Bernoulli **26**(3), 2020, 1832–1862
https://doi.org/10.3150/19-BEJ1171

Kernel and wavelet density estimators on manifolds and more general metric spaces

GALATIA CLEANTHOUS[1], ATHANASIOS G. GEORGIADIS[2], GERARD KERKYACHARIAN[3], PENCHO PETRUSHEV[4] and DOMINIQUE PICARD[5]

[1]*Department of Mathematics, Statistics and Physics, Newcastle University, Newcastle Upon Tyne, NE1 7RU, UK. E-mail: galatia.cleanthous@newcastle.ac.uk*
[2]*Department of Mathematics and Statistics, University of Cyprus, 1678 Nicosia, Cyprus. E-mail: gathana@ucy.ac.cy*
[3]*Université de Paris, LPSM, CREST, F-75013 Paris, France. E-mail: kerk@math.univ-paris-diderot.fr*
[4]*Department of Mathematics, University of South Carolina. E-mail: pencho@math.sc.edu*
[5]*Université de Paris, LPSM, F-75013 Paris, France. E-mail: picard@math.univ-paris-diderot.fr*

We consider the problem of estimating the density of observations taking values in classical or nonclassical spaces such as manifolds and more general metric spaces. Our setting is quite general but also sufficiently rich in allowing the development of smooth functional calculus with well localized spectral kernels, Besov regularity spaces, and wavelet type systems. Kernel and both linear and nonlinear wavelet density estimators are introduced and studied. Convergence rates for these estimators are established and discussed.

Keywords: adaptive density estimators; Ahlfors regularity; Besov space; heat kernel; non-parametric estimators; sample kernel density estimators; wavelet density estimators

1. Introduction

A great deal of efforts is nowadays invested in solving statistical problems, where the data are located in quite complex geometric domains such as matrix spaces or surfaces (manifolds). Geometric models are motivated by the fact that many real-world high dimensional data are affected by the "curse of dimensionality" phenomena. Although the data for a given data mining problem may have many features, in reality, viewed from a geometric perspective the intrinsic dimensionality of the data support of the full feature space may be low. A seminal example in this direction is the case of spherical data. Developments in this domain have been motivated by a number of important applications. We only mention here some of the statistical challenges posed by astrophysical data: investigating the fundamental properties of the cosmic microwave background (CMB) observation including its polarization, impainting of the CMB in zones on the sphere obstructed by other radiations, producing cosmological maps, exploring clusters of galaxies or point sources, investigating the true nature of ultra high energy cosmic rays (UHECR). We refer the reader to the overview by Starck, Murtagh, and Fadili [33] of the use of various wavelet tools in this domain as well as the work in [1] and [24] of some of the authors of this article.

Many more geometric objects have been analyzed in the statistical literature. For instance, landmark-based shape spaces have diverse applications in morphometrics, medical diagnostics,

1350-7265

machine vision (see, for instance, [2]); Pelletier [28,29] has long ago investigated nonparametric statistics on regular manifolds; the now emerging field of signal processing on graphs or networks is a mine for geometric investigation (see, for instance, [31]).

Dealing with complex data requires the development of more sophisticated tools and statistical methods. In particular, these tools should capture the natural topology and geometry of the application domain. Our contribution will be essentially theoretical, however, our statements will be illustrated by examples from various application fields.

Our aim in this article is to study the *density estimation problem*, namely, one observes $X_1, \ldots, X_n$ that are i.i.d. random variables defined on a space $\mathcal{M}$ and the problem is to find a good estimation to the common density function. This problem has a long history in mathematical statistics, especially when the set $\mathcal{M}$ is $\mathbb{R}^d$ or a cube in $\mathbb{R}^d$ (see e.g. the monograph [35] and the references herein). Here we will consider general spaces $\mathcal{M}$ such as Riemannian manifolds or spaces of matrices or graphs and prove that under some reasonable assumptions, we can develop an estimation theory with estimation procedures, regularity sets and upper bounds evaluations quite parallel to what has been done in $\mathbb{R}^d$. In particular, we intend to develop kernel methods with upper bounds and oracle properties as well as wavelet thresholding estimators with adaptive behavior.

There are two principle quantities that will dominate our setting and usually appear in the minimax rates of convergence. The first quantity d reflects the basic dimensional structure of the sets (here introduced with the aid of the doubling condition on the measure) and the second quantity s is associated to "regularity" and leads us to considering settings, where regularity spaces can be defined along with kernels.

The setting presented here is quite general. Naturally, the classical results on $\mathbb{R}^d$ and on the sphere are contained in this general framework, but also various other settings are covered. In particular, spaces of matrices, Riemannian manifolds, and convex subsets of Riemannian manifolds are covered.

This programm requires the development of new techniques and methods that break new ground in the density estimation problem. Our *main contributions* are as follows:

(a) In a general setting described below, we introduce *kernel density estimators* sufficiently concentrated to establish oracle inequalities and $\mathbb{L}^p$-error rates of convergence for probability density functions lying in *Besov spaces*.

(b) We develop linear *wavelet density estimators* and obtain $\mathbb{L}^p$-error estimates for probability density functions in general *Besov* smoothness spaces.

(c) We also establish $\mathbb{L}^p$-error estimates on *nonlinear wavelet density estimators with hard thresholding* and prove adaptivity properties up to logarithmic terms.

To put the results from this article in perspective, we next compare them with the results in [3]. The geometric settings in both articles are comparable and the two papers study adaptive methods. In [3], different standard statistical models (regression, white noise model, density estimation) are considered in a Bayesian framework. The methods are different (we do not consider here Bayesian estimators) and the results are also different (since, again, we are not interested here in a concentration result of the posterior distribution). It is noteworthy that the results in the so called *dense case* exhibit the same rates of convergence. It is also important to observe the wide adaptation properties of the thresholding estimates here which allow to obtain minimax rates of convergence in the so called *sparse case*, that was not possible in [3].

Outline. The organization of this article is as follows. In Section 2, we begin with a short toy-example. In Section 3, we describe our general setting and give some examples. We introduce kernel density estimators in Section 4 and establish $\mathbb{L}^p$-error estimates for probability density functions in general Besov spaces. In Section 5, we review some basic facts related to our setting such as the construction of wavelet frames, Besov spaces, and other background material. Section 6, we prove the main results on kernel density estimators from Section 4. Section 7 is devoted to linear wavelet estimators. In Section 8, we introduce and study adaptive wavelet threshold density estimators. We establish $\mathbb{L}^p$-error estimates for probability density functions in Besov spaces. In the Supplementary Material [4], Section 1, we present several additional examples covered by the setting, introduced in Section 3. In Section 2 of [4], we place the proofs of some claims from Section 5. Finally, in Section 3, of [4] we give the proof of the main Theorem 8.2 of Section 8.

Notation.

Throughout $\mathbb{I}_E$ will denote the indicator function of the set E and $\|\cdot\|_p := \|\cdot\|_{\mathbb{L}^p(\mathcal{M},\mu)}$. We denote by c, c' positive constants that may vary at every occurrence. Most of these constants will depend on some parameters that may be indicated in parentheses. We will also denote by $c_0, c_1, \ldots$ as well as $c_\star$, $c_\diamond$ constants that will remain unchanged throughout. The relation $a \sim b$ means that there exists a constant $c > 1$ such that $c^{-1}a \le b \le ca$. We will also use the notation $a \wedge b := \min\{a, b\}$, $a \vee b := \max\{a, b\}$, and $C^k(\mathbb{R}_+)$, $k \in \mathbb{N} \cup \{\infty\}$, will stand for the set of all functions with continuous derivatives of order up to k on $\mathbb{R}_+ := [0, \infty)$.

2. Illustrating example

Before going into the mathematical tools and results, let us study a toy example (although the problem has a scientific interest by itself, but will be treated here only from a methodological point of view): let us estimate a probability density on a spider web $\mathcal{M}$. A spider web is a good and emblematic example, since it reflects the geometric aspects of the problem. As well, most of the time, it is a very inhomogeneous medium, built of regions with varying characteristics and this is also an important issue that we will commented in the paper.

One observes $X_1, \ldots, X_n$ i.i.d. random variables defined on the space $\mathcal{M}$ with the common density f. The X_i's may be the places where preys are falling, and the density f might give indications on the regions of the web devoted to catching preys as opposed to others affected to others tasks such as informing the spider of the presence of predators.

To proceed to the estimation we first propose to identify $\mathcal{M}$ with a graph (simple, undirected, no loops) with T vertices and edges. As is standard in graph theory (and commonly used in clustering, see for instance [36]), we form the adjacency $T \times T$ matrix A, defined by $A_{ij} = 1$ if there is an edge between i and j, and $A_{ij} = 0$ otherwise. $\mathcal{M}$ is naturally equipped with the geodesic-distance and the Laplacian matrix $L_{T\times T}$ is defined as:

$$L = D - A,$$

where D is the diagonal degree matrix, $D_{ii} = \sum_{j\neq i} A_{i,j}$ is the degree of the vertex i.

Again, as is common in clustering procedures, we compute the spectral decomposition of L. L has $\lambda_1 \le \cdots \le \lambda_T$ as eigenvalues and $V^1, \ldots, V^T$ as (normed) eigenvectors: $V^j = (V_1^j, \ldots, V_T^j)$.

The kernel estimator that will be studied in the sequel is the following one: for an arbitrary point x of the graph:

$$\hat{K}_\delta(x) := \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{T}\Phi(\delta\sqrt{\lambda_j})V_{X_i}^j V_x^j.$$

Here δ is the bandwidth and Φ is a "Littlewood–Paley function", that is, $\Phi \in C^\infty(\mathbb{R})$ is a real-valued function with the following properties: $\operatorname{supp}\Phi \subset [0,1]$ and $\Phi(\lambda) = 1$ for $\lambda \in [0, 1/2]$.

3. Setting and motivation

We assume that $(\mathcal{M}, \rho, \mu)$ is a metric measure space equipped with a distance ρ and a positive Radon measure μ.

Let $X_1, \ldots, X_n$ be independent identically distributed (i.i.d.) random variables on $\mathcal{M}$ with common probability having a density function (pdf) f with respect to the measure μ. Our goal is to estimate the density f. To an estimator $\hat{f}_n$ of f we associate its risk:

$$R_n(\hat{f}, f, p) = \mathbb{E}_f\left(\int_{\mathcal{M}}\left|\hat{f}_n(x) - f(x)\right|^p \mu(dx)\right)^{\frac{1}{p}} = \mathbb{E}_f\|\hat{f}_n - f\|_p, \quad 1 \le p < \infty$$

as well as its $\mathbb{L}_\infty$ risk:

$$R_n(\hat{f}, f, \infty) = \mathbb{E}_f\left(\operatorname*{ess\,sup}_{x\in\mathcal{M}}\left|\hat{f}_n(x) - f(x)\right|\right) = \mathbb{E}_f\|\hat{f}_n - f\|_\infty.$$

We will operate in the setting described below. Most of the material related to the setting can be found in an extended form in papers [6,20]. Note that, depending on the results that will be established, some of the following conditions will be assumed, others will not.

3.1. Doubling and non-collapsing conditions

The following conditions are related to the "dimensional" structure of $\mathcal{M}$. This notion which is one of the important parameters in statistical estimation may be subtle to define in complex models. It has often been linked with entropy. We propose, here instead, to connect the dimension with a more flexible notion, the doubling condition that was introduced in the 70s by R. Coifman and G. Weiss [5].

C1. We assume that the metric space $(\mathcal{M}, \rho, \mu)$ satisfies the following *doubling volume condition*:

$$\mu\big(B(x,2r)\big) \le c_0\mu\big(B(x,r)\big) \quad \text{for all } x \in \mathcal{M} \text{ and } r > 0, \tag{3.1}$$

where $B(x,r) := \{y \in \mathcal{M} : \rho(x,y) < r\}$ and $c_0 > 1$ is a constant. The above implies that there exist constants $c_0' \ge 1$ and $d > 0$ such that

$$\mu\big(B(x,\lambda r)\big) \le c_0'\lambda^d\mu\big(B(x,r)\big) \quad \text{for all } x \in \mathcal{M},\ r > 0, \text{ and } \lambda > 1. \tag{3.2}$$

The least d such that (3.2) holds is the so called *homogeneous dimension* of $(\mathcal{M}, \rho, \mu)$.

From now on we will use the notation $|E| := \mu(E)$ for $E \subset \mathcal{M}$.

In developing adaptive density estimators in Section 8 we will additionally assume that $(\mathcal{M}, \rho, \mu)$ is a compact measure space with $\mu(\mathcal{M}) < \infty$ satisfying the following condition:

C1A. *Ahlfors regular volume condition:* There exist constants $c_1, c_2 > 0$ and $d > 0$ such that

$$c_1 r^d \leq \big|B(x,r)\big| \leq c_2 r^d, \quad \forall x \in \mathcal{M} \text{ and } 0 < r \leq \operatorname{diam}(\mathcal{M}). \tag{3.3}$$

Clearly, condition **C1A** implies conditions **C1** and condition **C2** below with d from (3.3) being the homogeneous dimension of $(\mathcal{M}, \rho, \mu)$.

It is interesting already to notice that d will in effect play the role of dimension in the statistical results as well. Condition **C1A** is obviously true for $\mathcal{M} = \mathbb{R}^d$ with μ the Lebesgue measure. Also, under **C1A**, the doubling condition is precisely related to the metric entropy using the following lemma whose elementary proof can be found for instance in [3], Proposition 1. Note that it might be not true in general. For $\epsilon > 0$, we define, as usual, the covering number $N(\epsilon, \mathcal{M})$ as the smallest number of balls of radius ϵ covering $\mathcal{M}$.

Lemma 3.1. *Under the condition* **C1A** *and if $\mathcal{M}$ is compact, there exist constants c' and c'' such that*

$$\frac{1}{c'}\left(\frac{1}{\epsilon}\right)^d \leq N(\epsilon, \mathcal{M}) \leq \frac{2^d}{c''}\left(\frac{1}{\epsilon}\right)^d, \tag{3.4}$$

for all $0 < \epsilon \leq \epsilon_0$.

The cases where **C1** is verified but not **C1A** are interesting but more complicated. They correspond to media $\mathcal{M}$ with non homogeneous behavior: think of the spider web, with different zones with different characteristics or medical images. More examples will be given in the sequel.

C2. *Non-collapsing condition:* There exists a constant $c_3 > 0$ such that

$$\inf_{x \in \mathcal{M}} \big|B(x,1)\big| \geq c_3 > 0. \tag{3.5}$$

This condition is not necessarily very restrictive. For instance, it is satisfied if $\mathcal{M}$ is compact. It is satisfied for $\mathbb{R}^d$ if μ is the Lebesgue measure, but untrue for $\mathbb{R}$ if μ is the Gaussian measure.

3.2. The underlying operator

Before delving into the specificity of our set of assumptions (described below), let us explain our motivation. A standard method in density estimation is the kernel estimation method. Namely, consider a family of functions $K_\delta \colon \mathcal{M} \times \mathcal{M} \to \mathbb{R}$, indexed by $\delta > 0$. Then the associated kernel density estimator is defined by

$$\widehat{K}_\delta(x) := \frac{1}{n}\sum_{i=1}^{n} K_\delta(X_i, x), \quad x \in \mathcal{M}. \tag{3.6}$$

In the classical case of $\mathbb{R}^d$, an important family is the family of translation kernels $K_\delta(x,y) = [\frac{1}{\delta}]^d G(\frac{x-y}{\delta})$, where G maps $\mathbb{R}^d$ into $\mathbb{R}$. When $\mathcal{M}$ is a more involved set such as a manifold or a set of graphs or matrices, the simple operations of translation and dilation may not be meaningful. Hence, even finding a family of kernels to start with might be challenging. As will be shown in Section 4.2 our assumptions will provide quite naturally a family of kernels.

When dealing with a kernel estimation method, it is standard to consider two quantities: $b_\delta(f) := \|\mathbb{E}_f \widehat{K}_\delta - f\|_p$, $\|\xi_f\|_p := \|\widehat{K}_\delta - \mathbb{E}_f \widehat{K}_\delta\|_p$. The analysis of the second (stochastic) term $\|\xi_f\|_p$ can be reduced via Rosenthal inequalities to proper bounds on norms of $K_\delta(\cdot,\cdot)$ and f (see the Lemmas 5.7, 5.8). The analysis of the first term $b_\delta(f)$ is linked to the approximation properties of the family $\mathbb{E}_f \widehat{K}_\delta$. One can stop at this level and express the performance of an estimator in terms of $\|\mathbb{E}_f \widehat{K}_\delta - f\|_p$. This is the purpose of oracle inequalities (see Theorem 4.4). However, it is more compelling if the rate of approximation of the family $\|\mathbb{E}_f \widehat{K}_\delta - f\|_p$ can be related to regularity properties of the function f. In $\mathbb{R}^d$ it is standardly proved (see, e.g., [17]) that if K is a translation family with mild conditions on K, then polynomial rates of approximation are obtained for functions with Besov regularity. Therefore, an important issue becomes finding regularity spaces associated to a possibly complex set $\mathcal{M}$. On a compact metric space (M,ρ) one can always define the scale of s-Lipschitz spaces by the following norm $\|f\|_{\mathrm{Lip}_s} := \|f\|_\infty + \sup_{x\neq y} \frac{|f(x)-f(y)|}{\rho(x,y)^s}$, $0 < s \leq 1$.

On Euclidian spaces a function can be much more regular than Lipschitz; for instance, it can be differentiable of an arbitrary order. When $\mathcal{M}$ is a set where there is no obvious notion of differentiability, one can introduce regularity spaces by means of an operator that is an analogue of the Laplace operator on $\mathbb{R}^d$ or a Riemannian manifold. To make this point more clear, we next discuss the classical case of Sobolev spaces W_2^k on $\mathbb{R}^d$. The space W_2^k is defined as the set of all functions f on $\mathbb{R}^d$ such that in weak sense $\partial_1^{k_1} \cdots \partial_d^{k_d} f \in \mathbb{L}^2$, $\sum_{i=1}^d k_i \leq k$, and

$$\|f\|_{W_2^k} = \sum_{\sum_{j=1}^d k_j \leq k} \left\|\partial_1^{k_1} \cdots \partial_d^{k_d} f\right\|_2 < \infty.$$

As is well known $\mathcal{F}(\partial_j^k f)(\xi) = (i\xi_j)^k \mathcal{F}(f)(\xi)$, where $\mathcal{F}(f)(\xi) := \int_{\mathbb{R}^d} f(x) e^{-ix\cdot\xi}\,dx$ is the Fourier transform of f. Now, applying Plancherel's identity we get, using the notation $|\xi|^2 := \sum_{i=1}^d |\xi_i|^2$,

$$\|f\|_{W_2^k} = \sum_{\sum_{j=1}^d k_j \leq k} \left\|\xi_l^{k_1} \cdots \xi_d^{k_d} \mathcal{F}(f)(\xi)\right\|_2 \sim \left\|\mathcal{F}(f)(\xi)\right\|_2 + \left\||\xi|^k \mathcal{F}(f)(\xi)\right\|_2.$$

Let A be the operator defined by the identity

$$\mathcal{F}\big(A(f)\big)(\xi) = |\xi| \mathcal{F}(f)(\xi), \quad \text{that is,} \quad A(f) = \mathcal{F}^{-1}\big(|\xi| \mathcal{F}(f)(\xi)\big).$$

Then

$$\|f\|_{W_2^k} \sim \sum_{0\leq l\leq k} \left\|A^l f\right\|_2 \quad \big(A^0 = I_d\big).$$

Clearly, the operator A is not a differential operator (even in dimension $d = 1$).

In this framework, the differential operator $\Delta = \sum_{j=1}^{d} \partial_j^2$ (Laplacian) plays a prominent role. We have

$$\mathcal{F}(-\Delta f)(\xi) = |\xi|^2 \mathcal{F}(f)(\xi) = A^2(f) \quad \text{and hence } A^2 = -\Delta \text{ or } A = \sqrt{-\Delta}.$$

The advantage of the operator $\sqrt{-\Delta}$ over Δ is that it is a first order operator, while Δ is of order two; $\sqrt{-\Delta}$ is a good substitute for differentiation. As a result the Sobolev spaces W_2^k, $k \in \mathbb{N}$, are naturally described in terms of the operator $\sqrt{-\Delta}$. The situation is quite similar if $\mathcal{M}$ is a Riemannian manifold.

The above discussion leads to the conclusion that regularity spaces can naturally be defined via operators that behave as the Laplace operator on $\mathbb{R}^d$. This is the underlying idea of our development in this article.

It is by now well understood that "good operators" are the ones that are self-adjoint, positive, and defined on a metric measure space $(\mathcal{M}, \rho, \mu)$ with the doubling property of the measure (see (3.1)) and whose heat kernel has Gaussian localization. This is precisely the setting introduced in [6,20] that we adopt in this article. This setting is rich enough to allow the development of the Littlewood–Paley theory in almost complete analogy with the classical case on $\mathbb{R}^d$ and at the same time it is sufficiently general to cover a number of interesting cases as will be shown in what follows. In particular, the setting allows to develop Besov spaces B_{pq}^s with complete set of indices. As will be seen it is also sufficiently flexible in allowing to develop kernel and wavelet density estimators.

Our main assumption is that the space $(\mathcal{M}, \rho, \mu)$ is complemented by an essentially self-adjoint non-negative operator L on $\mathbb{L}^2(\mathcal{M}, \mu)$, mapping real-valued to real-valued functions, such that the associated semigroup $P_t = e^{-tL}$ (see the formal definition in Section 4.1) consists of integral operators with the (heat) kernel $p_t(x, y)$ obeying the following conditions:

C3. *Gaussian localization:* There exist constants $c_4, c_5 > 0$ such that

$$\left|p_t(x, y)\right| \leq \frac{c_4 \exp(-\frac{c_5 \rho^2(x,y)}{t})}{(|B(x, \sqrt{t})||B(y, \sqrt{t})|)^{1/2}} \quad \text{for all } x, y \in \mathcal{M}, t > 0. \tag{3.7}$$

C4. *Hölder continuity:* There exists a constant $\alpha > 0$ such that

$$\left|p_t(x, y) - p_t\big(x, y'\big)\right| \leq c_4 \left(\frac{\rho(y, y')}{\sqrt{t}}\right)^{\alpha} \frac{\exp(-\frac{c_5 \rho^2(x,y)}{t})}{(|B(x, \sqrt{t})||B(y, \sqrt{t})|)^{1/2}} \tag{3.8}$$

for $x, y, y' \in \mathcal{M}$ and $t > 0$, whenever $\rho(y, y') \leq \sqrt{t}$.

C5. *Markov property:*

$$\int_{\mathcal{M}} p_t(x, y)\, d\mu(y) = 1 \quad \text{for all } x \in \mathcal{M} \text{ and } t > 0. \tag{3.9}$$

Above $c_0, c_1, c_2, c_3, c_4, c_5, d, \alpha > 0$ are structural constants. These technical assumptions express that fact that the Heat kernel associated with the operator L behaves as the standard Heat kernel of $\mathbb{R}^d$.

3.3. Typical examples

Here we present some examples of setups that are covered by the setting described above. We will use these examples in what follows to illustrate our theory. More involved examples will be given Section 1 of the Supplementary Material [4].

3.3.1. *Classical case on* $\mathcal{M} = \mathbb{R}^d$

Here μ is the Lebesgue measure and ρ is the Euclidean distance on $\mathbb{R}^d$. In this case, we consider the operator

$$-Lf(x) = \sum_{j=1}^{d} \partial_i^2 f(x) = \operatorname{div}(\nabla f)(x),$$

defined on the space $\mathcal{D}(\mathbb{R}^d)$ of C^∞ functions with compact support. As is well known the operator L is positive essentially self-adjoint and has a unique extension to a positive self-adjoint operator. The associate semigroup e^{-tL} is given by the operator with the Gaussian kernel: $p_t(x, y) = (4\pi t)^{-\frac{d}{2}} \exp(-\frac{|x-y|^2}{4t})$.

3.3.2. *Periodic case on* $\mathcal{M} = [-1, 1]$

Here μ is the Lebesgue measure and ρ is the Euclidean distance on the circle. The operator is $Lf = -f''$, defined on the set on infinitely differentiable periodic functions. It has eigenvalues $\lambda_k = k^2\pi^2$ for $k \in \mathbb{N}$ and eigenspaces

$$\ker(L) = E_0 = \operatorname{span}\left\{\frac{1}{\sqrt{2}}\right\}, \qquad \ker\bigl(L - k^2\pi^2 I_d\bigr) = E_{\lambda_k} = \operatorname{span}\{\cos k\pi x, \sin k\pi x\}.$$

3.3.3. *Non-periodic case on* $\mathcal{M} = [-1, 1]$ *with Jacobi weight*

Note that this example, further developed in Section 1 of the Supplemental Material [4], can arise when dealing with data issued from a density which itself has received a folding treatment such as in the Wicksell problem [18,21]. Now, the measure is

$$d\mu_{\alpha,\beta}(x) = w_{\alpha,\beta}(x)\,dx = (1-x)^\alpha (1+x)^\beta\,dx, \quad \alpha, \beta > -1,$$

the distance is $\rho(x, y) := |\arccos x - \arccos y|$, and L is the Jacobi operator

$$Lf(x) := -\frac{[w_{\alpha,\beta}(x)(1-x^2)f'(x)]'}{w_{\alpha,\beta}(x)}.$$

Conditions **C1–C5** are satisfied, but not the Ahlfors condition **C1A**, unless $\alpha = \beta = -\frac{1}{2}$. The discrete spectral decomposition of L is given by one dimensional spectral spaces:

$$\mathbb{L}^2(M, \mu_{\alpha,\beta}) = \bigoplus E_{\lambda_k^{\alpha,\beta}}, \quad E_{\lambda_k^{\alpha,\beta}} = \ker\bigl(L - \lambda_k^{\alpha,\beta} I_d\bigr) = \operatorname{span}\bigl\{P_k^{\alpha,\beta}(x)\bigr\},$$

where $P_k^{\alpha,\beta}(x)$ is the kth degree Jacobi polynomial and $\lambda_k^{\alpha,\beta} = k(k+\alpha+\beta+1)$.

3.3.4. *Riemannian manifold $\mathcal{M}$ without boundary*

If $\mathcal{M}$ is a Riemannian manifold, then the Laplace operator Δ_M is well defined on M (see [16]) and we consider

$$L = -\Delta_M.$$

If $\mathcal{M}$ is compact, then conditions **C1–C5** are verified, including the Ahlfors condition **C1A**. Furthermore, there exists an associated discrete spectral decomposition with finite dimensional spectral eigenspaces of L:

$$\mathbb{L}^2(\mathcal{M}, \mu) = \bigoplus E_{\lambda_k}, \quad E_{\lambda_k} = \ker(L - \lambda_k I_d), \lambda_0 = 0 < \lambda_1 < \lambda_1 < \cdots.$$

3.3.5. *Unit sphere $\mathcal{M} = \mathbb{S}^{d-1}$ in $\mathbb{R}^d$, $d \geq 3$*

This is the most famous Riemannian manifold with the induced structure from $\mathbb{R}^d$. Here μ is the Lebesgue measure on $\mathbb{S}^{d-1}$, ρ is the geodesic distance on $\mathbb{S}^{d-1}$: $\rho(\xi, \eta) = \arccos(\langle \xi, \eta \rangle_{\mathbb{R}^d})$, and $L := -\Delta_0$ with Δ_0 being the Laplace–Beltrami operator on $\mathbb{S}^{d-1}$. The spectral decomposition of the operator L can be described as follows:

$$\mathbb{L}^2\big(\mathbb{S}^{d-1}, \mu\big) = \bigoplus E_{\lambda_k}, \quad E_{\lambda_k} = \ker(L - \lambda_k I_d), \lambda_k = k(k + d - 2).$$

Here E_{λ_k} is the restriction to $\mathbb{S}^{d-1}$ of harmonic homogeneous polynomials of degree k (spherical harmonics), see [34]. We have $\dim(E_{\lambda_k}) = \binom{d-1}{d+k-1} - \binom{d-1}{d+k-3}$.

3.3.6. *Lie group of matrices*: $\mathcal{M} = \mathrm{SU}(2)$

This example is interesting in astrophysical problems, especially in the measures associated to the CMB, where instead of only measuring the intensity of the radiation we also measure its polarization. By definition

$$\begin{aligned} \mathrm{SU}(2) &:= \left\{ \begin{pmatrix} a & b \\ -\overline{b} & \overline{a} \end{pmatrix}, a, b \in \mathbb{C}, |a|^2 + |b|^2 = 1 \right\} \\ &= \left\{ \begin{pmatrix} \sin\theta e^{i\phi} & -\cos\theta e^{-i\psi} \\ \cos\theta e^{i\psi} & \sin\theta e^{-i\phi} \end{pmatrix}, 0 \leq \phi, \psi < 2\pi; 0 \leq \theta < \frac{\pi}{2} \right\}. \end{aligned}$$

The normalized Haar measure (see [9]) is given by

$$\int_{\mathrm{SU}(2)} f \, d\mu = \frac{1}{2\pi^2} \int_0^{2\pi} \int_0^{2\pi} \int_0^{\frac{\pi}{2}} f\left(\begin{pmatrix} \sin\theta e^{i\phi} & -\cos\theta e^{-i\psi} \\ \cos\theta e^{i\psi} & \sin\theta e^{-i\phi} \end{pmatrix} \right) \sin\theta \cos\theta \, d\phi \, d\psi \, d\theta.$$

Thus

$$q \in \mathrm{SU}(2) \quad \leftrightarrow \quad q \in M(2, \mathbb{C}), \qquad q^{-1} = -q^*, \qquad \det(q) = 1.$$

This is a compact group which topologically is the sphere $\mathbb{S}^3 \subset \mathbb{R}^4$. So, if

$$x = \begin{pmatrix} x_1 + ix_2 & x_3 + ix_4 \\ -(x_3 - ix_4) & x_1 - ix_2 \end{pmatrix}, \qquad y = \begin{pmatrix} y_1 + iy_2 & y_3 + iy_4 \\ -(y_3 - iy_4) & y_1 - iy_2 \end{pmatrix}$$

with $\|y\|^2 = \sum_i y_i^2 = 1 = \|x\|^2 = \sum_i x_i^2$, then

$$\langle x, y\rangle_4 = \sum_i x_i y_i = \frac{1}{2}\operatorname{Tr}\big[xy^*\big].$$

Therefore,

$$\rho_{\mathrm{SU}(2)}(x, y) = \arccos\frac{1}{2}\operatorname{Tr}\big[xy^*\big])$$

and for any $q, x, y \in \mathrm{SU}(2)$ we have $v = \rho_{\mathrm{SU}(2)}(qx, qy) = \rho_{\mathrm{SU}(2)}(xq, yq) = \rho_{\mathrm{SU}(2)}(x, y)$. The eigenvalues of $L = -\Delta$ are $\lambda_k = k(k+2)$ and the dimension of the respective eigenspaces E_{λ_k} is $(k+1)^2$.

Remark 3.2. Looking at some of these examples an important question already arises: how to choose in a given problem the distance ρ as well as the dominating measure before even choosing the operator L and a class of regularity? In $\mathbb{R}^d$, most often the Euclidean distance and the Lebesgue measure seem more or less unavoidable. In some other cases it might not be so obvious.

Let us take for instance, the simple case of $\mathcal{M} = [-1, 1]$. The cases of the ball, the simplex (see Section 1 of the Supplementary Material [4]) and more generally sets with boundaries give rise in fact to identical discussions. Therefore, we will focus on the case of the interval.

So, if $\mathcal{M} = [-1, 1]$, a possible choice and probably the most standard one in statistical examples could be taking ρ as the Euclidean distance and μ as the Lebesgue measure. Then standard kernel and wavelet statistical methods are available. However, "something" which generally is often swept under the carpet or not really detailed has to be "done" about the boundary points $\{-1, 1\}$. Often special regularity conditions are assumed about these boundary points such as $f(-1) = f(1) = 0$ (see Section 3.3.2), which de facto lead to different methods for representing the functions to be estimated.

Let us now look at the choices (again for the interval $[-1, 1]$, Section 3.3.3) that are made in the "Jacobi" case. The distance $\rho(x, y) = |\arccos x - \arccos y|$ suggests a one-to-one correspondence with the semi-circle. The measure $d\mu_{\alpha,\beta}(x) = (1-x)^\alpha(1+x)^\beta\,dx$, $\alpha, \beta > -1$, suggests that the points in the middle of the interval (say from $[-\frac{1}{2}, +\frac{1}{2}]$, where the measure behaves as the Lebesgue measure) will not be weighted in the same way as the points near the boundaries. In some cases, this makes perfect sense. For instance, if one needs to give a hard weight on these points because they require special attention, or on the contrary a small one.

Apart from these considerations, there are in fact two measures in the family $\mu_{\alpha,\beta}$ that are undeniable if $\mathcal{M} = [-1, 1]$ equipped with the distance $\rho(x, y) = |\arccos x - \arccos y|$. The first one is the Lebesgue measure (because Lebesgue is always undeniable), corresponding to $\alpha = \beta = 0$. The second one is $\mu_{-\frac{1}{2},-\frac{1}{2}}$, because in that case there is a one-to-one identification between $(\mathcal{M}, \rho, \mu_{-\frac{1}{2},-\frac{1}{2}})$ and the semi circle equipped with the Euclidean distance and Lebesgue measure.

If we look more precisely into these two choices, we see that for the last case all required conditions, including the Ahlfors condition are satisfied, and the dimension $d = 1$, which is intuitively expected. Let us now observe that the case of the Lebesgue measure $\mu_{0,0}$ would lead to a larger dimension $d = 2$.

4. Kernel density estimator on metric measure space

4.1. Functional calculus associated to L

A key trait of our setting is that it allows to develop a smooth functional calculus. Let E_λ, $\lambda \geq 0$, be the spectral resolution associated with the operator L in our setting. As L is non-negative, essentially self-adjoint and maps real-valued to real-valued functions, then for any real-valued, measurable, and bounded function h on $\mathbb{R}_+$, the operator

$$h(L) := \int_0^\infty h(\lambda)\, dE_\lambda, \tag{4.1}$$

is well defined on $\mathbb{L}^2(\mathcal{M})$. The operator $h(L)$, called *spectral multiplier*, is bounded on $\mathbb{L}^2(\mathcal{M})$, self-adjoint, and maps real-valued to real-valued functions [37]. We will be interested in integral spectral multiplier operators $h(L)$. If $h(L)(x, y)$ is the kernel of such an operator, it is real-valued and symmetric. From condition **C4** of our setting we know that e^{-tL} is an integral operator whose (heat) kernel $p_t(x, y)$ is symmetric and real-valued: $p_t(y, x) = p_t(x, y) \in \mathbb{R}$.

Observe that in the simple case (most common case in this paper) when the spectrum of L is discrete, L has eigenvalues $\lambda_0 < \lambda_1 < \cdots$, and $h(L)$ is an integral operator with kernel

$$h(L)(x, y) = \sum_k h(\lambda_k) P_k(x, y),$$

with $P_k(x, y) = \sum_i v_i^{\lambda_k}(x)\overline{v_i^{\lambda_k}(y)}$ (projector operator), where $v_i^{\lambda_k}(x)$, $i = 1, \ldots, \dim(E_{\lambda_k})$ is an orthonormal basis of E_{λ_k}, the eigenspace associated with the eigenvalue λ_k.

Our further development will heavily depend on the following result from the smooth functional calculus induced by the heat kernel, developed in [20], Theorem 3.4. It asserts that if a function $g(u)$ on $\mathbb{R}$ is even, compactly supported and sufficiently smooth, then the kernel of the operator $g(\delta\sqrt{L})$ is very well localized. Note that the operator with kernel $[\frac{1}{\delta}]^d G(\frac{x-y}{\delta})$ on $\mathbb{R}^d$, where $G : \mathbb{R}^d \to \mathbb{R}$ is a bounded compactly supported function, has a similar localization.

Theorem 4.1. *Under the conditions* **C1**–**C5**, *let* $g \in C^N(\mathbb{R})$, $N > d$, *be even, real-valued, and* $\operatorname{supp} g \subset [-R, R]$, $R > 0$. *Then* $g(\delta\sqrt{L})$, $\delta > 0$, *is an integral operator with kernel* $g(\delta\sqrt{L})(x, y)$ *satisfying*

$$\left|g(\delta\sqrt{L})(x, y)\right| \leq c\left|B(x, \delta)\right|^{-1}\left(1 + \delta^{-1}\rho(x, y)\right)^{-N+\frac{d}{2}}, \quad \forall x, y \in \mathcal{M}, \tag{4.2}$$

where $c > 0$ *is a constant depending on* $\|g\|_\infty$, $\|g^{(N)}\|_\infty$, N, R *and the constants* c_0, c_4, c_5 *from our setting. Furthermore, for any* $\delta > 0$ *and* $x \in \mathcal{M}$

$$\int_{\mathcal{M}} g(\delta\sqrt{L})(x, y)\, d\mu(y) = g(0). \tag{4.3}$$

This result is a building block for the localization properties of the kernel estimators considered in the sequel.

4.1.1. *Examples*

Here we revisit some of the examples from Section 3.3 and in each case we specify the form of the kernels of the spectral multipliers operators $h(L)$.

(a) Let $\mathcal{M} = [-1, 1]$ be in the periodic case (Section 3.3.2). It is readily seen that the projection operators are: $P_0(x, y) = \frac{1}{2}$, $P_k(x, y) = \cos k\pi(x - y)$. Hence, formally, $h(L)(x, y) = \frac{1}{2}h(0) + \sum_{k\geq 1} h(k^2\pi^2)\cos k\pi(x - y)$, $x, y \in [-1, 1]$.

(b) If $\mathcal{M}$ is a Riemanian manifold (Section 3.3.4), then $h(L)$ is a kernel operator with kernel $h(L)(x, y) = \sum_k h(\lambda_k)P_k(x, y)$, with $P_k(x, y) = \sum_i v_i^{\lambda_k}(x)\overline{v_i^{\lambda_k}(y)}$, where $v_i^{\lambda_k}(x)$, $i = 1, \ldots, \dim(E_{\lambda_k})$ is an orthonormal basis of E_{λ_k}.

(c) In the case of the sphere (Section 3.3.5), the orthogonal projector operator $P_{E_{\lambda_k}} : \mathbb{L}^2(\mathbb{S}^{d-1}) \mapsto E_{\lambda_k}$ is a kernel operator with kernel of the form $P_k(\xi, \eta) = L_k(\langle \xi, \eta\rangle_{\mathbb{R}^d})$, where $L_k(x) = \frac{1}{|\mathbb{S}^{d-1}|}(1 + \frac{k}{\nu})C_k^\nu(x)$, $\nu = \frac{d-2}{2}$. Here $C_k^\nu(x)$ is the Gegenbauer polynomials of degree k. Usually, the polynomials $\{C_k^\nu(x)\}$ are defined by the generating function $\frac{1}{(1-2rx+r^2)^\nu} = \sum_{k\geq 0} r^k C_k^\nu(x)$, $|r| < 1$, $|x| < 1$. Hence, formally $h(L)(\xi, \eta) = \sum_k h(k(k+d-2))L_k(\langle \xi, \eta\rangle_{\mathbb{R}^d})$, $\xi, \eta \in \mathbb{S}^{d-1}$.

(d) In the case of SU(2) (Section 3.3.6), the orthogonal projector $P_k : \mathbb{L}^2(\mathrm{SU}(2)) \mapsto E_{\lambda_k}$ is the operator with kernel $p_k(\xi, \eta) = L_k(\frac{1}{2}\mathrm{Tr}[\xi\eta^*])$, where $L_k(x) = \frac{1}{|\mathbb{S}^3|}(1 + k)C_k^1(x)$. Hence, formally $h(L)(\xi, \eta) = \sum_k h(k(k + 2))L_k(\frac{1}{2}\mathrm{Tr}[\xi\eta^*])$, $\xi, \eta \in \mathrm{SU}(2)$.

Observe that in each example from above, if $h(u) := g(\delta\sqrt{u})$, where $g(u)$ is even, compactly supported and sufficiently smooth, then $h(L)$ is an integral operator with kernel localized as specified in Theorem 4.1.

4.2. Kernel density estimators on the metric measure space $\mathcal{M}$

Our goal in this section is to introduce and study kernel density estimators (kde's) on a metric measure space $(\mathcal{M}, \rho, \mu)$ in setting described above. More precisely, we assume that conditions **C1–C5** are satisfied, and do not necessarily assume the Ahlfors regular volume condition **C1A**.

To explain our construction of kernel estimators, we begin by considering the classical example of the periodic case on $\mathcal{M} = [-1, 1]$, presented in Section 3.3.2. In this setting, the following nonparametric estimator is standard

$$\hat{f}_T(x) = \frac{1}{2} + \frac{1}{n}\sum_{i=1}^n \sum_{1\leq k\leq T} \cos k\pi(x - X_i).$$

It falls into the category of orthogonal series estimators. It is well known that these estimators have nice $\mathbb{L}^2$ properties but can drastically fail in $\mathbb{L}^p$, $p \neq 2$, or locally.

In this setting, we replace $\hat{f}_T$ by a "smoothed version" defined by

$$\hat{K}_\delta(x) = \frac{1}{2}K(0) + \frac{1}{n}\sum_{i=1}^n \sum_{k\geq 1} K(\delta k)\cos k\pi(x - X_i) =: \frac{1}{n}\sum_{i=1}^n K_\delta(x, X_i), \tag{4.4}$$

with $K_\delta(x,y) = \frac{1}{2}K(0) + \sum_{k\geq 1} K(\delta k)\cos k\pi(x-y)$, where K is a smooth and rapidly decaying (or compactly supported) function on $\mathbb{R}_+$.

In analogy to this case, replacing the circle by $\mathcal{M}$ and the Laplacian by the operator $-L$ we can naturally introduce kde's on $\mathcal{M}$ by means of the machinery of spectral multipliers.

We will use Theorem 4.1 to define a family of multiplier operators whose kernels are suitable for the construction of kernel density estimators on $\mathcal{M}$ and introduce the kde's in the general setting of this article.

Definition 4.2. Let $X_1, \ldots, X_n$ be i.i.d. random variables on $\mathcal{M}$. Let $K(\delta\sqrt{L})(x,y)$ with $0 < \delta \leq 1$ (the bandwidth) be the kernel of the integral operator $K(\delta\sqrt{L})$, where $K : \mathbb{R}_+ \to \mathbb{R}$. The associated kernel density estimator is defined by

$$\widehat{K}_\delta(x) := \frac{1}{n}\sum_{i=1}^{n} K(\delta\sqrt{L})(X_i, x), \quad x \in \mathcal{M}. \tag{4.5}$$

Remarks 4.3. Again the analogy with the torus case could lead to the choice of K to be the indicator function of the interval $[0,1]$, for instance. This choice would induce $\mathbb{L}^2$ properties, but not $\mathbb{L}^p$ because this function is not smooth enough to get the localization of $K(\delta\sqrt{L})(x,y)$ from Theorem 4.1.

If $K(\lambda) = e^{-\lambda^2}$, then $K(\delta\sqrt{L})(x,y) = p_{\delta^2}(x,y)$ (the "heat kernel") can be used to define a kernel density estimator. This choice relates to the Bayesian estimator provided in [3].

The kernel estimators from (4.5) although constructed using orthogonal projectors, because of the smoothing function K will have properties that are comparable to translation kernel estimators in $\mathbb{R}^d$. In $\mathbb{R}^d$ some properties such as a number of moment annulation (see for instance [35]) to get a correct bias are required which will be here replaced by the vanishing properties at infinity of the function K and its smoothness.

4.3. Upper bound estimates for kernel density estimators

We will especially study kernel density estimators induced by compactly supported C^∞ multipliers, often called Littlewood–Paley functions. In fact other type of kernels among the family of multipliers could lead to similar results. More explicitly, let Φ be an even $C^\infty(\mathbb{R})$ real-valued function with the following properties:

$$\operatorname{supp}\Phi \subset [-1,1] \quad \text{and} \quad \Phi(\lambda) = 1 \quad \text{for } \lambda \in [-1/2, 1/2]. \tag{4.6}$$

By Theorem 4.1, it follows that $\Phi(\delta\sqrt{L})$ is an integral operator with well localized symmetric kernel $\Phi(\delta\sqrt{L})(x,y)$ and the Markov property:

$$\int_{\mathcal{M}} \Phi(\delta\sqrt{L})(x,y)\, d\mu(y) = 1. \tag{4.7}$$

We will denote by $\mathbb{E} = \mathbb{E}_f$ the expectation with respect to the probability measure $\mathbb{P} = \mathbb{P}_f$. and we will consider the class of Litlewood–Paley kde's:

$$\widehat{\Phi}_\delta(x) = \widehat{\Phi}_\delta(x, X_1, \dots, X_n) := \frac{1}{n}\sum_{i=1}^{n} \Phi(\delta\sqrt{L})(x, X_i), \quad \forall x \in \mathcal{M}. \tag{4.8}$$

4.3.1. *Explicit examples of Littlewood Paley kernel density estimators*

We now present specific examples of kernel density estimators induced by the setups in Section 3.3 (see also Section 4.1.1). We have already discussed the periodic case on $[-1, 1]$ in (4.4).

(a) On $[-1, 1]$ in the Jacobi framework (Section 3.3.3), we get the following estimator:

$$\widehat{\Phi}_\delta(x) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k} \Phi\big(\delta\sqrt{k(k+\alpha+\beta+1)}\big) P_k^{\alpha,\beta}(x) P_k^{\alpha,\beta}(X_i) \quad \text{for } x \in [-1, 1],$$

where $P_k^{\alpha,\beta}(x)$ is the normalized kth degree Jacobi polynomial.

(b) For the sphere (Section 3.3.5), we get

$$\widehat{\Phi}_\delta(x) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k} \Phi\big(\delta\sqrt{k(k+d-2)}\big) L_k\big(\langle x, X_i\rangle_{\mathbb{R}^d}\big) \quad \text{for } x \in \mathbb{S}^{d-1},$$

where $L_k(x) = \frac{1}{|\mathbb{S}^{d-1}|}(1 + \frac{k}{\nu})C_k^\nu(x)$, $\nu = \frac{d-2}{2}$.

(c) For SU(2) (Section 3.3.6), we get

$$\widehat{\Phi}_\delta(x) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k} \Phi\big(\delta\sqrt{k(k+2)}\big) L_k\left(\frac{1}{2}\operatorname{Tr}\big[X_i x^*\big]\right) \quad \text{for } x \in \mathrm{SU}(2),$$

with $L_k(x) = \frac{1}{|\mathbb{S}^3|}(1+k)C_k^1(x)$.

4.3.2. *Upper bound results*

We first establish oracle inequalities for the kernel density estimators introduced in Definition 4.2.

Theorem 4.4. *Assume $1 \le p \le \infty$ and let Φ be a Littlewood–Paley function as above. In the setting described above and with $\widehat{\Phi}_\delta$ from (4.8), we have*:

(i) *If $2 \le p < \infty$, then*

$$\mathbb{E}\|\widehat{\Phi}_\delta - f\|_p \le \frac{c(p)}{(n\delta^d)^{1-\frac{1}{p}}} + \frac{c(p)}{(n\delta^d)^{\frac{1}{2}}}\|f\|_{\frac{p}{2}}^{\frac{1}{2}} + \big\|\Phi(\delta\sqrt{L})f - f\big\|_p, \quad 0 < \delta \le 1.$$

(ii) *If* $1 \le p < 2$ *and* $\operatorname{supp}(f) \subset B(x_0, R)$ *for some* $x_0 \in \mathcal{M}$ *and* $R > 0$, *then*

$$\mathbb{E}\|\widehat{\Phi}_\delta - f\|_p \le \frac{c(p)}{(n\delta^d)^{\frac{1}{2}}} \big|B(x_0, R)\big|^{\frac{1}{p}-\frac{1}{2}} + \big\|\Phi(\delta\sqrt{L})f - f\big\|_p, \quad 0 < \delta \le 1.$$

(iii) *There exists a constant* c *such that for any* $q \ge 2$ *and* $0 < \delta \le 1$ *we have*

$$\mathbb{E}\|\widehat{\Phi}_\delta - f\|_\infty \le c\delta^{-\frac{d}{q}} \left(\frac{q}{(n\delta^d)^{1-\frac{1}{q}}} + \frac{q^{1/2}}{(n\delta^d)^{\frac{1}{2}}} \|f\|_\infty^{\frac{1}{2}-\frac{1}{q}} \right) + \big\|\Phi(\delta\sqrt{L})f - f\big\|_\infty.$$

We next estimate the rates of $\mathbb{L}^p$-approximation of pdf's f belonging to Besov space balls by kernel estimators. The precise definition of the Besov spase $B^s_{p\tau}$ will be given in Section 5.4 below. Define

$$B^s_{p\tau}(m) := \big\{ f \text{ is pdf} : \|f\|_{B^s_{p\tau}} \le m \big\} \tag{4.9}$$

and

$$B^s_{p\tau}(m, x_0, R) := \big\{ f \in B^s_{p\tau}(m) : \operatorname{supp} f \subset B(x_0, R) \big\}, \quad x_0 \in \mathcal{M}, m, R > 0. \tag{4.10}$$

Here is our main result on the properties of our kernel estimators for density functions in Besov spaces, when the risk and the regularity classes are defined with the same norm.

Theorem 4.5. *Assume* $s > 0$, $1 \le p \le \infty$, $0 < \tau \le \infty$, $m > 0$, *and let* Φ *be a Littlewood–Paley function as above. In the setting described above and with* $\widehat{\Phi}_\delta$ *from* (4.8) *we have*:

(i) *If* $2 \le p < \infty$ *and* $\delta = n^{-\frac{1}{2s+d}}$, *then*

$$\sup_{f \in B^s_{p\tau}(m)} \mathbb{E}\|\widehat{\Phi}_\delta - f\|_p \le c n^{-\frac{s}{2s+d}}, \tag{4.11}$$

where $c = c(p, s, m, \tau) > 0$.

(ii) *If* $1 \le p < 2$, $x_0 \in \mathcal{M}$, $R > 0$, *and* $\delta = n^{-\frac{1}{2s+d}}$, *then*

$$\sup_{f \in B^s_{p\tau}(m, x_0, R)} \mathbb{E}\|\widehat{\Phi}_\delta - f\|_p \le c n^{-\frac{s}{2s+d}}, \tag{4.12}$$

where $c = c(p, s, m, \tau, x_0, R) > 0$.

(iii) *If* $\delta = (\frac{\log n}{n})^{\frac{1}{2s+d}}$, *then*

$$\sup_{f \in B^s_{\infty\tau}(m)} \mathbb{E}\|\widehat{\Phi}_\delta - f\|_\infty \le c \left(\frac{\log n}{n} \right)^{\frac{s}{2s+d}}, \tag{4.13}$$

where $c = c(s, m, \tau) > 0$.

The proofs of Theorems 4.4 and 4.5 will be preceded by several definitions and ancillary claims that we place in the next subsection. The actual proof will be given in Section 6.

Remarks 4.6. Note that we do not claim that the above rates are necessarily minimax, although they show similarities with the results established in $\mathbb{R}^d$.

The length of the paper does not allow to investigate the full lower bounds results. Let us only mention that if we add the Ahlfors condition **C1A** to the setting, then matching lower bounds can be obtained by direct adaptation of the proof given in the case of the sphere in [1]. In the case where the Ahlfors condition is not valid, the problem is more complex since not only the regularity might be non-homogeneous due to Besov conditions but also the dimension itself may vary spatially. In this case, the upper bounds might not be optimal.

It is interesting to compare the obtained upper bounds for $\mathcal{M} = [-1, 1]$ in different cases (torus or Jacobi). In the torus case, no surprise, the rate is the usual one with dimension $d = 1$. In the Jacobi case, the dimension is $d = 1 + (2\alpha + 1)_+ \vee (2\beta + 1)_+$, which in the particular case $\alpha = \beta = 0$ (corresponding to μ the Lebesgue measure), gives a slower rate than the usual one. This is due to the fact that the boundary affects the spaces and the approximation rate is not the same. In the case $\alpha = \beta = -\frac{1}{2}$ that corresponds to perfect identification of $[-1, 1]$ with the semi-circle (see Remark 3.2) the rate is the usual with dimension $d = 1$.

5. Besov spaces and wavelets in geometric setting

In this section, we collect some basic technical facts and results related to the setting described in Section 3 that will be crucial for the properties of the density estimators. Most of them can be found in [6,14,20].

5.1. Geometric properties

Conditions **C1** and **C2** yield

$$\big|B(x, r)\big| \geq (c_3/c_0)r^d, \quad x \in \mathcal{M}, 0 < r \leq 1. \tag{5.1}$$

To compare the volumes of balls with different centers $x, y \in \mathcal{M}$ and the same radius r we will use the inequality

$$\big|B(x, r)\big| \leq c_0\left(1 + \frac{\rho(x, y)}{r}\right)^d \big|B(y, r)\big|, \quad x, y \in \mathcal{M}, r > 0. \tag{5.2}$$

As $B(x, r) \subset B(y, \rho(y, x) + r)$ the above inequality is immediate from (3.2).

We will also need the following simple inequality (see [6], Lemma 2.3): If $\tau > d$, then for any $\delta > 0$

$$\int_{\mathcal{M}} \big(1 + \delta^{-1}\rho(x, y)\big)^{-\tau} d\mu(y) \leq c\big|B(x, \delta)\big|, \quad x \in \mathcal{M}, \tag{5.3}$$

where $c = (2^{-d} - 2^{-\tau})^{-1}$.

5.2. Spectral spaces

We recall the definition of the spectral spaces Σ_λ^p, $1 \le p \le \infty$, from [6]. Denote by $C_0^\infty(\mathbb{R})$ the set of all even real-valued compactly supported functions. We define

$$\Sigma_\lambda^p := \big\{ f \in \mathbb{L}^p(\mathcal{M}) : \theta(\sqrt{L}) f = f \text{ for all } \theta \in C_0^\infty(\mathbb{R}), \theta \equiv 1 \text{ on } [0, \lambda] \big\}, \quad \lambda > 0.$$

We will need the following proposition (Nikolski type inequality):

Proposition 5.1. *Let $1 \le p \le q \le \infty$. If $g \in \Sigma_\lambda^p$, $\lambda \ge 1$, then $g \in \Sigma_\lambda^q$ and*

$$\|g\|_q \le c_\star \lambda^{d(1/p-1/q)} \|g\|_p, \tag{5.4}$$

where the constant $c_\star > 1$ is independent of p and q.

This proposition was established in [6], Proposition 3.12 (see also [20], Proposition 3.11). We present its proof in the Supplementary Material [4] because we need to control the constant $c_\star$.

5.3. Wavelets

In the setting of this article, wavelet type frames for Besov and Triebel–Lizorkin spaces are developed in [20]. Here, we review the construction of the frames from [20] and their basic properties. Indeed, in this setting the 'wavelets' do not form an orthonormal basis but a frame. In this case, the construction of a 'dual wavelet system' is necessary to get a representation of type (5.10).

This construction is inspired by to the Littlewood–Paley construction of the standard wavelets introduced by [11–13].

The construction of frames involves a "dilation" constant $b > 1$ whose role is played by 2 in the wavelet theory on $\mathbb{R}$.

The construction starts with the selection of a function $\Psi_0 \in C^\infty(\mathbb{R}_+)$ with the properties: $\Psi_0(\lambda) = 1$ for $\lambda \in [0, 1]$, $0 \le \Psi_0(\lambda) \le 1$, and $\operatorname{supp} \Psi_0 \subset [0, b]$. Denote $\Psi(\lambda) := \Psi_0(\lambda) - \Psi_0(b\lambda)$ and set $\Psi_j(\lambda) := \Psi(b^{-j}\lambda)$, $j \in \mathbb{N}$. From this, it readily follows that

$$\sum_{j=0}^{J} \Psi_j(\lambda) = \Psi_0\big(b^{-J}\lambda\big), \quad \lambda \in \mathbb{R}_+. \tag{5.5}$$

For $j \ge 0$ we let $\mathcal{X}_j \subset \mathcal{M}$ be a maximal δ_j-net on $\mathcal{M}$ with $\delta_j := c_6 b^{-j}$. It is easy to see that for any $j \ge 0$ there exists a disjoint partition $\{A_{j\xi}\}_{\xi \in \mathcal{X}_j}$ of $\mathcal{M}$ consisting of measurable sets such that

$$B(\xi, \delta_j/2) \subset A_{j\xi} \subset B(\xi, \delta_j), \quad \xi \in \mathcal{X}_j.$$

Here $c_6 > 0$ is a sufficiently small constant (see [20]).

Lemma 5.2. *If $\mathcal{M}$ is compact, then there exists a constant $c_7 > 0$ such that*

$$\operatorname{card}(\mathcal{X}_j) \le c_7 b^{jd}, \quad j \ge 0. \tag{5.6}$$

Proof. This is a simple consequence of the proof of [6], Proposition 3.20. □

The jth level frame elements $\psi_{j\xi}$ are defined by

$$\psi_{j\xi}(x) := |A_{j\xi}|^{1/2} \Psi_j(\sqrt{L})(x, \xi), \quad \xi \in \mathcal{X}_j. \tag{5.7}$$

We will also use the more compact notation $\psi_\xi := \psi_{j\xi}$ for $\xi \in \mathcal{X}_j$.

Let $\mathcal{X} := \bigcup_{j\ge 0} \mathcal{X}_j$, where equal points from different sets $\mathcal{X}_j$ will be regarded as distinct elements of $\mathcal{X}$, so $\mathcal{X}$ can be used as an index set. Then $\{\psi_\xi\}_{\xi\in\mathcal{X}}$ is Frame #1.

The construction of a dual frame $\{\tilde{\psi}_\xi\}_{\xi\in\mathcal{X}} = \bigcup_j \{\tilde{\psi}_{j\xi}\}_{\xi\in\mathcal{X}_j}$ is much more involved; we refer the reader to Section 4.3 in [20] for the details.

By construction, the two frames satisfy

$$\Psi_j(\sqrt{L})(x, y) = \sum_{\xi\in\mathcal{X}_j} \psi_{j\xi}(y)\tilde{\psi}_{j\xi}(x), \quad j \ge 0. \tag{5.8}$$

A basic result from [20] asserts that for any $f \in \mathbb{L}^p(\mathcal{M}, \mu)$, $1 \le p < \infty$,

$$f = \sum_{j\ge 0} \Psi_j(\sqrt{L}) f \quad \big(\text{convergence in } \mathbb{L}^p\big) \tag{5.9}$$

and the same holds in $\mathbb{L}^\infty$ if f is uniformly continuous and bounded (UCB) on $\mathcal{M}$. As a consequence, for any $f \in \mathbb{L}^p(\mathcal{M}, \mu)$, $1 \le p \le \infty$, ($\mathbb{L}^\infty = \text{UCB}$) we have

$$f = \sum_{j=0}^{\infty} \sum_{\xi\in\mathcal{X}_j} \langle f, \tilde{\psi}_{j\xi}\rangle \psi_{j\xi} \quad \big(\text{convergence in } \mathbb{L}^p\big). \tag{5.10}$$

Furthermore, frame decomposition results are established in [20] for Besov and Triebel–Lizorkin spaces with full range of indices.

Properties of frames in the Ahlfors regularity case. We next present some properties of the frame elements in the case when condition **C1A** is stipulated (see [20]).

1. *Localization*: For every $k \in \mathbb{N}$, there exists a constant $c(k) > 0$ such that

$$\big|\psi_{j\xi}(x)\big|, \big|\tilde{\psi}_{j\xi}(x)\big| \le c(k) b^{jd/2} \big(1 + b^j \rho(x, \xi)\big)^{-k}, \quad x \in \mathcal{M}. \tag{5.11}$$

2. *Norm estimation*: For $1 \le p \le \infty$

$$c_\diamond^{-1} b^{jd(\frac{1}{2}-\frac{1}{p})} \le \|\psi_{j\xi}\|_p, \|\tilde{\psi}_{j\xi}\|_p \le c_\diamond b^{jd(\frac{1}{2}-\frac{1}{p})}, \quad \xi \in \mathcal{X}_j, j \ge 0. \tag{5.12}$$

3. For $1 \le p \le \infty$

$$\Big\| \sum_{\xi \in \mathcal{X}_j} \lambda_\xi \psi_{j\xi} \Big\|_p \le c_\diamond b^{jd(\frac{1}{2}-\frac{1}{p})} \Big(\sum_{\xi \in \mathcal{X}_j} |\lambda_\xi|^p \Big)^{1/p}, \quad j \ge 0, \tag{5.13}$$

with the usual modification when $p = \infty$. Above the constant $c_\diamond > 1$ depends only on p, b, Ψ_0, and the structural constants of the setting.

5.4. Besov spaces

We will deal with probability density functions (pdf's) in Besov spaces associated to the operator L in our setting. These spaces are developed in [6,20]. Definition 5.3 coincides in $\mathbb{R}^d$ with one the definitions of usual Besov spaces with L replaced by Laplacian ($-\Delta$ in fact to get a positive operator).

Here we present some basic facts about Besov spaces that will be needed later on.

Let $\Phi_0, \Phi \in C^\infty(\mathbb{R}_+)$ be real-valued functions satisfying the conditions:

$$\operatorname{supp} \Phi_0 \subset [0, b], \qquad \Phi_0(\lambda) = 1 \quad \text{for } \lambda \in [0, 1], \qquad \Phi_0(\lambda) \ge c > 0 \quad \text{for } \lambda \in \big[0, b^{3/4}\big], \tag{5.14}$$

$$\operatorname{supp} \Phi \subset \big[b^{-1}, b\big], \qquad \Phi(\lambda) \ge c > 0 \quad \text{for } \lambda \in \big[b^{-3/4}, b^{3/4}\big]. \tag{5.15}$$

Set $\Phi_j(\lambda) := \Phi(b^{-j}\lambda)$, for $j \ge 1$.

Definition 5.3. Let $s > 0$, $1 \le p \le \infty$, and $0 < q \le \infty$. The *Besov space* $B^s_{pq} = B^s_{pq}(\mathcal{M}, L)$ is defined as the set of all functions $f \in \mathbb{L}^p(\mathcal{M}, \mu)$ such that

$$\|f\|_{B^s_{pq}} := \Big(\sum_{j \ge 0} \big(b^{sj} \big\| \Phi_j(\sqrt{L}) f \big\|_p \big)^q \Big)^{1/q} < \infty, \tag{5.16}$$

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

Note that as shown in [20] the above definition of the Besov spaces B^s_{pq} is independent of the particular choice of Φ_0, Φ satisfying (5.14)–(5.15). For example with Ψ_j from the definition of the frame elements in Section 5.3, we have

$$\|f\|_{B^s_{pq}} \sim \Big(\sum_{j \ge 0} \big(b^{sj} \big\| \Psi_j(\sqrt{L}) f \big\|_p \big)^q \Big)^{1/q} \tag{5.17}$$

with the usual modification when $q = \infty$. The following useful inequality follows readily from above

$$\big\| \Psi_j(\sqrt{L}) f \big\|_p \le c b^{-sj} \|f\|_{B^s_{pq}}, \quad f \in B^s_{pq}, j \ge 0. \tag{5.18}$$

As in $\mathbb{R}^d$, we will need some embedding results involving Besov spaces. Recall the definition of embeddings: Let X and Y be two (quasi-)normed spaces. We say that X is continuously

embedded in Y and write $X \hookrightarrow Y$ if $X \subset Y$ and for each $f \in X$ we have $\|f\|_Y \le c\|f\|_X$, where $c > 0$ is a constant independent of f.

Proposition 5.4.

(i) *If* $1 \le q \le r \le \infty$, $0 < \tau \le \infty$, $s > 0$, *and* $\mu(\mathcal{M}) < \infty$, *then* $B^s_{r\tau} \hookrightarrow B^s_{q\tau}$.
(ii) *If* $1 \le r \le q \le \infty$, $0 < \tau \le \infty$ *and* $s > d(\frac{1}{r} - \frac{1}{q})$, *then* $B^s_{r\tau} \hookrightarrow B^{s-d(\frac{1}{r}-\frac{1}{q})}_{q\tau}$.
(iii) *If* $1 \le r \le \infty$, $0 < \tau \le \infty$, *and* $s > d/r$, *then* $B^s_{r\tau} \hookrightarrow \mathbb{L}^\infty$.
(iv) *If* $1 \le p \le r \le \infty$, $0 < \tau \le \infty$, $s > 0$, *and* $\mu(\mathcal{M}) < \infty$, *then* $B^s_{r\tau} \hookrightarrow \mathbb{L}^p$.

To streamline our presentation, we defer the proof of this proposition to the Supplementary Material [4].

Besov spaces in the Ahlfors regularity case. For the development of adaptive density estimators in Section 8, we will need some additional facts from the theory of Besov spaces when condition **C1A** is assumed. We first introduce the Besov bodies.

Definition 5.5. Assume $s > 0$, $1 \le p \le \infty$, $0 < q \le \infty$, and let $\mathcal{X} := \bigcup_{j\ge 0} \mathcal{X}_j$ be from the definition of the frames in Section 5.3. The Besov body $\mathfrak{b}^s_{pq} = \mathfrak{b}^s_{pq}(\mathcal{X})$ is defined as the set of all sequences $\{a_\xi\}_{\xi\in\mathcal{X}}$ of real (or complex) numbers such that

$$\|a\|_{\mathfrak{b}^s_{pq}} := \left(\sum_{j\ge 0} b^{jsq}\left(\sum_{\xi\in\mathcal{X}_j}\left[b^{-jd(\frac{1}{p}-\frac{1}{2})}|a_\xi|\right]^p\right)^{q/p}\right)^{1/q} < \infty, \tag{5.19}$$

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

One of the principle results in [20] asserts that the Besov spaces B^s_{pq} can be completely characterized in terms of the Besov bodies $\mathfrak{b}^s_{pq}$ of the frame coefficients of the respective functions. To be specific, denote

$$\beta_{j\xi}(f) := \langle f, \tilde{\psi}_{j\xi}\rangle, \quad \xi \in \mathcal{X}_j, j \ge 0. \tag{5.20}$$

We will also use the more compact notation: $\beta_\xi(f) := \beta_{j\xi}(f)$ for $\xi \in \mathcal{X}_j$. In the current setting, assume $s > 0$, $1 \le p \le \infty$, and $0 < q \le \infty$. In light of [20], Theorem 6.10, $f \in B^s_{pq}$ if and only if $\{\beta_\xi(f)\} \in \mathfrak{b}^s_{pq}$ with equivalent norms:

$$\|f\|_{B^s_{pq}} \sim \left\|\{\beta_\xi(f)\}\right\|_{\mathfrak{b}^s_{pq}}. \tag{5.21}$$

This implies that if $f \in B^s_{pq}$ for some $s > 0$, $p \ge 1$, and $0 < q \le \infty$, then

$$\left(\sum_{\xi\in\mathcal{X}_j}\left|\beta_{j\xi}(f)\right|^p\right)^{1/p} \le cb^{-j(s+d(\frac{1}{2}-\frac{1}{p}))}\|f\|_{B^s_{pq}}, \quad j \ge 0, \tag{5.22}$$

where $c = c(s, p, q) > 0$.

By (5.13) and (5.22) it follows that, if $f \in B^s_{pq}$ for some $s > 0$, $p \geq 1$, and $0 < q \leq \infty$, then

$$\Big\| \sum_{\xi \in \mathcal{X}_j} \beta_{j\xi}(f) \psi_{j\xi} \Big\|_p \leq c b^{-sj} \|f\|_{B^s_{pq}}, \quad j \geq 0. \tag{5.23}$$

We next assemble some additional facts we need about kernels in the setting of this article and then carry out the proof of Theorems 4.4 and 4.5.

5.4.1. *Spectral multiplier integral operators*

The operator $\Phi(\delta\sqrt{L})$ and its symmetric kernel $\Phi(\delta\sqrt{L})(x, y)$ from above have a number of useful properties that we describe and prove next.

(a) For any $k > d$ there exists a constant $c_k > 0$ such that

$$\big|\Phi(\delta\sqrt{L})(x, y)\big| \leq c(k)\big|B(x, \delta)\big|^{-1}\big(1 + \delta^{-1}\rho(x, y)\big)^{-k}, \quad x, y \in \mathcal{M}, 0 < \delta \leq 1, \tag{5.24}$$

where the constant $c(k) > 0$ depends only on k, Φ, and constant from the setting in Section 3. This inequality follows immediately from Theorem 4.1.

(b) For any $1 \leq p \leq \infty$

$$\big\|\Phi(\delta\sqrt{L})(x, \cdot)\big\|_p \leq c\big|B(x, \delta)\big|^{\frac{1}{p}-1} \leq c_\star \delta^{-d(1-\frac{1}{p})}, \quad x \in \mathcal{M}, 0 < \delta \leq 1, \tag{5.25}$$

where the constant $c_\star > 0$ is independent of p. This estimate follows readily by (5.24), (5.3), and (5.1).

(c) Let X be a random variable on $\mathcal{M}$ and $X \sim f(u)\, d\mu(u)$. Then

$$\mathbb{E}\big(\Phi(\delta\sqrt{L})(x, X)\big) = \int_{\mathcal{M}} \Phi(\delta\sqrt{L})(x, u) f(u)\, d\mu(u) = \Phi(\delta\sqrt{L}) f(x), \quad x \in \mathcal{M}. \tag{5.26}$$

This is a well known property of expected values.

We next estimate the bias term of the risk.

Proposition 5.6. *Let $s > 0$, $1 \leq p \leq \infty$, $0 < q \leq \infty$. If $f \in B^s_{pq}$, then $f \in \mathbb{L}^p$ and*

$$\big\|\Phi(\delta\sqrt{L}) f - f\big\|_p \leq c\delta^s \|f\|_{B^s_{pq}}, \quad 0 < \delta \leq 1, \tag{5.27}$$

where $c = c(s, p, q) > 0$.

This statement is quite standard. For completeness, we give its proof in the Supplementary Material [4].

We will also need the following two lemmas.

Lemma 5.7. *Let $2 \leq p < \infty$ and $0 < \delta \leq 1$. Then for any pdf f on $\mathcal{M}$ we have*

$$\left(\int_{\mathcal{M}} \int_{\mathcal{M}} \big|\Phi(\delta\sqrt{L})(x, u)\big|^p f(u)\, d\mu(u)\, d\mu(x)\right)^{1/p} \leq c_\star \delta^{-d(1-1/p)} \tag{5.28}$$

and

$$\left(\int_{\mathcal{M}}\left(\int_{\mathcal{M}}\left|\Phi(\delta\sqrt{L})(x,u)\right|^2 f(u)\,d\mu(u)\right)^{p/2} d\mu(x)\right)^{1/p} \le c_\star \delta^{-d/2}\|f\|_{p/2}^{1/2}, \tag{5.29}$$

where $c_\star > 0$ is the constant from (5.25); $c_\star$ is independent of p.

Lemma 5.8. *Let $1 \le p < 2$. Then there exists a constant $c = c(p) > 0$ such that for any $\delta > 0$ and any pdf f supported in a ball $B(x_0, R)$ with $x_0 \in \mathcal{M}$ and $R \ge \delta/2$ we have*

$$\left(\int_{\mathcal{M}}\left(\int_{\mathcal{M}}\left|\Phi(\delta\sqrt{L})(x,u)\right| f(u)\,d\mu(u)\right)^{p/2} d\mu(x)\right)^{\frac{1}{p}} \le c\left|B(x_0,R)\right|^{\frac{1}{p}-\frac{1}{2}}. \tag{5.30}$$

The proofs of these two lemmas are placed in the Supplementary Material [4].

6. Proof of Theorem 4.4 and Theorem 4.5

We will only prove Theorem 4.4. Theorem 4.5 follows readily.

By the triangle inequality we obtain the standard decomposition of the risk as the sum of stochastic and bias terms:

$$\mathbb{E}\|\widehat{\Phi}_\delta - f\|_p \le \mathbb{E}\left\|\widehat{\Phi}_\delta - \Phi(\delta\sqrt{L})f\right\|_p + \left\|\Phi(\delta\sqrt{L})f - f\right\|_p. \tag{6.1}$$

For the estimation of the bias term $\|\Phi(\delta\sqrt{L})f - f\|_p$ we will use estimate (5.27). We next focus on the estimation of the stochastic term $\mathbb{E}\|\widehat{\Phi}_\delta - \Phi(\delta\sqrt{L})f\|_p$. In the case $1 \le p < \infty$, using Jensen's inequality, we get

$$\begin{aligned}\mathbb{E}\left(\left\|\widehat{\Phi}_\delta - \Phi(\delta\sqrt{L})f\right\|_p\right) &\le \left(\mathbb{E}\left\|\widehat{\Phi}_\delta - \Phi(\delta\sqrt{L})f\right\|_p^p\right)^{\frac{1}{p}} \\ &= \left(\int_{\mathcal{M}}\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n \Phi(\delta\sqrt{L})(x,X_i) - \Phi(\delta\sqrt{L})f(x)\right|^p d\mu(x)\right)^{\frac{1}{p}}. \end{aligned}\tag{6.2}$$

(i) Assume the pdf $f \in B^s_{p\tau}(m)$ and let $X \sim X_i$. We first prove estimate (4.11) for $p = 2$. Clearly

$$\begin{aligned}\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n \Phi(\delta\sqrt{L})(x,X_i) - \Phi(\delta\sqrt{L})f(x)\right|^2 &\le \frac{1}{n}\mathbb{E}\left[\Phi(\delta\sqrt{L})(x,X)\right]^2 \\ &= \frac{1}{n}\int_{\mathcal{M}}\left|\Phi(\delta\sqrt{L})(x,u)\right|^2 f(u)\,d\mu(u).\end{aligned}$$

This coupled with (6.2) yields

$$\mathbb{E}\big\|\widehat{\Phi}_\delta - \Phi(\delta\sqrt{L})f\big\|_2 \leq \frac{1}{n^{1/2}}\left(\int_{\mathcal{M}}\int_{\mathcal{M}}\big|\Phi(\delta\sqrt{L})(x,u)\big|^2 f(u)\,d\mu(u)\,d\mu(x)\right)^{\frac{1}{2}} \leq \frac{c}{(n\delta^d)^{1/2}}, \tag{6.3}$$

where we used (5.28) with $p=2$. Combining (6.1), (5.27), and (6.3) we get

$$\mathbb{E}\|\widehat{\Phi}_\delta - f\|_2 \leq \frac{c}{(n\delta^d)^{1/2}} + cm\delta^s.$$

With $\delta = n^{-\frac{1}{2s+d}}$, that is, $\delta^s = \frac{1}{(n\delta^d)^{1/2}}$, this yields (4.11) when $p=2$.

Let $2 < p < \infty$. We will use the following version of Rosenthal's inequality that can be derived for instance from [17], page 245, inequality (C.5) with $\tau = \frac{p}{2}+1 \leq p+1$: If $Y_1,\dots,Y_n$ are i.i.d. random variables and $Y_i \sim Y$, then

$$\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n Y_i - \mathbb{E}Y\right|^p \leq \frac{(p+1)^p}{n^{p-1}}\mathbb{E}|Y|^p + \frac{p(p+1)^{p/2}e^{p/2+1}}{n^{p/2}}\big(\mathbb{E}|Y|^2\big)^{p/2}. \tag{6.4}$$

We get

$$\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n \Phi(\delta\sqrt{L})(x,X_i) - \Phi(\delta\sqrt{L})f(x)\right|^p \leq \frac{c}{n^{p-1}}\mathbb{E}\big|\Phi(\delta\sqrt{L})(x,X)\big|^p + \frac{c}{n^{p/2}}\big(\mathbb{E}\big|\Phi(\delta\sqrt{L})(x,X)\big|^2\big)^{p/2}.$$

This and (6.2) imply

$$\begin{aligned}\mathbb{E}\big\|\widehat{\Phi}_\delta - \Phi(\delta\sqrt{L})f\big\|_p &\leq \frac{c}{n^{1-1/p}}\left(\int_{\mathcal{M}}\int_{\mathcal{M}}\big|\Phi(\delta\sqrt{L})(x,u)\big|^p f(u)\,d\mu(u)\,d\mu(x)\right)^{1/p}\\ &\quad + \frac{c}{n^{1/2}}\left(\int_{\mathcal{M}}\left(\int_{\mathcal{M}}\big|\Phi(\delta\sqrt{L})(x,u)\big|^2 f(u)\,d\mu(u)\right)^{p/2} d\mu(x)\right)^{1/p}\\ &= \frac{c}{(n\delta^d)^{1/2}} + c\big(n\delta^d\big)^{-1}\|f\|_{p/2},\end{aligned} \tag{6.5}$$

where we used (5.28) and (5.28). Since $1 \leq \frac{p}{2} < p$ and $\|f\|_1 = 1$, we obtain by interpolation

$$\|f\|_{\frac{p}{2}} \leq \|f\|_1^{\frac{1}{p-1}}\|f\|_p^{\frac{p-2}{p-1}} = \|f\|_p^{\frac{p-2}{p-1}} \leq c\|f\|_{B^s_{p\tau}}^{\frac{p-2}{p-1}} \leq cm^{\frac{p-2}{p-1}}. \tag{6.6}$$

Here we also used Proposition 5.4(iv).

Combining (6.5)–(6.6) with (6.1) and (5.27), and taking into account that $\delta = n^{-\frac{1}{2s+d}}$, that is, $\delta^s = \frac{1}{(n\delta^d)^{1/2}}$ we arrive at

$$\mathbb{E}\|\widehat{\Phi}_\delta - f\|_p \le \frac{c}{(n\delta^d)^{1/2}} + cm\delta^s \le c' n^{-\frac{s}{2s+d}}. \tag{6.7}$$

The proof of part (i) of the theorem is complete.

(ii) Let $1 \le p < 2$ and $f \in B^s_{p\tau}(m, x_0, R)$. We use Jensen's inequality and the fact that $|\Phi(\delta\sqrt{L})(x,u)| \le c|B(x,\delta)|^{-1} \le c'\delta^{-d}$, using (5.24) and (5.1), to obtain

$$\begin{aligned}
&\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n \Phi(\delta\sqrt{L})(x, X_i) - \Phi(\delta\sqrt{L})f(x)\right|^p \\
&\quad\le \left(\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n \Phi(\delta\sqrt{L})(x, X_i) - \Phi(\delta\sqrt{L})f(x)\right|^2\right)^{\frac{p}{2}} \\
&\quad\le \frac{1}{n^{\frac{p}{2}}}\left(\mathbb{E}\left|\Phi(\delta\sqrt{L})(x, X)\right|^2\right)^{\frac{p}{2}} \\
&\quad= \frac{1}{n^{\frac{p}{2}}}\left(\int_{\mathcal{M}} \left|\Phi(\delta\sqrt{L})(x,u)\right|^2 f(u)\, d\mu(u)\right)^{\frac{p}{2}} \\
&\quad\le \frac{c}{(n\delta^d)^{\frac{p}{2}}}\left(\int_{\mathcal{M}} \left|\Phi(\delta\sqrt{L})(x,u)\right| f(u)\, d\mu(u)\right)^{\frac{p}{2}}.
\end{aligned}$$

This and (6.2) lead to

$$\mathbb{E}\left\|\widehat{\Phi}_\delta - \Phi(\delta\sqrt{L})f\right\|_p \le \frac{c}{(n\delta^d)^{\frac{1}{2}}}\left(\int_{\mathcal{M}}\left(\int_{\mathcal{M}} \left|\Phi(\delta\sqrt{L})(x,u)\right| f(u)\, d\mu(u)\right)^{\frac{p}{2}} d\mu(x)\right)^{\frac{1}{p}}.$$

We now invoke Lemma 5.8 to obtain

$$\mathbb{E}\left\|\widehat{\Phi}_\delta - \Phi(\delta\sqrt{L})f\right\|_p \le \frac{c}{(n\delta^d)^{\frac{1}{2}}}\left|B(x_0, R)\right|^{\frac{1}{p}-\frac{1}{2}} \le \frac{c'}{(n\delta^d)^{\frac{1}{2}}}.$$

Using this and (5.27) we complete the proof of (4.12) just as above in (6.7).

(iii) Assume the pdf $f \in B^s_{\infty\tau}(m)$ and let $q > 2$ be arbitrary. Since by construction $\operatorname{supp}\Phi \subset [-1, 1]$, the function $\widehat{\Phi}_\delta(x) - \Phi(\delta\sqrt{L})f(x)$ belongs to the spectral space $\Sigma_{1/\delta}$. Then by Proposition 5.1

$$\left\|\widehat{\Phi}_\delta - \Phi(\delta\sqrt{L})f\right\|_\infty \le c_\star \delta^{-\frac{d}{q}}\left\|\widehat{\Phi}_\delta - \Phi(\delta\sqrt{L})f\right\|_q,$$

where the constant $c_\star > 1$ is independent of q. This along with Jensen's inequality and Fubini's theorem lead to

$$\mathbb{E}\big\|\widehat{\Phi}_\delta - \Phi(\delta\sqrt{L})f\big\|_\infty \le c_\star \delta^{-\frac{d}{q}} \left(\int_{\mathcal{M}} \mathbb{E}\big|\hat{\Phi}_\delta(x,\cdot) - \Phi(\delta\sqrt{L})f(x)\big|^q \, d\mu(x)\right)^{\frac{1}{q}}. \tag{6.8}$$

We now apply Rosenthal's inequality (6.4) to obtain

$$\begin{aligned}
&\mathbb{E}\big|\widehat{\Phi}_\delta(x) - \Phi(\delta\sqrt{L})f(x)\big|^q \\
&\quad\le \frac{(q+1)^q}{n^{q-1}} \mathbb{E}\big|\Phi(\delta\sqrt{L})(x,X)\big|^q + \frac{q(q+1)^{\frac{q}{2}} e^{\frac{q}{2}+1}}{n^{\frac{q}{2}}} \Big(\mathbb{E}\big|\Phi(\delta\sqrt{L})(x,X)\big|^2\Big)^{\frac{q}{2}} \\
&\quad= \frac{(q+1)^q}{n^{q-1}} \int_{\mathcal{M}} \big|\Phi(\delta\sqrt{L})(x,u)\big|^q f(u)\, d\mu(u) \\
&\qquad+ \frac{q(q+1)^{\frac{q}{2}} e^{\frac{q}{2}+1}}{n^{\frac{q}{2}}} \left(\int_{\mathcal{M}} \big|\Phi(\delta\sqrt{L})(x,u)\big|^2 f(u)\, d\mu(u)\right)^{\frac{q}{2}}.
\end{aligned}$$

This coupled with (6.8) and the fact that $1/q < 1$ imply

$$\begin{aligned}
&\mathbb{E}\big\|\hat{\Phi}_\delta - \Phi(\delta\sqrt{L})f\big\|_\infty \\
&\quad\le c_\star \delta^{-\frac{d}{q}} \frac{q+1}{n^{1-\frac{1}{q}}} \left(\int_{\mathcal{M}}\int_{\mathcal{M}} \big|\Phi(\delta\sqrt{L})(x,u)\big|^q f(u)\, d\mu(u)\, d\mu(x)\right)^{\frac{1}{q}} \\
&\qquad+ c_\star \delta^{-\frac{d}{q}} \frac{q^{\frac{1}{q}}(q+1)^{\frac{1}{2}} e^{\frac{1}{2}+\frac{1}{q}}}{n^{\frac{1}{2}}} \left(\int_{\mathcal{M}} \left(\int_{\mathcal{M}} \big|\Phi(\delta\sqrt{L})(x,u)\big|^2 f(u)\, d\mu(u)\right)^{\frac{q}{2}} d\mu(x)\right)^{\frac{1}{q}} \\
&\quad\le c_\star \delta^{-\frac{d}{q}} \left(\frac{2c_\star q}{(n\delta^d)^{1-\frac{1}{q}}} + \frac{e^2 c_\star q^{1/2}}{(n\delta^d)^{1/2}} \|f\|_{q/2}^{1/2}\right),
\end{aligned} \tag{6.9}$$

where we used (5.28), (5.29), and the inequality $q^{\frac{1}{q}}(q+1)^{\frac{1}{2}} e^{\frac{1}{2}+\frac{1}{q}} \le e^2 q^{1/2}$, $(q > 2)$. Observe that the constant $c_\star$ above is from (5.25) and is independent of q.

By Proposition 5.4(iii) it follows that $f \in \mathbb{L}^\infty$ and since $\|f\|_1 = 1$ we obtain

$$\|f\|_{q/2} \le \|f\|_\infty^{1-2/q} \|f\|_1 \le \big(c\|f\|_{B^s_{\infty\tau}}\big)^{1-2/q} \le (cm)^{1-2/q} \le cm + 1.$$

Let $n \ge e^2$ and choose $q := \log n$. By assumption $\delta = (\frac{\log n}{n})^{1/(2s+d)}$. Now, it is easy to see that $n^{1/q} = e$, $\delta^{-d/q} \le e$, $\delta^s = \frac{q^{1/2}}{(n\delta^d)^{1/2}} = (\frac{\log n}{n})^{s/(2s+d)}$, and

$$\frac{q}{(n\delta^d)^{1-1/q}} \le \frac{q}{(n\delta^d)^{3/4}} \le \frac{\log n}{n^{\frac{3s/2}{2s+d}}} \le c\left(\frac{\log n}{n}\right)^{s/(2s+d)} \quad \text{if } n \ge e^2.$$

Putting all of the above together, we obtain

$$\mathbb{E}\big\|\hat{\Phi}_\delta - \Phi(\delta\sqrt{L})f\big\|_\infty \le c\left(\frac{\log n}{n}\right)^{s/(2s+d)}. \tag{6.10}$$

If $2 \le n < e^2$, then estimate (6.10) follows readily from (6.9) with $q = 2$.

As before we use (6.10) and (5.27) to obtain (4.13). The proof of Theorem 4.5 is complete.

A closer examination of the above proof shows that the oracle inequalities from Theorem 4.4 are valid.

7. Linear wavelet density estimators

In this section, we establish $\mathbb{L}^p$-error estimates for linear wavelet density estimators. Let $\{\psi_{j\xi}\}$, $\{\tilde{\psi}_{j\xi}\}$ be the pair of dual frames described in Section 5.3. We adhere to the notation from Section 5.3.

For any $j \ge 0$ and $\xi \in \mathcal{X}_j$ we define the *empirical coefficient estimators* by

$$\hat{\beta}_{j\xi} := \frac{1}{n}\sum_{i=1}^{n} \tilde{\psi}_{j\xi}(X_i). \tag{7.1}$$

Using this, we define the *linear wavelet density estimator* by

$$f^*(x) = \sum_{j=0}^{J}\sum_{\xi\in\mathcal{X}_j} \hat{\beta}_{j\xi}\psi_{j\xi}(x), \quad x \in \mathcal{M}, \tag{7.2}$$

where the parameter $J = J(n) \in \mathbb{N}$ is selected so that the factor b^{-J} de facto behaves as a bandwidth. More precisely, we define J as the unique positive integer such that

$$b^J \le n^{1/(2s+d)} < b^{J+1}. \tag{7.3}$$

It is easy to see that f^* can be written in the following way

$$\begin{aligned} f^*(x) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{J}\sum_{\xi\in\mathcal{X}_j} \psi_{j\xi}(x)\tilde{\psi}_{j\xi}(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{J} \Psi_j(\sqrt{L})(X_i, x) = \frac{1}{n}\sum_{i=1}^{n} \Psi_0\big(b^{-J}\sqrt{L}\big)(X_i, x). \end{aligned} \tag{7.4}$$

where we used (5.8) and (5.5).

Thus, this linear wavelet estimator is in fact a particular case of the linear estimators investigated in the previous subsection. This enables us to state the following upper bound theorem, which is an immediate consequence of Theorem 4.5.

Theorem 7.1. *Let $s > 0$, $0 < \tau \le \infty$, $m > 0$, $x_0 \in \mathcal{M}$ and $R > 0$.*

(i) *If $2 \le p < \infty$ and J is as in (7.3), then*

$$\sup_{f \in B^s_{p\tau}(m)} \mathbb{E}\left\| f^* - f \right\|_p \le cn^{-s/(2s+d)}, \tag{7.5}$$

where $c = c(p, \tau, s, m) > 0$.

(ii) *If $1 \le p < 2$ and J is as in (7.3), then*

$$\sup_{f \in B^s_{p\tau}(m, x_0, R)} \mathbb{E}\left\| f^* - f \right\|_p \le cn^{-s/(2s+d)}, \tag{7.6}$$

where $c = c(p, \tau, s, m, x_0, R) > 0$.

(iii) *If J is the unique integer satisfying $b^J \le (\frac{n}{\log n})^{1/(2s+d)} < b^{J+1}$, then*

$$\sup_{f \in B^s_{\infty\tau}(m)} \mathbb{E}\left\| f^* - f \right\|_\infty \le c\left(\frac{\log n}{n}\right)^{\frac{s}{2s+d}}, \tag{7.7}$$

where $c = c(\tau, s, m) > 0$.

8. Adaptive wavelet density estimators

As we want to parallel in our setting the main results in density estimation theory, say, on $[0, 1]^d$, we need to introduce adaptation, that is, to obtain up to logarithmic factors optimal rates of convergence without knowing the regularity. There are several techniques for this. For example, Lepski's method (see [15,27]) could be applied to our kernel estimators.

We choose to develop here nonlinear wavelet estimators, where we apply *hard thresholding*. This method has been developed in the classical case of $\mathbb{R}$ in [8] and on the sphere in [17]. We will operate in the general setting described in Section 3. Unlike the case of the kernel or linear wavelet density estimates considered in the previous section, here we assume that the space $\mathcal{M}$ is compact ($\mu(\mathcal{M}) < \infty$) and all conditions **C1**–**C5** (including the Ahlfors regularity condition **C1A**) are satisfied, see Section 3.

As before we assume that $X_1, \dots, X_n$ ($n \ge 2$) are i.i.d. random variables with values on $\mathcal{M}$ and with a common density function f with respect to the measure μ on $\mathcal{M}$. Let $X_j \sim X$. We denote by $\mathbb{E} = \mathbb{E}_f$ the expectation with respect to the probability measure $\mathbb{P} = \mathbb{P}_f$. In addition, we assume here that f is bounded. Denote

$$A := \max\left\{\|f\|_\infty, 4\right\} \quad \text{and set} \quad \kappa := c_\diamond (8A)^{1/2}, \tag{8.1}$$

where $c_\diamond > 1$ is the constant from the norm bounds of the frame elements in (5.12).

We will utilize the pair of frames $\{\psi_{j\xi}\}$, $\{\tilde{\psi}_{j\xi}\}$ described in Section 5.3. We adhere to the notation from Section 5.3. Recall that any $f \in \mathbb{L}^p(\mathcal{M}, d\mu)$ has the frame decomposition

$$f = \sum_{j=0}^{\infty} \sum_{\xi \in \mathcal{X}_j} \beta_{j\xi}(f)\psi_{j\xi}, \quad \beta_{j\xi}(f) := \langle f, \tilde{\psi}_{j\xi} \rangle \ \big(\text{convergence in } \mathbb{L}^p\big). \tag{8.2}$$

Assuming the pdf f fixed, we will use the abbreviated notation $\beta_{j\xi} := \beta_{j\xi}(f)$.

We introduce two parameters depending on n:

$$\lambda_n := \kappa \left(\frac{\log n}{n} \right)^{1/2} \tag{8.3}$$

and J_n uniquely defined by the following inequalities

$$b^{J_n} \leq \left(\frac{n}{\log n} \right)^{1/d} < b^{J_n+1}. \tag{8.4}$$

As in Section 7, we introduce the *empirical coefficient estimators*

$$\hat{\beta}_{j\xi} := \frac{1}{n} \sum_{i=1}^{n} \tilde{\psi}_{j\xi}(X_i), \quad j \geq 0, \xi \in \mathcal{X}_j. \tag{8.5}$$

We now define the *hard threshold coefficient estimators* by

$$\hat{\beta}^*_{j\xi} := \hat{\beta}_{j\xi} \mathbb{I}_{\{|\hat{\beta}_{j\xi}| > 2\lambda_n\}}, \quad j \geq 0, \xi \in \mathcal{X}_j. \tag{8.6}$$

Then the *wavelet threshold density estimator* is defined by

$$\hat{f}_n(x) := \sum_{0 \leq j \leq J_n} \sum_{\xi \in \mathcal{X}_j} \hat{\beta}^*_{j\xi} \psi_{j\xi}(x), \quad x \in \mathcal{M}. \tag{8.7}$$

Remark 8.1. Note that the density estimator $\hat{f}_n$ of the pdf f depends only on the number n of observations, the geometric constant $c_\diamond$, and the $\mathbb{L}^\infty$-norm of f.

We now state our main result on the adaptive wavelet threshold estimator defined above.

Theorem 8.2. *Let $1 \leq r \leq \infty$, $0 < \tau \leq \infty$, $1 \leq p < \infty$, $s > d/r$, and $m > 0$. Then there exists a constant $c = c(r, \tau, p, s, m) > 0$ such that in the setting described above and with $\hat{f}_n$ from (8.7) we have*:

(i)
$$\sup_{f \in B^s_{r\tau}(m)} \mathbb{E} \| \hat{f}_n - f \|_\infty \leq c \left(\frac{\log n}{n} \right)^{\frac{s - \frac{d}{r}}{2[s - d(\frac{1}{r} - \frac{1}{2})]}}. \tag{8.8}$$

(ii) *In the regular case $s \geq \frac{dp}{2}(\frac{1}{r} - \frac{1}{p})$*

$$\sup_{f \in B^s_{r\tau}(m)} \mathbb{E} \| \hat{f}_n - f \|_p \leq c \log n \left(\frac{\log n}{n} \right)^{\frac{s}{2s+d}}. \tag{8.9}$$

(iii) *In the sparse case* $s < \frac{dp}{2}(\frac{1}{r} - \frac{1}{p})$

$$\sup_{f \in B^s_{r\tau}(m)} \mathbb{E}\|\hat{f}_n - f\|_p \leq c \log n \left(\frac{\log n}{n}\right)^{\frac{s-d(\frac{1}{r}-\frac{1}{p})}{2[s-d(\frac{1}{r}-\frac{1}{2})]}}. \tag{8.10}$$

The proof of this theorem is quite long and involved. We place it in the Supplementary Material [4].

Remark 8.3. Several observations are in order:

(a) The assumption $s > d/r$ leads to $\|f\|_\infty \leq c\|f\|_{B^s_{r\tau}} \leq cm$, by Proposition 5.4(iii). In addition in the sparse case it implies $p > 2$.

(b) The geometry of the setting is represented by the dimension d. Note that the exponents of $\frac{\log n}{n}$ are the same as in the case of the sphere [1].

(c) In the regular case (modulo the logarithmic terms) we have the same rate of convergence $n^{-s/(2s+d)}$ as in the case of the linear wavelet estimator.

(d) Just as in the case of kernel density estimators (see Remark 4.6) we note that since we assume here all conditions **C1–C5** (including **C1A**) it would not be a problem to obtain lower bounds matching up to logarithmic terms the rates established above by a direct adaptation of the proof of the lower bounds in the case of the sphere from [1].

Acknowledgements

GK and DP have been supported by the ANR BASICS, ANR-17-CE40-0001. PP has been supported by NSF Grant DMS-1714369. The authors are grateful to an anonymous referee for very helpful remarks on an earlier version of this paper.

Supplementary Material

Supplement to "Kernel and wavelet density estimators on manifolds and more general metric spaces" (DOI: 10.3150/19-BEJ1171SUPP; .pdf). We present several examples of settings that are covered by the general framework from Section 3. The proofs of several claims from Section 5 as well as the proof of Theorem 8.2 are also given in the Supplementary Material.

References

[1] Baldi, P., Kerkyacharian, G., Marinucci, D. and Picard, D. (2009). Adaptive density estimation for directional data using needlets. *Ann. Statist.* **37** 3362–3395. MR2549563 https://doi.org/10.1214/09-AOS682

[2] Bhattacharya, A. and Bhattacharya, R. (2012). *Nonparametric Inference on Manifolds*: *With Applications to Shape Spaces*. *Institute of Mathematical Statistics* (*IMS*) *Monographs* **2**. Cambridge: Cambridge Univ. Press. MR2934285 https://doi.org/10.1017/CBO9781139094764

[3] Castillo, I., Kerkyacharian, G. and Picard, D. (2014). Thomas Bayes' walk on manifolds. *Probab. Theory Related Fields* **158** 665–710. MR3176362 https://doi.org/10.1007/s00440-013-0493-0

[4] Cleanthous, G., Georgiadis, A.G., Kerkyacharian, G., Petrushev, P. and Picard, D. (2020). Supplement to "Kernel and wavelet density estimators on manifolds and more general metric spaces." https://doi.org/10.3150/19-BEJ1171SUPP.

[5] Coifman, R.R. and Weiss, G. (1971). *Analyse Harmonique Non-commutative sur Certains Espaces Homogènes*: *Étude de certaines intégrales singulières*. *Lecture Notes in Mathematics* **242**. Berlin: Springer. MR0499948

[6] Coulhon, T., Kerkyacharian, G. and Petrushev, P. (2012). Heat kernel generated frames in the setting of Dirichlet spaces. *J. Fourier Anal. Appl.* **18** 995–1066. MR2970038 https://doi.org/10.1007/s00041-012-9232-7

[7] Dai, F. and Xu, Y. (2013). *Approximation Theory and Harmonic Analysis on Spheres and Balls*. *Springer Monographs in Mathematics*. New York: Springer. MR3060033 https://doi.org/10.1007/978-1-4614-6660-4

[8] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24** 508–539. MR1394974 https://doi.org/10.1214/aos/1032894451

[9] Faraut, J. (2008). *Analysis on Lie Groups*: *An Introduction*. *Cambridge Studies in Advanced Mathematics* **110**. Cambridge: Cambridge Univ. Press. MR2426516 https://doi.org/10.1017/CBO9780511755170

[10] Folland, G.B. (1999). *Real Analysis*: *Modern Techniques and Their Applications*, 2nd ed. *Pure and Applied Mathematics* (*New York*). New York: Wiley. A Wiley-Interscience Publication. MR1681462

[11] Frazier, M. and Jawerth, B. (1985). Decomposition of Besov spaces. *Indiana Univ. Math. J.* **34** 777–799. MR0808825 https://doi.org/10.1512/iumj.1985.34.34041

[12] Frazier, M. and Jawerth, B. (1990). A discrete transform and decompositions of distribution spaces. *J. Funct. Anal.* **93** 34–170. MR1070037 https://doi.org/10.1016/0022-1236(90)90137-A

[13] Frazier, M., Jawerth, B. and Weiss, G. (1991). *Littlewood–Paley Theory and the Study of Function Spaces*. *CBMS Regional Conference Series in Mathematics* **79**. Providence, RI: Amer. Math. Soc. MR1107300 https://doi.org/10.1090/cbms/079

[14] Georgiadis, A.G., Kerkyacharian, G., Kyriazis, G. and Petrushev, P. (2017). Homogeneous Besov and Triebel–Lizorkin spaces associated to non-negative self-adjoint operators. *J. Math. Anal. Appl.* **449** 1382–1412. MR3601596 https://doi.org/10.1016/j.jmaa.2016.12.049

[15] Goldenshluger, A. and Lepski, O. (2014). On adaptive minimax density estimation on R^d. *Probab. Theory Related Fields* **159** 479–543. MR3230001 https://doi.org/10.1007/s00440-013-0512-1

[16] Grigor'yan, A. (2009). *Heat Kernel and Analysis on Manifolds*. *AMS/IP Studies in Advanced Mathematics* **47**. Providence, RI: Amer. Math. Soc. MR2569498

[17] Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998). *Wavelets*, *Approximation*, *and Statistical Applications*. *Lecture Notes in Statistics* **129**. New York: Springer. MR1618204 https://doi.org/10.1007/978-1-4612-2222-4

[18] Johnstone, I.M. and Silverman, B.W. (1990). Speed of estimation in positron emission tomography and related inverse problems. *Ann. Statist.* **18** 251–280. MR1041393 https://doi.org/10.1214/aos/1176347500

[19] Juditsky, A. and Lambert-Lacroix, S. (2004). On minimax density estimation on $\mathbb{R}$. *Bernoulli* **10** 187–220. MR2046772 https://doi.org/10.3150/bj/1082380217

[20] Kerkyacharian, G. and Petrushev, P. (2015). Heat kernel based decomposition of spaces of distributions in the framework of Dirichlet spaces. *Trans. Amer. Math. Soc.* **367** 121–189. MR3271256 https://doi.org/10.1090/S0002-9947-2014-05993-X

[21] Kerkyacharian, G., Petrushev, P., Picard, D. and Willer, T. (2007). Needlet algorithms for estimation in inverse problems. *Electron. J. Stat.* **1** 30–76. MR2312145 https://doi.org/10.1214/07-EJS014

[22] Kerkyacharian, G., Petrushev, P. and Xu, Y. (2020). Gaussian bounds for the weighted heat kernels on the interval, ball and simplex. *Constr. Approx*. **51** 73–122.
[23] Kerkyacharian, G., Petrushev, P. and Xu, Y. (2020). Gaussian bounds for the heat kernels on the ball and the simplex: Classical approach. *Studia Math*. **250** 235–252. MR4034745 https://doi.org/10.4064/sm180423-13-10
[24] Kerkyacharian, G., Pham Ngoc, T.M. and Picard, D. (2011). Localized spherical deconvolution. *Ann. Statist*. **39** 1042–1068. MR2816347 https://doi.org/10.1214/10-AOS858
[25] Kerkyacharian, G. and Picard, D. (1992). Density estimation in Besov spaces. *Statist. Probab. Lett*. **13** 15–24. MR1147634 https://doi.org/10.1016/0167-7152(92)90231-S
[26] Lepski, O.V., Mammen, E. and Spokoiny, V.G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist*. **25** 929–947. MR1447734 https://doi.org/10.1214/aos/1069362731
[27] Lepskiĭ, O.V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatn. Primen*. **36** 645–659. MR1147167 https://doi.org/10.1137/1136085
[28] Pelletier, B. (2005). Kernel density estimation on Riemannian manifolds. *Statist. Probab. Lett*. **73** 297–304. MR2179289 https://doi.org/10.1016/j.spl.2005.04.004
[29] Pelletier, B. (2006). Non-parametric regression estimation on closed Riemannian manifolds. *J. Nonparametr. Stat*. **18** 57–67. MR2214065 https://doi.org/10.1080/10485250500504828
[30] Pollard, D. (1984). *Convergence of Stochastic Processes*. *Springer Series in Statistics*. New York: Springer. MR0762984 https://doi.org/10.1007/978-1-4612-5254-2
[31] Sandryhaila, A. and Moura, J.M.F. (2013). Discrete signal processing on graphs. *IEEE Trans. Signal Process*. **61** 1644–1656. MR3038378 https://doi.org/10.1109/TSP.2013.2238935
[32] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. *Monographs on Statistics and Applied Probability*. London: CRC Press. MR0848134 https://doi.org/10.1007/978-1-4899-3324-9
[33] Starck, J.-L., Murtagh, F. and Fadili, J.M. (2010). *Sparse Image and Signal Processing*: *Wavelets*, *Curvelets*, *Morphological Diversity*. Cambridge: Cambridge Univ. Press. MR2643260 https://doi.org/10.1017/CBO9780511730344
[34] Stein, E.M. and Weiss, G. (1971). *Introduction to Fourier Analysis on Euclidean Spaces*. *Princeton Mathematical Series* **32**. Princeton, NJ: Princeton Univ. Press. MR0304972
[35] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. New York: Springer. Revised and extended from the 2004 French original, translated by Vladimir Zaiats. MR2724359 https://doi.org/10.1007/b13794
[36] von Luxburg, U. (2007). A tutorial on spectral clustering. *Stat. Comput*. **17** 395–416. MR2409803 https://doi.org/10.1007/s11222-007-9033-z
[37] Yosida, K. (1980). *Functional Analysis*, 6th ed. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **123**. Berlin: Springer. MR0617913

Received March 2019 and revised August 2019