Probab. Theory Relat. Fields (2014) 158:665–710
DOI 10.1007/s00440-013-0493-0

Thomas Bayes' walk on manifolds

Ismaël Castillo · Gérard Kerkyacharian · Dominique Picard

Received: 15 May 2012 / Revised: 15 December 2012 / Published online: 2 March 2013

Abstract Convergence of the Bayes posterior measure is considered in canonical statistical settings where observations sit on a geometrical object such as a compact manifold, or more generally on a compact metric space verifying some conditions. A natural geometric prior based on randomly rescaled solutions of the heat equation is considered. Upper and lower bound posterior contraction rates are derived.

Keywords Bayesian nonparametrics · Gaussian process priors · Heat kernel

Mathematics Subject Classification (2000) 62G05 · 62G20

1 Introduction

Let $\mathcal{M}$ be a compact metric space, equipped with a Borel measure μ and the corresponding Borel-sigma field. Let $\mathbb{L}^p := \mathbb{L}^p(\mathcal{M}, \mu)$, $p \geq 1$ and $\mathcal{C}^0(\mathcal{M})$ respectively denote the real vector spaces of real-valued functions defined on $\mathcal{M}$ that are

I. Castillo's work is partly supported by ANR Grant 'Banhdits' ANR-2010-BLAN-0113-03.
G. Kerkyacharian and D. Picard's work is partly supported by ANR Grant 'Parcimonie' ANR-2009-BLAN-0128-01.

I. Castillo (✉)
CNRS-LPMA Universities Paris 6 and 7, Paris, France
e-mail: ismael.castillo@upmc.fr

G. Kerkyacharian · D. Picard
Université Paris Diderot-Paris 7, LPMA, Paris, France
e-mail: kerk@math.jussieu.fr

D. Picard
e-mail: picard@math.jussieu.fr

p-integrable with respect to μ and of real-valued continuous functions on $\mathcal{M}$. Also, denote by $\mathcal{D}(\mathbb{R})$ the algebra of real-valued infinitely differentiable functions on the real line.

In this paper we investigate rates of contraction of posterior distributions for nonparametric models on geometrical structures such as

1. Gaussian white noise on a compact metric space $\mathcal{M}$, where, for $n \geq 1$, one observes

$$dX^{(n)}(x) = f(x)dx + \frac{1}{\sqrt{n}}dZ(x), \quad x \in \mathcal{M},$$

   where f is in $\mathbb{L}^2$ and Z is a white noise on $\mathcal{M}$.
2. Fixed design regression where one observes, for $n \geq 1$,

$$Y_i = f(x_i) + \varepsilon_i, \quad 1 \leq i \leq n.$$

   The design points $\{x_i\}$ are fixed on $\mathcal{M}$ and the variables $\{\varepsilon_i\}$ are assumed to be independent standard normal.
3. Density estimation on a manifold where the observations are a sample

$$(X_i)_{1 \leq i \leq n} \sim f,$$

   $X_1, \ldots, X_n$ are independent identically distributed $\mathcal{M}$-valued random variables with positive density function f on $\mathcal{M}$.

Although an impressive amount of work has been done using frequentist approachs to estimation on manifolds, see [20] and the references therein, we focus in this paper on the Bayes posterior measure.

Works devoted to deeply understanding the behaviour of Bayesian nonparametric methods have recently experienced a considerable development in particular after the seminal works of A. W. van der Vaart, H. van Zanten, S. Ghosal and J. K. Ghosh [12], [27]. Especially, the class of Gaussian processes forms an important family of nonparametric prior distributions, for which precise rates have been obtained in [31], see also [5] for lower bound counterparts. In [33], the authors obtained adaptive performance up to logarithmic terms by introducing a random rescaling of a very smooth Gaussian random field. In these results, the considered rescaling corresponds to shrinking the paths of the process. These results have been obtained on $[0, 1]^d$, $d \geq 1$. Our point in this paper is to develop a Bayesian procedure adapted to the geometrical structure of the data. Among the examples covered by our results, we can cite directional data corresponding to the spherical case and more generally data supported by a compact manifold.

We follow the illuminating approach of [31] and [33] and use a fixed prior distribution, constructed by rescaling a smooth Gaussian random field. For a recent survey on Gaussian processes and their basic properties, see [19]. Basically our aim will be twofold:

First, because the 'shrinking of paths' approach from [33] has no natural analogue on a general manifold, this type of rescaling cannot be used. In our more general

setting, we show how a rescaling is made possible by introducing a notion of time decoupled from the underlying space and issued from the semigroup property of a family of operators. Another important difference brought by the geometrical nature of the problem is the underlying Gaussian process, which now originates from an harmonic analysis of the data space $\mathcal{M}$, with the rescaling naturally acting on the frequency domain. More precisely, we suppose that $\mathcal{M}$ is equipped with a positive self-adjoint operator L such that the associated semigroup e^{-tL}, $t > 0$, the *heat kernel*, allows a smooth functional calculus, which in turn allows the construction of the Gaussian random field. A central example of operator is $L = -\Delta$, where Δ is the Laplacian on $\mathcal{M}$. Our prior can then be interpreted as a randomly rescaled (random) solution of the heat equation.

This construction enables to obtain rates of contraction of the posterior distribution depending on the 'regularity' of the estimated function (defined in terms of approximation rates) and a 'dimension' of the geometrical object at hand.

Secondly, we prove a lower bound for the posterior contraction rate showing in particular that the logarithmic factor appearing in the upper evaluations of the posterior rate is necessary.

We also took inspiration on earlier work by [1], where the authors consider a symmetry-adaptive Bayesian estimator in a regression framework. Precise minimax rates in the $\mathbb{L}^2$-norm over Sobolev spaces of functions on compact connected orientable manifolds without boundary are obtained in [11]. We also mention a recent development by [2], where Bayesian consistency properties are derived for priors based on mixture of kernels over a compact manifold.

The paper is mostly self-contained and does not require prior knowledge of heat kernel theory. Definitions and notation for the heat kernel can be found in Sect. 2.4. To obtain sharp entropy bounds for some compact sets appearing in the proof, we use the existence of needlet-type basis on $\mathcal{M}$, as established in [9]. Standard general conditions to obtain rates for Bayesian posteriors are used following [12,13].

Here is an outline of the paper. We first detail in Sect. 2 the properties assumed on the structure $\mathcal{M}$ and the associated heat kernel allowing our construction. We then construct the associated Gaussian prior defining the procedure. Examples are considered in Sect. 3. The main results are stated in Sect. 4, that is: rates of contraction for the procedure, as well as a lower bound proving that the logarithmic factor present in the rates is, in fact, sharp. The rest of the paper is devoted to the proofs of these results. In Sects. 5 and 6 structural properties of the considered Gaussian processes are studied, and entropy estimates are stated. These properties enable one to check general sufficient conditions for upper-bound posterior rates, as we demonstrate in Sect. 7. Upper-bound rates are then derived in Sect. 8, as well as the corresponding lower bound result. Sections 9 and 10 contain respectively the definition of Besov spaces and the proofs of the sharp entropy results. Finally, in Sect. 10.5, a homogeneity property of measure of balls needed for our results is verified for compact Riemannian manifolds without boundary.

The notation $\lesssim$ means less than or equal to up to some universal constant. For any sequences of reals $(a_n)_{n\geq 0}$ and $(b_n)_{n\geq 0}$, the notation $a_n \sim b_n$ means that the sequences verify $c \leq \liminf_n (b_n/a_n) \leq \limsup_n (b_n/a_n) \leq d$ for some positive constants c, d, and $a_n \ll b_n$ stands for $\lim_n (b_n/a_n) = 0$. For any reals a, b, we denote

$\min(a, b) = a \wedge b$ and $\max(a, b) = a \vee b$. For a given differentiable function u, we denote by u' its derivative.

2 The geometrical framework and our method

2.1 Compact metric doubling space $\mathcal{M}$

Let ρ denote the metric on the space $\mathcal{M}$. The open ball of radius r centered at $x \in \mathcal{M}$ is denoted by $B(x, r)$ and to simplify the notation we put $\mu(B(x, r)) =: |B(x, r)|$. Without loss of generality, we impose in the abstract proofs that both the total mass $\mu(\mathcal{M})$ and the diameter of $\mathcal{M}$ are equal to 1. Although this is typically not the case in practical situations, see Sect. 3, considering the general case only changes constants in the proofs.

We assume that $\mathcal{M}$ has the so called *doubling property*: i.e. there exists $0 < D < \infty$ such that:

$$\text{for all } \ x \in \mathcal{M}, \ 0 < r, \quad 0 < |B(x, 2r)| \leq 2^D |B(x, r)| \tag{1}$$

We say that $\mathcal{M}$ verifies the *Ahlfors property* (see e.g. [16]) if there exist positive c_1, c_2, d such that

$$\text{for all } x \in \mathcal{M}, \ \text{for all } \ 0 < r \leq 1, \quad c_1 r^d \leq |B(x, r)| \leq c_2 r^d. \tag{2}$$

If (2) holds, then one must have $d \leq D$. Indeed, successive applications of (1) imply $|B(x, r)| \geq (r/2)^D$. In the rates obtained in the sequel as well as in the specific examples we consider, d plays the role of a dimension. Notice, however, that there is no need for d to be an integer.

2.2 Previous work on the real line

To motivate our approach, let us start with the simple case where the space $\mathcal{M}$ is a compact interval on the real line, say $\mathcal{M} = [0, 1]$. This is the case considered in [33]. The statistical goal, for anyone of the models considered in the introduction (white noise, regression and density), is to estimate the unknown function f. To do so, a Bayesian approach first has to put a prior distribution on f, that is a probability distribution on, say, continuous functions f on $\mathcal{M}$.

So, how does one build a prior on f? A possibility to model a random function on [0, 1] is to take realisations of a stochastic process on this interval. A natural class which comes to mind is the one of Gaussian processes. Any such process $(Z_t)_{t\in[0,1]}$ is characterised by a mean function, which here will be taken to be identically zero, and a covariance kernel $K(s, t) = \mathbb{E}(Z_s Z_t)$, for s, t in [0, 1]. In this case, choosing a prior reduces to choosing a covariance kernel K.

In [33], the authors make use of the so-called *squared-exponential* covariance kernel

$$K(s, t) = e^{-(s-t)^2}, \quad (s, t) \in (0, 1)^2. \tag{3}$$

It can be shown that the centered Gaussian process (Z_t) with such covariance has very smooth paths. To achieve minimax adaptation properties for the Bayesian posterior, the approach taken by [33] is to additionally allow for some extra freedom in the rescaling by considering (Z_{At}) with A random with properly chosen distribution.

There is a simple reason why the squared-exponential kernel cannot be used in a geometric context for general $\mathcal{M}$. Although (3) admits the immediate generalisation (recall that ρ is the metric on $\mathcal{M}$)

$$\kappa_\rho(s,t) = e^{-\rho(s,t)^2}, \quad (s,t) \in \mathcal{M}^2, \tag{4}$$

it can be shown that this function is *not* positive definite in general already for the simplest examples such as $\mathcal{M}$ taken to be the sphere in $\mathbb{R}^k$, $k \geq 2$.

2.3 Building a positive-definite kernel on $\mathcal{M}$ via an operator L

Following the previous idea of building a Gaussian process on $\mathcal{M}$, the question is now how to build an appropriate covariance kernel on $\mathcal{M}$.

Suppose one is given a decomposition of the space of square-integrable functions on $\mathcal{M}$

$$\mathbb{L}^2(\mathcal{M}) = \oplus_{k\geq 0}\mathcal{H}_k, \tag{5}$$

where the $\mathcal{H}_k$ are finite-dimensional subspaces of $\mathbb{L}^2$ consisting of continuous functions on $\mathcal{M}$, and orthogonal in $\mathbb{L}^2$. Then, the projector Q_k on $\mathcal{H}_k$ is actually a kernel operator $Q_k(x,y) := \sum_{1\leq i\leq dim(\mathcal{H}_k)} e_k^i(x)e_k^i(y)$, where $\{e_k^i\}$ is any orthonormal basis of $\mathcal{H}_k$; so it is obviously a positive-definite kernel. Also, given $\varphi : \mathbb{N} \to (0,+\infty)$ such that $\forall x \in M$, $\sum_{k\geq 0}\varphi(k)Q_k(x,x) < \infty$, the function $K_\varphi(x,y) = \sum_{k\geq 0}\varphi(k)Q_k(x,y)$ is a positive definite kernel which is the covariance kernel of a Gaussian process. Constructing explicitely a Gaussian process with this covariance is not difficult, this will be done in Sect. 2.6.

A simple way of obtaining a decomposition (5) is by diagonalisation of a self-adjoint positive operator L on $\mathcal{M}$ with discrete spectrum, finite dimension spectral spaces and eigenfunctions continuous on $\mathcal{M}$. In this case the subspaces $\mathcal{H}_k =: \mathcal{H}_{\lambda_k}$ can be taken to be the eigenspaces of L. Such an operator has non-negative eigenvalues λ_k that we order in an increasing way ($0 \leq \lambda_0 < \lambda_1 < \cdots$).

While many such operators L could in principle be used, we will be especially interested in the cases where L reflects quite well some geometrical properties of $\mathcal{M}$. A central example when $\mathcal{M}$ is a compact Riemannian manifold without boundary is $L = -\Delta$, where Δ is the *Laplacian* (or weighted Laplacian) on $\mathcal{M}$, see Sect. 3 for details.

The operator L being given, we still need to choose the function φ. For this purpose, we will concentrate on another important aspect of the approach developed in [31] and [33]: the rescaling At. In these papers, the rescaling drives the regularity and its proper choice is essential for the properties of the estimators. In the case of a general set $\mathcal{M}$, a multiplicative rescaling At has typically no sense. To find a meaningful

generalisation of the rescaling, we will use a standard tool of the theory of operators (see [10]), the semi-group associated to the operator L, which ultimately yields the choice $\varphi(k) = e^{-t\lambda_k}$. Note that a slightly similar point of view has been considered recently in statistical learning with Laplacian-based spectral methods (diffusion maps, diffusion wavelets...) propagating information on the data through a Markov kernel (see for instance [8,21]).

2.4 Heat kernel

Let us now give the properties on the operator L that we will need. These conditions arise naturally in the theory of heat kernels. Though no prerequisites on heat kernels are needed for the present paper, we refer the interested reader to [14,22,26] for standard expositions on heat kernel theory.

We suppose that the self-adjoint positive operator L is defined on a domain $D \subset \mathbb{L}^2$ dense in $\mathbb{L}^2$. Then $-L$ is the infinitesimal generator of a contraction self-adjoint semigroup e^{-tL}, see [10, Thm. 4.6].

We suppose in addition that e^{-tL} is a Markov kernel operator i.e. there exists a non negative continuous kernel $P_t(x, y)$ (*the 'heat kernel'*) such that:

$$e^{-tL} f(x) = \int_{\mathcal{M}} P_t(x, y) f(y) d\mu(y) \tag{6}$$

$$P_t(x, y) = P_t(y, x), \tag{7}$$

$$\int_{\mathcal{M}} P_t(x, y) d\mu(y) = 1, \tag{8}$$

$$\forall x, y \in M, \ \forall s, t > 0, \quad P_{t+s}(x, y) = \int_{\mathcal{M}} P_t(x, u) P_s(u, y) du. \tag{9}$$

Clearly, from (7), (9) and $P_t(x, y) \geq 0$, we have

$$P_t(x, y) = \int_{\mathcal{M}} P_{t/2}(x, u) P_{t/2}(y, u) du = \int_{\mathcal{M}} P_{t/2}(x, u) \overline{P_{t/2}(y, u)} du,$$

which immediately implies that $P_t(x, y)$ is a positive-definite kernel.

We will further assume the following bounds on the heat kernel, which are satisfied in a large variety of situations, in particular all the examples considered in Sect. 3 (see [14,22,26]):

Suppose that there exist $C_1, C_2 > 0, c_1, c_2 > 0$, such that, for all $t \in]0, 1[$, and any $x, y \in \mathcal{M}$,

$$\frac{C_2}{\sqrt{|B(x, \sqrt{t})||B(y, \sqrt{t})|}} e^{-\frac{c_2 \rho^2(x,y)}{t}} \leq P_t(x, y) \leq \frac{C_1}{\sqrt{|B(x, \sqrt{t})||B(y, \sqrt{t})|}} e^{-\frac{c_1 \rho^2(x,y)}{t}}. \tag{10}$$

It is also known ([9], and references therein), that $P_t(x, y)$ is a continuous function, the eigenspaces $\mathcal{H}_{\lambda_k}$ of L are of finite dimension and the eigenvectors are continuous. That is, $\mathbb{L}^2(\mathcal{M}) = \oplus_{k\geq 0}\mathcal{H}_{\lambda_k}$, and the orthogonal projectors $P_{\mathcal{H}_{\lambda_k}}$ on the eigenspaces $\mathcal{H}_{\lambda_k}$ are kernel operators $Q_k(x, y)$ with

$$Q_k(x, y) = \sum_{1\leq l\leq dim(\mathcal{H}_{\lambda_k})} e_k^l(x)e_k^l(y),$$

as soon as $\{e_k^l,\ 1 \leq l \leq dim(\mathcal{H}_{\lambda_k})\}$ is an orthonormal basis of $\mathcal{H}_{\lambda_k}$.

The Markov kernel P_t writes

$$P_t(x, y) = \sum_k e^{-t\lambda_k} Q_k(x, y). \tag{11}$$

A direct consequence of (10) is that for all $x \in \mathcal{M}$ and all $t \in]0, 1[$,

$$\frac{C_2}{|B(x, \sqrt{t})|} \leq P_t(x, x) \leq \frac{C_1}{|B(x, \sqrt{t})|}. \tag{12}$$

2.5 Why is the heat kernel a canonical kernel on $\mathcal{M}$?

Consider the already large class of compact Riemannian manifolds without boundary for $\mathcal{M}$. In that case we set $L = -\Delta$, where Δ is the Laplacian on $\mathcal{M}$, see Sect. 3 for details. We claim that in this case the associated heat kernel is a natural choice for the following reasons

1. The heat kernel P_t is a *positive definite* kernel on $\mathcal{M}$. Thus P_t is at least a possible candidate for use as a covariance kernel of a Gaussian process.
2. The heat kernel associated to $L = -\Delta$ on $\mathcal{M}$ verifies (10), see Sect. 3. In particular, up to constants, the heat kernel appears as a natural geometric generalisation of the squared-exponential kernel (3) on the real line ! Also, we see that the 'time' t is a natural candidate for a (inverse-) scale parameter. We will indeed allow t to vary in the definition of our prior below.
3. In the context of harmonic analysis on geometric spaces $\mathcal{M}$, the Laplacian on $\mathcal{M}$, or equivalently the associated heat kernel semi-group, are known as natural carriers of the information about the 'geometry' of $\mathcal{M}$. The collection of eigenspaces $\mathcal{H}_{\lambda_k}$ defined above can very much be interpreted as an harmonic analysis of $\mathcal{M}$. For instance, in the case of the circle $\mathbb{S}^1 \sim \mathbb{R}/2\pi\mathbb{Z}$, one has $\lambda_k = k^2$ and $\mathcal{H}_{\lambda_k} = \mathcal{H}_k$ is generated by $x \to \cos(kx)$ and $x \to \sin(kx)$, see Sect. 3.

2.6 Our method: Prior, definition

For the statistical results of the paper, we always assume that both conditions (1) and (2) on the set $\mathcal{M}$ hold. In particular, the prior depends on the 'dimension parameter' d

in (2). However, the key entropy property stated in Sect. 5 holds more generally under (1) only.

We consider a prior on functions from $\mathcal{M}$ to $\mathbb{R}$ constructed hierarchically as follows.

First, generate a collection of independent standard normal variables $\{X_k^l\}$ with indexes $k \geq 0$ and $1 \leq l \leq dim(\mathcal{H}_{\lambda_k})$. Set, for $x \in \mathcal{M}$ and any $t \in (0, 1]$,

$$W^t(x) = \sum_{k \geq 0} \sum_{1 \leq l \leq \dim \mathcal{H}_{\lambda_k}} e^{-\lambda_k t/2} X_k^l e_k^l(x). \tag{13}$$

To simplify the notation, and when no confusion is possible, we omit in the sequel the range for indexes k, l in summations. Equation (13) defines a Gaussian stochastic process $W^t(\cdot)$ indexed by $\mathcal{M}$. This process is centered and has covariance kernel precisely P_t, as follows from computing

$$\mathbb{E}(W^t(x)W^t(y)) = \sum_k e^{-\lambda_k t} \sum_l e_k^l(x) e_k^l(y) = P_t(x, y).$$

Also, W^t defines a Gaussian variable in various separable Banach spaces $\mathbb{B}$, see [32] for definitions. In particular, it is a Gaussian random element in both $\mathbb{B} = (\mathcal{C}^0(\mathcal{M}), \|\cdot\|_\infty)$ and $\mathbb{B} = (\mathbb{L}^2(\mathcal{M}, \mu), \|\cdot\|_2)$, the two Banach spaces we consider in the sequel. To check this, apply Theorem 4.2 in [32], where almost sure convergence of the series (13) in $\mathbb{B}$ follows from the properties of the Markov kernel (11).

Second, draw a positive random variable T according to a density g on $(0, 1]$. This variable can be interpreted as a random scaling, or 'time'. It turns out that convenient choices of g are deeply connected to the geometry of $\mathcal{M}$. We choose the density g of T such that, for a real $a > 1$ and positive constants C_1, C_2, q, with d defined in (2),

$$C_1 t^{-a} e^{-t^{-d/2} \log^q(1/t)} \leq g(t) \leq C_2 t^{-a} e^{-t^{-d/2} \log^q(1/t)}, \quad t \in (0, 1]. \tag{14}$$

We show below that the choice $q = 1 + d/2$ leads to sharp rates.

The full (non-Gaussian) prior we consider is W^T, where T is random with density g. Hence, this construction leads to a prior Π_w, which is the probability measure induced by

$$W^T(x) = \sum_k \sum_l e^{-\lambda_k T/2} X_k^l e_k^l(x). \tag{15}$$

Some comments are in order. First, one could be tempted to use the Gaussian prior W^t (with t fixed) as prior. Nevertheless, similar to what happens for the squared exponential prior on the real line [33], the paths of the corresponding Gaussian process are infinitely differentiable almost surely, which would lead to slow rates of convergence for the posterior, see [30]. This difficulty can be overcome by making t random, allowing the prior to adapt to regularity of the unknown function. The choice of the particular form for the prior on T is related to the form taken by the entropy of the Reproducing Kernel Hilbert Space (RKHS) of W^t, as will be seen in Sect. 5. For more discussion on this, see also Sect. 4.3.

3 Examples

This section is devoted to the presentation of a variety of examples which naturally fit into the framework introduced above. The three first examples reflect a situation with no boundary where Condition (2) is verified.

The last case gives an illustration of a more complicated situation, where a boundary is present and (2) may or may not be valid, depending on the type of measure μ and operator L.

It is interesting to observe in these examples that, in the situations where (2) is true, the constant d has a natural interpretation as the dimension of the problem.

Torus case Let $\mathcal{M} = \mathbb{S}^1$ be the torus, parameterised by $[-\pi, \pi]$, with identification of π and $-\pi$, and equipped with the normalised Lebesgue measure. The metric ρ reflects the previous identification.

$$\rho(x, y) = |x - y| \wedge (2\pi - |x - y|), \quad x, y \in [-\pi, \pi].$$

In particular, for any $0 < r \leq \pi$ one has $|B(x, r)| = r/\pi$, which ensures condition (2) with $d = 1$. The spectral decomposition of the Laplacian operator $\Delta = -L$ gives rise to the classical Fourier basis, with $\lambda_k = k^2$ and

$$\mathcal{H}_0 = span\{1\}; \quad \mathcal{H}_{\lambda_k} = span\{e^{ikx}, e^{-ikx}\} = span\{\sin kx, \cos kx\}.$$

From this one deduces that $Q_k(x, y) = 2\cos k(x - y)$ and

$$e^{-tL}(x, y) = e^{t\Delta}(x, y) = 1 + \sum_{k\geq 1} e^{-k^2 t} 2\cos k(x - y) = \sqrt{\frac{\pi}{t}} \sum_{l\in\mathbb{Z}} e^{-\frac{(x-y-2l\pi)^2}{4t}}.$$

Clearly, for all $t > 0$, $e^{t\Delta}(x, x) \geq 1$, and

$$\text{for all } 0 < t < 1, \quad x, y \in [-\pi, \pi], \quad C' \frac{1}{\sqrt{t}} e^{-c' \frac{\rho(x,y)^2}{t}} \leq e^{t\Delta}(x, y) \leq C \frac{1}{\sqrt{t}} e^{-c \frac{\rho(x,y)^2}{t}}.$$

Sphere case Let now $\mathcal{M} = \mathbb{S}^{d-1} \subset \mathbb{R}^d$. The geodesic distance on $\mathbb{S}^{d-1}$ is given by

$$\rho(x, y) = \cos^{-1}(\langle x, y\rangle), \qquad \langle x, y\rangle = \sum_{i=1}^{d} x_i y_i.$$

We take as μ the natural measure on $\mathbb{S}^{d-1}$ which is *rotation invariant*. It follows that

$$|B(x, r)| = |\mathbb{S}^{d-2}| \int_0^r (\sin t)^{d-2} dt,$$

From this one deduces the following inequalities ensuring (2) with 'dimension' $d-1$,

$$\text{for all } x \in \mathcal{M}, \text{ for all } 0 \leq r \leq \pi = diam(\mathbb{S}^{d-1}), \quad c_1 r^{d-1} \leq |B(x,r)| \leq c_2 r^{d-1}.$$

As $\mathcal{M} = \mathbb{S}^{d-1}$ is a Riemannian manifold, there is a natural Laplacian on $\mathbb{S}^{d-1}$, $\Delta_{\mathbb{S}^{d-1}} = -L$, which is a negative self-adjoint operator, whose spectral decomposition we describe next. The eigenspaces $\mathcal{H}_{\lambda_k}$ turn out to be the restriction to $\mathbb{S}^{d-1}$ of polynomials of degree k which are *homogeneous* (i.e. $P(x) = \sum_{|\alpha|=k} a_\alpha x^\alpha$, $\alpha = (\alpha_1, \ldots, \alpha_n)$, $|\alpha| = \sum \alpha_i$, $\alpha_i \in \mathbb{N}$) and *harmonic on* $\mathbb{R}^d$ (i.e. $\Delta P = \sum_{i=1}^d \frac{\partial^2 P}{\partial x_i^2} = 0$.) We have,

$$P \in \mathcal{H}_{\lambda_k} \iff LP = -\Delta_{\mathbb{S}^{d-1}} P = k(k+d-2)P := \lambda_k P$$

Moreover $dim(\mathcal{H}_{\lambda_k}) = \frac{2k+d-2}{d-2}\binom{d+k-3}{k} := N_k(d)$. This space is called the space of spherical harmonics of order k. Moreover, if $(Y_{ki})_{1 \leq i \leq N_k}$ is any orthonormal basis of $\mathcal{H}_{\lambda_k}$, the projector writes

$$Q_k(x,y) = \sum_{1 \leq i \leq N_k} Y_{ki}(x) Y_{ki}(y).$$

In fact, Q_k further has an explicit expression in terms of Gegenbauer polynomials, see [28]. Now for all $0 < t \leq 1$, it can be proved (as a consequence of a general result on compact Riemannian manifolds, see below) that

$$\frac{C'}{t^{(d-1)/2}} e^{-c' \frac{\rho(x,y)^2}{t}} \leq e^{-tL}(x,y) = \sum_k e^{-tk(k+d-2)} Q_k(x,y) \leq \frac{C}{t^{(d-1)/2}} e^{-c \frac{\rho(x,y)^2}{t}}.$$

Compact Riemannian manifold, without boundary This case obviously generalises the two examples above. Let $\mathcal{M}$ be a compact Riemannian manifold of dimension d without boundary. Associated to the Riemannian structure we have a measure dx, a metric ρ, and a Laplacian Δ, such that

$$\int_{\mathcal{M}} \Delta f(x) g(x) dx = -\int \nabla(f)(x) \cdot \nabla(g)(x) dx.$$

So $L = -\Delta$ is a symmetric nonnegative operator. Now the associated semigroup $e^{t\Delta}$ is a positive kernel operator verifying: for all $t \in]0, 1[$, (see [7])

$$C' \frac{1}{\sqrt{|B(x,\sqrt{t})||B(y,\sqrt{t})|}} e^{-c' \frac{\rho^2(x,y)}{t}} \leq e^{t\Delta}(x,y) \leq C \frac{1}{\sqrt{|B(x,\sqrt{t})||B(y,\sqrt{t})|}} e^{-c \frac{\rho^2(x,y)}{t}}.$$

The main property proved in Sect. 10.5 is that there exist $0 < c_1 < c_2 < \infty$, such that

$$\text{for all } x \in \mathcal{M}, \quad \text{for all } r \leq diam(\mathcal{M}), \quad c_1 r^d \leq |B(x,r)| \leq c_2 r^d,$$

which is exactly (2) with dimension d.

Jacobi case Consider $\mathcal{M} = [-1, 1]$, equipped with the measure $\omega(x)dx$ with $\omega(x) = (1-x)^\alpha (1+x)^\beta$ and $\alpha > -1,\ \beta > -1$. (So we have, in fact, a family of measures.) Consider the metric

$$\rho(x, y) = |\arccos x - \arccos y| = \arccos(xy + \sqrt{1-x^2}\sqrt{1-y^2}).$$

If $\sigma(x) = (1-x)^2$, then $\tau := \frac{(\sigma\omega)'}{\omega}$ is a polynomial of degree 1, and we set

$$-L(f) = D_J(f) = \frac{(\sigma\omega f')'}{\omega} = \sigma f'' + \tau f'.$$

The operator L is a nonnegative symmetric second order differential operator in $\mathbb{L}_2(\omega(x)dx)$. Using Gram Schmidt orthonormalisation (again, in $\mathbb{L}_2(\omega(x)dx)$) of $\{x^k,\ k \in \mathbb{N}\}$ we get a family of orthonormal polynomials $\{\pi_k,\ k \in \mathbb{N}\}$ called Jacobi polynomials, which coincides with the spectral decomposition of D_J. More precisely,

$$D_J \pi_k := -\lambda_k \pi_k = -k(k+1+\alpha+\beta))\pi_k$$

Then, for any $k \in \mathbb{N}$, $\mathcal{H}_{\lambda_k} = span\{\pi_k\},\ dim(\mathcal{H}_{\lambda_k}) = 1$ and

$$Q_k(x, y) = \pi_k(x)\pi_k(y); \quad \lambda_k = k(k+\alpha+\beta+1).$$

$$e^{-tL}(x, y) = c_{\alpha,\beta} + \sum_{k \geq 1} e^{-tk(k+1+\alpha+\beta))}\pi_k(x)\pi_k(y), \quad c_{\alpha,\beta} \int_M \omega(x)dx = 1.$$

It can be proved, see [9], that for any x, y on $\mathcal{M}$ and any $t \in]0, 1[$,

$$C' \frac{1}{\sqrt{|B(x,\sqrt{t})||B(y,\sqrt{t})|}} e^{-c' \frac{\rho^2(x,y)}{t}} \leq e^{-tL}(x, y) \leq C \frac{1}{\sqrt{|B(x,\sqrt{t})||B(y,\sqrt{t})|}} e^{-c \frac{\rho^2(x,y)}{t}}.$$

Furthermore it can be checked, see [23], that for all $x \in [-1, 1]$ and $0 < r \leq \pi$,

$$|B(x,r)| \sim r((1-x) \vee r^2)^{\alpha+1/2}((1+x) \vee r^2)^{\beta+1/2}.$$

So, in this case, condition (2) is true for $\alpha = \beta = -1/2$ and not fulfilled for other values of α and β.

4 Main results

Before stating the main results, we briefly present the general Bayesian framework.

4.1 Bayesian framework and general result

Data Given a metric space $\mathscr{F}$ equipped with a σ-field $\mathcal{T}$, consider a sequence of statistical experiments $(\mathcal{X}_n, \mathcal{A}_n, \{P_f^{(n)}\}_{f\in\mathscr{F}})$. Suppose there exists a common (σ-finite) dominating measure $\mu^{(n)}$ to all probability measures $\{P_f^{(n)}\}_{f\in\mathscr{F}}$, that is

$$dP_f^{(n)}(x^{(n)}) = p_f^{(n)}(x^{(n)})d\mu^{(n)}(x^{(n)}).$$

We assume that the map $(x^{(n)}, f) \to p_f^{(n)}(x^{(n)})$ is jointly measurable relative to $\mathcal{A}_n \otimes \mathcal{T}$.

Prior We equip the space $(\mathscr{F}, \mathcal{T})$ with a probability measure Π that is called prior. Then the space $\mathcal{X}_n \times \mathscr{F}$ can be naturally equipped of the σ-field $\mathcal{A}_n \otimes \mathcal{T}$ and of the probability measure

$$P(A_n \times T) = \int\int_{A_n\times T} p_f^{(n)}(x^{(n)})d\mu^{(n)}(x^{(n)})d\Pi(f).$$

The marginal in f of this measure is the prior Π. The law $X|f$ is $P_f^{(n)}$.

Bayes formula Under the preceding framework, the conditional distribution of f given the data $X^{(n)}$ is absolutely continuous with respect to Π and is given by, for any measurable set $A \in \mathcal{T}$,

$$\Pi(A\,|X^{(n)}) = \frac{\int_A p_f^{(n)}(X^{(n)})d\Pi(f)}{\int p_f^{(n)}(X^{(n)})d\Pi(f)}.$$

We study the convergence of the posterior measure in a frequentist sense in that we suppose that there exists a 'true' parameter, here an unknown function, denoted f_0. That is, we consider convergence under the law $P_{f_0}^{(n)}$. The expectation under this distribution is denoted $\mathbb{E}_{f_0}$. Theorem 1 in [13] gives sufficient conditions for the posterior to concentrate at rate $\varepsilon_n \to 0$ towards f_0, when n goes to infinity,

$$\mathbb{E}_{f_0}\Pi(f : d_n(f, f_0) \le M\varepsilon_n \,|\, X^{(n)}) \to 1, \tag{16}$$

where d_n is a semi-distance for which certain exponential tests exist, see [13] for details. In [31,33], the case of plain and randomly rescaled Gaussian priors is considered and the authors establish that, as soon as the statistical distance d_n of the problem properly relates to the Banach space norm $\mathbb{B}$, then Conditions (23), (24), (25) defined in the sequel imply the convergence (16).

4.2 Concentration results

Let us recall that we assume that the compact metric space $\mathcal{M}$ satisfies the conditions of Sect. 2, that is the doubling property (1) together with the polynomial-type growth (2) of volume of balls. The operator L is also supposed to verify the properties listed in Sect. 2.4.

Gaussian white noise One observes

$$dX^{(n)}(x) = f(x)dx + \frac{1}{\sqrt{n}}dZ(x), \quad x \in \mathcal{M}.$$

In this case, set $(\mathbb{B}, \|\cdot\|) = (\mathbb{L}_2, \|\cdot\|_2)$. We set the prior Π to be the law Π_w induced by W^T, see (15). Here Π serves directly as a prior on f (so $w = f$ here). Besov spaces are defined in Sect. 9.

Theorem 1 (Gaussian white noise on $(\mathcal{M}, \rho)$, upper-bound) *Let the set $\mathcal{M}$ and the operator L satisfy the properties listed above. Suppose that f_0 is in the Besov space $B^s_{2,\infty}(\mathcal{M})$ with $s > 0$ and that the prior Π on f is W^T given by (15). Let $q = 1 + d/2$ in (14). Set $\varepsilon_n \sim \bar{\varepsilon}_n \sim (\log n/n)^{2s/(2s+d)}$. For M large enough, as $n \to \infty$,*

$$\mathbb{E}_{f_0}\Pi(\|f - f_0\|_2 \geq M\varepsilon_n \mid X^{(n)}) \to 0.$$

The proof of Theorem 1 is given in Sect. 8. For the next two examples, we only give a sketch of the argument.

Fixed design regression With the notation from Sect. 1, the observations are

$$Y_i = f(x_i) + \varepsilon_i, \quad 1 \leq i \leq n.$$

The prior Π is, as above, the law induced by W^T, see (15), and serves directly as a prior on f (so $w = f$ here).

If f_0 is in $B^s_{\infty,\infty}(\mathcal{M})$, with $s > 0$, and $q = 1 + d/2$ in (14), it follows from Sect. 7 that (23), (24), (25) are satisfied with $\varepsilon_n \sim \bar{\varepsilon}_n \sim (\log n/n)^{2s/(2s+d)}$ and $(\mathbb{B}, \|\cdot\|) = (\mathcal{C}^0(\mathcal{M}), \|\cdot\|_\infty)$. This implies, as in [31]-Section 3.3, that if d_n is the semi-distance defined by, for f_1, f_2 in $\mathscr{F}$,

$$d_n(f_1, f_2)^2 = \int (f_1 - f_2)^2 d\mathbb{P}^t_n = \frac{1}{n}\sum_{i=1}^n (f_1 - f_2)^2(x_i),$$

then posterior concentration (16) holds at rate ε_n.

Density estimation The observations are a sample

$$(X_i)_{1 \leq i \leq n} \quad \text{i.i.d.} \quad \sim f,$$

for a density f on $\mathcal{M}$. The true density f_0 is assumed to be continuous and bounded away from 0 and infinity on $\mathcal{M}$. In order to build a prior on densities, we consider the transformation, for any given continuous function $w : \mathcal{M} \to \mathbb{R}$,

$$f_w^{\Lambda}(x) := \frac{\Lambda(w(x))}{\int_{\mathcal{M}} \Lambda(w(u))d\mu(u)}, \quad x \in \mathcal{M},$$

where $\Lambda : \mathbb{R} \to (0, +\infty)$ is such that $\log \Lambda$ is Lipschitz on $\mathbb{R}$ and has an inverse $\Lambda^{-1} : (0, +\infty) \to \mathbb{R}$. For instance, one can take the exponential function as Λ. Here, the function w_0 is taken to be $w_0 := \Lambda^{-1} f_0$. The law of W^T, see (15), here serves as a prior Π_w on w's, which induces a prior Π on densities via the transformation f_w^{Λ}. That is, the final prior Π on densities we consider is $f_{W^T}^{\Lambda}$. In this case we set $(\mathbb{B}, \|\cdot\|) = (\mathcal{C}^0(\mathcal{M}), \|\cdot\|_\infty)$, the Banach space in which the function w and the prior Π_w live.

If f_0 is in $B^s_{\infty,\infty}(\mathcal{M})$, with $s > 0$, and $q = 1 + d/2$ in (14), it follows from Sect. 7 that (23), (24), (25) are satisfied with $\varepsilon_n \sim \bar{\varepsilon}_n \sim (\log n/n)^{2s/(2s+d)}$ and $(\mathbb{B}, \|\cdot\|) = (\mathcal{C}^0(\mathcal{M}), \|\cdot\|_\infty)$. This implies, as in [31, Section 3.1] (see also [33, Thm. 3.1]), that (16) holds with the previous rate, where d_n is the Hellinger distance between densities. The verification is as in [31], extending their Lemma 3.1 to the case of a general Λ with $\log \Lambda$ Lipschitz. The proof is not difficult and is left to the reader.

Discussion In the case that $\mathcal{M}$ is a compact connected orientable manifold without boundary, minimax rates of convergence have been obtained in [11], where Sobolev balls of smoothness index s are considered and data are generated from a regression setting. In particular, in this framework, our procedure is adaptive in the minimax sense for Besov regularities, up to a logarithmic factor.

We have obtained convergence rates for the posterior distribution associated to the geometrical prior in a variety of statistical frameworks. Obtaining these rates does not presuppose any a priori knowledge of the regularity of the function f_0. Therefore our procedure is not only nearly minimax, but also nearly adaptive.

Note also that another attractive property of the method is that it does not assume a priori any (upper or lower) bound on the regularity index $s > 0$. This is related to the fact that approximation is via the spaces $\mathbb{H}_t$, which are made of (super)-smooth functions.

4.3 Lower bound for the rate

Works obtaining (nearly-)adaptive rates of convergence for posterior distributions are relatively recent and so far were obtained for density or regression on subsets of the real line or the Euclidian space. Often, logarithmic factors are reported in the (upper-bound) rates, but it is unclear whether the rate must include such a logarithmic term. We aim at giving an answer to this question in our setting by providing a lower bound for the rate of convergence of our general procedure. This lower bound implies that the rates obtained in Sect. 4 are, in fact, sharp. One can conjecture that the same phenomenon appears for hierarchical Bayesian procedures with randomly rescaled Gaussian priors when the initial Gaussian prior has a RKHS which is made of super-smooth functions (e.g. infinitely differentiable functions), for instance the priors considered in [25,33].

For simplicity we consider the Gaussian white noise model

$$dX^{(n)}(x) = f(x)dx + \frac{1}{\sqrt{n}}dZ(x), \quad x \in \mathcal{M}.$$

We set $(\mathbb{B}, \|\cdot\|) = (\mathbb{L}_2(\mathcal{M}), \|\cdot\|_2)$. As before, for this model the prior sits on the same space as the function f to be estimated, so $w = f$. Again, the set $\mathcal{M}$ and the operator L are as in Sect. 4.2.

Theorem 2 (Gaussian white noise on $(\mathcal{M}, \rho)$, lower bound) *Let $\varepsilon_n = (\log n/n)^{s/(2d+s)}$ for $s > 0$ and let the prior on f be the law induced by W^T, see (15), with $q > 0$ in (14). Then there exist f_0 in the unit ball of $B^s_{2,\infty}(\mathcal{M})$ and a constant $c > 0$ such that*

$$\mathbb{E}_{f_0}\Pi(\|f - f_0\|_2 \le c(\log n)^{0\vee(q-1-\frac{d}{2})}\varepsilon_n \mid X^{(n)}) \to 0.$$

As a consequence, for any prior of the type (15) with any $q > 0$ in (14), the posterior convergence rate cannot be faster than ε_n above. If q is larger than $1 + d/2$, the rate becomes slower than ε_n.

Remark 1 More generally, an adaptation of the proof of Theorem 2 yields that, for any 'reasonable' prior on T, in that, for $\varepsilon_n \sim (\log n/n)^{s/(2d+s)}$, it holds

$$\Pi[\|f - f_0\|_2 \le \varepsilon_n] \ge e^{-Cn\varepsilon_n^2},$$

then $\Pi[\|f - f_0\|_2 \le c\varepsilon_n|X] \to 0$ for small enough $c > 0$. This condition is the standard 'prior mass' condition in checking upper-bound rates, see (23). Note that the previous display is automatically implied if the prior satisfies $\Pi[\|f - f_0\|_2 \le \varepsilon_n^*] \ge e^{-Cn\varepsilon_n^{*2}}$ for $\varepsilon_n^* = n^{-s/(2d+s)}$, or more generally for any rate at least as fast as ε_n. For instance, this can be used to check that taking a uniform prior on $(0, 1)$ as law for T leads to the same lower bound rate.

4.4 Structure of the proofs

The proof of Theorem 1 is split in two different parts. The first part considers the properties of the heat kernel Gaussian process and the concentration of the corresponding posterior measure. A key result for this part lies in sharp entropy estimates for the process and is stated in Sect. 5. Establishing such sharp estimates is considered in the second part of the proofs.

5 RKHS and heat kernel Gaussian process

This section focuses on the analysis of the underlying Gaussian process W^t in (13), for $t > 0, x \in \mathcal{M}$,

$$W^t(x) = \sum_k \sum_l e^{-\lambda_k t/2} X_k^l e_k^l(x).$$

The Gaussian process $(W^t(x))_{x\in\mathcal{M}}$ is centered and its covariance kernel precisely coincides with the heat kernel on $\mathcal{M}$, as noted above. Since $P_t(\cdot,\cdot)$ is a covariance kernel for any fixed $t > 0$, it is associated to a Reproducing Kernel Hilbert Space (RKHS) $\mathbb{H}_t$, which is also the RKHS of the Gaussian process W^t, see [32] for definition and properties of the RKHS. Also, $\mathbb{H}_t$ is the isometric image of $\mathbb{L}^2$ by $P_t^{1/2} = P_{t/2}$. So the family $\{e^{-\lambda_k t/2}e_k^l,\ k\in\mathbb{N},\ 1\le l\le dim(\mathcal{H}_{\lambda_k})\}$ is a 'natural' orthonormal basis of $\mathbb{H}_t$.

The RKHS $\mathbb{H}_t$ has also the following description, for any $t > 0$:

$$\mathbb{H}_t = \left\{ h^t = \sum_k \sum_l a_k^l e^{-\lambda_k t/2} e_k^l, \qquad \sum_{k,l} |a_k^l|^2 < +\infty \right\},$$

equipped with the inner product

$$\left\langle \sum_k \sum_l a_k^l e^{-\lambda_k t/2} e_k^l \ , \ \sum_k \sum_l b_k^l e^{-\lambda_k t/2} e_k^l \right\rangle_{\mathbb{H}_t} = \sum_k \sum_l a_k^l b_k^l.$$

Hence, if we denote by $\mathbb{H}_t^1$ the unit ball of $\mathbb{H}_t$:

$$f\in\mathbb{H}_t^1 \iff f = \sum_k \sum_l a_k^l e^{-\lambda_k t/2} e_k^l(x), \qquad \sum_{k,l} |a_k^l|^2 \le 1.$$

5.1 Entropy of the RKHS unit ball

Let (X,ρ) be a metric space. For $\varepsilon > 0$, we define, as usual, the covering number $N(\varepsilon, X)$ as the smallest number of balls of radius ε covering X. The entropy $H(\varepsilon, X)$ is by definition $H(\varepsilon, X) = \log_2 N(\varepsilon, X)$.

An important result of this section is the link between the covering number $N(\varepsilon,\mathcal{M},\rho)$ of the space $\mathcal{M}$, and $H(\varepsilon,\mathbb{H}_t^1,\mathbb{L}^p)$ for $p = 2,\ \infty$, where $\mathbb{H}_t^1$ is the unit ball of the RKHS defined above. More precisely we prove in Sect. 10 the following theorem:

Theorem 3 *Let $\mathcal{M}$ be a compact metric space satisfying (1), on which a self-adjoint positive operator L exists such that e^{-tL} has kernel P_t satisfying the properties listed in Sect. 2.4. Let us fix $\nu > 0,\ a > 0$. There exists $\varepsilon_0 > 0$ such that for ε, t with $\varepsilon^\nu \le at$ and $0 < \varepsilon \le \varepsilon_0$,*

$$H(\varepsilon,\mathbb{H}_t^1,\mathbb{L}^2) \sim H(\varepsilon,\mathbb{H}_t^1,\mathbb{L}^\infty) \sim N(\delta(t,\varepsilon),\mathcal{M})\cdot\log\frac{1}{\varepsilon} \quad \textit{where} \quad \frac{1}{\delta(t,\varepsilon)} := \sqrt{\frac{1}{t}\log\left(\frac{1}{\varepsilon}\right)}.$$

Remark 2 Theorem 3 gives the precise behaviour up to constants, from above and below, of the entropy of the RKHS unit ball $\mathbb{H}_t^1$. The constants involved depend only on $\mathcal{M}, a, \nu$. The mild restriction on the range of t arises from technical reasons in the

proof of the upper-bound. As an examination of the proof reveals, this restriction is not needed in the proof of the lower bound.

5.2 Entropy under Ahlfors' condition

In Sect. 10, the general case is considered, but for sake of simplicity, in the sequel we focus on the case where Ahlfors' condition (2) is fullfilled. In this case, Theorem 3 takes the following form.

Proposition 1 *Under the conditions of Theorem 3, suppose additionally that $\mathcal{M}$ satisfies (2). If Λ_ε is a maximal ε-net of $\mathcal{M}$, then*

$$\frac{1}{c_2}\left(\frac{1}{\varepsilon}\right)^d \leq N(\varepsilon, \mathcal{M}) \leq card(\Lambda_\varepsilon) \leq \frac{2^d}{c_1}\left(\frac{1}{\varepsilon}\right)^d \tag{17}$$

$$H(\varepsilon, \mathbb{H}_t^1, \mathbb{L}^2) \sim H(\varepsilon, \mathbb{H}_t^1, \mathbb{L}^\infty) \sim \left(\frac{1}{\delta(t,\varepsilon)}\right)^d \log\frac{1}{\varepsilon}, \tag{18}$$

for all $0 < \varepsilon \leq \varepsilon_0$, where we suppose that for some $\nu > 0$ and $a > 0$, it holds $\varepsilon^\nu \leq at$.

Proof Let $(B(x_i, \varepsilon))_{i\in I}$ be a minimal covering of $\mathcal{M}$. Then,

$$1 = |\mathcal{M}| \leq \sum_{i\in I} |B(x_i, \varepsilon)| \leq N(\varepsilon, \mathcal{M}) c_2 \varepsilon^d.$$

On the other hand, if Λ_ε is any maximal $\varepsilon-$net, we have:

$$1 = |\mathcal{M}| \geq \sum_{\xi\in\Lambda_\varepsilon} |B(\xi, \varepsilon/2)| \geq card(\Lambda_\varepsilon) c_1 (\varepsilon/2)^d.$$

□

6 Proofs I: Geometrical Prior, concentration function

6.1 Approximation and small ball probabilities

The so-called concentration function of a Gaussian process defined below turns out to be fundamental to prove sharp concentration of the posterior measure. For this reason we focus now on the detailed study of this function for the geometrical prior.

In the sequel, the notation $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$ is used for anyone of the two spaces

$$(\mathbb{B}, \|\cdot\|_{\mathbb{B}}) = (\mathcal{C}^0(\mathcal{M}), \|\cdot\|_\infty) \quad \text{or} \quad (\mathbb{B}, \|\cdot\|_{\mathbb{B}}) = (\mathbb{L}^2, \|\cdot\|_2).$$

Any property stated below with a $\|\cdot\|_{\mathbb{B}}$-norm holds for both spaces.

Concentration function Consider the Gaussian process W^t defined in (13), for a fixed $t \in (0, 1]$. Its concentration function within $\mathbb{B}$ is defined, for any function w_0 in $\mathbb{B}$, as the sum of two terms

$$\begin{aligned}\varphi^t_{w_0}(\varepsilon) &= \inf_{h_t \in \mathbb{H}_t,\ \|w_0 - h_t\|_{\mathbb{B}} < \varepsilon} \frac{\|h_t\|^2_{\mathbb{H}_t}}{2} - \log \mathbb{P}(\|W^t\|_{\mathbb{B}} < \varepsilon) \\ &:= \quad A^t_{w_0}(\varepsilon) \quad + \quad S^t(\varepsilon).\end{aligned}$$

The approximation term $A^t_{w_0}(\varepsilon)$ quantifies how well w_0 is approximable by elements of the RKHS $\mathbb{H}_t$ of the prior while keeping the 'complexity' of those elements, quantified in terms of RKHS-norm, as small as possible. The term $A^t_{w_0}(\varepsilon)$ is finite for all $\varepsilon > 0$ if and only if w_0 lies in the closure in $\mathbb{B}$ of $\mathbb{H}_t$ (which can be checked to coincide with the support of the Gaussian prior, see [32, Lemma 5.1]) It turns out that for the prior W^t, this closure is $\mathbb{B}$ itself, as follows from the approximation results below.

In order to have a precise calibration of $A^t_{w_0}(\varepsilon)$, we will assume regularity conditions on the function w_0, which in turn will yield the rate of concentration. Namely we shall assume that w_0 belongs to a regularity class $\mathscr{F}_s(\mathcal{M})$, $s > 0$, taken equal to a Besov space

$$\mathscr{F}_s(\mathcal{M}) = B^s_{\infty,\infty}(\mathcal{M}) \text{ if } \mathbb{B} = \mathcal{C}^0(\mathcal{M}) \qquad (\text{resp. } B^s_{2,\infty}(\mathcal{M}) \text{ if } \mathbb{B} = \mathbb{L}^2).$$

The problem of the regularity assumption in a context like here is not a simple one. We took here a natural generalisation of the definition of usual spaces on the real line, by means of approximation properties. For more details we refer to Sect. 9.

Approximation term $A^t_{w_0}(\varepsilon)$-regularity assumption on $\mathcal{M}$. For any w_0 in the Banach space $\mathbb{B}$, consider the following sequence of approximations. For $\delta \to 0$, the operator L and Φ a Littlewood–Paley function, see (33) in Sect. 9, define

$$\Phi(\delta\sqrt{L})w_0 := \sum_{\lambda_k \le \delta^{-2}} \Phi(\delta\sqrt{\lambda_k}) P_{\mathcal{H}_{\lambda_k}} w_0,$$

where $P_{\mathcal{H}_{\lambda_k}}$ is the projector onto the eigenspace $\mathcal{H}_{\lambda_k}$. For any $\delta > 0$, the sum in the last display is finite thus $\Phi(\delta\sqrt{L})w_0$ belongs to $\mathbb{H}_t$. It directly follows from the definition of the considered Besov spaces, see (36) in Sect. 9, that, for $w_0 \in \mathscr{F}_s(\mathcal{M})$,

$$\|\Phi(\delta\sqrt{L})w_0 - w_0\|_{\mathbb{B}} \le C\delta^s =: C\varepsilon.$$

On the other hand, making use of the choice $\delta^s =: \varepsilon$,

$$\begin{aligned}\|\Phi(\delta\sqrt{L})w_0\|^2_{\mathbb{H}_t} &= \sum_{\lambda_k \le \delta^{-2}} |\Phi(\delta\sqrt{\lambda_k})|^2 e^{\lambda_k t} \|P_{\mathcal{H}_{\lambda_k}} w_0\|^2_2 \\ &\le C \sum_{\lambda_k \le \delta^{-2}} e^{\lambda_k t} \|P_{\mathcal{H}_{\lambda_k}} w_0\|^2_2 \\ &\le C e^{t\varepsilon^{-2/s}} \|w_0\|^2_2.\end{aligned}$$

Note that $\|w_0\|_2 \leq 1$ if we suppose that w_0 is in the unit ball of $\mathscr{F}_s(\mathcal{M})$ (since necessarily $\|w_0\|_{\mathbb{B}}$ is bounded by 1 and, for the case of the sup-norm, since $\mathcal{M}$ is compact with μ-measure 1). Hence,

$$A^t_{w_0}(\varepsilon) \leq Ce^{t\varepsilon^{-2/s}}, \quad \text{if } w_0 \in \mathscr{F}_s(\mathcal{M}). \tag{19}$$

Note that this is precisely the place where the regularity of the function plays a role.

Small ball probability $S^t(\varepsilon)$. Let us now show in successive steps that the following upper-bound on the small ball probability of the Gaussian process W^t viewed as a random element in $\mathbb{B}$ holds.

Proposition 2 *Fix $A > 0$. There exists a universal constant $\varepsilon_0 > 0$, and constants $C_0, C_1 > 0$ which depend on $d, A, \mathbb{B}$ only, such that, for any $\varepsilon \leq \varepsilon_0$ and any $t \in [C_1\varepsilon^A, 1]$,*

$$-\log \mathbb{P}\left(\|W^t\|_{\mathbb{B}} \leq \varepsilon\right) \leq C_0 \left(\frac{1}{\sqrt{t}}\right)^d \left(\log \frac{1}{\varepsilon}\right)^{1+\frac{d}{2}}. \tag{20}$$

The steps of the proof follow the method proposed by [33]. The starting point is a bound on the entropy of the sunit ball $\mathbb{H}^1_t$ of $\mathbb{H}_t$ with respect to the sup-norm, which is a direct consequence of (18) and is summarised by the following:

There exists a universal constant $\varepsilon_1 > 0$, and constants $C_2, C_3 > 0$ which depend on d, A only, such that, for any $\varepsilon \leq \varepsilon_1$ and any $t \in [C_2\varepsilon^A, 1]$,

$$\log N(\varepsilon, \mathbb{H}^1_t, \|\cdot\|_{\mathbb{B}}) \leq C_3 \left(\frac{1}{\sqrt{t}}\right)^d \left(\log \frac{1}{\varepsilon}\right)^{1+\frac{d}{2}}. \tag{21}$$

Step 1, crude bound. Let u_t be the mapping canonically associated to W^t considered in [18] and, as in this paper, set

$$\begin{aligned} e_n(u_t) &:= \inf\{\eta > 0,\ N(\eta, \mathbb{H}^1_t, \|\cdot\|_{\mathbb{B}}) \leq 2^{n-1}\} \\ &\leq \inf\{0 < \eta < t,\ \log N(\eta, \mathbb{H}^1_t, \|\cdot\|_{\mathbb{B}}) \leq (n-1)\log 2\}. \end{aligned}$$

By definition, the previous quantity is smaller than the solution of the following equation in η, where we use the bound (21),

$$Ct^{-\frac{d}{2}} \log^{1+\frac{d}{2}} \frac{1}{\eta} = n$$

that is $\eta = \exp\{-Cn^{\frac{2}{2+d}} t^{\frac{d}{2+d}}\}$. Thus

$$e_n(u_t) \leq \exp\{-Cn^{\frac{2}{2+d}} t^{\frac{d}{2+d}}\}, \quad n \geq 1.$$

The first equation of [29], p. 300 can be written

$$\sup_{k\le n} k^{\alpha} e_k(u_t^*) \le 32 \sup_{k\le n} k^{\alpha} e_k(u_t).$$

We have, for any $k \ge 1$ and any $m \ge 1$,

$$\begin{aligned} k^{2m} e_k(u_t) &\le k^{2m} \exp\{-Ck^{\frac{2}{2+d}} t^{\frac{d}{2+d}}\} \\ &\le t^{-md} (k^2 t^d)^m \exp\{-C(k^2 t^d)^{\frac{1}{2+d}}\} \\ &\le t^{-md} V_m(k^2 t^d), \end{aligned}$$

where $V_m : x \to x^m e^{-Cx^{\frac{1}{2+d}}}$ is uniformly bounded on $(0, +\infty)$ by a finite constant c_m (we omit the dependence in d in the notation). It follows that for any $n \ge 1$,

$$\begin{aligned} n^{2m} e_n(u_t^*) &\le \sup_{k\le n} k^{2m} e_k(u_t^*) \\ &\le 32 \sup_{k\le n} k^{2m} e_k(u_t) \\ &\le 32 c_m t^{-md}. \end{aligned}$$

We have obtained $e_k(u_t) \le 32 c_m t^{-md} k^{-2m}$ for any $k \ge 1$. Lemma 2.1 in [18], itself cited from [24], can be written as follows. If $\ell_n(u_t)$ denotes the n-th approximation number of u_t as defined in [18, p. 1562],

$$\ell_n(u_t) \le c_1 \sum_{k \ge c_2 n} e_k(u_t^*) k^{-1/2} (1 + \log k).$$

From the bound on $e_k(u_t^*)$ above one deduces, for some constant c_m' depending only on m, for any $n \ge 1$,

$$\ell_n(u_t) \le c_m' t^{-d} n^{1-2m}.$$

Consider the definitions, for any $\varepsilon > 0$ and $t > 0$,

$$n_t(\varepsilon) := \max\{n : 4\ell_n(u_t) \ge \varepsilon\}, \quad \sigma(W^t) = \mathbb{E}\left[\|W^t\|^2\right]^{1/2}.$$

A sufficient condition for $n_t(\varepsilon)$ to exist is $4\sigma(W^t) \ge \varepsilon$, since $\ell_n(u_t) \le \ell_1(u_t) = \sigma(W^t)$. So, provided $\varepsilon \le 4\sigma(W^t)$, the bound on ℓ_n implies $n_t(\varepsilon) \le C_m(\varepsilon^{-1} t^{-d})^{1/(2m-1)}$.

The following result makes Proposition 2.3 in [18] precise with respect to constants involving the process under consideration. This is important in our context since we consider a collection of processes $\{W_t\}$ indexed by t and need to keep track of the dependence in t.

Proposition 3 *Let X be centered Gaussian in a real separable Banach space $(E, \|\cdot\|)$. Define $n(\varepsilon)$ and $\sigma(X)$ as above. Then for a universal constant $C_4 > 0$, any $\varepsilon \leq 1 \wedge (4\sigma(X))$,*

$$-\log \mathbb{P}\left[\|X\| < \varepsilon\right] \leq C_4 n(\varepsilon) \log\left[\frac{6n(\varepsilon)(\sigma(X) \vee 1)}{\varepsilon}\right].$$

Explicit upper and lower bounds for $\sigma(W^t)$ are given in Sect. 10, see (50). In the 'polynomial case', see (2), these bounds imply, uniformly in the interval of t's considered, that $1 \lesssim \sigma(W^t) \lesssim \varepsilon^{-B}$ for some $B > 0$.

Combining this fact with Proposition 3 and the previous bound on n_t, we obtain that for some positive constants $C_7, \varepsilon_3, \zeta$, for any $\varepsilon \leq \varepsilon_3$ and $t \in [C_2\varepsilon^A, 1]$

$$S^t(\varepsilon) = -\log \mathbb{P}\left(\|\,|W^t\|_{\mathbb{B}} \leq \varepsilon\right) \leq C_7 \varepsilon^{-\zeta}. \tag{22}$$

Step 2, general link between entropy and small ball. According to Lemma 1 in [17], we have, if G is the distribution function of the standard Gaussian distribution (see their formula (3.19), or (3.2)),

$$S^t(2\varepsilon) + \log G(\lambda + G^{-1}(e^{-S^t(\varepsilon)})) \leq \log N\left(\frac{\varepsilon}{\lambda}, \mathbb{H}_t^1, \|\cdot\|_{\mathbb{B}}\right).$$

Lemma 4.10 in [33] implies, for every $x > 0$,

$$G(\sqrt{2x} + G^{-1}(e^{-x})) \geq 1/2.$$

Take $\lambda = \sqrt{2S^t(\varepsilon)}$ in the previous display. Then for values of t, ε such that (21) holds,

$$S^t(2\varepsilon) + \log\frac{1}{2} \leq C\left(\frac{1}{\sqrt{t}}\right)^d \left(\log\frac{S^t(\varepsilon)}{\varepsilon}\right)^{1+\frac{d}{2}}.$$

Finally combine this with (22) to obtain the desired Equation (20) that is

$$S^t(\varepsilon) \leq C\left(\frac{1}{\sqrt{t}}\right)^d \left(\log\frac{1}{\varepsilon}\right)^{1+\frac{d}{2}}.$$

under the conditions $\varepsilon \leq \varepsilon_3$ and $C_2\varepsilon^A \leq t \leq 1$.

7 Proofs I: General conditions for posterior rates

A general theory to obtain convergence rates for posterior distributions for some distances is presented in [12,13]. The object of interest is a function f_0 (e.g. a regression function, a density function etc.). In some cases, for instance density estimation with Gaussian priors, one cannot directly put the prior on the density itself (a Gaussian prior does not lead to positive paths). This is why we parameterise the considered statistical

problem with the help of a function w_0 in some separable Banach space $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$ of functions defined over $(\mathcal{M}, \rho)$. As already noticed in Sect. 4, in some cases (e.g. regression) w_0 and f_0 coincide, in others not (e.g. density estimation). As before, $\mathbb{B}$ is either $\mathcal{C}^0(\mathcal{M})$ or $\mathbb{L}^2$.

In this section we check that there exist Borel measurable subsets B_n in $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$ such that, for some vanishing sequences ε_n and $\bar{\varepsilon}_n$, some $C > 0$ and n large enough,

$$\mathbb{P}(\|W^T - w_0\|_{\mathbb{B}} \leq \varepsilon_n) \geq e^{-Cn\varepsilon_n^2} \tag{23}$$

$$\mathbb{P}(W^T \notin B_n) \leq e^{-(C+4)n\varepsilon_n^2} \tag{24}$$

$$\log N(\bar{\varepsilon}_n, B_n, \|\cdot\|_{\mathbb{B}}) \leq n\bar{\varepsilon}_n^2 \tag{25}$$

This will imply, as in [31], that the posterior concentrates at rate ε_n around f_0, see Sect. 8. In [33], the authors also follow this approach. One advantage of the prior considered here is that, contrary to [33], the RKHS unit balls are precisely nested as the time parameter t varies, see (28). This leads to slightly simplified proofs.

Prior mass For any fixed function w_0 in $\mathbb{B}$ and any $\varepsilon > 0$, by conditioning on the value taken by the random variable T,

$$\mathbb{P}(\|W^T - w_0\|_{\mathbb{B}} < 2\varepsilon) = \int_0^1 \mathbb{P}(\|W^t - w_0\|_{\mathbb{B}} < 2\varepsilon) g(t) dt.$$

The following inequality links mass of Banach-space balls for Gaussian priors with their concentration function in $\mathbb{B}$, see [32, Lemma 5.3],

$$e^{-\varphi_{w_0}^t(\varepsilon)} \leq \mathbb{P}(\|W^t - w_0\|_{\mathbb{B}} < 2\varepsilon) \leq e^{-\varphi_{w_0}^t(2\varepsilon)},$$

for any w_0 in the support of W^t. We have seen above that any f_0 in $\mathscr{F}^s(\mathcal{M})$ belongs to the support of the prior. It is not hard to adapt the argument to check that in fact any f_0 in $\mathbb{B}$ can be approximated in $\mathbb{B}$ by a sequence of elements in the RKHS $\mathbb{H}_t$ and thus belongs to the support in $\mathbb{B}$ of the prior by Lemma 5.1 in [32]. Then

$$\begin{aligned}\mathbb{P}(\|W^T - w_0\|_{\mathbb{B}} < 2\varepsilon) &\geq \int_0^1 e^{-\varphi_{w_0}^t(\varepsilon)} g(t) dt \\ &\geq \int_{t_\varepsilon^*}^{2t_\varepsilon^*} e^{-\varphi_{w_0}^t(\varepsilon)} g(t) dt,\end{aligned}$$

for some t_ε^* to be chosen.

The concentration function is bounded from above, assuming $\varepsilon \leq \varepsilon_3$ and $t \in [C_2\varepsilon^A, 1]$, by

$$\varphi^t_{w_0}(\varepsilon) \leq C\left[e^{\varepsilon^{-2/s}t} + \left(\frac{1}{\sqrt{t}}\right)^d \left(\log\frac{1}{\varepsilon}\right)^{1+\frac{d}{2}}\right]$$

Set $t^*_\varepsilon = \delta\varepsilon^{\frac{2}{s}} \log\frac{1}{\varepsilon}$ with δ small enough to be chosen. This is compatible with the above conditions provided $A > 2/s$. Then for ε small enough and any $t \in [t^*_\varepsilon, 2t^*_\varepsilon]$,

$$\varphi^t_{w_0}(\varepsilon) \leq C\left[\varepsilon^{-2\delta} + \delta^{-d}\varepsilon^{-\frac{d}{s}}\left(\log\frac{1}{\varepsilon}\right)\right].$$

Set $\delta = d/(4s)$. One obtains, for any $t \in [t^*_\varepsilon, 2t^*_\varepsilon]$,

$$\varphi^t_{w_0}(\varepsilon) \leq C_d\varepsilon^{-\frac{d}{s}}\left(\log\frac{1}{\varepsilon}\right).$$

Inserting this estimate in the previous bound on the prior mass, one gets, together with (14), for ε small enough and $q \leq 1 + d/2$,

$$\begin{aligned}
\mathbb{P}(\|W^T - w_0\|_{\mathbb{B}} < 2\varepsilon) &\geq t^*_\varepsilon e^{-C\varepsilon^{-\frac{d}{s}}\left(\log\frac{1}{\varepsilon}\right)}\left[\inf_{t\in[t^*_\varepsilon, 2t^*_\varepsilon]} g(t)\right] \\
&\geq Ct^{*1-a}_\varepsilon e^{-t^{*-\frac{d}{2}}_\varepsilon\left(\log\frac{1}{t^*_\varepsilon}\right)^q - C\varepsilon^{-\frac{d}{s}}\left(\log\frac{1}{\varepsilon}\right)} \\
&\geq C\varepsilon^{2(1-a)/s}(\log\frac{1}{\varepsilon})^{1-a} e^{-C\varepsilon^{-\frac{d}{s}}\left(\log\frac{1}{\varepsilon}\right)^{q-\frac{d}{2}} - C\varepsilon^{-\frac{d}{s}}\left(\log\frac{1}{\varepsilon}\right)} \\
&\geq Ce^{-C'\varepsilon^{-\frac{d}{s}}\left(\log\frac{1}{\varepsilon}\right)}. \qquad (26)
\end{aligned}$$

Condition (23) is satisfied for the choice

$$\varepsilon_n \sim \left(\frac{n}{\log n}\right)^{-\frac{s}{2s+d}}. \qquad (27)$$

Sieve The idea is to build sieves using Borell's inequality. Recall here that $\mathbb{H}^1_r$ is the unit ball of the RKHS of the centered Gaussian process W^r, viewed as a process on the Banach space $\mathbb{B}$. The notation $\mathbb{B}_1$ (as well as $\mathbb{H}^1_r$) stands for the unit ball of the associated space.

First, notice that from the explicit form of the RKHS of W^t, we have

$$\text{If } t_2 \geq t_1, \quad \text{then} \quad \mathbb{H}^1_{t_2} \subset \mathbb{H}^1_{t_1}. \qquad (28)$$

Let us set for $M = M_n$, $\varepsilon = \varepsilon_n$ and $r > 0$ to be chosen later,

$$B_n = M\mathbb{H}^1_r + \varepsilon\mathbb{B}_1,$$

Consider the case $t \geq r$, then using (28)

$$\begin{aligned}\mathbb{P}(W^t \notin B_n) &= \mathbb{P}(W^t \notin M\mathbb{H}_r^1 + \varepsilon\mathbb{B}_1) \\ &\leq \mathbb{P}(W^t \notin M\mathbb{H}_t^1 + \varepsilon\mathbb{B}_1) \\ &\leq 1 - G(G^{-1}(e^{-S^t(\varepsilon)}) + M). \end{aligned} \tag{29}$$

where the last line follows from Borell's inequality.

Choices of ε, r and M. Let us set $\varepsilon = \varepsilon_n$ given by (27) and

$$r^{-\frac{d}{2}} \sim \frac{n\varepsilon_n^2}{(\log n)^{1+\frac{d}{2}}} \quad \text{and} \quad M^2 \sim n\varepsilon_n^2. \tag{30}$$

First, one checks that r belongs to $[C_2\varepsilon^A, 1]$. This is clear from the definition since we have assumed $A > 2/s$. Then any $t \in [r, 1]$ also belongs to $[C_2\varepsilon^A, 1]$ so we can use the entropy bound and write

$$S^t(\varepsilon) \leq Ct^{-\frac{d}{2}} \left(\log \frac{1}{\varepsilon_n}\right)^{1+\frac{d}{2}} \leq Cr^{-\frac{d}{2}} \left(\log \frac{1}{\varepsilon_n}\right)^{1+\frac{d}{2}} =: S_n^*.$$

Now the bounds $-\sqrt{2\log(1/u)} \leq G^{-1}(u) \leq -\frac{1}{2}\sqrt{\log(1/u)}$ valid for $u \in (0, 1/4)$ imply that

$$1 - G(G^{-1}(e^{-S^t(\varepsilon)}) + M) \leq 1 - G(G^{-1}(e^{-S_n^*}) + M) \leq e^{-M^2/8},$$

as soon as $M \geq 4\sqrt{S_n^*}$ and $e^{-S_n^*} < 1/4$.

To check $e^{-S_n^*} < 1/4$ note that $S_n^* \geq S^r(\varepsilon)$ which can be further bounded from below using Equation (3.1) in [17] which leads to, for any $\varepsilon, \lambda > 0$,

$$S^r(\varepsilon) \geq H(2\varepsilon, \lambda\mathbb{H}_r^1) - \frac{\lambda^2}{2} \geq Cr^{-\frac{d}{2}} \left(\log \frac{\lambda}{\varepsilon}\right)^{1+\frac{d}{2}} - \frac{\lambda^2}{2}.$$

Here we have used the bound from below of the entropy see (18). Then take $\lambda = 1$ to obtain $S_n^*(\varepsilon) \geq \log(4)$ for ε small enough.

The first inequality $M \geq 4\sqrt{S_n^*}$ is satisfied if

$$M^2 \geq 16r^{-\frac{d}{2}} \left(\log \frac{1}{\varepsilon_n}\right)^{1+\frac{d}{2}},$$

and this holds for the choices of r and M given by (30). Hence for large enough n,

$$\begin{aligned}\mathbb{P}(W^t \notin B_n) &\leq e^{-M^2/8} \\ &\leq e^{-Cn\varepsilon_n^2}.\end{aligned}$$

Then we can write, if $q \geq 1 + d/2$,

$$\begin{aligned}\mathbb{P}(W^T \notin B_n) &= \int_0^1 \mathbb{P}(W^t \notin B_n) g(t) dt \\ &\leq \mathbb{P}(T < r) + \int_r^1 \mathbb{P}(W^t \notin B_n) g(t) dt \\ &\leq C r^{-c} e^{-C' r^{-d/2} \log^q(\frac{1}{r})} + e^{-M^2/8} \leq e^{-Cn\varepsilon_n^2}.\end{aligned}$$

Entropy It is enough to bound from above

$$\begin{aligned}\log N(2\varepsilon_n, M\mathbb{H}_r^1 + \varepsilon_n \mathbb{B}_1, \|\cdot\|_{\mathbb{B}}) &\leq \log N(\varepsilon_n, M\mathbb{H}_r^1, \|\cdot\|_{\mathbb{B}}) \\ &\leq r^{-d/2} \left(\log \frac{M}{\varepsilon_n}\right)^{1+\frac{d}{2}} \\ &\leq Cn\varepsilon_n^2,\end{aligned}$$

where we have used (21) to obtain the one but last inequality.

8 Proofs I: Posterior concentration

Proof of Theorem 1 We check that (23), (24) and (25) are satisfied with $(\mathbb{B}, \|\cdot\|_{\mathbb{B}}) = (\mathbb{L}^2, \|\cdot\|_2)$. With the considered prior, it follows from Sect. 7 that, since $q \leq 1 + d/2$, Condition (23) holds and, since $q \geq 1 + d/2$, Condition (24) holds. Also, (25) holds with the choice of B_n from Sect. 7, regardless of the value of q. One can then apply the general rate result (Theorem 1 in [13]), with the distance d_n in (16) chosen to be the $\mathbb{L}^2$-norm, see [13] Sect. 5. The end of the proof is as in [31], Theorem 3.1 and 3.4, and is omitted. □

Proof of Theorem 2 We use a general approach to prove lower bounds for posterior measures introduced in [5] (see [5,6] for examples). The idea is to apply the following lemma (Lemma 1 in [13]) to the sets $\{f \in \mathbb{B},\ \|f - f_0\|_{\mathbb{B}} \leq \zeta_n\}$, for some rate $\zeta_n \to 0$ and f_0 in $B^s_{2,\infty}$, with $s > 0$.

Lemma 1 *Let $\alpha_n \to 0$ such that $n\alpha_n^2 \to +\infty$ as $n \to \infty$ and let B_n be a measurable set such that*

$$\Pi(B_n)/\Pi(B_{KL}(f_0, \alpha_n)) \leq e^{-2n\alpha_n^2},$$

where, in the white noise model, $B_{KL}(f_0, \alpha_n) = \{f : \|f - f_0\|_2 \leq \alpha_n\}$. Then $\mathbb{E}_{f_0} \Pi(B_n \mid X^{(n)}) \to 0$.

In our context this specialises as follows. Let $\alpha_n \to 0$ and $n\alpha_n^2 \to +\infty$. Suppose that, as $n \to +\infty$,

$$\frac{\Pi(\|f - f_0\|_2 \leq \zeta_n)}{\Pi(\|f - f_0\|_2 \leq \alpha_n)} = o(e^{-2n\alpha_n^2}).$$

Then ζ_n is a lower bound for the rate of the posterior in that, as $n \to +\infty$,

$$\mathbb{E}_{f_0} \Pi(\|f - f_0\|_2 \leq \zeta_n) \mid X^{(n)}) \to 0.$$

We first deal with the case where $q \leq 1 + d/2$. In this case let us choose $\alpha_n = 2\varepsilon_n$, where $\varepsilon_n = (\log n/n)^{2s/(2s+d)}$. In Sect. 7, we have established in (26) that, for the prior W^T with $q \leq 1 + d/2$ in (14), there exists $C > 0$ with

$$\Pi(\|f - f_0\|_2 \leq \varepsilon_n) = \mathbb{P}(\|W^T - w_0\|_{\mathbb{B}} \leq \varepsilon_n) \geq e^{-Cn\varepsilon_n^2}.$$

So it is enough to show that, for some well-chosen $\zeta_n \to 0$,

$$\Pi(\|f - f_0\|_2 \leq \zeta_n)) = o(e^{-(8+C)n\varepsilon_n^2}). \tag{31}$$

We would like to take $\zeta_n = c\varepsilon_n$, for some (small) constant $c > 0$. In order to bound from above the previous probability, we write

$$\begin{aligned}\Pi[\|f - f_0\|_2 \leq \zeta_n] &= \int_0^1 \Pi[\|W^t - f_0\|_2 \leq \zeta_n] g(t) dt \\ &\leq \int_0^1 \exp\left[-\varphi_{f_0}^t(\zeta_n)\right] g(t) dt.\end{aligned}$$

We separate the above integral in two parts. The first one is $\mathcal{T}_1 := \{\mu_n \leq t \leq Bt_n^*\}$, where t_n^* is a similar cut-off as in the upper-bound proof $t_n^* = \zeta_n^{2/s} \log(1/\zeta_n)$. On $\mathcal{T}_1$, one can bound from below $\varphi_{f_0}^t(\zeta_n)$ by its small ball probability part $\varphi_0^t(\zeta_n)$. Moreover, thanks to relation (3.1) in [17], we have, for any $\lambda > 0$ and $t \in (0, 1]$,

$$\varphi_0^t(\zeta_n) = -\log \mathbb{P}[\|W^t\|_2 < \zeta_n] \geq H(2\zeta_n, \lambda \mathbb{H}_t^1, \|\cdot\|_2) - \frac{\lambda^2}{2}.$$

Set $\lambda = 1$ and recall from Remark 2 that the lower bound on the entropy can be used for any t regardless of the value of ε. This yields, for large enough n, if $\zeta_n = o(1)$,

$$\varphi_0^t(\zeta_n) \geq C(Bt_n^*)^{-d/2} \log^{1+d/2}(1/\zeta_n) - \frac{1}{2} \geq CB^{-d/2} \zeta_n^{-d/s} \log(1/\zeta_n).$$

Thus we obtain

$$\int_0^{Bt_n^*} \exp\left[-\varphi_{f_0}^t(\zeta_n)\right] g(t)dt \le e^{-CB^{-d/2}\zeta_n^{-d/s}\log(1/\zeta_n)} \int_0^{Bt_n^*} g(t)dt \le e^{-CB^{-d/2}\zeta_n^{-d/s}\log(1/\zeta_n)}.$$

This is less than $e^{-(8+C)n\varepsilon_n^2}$ provided $\zeta_n = \kappa\varepsilon_n$ and $\kappa > 0$ is small enough.

It remains to bound the integral from above on $\mathcal{T}_2 := \{Bt_n^* \le t \le 1\}$. Here we bound $\varphi_{f_0}^t(\zeta_n)$ from below by its approximation part. For any $t \in \mathcal{T}_2$,

$$\varphi_{f_0}^t(\zeta_n) \ge \frac{1}{2} \cdot \inf_{h\in\mathbb{H}_t,\ \|h-f_0\|_2<\zeta_n} \|h\|_{\mathbb{H}_t}^2.$$

We prove in Sect. 9, see Theorem 4, that there exist constants c, C and f_0 in $B_{2,\infty}^s(\mathcal{M})$ such that

$$\varphi_{f_0}^t(\zeta_n) \ge C\zeta_n^2 e^{c\zeta_n^{-2/s}t}. \tag{32}$$

Now, under (32) for the previous fixed function f_0, taking $\zeta_n = \kappa\varepsilon_n$ for small (but fixed) enough κ, it holds, when t belongs to $\mathcal{T}_2$,

$$\varphi_{f_0}^t(\zeta_n) \ge C(\kappa\varepsilon_n)^a e^{c\kappa^{-2/s}B\log(1/\zeta_n)}.$$

For κ small enough, this is larger that any given power of ε_n. In particular, it is larger than $(8+C)n\varepsilon_n^2$ if the (upper-bound) rate ε_n is no more than polynomial in n, which is the case here since $\varepsilon_n = (\log n/n)^{s/(2s+p)}$. We have verified that (31) is satisfied, which gives the desired lower bound result when $q \le 1 + d/2$ using Lemma 1.

In the case that $q > 1 + d/2$, the proof is similar, except that the exponent of the logarithmic factor in (26) has now the power $q - d/2$, due to the assumption on the prior density g, and that ε_n is now replaced by $\widetilde{\varepsilon}_n = (\log n)^{q-1-\frac{d}{2}}\varepsilon_n$. □

9 Proofs II: Besov spaces and needlets

In this section we start by introducing some standard notation useful in the context of Besov spaces, mainly the concepts of Littlewood–Paley function and decomposition into low-frequency subspaces. Let L be the operator whose properties are listed in Sect. 2.4.

A *Littlewood–Paley function* is any even function Φ in $\mathcal{D}(\mathbb{R})$ with

$$0 \le \Phi, \quad 1 = \Phi(x) \text{ for } |x| \le 1/2, \quad supp(\Phi) \subset [-1, 1]. \tag{33}$$

Given a Littlewood–Paley function, let us also define

$$\Psi(x) = \Phi(\frac{x}{2}) - \Phi(x).$$

From this it follows that $0 \leq \Psi(x) \leq 1$, that the support of Ψ is included in $\{\frac{1}{2} \leq |x| \leq 2\}$ and that

$$\forall\, \delta > 0, \quad 1 \equiv \Phi(\delta x) + \sum_{j \geq 0} \Psi(2^{-j}\delta x).$$

For any even Θ in $\mathcal{D}(\mathbb{R})$ (in the sequel we apply (34) below for $\Theta = \Phi$, $\Theta = \Psi$ or rescaled versions of them) and $0 < \delta \leq 1$, define the kernel operator $\Theta(\delta\sqrt{L})$ using the spectral decomposion of L, by setting

$$\Theta(\delta\sqrt{L})(x, y) = \sum_k \Theta(\delta\sqrt{\lambda_k})\, Q_k(x, y). \tag{34}$$

So any square-integrable function f on $\mathcal{M}$ can be expanded $f = \Phi(\delta\sqrt{L})f + \sum_{j \geq 0} \Psi(2^{-j}\delta\sqrt{L})f$, where

$$\Phi(\delta\sqrt{L})f(x) = \int_M \Phi(\delta\sqrt{L})(x, y) f(y) d\mu(y),$$
$$\Psi(\delta 2^{-j}\sqrt{L})f(x) = \int_M \Psi(\delta 2^{-j}\sqrt{L})(x, y) f(y) d\mu(y).$$

Moreover, if we define the 'low frequency' functions using the eigenspaces $\mathcal{H}_{\lambda_k}$ of the operator L by

$$\Sigma_t = \bigoplus_{\lambda \leq \sqrt{t}} \mathcal{H}_\lambda, \tag{35}$$

for any $t > 0$, we have from the definitions of Φ and Ψ that

$$\Phi(\delta\sqrt{L})f \in \Sigma_{\frac{1}{\delta}}, \quad \Psi(2^{-j}\delta\sqrt{L})f \in \Sigma_{\frac{2j+1}{\delta}} \cap [\Sigma_{\frac{2j+1}{\delta}}]^{\perp}.$$

Also recall that an $\varepsilon-$ *net* $\Lambda \subset \mathcal{M}$ is a set such that $\xi \neq \xi',\ \xi, \xi' \in \Lambda$ implies $\rho(\xi, \xi') > \varepsilon$. A *maximal* $\varepsilon-$*net* Λ, is a an $\varepsilon-$net such that for all $x \in X \setminus \Lambda,\ \Lambda \cup \{x\}$ is no more an $\varepsilon-$net.

9.1 Definition of Besov spaces

We follow [9] to introduce the Besov spaces B^s_{pq} in this setting with $s > 0, 1 \le p \le \infty$ and $0 < q \le \infty$. To do so, let us choose any Littlewood–Paley function Φ as in (33) and let $\Phi_j(\lambda) := \Phi(2^{-j}\lambda)$ for $j \ge 1$. Again, L is the operator from Sect. 2.4.

Definition 1 Let $s > 0$, $1 \le p \le \infty$, and $0 < q \le \infty$. The Besov space $B^s_{pq} = B^s_{pq}(\mathcal{M}) = B^s_{pq}(\mathcal{M}, L)$ is defined as the set of all $f \in \mathbb{L}^p(\mathcal{M}, \mu)$ such that

$$\|f\|_{B^s_{pq}} := \left(\sum_{j \ge 0} \left(2^{sj} \|\Phi_j(\sqrt{L}) f(\cdot) - f(\cdot)\|_{\mathbb{L}_p} \right)^q \right)^{1/q} < \infty. \tag{36}$$

Here the ℓ^q-norm is replaced by the sup-norm if $q = \infty$.

Remark 3 One can prove, as a consequence of the Gaussian estimate (10), see [9], that this definition is independent of the choice of Φ and that the Besov spaces can also be introduced via the following approximation properties: If $\mathbb{E}_t(f)_p$ denotes the best approximation of $f \in \mathbb{L}^p$ from Σ_t, see (35), i.e.

$$\mathbb{E}_t(f)_p := \inf_{g \in \Sigma_t} \|f - g\|_p,$$

(where, here $\mathbb{L}^\infty$ is identified as the space UCB of all uniformly continuous and bounded functions on M) then it is proved in [9] that

$$B^s_{pq} := \left\{ f \in \mathbb{L}_p, \quad \|f\|_{A^s_{pq}} := \|f\|_p + \left(\sum_{j \ge 0} \left(2^{sj} \mathbb{E}_{2^j}(f)_p \right)^q \right)^{1/q} < \infty \right\}.$$

9.2 Smooth functional calculus and 'sampling-father-wavelets'

In addition to the orthogonal analysis provided by the projectors $P_{\mathcal{H}_{\lambda_k}}$ onto eigenspaces of the operator L, one can build, following [9], a wavelet-type analysis on $\mathcal{M}$ associated to L. The properties of the operator L given in Sect. 2.4 have the following important consequences, see [9],

Localisation [9, Section 3] For any even Θ in $\mathcal{D}(\mathbb{R})$, there exists a constant $C(\Theta)$ such that

$$\text{for all } 0 < \delta \le 1, \forall\, x, y \in \mathcal{M}, \ |\Theta(\delta\sqrt{L})(x, y)| \le \frac{1}{|B(x, \delta)|} \frac{C(\Theta)}{(1 + \frac{\rho(x,y)}{\delta})^{D+1}}. \tag{37}$$

From (37) one can easily deduce the symmetrical bound $|\Theta(\delta\sqrt{L})(x, y)| \le \frac{1}{\sqrt{|B(x,\delta)||B(y,\delta)|}} \frac{C(\Theta)}{(1+\frac{\rho(x,y)}{\delta})^{D+1}}$.

Father wavelet One can deduce from [9] (Lemmas 5.2 and 5.4) that there exist $0 < C_0 < \infty$, $0 < \gamma$ structural constants such that for any $0 < \delta \leq 1$, for any $\Lambda_{\gamma\delta}$ maximal $\gamma\delta-$net, there exists a family of functions : $(D_\xi^\delta)_{\xi\in\Lambda_{\gamma\delta}}$ such that

$$|D_\xi^\delta(x)| \leq \frac{1}{|B(\xi,\delta)|}\frac{C_0}{(1+\frac{\rho(x,\xi)}{\delta})^{D+1}}, \quad \forall\, x \in \mathcal{M} \tag{38}$$

we have the following wavelet-type representation:

$$\forall\, \varphi \in \Sigma_{1/\delta}, \quad \varphi(x) = \sum_{\xi\in\Lambda_{\gamma\delta}} \varphi(\xi)|B(\xi,\delta)|D_\xi^\delta(x), \tag{39}$$

$$\forall\, (\alpha_\xi)_{\xi\in\Lambda_{\gamma\delta}}, \left\| \sum_{\xi\in\Lambda_{\gamma\delta}} \alpha_\xi |B(\xi,\delta)| D_\xi^\delta(x) \right\|_\infty \lesssim \sup_{\xi\in\Lambda_{\gamma\delta}} |\alpha_\xi|. \tag{40}$$

We see on the formulae (39) and (40) that the functions $|B(\xi,\delta)|D_\xi^\delta$ behave like father-wavelets, with coefficients directly obtained by sampling. We will see in Sect. 10 that these functions play an important role for instance to bound the entropy of various functional spaces.

10 Proofs II: Entropy properties

10.1 Covering number, entropy, ε-net

Let Λ be a maximal $\varepsilon-$net over a metric space (X,ρ). We have :

$$X \subset \cup_{\xi\in\Lambda} B(\xi,\varepsilon),\ \xi \neq \xi',\ \xi,\xi' \in \Lambda \Rightarrow B(\xi,\varepsilon/2) \cap B(\xi',\varepsilon/2) = \emptyset.$$

Hence, for Λ_ε a maximal $\varepsilon-$net it holds

$$N(\varepsilon/2, X) \geq card(\Lambda_\varepsilon) \geq N(\varepsilon, X).$$

Now if (X,ρ) is a doubling metric space then we have the following property : If $x_1,\ldots,x_N \in B(x,r)$ are such that, $\rho(x_i,x_j) > r2^{-l}$ $(l \in \mathbb{N})$ clearly $B(x,r) \subset B(x_i,2r) = B(x_i, 2^{l+2}(r2^{-l-1}))$ and the balls $B(x_i, r2^{-l-1})$ are disjoint and contained in $B(x,2r)$. so:

$$N2^{-(l+2)D}|B(x,r)| \leq \sum_{i=1}^{N} |B(x_i, r2^{-l-1})| \leq B(x,2r)| \leq 2^D|B(x,r)| \tag{41}$$

If $\Lambda_{r2^{-l}}$ is any $r2^{-l}-$net then: $Card(\Lambda_{r2^{-l}}) \leq 2^{(l+3)d}N(X,r)$.

So if Λ_ε is any maximal $\varepsilon-$net and for $l \in \mathbb{N}$, $\Lambda_{2^l\varepsilon}$ is any maximal $2^l\varepsilon-$net then :

$$N(X, \varepsilon 2^l) \leq N(X, \varepsilon) \leq Card(\Lambda_\varepsilon) \leq 2^{(l+3)D} N(X, 2^l\varepsilon) \leq 2^{(l+3)D} Card(\Lambda_{2^l\varepsilon}). \tag{42}$$

For $l = 0$

$$2^{-3D} Card(\Lambda_\varepsilon) \leq N(X, \varepsilon) \leq Card(\Lambda_\varepsilon).$$

So for any $\varepsilon > 0$, and for any maximal $\varepsilon-$net Λ_ε, $Card(\Lambda_\varepsilon)$ and $N(X, \varepsilon)$ are of the same order.
Moreover clearly, taking $r = 1$ in (41), so that $B(x, 1) = \mathcal{M}$, we get:

$$N(\delta, \mathcal{M}) \leq 4^D \left(\frac{1}{\delta}\right)^D \tag{43}$$

10.2 Dimension of spectral spaces, covering number, and trace of P_t

Let us now use the heat kernel' assumptions. The following proposition gives the link between the covering number $N(\delta, \mathcal{M})$ of the underlying space $\mathcal{M}$, the behavior of the trace of e^{-tL} and the dimension of the spectral spaces. Let us recall:

$$\Sigma_\lambda = \oplus_{\sqrt{\lambda_k} \leq \lambda} \mathcal{H}_{\lambda_k}.$$

Denote by P_{Σ_λ} the orthogonal projector onto this space and also, with a slight abuse of notation, the associated kernel

$$P_{\Sigma_\lambda}(x, y) = \sum_{\sqrt{\lambda_k} \leq \lambda} Q_k(x, y).$$

Then one can prove the following bounds (see [9], Lemma 3.19): For any $\lambda \geq 1$, and $\delta = \frac{1}{\lambda}$,

$$\exists C_2, C_2', \quad \text{such that} \quad \frac{C_2'}{|B(x, \delta)|} \leq P_{\Sigma_\lambda}(x, x) \leq \frac{C_2}{|B(x, \delta)|} \tag{44}$$

Let us recall that $Tr(e^{-tL}) = \sum_k e^{-\lambda_k t} dim(\mathcal{H}_{\lambda_k})$. In addition we have $\int_{\mathcal{M}} P_t(x, x)$ $d\mu(x) = Tr(e^{-tL})$. Moreover, as

$$P_t(x, x) = \int_{\mathcal{M}} P_{t/2}(x, u) P_{t/2}(u, x) d\mu(u) = \int_{\mathcal{M}} (P_{t/2}(x, u))^2 d\mu(u)$$

we have, if $\| \ \|_{HS}$ stands for the Hilbert-Schmidt norm,

$$Tr(e^{-tL}) = \int_{\mathcal{M}} P_t(x, x) d\mu(x) = \int_{\mathcal{M}} \int_{\mathcal{M}} (P_{t/2}(x, u))^2 d\mu(u) d\mu(x) = \|e^{-\frac{t}{2}L}\|_{HS}^2.$$

Proposition 4 1. *For* $\lambda \geq 1, \quad \delta = \frac{1}{\lambda}$,

$$C_2' \int_{\mathcal{M}} \frac{1}{|B(x,\delta)|} d\mu(x) \leq \dim(\Sigma_\lambda) = \int_{\mathcal{M}} P_{\Sigma_\lambda}(x,x) d\mu(x) \leq C_2 \int_{\mathcal{M}} \frac{1}{|B(x,\delta)|} d\mu(x) \tag{45}$$

2.

$$\begin{aligned} 2^{-2D} N(\delta, \mathcal{M}) \leq 2^{-2D} card(\Lambda_\delta) \leq \int_{\mathcal{M}} \frac{1}{|B(x,\delta)|} d\mu(x) \leq 2^D card(\Lambda_\delta) \\ \leq 2^{4D} N(\delta, \mathcal{M}) \end{aligned} \tag{46}$$

where Λ_δ *is any* $\delta-$*maximal net.*

3.

$$C_1' \int_{\mathcal{M}} \frac{1}{|B(x,\sqrt{t})|} d\mu(x) \leq Tr(e^{-tL}) \leq C_1 \int_{\mathcal{M}} \frac{1}{|B(x,\sqrt{t})|} d\mu(x)$$

Proof of the Proposition Point 1. is a consequence of (44) while 3. is a consequence of (12). Let us now prove 2. Let Λ_δ be any $\delta-$maximal net.

$$\sum_{\xi \in \Lambda_\delta} \int_{B(\xi,\delta/2)} \frac{1}{|B(x,\delta)|} d\mu(x) \leq \int_M \frac{1}{|B(x,\delta)|} d\mu(x) \leq \sum_{\xi \in \Lambda_\delta} \int_{B(\xi,\delta)} \frac{1}{|B(x,\delta)|} d\mu(x)$$

But:

$$x \in B(\xi, \delta/2) \Longrightarrow B(x,\delta) \subset B(\xi, 2\delta), \quad \text{so} \quad \frac{1}{|B(x,\delta)|} \geq \frac{2^{-2D}}{|B(\xi,\delta/2)|}$$

and in the same way:

$$x \in B(\xi, \delta) \Longrightarrow B(\xi,\delta) \subset B(x, 2\delta), \quad \text{so} \quad \frac{1}{|B(x,\delta)|} \leq \frac{2^{D}}{|B(\xi,\delta)|}$$

This implies:

$$2^{-2D} card(\Lambda_\delta) \leq \int_{\mathcal{M}} \frac{1}{|B(x,\delta)|} d\mu(x) \leq 2^D card(\Lambda_\delta).$$

□

The former results can be summarised in the following corollary:

Corollary 1

$$Trace(e^{-\delta^2 L}) \sim dim(\Sigma_\lambda) \sim N(\delta, \mathcal{M}); \quad \delta = \frac{1}{\lambda}$$

10.3 Connection between the covering number of $\mathcal{M}$ and the entropy of $\mathbb{H}_t^1$

In this section we establish the link between the covering number $N(\varepsilon, \mathcal{M})$ of the space $\mathcal{M}$, and $H(\varepsilon, \mathbb{H}_t^1, \mathbb{L}^p)$ for $p = 2,\ \infty$ stated in Theorem 3.

Notice, of course, that, using the previous section, one can replace $N(\delta(t, \varepsilon), \mathcal{M})$ at any place by $card(\Lambda_{\delta(t,\varepsilon)})$, where $\Lambda_{\delta(t,\varepsilon)}$ is a maximal $\delta(t, \varepsilon)-$net. Also, since $\mu(\mathcal{M}) = 1$, we have

$$H(\varepsilon, \mathbb{H}_t^1, \mathbb{L}^2) \leq H(\varepsilon, \mathbb{H}_t^1, \mathbb{L}^\infty).$$

So the proof will be done in two steps:
1-We prove the lower bound for $H(\varepsilon, \mathbb{H}_t^1, \mathbb{L}^2)$ in the next subsection, using Carl's inequality.
2-We prove next the upper bound for $H(\varepsilon, \mathbb{H}_t^1, \mathbb{L}^\infty)$.

10.3.1 Proof of Theorem 3: Lower estimates for $H(\varepsilon, \mathbb{H}_t^1, \mathbb{L}^2)$

Let us recall some classical facts: see the following references [3,4].

For any subset X of a metric space, we define, for any $k \in \mathbb{N}$,

$$e_k(X) = \inf\{\varepsilon \geq 0,\ \exists\, 2^k \quad \text{balls of radius } \ \varepsilon,\ \text{covering } X.\}.$$

Clearly

$$\varepsilon < e_k(X) \Longrightarrow H(\varepsilon, X) > k$$

Now for the special case of a compact positive selfadjoint operator $T : \mathbb{H} \mapsto \mathbb{H}$ we have the following Carl inequality (see [3]) relating $e_k(T(B))$ where B is the unit ball of $\mathbb{H}$ and the eigenvalues $0 \leq \mu_1 \leq \mu_2, \ldots$ (possibly repeated with their multiplicity order) of T:

$$\text{for all } k \in \mathbb{N}^*,\ n \in \mathbb{N}^*, \quad e_k(T(B)) \geq 2^{-\frac{k}{2n}} \prod_{i=1}^{n} \mu_i^{1/n}. \tag{47}$$

In our case, let us take: $T = P_{t/2},\ \mu_i = e^{-(t/2)\lambda_i}$, $T(B) = \mathbb{H}_t^1$. Let us fix:

$$\lambda = \sqrt{\frac{1}{t} \log \frac{1}{\varepsilon}} = \frac{1}{\delta} = \frac{1}{\delta(t, \varepsilon)}.$$
$$n = dim(\Sigma_\lambda); \quad k \sim n \log \frac{1}{\varepsilon} \frac{1}{\log 2}.$$

Carl's inequality gives:

$$e_k \geq 2^{-\frac{k}{2n}} e^{-\frac{1}{n}\sum_{t\lambda_i \leq \log\frac{1}{\varepsilon}}(t/2)\lambda_i} \geq \varepsilon.$$

So

$$H(\varepsilon, \mathbb{H}_t^1, \mathbb{L}^2) \geq k \sim n \log\frac{1}{\varepsilon}\frac{1}{\log 2} \sim dim(\Sigma_\lambda)\log\frac{1}{\varepsilon},$$

but by Corollary 1, it holds $dim(\Sigma_\lambda) \sim N(\delta, \mathcal{M})$, with $\delta = \frac{1}{\lambda}$. So,

$$H(\varepsilon, \mathbb{H}_t^1, \mathbb{L}^2) \gtrsim \log\frac{1}{\varepsilon} N(\delta, \mathcal{M}), \quad \frac{1}{\delta} = \lambda = \sqrt{\frac{1}{t}\log\frac{1}{\varepsilon}}.$$

10.3.2 Proof of Theorem 3: Upper estimate for $H(\varepsilon, H_t^1, \mathbb{L}_\infty)$

Recall the notation introduced in Sect. 9 (especially 9.2). Let us suppose : $\varepsilon^\nu \leq at$, $\nu > 0, a > 0$.

First, we prove that for all $\varepsilon > 0$, small enough, there exists δ $(\sim \delta(t,\varepsilon) := \sqrt{\frac{1}{t}\log\frac{1}{\varepsilon}})$ such that

$$\text{for all } f \in \mathbb{H}_t^1, \quad \|\Phi(\delta\sqrt{L})f - f\|_\infty \leq \frac{\varepsilon}{2}.$$

In a second step, we use (39) to expand on the $|B(\xi,\delta)|D_\xi^\delta$'s:

$$\Phi(\delta\sqrt{L})f(x) = \sum_{\xi\in\Lambda_{\gamma\delta}} \Phi(\delta\sqrt{L})f(\xi)|B(\xi,\delta)|D_\xi^\delta(x).$$

In a third step, we use a family of points of $\Sigma_{\frac{1}{\delta}}$ as centers of balls of radius $\varepsilon/2$ covering $\Phi(\delta\sqrt{L})(\mathbb{H}_t^1)$ so that the balls centered in these points is an $\varepsilon-$ covering in $\mathbb{L}^\infty$ norm of $\mathbb{H}_t^1$.

The next lemma gives evaluations of $\|\Phi(\delta\sqrt{L})f - f\|_\infty$ and $\|\Phi(\delta\sqrt{L})(\mathbb{H}_t^1)\|_\infty$.

Lemma 2 *for all $f \in \mathbb{H}_t^1$*

1.

$$\|\Phi(\delta\sqrt{L})f\|_\infty \lesssim \frac{1}{t^{D/4}}$$

2.

$$\|\Psi(\delta\sqrt{L})f\|_\infty \lesssim e^{-\frac{t}{8\delta^2}}\frac{1}{\delta^{D/2}}$$

3.

$$\|\Phi(\delta\sqrt{L})f - f\|_\infty \le \sum_{j\ge 0} \|\Psi(2^{-j}\delta\sqrt{L})f\|_\infty \lesssim \frac{1}{\delta^{D/2}} e^{-\frac{A}{4}} A^{-1}, \quad A = \frac{t}{8\delta^2}$$

Proof of the Lemma First, $f \in \mathbb{H}_t^1$ so $f = \sum_k \sum_l a_k^l e_k^l(.) e^{-\lambda_k t/2}$, $\sum_k \sum_l |a_k^l|^2 \le 1$. As $\Phi(\delta\sqrt{L})(x,y) = \sum_k \Phi(\delta\sqrt{\lambda_k}) P_k(x,y)$,

$$\Phi(\delta\sqrt{L})f(x) = \langle \Phi(\delta\sqrt{L})(x,.), f(.)\rangle = \sum_k \sum_l \Phi(\delta\sqrt{\lambda_k}) a_k^l e_k^l(x) e^{-\lambda_k t/2}, \text{ hence}$$

$$\begin{aligned}
|\Phi(\delta\sqrt{L})f(x)| &\le \left(\sum_k \sum_l |a_k^l|^2\right)^{1/2} \left(\sum_k e^{-\lambda_k t} \Phi^2(\delta\sqrt{\lambda_k}\right) \sum_l (e_k^l(x)^2)^{1/2} \\
&\le \left(\sum_k e^{-\lambda_k t} \Phi^2(\delta\sqrt{\lambda_k}) P_k(x,x)\right)^{1/2} \\
&\le \left(\sum_k \Phi^2(\delta\sqrt{\lambda_k}) P_k(x,x)\right)^{1/2} \wedge \left(\sum_k e^{-\lambda_k t} P_k(x,x)\right)^{1/2} \\
&= [\Phi^2(\delta\sqrt{L})(x,x) \wedge P_t(x,x)]^{1/2} \\
&\le \frac{\sqrt{C(\Phi^2)}}{\sqrt{|B(x,\delta)|}} \wedge \frac{\sqrt{C_1}}{\sqrt{|B(x,\sqrt{t})|}} \lesssim \frac{1}{t^{D/4}}
\end{aligned}$$

using (12), (37) and the lower bound $|B(x,r)| \ge (r/2)^D$ obtained in Sect. 2.1. In the same way,

$$\Psi(\delta\sqrt{L})f(x) = \langle \Psi(\delta\sqrt{L})(x,.), f(.)\rangle = \sum_k \sum_l \Psi(\delta\sqrt{\lambda_k}) a_k^l e_k^l(x) e^{-\lambda_k t/2}, \text{ hence}$$

$$\begin{aligned}
|\Psi(\delta\sqrt{L})f(x)| &\le \left(\sum_k \sum_l |a_k^l|^2\right)^{1/2} \left(\sum_k e^{-\lambda_k t}\ \Psi^2(\delta\sqrt{\lambda_k}) \sum_l (e_k^l(\xi))^2\right)^{1/2} \\
&\le e^{-\frac{1}{4\delta^2}t/2} \left(\sum_k \Psi^2(\delta\sqrt{\lambda_k}) P_k(x,x)\right)^{1/2} \\
&= e^{-\frac{t}{8\delta^2}} [\Psi^2(\delta\sqrt{L})(x,x)]^{1/2} \\
&\le C(\Psi^2) e^{-\frac{t}{8\delta^2}} \frac{1}{|B(x,\delta)|^{1/2}} \\
&\le 2^{D/2} C(\Psi^2) e^{-\frac{t}{8\delta^2}} \frac{1}{\delta^{D/2}} \\
&\lesssim e^{-\frac{t}{8\delta^2}} \frac{1}{\delta^{D/2}}.
\end{aligned}$$

So, we have

$$\sum_{j\geq 0}\|\Psi(2^{-j}\delta\sqrt{L})f\|_\infty \lesssim \frac{1}{\delta^{D/2}}\sum_{j\geq 0} e^{-2^{2j}\frac{t}{8\delta^2}}2^{jD/2}.$$

Put $A=\frac{t}{8\delta^2}$; as:

$$\int_{2^j}^{2^{j+1}} x^{D/2}e^{-\frac{A}{4}x^2}\frac{Dx}{x} \geq 2^{\frac{D}{2}j}e^{-A2^{2j}}\log 2$$

$$\sum_{j=0}^{\infty}2^{\frac{D}{2}j}e^{-A2^{2j}} \leq \frac{1}{\log 2}\int_1^\infty x^{D/2}e^{-\frac{A}{4}x^2}\frac{Dx}{x} = \frac{1}{\log 2}\frac{1}{2}(\frac{4}{A})^{D/4}\int_{A/4}^\infty u^{D/4}e^{-u}\frac{Du}{u}$$

as

$$\text{for all } a\in\mathbb{R}, X>0, \quad \int_X^\infty t^{a-1}e^{-t}dt \leq 2e^{-X}X^{a-1}, \quad \text{if } X\geq 2(a-1)$$

$$\sum_{j=0}^{\infty}2^{\frac{D}{2}j}e^{-A2^{2j}} \leq \frac{4}{\log 2}e^{-\frac{A}{4}}A^{-1}, \quad \text{if } A\geq 8(D-2)$$

Conclude that

$$\sum_{j\geq 0}\|\Psi(2^{-j}\delta\sqrt{L})f\|_\infty \leq Ct^{-D/4}(\frac{A}{4})^{D/4}e^{-\frac{A}{4}}(\frac{A}{4})^{-1}. \quad \square$$

$\square$

First step Fix δ such that $\|f-\Phi(\delta\sqrt{L})f\|_\infty < \frac{\varepsilon}{2}$

Using the previous lemma, we need to choose δ so that

$$\frac{\varepsilon}{2} > Ct^{-D/4}\left(\frac{t}{32\delta^2}\right)^{D/4}e^{-\frac{A}{4}}\left(\frac{A}{4}\right)^{-1}, \quad \frac{A}{4}=\frac{t}{32\delta^2}. \tag{48}$$

Let us take

$$\frac{A}{4}=\frac{t}{32\delta^2}=\alpha\log\frac{1}{\varepsilon}.$$

Then, as $\varepsilon^{\nu} \leq at$,

$$Ct^{-D/4}\left(\frac{A}{4}\right)^{D/4} e^{-\frac{A}{4}}\left(\frac{A}{4}\right)^{-1} = Ct^{-D/4}\left(\alpha \log\frac{1}{\varepsilon}\right)^{D/4-1}\varepsilon^{\alpha}$$
$$\leq C(a\varepsilon^{\nu})^{-D/4}\left(\alpha \log\frac{1}{\varepsilon}\right)^{D/4-1}\varepsilon^{\alpha} \leq \frac{\varepsilon}{2}$$

if α is suitably chosen. So for $\frac{1}{\delta} \sim \sqrt{\frac{1}{t}\log\frac{1}{\varepsilon}}$,

$$\|f - \Phi(\delta\sqrt{L})f\|_{\infty} < \frac{\varepsilon}{2}$$

Second step $\varepsilon-$ covering of $\mathbb{H}_t^1$.

Now if $f \in \mathbb{H}_t^1$, using Lemma 2, $\|\Phi(\delta\sqrt{L})f\|_{\infty} \lesssim t^{-D/4}$. Moreover $\Phi(\delta\sqrt{L})f \in \Sigma_{1/\delta}$, so, using (39)

$$\Phi(\delta\sqrt{L})f(x) = \sum_{\xi\in\Lambda_{\gamma\delta}} \Phi(\delta\sqrt{L})f(\xi)|B(\xi,\delta)|D_{\xi}^{\delta}(x).$$

Let us consider the following family :

$$f_{(k.)} = C\sum_{\xi\in\Lambda_{\gamma\delta}} k_{\xi}\varepsilon|B(\xi,\delta)|D_{\xi}^{\delta}(x), \quad k_{\xi} \in \mathbb{N}, |k_{\xi}| \leq K \in \mathbb{N}, \quad KC\varepsilon \leq \frac{1}{t^{D/4}}.$$

Certainly for all $f \in \mathbb{H}_t^1$, there exists (k_{ξ}) in the previous family such that

$$\|\Phi(\delta\sqrt{L})f - \sum_{\xi\in\Lambda_{\gamma\delta}} k_{\xi}C\frac{\varepsilon}{2}|B(\xi,\delta)|D_{\xi}^{\delta}(x)\|_{\infty}$$
$$= \|\sum_{\xi\in\Lambda_{\gamma\delta}} (\Phi(\delta\sqrt{L})f(\xi) - Ck_{\xi}\varepsilon)|B(\xi,\delta)|D_{\xi}^{\delta}(x)\|_{\infty}$$
$$\lesssim \sup_{\xi\in\Lambda_{\gamma\delta}} |\Phi(\delta\sqrt{L})f(\xi) - Ck_{\xi}\varepsilon| < \frac{\varepsilon}{2}.$$

As $\|\Phi(\delta\sqrt{L})f - f\|_{\infty} \leq \frac{\varepsilon}{2}$, one can cover $\mathbb{H}_t^1$ by balls centered in the $f_{(k.)}$ of radius ε.

The cardinality of this family of balls is: $(2K+1)^{card(\Lambda_{\gamma\delta})}$. As γ is a structural constant, $\varepsilon^{\nu} \leq at$ and $\delta \sim \delta(t,\varepsilon)$, clearly

$$H(\varepsilon, \mathbb{H}_t^1, \mathbb{L}^{\infty}) \lesssim \mathcal{N}(\delta(t,\varepsilon), \mathcal{M}) \cdot \log\frac{1}{\varepsilon}.$$

□

10.4 Bounds for $\mathbb{E}(\|W^t\|_{\mathbb{B}}^2)$

In the two remaining subsections, we prove some useful results used in Sect. 6 and the proof of Theorem 2 respectively. The next Proposition provides a control on the expectation of the squared sup-norm of W^t. Similar bounds in the $\mathbb{L}^2$ norm are obtained along the way (even slightly more precise).

Proposition 5 *There exist universal constants C_1 and C_2 such that*

$$C_1 \sup_{x\in\mathcal{M}} \frac{1}{|B(x,\sqrt{t})|} \leq \mathbb{E}\|W^t\|_\infty^2 \leq C_2 N(\sqrt{t},\mathcal{M}) \sup_{x\in\mathcal{M}} \frac{1}{|B(x,\sqrt{t})|}. \tag{49}$$

We recall that W^t writes

$$W^t(x) = \sum_k \sum_{1\leq l\leq \dim \mathcal{H}_{\lambda_k}} e^{-\lambda_k t/2} X_k^l e_k^l(x)$$

where X_k^l is a family of independent $N(0,1)$ Gaussian variables. Clearly since $\mathcal{M}$ is supposed to have measure 1,

$$\mathbb{E}(\|W^t\|_2^2) \leq \mathbb{E}(\|W^t\|_\infty^2).$$

As $\|W^t\|_2^2 = \sum_k e^{-\lambda_k t} \sum_{1\leq l\leq \dim \mathcal{H}_{\lambda_k}} (X_k^l)^2$, we get

$$\mathbb{E}(\|W^t\|_2^2) = \sum_k e^{-\lambda_k t} \dim \mathcal{H}_{\lambda_k} = Trace(e^{-tL}) = \int_{\mathcal{M}} P_t(u,u)d\mu(u).$$

Hence using Proposition 4, one obtains

$$\int_{\mathcal{M}} \frac{1}{|B(x,\sqrt{t})|} d\mu(x) \lesssim \mathbb{E}(\|W^t\|_2^2) = \int_M P_t(u,u)d\mu(u) \lesssim \int_{\mathcal{M}} \frac{1}{|B(x,\sqrt{t})|} d\mu(x). \tag{50}$$

Now, let us first observe, using again Proposition 4, that

$$\begin{aligned}
\mathbb{E}(\|W^t\|_\infty^2) &= \mathbb{E}(\sup_{x\in\mathcal{M}} |W^t(x)|^2) \\
&\geq \sup_{x\in\mathcal{M}} \mathbb{E}(|W^t(x)|^2) \\
&= \sup_{x\in\mathcal{M}} \mathbb{E}\left(\left|\sum_k \sum_{1\leq l\leq \dim \mathcal{H}_{\lambda_k}} e^{-\lambda_k t/2} X_k^l e_k^l(x)\right|^2\right) \\
&= \sup_{x\in\mathcal{M}} \sum_k \sum_{1\leq l\leq \dim \mathcal{H}_{\lambda_k}} e^{-\lambda_k t} (e_k^l(x))^2 \\
&= \sup_{x\in\mathcal{M}} \sum_k e^{-\lambda_k t} P_k(x,x) \\
&= \sup_{x\in\mathcal{M}} P_t(x,x) \sim \sup_{x\in\mathcal{M}} \frac{1}{|B(x,\sqrt{t})|}
\end{aligned}$$

On the other side, using Cauchy-Schwarz inequality,

$$\begin{aligned}
|W^t(x)|^2 &= |\sum_k \sum_{1\leq l\leq \dim \mathcal{H}_{\lambda_k}} e^{-\lambda_k t/2} X_k^l e_k^l(x)|^2 \\
&\leq \left\{\sum_k \sum_{1\leq l\leq \dim \mathcal{H}_{\lambda_k}} e^{-\lambda_k t/2} (X_k^l)^2\right\} \left\{\sum_k \sum_{1\leq l\leq \dim \mathcal{H}_{\lambda_k}} e^{-\lambda_k t/2} (e_k^l(x))^2\right\} \\
&= \left\{\sum_k \sum_{1\leq l\leq \dim \mathcal{H}_{\lambda_k}} e^{-\lambda_k t/2} (X_k^l)^2\right\} \left\{\sum_k e^{-\lambda_k t/2} P_k(x,x)\right\} \\
&= \left\{\sum_k \sum_{1\leq l\leq \dim \mathcal{H}_{\lambda_k}} e^{-\lambda_k t/2} (X_k^l)^2\right\} P_{t/2}(x,x).
\end{aligned}$$

So

$$\begin{aligned}
\mathbb{E}(\|W^t\|_\infty^2) &\leq \mathbb{E}\left\{\sum_k \sum_{1\leq l\leq \dim \mathcal{H}_{\lambda_k}} e^{-\lambda_k t/2} (X_k^l)^2\right\} . \sup_{x\in\mathcal{M}} P_{t/2}(x,x) \\
&= Trace(e^{-t/2L}) \sup_{x\in\mathcal{M}} P_{t/2}(x,x) \\
&= \left(\int_{\mathcal{M}} P_{t/2}(u,u) d\mu(u)\right) \left(\sup_{x\in\mathcal{M}} P_{t/2}(x,x)\right).
\end{aligned}$$

Hence, we get

$$\sup_{x\in\mathcal{M}} \frac{1}{|B(x,\sqrt{t})|} \sim \sup_{x\in\mathcal{M}} P_t(x,x) \leq \mathbb{E}(\|W^t\|_\infty^2)$$
$$\leq \left(\int_{\mathcal{M}} P_{t/2}(u,u)d\mu(u)\right)\left(\sup_{x\in\mathcal{M}} P_{t/2}(x,x)\right)$$
$$\sim \mathcal{N}(\sqrt{t},\mathcal{M}) \sup_{x\in\mathcal{M}} \frac{1}{|B(x,\sqrt{t})|}.$$

And we have in addition: $N(\sqrt{t},\mathcal{M}) \sim \int_{\mathcal{M}} \frac{1}{|B(x,\sqrt{t})|}d\mu(x) \ll \sup_{x\in\mathcal{M}} \frac{1}{|B(x,\sqrt{t})|}$.

10.5 Lower bound for $A_f^t(\varepsilon)$

Theorem 4 *For $s > 0$ fixed, there exists $f \in B^s_{2,\infty}(\mathcal{M})$, (the unit ball of the Besov space) with $\|f\|_2^2 = 1$ and constants $c > 0, C > 0$ such that:*

$$\textit{for all } 1 \geq t > 0, \ \textit{ for all } 1 > \varepsilon > 0, \quad \inf_{\|f-h\|_2\leq\varepsilon} \|h\|^2_{\mathbb{H}_t} \geq C\varepsilon^2 e^{ct\varepsilon^{-2/s}}$$

Let us take f such that

$$\|f\|_2 = 1 > \varepsilon > 0.$$

We are interested in:

$$\inf_{\|f-P_{t/2}g\|_2=\varepsilon} \|g\|_2^2.$$

Let us put

$$\Phi(g) = \|f - P_{t/2}g\|_2^2 = \|f\|_2^2 - 2\langle P_{t/2}f, g\rangle + \langle P_t g, g\rangle = \varepsilon^2, \ \Psi(g) = \|g\|_2^2. \quad (51)$$

We have,

$$D\Phi(g) = -2P_{t/2}f + 2P_t(g), \quad D\Psi(g) = 2g$$

$$\text{So, } \inf_{\Phi(g)=\varepsilon^2} \Psi(g) = \Psi(g_0) \Longrightarrow D\Psi(g_0) = -\mu D\Phi(g_0)$$

$$\text{with } g_0 = -\mu P_t(g_0) + \mu P_{t/2}f.$$

Necessarily $\mu \neq 0$, otherwise $g_0 = 0$ and $\Phi(g_0) = \|f\|_2^2 \gg \varepsilon^2$. Let us put $\lambda = \frac{1}{\mu}$. We necessarily have $\lambda g_0 = P_{t/2} f - P_t(g_0)$, hence $(\lambda + P_t)(g_0) = P_{t/2} f$, so

$$g_0 = (\lambda + P_t)^{-1} P_{t/2} f.$$

Let us now write the constraint:

$$\begin{aligned} \varepsilon^2 = \|f - P_{t/2} g\|_2^2 &= \|f - P_{t/2}(\lambda + P_t)^{-1} P_{t/2} f\|_2^2 \\ &= \|f - (\lambda + P_t)^{-1} P_t f\|_2^2 = \|\lambda(\lambda + P_t)^{-1} f\|_2^2. \end{aligned}$$

Clearly:

$$\lambda \mapsto \|\lambda(\lambda + P_t)^{-1} f\|_2^2$$

is increasing from 0 to $\|f\|_2^2$. As well,

$$\lambda \mapsto \|(\lambda + P_t)^{-1} P_{t/2} f\|_2^2$$

is decreasing.

On the other way: if $L = \int x dE_x$, and

$$\|\lambda(\lambda + P_t)^{-1} f\|_2^2 = \int_0^\infty \left(\frac{\lambda}{\lambda + e^{-tx}}\right)^2 d\langle E_x f, f\rangle \geq \varepsilon^2$$

and

$$\|g_0\|_2^2 = \|(\lambda + P_t)^{-1} P_{t/2} f\|_2^2 = \int_0^\infty \left(\frac{1}{\lambda + e^{-tx}}\right)^2 e^{-tx} d\langle E_x f, f\rangle.$$

Let us recall the following result from [9, Lemma 3.19].

Theorem 5 *There exists* $b > 1$, $C_1'' > 0$, $C_2'' > 0$, *such that for all* $\lambda \geq 1$, $\delta = \frac{1}{\lambda}$, *then*

$$\begin{aligned} &dim(\Sigma_{b\lambda}) - dim(\Sigma_\lambda) = dim(\Sigma_{b\lambda} \ominus \Sigma_\lambda) \\ &= \int_M P_{\Sigma_{b\lambda}}(x, x) d\mu(x) - \int_{\mathcal{M}} P_{\Sigma_\lambda}(x, x) d\mu(x) \neq 0 \end{aligned}$$

and more precisely:

$$C_1'' \int_{\mathcal{M}} \frac{1}{|B(x, \delta)|} d\mu(x) \leq dim(\Sigma_{b\lambda} \ominus \Sigma_\lambda) \leq C_2'' \int_{\mathcal{M}} \frac{1}{|B(x, \delta)|} d\mu(x). \tag{52}$$

As $P_{\Sigma_{\sqrt{a}}} = E_a$, one can build a function $f \in \mathbb{L}^2$ such that:

$$\|f - P_{\Sigma_{\sqrt{a}}} f\|_2^2 = \int_a^\infty \langle E_x f, f\rangle = \|f\|_2^2 - \|E_a f\|_2^2 = \|f - E_a f\|_2^2 = a^{-s}$$

for $a = b^{2j}$, and $j \in \mathbb{N}$. It is enough to have:

$$\|P_{\Sigma_{b^{j+1}} \ominus \Sigma_{b^j}}(f)\|_2^2 = b^{-2js} - b^{-2(j+1)s}$$

and this could be done by the previous theorem.

Let us choose for $\varepsilon > 0$, $b^{-2js} \geq 4\varepsilon^2 \geq b^{-2(j+1)s}$. So

$$\int_{b^{2j}}^\infty \langle E_x f, f\rangle = b^{-2js} \geq 4\varepsilon^2 \geq b^{-2(j+1)s} = \int_{b^{2(j+1)}}^\infty \langle E_x f, f\rangle$$

so, if $\lambda = e^{-ta}$, $a = b^{2j}$,

$$\int_0^\infty \left(\frac{\lambda}{\lambda + e^{-tx}}\right)^2 d\langle E_x f, f\rangle \geq \int_a^\infty \left(\frac{\lambda}{\lambda + e^{-tx}}\right)^2 d\langle E_x f, f\rangle$$
$$\geq \int_a^\infty \left(\frac{e^{-ta}}{e^{-ta} + e^{-tx}}\right)^2 d\langle E_x f, f\rangle = \frac{1}{4}\int_a^\infty d\langle E_x f, f\rangle \geq \varepsilon^2.$$

But

$$\|g_0\|_2^2 \geq \|(\lambda + P_t)^{-1} P_{t/2} f\|_2^2 = \int_0^\infty \left(\frac{1}{\lambda + e^{-tx}}\right)^2 e^{-tx} d\langle E_x f, f\rangle$$
$$= e^{ta}\int_0^\infty \left(\frac{1}{e^{-ta} + e^{-tx}}\right)^2 e^{-ta} e^{-tx} d\langle E_x f, f\rangle$$
$$= e^{ta}\int_0^\infty \left(\frac{e^{-t/2x} e^{-t/2a}}{e^{-ta} + e^{-tx}}\right)^2 d\langle E_x f, f\rangle$$
$$= e^{ta}\int_0^\infty \left(\frac{1}{e^{-t/2(a-x)} + e^{-t/2(x-a)}}\right)^2 d\langle E_x f, f\rangle$$
$$\geq e^{ta}\frac{1}{4}\int_0^\infty e^{-t|a-x|} d\langle E_x f, f\rangle \geq e^{ta}\frac{1}{4}\int_{\frac{a}{b^2}}^a e^{-t(a-x)} d\langle E_x f, f\rangle$$

$$\geq e^{t\frac{a}{b^2}}\frac{1}{4}\int_{\frac{a}{b^2}}^{a} d\langle E_x f, f\rangle = e^{t\frac{a}{b^2}}\frac{1}{4}\int_{b^{2j-2}}^{b^{2j}} d\langle E_x f, f\rangle = e^{t\frac{a}{b^2}}\frac{1}{4}(b^{-(2j-2)s} - b^{-2js})$$

$$= e^{t\frac{a}{b^2}}\frac{1}{4}b^{-2js}(b^{2s}-1) \geq \varepsilon^2(b^{2s}-1)e^{tb^{2j-2}}$$

$$\geq \varepsilon^2(b^{2s}-1)e^{tc\varepsilon^{-2/s}}; \quad c = 4^{-1/s}b^{-4}.$$

□

Acknowledgments The authors would like to thank Richard Nickl, Aad van der Vaart and Harry van Zanten for insightful comments on this work.

Appendix: Compact Riemannian manifold

We investigate now the case described in Sect. 3 where $\mathcal{M}$ is a compact Riemannian manifold of dimension d without boundary. Our aim here is to prove Ahlfors' condition (2) for this special case.

Proposition 6 *Let $\mathcal{M}$ be a compact Riemannian manifold of dimension d without boundary. Then there exist $0 < c \leq C < \infty$ such that,*

$$\text{for all } x \in \mathcal{M}, \text{ for all } 0 < r < Diam(\mathcal{M}), \quad cr^d \leq |B(x,r)| \leq Cr^d.$$

Proof Let μ and ρ be the (non normalised) Riemannian measure and metric on $\mathcal{M}$. The proposition is a consequence of the Bishop-Gromov comparison Theorem, see [15] and [7].

As $\mathcal{M}$ is compact, clearly

$$\exists \kappa \in \mathbb{R}, \quad \text{such that : for all } x \in \mathcal{M}, \ Ricc_x \geq (d-1)\kappa g_x$$

where $Ricc$ is the Ricci tensor and g is the metric tensor. Let $V_\kappa(r)$ be the volume of the (any) ball of radius r in the model space of dimension d and constant sectional curvature κ. Let V_d be the volume of the unit ball of $\mathbb{R}^d$.

1. For $\kappa > 0$, the model space is the sphere $\frac{1}{\sqrt{\kappa}}\mathbb{S}_d$ of $\mathbb{R}^{d+1}$ of radius $\frac{1}{\sqrt{\kappa}}$ and

$$V_\kappa(r) = dV_d\int_0^r \left(\frac{\sin\sqrt{\kappa}t}{\sqrt{\kappa}}\right)^{d-1} dt; \text{ so } \left(\frac{2}{\pi}\right)^{d-1} V_d r^d \leq V_\kappa(r) \leq V_d r^d$$

2. For $\kappa = 0$, the model space is $\mathbb{R}^d$ and

$$V_\kappa(r) = V_d r^d$$

3. For $\kappa < 0$ the model space is the hyperbolic space of constant sectional curvature κ.

$$V_\kappa(r) = dV_d \int_0^r \left(\frac{\sinh \sqrt{|\kappa|}t}{\sqrt{|\kappa|}} \right)^{d-1} dt; \text{ so } \quad V_d r^d \leq V_\kappa(r) \leq V_d r^d e^{(d-1)\sqrt{|\kappa|}r}$$

as $s \leq \sinh(s) \leq se^s$.

Moreover by the Bishop-Gromov comparaison comparison Theorem: $r \mapsto \frac{|B(x,r)|}{V_\kappa(r)}$ is non increasing. So if $0 < \varepsilon < r < s \leq R = diam(M)$:

$$\frac{\mu(\mathcal{M})}{V_\kappa(R)} = \frac{|B(x,R)|}{V_\kappa(R)} \leq \frac{|B(x,s)|}{V_\kappa(s)} \leq \frac{|B(x,r)|}{V_\kappa(r)} \leq \frac{|B(x,\varepsilon)|}{V_\kappa(\varepsilon)} \mapsto 1, \text{ when } \varepsilon \mapsto 0.$$

So

$$\frac{V_\kappa(s)}{V_\kappa(r)} \leq \frac{|B(x,r)|}{|B(x,s)|}; \quad \mu(M)\frac{V_\kappa(r)}{V_\kappa(R)} \leq |B(x,r)| \leq V_\kappa(r)$$

So

$$A(\frac{r}{s})^d \leq \frac{|B(x,r)|}{|B(x,s)|} \text{ (doubling)}; \quad cr^d \leq |B(x,r)| \leq CV_d r^d \text{ (homogeneity)};$$

$$\text{for } \kappa > 0, C = 1, \ c = \left(\frac{2}{\pi}\right)^{d-1} \frac{\mu(\mathcal{M})}{R^d}. \quad A = \left(\frac{2}{\pi}\right)^{d-1}.$$

$$\text{for } \kappa = 0, C = 1, \ c = \frac{\mu(\mathcal{M})}{R^d}. \quad A = 1.$$

$$\text{for } \kappa < 0, \ C = e^{(d-1)\sqrt{|\kappa|}R}; \ c = \frac{\mu(\mathcal{M})}{R^d e^{(d-1)\sqrt{|\kappa|}R}}. \quad A = \frac{1}{e^{(d-1)\sqrt{|\kappa|}R}}. \qquad \square$$

Remark 4 If $(\mathcal{M}, \mu, \rho)$ is a compact metric space with a Borel measure μ, then if we have the doubling condition:

$$0 < r < s \Longrightarrow |B(x,s)| \leq \frac{1}{A}\left(\frac{s}{r}\right)^m |B(x,r)|$$

then

$$\text{for all } r \leq R = diam(M), \ Cr^m \leq |B(x,r)|, \quad \text{with} \quad C = \frac{A|\mathcal{M}|}{R^m}.$$

References

1. Angers, J.-F., Kim, P.T.: Multivariate Bayesian function estimation. Ann. Stat. **33**(6), 2967–2999 (2005)
2. Bhattacharya, A., Dunson, D.B.: Nonparametric Bayesian density estimation on manifolds with applications to planar shapes. Biometrika **97**(4), 851–865 (2010)
3. Carl, B.: Entropy numbers, s-numbers, and eigenvalue problems. J. Funct. Anal. **41**(3), 290–306 (1981)
4. Carl, B., Stephani, I.: Entropy, Compactness and the Approximation of Operators. Cambridge Tracts in Mathematics, vol. 98. Cambridge University Press, Cambridge (1990)
5. Castillo, I.: Lower bounds for posterior rates with Gaussian process priors. Electr. J. Stat. **2**, 1281–1299 (2008)
6. Castillo, I., van der Vaart, A.: Needles and straw in a haystack: posterior contraction for possibly sparse sequences. Ann. Stat. **40**(4), 2069–2101 (2012)
7. Chavel, I.: Eigenvalues in Riemannian Geometry. Pure and Applied Mathematics, vol. 115. Academic Press Inc., Orlando (1984); Including a chapter by Burton Randol, With an appendix by Jozef Dodziuk
8. Coifman, R.R., Maggioni, M.: Diffusion wavelets. Appl. Comput. Harm. Anal. **21**(1), 53–94 (2006)
9. Coulhon, T., Kerkyacharian, G., Petrushev, P.: Heat kernel generated frames in the setting of Dirichlet spaces. J. Fourier Anal. Appl. **18**, 995–1066 (2012)
10. Davies, E.B.: One-parameter semigroups. London Mathematical Society Monographs, vol. 15. Academic Press Inc. (Harcourt Brace Jovanovich Publishers), London (1980)
11. Efromovich, S.: On sharp adaptive estimation of multivariate curves. Math. Methods Stat. **9**(2), 117–139 (2000)
12. Ghosal, S., Ghosh, J.K., van der Vaart, A.W.: Convergence rates of posterior distributions. Ann. Stat. **28**(2), 500–531 (2000)
13. Ghosal, S., van der Vaart, A.W.: Convergence rates of posterior distributions for noniid observations. Ann. Stat. **35**(1) (2007)
14. Grigor'yan, A.: Heat kernel and analysis on manifolds. AMS/IP Studies in Advanced Mathematics, vol. 47. American Mathematical Society, Providence (2009)
15. Gromov, M.: Metric structures for Riemannian and non-Riemannian spaces. Progress in Mathematics, vol. 152. Birkhäuser Boston Inc., Boston (1999)
16. Heinonen, J.: Lectures on Analysis on Metric Spaces. Universitext. Springer, New York (2001)
17. Kuelbs, J., Li, W.V.: Metric entropy and the small ball problem for Gaussian measures. J. Funct. Anal. **116**(1), 133–157 (1993)
18. Li, W.V., Linde, W.: Approximation, metric entropy and small ball estimates for Gaussian measures. Ann. Probab. **27**(3), 1556–1578 (1999)
19. Lifshits, M.: Lectures on Gaussian processes. SpringerBriefs in Mathematics. Springer, Berlin (2012)
20. Mardia, K.V., Jupp, P.E.: Directional statistics. Wiley Series in Probability and Statistics. Wiley, Chichester (2000); Revised reprint of Statistics of directional data by Mardia [MR0336854 (49 #1627)]
21. Nadler, B., Lafon, S., Coifman, R.R., Kevrekidis, I.G.: Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. App. Comput. Harm. Anal. **21**(1), 113–127 (2006)
22. Ouhabaz, E.M.: Analysis of heat equations on domains. London Mathematical Society Monographs Series, vol. 31. Princeton University Press, Princeton (2005)
23. Petrushev, P., Xu, Y.: Localized polynomial frames on the interval with Jacobi weights. J. Fourier Anal. Appl. **11**(5), 557–575 (2005)
24. Pisier, G.: The volume of convex bodies and Banach space geometry. Cambridge Tracts in Mathematics, vol. 94. Cambridge University Press, Cambridge (1989)
25. Rivoirard, V., Rousseau, J.: Posterior concentration rates for infinite dimensional exponential families. Bayesian Anal. **7**(2), 1–24 (2012)
26. Saloff-Coste, L.: Aspects of Sobolev-type inequalities. London Mathematical Society Lecture Note Series, vol. 289. Cambridge University Press, Cambridge (2002)
27. Shen, X., Wasserman, L.: Rates of convergence of posterior distributions. Ann. Stat. **29**(3), 687–714 (2001)
28. Stein, E.M., Weiss, G.: Introduction to Fourier analysis on Euclidean spaces. Princeton Mathematical Series, vol. 32. Princeton University Press, Princeton (1971)
29. Tomczak-Jaegermann, N.: Dualité des nombres d'entropie pour des opérateurs à valeurs dans un espace de Hilbert. C. R. Acad. Sci. Paris Sér. I Math. **305**(7):299–301 (1987)
30. van der Vaart, A., van Zanten, H.: Information rates of nonparametric Gaussian process methods. J. Mach. Learn. Res. **12**, 2095–2119 (2011)

31. van der Vaart, A.W., van Zanten, H.: Rates of contraction of posterior distributions based on Gaussian process priors. Ann. Stat. **36**(3), 1435–1463 (2008)
32. van der Vaart, A.W., van Zanten, H.: Reproducing kernel Hilbert spaces of Gaussian priors. IMS Collect. **3**, 200–222 (2008)
33. van der Vaart, A.W., van Zanten, J.H.: Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. Ann. Stat. **37**(5B):2655–2675 (2009)