proposition 11.17 (Borell's inequality) *For any mean zero Gaussian random element W in a separable Banach space and every $\epsilon, M > 0$,*

$$\mathrm{P}\left(W \in \epsilon\,\mathbb{B}_1 + M\mathbb{H}_1\right) \geq \Phi\left(\Phi^{-1}(e^{-\varphi_0(\epsilon)}) + M\right).$$

For $M = 0$ the inequality in the proposition is an equality, and just reduces to the definition of the small ball exponent. For $M \to \infty$ and fixed $\epsilon > 0$, the right side tends to 1 like the tail of the normal distribution. This shows that the bulk of the Gaussian distribution is contained in an ϵ-shell of a big multiple of the unit ball of the RKHS. We should keep in mind here the ellipsoid shapes found in Examples 11.15 and 11.16, which is coded in general in the shape of $\mathbb{H}_1$ within $\mathbb{B}$. Not only does the Gaussian distribution concentrate most mass close to zero (it has small tails), but also it distributes the mass unevenly in the infinitely many possible directions, as determined by the shape of the RKHS.

The addition of the small ball $\epsilon\mathbb{B}_1$ creates an ϵ-cushion around the multiple $M\mathbb{H}_1$. This is necessary to capture the mass of W, because the RKHS itself may have probability zero. Since $M\mathbb{H}_1 \uparrow \mathbb{H}$ as $M \uparrow \infty$, we have the equality $\mathrm{P}(W \in \epsilon\mathbb{B}_1 + \mathbb{H}) = 1$, for any $\epsilon > 0$, and hence W is supported on the closure $\bar{\mathbb{H}}$ of the RKHS in $\mathbb{B}$.

The second property of a Gaussian variable relates to the change in measure under changes of location. Zero-mean Gaussian variables put most mass near 0; in the Euclidean case this results from the mode of the density being at the mean, while *Anderson's lemma* (see Lemma K.12) expresses this in general. The decrease of mass in a ball of a given radius if the center of this ball is moved away from 0 may be studied quantitatively using Radon-Nikodym derivatives.

The distributions of the two Gaussian variables W and $W + h$ can be shown to be either mutually absolutely continuous or orthogonal, depending on whether the shift h is contained in the RKHS or not. In the first case, when $h \in \mathbb{H}$, a density of the law of $W + h$ relative to the law of W can be shown to take the form

$$\frac{dP_{W+h}}{dP_W}(W) = e^{Uh - \frac{1}{2}\|h\|_{\mathbb{H}}^2},$$

where $U\colon \mathbb{H} \to \overline{\mathrm{lin}}(W)$ is a linear isometry that extends the map defined by $U(Sb^*) = b^*(W)$, for $b^* \in \mathbb{B}^*$. The variable Uh is normally distributed, and the formula should recall the formula for the quotient of two normal densities with different means and equal variances. The formula allows us to compute the small ball probability $\mathrm{P}(\|W + h\| < \epsilon)$ of the shifted variable (a *decentered small ball probability*) in terms of the distribution of W, and leads to the following lemma (see Lemma I.27 for a proof of a stronger result).

Lemma 11.18 (Decentered small ball) *For any $h \in \mathbb{H}$ and every $\epsilon > 0$ we have*

$$\mathrm{P}\left(\|W + h\| < \epsilon\right) \geq e^{-\frac{1}{2}\|h\|_{\mathbb{H}}^2}\mathrm{P}\left(\|W\| < \epsilon\right).$$

The lemma only applies to shifts within the reproducing kernel Hilbert space, but it can be extended to general shifts by approximation. Here we restrict to shifts w_0 inside the closure of the RKHS, as otherwise a small enough ball around w_0 will have probability zero. Define the *concentration function* of W at w by

$$\varphi_w(\epsilon) = \inf_{h\in\mathbb{H}:\|h-w\|\le\epsilon} \tfrac{1}{2}\|h\|_{\mathbb{H}}^2 - \log \mathrm{P}(\|W\| < \epsilon). \tag{11.11}$$

For $w = 0$ this reduces to the small ball exponent $\varphi_0(\epsilon)$ defined in (11.10) (the infimum is achieved for $h = 0$), and it measures concentration of W at 0. The extra term if $w \neq 0$, which will be referred to as the *decentering function*, measures the decrease in mass when shifting from the origin to w. Indeed, up to constants, the concentration function is the exponent of the small ball around w, for every w in the support of W. See Lemma I.28 for a proof of the following result.

Proposition 11.19 (Small ball exponent) *For any mean zero Gaussian random element W in a separable Banach space, and any w in the closure of its RKHS, and any $\epsilon > 0$,*

$$\varphi_w(\epsilon) \le -\log \mathrm{P}(\|W - w\| < \epsilon) \le \varphi_w(\epsilon/2).$$

11.3 Posterior Contraction Rates

In this section we first give a generic result on Gaussian priors, which characterizes rates such that the prior has sufficient mass near a given "true function" w_0 and almost all its mass in a set of bounded complexity. The formulation of the result reminds of Theorems 8.9 or 8.19 (and Lemma 8.20) on posterior contraction rates. However, the result is purely in terms of the norm of the Banach space in which the prior process lives. In subsequent sections we apply the generic result to obtain rates of posterior contraction for Gaussian process priors in standard statistical settings by relating the statistically relevant norms and discrepancies to the Banach space norm.

The theorem is based on Borell's inequality and the concentration function, whence we assume that the prior is the law of a random element in an appropriate separable Banach space $(\mathbb{B}, \|\cdot\|)$. We write its RKHS as $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$, and denote by w_0 a fixed element of the Banach space, considered to be the "true" parameter, where posterior concentration is expected to occur. As the Gaussian prior is supported on the closure of the RKHS in $\mathbb{B}$, it is necessary that this parameter belongs to this closure.

Theorem 11.20 (Gaussian contraction rate) *Let W be a mean zero Gaussian random element in a separable Banach space $\mathbb{B}$ with RKHS $\mathbb{H}$ and let $w_0 \in \bar{\mathbb{H}}$. If $\epsilon_n > 0$ is such that*

$$\varphi_{w_0}(\epsilon_n) \le n\epsilon_n^2, \tag{11.12}$$

then for any $C > 1$ such that $Cn\epsilon_n^2 > \log 2$, there exists a measurable set $B_n \subset \mathbb{B}$ such that

$$\log N(3\epsilon_n, B_n, \|\cdot\|) \le 6Cn\epsilon_n^2, \tag{11.13}$$

$$\mathrm{P}(W \notin B_n) \le e^{-Cn\epsilon_n^2}, \tag{11.14}$$

$$\mathrm{P}(\|W - w_0\| < 2\epsilon_n) \ge e^{-n\epsilon_n^2}. \tag{11.15}$$

Proof Inequality (11.15) is an immediate consequence of (11.12) and Proposition 11.19. We need to prove existence of the set B_n such that the first and second inequalities in the theorem hold.

If $B_n = \epsilon_n \mathbb{B}_1 + M_n \mathbb{H}_1$, where $\mathbb{B}_1$ and $\mathbb{H}_1$ are the unit balls of $\mathbb{B}$ and $\mathbb{H}$, respectively, and M_n is a positive constant, then $\mathrm{P}(W \notin B_n) \leq 1 - \Phi(\alpha_n + M_n)$ for α_n given by $\Phi(\alpha_n) = \mathrm{P}(W \in \epsilon_n \mathbb{B}_1) = e^{-\varphi_0(\epsilon_n)}$, by Borell's inequality. Since $\varphi_0(\epsilon_n) \leq \varphi_{w_0}(\epsilon_n) \leq n\epsilon_n^2$ and $C > 1$, we have that $\alpha_n \geq -M_n/2$ if $M_n = -2\Phi^{-1}(e^{-Cn\epsilon_n^2})$. It follows that for this choice $\mathrm{P}(W \notin B_n) \leq 1 - \Phi(M_n/2) = e^{-Cn\epsilon_n^2}$.

It remains to verify the complexity estimate (11.13). If $h_1, \ldots, h_N \in M_n \mathbb{H}_1$ are $2\epsilon_n$-separated in terms of the Banach space norm $\|\cdot\|$, then the ϵ_n-balls $h_1 + \epsilon_n \mathbb{B}_1, \ldots, h_1 + \epsilon_n \mathbb{B}_1$ are disjoint and hence, by Lemma 11.18,

$$1 \geq \sum_{j=1}^{N} \mathrm{P}(W \in h_j + \epsilon_n \mathbb{B}_1) \geq \sum_{j=1}^{N} e^{-\|h_j\|_{\mathbb{H}}^2/2} \mathrm{P}(W \in \epsilon_n \mathbb{B}_1) \geq N e^{-M_n^2/2} e^{-\varphi_0(\epsilon_n)}.$$

For a maximal $2\epsilon_n$-separated set $h_1, \ldots, h_N$, the balls around $h_1, \ldots, h_N$ of radius $2\epsilon_n$ cover the set $M_n \mathbb{H}_1$ and hence we obtain the estimates $N(2\epsilon_n, M_n \mathbb{H}_1, \|\cdot\|) \leq N \leq e^{M_n^2/2} e^{\varphi_0(\epsilon_n)}$. Since any point of B_n is within ϵ_n of an element of $M_n \mathbb{H}_1$, this is also a bound on $N(3\epsilon_n, B_n, \|\cdot\|)$. To complete the proof, observe that $M_n^2/2 \leq 5Cn\epsilon_n^2$, in view of Lemma K.6. □

Ignoring multiplicative constants in the rate, we can rewrite relation (11.12) as the pair of inequalities

$$-\log \mathrm{P}(\|W\| < \epsilon_n) \leq n\epsilon_n^2, \qquad \inf_{h \in \mathbb{H}: \|h - w_0\| \leq \epsilon_n} \|h\|_{\mathbb{H}}^2 \leq n\epsilon_n^2.$$

Both inequalities have a minimal solution ϵ_n, and the final rate ϵ_n satisfying (11.12) is the maximum of the two minimal solutions, up to a constant. The first inequality in the preceding display concerns the small ball probability at 0. It depends on the prior, but not on the true parameter w_0: priors that put little mass near 0 will give slow rates ϵ_n, whatever the true parameter w_0. The second inequality measures the decrease of prior mass around the true parameter w_0 relative to the zero parameter (on a logarithmic scale). A prior that puts much mass around 0 may still give bad performance for a nonzero w_0, depending on its position relative to the RKHS. The most favorable situation is that the true parameter is contained in the RKHS: then the choice $h = w_0$ is eligible in the infimum, whence the infimum is bounded by $\|w_0\|_{\mathbb{H}}^2$, and the condition merely says that ϵ_n must not be smaller than a multiple of the "parametric rate" $n^{-1/2}$. However, the RKHS can be a very small space, and hence this favorable situation is rare.

By Proposition 11.19 the concentration function $\varphi_{w_0}(\epsilon)$ measures the prior mass around w_0. Thus the theorem shows that for Gaussian priors the rate of contraction is driven by the prior mass condition only. The existence of sieves of a prescribed complexity, required in (8.5) and (8.4) of the general rate theorems, is implied by the prior mass condition. The preceding theorem establishes this only for entropy, neighborhoods and metric of convergence all described in terms of the Banach space norm, but in the sequel we show that for the standard inference problems this can be translated into the statistically relevant quantities used in Theorems 8.9 and 8.19.

Lower Bounds

The concentration function φ_{w_0} delivers both lower and upper bounds on the concentration of a Gaussian prior. The lower bound is instrumental to verify the prior mass condition and derive an upper bound on the rate of contraction. The upper bound may be used to show that this rate of contraction is sharp, through an application of Theorem 8.35. By Lemma 11.19 the probability $\mathrm{P}(\|W - w_0\| < \epsilon)$ is sandwiched between $e^{-\varphi_{w_0}(\epsilon/2)}$ and $e^{-\varphi_{w_0}(\epsilon)}$. Therefore, inequalities (8.45) with d the metric induced by the norm are satisfied if

$$\phi_{w_0}(\epsilon_n/2) \leq C_1 n\epsilon_n^2 \leq C_2 n\epsilon_n^2 \leq \phi_{w_0}(\delta_n).$$

Then Theorem 8.35 says that δ_n is a lower bound for the contraction rate. If the concentration function is "regularly varying", then these inequalities ought to be true for δ_n, a multiple of ϵ_n. If then the Banach space norm also relates properly to the Kullback-Leibler discrepancies, the bounds on the contraction rates obtained in the following will be sharp.

11.3.1 Density Estimation

Consider estimating a probability density p relative to a σ-finite measure ν on a measurable space $(\mathfrak{X}, \mathscr{X})$, based on a sample of observations $X_1, \ldots, X_n | p \overset{\text{iid}}{\sim} p$. Construct a prior Π for p as the exponential transform of a Gaussian process $W = (W_x : x \in \mathfrak{X})$

$$p(x) = \frac{e^{W_x}}{\int_{\mathfrak{X}} e^{W_y}\, d\nu(y)}.$$

This transform was considered in Section 2.3.1, and statistical distances were seen to be controlled by the uniform norm on the exponent.

Theorem 11.21 *Let W be a mean-zero Gaussian random element in a separable subspace of $\mathfrak{L}_\infty(\mathfrak{X})$ with measurable sample paths. If $w_0 = \log p_0$ belongs to the support of W, and ϵ_n satisfies the rate equation $\varphi_{w_0}(\epsilon_n) \leq n\epsilon_n^2$, then $\Pi_n(p : d_H(p, p_0) > M\epsilon_n | X_1, \ldots, X_n) \to 0$ in P_0^n-probability, for some sufficiently large constant M. Furthermore, if $\tilde{\varphi}_{\tilde{w}_0}(\delta_n) \geq C_2 n\epsilon_n^2$, for a sufficiently large constant C_2, where $\tilde{\varphi}_{\tilde{w}_0}$ is the concentration function of the process $W - W(x_0)$ at $\tilde{w}_0 = w_0 - w_0(x_0)$ for some $x_0 \in \mathfrak{X}$, then $\Pi_n(p : \|p - p_0\|_\infty \leq m\delta_n | X_1, \ldots, X_n) \to 0$ in P_0^n-probability, for a sufficiently small constant m.*

Proof The Kullback-Leibler divergence and variation, and the square Hellinger distance between densities $p_{w_1} \propto e^{w_1}$ and $p_{w_2} \propto e^{w_2}$ are bounded by the square of the uniform norm of the difference between the exponents w, by Lemma 2.5. Therefore the prior mass, remaining mass and entropy conditions of Theorem 8.9 are verified at a multiple of ϵ_n in view of Theorem 11.20.

By the same lemma, a ball $\{w : \|p_w - p_{w_0}\|_\infty < \delta\}$ is contained in a ball $\{w : \|w - w_0\|_\infty < 2D(w, w_0)\delta\}$, if all functions w and w_0 are restricted to take the value zero at a given point $x_0 \in \mathfrak{X}$. Here $2D(w, w_0)$ is bounded above by a constant D_0 that depends only on w_0, for every w with $\|p_w - p_{w_0}\|_\infty < 1/2$. Because a constant shift in w cancels under forming the density p_w, the (prior) distributions of the

densities $p_{\tilde{W}}$ and p_W are identical. Thus $\mathrm{P}\big(\|p_W - p_{w_0}\|_\infty < \delta_n/D_0\big)$ is bounded above by $\mathrm{P}\left(\|W - W(x_0) - w_0 - w_0(x_0)\|_\infty < \delta_n\right) \le e^{-\tilde{\varphi}_{\tilde{w}_0}(\delta_n)}$, by Proposition 11.19. For δ_n chosen such that the right side is bounded above by $e^{-C_2 n\epsilon_n^2}$, the posterior mass of this set tends to zero, by Theorem 8.35. □

11.3.2 Nonparametric Binary Regression

Consider estimating a binary regression function $p(x) = \mathrm{P}(Y = 1 | X = x)$ based on an i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the distribution of (X, Y), where $Y \in \{0, 1\}$, and X takes its values in some measurable space $(\mathfrak{X}, \mathscr{X})$ following a distribution G. Given a fixed link function $\Psi: \mathbb{R} \to (0, 1)$ construct a prior for p through a Gaussian process $W = (W_x : x \in \mathfrak{X})$ by

$$p(x) = \Psi(W_x).$$

We shall assume that Ψ is strictly increasing from $\Psi(-\infty) = 0$ to $\Psi(\infty) = 1$, and differentiable with bounded derivative.

The likelihood for (X, Y) factorizes as $p(x)^y\,(1 - p(x))^{1-y}\,dG(x)$. Because the contribution of G cancels out from the posterior distribution for p, we can consider G known and need not specify a prior for it. Assume that p_0 is never zero and let $w_0 = \Psi^{-1}(p_0)$, for p_0 the true value of p.

The following theorem considers two settings: the first assumes that w_0 is bounded and the second not. In the second case the theorem is true under an additional condition on Ψ, which is satisfied by the logistic link function, but rules out the normal link. Of course, on compact domains $\mathfrak{X}$, boundedness is implied by continuity of w_0, and the first case suffices.

Theorem 11.22 *Let W be a mean-zero Gaussian random element W in a separable Banach space $\mathbb{B}$. If w_0 belongs to the support of W and ϵ_n satisfies the rate equation $\varphi_{w_0}(\epsilon_n) \le n\epsilon_n^2$, then $\Pi_n(\|p - p_0\|_{2,G} > M\epsilon_n | X_1, Y_1, \ldots, X_n, Y_n) \to 0$ in P_0^n-probability, for some $M > 0$, in the following cases:*

(i) $\mathbb{B} \subset \mathfrak{L}_\infty(\mathfrak{X})$ *and w_0 is bounded.*
(ii) $\mathbb{B} = \mathbb{L}_2(G)$ *and the function $\psi/(\Psi(1 - \Psi))$ is bounded.*

Proof This follows from combining Lemma 2.8, Theorem 8.9, and Theorem 11.20. □

11.3.3 Nonparametric Normal Regression

Consider estimating a regression function f based on observations $Y_1, \ldots, Y_n$ in the normal regression model with fixed covariates $Y_i = f(x_i) + \varepsilon_i$, where $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{Nor}(0, \sigma_0^2)$ and the covariates $x_1, \ldots, x_n$ are fixed elements from a set $\mathfrak{X}$.

A prior on f is induced by setting $f(x) = W_x$, for a Gaussian process $(W_x : x \in \mathfrak{X})$. If σ is unknown, then we also put a prior on σ, which we assume to be supported on an interval $[a, b] \subset (0, \infty)$ with a density π that is bounded away from zero.

Let $\|\cdot\|_n$ be the $\mathbb{L}_2(\mathbb{P}_n^x)$-norm for the empirical measure $\mathbb{P}_n^x$ of the design points $x_1, \ldots, x_n$, and let $\varphi_{w_0,n}$ be the concentration function of W viewed as a map in $\mathbb{L}_2(\mathbb{P}_n^x)$. The inconvenience that this depends on n can be removed by bounding the empirical norm by the uniform norm, which gives a corresponding bound on the concentration function.

Theorem 11.23 *Let W be a mean-zero Gaussian random element in $\mathbb{L}_2(\mathbb{P}_n^x)$, for every n, and suppose that the true values f_0 of f and σ_0 of σ belong to the supports of W and π. If ϵ_n satisfies the rate equation $\varphi_{w_0,n}(\epsilon_n) \leq n\epsilon_n^2$, then $\Pi_n\left((w,\sigma)\colon \|w - w_0\|_n + |\sigma - \sigma_0| > M\epsilon_n \,|\, Y_1, \ldots, Y_n\right) \to 0$ in P_0^n-probability, for some $M > 0$. Furthermore, if $\varphi_{w_0,n}(\delta_n) \geq C_2 n\epsilon_n^2$ for a sufficiently large constant C_2, then $\Pi_n\left(w\colon \|w - w_0\|_n \leq m\delta_n \,|\, Y_1, \ldots, Y_n\right) \to 0$ in P_0^n-probability, for sufficiently small $m > 0$.*

Proof It is noted in Section 8.3.2 that the Kullback-Leibler divergence and variation are equivalent to the empirical squared distance $\|\cdot\|_{n,2}^2$. Furthermore, likelihood ratio tests have the appropriate properties relative to this distance. Since σ is restricted to a compact interval containing the truth and its prior is bounded away from zero, it plays a minor role in the rate of contraction. Thus the upper bound of the theorem follows from combining an extension of Theorem 8.26 to include σ and Theorem 11.20.

The lower bound follows from combining the lower bounds from Theorem 8.35 and Proposition 11.19. □

11.3.4 White Noise Model

Consider estimation of the signal θ based on the observation $X^{(n)} = (X_t^{(n)}\colon 0 \leq t \leq 1)$ in the white noise model described in Section 8.3.4. As a prior on θ we take a Gaussian random element W in $\mathbb{L}_2[0,1]$, which is conjugate with the Gaussian likelihood.

Theorem 11.24 *Let W be a mean zero Gaussian random element in $\mathbb{L}_2[0,1]$. If the true value θ_0 of θ is contained in the support of W, and ϵ_n satisfies the rate equation $\varphi_{\theta_0}(\epsilon_n) \leq n\epsilon_n^2$ with respect to the $\mathbb{L}_2[0,1]$-norm, then $\Pi_n\left(\|\theta - \theta_0\|_2 > M\epsilon_n \,|\, X^{(n)}\right) \to 0$ in P_0^n-probability, for some $M > 0$. Furthermore, if $\varphi_{w_0}(\delta_n) \geq C_2 n\epsilon_n^2$ for a sufficiently large constant C_2, then $\Pi_n\left(\|\theta - \theta_0\|_2 \leq m\delta_n \,|\, X^{(n)}\right) \to 0$ in P_0^n-probability, for sufficiently small $m > 0$.*

Proof The upper bound follows from combining Theorems 8.31 and 11.20; the lower bound from combining Theorem 8.35 and Proposition 11.19. □

11.4 Specific Gaussian Processes as Priors

In this section we calculate posterior contraction rates associated with specific Gaussian process priors. In each example this involves an estimate of the small ball exponent $\varphi_0(\epsilon)$, a characterization of the RKHS of the process, and the approximation of a true function w_0 by elements of the RKHS. The small ball exponent can be calculated by a probabilistic method,

but can also be obtained from the metric entropy of the unit ball $\mathbb{H}_1$ of the RKHS. Under regularity conditions it is true that, as $\epsilon \downarrow 0$,[4]

$$\varphi_0(\epsilon) \asymp \log N\Big(\frac{\epsilon}{\sqrt{2\varphi_0(\epsilon)}}, \mathbb{H}_1, \|\cdot\|_{\mathbb{B}}\Big).$$

Given an upper bound for the covering number $N(\epsilon, \mathbb{H}_1, \|\cdot\|_{\mathbb{B}})$, this may be solved for the small ball probability. The following lemma gives the solution in a common case.

Lemma 11.25 (Small ball and entropy) *For $\alpha > 0$ and $\beta \in \mathbb{R}$, as $\epsilon \downarrow 0$, $\varphi_0(\epsilon) \asymp \epsilon^{-\alpha}(\log_- \epsilon)^{\beta}$ if and only if $\log N(\epsilon, \mathbb{H}_1, \|\cdot\|) \asymp \epsilon^{-2\alpha/(2+\alpha)}(\log_- \epsilon)^{2\beta/(2+\alpha)}$.*

11.4.1 Brownian Motion and Its Primitives

Brownian motion B is a good starting point for modeling functions on $[0, 1]$. The RKHS and small ball probabilities of Brownian motion are classical. Let $\mathfrak{W}^k[0, 1]$ be the *Sobolev space* of all functions $f \in \mathbb{L}_2[0, 1]$ that are $k-1$ times differentiable with $f^{(k-1)}$ absolutely continuous with derivative $f^{(k)}$ belonging to $\mathbb{L}_2[0, 1]$ (see Definition C.6).

Lemma 11.26 (RKHS) *The RKHS of Brownian motion is equal to $\{f \in \mathfrak{W}^1[0, 1]: f(0) = 0\}$ with the inner product $\langle f, g\rangle_{\mathbb{H}} = \int_0^1 f'(t)g'(t)\,dt$.*

Lemma 11.27 (Small ball) *The small ball exponent of Brownian motion viewed as a map into $\mathfrak{C}[0, 1]$ or $\mathbb{L}_r[0, 1]$, for some $r \geq 1$, satisfies, as $\epsilon \downarrow 0$,*

$$\varphi_0(\epsilon) \asymp \epsilon^{-2}.$$

Proof For a proof of the second lemma apply Lemma 11.25 together with the entropy estimate of the unit ball of the RKHS given by Proposition C.7, or see Li and Shao (2001).

Since $K(s, t) = s \wedge t$, the RKHS is equal to the completion of the linear span of the functions $\{t \mapsto s \wedge t: s \in [0, 1]\}$ under the inner product determined by

$$\langle s_1 \wedge \cdot, s_2 \wedge \cdot\rangle_{\mathbb{H}} = s_1 \wedge s_2 = \int (s_1 \wedge t)'(s_2 \wedge t)'\,dt.$$

Here $t \mapsto (s \wedge t)' = \mathbb{1}\{[0, s]\}(t)$ is the Radon-Nikodym derivative of the function $s \wedge \cdot$. The equation shows that the RKHS inner product is indeed the inner product of $\mathfrak{W}^1[0, 1]$. It suffices to show that the linear span of all functions of this type is dense in $\{f \in \mathfrak{W}^1[0, 1]: f(0) = 0\}$. Now the linear span contains every function that is 0 at 0, continuous, and piecewise linear on a partition $0 = s_0 < s_1 \cdots < s_N = 1$, since such a function with slopes α_j on the intervals (s_{j-1}, s_j), for $j = 1, \ldots, N$, can be constructed as a linear combination by first determining the coefficient of $(s_N \wedge \cdot)$ to have the correct slope on (s_{N-1}, s_N), next determining the coefficient of $(s_{N-1} \wedge \cdot)$ to have the correct slope on (s_{N-2}, s_{N-1}), etc. The derivatives of these piecewise linear functions are piecewise constant, and the set of piecewise constant functions is dense in $\mathbb{L}_2[0, 1]$. Thus the completion of the linear span is as claimed. □

[4] See Lemma I.29 for a precise statement.

Standard Brownian motion is zero at zero, and hence for use as a prior it is preferable to "release it at zero," by adding a variable that gives a prior for the unknown function at zero. Adding an independent Gaussian variable $Z \sim \mathrm{Nor}(0, 1)$ yields another Gaussian process of the form $t \mapsto Z + B_t$. The RKHS of the constant process $t \mapsto Z$ consists of the constant functions, and general rules for the formation of reproducing Hilbert spaces (see Lemma I.18) shows that the RKHS of $Z + B$ is the direct sum of the constant functions and the RKHS of B. In other words, the point zero is also "released" in the RKHS: the RKHS is equal to the Sobolev space $\mathfrak{W}^1[0, 1]$ with inner product

$$\langle f, g \rangle_{\mathbb{H}} = f(0)g(0) + \int_0^1 f'(s)g'(s)\, ds.$$

The addition of the single variable Z makes the small ball probability at 0 smaller, but it does not change the rate ϵ^{-2} obtained in the preceding lemma, as $-\log \mathrm{P}(|Z| < \epsilon) \asymp \log_- \epsilon \ll \epsilon^{-2}$.

The concentration function $\varphi_{w_0}(\epsilon)$ of $Z + B$ depends on the position of the true parameter w_0 relative to the RKHS. It can be computed using a kernel smoother $w_0 * \psi_\sigma$, which is contained in the RKHS if ψ is a smooth kernel.

Lemma 11.28 (Decentering) *If $w_0 \in \mathfrak{C}^\beta[0, 1]$ for some $\beta \in (0, 1]$, then the decentering of the concentration function of Brownian motion released at zero viewed as map in $\mathfrak{C}[0, 1]$ satisfies, for $\epsilon \downarrow 0$,*

$$\inf_{h:\|h-w_0\|_\infty<\epsilon} \|h'\|_2^2 \lesssim \epsilon^{-(2-2\beta)/\beta}.$$

Proof If ψ_σ is the density of the $\mathrm{Nor}(0, \sigma^2)$-distribution, then $\|w_0 * \psi_\sigma - w_0\| \lesssim \sigma^\beta$, as $\sigma \to 0$, and the squared RKHS-norm of $w_0 * \psi_\sigma$ is given by $(w_0 * \psi_\sigma)(0)^2 + \|(w_0 * \psi_\sigma)'\|_2^2 \asymp \sigma^{-(2-2\beta)}$. Choosing $\sigma \asymp \epsilon^{1/\beta}$, we obtain the assertion. □

Combining the preceding we see that the concentration function of released Brownian motion satisfies, if $w_0 \in \mathfrak{C}^\beta[0, 1]$,

$$\varphi_{w_0}(\epsilon) \lesssim \epsilon^{-(2-2\beta)/\beta} + \epsilon^{-2}.$$

For $\beta \geq 1/2$, the second term ϵ^{-2} dominates, and hence the minimal solution to the rate equation $\varphi_{w_0}(\epsilon_n) \leq n\epsilon_n^2$ satisfies $\epsilon_n^{-2} \asymp n\epsilon_n^2$, or $\epsilon_n \asymp n^{-1/4}$. For $\beta \in (0, 1/2)$, the first term dominates, leading to $\epsilon_n^{-(2-2\beta)/\beta} \lesssim n\epsilon_n^2$, with minimal solution $\epsilon_n \asymp n^{-\beta/2}$.

The resulting contraction rate can be summarized as $n^{-(\beta \wedge 1/2)/2}$. It is equal to the minimax rate of estimation $n^{-\beta/(2\beta+1)}$ for functions in a Hölder space of order β if and only if $\beta = 1/2$. This is intuitively understandable, as the sample paths of a Brownian motion are regular of that order: only matching of prior and true smoothness yields optimal results. For any positive $\beta \neq 1/2$ the posterior distribution is consistent, but the performance of the Brownian motion prior is suboptimal. The discrepancy is most felt for smooth w_0: the contraction rate is $n^{-1/4}$, no matter how smooth w_0 is. Technically this stems from the small ball probability of Brownian motion. This prior places a tiny fraction $\exp(-C\epsilon^{-2})$ of its mass in a ball of radius ϵ around the zero function, which may be viewed as the smoothest function of all. No amount of data will fully wash out this prior preference for non-smooth functions.

This shortcoming of Brownian motion as a prior for smooth functions can be remedied by integrating its sample paths. The k-fold integrated Brownian motion $I_{0+}^k B$ is smooth of order (nearly) $k+1/2$. Its k vanishing derivatives at zero can be released by adding a random polynomial, yielding the process

$$W_t = \sum_{i=0}^{k} Z_i \frac{t^i}{i!} + (I_{0+}^k B)_t, \qquad Z_0, \ldots, Z_k \overset{\text{iid}}{\sim} \text{Nor}(0,1) \perp\!\!\!\perp B. \tag{11.16}$$

Lemma 11.29 (RKHS) *The RKHS of the process W given in* (11.16)*, for B a Brownian motion independent of $Z_1, \ldots, Z_k \overset{iid}{\sim} \text{Nor}(0,1)$, is the Sobolev space $\mathfrak{W}^{k+1}[0,1]$, with inner product*

$$\langle f, g \rangle_{\mathbb{H}} = \sum_{i=0}^{k} f^{(i)}(0) g^{(i)}(0) + \int_0^1 f^{(k+1)}(s) g^{(k+1)}(s)\, ds.$$

Lemma 11.30 (Small ball) *The small ball exponents of k-fold integrated Brownian motion $I_{0+}^k B$ and the process W given in* (11.16)*, for B a Brownian motion independent of $Z_1, \ldots, Z_k \overset{iid}{\sim} \text{Nor}(0,1)$, viewed as maps into $\mathfrak{C}[0,1]$ satisfy, as $\epsilon \downarrow 0$,*

$$\varphi_0(\epsilon) \asymp \epsilon^{-2/(2k+1)}.$$

Proof The RKHS of $I_{0+}^k B$ can be deduced from the RKHS of Brownian motion and the general principle for the transformation of a RKHS under a continuous, linear transformation given in Lemma I.16. Because I_{0+}^k is a continuous, linear, one-to-one map from $\mathfrak{C}[0,1] \to \mathfrak{C}[0,1]$, the RKHS of $I_{0+}^k B$ is given by $\mathbb{H} = \{I_{0+}^k f : f \in \mathbb{H}_B\}$, where $\mathbb{H}_B$ is the RKHS of the Brownian motion given in Lemma 11.26, and the inner product satisfies $\langle I_{0+}^k f, I_{0+}^k g \rangle_{\mathbb{H}} = \int_0^1 f'(s) g'(s)\, ds$. In other words, the RKHS is the subset of the Sobolev space $\mathfrak{W}^{k+1}[0,1]$ of functions f with $f(0) = \cdots = f^{(k)}(0) = 0$, under its natural inner product.

The RKHS of the process W can likewise be obtained by applying a general result, which gives the RKHS of a sum of independent Gaussian random elements (see Lemma I.18). Since the supports of the polynomial process $t \mapsto \sum_{i=0}^k Z_i t^i / i!$ and $I_{0+}^k B$ in $\mathfrak{C}[0,1]$ intersect nontrivially, the lemma does not apply directly. However, we may also consider these processes as Borel measurable random elements in the space $\mathfrak{C}^k[0,1]$. The RKHSs remain the same in view of Lemma I.17. Since the only kth degree polynomial with k vanishing derivatives at zero is the zero function, the supports of the two processes in $\mathfrak{C}^k[0,1]$ have trivial intersection, and hence the RKHS of W is the direct sum of the RKHSs of the polynomial process and $I_{0+}^k B$, by Lemma I.18. The first process is a finite series and hence its RKHS are the polynomials with norm the Euclidean norm of the coefficients. The norm can alternatively be written in terms of the derivates at zero; see Example I.24.

The small ball probability is obtained in Li and Linde (1998), or follows from Lemma I.29 together with the entropy estimate of the unit ball of the RKHS given in Proposition C.7. □

Lemma 11.31 (Decentering) *If $w_0 \in \mathfrak{C}^\beta[0,1]$ for $\beta \le k+1$, then the decentering function of the process in* (11.16)*, for B a Brownian motion independent of $Z_1, \ldots, Z_k \overset{iid}{\sim} \mathrm{Nor}(0,1)$, viewed as map in $\mathfrak{C}[0,1]$ satisfies, as $\epsilon \downarrow 0$,*

$$\inf_{h::\|h-w_0\|_\infty<\epsilon} \|h\|_{\mathbb{H}}^2 \lesssim \epsilon^{-(2k-2\beta+2)/\beta}.$$

Proof The convolution $w_0 * \psi_\sigma$, where ψ_σ is the scaled version of a smooth kth order kernel (an integrable function ψ satisfying $\int \psi(t)\,dt = 1$ and $\int t^r \psi(t)\,dt = 0$ for $r = 1, \ldots, k$, and $\int |t|^{k+1}\psi(t)\,dt < \infty$), satisfies $\|w_0 * \psi_\sigma - w_0\|_\infty \lesssim \sigma^\beta$, as follows by the well known estimates from the literature on kernel density estimation. The function $w_0 * \psi_\sigma$ belongs to the RKHS and satisfies $(w_0 * \psi_\sigma)^{(l)} = w_0^{(\beta)} * \psi_\sigma^{(l-\beta)}$, for $\underline{\beta}$ the largest integer strictly smaller than β. Hence $\|(w_0 * \psi_\sigma)^{(l)}\|_\infty \lesssim \sigma^{-(l-\beta)}$ if $w_0 \in \mathfrak{C}^\beta[0,1]$ and $l \ge \beta$. Consequently $\int (w_0 * \psi_\sigma)^{(k+1)}(t)^2\,dt \lesssim \sigma^{-(2k-2\beta+2)}$ and the square derivatives at 0 up to order k are bounded or of smaller order in $1/\sigma$. It follows that $\|w_0 * \psi_\sigma\|_{\mathbb{H}} \lesssim \sigma^{-(k-\beta+1)}$, if $w_0 \in \mathfrak{C}^\beta[0,1]$. The choice $\sigma \asymp \epsilon^{1/\beta}$ leads to $\|w_0 * \psi_\sigma - w_0\|_\infty \lesssim \epsilon$ and $\|w_0 * \psi_\sigma\|_{\mathbb{H}}^2 \lesssim \epsilon^{-(2k-2\beta+2)/\beta}$. □

It follows that the concentration function of "released integrated Brownian motion" (11.16) takes the form

$$\varphi_{w_0}(\epsilon) \lesssim \epsilon^{-(2k-2\beta+2)/\beta} + \epsilon^{-2/(2k+1)}.$$

For $\beta \ge k + 1/2$ the second term dominates, and the rate inequality becomes $\epsilon_n^{-2/(2k+1)} \le n\epsilon_n^2$, with as minimal solution $\epsilon_n \asymp n^{-(2k+1)/(4k+4)}$. For $\beta \le k + 1/2$ the first term in the concentration function dominates, and the rate inequality $\epsilon_n^{-(2k-2\beta+2)/\beta} \lesssim n\epsilon_n^2$ has the minimal solution $\epsilon_n \asymp n^{-\beta/(2k+2)}$. The posterior contraction rate can be summarized as the maximum $n^{-(\beta\wedge k+1/2)/(2k+2)}$ of the two rates. This is the minimax rate for β-regular functions if and only if $\beta = k + 1/2$.

11.4.2 Riemann-Liouville Process

The Riemann-Liouville process with Hurst parameter $\alpha > 0$, defined in Example 11.6, is a random element in $\mathfrak{C}[0,1]$. We add an independent Gaussian polynomial of degree the smallest integer $\lfloor\alpha\rfloor + 1$ strictly greater than α to "release" it at zero, and consider, with $Z_1, \ldots, Z_n \overset{\mathrm{iid}}{\sim} \mathrm{Nor}(0,1)$ independent of R^α,

$$W_t = \sum_{k=0}^{\lfloor\alpha\rfloor+1} Z_k t^k + R_t^\alpha. \tag{11.17}$$

Let I_{0+}^α be the fractional integral, as defined in (11.1).

Lemma 11.32 (RKHS) *The RKHS of the Riemann-Liouville process of order α is the space $I_{0+}^{\alpha+1/2}(\mathbb{L}_2[0,1])$ with the inner product*

$$\langle I_{0+}^{\alpha+1/2} f, I_{0+}^{\alpha+1/2} g\rangle_{\mathbb{H}} = \langle f, g\rangle_2.$$

Lemma 11.33 (Small ball) *The small ball exponent of the Riemann-Liouville process released at zero viewed as a map into* $\mathfrak{C}[0,1]$*, satisfies, as* $\epsilon \downarrow 0$*,*

$$\varphi_0(\epsilon) \asymp \epsilon^{-1/\alpha}.$$

Proof The Riemann-Liouville process is $R_t^\alpha = \int_0^t f_t(u)\, dB_u$, for the function $f_t(u) = (t-u)_+^{\alpha-1/2}/\Gamma(\alpha+1/2)$. The isometry property of stochastic integrals gives that

$$\mathrm{E}(R_s^\alpha R_t^\alpha) = \langle f_s, f_t\rangle_2 = I_{0+}^{\alpha+1/2} f_s(t).$$

It follows that the function $I_{0+}^{\alpha+1/2} f_s$ is contained in the RKHS, for every $s \in [0,1]$, and has RKHS norm $\|f_s\|_2$. By definition the RKHS is the completion of the linear span of these functions under this norm. It suffices to show that the linear span of the functions f_s is dense in $\mathbb{L}_2[0,1]$. Now $f \perp f_s$, for every s, implies that $I_{0+}^{\alpha-1/2} f(t) = \int f(u)(t-u)^{\alpha-1/2}\, du = 0$, and this implies that $f = 0$, by the injectivity of the operator $I_{0+}^{\alpha-1/2}\colon \mathbb{L}_2[0,1] \to \mathbb{L}_2[0,1]$.

The small ball estimate follows from Theorem 2.1 of Li and Linde (1998), or a calculation on the entropy of the unit ball of the RKHS. □

Lemma 11.34 (Decentering) *If* $w_0 \in \mathfrak{C}^\beta[0,1]$*, then the concentration function of the Riemann-Liouville process released at zero viewed as a map into* $\mathfrak{C}[0,1]$*, satisfies, as* $\epsilon \downarrow 0$*,*

$$\varphi_{w_0}(\epsilon) \lesssim \begin{cases} \epsilon^{-1/\alpha}, & \text{if } 0 < \alpha \le \beta, \\ \epsilon^{-(2\alpha-2\beta+1)/\beta}, & \text{if } \alpha > \beta \text{ and } (\langle\alpha\rangle = \tfrac12 \text{ or } \alpha \notin \beta + \tfrac12 + \mathbb{N}), \\ \epsilon^{-(2\alpha-2\beta+1)/\beta} \log_- \epsilon, & \text{otherwise.} \end{cases}$$

Proof For $\alpha < \beta$ the small ball probability dominates the concentration function. For $\alpha \ge \beta$, the proof must use properties of fractional integrals, which makes it technical. We refer to van der Vaart and van Zanten (2008a) or Castillo (2008) for details. □

In view of the preceding lemma the rate equation $\varphi_{w_0}(\epsilon_n) \le n\epsilon_n^2$ is satisfied by $\epsilon_n \ge n^{-(\alpha\wedge\beta)/(2\alpha+1)}$ if either $\langle\alpha\rangle = \frac12$ or $\alpha - \beta - \frac12 \notin \mathbb{N}$, and by $\epsilon_n \gtrsim n^{-\beta/(2\alpha+1)} \log n$ otherwise. Up to the logarithmic factor in the last case, this is the minimax rate for β-regular functions if and only if $\alpha = \beta$. (The extra $\log n$-factor in the case that $\langle\alpha\rangle \ne \frac12$ or $\alpha - \beta - \frac12 \in \mathbb{N}$ appears, because the fractional integral I_{0+}^α maps $\mathfrak{C}^\lambda[0,1] \to \mathfrak{C}^{\lambda+\alpha}[0,1]$ only if $\alpha + \lambda \ne 1$, and appears not to be an artifact of the method of proof.)

11.4.3 Fractional Brownian Motion

The fractional Brownian motion fBm^α of Hurst index α is related to the Riemann-Liouville process R^α as $\mathrm{fBm}_t^\alpha = c_\alpha R_t^\alpha + c_\alpha Z_t$, where $c_\alpha > 0$ is a constant, R^α and Z are independent Gaussian processes in $\mathfrak{C}[0,1]$, and Z possesses small ball probability, as $\epsilon \downarrow 0$,

$$\varphi_0(\epsilon; Z) = -\log \mathrm{P}(\|Z\| < \epsilon) = o(\epsilon^{-1/\alpha}). \tag{11.18}$$

(See Mandelbrot and Van Ness 1968 and Lemma 3.2 of Li and Linde 1998.) Because this is of smaller order than the small ball probability of the Riemann-Liouville process, the concentration functions φ_{w_0} of fBm^α and R^α behave similarly (for any w_0), and hence the results on the Riemann-Liouville process extend to the fractional Brownian motion.

11.4.4 Stationary Processes

The RKHS of a stationary Gaussian process $(W_t: t \in T)$ indexed by a subset $T \subset \mathbb{R}^d$, as in Example 11.8, can be characterized in terms of its spectral measure μ.

For $t \in T$ let $e_t: \mathbb{R}^d \to \mathbb{C}$ denote the function $e_t(\lambda) = e^{i\lambda^\top t}$, and for $\psi \in \mathbb{L}_2(\mu)$, let $H\psi: T \to \mathbb{R}$ be the function defined by $(H\psi)(t) = \int e_t \psi \, d\mu$. If μ has Lebesgue density m, then this function is the Fourier transform of the (integrable) function ψm.

Lemma 11.35 (RKHS) *The RKHS of the stationary Gaussian process with spectral measure μ is the set of functions $H\psi$ for ψ ranging over $\mathbb{L}_2(\mu)$, with RKHS-norm equal to $\|H\psi\|_{\mathbb{H}} = \|P\psi\|_{2,\mu}$, where P is the projection onto the closed linear span of the set of functions $(e_t: t \in T)$ in $\mathbb{L}_2(\mu)$. If $T \subset \mathbb{R}^d$ has an interior point and $\int e^{\delta\|\lambda\|} \, d\mu(\lambda) < \infty$ for some $\delta > 0$, then this closed linear span is $\mathbb{L}_2(\mu)$ and the RKHS norm is $\|H\psi\|_{\mathbb{H}} = \|\psi\|_{2,\mu}$.*

Proof The spectral representation (11.3) can be written $\mathrm{E} W_s W_t = \langle e_t, e_s \rangle_{\mu,2}$. By definition the RKHS is therefore the set of functions $H\psi$ with ψ running through the closure $\mathbb{L}_T$ in $\mathbb{L}_2(\mu)$ of the linear span of the set of functions $(e_s: s \in T)$, and the norm equal to the norm of ψ in $\mathbb{L}_T \subset \mathbb{L}_2(\mu)$. Here the "linear span" is taken over the reals. If instead we take the linear span over the complex numbers, we obtain complex functions whose real parts give the RKHS.

The set of functions obtained by letting ψ range over the full space $\mathbb{L}_2(\mu)$ rather than $\mathbb{L}_T$ is precisely the same, as a general element $\psi \in \mathbb{L}_2(\mu)$ gives exactly the same function as its projection $P\psi$ onto $\mathbb{L}_T$. However, the associated norm is the $\mathbb{L}_2(\mu)$ norm of $P\psi$. This proves the first assertion of the lemma. For the second we must show that $\mathbb{L}_T = \mathbb{L}_2(\mu)$ under the additional conditions.

The partial derivative of order $(k_1, \ldots, k_d)$ with respect to $(t_1, \ldots, t_d)$ of the map $t \mapsto e_t$ at t_0 is the function $\lambda \mapsto (i\lambda_1)^{k_1} \cdots (i\lambda_d)^{k_d} e_{t_0}(\lambda)$. Appealing to the dominated convergence theorem, we see that this derivative exists as a derivative in $\mathbb{L}_2(\mu)$. Because t_0 is an interior point of T by assumption, we conclude that the function $\lambda \mapsto (i\lambda)^k e_{t_0}(\lambda)$ belongs to $\mathbb{L}_T$ for any multi-index k of nonnegative integers. Consequently, the function $p e_{t_0}$ belongs to $\mathbb{L}_T$ for any polynomial $p: \mathbb{R}^d \to \mathbb{C}$ in d arguments. It suffices to show that these functions are dense in $\mathbb{L}_2(\mu)$.

Equivalently, it suffices to prove that the polynomials themselves are dense in $\mathbb{L}_2(\mu)$. Indeed, if $\psi \in \mathbb{L}_2(\mu)$ is orthogonal to all functions of the form $p e_{t_0}$, then $\psi \overline{e_{t_0}}$ is orthogonal to all polynomials. Denseness of the set of polynomials then gives that $\psi \overline{e_{t_0}}$ vanishes μ-almost everywhere, whence ψ vanishes μ-almost everywhere.

Finally we prove that the polynomials are dense in $\mathbb{L}_2(\mu)$ if μ has an exponential moment. Suppose that $\psi \in \mathbb{L}_2(\mu)$ is orthogonal to all polynomials. Since μ is a finite measure, the complex conjugate $\overline{\psi}$ is μ-integrable, and hence we can define a complex measure ν by

$$\nu(B) = \int_B \overline{\psi(\lambda)} \, \mu(d\lambda).$$

It suffices to show that ν is the zero measure, so that $\psi = 0$ almost everywhere relative to μ.

By the Cauchy-Schwarz inequality and the assumed exponential integrability of the spectral measure, with $|\nu|$ the (total) variation measure of ν, $\int e^{\delta\|\lambda\|/2} \, |\nu|(d\lambda) < \infty$. By a

standard argument, based on the dominated convergence theorem (see e.g. Bauer 2001, Theorem 8.3.5), this implies that the function $z \mapsto \int e^{\langle \lambda, z\rangle}\, \nu(d\lambda)$ is analytic on the strip $\Omega = \{z \in \mathbb{C}^d : |\operatorname{Re} z_1| < \delta/(2\sqrt{d}), \ldots, |\operatorname{Re} z_d| < \delta/(2\sqrt{d})\}$. Also for z real and in this strip, by the dominated convergence theorem,

$$\int e^{\langle \lambda, z\rangle}\, \nu(d\lambda) = \int \sum_{n=0}^{\infty} \frac{\langle \lambda, z\rangle^n}{n!}\, \nu(d\lambda) = \sum_{n=0}^{\infty} \int \frac{\langle \lambda, z\rangle^n}{n!}\, \overline{\psi}(\lambda)\, \mu(d\lambda).$$

The right side vanishes, because ψ is orthogonal to all polynomials by assumption.

We conclude that the function $z \mapsto \int e^{(\lambda, z)}\, \nu(d\lambda)$ vanishes on the set $\{z \in \Omega : \operatorname{Im} z = 0\}$. Because this set contains a nontrivial interval in $\mathbb{R}$ for every coordinate, we can apply (repeated) analytic continuation to see that this function vanishes on the complete strip Ω. In particular the Fourier transform $t \mapsto \int e^{i\lambda^\top t}\, \nu(d\lambda)$ of ν vanishes on all of $\mathbb{R}^d$, whence ν is the zero-measure. □

The functions $H\psi$ in the RKHS are always uniformly continuous, as ψ is μ-integrable. Their exact regularity is determined by the tails of the spectral measure. If these are light, then $H\psi$ is a mixture of mainly low frequency trigonometric functions, and hence smooth.

To characterize the small ball probability and concentration function we must specialize to particular spectral measures. We consider the Matérn and square exponential measures in the following subsections.

Matérn Process

The spectral measure of the Matérn process is given in (11.5). Its polynomial tails make the sample paths of the process regular of finite order. The small ball exponent of the process is similar to that of the Riemann-Liouville process of the same regularity.

Lemma 11.36 (Small ball) *The small ball exponent of the Matérn process W viewed as a map in $\mathfrak{C}[0, 1]$ satisfies, as $\epsilon \downarrow 0$,*

$$\varphi_0(\epsilon) \lesssim \epsilon^{-d/\alpha}.$$

Proof The Fourier transform of $H\psi$ is, up to a constant, the function $\phi = \psi m$, if $d\mu(\lambda) = m\, d\lambda$. For ψ the choice of minimal norm in the definition of $H\psi$, this function satisfies

$$\int |\phi(\lambda)|^2 (1 + \|\lambda\|^2)^{\alpha + d/2}\, d\lambda = \|H\psi\|_{\mathbb{H}}^2.$$

In other words, the unit ball $\mathbb{H}_1$ of the RKHS is contained in a Sobolev ball of order $\alpha + d/2$. The metric entropy relative to the uniform norm of such a Sobolev ball is bounded by a constant times $\epsilon^{-d/(\alpha + d/2)}$, by Proposition C.7. The lemma next follows from Lemma 11.25, which characterizes the small ball probability in terms of the entropy of the RKHS-unit ball. □

To estimate the infimum in the definition of the concentration function φ_{w_0} for a nonzero response function w_0, we approximate w_0 by elements of the RKHS. The idea is to write w_0 in terms of its Fourier inverse $\hat{w}_0$ as, with $d\mu(\lambda) = m(\lambda)\, d\lambda$,

$$w_0(x) = \int e^{i\lambda^\top x}\hat{w}_0(\lambda)\,d\lambda = \int e^{i\lambda^\top x}\frac{\hat{w}_0}{m}(\lambda)\,d\mu(\lambda). \tag{11.19}$$

If $\hat{w}_0/m$ were contained in $\mathbb{L}_2(\mu)$, then w_0 would be contained in the RKHS, with RKHS-norm bounded by the $\mathbb{L}_2(\mu)$-norm of $\hat{w}_0/m$, i.e. the square root of $\int(|\hat{w}_0|^2/m)(\lambda)\,d\lambda$. In general this integral may be infinite, but we can remedy this by truncating the tails of $\hat{w}_0/m$.

A natural a priori condition on the true response function $w_0\colon [0,1]^d \to \mathbb{R}$ is that this function is contained in a Sobolev space of order β. The *Sobolev space* $\mathfrak{W}^\alpha[0,1]^d$ is here defined as the set of functions $w\colon [0,1]^d \to \mathbb{R}$ that are restrictions of a function $w\colon \mathbb{R}^d \to \mathbb{R}$ with Fourier transform $\hat{w}(\lambda) = (2\pi)^{-d}\int e^{i\lambda^\top t}w(t)\,dt$ such that

$$\|w\|_{2,2,\alpha}^2 := \int (1+\|\lambda\|^2)^\alpha |\hat{w}(\lambda)|^2\,d\lambda < \infty.$$

Roughly speaking, for integer α, a function belongs to $\mathfrak{W}^\alpha([0,1]^d)$ if it has partial derivatives up to order α that are all square integrable. This follows, because the αth derivative of a function w has Fourier transform $\lambda \mapsto (i\lambda)^\alpha\hat{w}(\lambda)$,

Lemma 11.37 (Decentering) *If $w_0 \in \mathfrak{C}^\beta([0,1]^d) \cap \mathfrak{W}^\beta([0,1]^d)$ for $\beta \le \alpha$, then the decentering function of the Matérn process satisfies, for $\epsilon < 1$,*

$$\inf_{h:\|h-w_0\|_\infty<\epsilon} \|h\|_{\mathbb{H}}^2 \lesssim \epsilon^{-(2\alpha+d-2\beta)/\beta}.$$

Proof Let $\kappa\colon \mathbb{R} \to \mathbb{R}$ be a function with a real, symmetric Fourier transform $\hat{\kappa}$, which equals $1/(2\pi)$ in a neighborhood of 0 and which has compact support. From $\hat{\kappa}(\lambda) = (2\pi)^{-1}\int e^{i\lambda t}\kappa(t)\,dt$ it then follows that $\int \kappa(t)\,dt = 1$ and $\int (it)^k\kappa(t)\,dt = 0$ for $k \ge 1$. For $t = (t_1,\ldots,t_d)$, define $\phi(t) = \kappa(t_1)\cdots\kappa(t_d)$. Then ϕ integrates to 1, has finite absolute moments of all orders, and vanishing moments of all orders bigger than 0.

For $\sigma > 0$ set $\phi_\sigma(x) = \sigma^{-d}\phi(x/\sigma)$ and $h = \phi_\sigma * w_0$. Because ϕ is a higher order kernel, standard arguments from the theory of kernel estimation show that $\|w_0 - \phi_\sigma * w_0\|_\infty \lesssim \sigma^\beta$.

The Fourier transform of h is the function $\lambda \mapsto \hat{h}(\lambda) = \hat{\phi}(\sigma\lambda)\hat{w}_0(\lambda)$, and therefore, by (11.19),

$$\begin{aligned}\|h\|_{\mathbb{H}}^2 &\lesssim \int |\hat{\phi}(\sigma\lambda)\hat{w}_0(\lambda)|^2\frac{1}{m(\lambda)}\,d\lambda \lesssim \sup_\lambda\Big[(1+\|\lambda\|^2)^{\alpha+d/2-\beta}|\hat{\phi}(\sigma\lambda)|^2\Big]\|w_0\|_{2,2,\beta}^2 \\ &\lesssim C(\sigma)\sup_\lambda\Big[(1+\|\lambda\|^2)^{\alpha+d/2-\beta}|\hat{\phi}(\lambda)|^2\Big]\|w_0\|_{2,2,\beta}^2,\end{aligned}$$

for

$$C(\sigma) = \sup_\lambda\Big(\frac{1+\|\lambda\|^2}{1+\|\sigma\lambda\|^2}\Big)^{\alpha+d/2-\beta} \lesssim \Big(\frac{1}{\sigma}\Big)^{2\alpha+d-2\beta},$$

if $\sigma \le 1$. The assertion of the lemma follows upon choosing $\sigma \sim \epsilon^{1/\beta}$. □

It follows that the rate equation $\varphi_{w_0}(\epsilon_n) \le n\epsilon_n^2$ has the minimal solution $\epsilon_n \asymp n^{-\beta/(2\alpha+d)}$, for $\beta \le \alpha$. Thus again the rate is minimax if and only if prior and smoothness match.

Square Exponential Process

The spectral measure of the *square exponential process* is the Gaussian measure given in (11.4). The analytic sample paths of the process make the small ball probability of this process much larger than that of the processes considered so far: it is nearly "parametric."

Lemma 11.38 (Small ball) *The small ball exponent of the square exponential process viewed as a map in* $\mathfrak{C}([0,1]^d)$ *satisfies, for a constant* C *depending only on* d*, as* $\epsilon \downarrow 0$*,*

$$\varphi_0(\epsilon) \le C(\log_- \epsilon)^{1+d/2}.$$

Proof Every function $H\psi$ in the (complex) RKHS as described in Lemma 11.35 can be extended to an entire function $H\psi\colon \mathbb{C} \to \mathbb{C}$ defined by $H\psi(z) = \int e^{\langle \lambda, z\rangle}\psi(\lambda)\,d\mu(\lambda)$. If the function $H\psi$ is contained in the unit ball of the RKHS, then $\|\psi\|_{2,\mu} \le 1$, and an application of the Cauchy-Schwarz inequality gives that $|H\psi(z)|^2 \le \int e^{2\|\lambda\|\,|z|}\,d\mu(\lambda) \le e^{2C|z|^2}$, for some universal constant C. In particular, the functions $H\psi$ can be extended to analytic functions on the strip $\{z \in \mathbb{C}\colon \|z\|_\infty \le A\}$ that are uniformly bounded by e^{CA^2}, for any A. It follows by Proposition C.9 that $N(\epsilon, \mathbb{H}_1, \|\cdot\|_\infty) \le A^{-d}\log(e^{CA^2}/\epsilon)^{1+d}$. Choosing A of the order $\log_-\epsilon$ leads to the bound $N(\epsilon, \mathbb{H}_1, \|\cdot\|_\infty) \le (\log_-\epsilon)^{1+d/2}$.

The result now follows from the characterization of the small ball exponent by the entropy of the RKHS unit ball, Lemma I.29. □

Lemma 11.39 (Decentering) *If* $w_0 \in \mathfrak{W}^\beta([0,1]^d)$ *for* $\beta > d/2$*, then the concentration function of the square exponential process satisfies, for* $\epsilon < 1$*, a constant* C *that depends only on* w_0*,*

$$\inf_{h:\|h-w_0\|_\infty<\epsilon} \|h\|_{\mathbb{H}}^2 \lesssim \exp\left(C\epsilon^{-2/(\beta-d/2)}\right).$$

Proof For given $K > 0$ let $\psi(\lambda) = (\hat{w}_0/m)(\lambda)\mathbb{1}_{\|\lambda\|\le K}$, for m the density in (11.4). The function $H\psi$ satisfies

$$\begin{aligned}\|H\psi - w_0\|_\infty &\le \int_{\|\lambda\|>K} |\hat{w}_0(\lambda)|\,d\lambda \le \|w_0\|\left[\int_{\|\lambda\|>K}(1+\|\lambda\|^2)^{-\beta}\,d\lambda\right]^{1/2}\\ &\lesssim \|w_0\|_{2,2,\beta}K^{-(\beta-d/2)}.\end{aligned}$$

Furthermore, the squared RKHS-norm of $H\psi$ is given by

$$\begin{aligned}\|H\psi\|_{\mathbb{H}}^2 &= \int_{\|\lambda\|\le K}\frac{|\hat{w}_0|^2}{m}(\lambda)\,d\lambda \le \sup_{\|\lambda\|\le K}\left[m(\lambda)^{-1}(1+\|\lambda\|^2)^{-\beta}\right]\\ &\quad\times \|w_0\|_{2,2,\beta}^2 \lesssim e^{K^2/4}\|w_0\|_{2,2,\beta}^2.\end{aligned}$$

We conclude the proof by choosing $K \asymp \epsilon^{-1/(\beta-d/2)}$. □

Combining the preceding we see that the concentration function of the square exponential process satisfies

$$\varphi_{w_0}(\epsilon) \lesssim \exp\left(C\epsilon^{-2/(\beta-d/2)}\right) + (\log_-\epsilon)^{1+d/2}.$$

The first term (decentering function) dominates the second (centered small ball exponent) for any $\beta > 0$, and the contraction rate for a β-smooth function satisfies $\epsilon_n \asymp (\log n)^{-(\beta/2-d/4)}$. This extremely slow rate is the result of the discrepancy between the infinite smoothness of the prior and the finite smoothness of the true parameter. A remedy for this mismatch is to rescale the sample paths and is discussed in Sections 11.5 and 11.6.

Actually, the preceding establishes only an upper bound on the contraction rate. The next lemma can be used to show that the logarithmic rate is real: balls of logarithmic radius around a parameter w_0 that is exactly of finite smoothness asymptotically receive zero posterior mass. The lemma shows that the upper bound on the decentered concentration function obtained in Lemma 11.39 can be reversed for w_0 that are exactly finitely smooth, at least for the $\mathbb{L}_2$-norm. The implication for the posterior contraction rate can then be obtained from Theorem 8.35. The "exact" finite smoothness of w_0 is operationalized by assuming that its Fourier transform has polynomial tails. (See van der Vaart and van Zanten 2011, Theorem 8, for a proof of the lemma, and the details of an application to posterior rates in the regression model.)

Lemma 11.40 (Lower bound) *If w_0 is contained in $\mathfrak{W}^\beta([0,1]^d)$ for some $\beta > d/2$, has support within $(0,1)^d$ and possesses a Fourier transform satisfying $|\hat{w}_0(\lambda)| \gtrsim \|\lambda\|^{-k}$ for some $k > 0$ and every $\|\lambda\| \geq 1$, then there exists a constants $b, v > 0$ such that*

$$\inf_{h::\|h-w_0\|_2<\epsilon} \|h\|_{\mathbb{H}}^2 \geq e^{b\epsilon^{-v}}.$$

As the square exponential process prior puts all of its mass on analytic functions, perhaps it is not fair to study its performance only for β-regular functions. We shall now show that for "supersmooth," analytic true parameters the prior works very well.

For $r \geq 1$ and $\lambda > 0$, we define $\mathcal{A}^{\gamma,r}(\mathbb{R}^d)$ as the space of functions $w: \mathbb{R}^d \to \mathbb{R}$ with Fourier transform $\hat{w}$ satisfying

$$\|w\|_{\mathcal{A}}^2 := \int e^{\gamma\|\lambda\|^r} |\hat{w}|^2(\lambda)\, d\lambda < \infty.$$

Finiteness of this norm requires exponential decrease of the Fourier transform, in contrast to polynomial decrease for Sobolev smoothness. The functions in $\mathcal{A}^{\gamma,r}(\mathbb{R}^d)$ are infinitely often differentiable and "increasingly smooth" as γ or r increase. They extend to functions that are analytic on a strip in $\mathbb{C}^d$ containing $\mathbb{R}^d$ if $r = 1$, and to entire functions if $r > 1$ (see e.g. Bauer 2001, 8.3.5).

Lemma 11.41 (Decentering) *Suppose that w_0 is the restriction to $[0,1]^d$ of an element of $\mathcal{A}^{\gamma,r}(\mathbb{R}^d)$.*

(i) *If $r > 2$ or ($r \geq 2$ and $\gamma \geq 1/4$), then $w_0 \in \mathbb{H}$.*
(ii) *If $r < 2$, then there exists a constant C depending on w_0 such that, for $\epsilon < 1$,*

$$\inf_{h:\|h-w\|_\infty \leq \epsilon} \|h\|_{\mathbb{H}}^2 \leq C \exp\left(\frac{(\log_- \epsilon)^{2/r}}{4\gamma^{2/r}}\right).$$

Proof The first assertion follows, because $\psi = \hat{w}_0/m$ is contained in $\mathbb{L}_2(\mu)$, if w_0 satisfies the conditions, and $H\psi = w_0$.

The second assertion is proved in the same way as Lemma 11.39, where this time, with $\|w_0\|_{\mathcal{A}}$ the norm of w_0 in $\mathcal{A}^{\gamma,r}(\mathbb{R}^d)$,

$$\|H\psi - w_0\|_\infty^2 \le \int_{\|\lambda\|>K} e^{-\gamma\|\lambda\|^r}\, d\lambda\, \|w_0\|_{\mathcal{A}}^2 \le e^{-\gamma K^r} K^{-r+1} \|w_0\|_{\mathcal{A}}^2,$$
$$\|H\psi\|_{\mathbb{H}}^2 \le \sup_{\|\lambda\|\le K} e^{\|\lambda\|^2/4 - \gamma\|\lambda\|^r} \|w_0\|_{\mathcal{A}}^2 \le e^{K^2/4} \|w_0\|_{\mathcal{A}}^2.$$

We finish by choosing $K \asymp (\gamma^{-1} \log_- \epsilon)^{1/r}$. □

Combination of Lemmas 11.38 and 11.41 leads to the rate equation $\varphi_{w_0}(\epsilon_n) \le n\epsilon_n^2$ of the form

$$\big(\log_- \epsilon_n\big)^{1+d/2} + \exp\Big(C(\log_- \epsilon_n)^{2/r}\Big) \le n\epsilon_n^2.$$

Solving this leads to a posterior contraction rate $n^{-1/2}(\log n)^{1/r \vee (1/2+d/4)}$ for any super-smooth true parameter $w_0 \in \mathcal{A}^{\gamma,r}(\mathbb{R}^d)$. This is up to a logarithmic factor equal to the posterior rate of contraction $n^{-1/2}$ for finite-dimensional models. This "almost parametric rate" is explainable from the fact that spaces of analytic functions are only slightly bigger than finite-dimensional spaces in terms of their metric entropy.

11.4.5 Series Priors

Any Gaussian process can be represented (in many ways) as an infinite series with independent standard normal coefficients and deterministic "coordinate functions." Since the RKHS can be characterized using the same coordinate functions (see Example 11.16) this may be a useful device to derive properties of the prior. Conversely, we may fix our favorite coordinate functions from the beginning, and form the process as a random series. The properties of the resulting prior will depend on the coordinate functions. In this section we consider two examples, one with a truncated wavelet basis and one with an infinite, single-indexed sequence of basis functions. Finite series are attractive for computation, but for good approximation their truncation point must increase to infinity.

We start with a general observation that relates a truncated and an infinite series prior. A truncated Gaussian series is another Gaussian process, to which the general theory on Gaussian processes applies. Naturally, if the truncation point is sufficiently high, the concentration functions of the truncated and full series are equivalent, and the resulting theory for the two prior processes is identical. The following lemma quantifies the truncation point.

The lemma applies more generally to approximation of a Gaussian variable W by a sequence W_n, defined on the same probability space.

Lemma 11.42 *Let W be a mean-zero Gaussian random element in a separable Banach space $\mathbb{B}$ with RKHS $\mathbb{H}$ and concentration function φ_{w_0} at $w_0 \in \bar{\mathbb{H}}$. If W_n are mean-zero Gaussian random elements in $\mathbb{B}$ defined on the same probability space such that $10\mathrm{E}\|W_n - W\|^2 \le n^{-1}$, then for $\epsilon_n > 0$ satisfying $n\epsilon_n^2 \ge 4\log 4$ and $\varphi_{w_0}(\epsilon_n) \le n\epsilon_n^2$ and any $C > 4$,*

then there exist measurable sets $B_n \subset \mathbb{B}$ such that (11.13)–(11.15) *hold with W replaced by W_n and ϵ_n replaced by $2\epsilon_n$.*

Proof By Borell's inequality for the norm, Proposition I.8, applied to $W_n - W$,

$$\mathrm{P}(\|W_n - W\| \geq \epsilon_n) \leq 2e^{-\epsilon_n^2/8\mathrm{E}\|W_n - W\|^2} \leq 2e^{-5n\epsilon_n^2/4}.$$

In view of the inequalities $\mathrm{P}(\|W_n - w_0\| < 3\epsilon_n) \geq \mathrm{P}(\|W - w_0\| < 2\epsilon_n) - \mathrm{P}(\|W_n - W\| \geq \epsilon_n)$, the small ball probability of W_n can be bounded in terms of that of W. Next we choose the sets $B_n = 2\epsilon_n \mathbb{B}_1 + M_n \mathbb{H}_1^n$, where $\mathbb{H}_1^n$ is the unit ball of the RKHS of W_n, and follow steps as in the proof of Theorem 11.20. □

Truncated Wavelet Expansions

Let $\{\psi_{j,k} \colon j \in \mathbb{N}, k = 1, \ldots, 2^{jd}\}$ be an orthonormal basis of $\mathbb{L}_2([0,1]^d)$ of compactly supported wavelets, with j referring to the resolution level and k to the dilation. Let $w = \sum_{j=1}^{\infty} \sum_{k=1}^{2^{jd}} w_{j,k}\psi_{j,k}$ be the expansion of a given function $w \in \mathbb{L}_2([0,1]^d)$, and consider the norms

$$\|w\|_{1,2} = \sum_{j=1}^{\infty} \Big(\sum_{1 \leq k \leq 2^{jd}} |w_{j,k}|^2 \Big)^{1/2},$$

$$\|w\|_{1,\infty} = \sum_{j=1}^{\infty} 2^{jd/2} \max_{1 \leq k \leq 2^{jd}} |w_{j,k}|,$$

$$\|w\|_{\infty,\infty,\beta} = \sup_{1 \leq j < \infty} 2^{j\beta} 2^{jd/2} \max_{1 \leq k \leq 2^{jd}} |w_{j,k}|.$$

For a suitable smooth basis these norms are upper bounds on the $\mathbb{L}_2$-norm and the supremum norm, and equivalent to the Besov (∞, ∞, β)-norm of w, respectively. Actually, that the functions $\psi_{j,k}$ are a wavelet basis is only important to ensure this interpretation. All the following is valid for arbitrary bases, provided the norms of the functions are defined by the preceding formulas.

Consider a Gaussian prior of the type, for given positive constants μ_j, and i.i.d. random variables $Z_{j,k} \sim \mathrm{Nor}(0,1)$,

$$W = \sum_{j=1}^{J_\alpha} \sum_{k=1}^{2^{jd}} \mu_j Z_{j,k} \psi_{j,k}. \tag{11.20}$$

The number of terms in the sum is $(2^{J_\alpha d} - 1)/(1 - 2^{-d})$. For interpretability of α we set J_α so that this is the usual dimension for estimating an α-regular function: determine J_α to be the integer that is closest to the solution of the equation $2^{Jd} = n^{d/(2\alpha+d)}$. We next study the posterior rate of contraction when the true parameter has regularity β, which may be different from the "nominal smoothness level" α.

Since $W_j := \sum_{k=1}^{2^{jd}} \mu_j Z_{j,k} \psi_{j,k}$ contributes the variance $\mathrm{E}\|W_j\|_2^2 = \mu_j^2 2^{jd}$ to the prior, the choice $\mu_j = 2^{-jd/2}$ gives all resolution levels the same prior variation. If $\mu_j 2^{jd/2} \downarrow 0$ as $j \to \infty$, then higher resolution levels receive less weight and hence the prior gives increasingly more weight to dimensions lower than the nominal dimension $2^{J_\alpha d}$. This is

advantageous if the true regularity satisfies $\beta > \alpha$, but in the opposite case $\beta < \alpha$ the nominal dimension $2^{J_\alpha d}$ is already too small, and this would be exacerbated by putting lower weight on the higher levels. The following results show that the choice $\mu_j 2^{jd/2} = 2^{-j\beta}$ is a good compromise: it (nearly) obtains the optimal rate $n^{-\beta/(2\beta+d)}$ for $\beta \geq \alpha$, and the "optimal rate $n^{-\beta/(2\alpha+d)}$ when using a $2^{J_\alpha d}$-dimensional model" in the other case.

Lemma 11.43 (RKHS) *The RKHS of W given by* (11.20) *is the set of functions $w = \sum_{j=1}^{J_\alpha} \sum_k w_{j,k}\psi_{j,k}$ for which the norm $\|w\|_{\mathbb{H}}^2 = \sum_{j=1}^{J_\alpha} \sum_k \mu_j^{-2} w_{j,k}^2$ is finite.*

Lemma 11.44 (Small ball) *The centered small ball probability of W given by* (11.20) *with $\mu_j 2^{jd/2} = 2^{-ja}$ for some $a \geq 0$, viewed as a map in $\mathfrak{L}_\infty([0,1]^d)$, satisfies, as $\epsilon_n \to 0$*

$$\varphi_0(\epsilon_n) \lesssim \begin{cases} 2^{J_\alpha d}(\log_- \epsilon_n + J_\alpha), & \text{if } \epsilon_n 2^{J_\alpha a} \lesssim J_\alpha^2, \\ \epsilon_n^{-d/a}(\log_- \epsilon_n)^{2d/a}, & \text{if } \epsilon_n 2^{J_\alpha a} \gtrsim J_\alpha^2. \end{cases}$$

Proof The first lemma is immediate from Example 11.16. For the proof of the second, take any numbers $\alpha_j \geq 0$ with $\sum_{j=1}^{J_\alpha} \alpha_j \leq 1$, and note that

$$\mathrm{P}(\|W\|_\infty < \epsilon) = \mathrm{P}\Big(\sum_{j=1}^{J_\alpha} 2^{jd/2} \max_{1\leq k\leq 2^{jd}} |\mu_j Z_{j,k}| < \epsilon\Big) \geq \prod_{j=1}^{J_\alpha} \prod_{k=1}^{2^{jd}} \mathrm{P}(|\mu_j 2^{jd/2} Z_{j,k}| < \alpha_j \epsilon).$$

Therefore for $\mu_j 2^{jd/2} = 2^{-ja}$, and $\alpha_j = (K + d^2 j^2)^{-1}$, it follows that

$$\varphi_0(\epsilon_n) \leq -\sum_{j=1}^{J_\alpha} 2^{jd} \log\Big(2\Phi(\alpha_j \epsilon_n 2^{ja}) - 1\Big) \lesssim \int_1^{2^{J_\alpha d}} -\log\left(2\Phi\Big(\frac{\epsilon_n x^{a/d}}{K + \log_2^2 x}\Big) - 1\right) dx,$$

provided that K is large enough to make the map $x \mapsto x^{a/d}/(K + \log_2^2 x)$ nondecreasing on $[1, \infty)$. It may be verified that the function f given by $f(y) = -\log(2\Phi(y) - 1)$ appearing in the right side is decreasing, with $f(0) = \infty$, $f(\infty) = 0$, $f(y) \leq 1 + |\log y|$ on $[0, c]$, and $f(y) \leq e^{-y^2/2}$ for $y \geq c$.

We bound the right side of the preceding display separately for the two cases considered in the lemma. If $\epsilon_n 2^{J_\alpha a} \leq K + J_\alpha^2 d^2$, then $\epsilon_n x^{a/d}/(K + \log_2^2 x) = O(1)$ on $[1, 2^{J_\alpha d}]$ and hence the right side is bounded by a multiple of

$$\int_1^{2^{J_\alpha d}} \left(1 + \left|\log\Big(\frac{\epsilon_n x^{a/d}}{K + \log_2^2 x}\Big)\right|\right) dx \lesssim 2^{J_\alpha d}(\log_- \epsilon_n + J_\alpha).$$

Alternatively, if $\epsilon_n 2^{J_\alpha a} > (K + J_\alpha^2 d^2)$, then the right side is bounded by

$$\begin{aligned} &\epsilon_n^{-d/a} \int_{\epsilon_n}^{\epsilon_n 2^{J_\alpha a}} f\Big(\frac{y}{K + (d/a)^2(\log_2 y + \log_2 \epsilon_n^{-1})^2}\Big) \frac{d}{a} y^{d/a-1}\, dy \\ &\leq \Big[\int_0^{\epsilon_n^{-1}} f\Big(\frac{y}{K + (2d/a)^2 \log_2^2 \epsilon_n^{-1}}\Big) + \int_{\epsilon_n^{-1}}^{\infty} f\Big(\frac{y}{K + (2d/a)^2 \log_2^2 y}\Big)\Big] \frac{d}{a} y^{d/a-1}\, dy \\ &\leq \mu_n^{d/a} \int_0^{\epsilon_n^{-1}\mu_n^{-1}} f(x) \frac{d}{a} x^{d/a-1}\, dx + \int_0^{\infty} f\Big(\frac{y}{K + (2d/a)^2 \log_2^2 y}\Big) \frac{d}{a} y^{d/a-1}\, dy, \end{aligned}$$

for $\mu_n = K + (2d/a)^2(\log_2 \epsilon_n^{-1})^2$. The first term on the right side is bounded by a constant, whence the whole expression is bounded by a multiple of $(\log_- \epsilon_n)^{2d/a}$. □

Lemma 11.45 (Decentering) *If $\|w\|_{\infty,\infty,\beta} < \infty$, then the decentering function of W given by* (11.20) *with $\mu_j 2^{jd/2} = 2^{-ja}$ for some $a \geq 0$ satisfies, for any $J \leq J_\alpha$ and $\epsilon \geq 2^{-J\beta}\|w\|_{\infty,\infty,\beta}/(2^\beta - 1)$,*

$$\inf_{h:\|h-w\|_\infty<\epsilon} \|h\|_{\mathbb{H}}^2 \lesssim 2^{J(2a-2\beta+d)}\|w\|_{\infty,\infty,\beta}^2.$$

Proof If w^J stands for the projection of w on the space spanned by the wavelet basis up to resolution level J, then

$$\|w - w^J\|_\infty \leq \sum_{j=J+1}^{\infty} 2^{jd/2} \max_k |w_{j,k}| \leq \sum_{j=J+1}^{\infty} 2^{-j\beta}\|w\|_{\infty,\infty,\beta} \lesssim \frac{2^{-J\beta}}{2^\beta - 1}\|w\|_{\infty,\infty,\beta}.$$

For $J \leq J_\alpha$ the function w^J belongs to the RKHS, and hence the infimum in the lemma is smaller than the square RKHS norm of w^J if $2^{-J\beta}\|w\|_{\infty,\infty,\beta} \leq \epsilon(2^\beta - 1)$. This norm is equal to $\sum_{j=1}^J \sum_{k=1}^{2^{jd}} w_{j,k}^2/\mu_j^2 \leq \sum_{j=1}^J 2^{j(2a-2\beta+d)}\|w\|_{\infty,\infty,\beta}^2$. □

To solve the rate equation $\varphi_{w_0}(\epsilon_n) \leq n\epsilon_n^2$ for w_0 with finite norm $\|w_0\|_{\beta,\infty,\infty}$, we determine ϵ_n and J such that $2^{-J\beta} \lesssim \epsilon_n$ and $2^{J(2a-2\beta+d)} \leq n\epsilon_n^2$ and $\varphi_0(\epsilon_n) \leq n\epsilon_n^2$. It can be verified that the minimal value of ϵ_n is obtained when choosing $J = J_a \wedge J_\alpha$, and results in the following lower bounds on ϵ_n:

- $n^{-\beta/(2\alpha+d)} \log n$ if $a \leq \beta \leq \alpha$;
- $n^{-\alpha/(2\alpha+d)} \log n$ if $a \leq \alpha \leq \beta$;
- $n^{-a/(2a+d)}(\log n)^{d/(2a+d)}$ if $\alpha \leq a \leq \beta$;
- $n^{-\beta/(2a+d)}(\log n)^{d/(2a+d)}$ if $\alpha \leq \beta \leq a$.

Infinite Series

Let $\phi_1, \phi_2, \ldots$ be a sequence of basis functions, such that the series $w = \sum_{j=1}^\infty w_j\phi_j$ is a well-defined function for every $w \in \ell_2$, where the correspondence between the function and the coefficients is one-to-one. We define the square norms

$$\|w\|_2^2 = \sum_{j=1}^\infty w_j^2, \qquad \|w\|_{2,2,\beta}^2 = \sum_{j=1}^\infty j^{2\beta} w_j^2.$$

For an orthonormal basis of $\mathbb{L}_2[0, 1]$, the first norm is simply the $\mathbb{L}_2$-norm. An example is given by the trigonometric functions $\phi_1(t) = 1$, $\phi_{2j}(t) = \cos(2\pi jt)$, and $\phi_{2j+1}(t) = \sin(2\pi jt)$, for $j \in \mathbb{N}$. For this basis the norm $\|\cdot\|_{2,2,\beta}$ is a *Sobolev norm*, which measures smoothness of the function w. For a general basis (ϕ_j), we may think of this norm as measuring regularity of order β in a general sense.

We consider the Gaussian random element $W = \sum_{j=1}^\infty j^{-\alpha-1/2} Z_j\phi_j$, and $Z_j \overset{\text{iid}}{\sim} \text{Nor}(0, 1)$. Then $\mathrm{E}\|W\|_{2,2,\beta}^2 = \sum_j j^{2\beta-2\alpha-1} < \infty$ for every $\beta < \alpha$, and hence the prior can be viewed as (almost) regular of order β. In particular, the prior is well defined as an almost surely converging sequence whenever $\alpha > 0$.

Lemma 11.46 (RKHS) *The RKHS of the variable* $W = \sum_{j=1}^{\infty} j^{-\alpha-1/2} Z_j \phi_j$ *is the set of functions* $\sum_j w_j \phi_j$ *with* $w \in \mathfrak{W}^{\alpha+1/2}[0,1]$ *and the RKHS norm is* $\|w\|_{2,2,\alpha+1/2}$.

Lemma 11.47 (Small ball) *The small ball probability of* $W = \sum_{j=1}^{\infty} j^{-\alpha-1/2} Z_j \phi_j$ *relative to the norm* $\|\cdot\|_2$ *satisfies, for universal constants* $0 < c < C < \infty$ *and* $d > 0$, *and every* $\epsilon^{-1/\alpha} \geq d$,

$$(c2^{-1/(2\alpha)})\,\epsilon^{-1/\alpha} \leq \varphi_0(\epsilon) \leq (C2^{1/(2\alpha)})\,\epsilon^{-1/\alpha}.$$

Lemma 11.48 (Decentering) *If* $\|w\|_{2,2,\beta} < \infty$ *for* $\beta \leq \alpha + 1/2$, *then, for* $\epsilon \leq \|w\|_{2,2,\beta}$,

$$\inf_{h:\|h-w\|_2<\epsilon} \|h\|_{\mathbb{H}}^2 \leq \|w\|_{2,2,\beta}^{(2\alpha+1)/\beta} \epsilon^{-(2\alpha-2\beta+1)/\beta}.$$

Proof The first lemma is immediate from Example 11.16.

For the proof of the upper bound in the second lemma we first note that, because normal densities with standard deviations $\sigma \geq \tau$ satisfy $\phi_\sigma(x)/\phi_\tau(x) \geq \tau/\sigma$, for every x,

$$\mathrm{P}\Big(\sum_{j\leq J} Z_j^2 j^{-2\alpha-1} < \epsilon^2\Big) \geq \prod_{j=1}^{J} \Big(\frac{J^{-\alpha-1/2}}{j^{-\alpha-1/2}}\Big) \mathrm{P}\Big(\sum_{j=1}^{J} Z_j^2 J^{-2\alpha-1} < \epsilon^2\Big).$$

The leading factor is $(J!/J^J)^{\alpha+1/2} \geq e^{-J(\alpha+1/2)}$ and the probability on the far right tends to a number not smaller than $1/2$ if $\epsilon^2 J^{2\alpha} \geq 1$, by the central limit theorem, as $J \to \infty$. Second, by Markov's inequality,

$$\mathrm{P}\Big(\sum_{j>J} Z_j^2 j^{-2\alpha-1} < \epsilon^2\Big) \geq 1 - \frac{1}{\epsilon^2} \sum_{j>J} \mathrm{E}(Z_j^2 j^{-2\alpha-1}) \geq 1 - \frac{1}{2\alpha\epsilon^2 J^{2\alpha}}.$$

The right side is at least $1/2$ if $J \geq \epsilon^{-1/\alpha}\alpha^{-2\alpha}$. Since $\alpha^{-2\alpha} \geq e^{2e^{-1}}$, for any $\alpha > 0$, the inequalities in both preceding displays are satisfied for $J \geq \epsilon^{-1/\alpha} e^{2e^{-1}}$, and then $\mathrm{P}(\sum_j Z_j^2 j^{-2\alpha-1} < 2\epsilon^2)$ is bounded below by a multiple of $e^{-J(\alpha+1/2)} \geq e^{-J/2}$.

The small ball probability is upper bounded by $\mathrm{P}(\sum_{j=1}^{J} Z_j^2 J^{-2\alpha-1} < \epsilon^2)$, for every J. If $\epsilon^2 J^{2\alpha} \leq 1/2$, then this is bounded above by the probability that a chi-squared variable with J degrees of freedom is bounded above by $J/2$, which can be written as, for $I = J/2$,

$$\frac{1}{\Gamma(I)} \int_0^{I/2} x^{I-1} e^{-x}\, dx = \frac{e^{-I/2} I^{I-1}}{\Gamma(I) 2^I} \int_0^I \Big(1 - \frac{y}{I}\Big)^{I-1} e^{y/2}\, dy,$$

by the substitution $2x = I - y$. By Stirling's formula the leading factor is bounded above by $(\sqrt{e}/2)^I/\sqrt{\pi I}$, while for $I \geq 1$ the integrand is bounded above by $e^{-y/2+y/I}$ whence the integral is bounded by $1/(1/2 - 1/I)$ if $I > 2$. We conclude by choosing $2I = J$ equal to the biggest integer smaller than $(\sqrt{2}\epsilon)^{-1/\alpha}$.

For every $J \in \mathbb{N}$ the truncated function $w^J := \sum_{j=1}^{J} w_j \phi_j$ is contained in the RKHS. Its square distance to w and square RKHS-norm satisfy

$$\|w^J - w\|_2^2 = \sum_{j>J} w_j^2 \le J^{-2\beta}\|w\|_{2,2,\beta}^2,$$

$$\|w^J\|_{\mathbb{H}}^2 = \sum_{j=1}^{J} j^{2\alpha+1} w_j^2 \le \|w\|_{2,2,\beta}^2 \max_{1\le j\le J} j^{2\alpha-2\beta+1}.$$

Choosing a minimal integer J such that $J^\beta \ge \|w\|_{2,2,\beta}/\epsilon$ readily gives the third lemma. □

If the true parameter w_0 has finite norm $\|w_0\|_{2,2,\beta}$, then the minimal solution to the rate equation $\varphi_{w_0}(\epsilon_n) \le n\epsilon_n^2$ can be seen to be $\epsilon_n \asymp n^{-(\alpha\wedge\beta)/(2\alpha+1)}$. This is the minimax rate $n^{-\beta/(2\beta+1)}$ if and only if prior and true function match (i.e. $\alpha = \beta$), as usual. We shall now show that the suboptimal rates in the other cases are sharp, separately in the case that $\beta > \alpha$ (rate $n^{-\alpha/(2\alpha+1)}$) and the case that $\beta < \alpha$ (rate $n^{-\beta/(2\alpha+1)}$).

Because the centered small ball exponent is of the exact order $\epsilon^{-1/\alpha}$, we have $\varphi_{w_0}(\delta_n) \gtrsim \delta_n^{-1/\alpha}$, for any w_0. For $\delta_n = C_2^{-\alpha} n^{-\alpha/(2\alpha+1)}$ this is equal to $C_2 n^{1/(2\alpha+1)} = C_2 n\epsilon_n^2$ if $\alpha \le \beta$. By Theorem 8.35 we conclude that the posterior rate of contraction is not faster than δ_n for any w_0 that is β-regular of order $\beta \ge \alpha$. Thus for any w_0 that is regular of order $\beta \ge \alpha$ bigger than the prior regularity α, the posterior rate of contraction is of the exact order $n^{-\alpha/(2\alpha+1)}$. The suboptimality of this rate (relative to the minimax rate $n^{-\beta/(2\beta+1)}$) when β is strictly bigger than α is due to the roughness of the prior, which puts less mass near 0 than is desirable for smooth true functions.

For w_0 of regularity $\beta < \alpha$ the rate $\epsilon_n \asymp n^{-\beta/(2\alpha+1)}$ may be thought of as sharp as well, but only for w_0 that are *exactly* of regularity β. Finiteness of the Sobolev norm $\|w_0\|_{2,2,\beta}$ allows w_0 also to be *more* regular than β and then the posterior will contract at a faster rate. Unfortunately, in the Sobolev case the concept of "exact regularity" appears difficult to define. In the following examples we obtain nearly sharp rates for certain fixed w_0 "near" the boundary of the Sobolev ball, and exhibit a sequence $w_{0,n}$ that "moves to the boundary" for which the rate $n^{-\beta/(2\alpha+1)}$ is sharp.

Example 11.49 (Lower bound) Let $\beta < \alpha$ and $a > 0$, and consider the function w with Fourier coefficients $w_1 = 0$ and $w_j = j^{-\beta-1/2}(\log j)^{-a}$, for $j \in \mathbb{N}$ and $j \ge 2$. For $a > 1/2$ the function satisfies $\|w\|_{2,2,\beta} < \infty$, whereas $\|w\|_{2,2,b} = \infty$, for every $b > \beta$. Hence w can be interpreted as being of "nearly exact" regularity β (up to a logarithmic factor).

The decentering part of the concentration function, as in Lemma 11.48, takes the form

$$d(\epsilon) := \inf_{\sum_{j=1}^\infty (h_j - w_j)^2 \le \epsilon^2} \sum_{j=1}^{\infty} j^{2\alpha+1} h_j^2.$$

By introducing a Lagrange multiplier λ, we can see that the solution of this problem takes the form $h_j = \lambda w_j/(j^{2\alpha+1} + \lambda)$, where $\lambda = \lambda(\epsilon)$ solves $\sum_j (h_j - w_j)^2 = \epsilon^2$, for small $\epsilon > 0$. For $\{w_j\}$ as given, some calculus (see Lemma K.8) gives the solution

$$d(\epsilon) \asymp \lambda(\epsilon)\epsilon^2, \quad \lambda(\epsilon) \asymp \epsilon^{-(2\alpha+1)/\beta}(\log_- \epsilon)^{-a(2\alpha+1)/\beta}, \quad \epsilon \asymp \lambda(\epsilon)^{-\beta/(2\alpha+1)}(\log \lambda(\epsilon))^{-a}.$$

The equation $d(\epsilon_n) \asymp n\epsilon_n^2$ is solved for $\lambda(\epsilon_n) \asymp n$ and $\epsilon_n \asymp n^{-\beta/(2\alpha+1)}(\log n)^{-a}$. Since $\epsilon_n^{-1/\alpha} \ll d(\epsilon_n)$, the rate equation $\varphi_0(\epsilon_n) \asymp n\epsilon_n^2$ gives the same solution, whence we obtain the rate of contraction $n^{-\beta/(2\alpha+1)}$ found in the preceding up to a logarithmic factor. By the

same calculation it also follows that for every constant C_2 there exists a constant m such that $d(m\epsilon_n) \geq C_2 n\epsilon_n^2$. This shows that $m\epsilon_n$ is a lower bound for the rate of contraction, and hence the rate $n^{-\beta/(2\alpha+1)}(\log n)^{-a}$ is sharp.

Example 11.50 (Lower bound, sequence) Let $\beta < \alpha$ and consider the sequence w^n with a single nonzero coordinate $w_j = j^{-\beta}$ for $j = J_n \sim Cn^{1/(2\alpha+1)}$ (and $w_j = 0$ for $j \neq J_n$). The sequence w^n satisfies $\|w^n\|_{2,2,\beta} = 1$, for every n, and hence is uniformly of regularity β. The minimizing h in the left side of Lemma 11.48 has $h_j = (w_j - \epsilon)_+$ for $j = J_n$ and $h_j = 0$ otherwise and hence the decentering function is given by $d(\epsilon) = J_n^{2\alpha+1}(J_n^{-\beta} - \epsilon)_+^2$.

The corresponding concentration function is $\varphi_{w^n}(\epsilon) \asymp \epsilon^{-1/\alpha} + d(\epsilon)$. For $\epsilon_n = J_n^{-\beta}$ we have $\varphi_{w^n}(\epsilon_n) \lesssim \epsilon_n^{-1/\alpha} = J_n^{\beta/\alpha} \lesssim n^{(\beta/\alpha)/(2\alpha+1)}$, and for $\delta_n = J_n^{-\beta}/2$ we have $\varphi_{w^n}(\delta_n) \geq d(\delta_n) \gtrsim J_n^{2\alpha+1} J_n^{-2\beta} \sim C^{2\alpha+1-2\beta} n^{(2\alpha-2\beta+1)/(2\alpha+1)}$. Since $n\epsilon_n^2 \sim C^{-2\beta} n^{(2\alpha-2\beta+1)/(2\alpha+1)}$ we have $\varphi_{w^n}(\epsilon_n) \ll n\epsilon_n^2$ and $C_2 n\epsilon_n^2 \leq \varphi_{w^n}(\delta_n)$, if C is chosen sufficiently large.

It follows that the rate $\delta_n \asymp n^{-\beta/(2\alpha+1)}$ is sharp along the sequence w^n.

11.5 Rescaled Gaussian Processes

The general finding in Section 11.3 is that priors based on Gaussian processes lead to optimal or near-optimal posterior contraction rates provided the smoothness of the Gaussian process matches that of the target function. Both oversmoothing and undersmoothing lead to suboptimal contraction rates. Nevertheless, it is common practice to use a Gaussian process of a specific form, for instance a supersmooth stationary process such as the square exponential process. The properties of this process are then adapted to the target function by hyperparameters in the covariance kernel. Inserting a scale parameter a in the covariance function is equivalent to scaling the time parameter of the process, thus considering a prior process $t \mapsto W_t^a := W_{at}$ instead of the original process, and is known as changing the *length scale* of the process. If the scale parameter is limited to a compact subset of $(0, \infty)$, then the contraction rate does not change (see e.g. Theorem 2.4 in van der Vaart and van Zanten 2008a). However, while the qualitative smoothness of the sample paths will not change for any a, a dramatic impact on the posterior contraction rate can be observed when $a = a_n$ decreases to 0 or increases to infinity with the sample size n. This is illustrated in this section for the classes of stationary and self-similar prior processes, for deterministic rescaling rates. In Section 11.6 we follow this up by considering a hyperprior on the rescaling rates, thus allowing for random, data-dependent rescaling.

We take the index set for the Gaussian process prior W^a equal to $[0, 1]^d$, and construct this process as $W_t^a = W_{at}$, given a fixed process $W = (W_t : t \in [0, 1/a]^d)$. For $a < 1$ this entails shrinking a process on a bigger time set to the time set $[0, 1]^d$, whereas $a > 1$ corresponds to stretching. Intuitively shrinking makes the sample paths more variable, as the randomness on a bigger time set is packed inside $[0, 1]^d$, whereas stretching creates a smoother process. One finding in the following is that shrinking can make a given process arbitrarily rough, but the smoothing effect of stretching is limited to the smoothness of functions in the RKHS of the initial process. This may motivate to start with a very smooth process.

The scaling map $w \mapsto (t \mapsto w(at))$ is typically a continuous, linear map between the encompassing Banach spaces (e.g. $\mathfrak{C}(T)$, $\mathbb{L}_2(T)$, $\mathfrak{L}_\infty(T)$). Then by general arguments (see

Lemma I.16) the same map is an isometry between the RKHSs of the rescaled and original processes W^a and W.

Lemma 11.51 (RKHS) *If the map $w \mapsto (t \mapsto w(at))$ is a continuous, linear map from the Banach spaces $\mathbb{B}_a$ into $\mathbb{B}$, and W is a random element in $\mathbb{B}_a$, then the RKHS of the process W^a given by $W^a_t = W_{at}$ consists of the functions $t \mapsto h(at)$ for h in the RKHS of W, with identical norms.*

11.5.1 Self-Similar Processes

A stochastic process W is self-similar of order α if the processes $(W_{at}: 0 \leq t \leq 1)$ and $(a^\alpha W_t: 0 \leq t \leq 1)$ are equal in distribution. We shall understand the latter as referring to the Borel laws of the two processes in the encompassing Banach spaces, and assume that the stochastic process and Banach space RKHSs are equivalent. Thus the rescaling of the time-axis is equivalent to a rescaling of the vertical axis. This observation makes the following two lemmas evident.

Lemma 11.52 (RKHS) *The RKHS $\mathbb{H}^a$ of the rescaled process W^a corresponding to a self-similar process W of order α is the RKHS $\mathbb{H}$ of W, but equipped with the norm $\|h\|_{\mathbb{H}^a} = a^{-\alpha}\|h\|_{\mathbb{H}}$.*

Lemma 11.53 (Small ball) *The small ball exponent φ^a_0 of the rescaled process W^a corresponding to a self-similar process W of order α satisfies $\varphi^a_0(\epsilon) = \varphi_0(a^{-\alpha}\epsilon)$, for φ_0 the small ball exponent of W.*

Proofs The function $t \mapsto \mathrm{E}[W_{at}Z] = a^\alpha \mathrm{E}[W_t Z]$ is contained in both $\mathbb{H}^a$ and $\mathbb{H}$ and has square norm $\mathrm{E}[Z^2]$ in $\mathbb{H}^a$ and $a^{2\alpha}\mathrm{E}[Z^2]$ in $\mathbb{H}$. The second lemma is immediate from the fact that the events $\{\|W^a\| < \epsilon\}$ and $\{\|a^\alpha W\| < \epsilon\}$ are equal in probability. □

It follows that the concentration function $\varphi^a_{w_0}$ of the rescaled process W^a can be written as

$$\varphi^a_{w_0}(\epsilon) = \varphi_0(a^{-\alpha}\epsilon) + a^{-2\alpha} \inf_{\|h-w_0\|\leq\epsilon} \|h\|^2_{\mathbb{H}}.$$

Increasing the scaling factor a makes the first term on the right side bigger, but decreases the second term. We may now choose $a = a_n$ such that the solution ϵ_n of the rate equation $\varphi^{a_n}_{w_0}(\epsilon_n) \asymp n\epsilon_n^2$ is smallest. In the case that $\varphi_0(\epsilon) \asymp \epsilon^{-r}$ and $\inf\{\|h\|^2_{\mathbb{H}}: \|h-w_0\| \leq \epsilon\} \asymp \epsilon^{-s}$, for some $r, s > 0$, the optimal scaling value and resulting contraction rate can be seen to be

$$a_n = n^{(s-r)/(4\alpha+4r\alpha+rs\alpha)}, \qquad \epsilon_n = n^{-(2+r)/(4+4r+rs)}.$$

It may be noted that the exponent of self-similarity appears in the scaling factor, but not in the contraction rate.

Example 11.54 (Rescaled integrated Brownian motion) The k-fold integrated Brownian motion $I^k_{0+}B$ of Example 11.6 is self-similar of order $k + 1/2$. Its small ball probability is of the order $\epsilon^{-1/(k+1/2)}$, and its decentering function for a function w_0 belonging to the Hölder space of order $\beta \leq k+1$ and having vanishing derivatives at 0 is of the order $\epsilon^{-(2k-2\beta+2)/\beta}$.

Substitution of $\alpha = k + 1/2$, $r = 1/(k + 1/2)$ and $s = (2k - 2\beta + 2)/\beta$ in the preceding display yields the rescaling rate $a_n = n^{(k+1/2-\beta)/((k+1/2)(2\beta+1))}$ and the contraction rate $n^{-\beta/(2\beta+1)}$, for $\beta \leq k + 1$. Thus the minimax rate is obtained for all $\beta \leq k + 1$. For $\beta < k + 1/2$ the integrated Brownian motion is shrunk, and every possible smoothness level is attained. For $\beta > k + 1/2$ the process is stretched, but this is successful only up to smoothness level $k + 1$.

In order to drop the restriction that w_0 has vanishing derivatives at 0 we added a random polynomial of order k to the process in Example 11.6. Because this polynomial process is not self-similar, the preceding argument does not apply to this extension. Actually rescaling the polynomial part in the same way as the integrated Brownian motion may also not be natural. Now it may be checked that the decentering function of the process $a_n^{k+1/2} I_{0+}^k B + b_n \sum_{i=1}^k Z_i t^i$ satisfies, for $w_0 \in \mathfrak{C}^\beta[0, 1]$ and $\beta \leq k + 1$,

$$\inf_{\|h-w\|_\infty \leq \epsilon} \|h\|_{\mathbb{H}^{a,b}}^2 \lesssim a_n^{-(2k+1)} \epsilon^{-(2k-2\beta+2)/\beta} + b_n^{-2}[\epsilon^{-(2k-2\beta)/\beta} \vee 1].$$

This is dominated by the first term (arising from the integrated Brownian motion) if $b_n \geq a_n^{k+1/2} \epsilon_n^{((k+1-\beta)\wedge 1)/\beta}$. Since the small ball exponent of the process is hardly determined by the polynomial part, under the latter condition the preceding derivation goes through without essential changes for any w_0 in the Hölder space of order $\beta \leq k + 1$.

11.5.2 Stationary Gaussian Processes

In the preceding section it is seen that certain finitely smooth processes may be roughened to an arbitrary degree to make it a suitable prior for a function of lesser smoothness than the process, but cannot be smoothened a lot. This may motivate to take a process with infinitely smooth sample paths as the base prior. In this section we discuss the example of a stationary process with a spectral measure that possesses exponential moments.

If W is a stationary Gaussian process with spectral measure μ, then the rescaled process W^a is a stationary Gaussian process with spectral measure μ^a given by $\mu^a(B) = \mu(a^{-1}B)$. Furthermore, if μ is absolutely continuous with spectral density m, then μ^a has density $t \mapsto m(t/a)/a^d$.

The RKHS of W^a is already described in Lemma 11.35, where μ must be replaced by μ^a. The small ball probability and decentering function require new calculations that take the scaling into account.

Lemma 11.55 (Small ball) *Let W be a stationary process with spectral measure μ such that $\int e^{\delta\|\lambda\|}\, \mu(\lambda) < \infty$, for some $\delta > 0$. For any $a_0 > 0$ there exist constants C and ϵ_0 depending only on a_0, μ and d only such that the small ball exponent of the process W^a satisfies, for $a \geq a_0$ and $\epsilon < \epsilon_0$,*

$$\varphi_0^a(\epsilon) \leq C a^d (\log(a/\epsilon))^{1+d}.$$

Proof This is similar to the proof of Lemma 11.38. The entropy of the unit ball of the RKHS $\mathbb{H}^a$ of W^a, can be seen to satisfy $\log N(\epsilon, \mathbb{H}_1^a, \|\cdot\|_\infty) \lesssim a^d (\log_- \epsilon)^{1+d}$. This upper bound has exponent $1 + d$ instead of $1 + d/2$ in Lemma 11.38, because presently only exponential and not square exponential moments on μ are assumed finite. The proof must take

proper care of the dependence of all entities on a. In the application of the characterization of the small ball exponent through entropy this requires following the steps in the proof of Lemma I.29 in Kuelbs and Li (1993) and Li and Linde (1998). See van der Vaart and van Zanten (2007) for details. □

Lemma 11.56 (Decentering) *If the spectral density m is bounded away from 0 in a neighborhood of 0 and $\int e^{\delta\|\lambda\|}\, m(\lambda)\, d\lambda < \infty$ for some $\delta > 0$, then for any $\beta > 0$ and $w \in \mathfrak{C}^\beta([0,1]^d)$ there exist constants C and D depending only on m and w such that, for $a > 0$,*

$$\inf_{h:\|h-w\|_\infty \le Ca^{-\beta}} \|h\|_{\mathbb{H}^a}^2 \le Da^d.$$

Proof Let $\underline{\beta}$ be the biggest integer strictly smaller than β, and let G be a bounded neighborhood of the origin on which m is bounded away from 0. Take a function $\psi\colon \mathbb{R} \to \mathbb{C}$ with a symmetric, real-valued, infinitely smooth Fourier transform $\hat\psi$ that is supported on an interval I such that $I^d \subset G$ and which equals $1/(2\pi)$ in a neighborhood of zero. Then ψ has moments of all orders and

$$\int (it)^k \psi(t)\, dt = 2\pi \hat\psi^{(k)}(0) = \begin{cases} 0, & k \ge 1, \\ 1, & k = 0. \end{cases}$$

Define $\phi\colon \mathbb{R}^d \to \mathbb{C}$ by $\phi(t) = \psi(t_1) \times \cdots \times \psi(t_d)$. Then we have that $\int \phi(t)\, dt = 1$, and $\int t^k \phi(t)\, dt = 0$, for any nonzero multi-index $k = (k_1, \ldots, k_d)$ of nonnegative integers. Moreover, we have that $\int \|t\|^\beta |\phi(t)|\, dt < \infty$, and the functions $|\hat\phi|/m$ and $|\hat\phi|^2/m$ are uniformly bounded.

By Whitney's theorem we can extend $w\colon [0,1]^d \to \mathbb{R}$ to a function $w\colon \mathbb{R}^d \to \mathbb{R}$ with compact support and $\|w\|_\beta < \infty$. (See Whitney 1934 or Stein 1970, Chapter VI; we can multiply an arbitrary smooth extension by an infinitely smooth function that vanishes outside a neighborhood of $[0,1]^d$ to ensure compact support.)

By Taylor's theorem we can write, for $s, t \in \mathbb{R}^d$,

$$w(t+s) = \sum_{j:j.\le\underline{\beta}} D^j w(t) \frac{s^j}{j!} + S(t,s),$$

where $|S(t,s)| \le C\|s\|^\beta$, for a positive constant C that depends on w but not on s and t. If we set $\phi_a(t) = \phi(at)$ we get, in view of the fact that ϕ is a higher order kernel, for any $t \in \mathbb{R}^d$,

$$a^d(\phi_a * w)(t) - w(t) = \int \phi(s)(w(t - s/a) - w(t))\, ds = \int \phi(s) S(t, -s/a)\, ds.$$

Combining the preceding displays shows that $\|a^d \phi_a * w - w\|_\infty \le KCa^{-\beta}$, for $K = \int \|s\|^\beta |\phi|(s)\, ds$.

For $\hat w$ the Fourier transform of w, we can write

$$\frac{1}{(2\pi)^d}(\phi_a * w)(t) = \int e^{-i(t,\lambda)} \hat w(\lambda)\hat\phi_a(\lambda)\, d\lambda = \int e^{-i(t,\lambda)} \frac{\hat w(-\lambda)\hat\phi_a(\lambda)}{m_a(\lambda)}\, d\mu_a(\lambda).$$

Therefore, by Lemma 11.35 the function $a^d \phi_a * w$ is contained in the RKHS $\mathbb{H}^a$, with square norm a multiple of

$$a^{2d} \int \left| \frac{\hat{w}\hat{\phi}_a}{m_a} \right|^2 d\mu_a \le a^d \int |\hat{w}(\lambda)|^2 \, d\lambda \, \left\| \frac{|\hat{\phi}|^2}{m} \right\|_\infty.$$

Here $(2\pi)^d \int |\hat{w}(\lambda)|^2 \, d\lambda = \int |w(t)|^2 \, dt$ is finite, and $|\hat{\phi}|^2/m$ is bounded by the construction of $\hat{\phi}$. □

Combining the preceding we see that for $\epsilon \gtrsim a^{-\beta}$ the concentration function of W^a at $w \in \mathfrak{C}^\beta([0,1]^d)$ satisfies

$$\varphi_w^a(\epsilon) \lesssim Da^d + a^d (\log(a/\epsilon))^{1+d}.$$

Thus for $a = a_n$ the rate equation $\varphi_w^a(\epsilon_n) \le n\epsilon_n^2$ is satisfied by ϵ_n such that

$$\epsilon_n \gtrsim a_n^{-\beta}, \qquad a_n^d (\log(a_n/\epsilon_n))^{1+d} \lesssim n\epsilon_n^2, \qquad a_n^d \lesssim n\epsilon_n^2.$$

For a_n bounded away from zero the third inequality is redundant and the first two inequalities can be seen to be satisfied by

$$a_n = n^{1/(2\beta+d)} (\log n)^{-(1+d)/(2\beta+d)}, \qquad \epsilon_n = n^{-\beta/(2\beta+d)} (\log n)^{\beta(1+d)/(2\beta+d)}.$$

The contraction rate ϵ_n is the minimax rate for estimating a β-regular function up to a logarithmic factor, while the inverse scaling rate a_n^{-1} agrees with the usual bandwidth for kernel smoothing up to a logarithmic factor.

11.6 Adaptation

In the preceding section it is seen that an appropriate length scale can turn a Gaussian process with smooth sample paths into an appropriate prior for true functions of arbitrary smoothness. However, inspection of the formulas reveals that the optimal length scale depends on the smoothness of the function of interest, which will typically not be known a priori. It is natural to try and choose the length scale using the data. A popular *empirical Bayes* method is to maximize the marginal (Bayesian) likelihood for the rescaling constant a, but the theoretical properties of these procedures have become known only recently and for the Gaussian white noise model only; see Szabó et al. (2013). In this section we consider the full Bayes alternative to put a (hyper) prior on a, and show that the resulting mixture prior gives adaptation in the sense of Chapter 10.

We adopt the setting of Section 11.5.2, with a stationary Gaussian process $W = (W_t : t \in \mathbb{R}^d)$ with spectral density with exponentially small tails, so that the process has analytic sample paths. Rather than scaling the process deterministically, we now consider the process $W^A = (W_{At} : t \in [0,1]^d)$, for A a random variable independent of W. The resulting prior is a mixture of Gaussian processes. Specifically we study the case that the variable A^d follows a gamma distribution. The parameters of this gamma distribution are inessential, and the gamma distribution can be replaced by another distribution with tails of the same weights, but the power d in A^d appears important.

The following theorem gives the contraction rate for the prior process W^A in the abstract setting of a mixed Gaussian prior, and may be compared with Theorem 11.20. The theorem

can be translated to contraction rates in concrete settings, such as density estimation, classification, regression and white noise, in the same way as Theorems 11.21–11.24 are derived from Theorem 11.20. The theorem gives the existence of sets B_n and rates ϵ_n and $\bar{\epsilon}_n$ such that

$$\log N(\epsilon_n, B_n, \|\cdot\|_\infty) \le n\epsilon_n^2, \tag{11.21}$$

$$\mathrm{P}(W^A \notin B_n) \le e^{-4n\bar{\epsilon}_n^2}, \tag{11.22}$$

$$\mathrm{P}(\|W^A - w_0\|_\infty \le \bar{\epsilon}_n) \ge e^{-n\bar{\epsilon}_n^2}. \tag{11.23}$$

The rates are specified for two situations: the ordinary smooth case, where the true function w_0 belongs to a Hölder class, and the supersmooth case, where w_0 is analytic. For the first situation it was seen in Section 11.5.2 that a suitable deterministic shrinking rate turns the process W in an appropriate prior. For the second situation the unscaled process W was seen to be an appropriate prior at the end of Section 11.4.4. The following theorem shows that choosing the length scale according to dth root of a gamma variable, keeps the best of both worlds.

Theorem 11.57 (Mixed Gaussian contraction rate) *If $(W_t : t \in \mathbb{R}^d)$ is a stationary Gaussian process with spectral density m such that $a \mapsto m(a\lambda)$ is decreasing on $(0, \infty)$, for every $\lambda \in \mathbb{R}^d$, and $\int e^{\delta\|\lambda\|} m(\lambda)\, d\lambda < \infty$, for some $\delta > 0$, and A^d is an independent gamma variable, then there exist measurable sets $B_n \subset \mathfrak{C}([0,1]^d)$ such that* (11.21)–(11.23) *hold for the process $W_t^A = W_{At}$ and ϵ_n and $\bar{\epsilon}_n$ defined as follows.*

(i) *If $w_0 \in \mathfrak{C}^\beta([0,1]^d)$, then $\bar{\epsilon}_n \asymp n^{-\beta/(2\beta+d)}(\log n)^{(1+d)\beta/(2\beta+d)}$ and $\epsilon_n \asymp \bar{\epsilon}_n(\log n)^{(1+d)/2}$.*

(ii) *If w_0 is the restriction of a function in $\mathcal{A}^{\gamma,r}(\mathbb{R}^d)$ to $[0,1]^d$ and $m(\lambda) \ge C_3 \exp(-D_3\|\lambda\|^\nu)$ for constants $C_3, D_3, \nu > 0$, then $\bar{\epsilon}_n \asymp n^{-1/2}(\log n)^{(d+1)/2}$ and $\epsilon_n \asymp \bar{\epsilon}_n(\log n)^\kappa$, where $\kappa = 0$ for $r \ge \nu$, and $\kappa = d/(2r)$ for $r < \nu$.*

Proof If $\varphi_{w_0}^a$ is the concentration function of W^a and g the density of A, then by applying Lemma 11.19 we see

$$\mathrm{P}(\|W^A - w_0\| \le 2\epsilon) \ge \int_0^\infty e^{-\varphi_{w_0}^a(\epsilon)}\, g(a)\, da. \tag{11.24}$$

By Lemma 11.55 we have that $\varphi_0^a(\epsilon) \le C_4 a^d (\log(a/\epsilon))^{1+d}$, for $a > a_0$ and $\epsilon < \epsilon_0$, where the constants a_0, ϵ_0, C_4 depend only on μ and w.

For $\mathbb{B}_1$ the unit ball of $\mathfrak{C}([0,1]^d)$ and given constants $M, r, \delta, \epsilon > 0$, set

$$B = \Big(M\sqrt{\frac{r}{\delta}}\mathbb{H}_1^r + \epsilon\mathbb{B}_1\Big) \bigcup \Big(\bigcup_{a<\delta} (M\mathbb{H}_1^a) + \epsilon\mathbb{B}_1\Big). \tag{11.25}$$

By Lemma 11.59, $B \supset M\mathbb{H}_1^a + \epsilon\mathbb{B}_1$ for any $a \in [\delta, r]$, and also for $a < \delta$, by definition. Thus by Borell's inequality, for any $a \le r$,

$$\begin{aligned}\mathrm{P}(W^a \notin B) &\le \mathrm{P}(W^a \notin M\mathbb{H}_1^a + \epsilon\mathbb{B}_1) \le 1 - \Phi(\Phi^{-1}(e^{-\varphi_0^a(\epsilon)}) + M) \\ &\le 1 - \Phi(\Phi^{-1}(e^{-\varphi_0^r(\epsilon)}) + M),\end{aligned}$$

because $e^{-\varphi_0^a(\epsilon)} = \mathrm{P}(\sup_{t\in[0,a]^d}|W_t| \le \epsilon)$ is decreasing in a. Let ϵ_1 be such that $e^{-\varphi_0^1(\epsilon_1)} < 1/4$, which ensures $e^{-\varphi_0^r(\epsilon)} \le e^{-\varphi_0^1(\epsilon)} < 1/4$ for all $r > 1$ and $\epsilon < \epsilon_1$. Now using Lemma K.6 (ii), for $M \ge 4\sqrt{\varphi_0^r(\epsilon)}$, $r > 1$ and $\epsilon < \epsilon_1$, we have that $M \ge -2\Phi^{-1}(e^{-\varphi_0^r(\epsilon)})$. Consequently, the right-hand side of the preceding display is bounded by $1 - \Phi(M/2) \le e^{-M^2/8}$, in view of Lemma K.6 (i). By Lemma 11.55, we conclude that $\mathrm{P}(W^a \notin B) \le e^{-M^2/8}$ for any $a \le r$, provided that

$$M^2 \ge 16C_4 r^d (\log(r/\epsilon))^{1+d}, \qquad r > 1, \qquad \epsilon < \epsilon_1 \wedge \epsilon_0, \tag{11.26}$$

where ϵ_0 is obtained from Lemma 11.55. Since $A^d \sim \mathrm{Ga}(p, q)$, for $r > r_0$ and a constant r_0 that depends on d, p, q only,

$$\mathrm{P}(W^A \notin B) \le \mathrm{P}(A > r) + \int_0^r \mathrm{P}(W^a \notin B)\, g(a)\, da \lesssim r^{d(p-1)} e^{-qr^d} + e^{-M^2/8}. \tag{11.27}$$

This holds for any $B = B_{M,r,\delta,\epsilon}$ with M, r, δ, ϵ satisfying (11.26).

By the proof of Lemma 11.38, for $M\sqrt{r/\delta} > 2\epsilon$ and $r > a_0$,

$$\begin{aligned}\log N\Big(2\epsilon, M\sqrt{\frac{r}{\delta}}\mathbb{H}_1^r + \epsilon\mathbb{B}_1, \|\cdot\|_\infty\Big) &\le \log N\Big(\epsilon, M\sqrt{\frac{r}{\delta}}\mathbb{H}_1^r, \|\cdot\|_\infty\Big)\\ &\le Kr^d\bigg(\log\Big(\frac{M\sqrt{r/\delta}}{\epsilon}\Big)\bigg)^{1+d}.\end{aligned}$$

By Lemma 11.58, any function $h \in M\mathbb{H}_1^a$, for $a < \delta$ is within uniform distance $\delta\sqrt{d}\tau M$ of a constant function, where the constant value belongs to the interval $[-M\sqrt{\|\mu\|}, M\sqrt{\|\mu\|}]$, and $\tau^2 = \int \|\lambda\|^2\, d\mu(\lambda)$. It follows that, for $\epsilon > \delta\sqrt{d}\tau M$,

$$N\Big(3\epsilon, \bigcup_{a<\delta}(M\mathbb{H}_1^a) + \epsilon\mathbb{B}_1, \|\cdot\|_\infty\Big) \le \frac{2M\sqrt{\|\mu\|}}{\epsilon}.$$

Choose $\delta = \epsilon/(2\sqrt{d}\tau M)$. Combining the last two displays, and using the elementary inequality $\log(x + y) \le \log x + 2\log y$ for $x \ge 1$, $y \ge 2$, we obtain, for $\epsilon \le M\sqrt{\|\mu\|}$,

$$\log N(3\epsilon, B, \|\cdot\|_\infty) \le Kr^d\bigg(\log\Big(\frac{M^{3/2}\sqrt{2\tau r}d^{1/4}}{\epsilon^{3/2}}\Big)\bigg)^{1+d} + 2\log\frac{2M\sqrt{\|\mu\|}}{\epsilon}. \tag{11.28}$$

This inequality is valid for any $B = B_{M,r,\delta,\epsilon}$ with $\delta = \epsilon/(2\sqrt{d}\tau M)$, and any M, r, ϵ with

$$M^{3/2}\sqrt{2\tau r}d^{1/4} > 2\epsilon^{3/2}, \qquad r > a_0, \qquad M\sqrt{\|\mu\|} > \epsilon. \tag{11.29}$$

In the remainder of the proof we make special choices for these parameters, depending on the assumption on w_0, to complete the choice of B_n and obtain $\bar{\epsilon}_n$ and ϵ_n.

(i). *Hölder smoothness.* Let $w_0 \in \mathfrak{C}^\beta([0,1]^d)$ for some $\beta > 0$. In view of Lemmas 11.56 and 11.55, for every a_0, there exist $\epsilon_0 < \frac{1}{2}$, C, D, $K > 0$ depending on w and μ only such that, for $a > a_0$, $\epsilon < \epsilon_0$, and $\epsilon > Ca^{-\beta}$,

$$\varphi_{w_0}^a(\epsilon) \le Da^d + C_4 a^d\Big(\log\frac{a}{\epsilon}\Big)^{1+d} \le K_1 a^d\Big(\log\frac{a}{\epsilon}\Big)^{1+d},$$

for K_1 depending on a_0, μ and d only. Therefore, for $\epsilon < \epsilon_0 \wedge Ca_0^{-\beta}$ (so that $(C/\epsilon)^{1/\beta} > a_0$), by (11.24),

$$\mathrm{P}(\|W^A - w_0\|_\infty \le 2\epsilon) \ge \int_{(C/\epsilon)^{1/\beta}}^{2(C/\epsilon)^{1/\beta}} e^{-K_1 a^d \log^{1+d}(a/\epsilon)}\, g(a)\, da$$
$$\ge C_1 e^{-K_2 \epsilon^{-d/\beta} (\log_-\epsilon)^{(1+d)}} \left(\frac{C}{\epsilon}\right)^{(dp-1)/\beta} \left(\frac{C}{\epsilon}\right)^{1/\beta},$$

for a constant K_2 that depends only on K_1, C, D_1, d, β. Therefore, for all sufficiently large n, we have that $\mathrm{P}(\|W^A - w_0\|_\infty \le \bar\epsilon_n) \ge e^{-n\bar\epsilon_n^2}$ for $\bar\epsilon_n$ a large multiple of $n^{-1/(2+d/\beta)}(\log n)^\gamma$, where $\gamma = (1+d)/(2+d/\beta)$.

Inequalities (11.26)–(11.27) give that $\mathrm{P}(W^A \notin B) \lesssim e^{-C_0 n\bar\epsilon_n^2}$ for an arbitrarily large constant C_0 if (11.26) holds and

$$D_2 r^d (\log r)^q \ge 2C_0 n\bar\epsilon_n^2, \qquad r^{p-d+1} \le e^{C_0 n\bar\epsilon_n^2}, \qquad M^2 \ge 8C_0 n\bar\epsilon_n^2. \tag{11.30}$$

Given C_0, choose $r = r_n$ equal to the minimal solution of the first equation in (11.30), and then choose $M = M_n$ to satisfy the third. The second equation is then automatically satisfied, for large n.

Let B_n be the set B from (11.25) with the preceding choices of M_r and r_n, and ϵ_n bounded below by a power of n. Then the right side of (11.28) is bounded above by a multiple of $r_n^d(\log n)^{1+d} + \log n \le n\epsilon_n^2$ for ϵ_n^2 a large multiple of $(r_n^d/n)(\log n)^{1+d}$. Inequalities (11.29) are clearly satisfied.

(ii)-1. *Infinite smoothness: $r \ge \nu$.*

By combining the first part of Lemma 11.41 and Lemma 11.55, we see that there exist positive constants $a_0 < a_1, \epsilon_0, K_1$ and C_4 that depend on w and μ only such that $\varphi^a_{w_0}(\epsilon) \le K_1 + C_4 a^d (\log(a/\epsilon))^{1+d}$, for $a \in [a_0, a_1]$ and $\epsilon < \epsilon_0$. Consequently, by (11.24),

$$\mathrm{P}(\|W^A - w_0\|_\infty \le 2\epsilon) \ge e^{-K_1 - C_4 a_1^d (\log(a_1/\epsilon))^{1+d}}\, \mathrm{P}(a_0 < A < a_1).$$

This gives $\mathrm{P}(\|W^A - w_0\|_\infty \le \bar\epsilon_n) \ge e^{-n\bar\epsilon_n^2}$ for $\bar\epsilon_n$ a large multiple of $n^{-1/2}(\log n)^{(d+1)/2}$, provided that n is sufficiently large.

Next we choose B_n as before, with r and M solving (11.30) and satisfying (11.26), i.e. r_n^d and M_n^2 large multiples of $(\log n)^{d+1}$. Then (11.26)–(11.27) imply $\mathrm{P}(W^A \notin B_n) \gtrsim e^{-C_0 n\bar\epsilon_n^2}$, and the right side of (11.28) is bounded by a multiple of $r_n^d(\log_-\epsilon + \log\log n)^{1+d} + \log_-\epsilon + \log\log n$. For $\epsilon = \epsilon_n$ a large multiple of $n^{-1/2}(\log n)^{d+1}$, this is bounded above by $n\epsilon_n^2$.

(ii)-2. *Infinite smoothness: $r < \nu$.*

Combining the second part of Lemma 11.41 and Lemma 11.55, we see that there exist $a_0, \epsilon_0, C, D, K_1, C_4 > 0$ depending only on w and μ, and $\gamma' > \gamma$ such that, for $a > a_0$, $\epsilon < \epsilon_0$ and $Ce^{-\gamma' a^r} < \epsilon$, we have $\varphi^a_{w_0}(\epsilon) \le Da^d + C_4 a^d (\log(a/\epsilon))^{1+d}$. Consequently, by (11.24), with D_1, D_2 depending on w and μ only,

$$\mathrm{P}(\|W^A - w_0\|_\infty \le 2\epsilon) \ge \int_{(\log(C/\epsilon)/\gamma')^{1/r}}^{\infty} e - \varphi^a_{w_0}(\epsilon) g(a)\, da \ge D_2 e^{-D_1(\log_-\epsilon)^{d/r+d+1}}.$$

For all sufficiently large n this gives that $\mathrm{P}(\|W^A - w_0\|_\infty \le \bar\epsilon_n) \ge e^{-n\bar\epsilon_n^2}$, for $\bar\epsilon_n$ a large multiple of $n^{-1/2}(\log n)^{d/(2r)+(d+1)/2}$.

Next we choose B_n of the form as before, with $r = r_n$ and $M = M_n$ solving (11.30), i.e., r_n^d and M_n^2 chosen equal to large multiples of $(\log n)^{d/r+d+1}$. Then (11.26)–(11.27) imply that $\mathrm{P}(W^A \notin B) \le e^{-C_0 n\epsilon_n^2}$, and the right side of (11.28) is bounded above by a multiple of $r_n^d(\log_- \epsilon + \log\log n)^{1+d} + \log_- \epsilon + \log\log n$. For $\epsilon = \epsilon_n$ a large multiple of $n^{-1/2}(\log n)^{d+1+d/(2r)}$, this is bounded above by $n\epsilon_n^2$. □

The following three lemmas are used in the preceding proof. The third is similar to Lemma 11.41, and extends this to the case of analytic true functions.

Lemma 11.58 *For any $h \in \mathbb{H}_1^a$ and $t \in \mathbb{R}^d$, we have $|h(0)|^2 \le \|\mu\|$ and $|h(t) - h(0)|^2 \le a^2\|t\|^2 \int \|\lambda\|^2\, d\mu(\lambda)$.*

Proof An element $H\psi$ of $\mathbb{H}^a$ takes the form $H\psi(t) = \int e^{i\langle t,\lambda\rangle}\psi(\lambda)\, d\mu_a(\lambda)$ and has square norm $\|H\psi\|_{\mathbb{H}^a}^2 = \int |\psi|^2\, d\mu_a$. It follows that $|H\psi(0)| \le \int |\psi|\, d\mu_a$ and

$$|H\psi(t) - H\psi(0)| \le \int |e^{i\langle t,\lambda\rangle} - 1|\, |\psi(\lambda)|\, d\mu_a(\lambda) \le \int |\langle t, \lambda\rangle|\, |\psi(\lambda)|\, d\mu_a(\lambda).$$

By two successive applications of the Cauchy-Schwarz inequality, the square of the right side is bounded above by $\|t\|^2 \int \|\lambda\|^2 d\mu_a(\lambda) \int |\psi|^2\, d\mu_a$. This gives the result. □

Lemma 11.59 *If the spectral density of W is radially decreasing and $\mathbb{H}^a$ is the RKHS of the rescaled stationary Gaussian process W^a, then $\sqrt{a}\,\mathbb{H}_1^a \subset \sqrt{b}\,\mathbb{H}_1^b$, for any $a \le b$.*

Proof The assumption that $a \mapsto m(a\lambda)$ is decreasing in $a \in (0,\infty)$ implies that $m_a(\lambda) \le (b/a)m_b(\lambda)$, for every λ. By Lemma 11.35 an arbitrary element of $\mathbb{H}^a$ takes the form

$$(H^a\psi)(t) = \int e^{i\langle t,\lambda\rangle}\psi(\lambda)\, d\mu_a(\lambda) = \int e^{i\langle t,\lambda\rangle}\Big(\psi\frac{m_a}{m_b}\Big)(\lambda)\, d\mu_b(\lambda).$$

Because $(\psi m_a/m_b) \in \mathbb{L}_2(\mu_b)$ if $\psi \in \mathbb{L}_2(\mu_a)$, it follows that $H^a\psi = H^b(\psi m_a/m_b)$ is contained in $\mathbb{H}^b$. The square norms of this function in the two spaces satisfy $\|H^a\psi\|_{\mathbb{H}^b}^2 = \|\psi m_a/m_b\|_{\mu_b,2}^2 \le (b/a)\|\psi\|_{\mu_a,2}^2 = \|H\psi\|_{\mathbb{H}^a}^2$. This gives the nesting of the two scaled unit balls. □

Lemma 11.60 (Decentering) *Suppose that $m(\lambda) \ge C_3\exp(-D_3\|\lambda\|^\nu)$ for some constants $C_3, D_3, \nu > 0$, and let w be the restriction to $[0,1]^d$ of a function in $\mathcal{A}^{\gamma,r}(\mathbb{R}^d)$.*

(a) *If $r \ge \nu$, then $w \in \mathbb{H}^a$ for all sufficiently large a with uniformly bounded norm $\|w\|_{\mathbb{H}^a}$.*
(b) *If $r < \nu$, then there exist $a_0, C, D > 0$ depending only on μ and w such that, for $a > a_0$,*

$$\inf_{\|h-w\|_\infty \le Ce^{-\gamma a^r}/a^{-r+1}} \|h\|_{\mathbb{H}^a}^2 \le Da^d.$$

11.7 Computation

This section provides a brief review of algorithms to compute or approximate aspects of the posterior distribution corresponding to a Gaussian prior.

11.7.1 Kernel Methods and the Posterior Mode

In regression problems the posterior mode corresponding to a Gaussian process prior can be computed as the solution to a minimization problem involving the corresponding RKHS-norm. In particular, for the integrated Brownian motion prior the posterior mode is a penalized spline estimator.

Consider a regression model with independent observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ following a density $(x, y) \mapsto p_w(x, y)$ that depends on the unknown parameter w only through its value $w(x)$. Define a loss $L(y, w(x)) = -2 \log p_w(x, y)$ and consider the criterion

$$x \mapsto \sum_{i=1}^{n} L(Y_i, w(X_i)) + n\lambda \|w\|_{\mathbb{H}}^2. \tag{11.31}$$

Here the *smoothing parameter* λ is fixed, and $\|\cdot\|_{\mathbb{H}}$ is the RKHS-norm corresponding to a given Gaussian process W.

Theorem 11.61 *The minimizer $\hat{w}$ of the criterion* (11.31) *over $w \in \mathbb{H}$ is the posterior mode under the prior $w \sim (\sqrt{n}\lambda)^{-1} W$, for W the Gaussian process with RKHS $\mathbb{H}$. This minimizer can be written as a linear combination of the functions $x \mapsto K(X_i, x)$, for $i = 1, \ldots, n$ and K the covariance kernel of W.*

Proof We shall show that $\hat{w}$ and the posterior mode agree at the points $X_1, \ldots, X_n$ and also at an arbitrary additional point X_0, whence at all points.

The reproducing formula gives that $w(X_i) = \langle K(X_i, \cdot), w\rangle_{\mathbb{H}}$, for every $i = 0, 1, \ldots, n$. If Pw is the orthogonal projection of w in $\mathbb{H}$ onto the linear span of the functions $K(X_i, \cdot)$, for $i = 0, \ldots, n$, then $w(X_i) = \langle K(X_i, \cdot), Pw\rangle_{\mathbb{H}}$, and hence the function Pw attains the same value on the loss function $\sum_{i=1}^n L(Y_i, w(X_i))$ as the function w. Since $\|Pw\|_{\mathbb{H}} \le \|w\|_{\mathbb{H}}$ it follows that Pw attains a smaller value in the criterion (11.31) than w. Hence the minimizer of (11.31) can be written $\hat{w} = \sum_{j=0}^n \hat{\nu}_j K(X_j, \cdot)$, for numbers $\hat{\nu}_0, \ldots, \hat{\nu}_n$ that minimize

$$\begin{aligned}\sum_{i=1}^{n} L\Big(Y_i, \sum_{j=0}^{n} \nu_j K(X_j, X_i)\Big) + n\lambda \Big\| \sum_{j=0}^{n} \nu_j K(X_j, \cdot)\Big\|_{\mathbb{H}}^2 \\ = \sum_{i=1}^{n} L\big(Y_i, (K_{n+1}\nu)_i\big) + n\lambda \nu^\top K_{n+1}\nu,\end{aligned}$$

where K_{n+1} is the matrix with (i, j)th element $K(X_i, X_j)$ and $\nu = (\nu_0, \ldots, \nu_n)^\top$.

By assumption, the likelihood depends on w only through (the last n coordinates) of the vector $w_{n+1} = (w(X_0), w(X_1), \ldots, w(X_n))^\top$. Under the prior the vector w_{n+1} possesses a $\mathrm{Nor}_{n+1}(0, K_{n+1}/(n\lambda))$-distribution, and hence is distributed as $K_{n+1}\nu$ for $\nu \sim \mathrm{Nor}_{n+1}(0, K_{n+1}^{-1}/(n\lambda))$. The corresponding posterior for ν is proportional to, with w the vector $K_{n+1}\nu$ minus its first coordinate,

$$\prod_{i=1}^{n} p_w(X_i, Y_i) e^{-\frac{1}{2} n\lambda \nu^\top K_{n+1}\nu} = \exp\Big(-\tfrac{1}{2}\Big[\sum_{i=1}^{n} L(Y_i, (K_{n+1}\nu)_i) + n\lambda \nu^\top K_{n+1}\nu\Big]\Big).$$

Clearly, the posterior mode for ν is the same as the minimizer $\hat{\nu}$ found previously. The resulting mode for w_{n+1} is $K_{n+1}\hat{\nu}$.

Thus the penalized estimator $\hat{w}$ and the posterior mode agree at the points $X_0, X_1, \ldots, X_n$. □

Example 11.62 The preceding theorem applies to the Gaussian regression problem, with loss function $L(y, w(x)) = (y - w(x))^2$. The posterior distribution is then Gaussian, and hence the posterior mode is identical to the posterior mean.

The theorem also applies to the binary regression model. For instance, for logistic regression with responses coded as $y \in \{-1, 1\}$, we can set $L(y, w(x)) = \log(1 + e^{-w(x)y})$.

In both cases the theorem shows that the computation of the posterior mode can be reduced to solving a finite-dimensional optimization problem.

The RKHS norm of $(k-1)$ times integrated Brownian motion $I_{0+}^{k-1} B$ is the Sobolev norm of order k (see Lemma 11.29). For this prior the criterion in (11.31) takes the form

$$\sum_{i=1}^{n} L(Y_i, w(X_i)) + n\lambda \int_0^1 |w^{(k)}(t)|^2 \, dt. \tag{11.32}$$

By the preceding theorem, the posterior mode for this prior is the minimizer of this expression over the RKHS of integrated Brownian motion. This consists of all functions in the Sobolev space $\mathfrak{W}^k[0, 1]$ with $w(0) = w'(0) = \cdots = w^{(k-1)}(0) = 0$. We may also minimize the preceding display over all functions in the Sobolev space, not necessarily with derivatives tied to zero at zero. This minimizer can be related to the integrated Brownian motion prior released at zero.

Theorem 11.63 *The minimizer of* (11.32) *is the limit as* $b \to \infty$ *of the posterior mode under the prior* $w \sim (\sqrt{n}\lambda)^{-1} W$, *for* $W_t = I_{0+}^{k-1} B_t + b \sum_{j=0}^{k-1} Z_j t^j$, *where* B *is a Brownian motion independent of* $Z_j \overset{iid}{\sim} \mathrm{Nor}(0, 1)$.

Proof As in the preceding proof let $w_{n+1} = (w(X_0), w(X_1), \ldots, w(X_n))^\top$. The first part of the preceding theorem still applies, but in the second part we introduce the extra parameter vector $\mu \sim \mathrm{Nor}_k(0, b^2 I)$ to form the prior for w_{n+1} through $w_{n+1} = P_{n+1}\mu + K_{n+1}\nu$, for P_{n+1} the matrix with ith row $(1, X_i, \ldots, X_i^{k-1})$, for $i = 0, 1, \ldots, n$. The posterior distribution for (μ, ν) is proportional to

$$\exp\Big(-\tfrac{1}{2}\Big[\sum_{i=1}^{n} L(Y_i, (K_{n+1}\nu)_i) + \mu^\top \mu / b^2 + n\lambda \nu^\top K_{n+1} \nu\Big]\Big).$$

For $b \to \infty$ the second term on the right disappears. The maximizer of the expression then converges to the minimizer of (11.32). □

Example 11.64 (Integrated Brownian motion and splines) The kernel of Brownian motion can be written $\mathrm{E}[B_s B_t] = s \wedge t = \int_0^1 \mathbb{1}_{[0,s]}(u) \mathbb{1}_{[0,t]}(u) \, du$. By Fubini's theorem the kernel of integrated Brownian motion is then

$$\mathrm{E}(I_{0+}B_s)(I_{0+}B_t) = \int_0^1 \int_0^s \mathbb{1}_{[0,x]}(u)\,dx \int_0^t \mathbb{1}_{[0,y]}(u)\,dy\,du = \int_0^1 (s-u)_+(t-u)_+\,du.$$

Repeating this argument we find that the kernel of the RKHS of $I_{0+}^{k-1}B$ is equal to

$$K(s,t) = \int_0^1 \frac{(s-u)_+^{k-1}}{(k-1)!}\frac{(t-u)_+^{k-1}}{(k-1)!}\,du.$$

This shows that the function $t \mapsto K(s,t)$ for fixed s is a spline of order $2k$ (i.e. pieces of polynomial of degree $2k-1$) with knot s. By Theorem 11.61 the posterior mode is a linear combination of these functions with $s \in \{X_1, \ldots, X_n\}$, and hence itself a spline function with knots $X_1, \ldots, X_n$.

11.7.2 Density Estimation

Suppose the data consists of a random sample $X_1, \ldots, X_n$ from the density p_w considered in Section 11.3.1. The likelihood function is given by

$$L(w|\, X_1, \ldots, X_n) = \prod_{i=1}^n \frac{e^{w(X_i)}}{\int e^w\,d\nu}.$$

Importance Sampling

The posterior expectation of a function $f(W)$, such as the density $p_W(x)$ at a point, is given by

$$\frac{\mathrm{E}_W\big[f(W)L(W|\, X_1, \ldots, X_n)\big]}{\mathrm{E}_W\big[L(W|\, X_1, \ldots, X_n)\big]} = \frac{\mathrm{E}_{W^*}\big[f(W^*)(\int e^{W^*(x)}\,d\nu(x))^{-n}\big]}{\mathrm{E}_{W^*}\big[(\int e^{W^*(x)}\,d\nu(x))^{-n}\big]},$$

where E_W denotes the expectation relative to the prior, and E_{W^*} the expectation with respect to the process W^* whose law has density proportional to $\prod_{i=1}^n e^{W(X_i)}$ relative to the law of W.

An elementary calculation shows that W^* is also a Gaussian process, with mean function $\mathrm{E}[W^*(x)] = \mathrm{E}[W(x)] + \sum_{i=1}^n K(x, X_i)$ and the same covariance kernel K as W. The right side of the preceding display can be approximated by generating a large number of copies of W^*s and replacing the two expectations by averages. In practice we generate the Gaussian processes on a grid, and approximate the integrals by sums.

Finite Interpolation

Let W^* be the kriging of W on a fine grid, as discussed in Example 11.10. If the original process W has continuous sample paths, and the mesh size of the grid is fine enough, then W^* will be close to W, and hence p_{W*} close to p_W, by Lemma 2.5.

The kriging W^* is a finite linear combination $W_t^* = \sum_{j=1}^m a_j(t)W_{t_j}$ of the values of W at the grid points, and hence to compute the posterior distribution with respect to the prior W^* it suffices to update the prior distribution of the vector $W^m = (W_{t_1}, \ldots, W_{t_m})$ with the data. For K_m the covariance matrix of W^m, its posterior distribution satisfies

$$p(W^m \mid X_1, \ldots, X_n) \propto \prod_{i=1}^{n} p_{\sum_j a_j W_{t_j}}(X_i)\, \pi(W^m)$$
$$= \frac{e^{\sum_{i=1}^n \sum_j a_j(X_i) W_{t_j}}}{\prod_{i=1}^n \int e^{\sum_j a_j(x) W_{t_j}}\, d\nu(x)}\, e^{-(W^m)^\top K_m W^m/2}.$$

This can be the basis for execution of a Metropolis-Hastings algorithm, where at each step the right side is evaluated by (numerically) computing the integrals. A random walk Metropolis-Hastings step is particularly easy to implement. The formula can be adapted to include a tuning parameter in the kernel function, thus allowing a conditional Gaussian prior of W^m given this tuning parameter.

For increased accuracy and efficiency an adaptive choice of the grid can be incorporated into the algorithm by treating also the locations $t_1, \ldots, t_m$ as variables, which then are updated in addition to W^m and tuning parameters. It is natural to move through the grid points by birth, death or swap moves, with corresponding proposal probabilities. Acceptance probabilities for movements across spaces of different dimensions can be calculated using the general recipe of reversible jump MCMC.

11.7.3 Nonparametric Binary Regression

The posterior distribution in the binary regression model of Section 2.5 is not analytically tractable, due to the nonlinearity of the link function. This can be overcome by an MCMC procedure based on data augmentation.

Because the likelihood depends on the process W only through its values at the observed covariates $X^n = (X_1, \ldots, X_n)$, the posterior distribution of W given $W^n = (W_{X_1}, \ldots, W_{X_n})$ and the data $Y^n = (Y_1, \ldots, Y_n)$ is the same as the prior of W given W^n. Since this is Gaussian and can be computed from the joint multivariate normal distribution, we may focus attention on the posterior distribution of W^n. Let μ_n and K_n be the prior mean vector and covariance matrix of this vector.

The case of the *probit link* function is the most straighforward. In this case the data Y^n can be viewed as arising from the following hierarchical scheme:

$$Y_i = \mathbb{1}\{Z_i > 0\}, \qquad Z_i \mid X^n, W^n \overset{\text{ind}}{\sim} \text{Nor}(W_{X_i}, 1), \qquad W^n \mid X^n \sim \text{Nor}_n(\mu_n, K_n).$$

Given (X^n, W^n) the resulting variables $Y_1, \ldots, Y_n$ are independent with $\mathrm{P}(Y_i = 1 \mid X^n, W^n) = \Phi(W_{X_i})$, as in the original model. The advantage of the latent variables $Z^n = (Z_1, \ldots, Z_n)$ (and the probit link) is that their joint distribution with W^n is Gaussian, so that the conditional distributions needed for a Gibbs sampler are Gaussian as well, and given by standard formulas. In its simplest form, with distinct covariates X_i, the steps of this sampler are:

(i) $W^n \mid X^n, Y^n, Z^n \sim \text{Nor}_n(\mu^*, K^*)$, for $K^* = (I + K_n^{-1})^{-1}$, $\mu^* = K^*(Z^n - \mu_n) + \mu_n$.
(ii) $Z_1, \ldots, Z_n \mid W^n, X^n, Y^n$ are independent and distributed according to the $\text{Nor}(W_{X_i}, 1)$-distribution truncated to $[0, \infty)$ if $Y_i = 1$, and according the same distribution truncated to $(-\infty, 0]$ if $Y_i = 0$.

If a covariate X_i appears multiple times, then the corresponding observations and values W_{X_i} must be grouped. In the computation of K^* for large n, direct computation of K^{-1} must be avoided but instead the spectral decomposition should be used.

If the mean function and the covariance kernel have additional hyperparameters, these should be updated also in additional Gibbs steps, possibly using nested Metropolis-Hastings samplers.

The case of the *logit link* function is often resolved by approximating the logistic distribution by a scale mixture of Gaussian distributions. The distribution function of a mixture with, for instance, five (well-chosen and fixed) normal components is practically indistinguishable from the logistic distribution function. The advantage of data augmentation with a Gaussian variable can then be retained.

11.7.4 Expectation Propagation

In many problems the parameter is a (high-dimensional) vector $\theta = (\theta_1, \ldots, \theta_n)$, and the likelihood depends on the parameter as a product of terms depending on the individual coordinates θ_i of the vector. The *expectation propagation* (EP) algorithm tries to approximate the marginal posterior distributions of the θ_i.

If $\theta \sim \mathrm{Nor}_n(\mu, \Sigma)$ under the prior, with prior density denoted by ϕ, and the likelihood takes the form $\prod_{j=1}^n t_j(\theta_j)$, where the observations are suppressed from the notation, then the posterior density p satisfies

$$p(\theta) \propto \prod_{j=1}^{n} t_j(\theta_j)\, \phi(\theta). \tag{11.33}$$

We assume that the numerical value of the right side of (11.33) is computable, for every θ. The problem we wish to solve is that its normalizing constant is an n-dimensional integral and hence is expensive to compute, unless it is accessible to analytical computation.

The main interest is in computing the marginal densities

$$p_i(\theta_i) \propto \int \prod_{j=1}^{n} t_j(\theta_j)\, \phi(\theta)\, d\theta_{-i} = t_i(\theta_i)\phi_i(\theta_i) \int \prod_{j\neq i} t_j(\theta_j)\, \phi_{-i|i}(\theta_{-i}\,|\,\theta_i)\, d\theta_{-i}. \tag{11.34}$$

Here θ_{-i} denotes the vector $\theta \in \mathbb{R}^n$ without its ith coordinate, and ϕ_i and $\phi_{-i|i}$ are the marginal and conditional (prior) density of θ_i, and of θ_{-i} given θ_i. These are a univariate and multivariate normal density, respectively. The integral is $(n-1)$-dimensional and hence numerically expensive to compute. On the other hand, the normalizing constant for the right side as a function of θ_i is a low-dimensional integral, whose computation should be feasible.

The EP-algorithm can be viewed as a method to compute an approximation to the normalizing constant in (11.33). If applied to the "conditional" density $\theta_{-i} \mapsto \prod_{j\neq i} t_j(\theta_j)\, \phi_{-i|i}(\theta_{-i}\,|\,\theta_i)$ it gives an approximation to the integral in (11.34).

The approximation is based on a *Gaussian approximation* to the posterior (11.33), i.e. a function q that is proportional to a Gaussian density that hopefully is close to (11.33). Because the integral of such a "Gaussian function" is easy to compute analytically, it is irrelevant whether q is normalized or not.

The approximation is computed recursively, where we cycle repeatedly through the index $i = 1, \ldots, n$. Every approximation q is of the form, for functions $\tilde{t}_i$ (called *term approximations*) that are proportional to univariate Gaussian densities,

$$q(\theta) \propto \prod_{i=1}^{n} \tilde{t}_i(\theta_i)\, \phi(\theta). \tag{11.35}$$

Given a Gaussian approximation q of the form (11.35) and some i we form a new density $\tilde{q}$ by replacing the current $\tilde{t}_i$ by the true t_i:

$$\tilde{q}(\theta) \propto q(\theta) \frac{t_i(\theta_i)}{\tilde{t}_i(\theta_i)}.$$

Unless the factors t_i are "Gaussian", the resulting density $\tilde{q}$ will not be Gaussian. We define q^{new} as the multivariate Gaussian density with the same moments as the distribution with density proportional to $\theta \mapsto \tilde{q}(\theta)$. Depending on the functions t_i, the appropriate moments may have to be computed numerically. To keep the validity of (11.35) we update $\tilde{t}_i$ to $\tilde{t}_i^{\text{new}}$ so that $q^{\text{new}}(\theta) \propto \prod_{j \neq i} \tilde{t}_j(\theta_j) \tilde{t}_i^{\text{new}}(\theta_i) \phi(\theta)$, i.e., in view of (11.35),

$$\tilde{t}_i^{\text{new}}(\theta_i) \propto \frac{q^{\text{new}}}{q}(\theta) \tilde{t}_i(\theta_i). \tag{11.36}$$

Here there is a function of θ_i on the left and seemingly a function of the full vector θ on the right. However, for the (Gaussian) $q/\tilde{t}_i$ satisfying (11.35) it turns out that the function on the right depends on θ_i only. (See the lemma below, with the $\text{Nor}(\mu, \Sigma)$-density taken equal to the density propertional to $q/\tilde{t}_i$ and $U\theta = \theta_i$; also see Problem 11.8 for explicit formulas.) This algorithm is iterated until "convergence," although apparently it is not known whether convergence will always take place.

Lemma 11.65 *Let $Z \sim \text{Nor}_n(\mu, \Sigma)$, $U: \mathbb{R}^n \to \mathbb{R}^m$ be a linear map of full rank and $t: \mathbb{R}^m \to [0, \infty)$ be a measurable map such that $t(UZ)$ has finite second moment. If μ_1 and Σ_1 are the mean and covariance matrix of the distribution with density proportional to $z \mapsto t(Uz)\phi_{\mu,\Sigma}(z)$, then $(\phi_{\mu_1,\Sigma_1}/\phi_{\mu,\Sigma})(z)$ is a function of Uz.*

Proof Minimization of the Kullback-Leibler divergence $K(Q; N)$ between a given measure Q on $\mathbb{R}^n$ over the set of all n-variate normal distributions can be shown to yield the normal distribution with mean and covariance equal to the mean and covariance of Q. Therefore the normal distribution $\text{Nor}(\mu_1, \Sigma_1)$ is closest in Kullback-Leibler divergence to the distribution with density proportional to $z \mapsto t(Uz)\phi_{\mu,\Sigma}(z)$, i.e. (μ_1, Σ_1) minimizes

$$\int \Big[\log \frac{t(Uz)\phi_{\mu,\Sigma}(z)}{\phi_{\mu_1,\Sigma_1}(s)}\Big] t(Uz)\, \phi_{\mu,\Sigma}(z)\, dz.$$

Let ψ_0 and ψ_1 be the densities of $Y = UZ$ under (μ, Σ) and (μ_1, Σ_1), respectively. We can factorize $\phi_{\mu,\Sigma}(z) = \psi_0(y)\phi_0(z|\, y)$ and $\phi_{\mu_1,\Sigma_1}(z) = \psi_1(y)\phi_1(z|\, y)$, for $y = Uz$ and ϕ_0 and ϕ_1 conditional densities given a suitable dominating measure ν. The preceding display can be written as $\int \big[\log t(y)\big] t(y)\, \psi_0(y)\, dy$ plus

$$\int \Big[\log \frac{\psi_0(y)}{\psi_1(y)}\Big] t(y)\, \psi_0(y)\, dy + \int \Big(\int \Big[\log \frac{\phi_0(z|\, y)}{\phi_1(z|\, y)}\Big] \phi_0(z|\, y)\, d\nu(z)\Big) t(y)\psi_0(y)\, dy.$$

The marginal and conditional densities ψ_1 and ϕ_1 range over all Gaussian (conditional) densities if (μ_1, Σ_1) ranges over all possible combinations of a mean vector and covariance matrix. Thus we can minimize the sum in the display over (μ_1, Σ_1) by minimizing both terms separately. The inner integral in the second term is nonnegative for every fixed y and hence it can be minimized (to 0) by choosing $\phi_1(\cdot\,|\,y) = \phi_0(\cdot\,|\,y)$. Thus minimization is equivalent to minimizing the first term. The minimizing values satisfy $(\phi_{\mu_1,\Sigma_1}/\phi_{\mu,\Sigma})(z) = (\psi_1/\psi_0)(y)$. □

Since q approximates the posterior, its marginal density is an approximation to the marginal density $\theta_i \mapsto p_i(\theta_i)$ in (11.34). This gives a Gaussian approximation, but it appears to be not accurate.

An alternative is to apply EP as described in the preceding to compute the norming constant not of the full posterior, but of the conditional density $\theta_{-i} \mapsto \prod_{j\neq i} t_j(\theta_j)\,\phi_{-i|i}(\theta_{-i}\,|\,\theta_i)$. This may next be substituted in the far right side of (11.34). (The "conditional prior" $\psi_{-i|i}$ is Gaussian, so that the preceding applies.)

Other approximations are suggested by the identities (for the first one use (11.35) to eliminate ϕ)

$$p_i(\theta_i) \propto \int \prod_{j=1}^n t_j(\theta_j)\,\phi(\theta)\,d\theta_{-i} \propto \int \prod_{j=1}^n \frac{t_j(\theta_j)}{\tilde{t}_j(\theta_j)}\,q(\theta)\,d\theta_{-i}$$
$$= \frac{t_i(\theta_i)}{\tilde{t}_i(\theta_i)}\,q_i(\theta_i) \int \prod_{j\neq i} \frac{t_j(\theta_j)}{\tilde{t}_j(\theta_j)}\,q_{-i|i}(\theta_{-i}\,|\,\theta_i)\,d\theta_{-i}.$$

Here q_i and q_{-i} are the marginal and conditional density resulting from q. Both are Gaussian and can be computed analytically from q. The function $\theta_i \mapsto q_{-i|i}(\theta_{-i}\,|\,\theta_i)$ is proportional to a univariate Gaussian density, and hence the integral is a mixture of Gaussian densities. A crude approximation is to approximate every $t_j/\tilde{t}_j$ by 1 and evaluate the integral also to 1. This gives the approximation

$$\frac{t_i(\theta_i)}{\tilde{t}_i(\theta_i)}\,q_i(\theta_i).$$

This is the marginal density of the Gaussian approximation q, but with the "correct" t_i put back.

Better would be to approximate the integral. This is possible using EP applied to the density that is proportional to

$$\theta_{-i} \mapsto \prod_{j\neq i} \frac{t_j(\theta_j)}{\tilde{t}_j(\theta_j)}\,q_{-i|i}(\theta_{-i}\,|\,\theta_i).$$

In other words, the original ϕ is replaced by $q_{-i|i}$ and the original t_j are replaced by the quotients $t_j/\tilde{t}_j$. Because this procedure requires nested approximations, it is expensive. It is therefore recommended not to run the iterations of EP to convergence, but perform only one round of updates of the coordinates $t_j/\tilde{t}_j$.

Several other modifications have been suggested in the literature.

Many Gaussian priors come with a low-dimensional hyperparameter τ. If this is determined by an empirical Bayes method, then this causes no additional difficulty. In a full Bayes approach the marginal densities take the form

$$p_i(\theta_i) = \int p_i(\theta_i \mid \tau)\, p_\tau(\tau)\, d\tau. \tag{11.37}$$

Here $\theta_i \mapsto p_i(\theta_i \mid \tau)$ is the marginal density given τ, and p_τ is the *posterior* density of τ (dependence on the data is not shown!). The first was approximated in the preceding, with τ fixed to a particular value. The second is a conditional density given the data, and has the form, for π the prior on τ and the likelihood proportional to $\prod_{j=1}^n t_j(\theta_j)$,

$$p_\tau(\tau) \propto \int \prod_{j=1}^n t_j(\theta_j)\, \phi(\theta \mid \tau)\, d\theta\, \pi(\tau).$$

The integral is the normalizing constant of the right-hand side in the equation (11.33) for the posterior density. Hence up to the normalizing constant the expression can be approximated using the EP approximation.

Thus we can approximate the two terms in the integrand of (11.37), up to normalizing constants, for every τ. If τ is low-dimensional, then the integral can be approximated by a discretization. The normalizing constant to (11.37) can next also be determined by numerical integration.

11.7.5 Laplace Approximation

The *Laplace approximation* to the posterior distribution is the Gaussian approximation that also appears in the Bernstein–von Mises theorem. The approximation to the posterior density can be obtained by exponentiating and renormalizing a second order Taylor expansion of the sum of the log likelihood and log prior density around its mode.

Consider this in detail for the situation of a likelihood that factorizes in the coordinates of a parameter vector, as in (11.33), combined with a Gaussian prior with density ϕ. Let $\underline{p}$ be the right side of (11.33), so that $p \propto \underline{p}$, and compute a second order Taylor expansion in θ to

$$\log \underline{p}(\theta) = \sum_{j=1}^n \log t_j(\theta_j) + \log \phi(\theta),$$

around its point of maximum (the *mode* of p). Putting the quadratic in the exponent gives a Gaussian approximation to the full posterior, up to the normalizing constant.

The Gaussian term $\log \phi$ is quadratic already, so it suffices to expand the univariate functions $\log t_j$. The expansion of $\log \underline{p}$ around its mode takes the form

$$\log \underline{p}(\theta) = \log \underline{p}(\hat\theta) - \tfrac{1}{2}(\theta - \hat\theta)^\top \hat A (\theta - \hat\theta) + o(\|\theta - \hat\theta\|^2),$$

where the matrix $\hat A$ is minus the second derivative matrix

$$\hat A = -\frac{\partial^2}{\partial \theta^2} \log \underline{p}(\theta)|_{\theta=\hat\theta} = \operatorname{diag}(\hat c) + \Sigma^{-1}, \tag{11.38}$$

for Σ the covariance matrix of the prior, and $\operatorname{diag}(\hat c)$ the diagonal matrix with minus the second derivatives of the functions $\log t_j$ at the mode $\hat\theta$:

$$\hat{c}_j = -\frac{\partial^2}{\partial \theta_j^2} \log t_j(\theta_j)|_{\theta_j = \hat{\theta}_j}.$$

The mode $\hat{\theta}$ can be computed by Newton-Raphson: maximize a quadratic expansion (with linear term) of $\log \underline{p}$ around an initial guess $\tilde{\theta}$ to obtain a new guess, and iterate. The iterations are $\tilde{\theta}^{\text{new}} = \tilde{\theta} + \tilde{A}^{-1}\tilde{b}$, for $\tilde{b}$ the vector of first derivatives of the functions $\log t_j$ at $\tilde{\theta}$, and $\tilde{A} = \tilde{c} + \Sigma^{-1}$ the second derivative matrix at this point.

The normalizing constant of $\underline{p}$ can be approximated by, with d the dimension of θ,

$$\int \underline{p}(\theta)\, d\theta \approx \underline{p}(\hat{\theta}) \int e^{-\frac{1}{2}(\theta-\hat{\theta})^\top \hat{A}(\theta-\hat{\theta})}\, d\theta = \underline{p}(\hat{\theta})\, (\det \hat{A})^{-1/2}\, (2\pi)^{d/2}. \tag{11.39}$$

This requires only the computation of the determinant.

The Gaussian approximation to the full posterior may be marginalized to give Gaussian approximations of the marginals, but these appear to be inaccurate in many situations. A better approximation is obtained by noting that the marginal density $\theta_i \mapsto p_i(\theta_i)$ is the denominator in the definition of the conditional density $\theta_{-i} \mapsto p_{-i|i}(\theta_{-i}\,|\,\theta_i)$, i.e. the norming constant of the function $\theta_{-i} \mapsto \underline{p}(\theta)$, with θ_i fixed. We can apply the Laplace approximation to the norming constant (11.39) with $\underline{p}$ in that equation replaced by $\underline{p}$ viewed as a function of θ_{-i} (and integrated with respect to this variable), for fixed θ_i. This leads to

$$\underline{p}(\hat{\theta}_{-i}(\theta_i), \theta_i) \left(\det\Big(-\frac{\partial^2}{\partial \theta_{-i}^2} \log \underline{p}(\theta_{-i}, \theta_i) \Big)|_{\theta_{-i} = \hat{\theta}_{-i}(\theta_i)} \right)^{-1/2} (2\pi)^{d_1/2}. \tag{11.40}$$

Here $\hat{\theta}_{-i}(\theta_i)$ is the mode of $\theta_{-i} \mapsto \underline{p}(\theta)$, and we write $\underline{p}(\theta_{-i}, \theta_i)$ for $\underline{p}(\theta)$.

A different way to derive this approximation, due to Tierney and Kadane (1986), starts from the identity

$$p_i(\theta_i) = \frac{p(\theta_{-i}, \theta_i)}{p_{-i|i}(\theta_{-i}\,|\,\theta_i)}. \tag{11.41}$$

This is the definition of the conditional density $p_{-i|i}$, written "upside down," and it is an identity for *all* θ_{-i} (θ_{-i} appears on the right, but not on the left side!). We obtain an approximation of p_i by replacing the denominator $p_{-i|i}(\theta_{-i}\,|\,\theta_i)$ by an approximation, and next evaluating the resulting quotient at a suitable value θ_{-i}, for instance the mode $\hat{\theta}_{-i}(\theta_i)$ of $\theta_{-i} \mapsto p_{-i|i}(\theta_{-i}\,|\,\theta_i) \propto \underline{p}(\theta_{-i}, \theta_i)$. Because the constant cancels in the quotients (11.41) the approximation may be up to the normalizing constant. The value of the density of the multivariate distribution $\text{Nor}_d(\mu, \Lambda)$ at its mode is equal to $(\det \Lambda)^{-1/2}(2\pi)^{-d/2}$, and the Laplace approximation to $\theta_{-i} \mapsto \log p(\theta_{-i}, \theta_i)$ is a Gaussian with the second derivative matrix in the right side of (11.40) as its inverse covariance matrix. Therefore, if we use the Laplace approximation for $\theta_{-i} \mapsto p_{-i|i}(\theta_{-i}\,|\,\theta_i)$, we are lead back to (11.40).

A hyperparameter in the prior can be handled in the same way as described in Section 11.7.4.

11.8 Historical Notes

Gaussian processes appear to have first been considered as priors by Kimeldorf and Wahba (1970) and Wahba (1978), in connection to spline smoothing, as explained in Section 11.7.1.

Wood and Kohn (1998) and Shively et al. (1999) extended the idea to binary regression. Gaussian processes as priors for density estimation were first used by Leonard (1978) and later by Lenk (1988), who introduced the computational method based on importance sampling in Section 11.7.2. The computational method based on kriging presented here and its convergence properties were obtained by Tokdar (2007). The MCMC technique for (parametric) probit regression using latent Gaussian variables was introduced by Albert and Chib (1993), and modified to the nonparametric context in Choudhuri et al. (2007). Expectation propagation has its roots in mathematical phyics; we learned it from Cseke and Heskes (2011) and the thesis by Cseke. The Laplace approximation is the basis for Integrated Nested Laplace Approximation or *INLA* (see Rue et al. 2009), which is available as an R-package. An extensive review of Gaussian process priors from the point of view of machine learning is given in Rasmussen and Williams (2006). Appendix I gives an overview of Gaussian process theory, including many references. The key results mentioned in the present chapter are Borell's inequality, discovered by Borell (1975), and the bounds on small ball probabilities and shifted normal distributions due to Kuelbs and Li (1993) and Li and Linde (1998). Posterior consistency for Gaussian process priors was studied by Tokdar and Ghosh (2007), Ghosal and Roy (2006) and Choi and Schervish (2007), respectively, for density estimation, nonparametric binary regression and nonparametric normal regression models. The theory of contraction rates for Gaussian process priors was developed in the paper van der Vaart and van Zanten (2008a), and elaborated for various particular examples and rescaled processes in van der Vaart and van Zanten (2007, 2009, 2011). Lemma 11.34 is taken from Castillo (2008), who generalized a similar result by van der Vaart and van Zanten (2008a) for the special case $\alpha = \beta$. The lower bound statements in the rate theorems are also due to Castillo (2008). From many recent papers, many of which concerned with adaptation using Gaussian priors, we mention De Jonge and van Zanten (2010) on Gaussian mixtures, Panzar and van Zanten (2009) on diffusion models, van Waaij and van Zanten (2016) on adaptation, Yang and Tokdar (2015) on choice of variables in Gaussian regression models, Sniekers and van der Vaart (2015b,a) on adaptive regression and Knapik et al. (2016) on inverse problems.

Problems

11.1 (van der Vaart and van Zanten 2008a) A version of Theorem 11.22 is possible even if w_0 and $\psi/(\Psi(1-\Psi))$ are both unbounded, by using appropriate norms on the Gaussian process. Using Problem 2.6, show that for the probit link function, contraction rate holds for the sum of $\mathbb{L}_2((w_0^2 \vee 1)\cdot G)$ and $\mathbb{L}_4(G)$-norms, provided that $w_0 \in \mathbb{L}_4(G)$.

11.2 We know that the support of Brownian motion in $\mathfrak{C}[0,1]$ is the set of all functions with $w(0) = 0$. Show that the support of the standard Brownian motion as a random element in $\mathbb{L}_r[0,1]$ is the full space $\mathbb{L}_r[0,1]$ for any $r < \infty$.

11.3 (van der Vaart and van Zanten 2008a) If $w_0 \in \mathfrak{C}^\beta[0,1]$, $0 < \beta \leq 1$, and ϕ is a differentiable kernel function, then show that $(w_0 * \phi_\sigma)(0)^2 + \|(w_0 * \phi_\sigma)'\|_2^2 \lesssim \sigma^{-(2-2\beta)}$.

11.4 (van der Meulen et al. 2006) Consider a (sequence of) diffusion process defined by the stochastic differential equation

$$dX_t^{(n)} = \beta_\theta^{(n)}(t, X^{(n)})\,dt + \sigma^{(n)}(t, X^{(n)})\,dB_t^{(n)}, \quad t \in [0, T_n], \quad X_0^{(n)} = X_0, \tag{11.42}$$

where $B^{(n)}$ is a (sequence of) Brownian motion, and the functional forms of the drift coefficient $\beta_\theta^{(n)}(t, X^{(n)})$ and diffusion coefficient $\sigma^{(n)}(t, X^{(n)})$ are given continuous functions and $\theta \in \Theta$ is the unknown parameter. Let $\theta \sim \Pi$ and $\theta_0 \in \Theta$ stand for the true value of the parameter. Let $h_n^2(\theta, \theta_0) = \int_0^{T_n} (\beta_\theta^{(n)}(t, X^{(n)}) - \beta_{\theta_0}^{(n)}(t, X^{(n)})^2 (\sigma^{(n)}(t, X^{(n)}))^{-2} dt$. (This is equal to the squared Hellinger distance between the Gaussian process distributions of $(X_t^{(n)} : t \in T_n)$ under θ and θ_0, respectively.) Let $B_n(\theta_0, \epsilon) = \{\theta : h_n(\theta, \theta_0) < \epsilon\}$. Suppose that for some sequence ϵ_n, $\log N(a\epsilon, B_n(\theta_0, \epsilon), h_n) \lesssim \epsilon_n^2$ for all $\epsilon > \epsilon_n$ and that for any $\xi > 0$, there is $J \in \mathbb{N}$ such that $\Pi(B_n(\theta_0, j\epsilon_n))/\Pi(B_n(\theta_0, \epsilon_n)) \le e^{\xi j^2 \epsilon_n^2}$ for all $j \ge J$. Show that the posterior contracts at θ_0 at the rate ϵ_n with respect to h_n.

11.5 (van der Meulen et al. 2006) Consider a special case of Problem 11.4 given by signal plus white noise model $dX_t^{(n)} = \theta(t)\, dt + \sigma_n dB_t$, $t \in [0, T]$, $X_0^{(n)} = x_0$, $\sigma_n \to 0$. Assume that for some sequence ϵ_n, $\log N(\epsilon/8, \{\theta : \|\theta - \theta_0\|_2 < \epsilon\}, \|\cdot\|_2) \lesssim \sigma_n^{-2}\epsilon_n^2$ for all $\epsilon \ge \epsilon_n$ and for some $J \in \mathbb{N}$, $\Pi(\|\theta - \theta_0\|_2 < j\epsilon_n))/\Pi(\|\theta - \theta_0|_2 < \epsilon_n) \le e^{j^2\epsilon_n^2\sigma_n^{-2}/9216}$ for all $j \ge J$. Show that the posterior contracts at θ_0 at the rate ϵ_n with respect to $\|\cdot\|_2$.

11.6 (van der Meulen et al. 2006) Consider a special case of Problem 11.4 given by the perturbed dynamical system $dX_t^{(n)} = \theta(X_t^{(n)})\, dt + \sigma_n dB_t^{(n)}$, $t \in [0, T]$, $X_0^{(n)} = X_0$, $\sigma_n \to 0$. Let $d^2(\theta, \theta_0) = \int |\theta(x_t) - \theta_0(x_t)|^2 dt$, where x_t is the unique solution of the ordinary differential equation $dx_t = \theta_0(x_t)$, $x_t = x_0$ for $t = 0$. Suppose that all θ are uniformly bounded and uniformly Lipschitz continuous. Assume that for some sequence ϵ_n, $\log N(\epsilon/24, \{\theta : d(\theta, \theta_0) < \epsilon\}, d) \lesssim \sigma_n^{-2}\epsilon_n^2$ for all $\epsilon \ge \epsilon_n$, and for some $J \in \mathbb{N}$, $\Pi(d(\theta, \theta_0) < j\epsilon_n))/\Pi(d(\theta, \theta_0) < \epsilon_n) \le e^{j^2\epsilon_n^2\sigma_n^{-2}/20736}$ for all $j \ge J$. Show that the posterior contracts at θ_0 at the rate ϵ_n with respect to d.

If θ lies in a bounded subset of the Besov space $\mathfrak{B}_{p,\infty}^\alpha$ of α-smooth functions, $p\alpha > 1$, construct a prior based on an $\sigma_n^{2\alpha/(2\alpha+1)}$-net in terms of the uniform distance as in Subsection 8.2.2. Show that the posterior contracts at the rate $\sigma_n^{2\alpha/(2\alpha+1)}$ with respect to d.

If θ lies in a bounded subset of the Besov space $\mathfrak{B}_{\infty,\infty}^\alpha$ of α-smooth functions, construct a prior based on a wavelet expansion $\theta = \sum_{j=1}^J \sum_{k=1}^{2^j} 2^{-j/2} Z_{j,k} \psi_{j,k}$, where $\psi_{j,k}$ are wavelet functions and $Z_{j,k} \overset{\text{iid}}{\sim} \text{Nor}(0, 1)$. Show that the posterior contracts at the rate $\sigma_n^{2\alpha/(2\alpha+1)}$ with respect to d.

11.7 (van der Meulen et al. 2006) Consider a special case of Problem 11.4 given by the ergodic diffusion model $dX_t = \theta(X_t)\, dt + \sigma(X_t) dB_t$, $t \in [0, T_n]$, $T_n \to \infty$. Let m_{θ_0} be the speed measure defined by having a density $\sigma^{-2}(x) \exp\{2 \int_{x_0}^x \theta_0(z)\sigma^{-2}(z)\, dz\}$, where x_0 is fixed but arbitrary. Let I be a compact subinterval of $\mathbb{R}$ and $\|f\|_{2,\mu_0,I}^2 = \int_I |f(x)|^2 d\mu_0(x)$ and $\|\cdot\|_{2,\mu_0} = \|\cdot\|_{2,\mu_0,\mathbb{R}}$. Assume that $m_{\theta_0}(I) < \infty$ and $\mu_0 = m_0/m_{\theta_0}(I)$. Suppose that for some sequence $\epsilon_n > 0$, $T_n\epsilon_n^2 \gg 0$, we have $\{\log N(a\epsilon, \{\theta : \|(\theta - \theta_0)/\sigma\|_{2,I} < \epsilon\}, \|\cdot\|_{2,\mu_0}) : \epsilon \ge \epsilon_n\} \lesssim T_n\epsilon_n^2$ for every $a > 0$ and for all $\xi > 0$, there exists some $J \in \mathbb{N}$, $\Pi(\|\theta - \theta_0\|_{2,\mu_0,I} < j\epsilon_n)/\Pi(\|\theta - \theta_0\|_{2,\mu_0} < \epsilon_n) \le e^{\xi j^2 T_n \epsilon_n^2}$ for all $j \ge J$. Show that the posterior contracts at θ_0 at the rate ϵ_n with respect to $\|\cdot\|_{2,\mu_0}$.

In the above setting, let there be a k-dimensional parameter θ defining the process by $dX_t = \beta_\theta(X_t)\,dt + \sigma(X_t)dB_t$, where $\underline{\beta}(x)\|\theta - \theta^*\| \le |\beta_\theta(x) - \beta_{\theta^*}(x)| \le \bar{\beta}(x)$ for some functions $\underline{\beta}$ and $\bar{\beta}$ satisfying $0 < \int(\underline{\beta}/\sigma)^2 d\mu_0(x) \le \int(\bar{\beta}/\sigma)^2 d\mu_0(x) < \infty$. Let θ have a prior density bounded and bounded away from zero, and let θ_0 be the true value of θ. Show that the posterior for θ contracts at θ_0 at the rate $T_n^{-1/2}$ with respect to the Euclidean distance.

11.8 The parameters μ_1 and Σ_1 in Lemma 11.65 can be computed from the factorization $\phi_{\mu_1,\Sigma_1}(z) = \psi_1(Uz)\phi_1(z|\,Uz)$, where ψ_1 is the normal density with mean vector $\bar{\mu}_1 = \mathrm{E}Yt(Y)/\mathrm{E}t(Y)$ and covariance matrix $\mathrm{E}(Y-\bar{\mu}_1)(Y-\bar{\mu}_1)^\top t(Y)/\mathrm{E}t(Y)$, for $Y \sim \mathrm{Nor}(U\mu, U\Sigma U^\top)$, and $\phi_1 = \phi_0$ is the conditional density of Z given $Y = UZ$. The latter is a normal distribution with mean $\mathrm{E}(Z|\,Y) = AU(Z-\mu)+\mu$, for $A = \Sigma U^\top(U\Sigma U^\top)^{-1}$ and covariance matrix $\mathrm{Cov}\,(Z - \mathrm{E}(Z|\,Y)) = \Sigma - AU\Sigma U^\top A^\top$. Alternatively these parameters can be computed directly using

$$\mu_1 = \frac{\mathrm{E}Zt(UZ)}{\mathrm{E}t(UZ)} = \frac{\mathrm{E}\mathrm{E}(Z|\,UZ)t(UZ)}{\mathrm{E}t(UZ)},$$

$$\Sigma_1 = \frac{\mathrm{E}(Z-\mu_1)(Z-\mu_1)^\top t(UZ)}{\mathrm{E}t(UZ)} = \frac{\mathrm{E}\mathrm{E}((Z-\mu_1)(Z-\mu_1)^\top|\,UZ)t(UZ)}{\mathrm{E}t(UZ)}$$

$$= \mathrm{Cov}(Z|\,UZ) + \frac{\mathrm{E}(\mathrm{E}(Z|\,UZ)-\mu_1)(\mathrm{E}(Z|\,UZ)-\mu_1)^\top t(UZ)}{\mathrm{E}t(UZ)}.$$

The last term on the right is the covariance matrix of the vector $AU(Z^*-\mu)+\mu$, with Z^* given the tilted normal distribution with density proportional to $z \mapsto t(Uz)\phi_{\mu,\Sigma}(z)$.