Maynooth University
National University of Ireland Maynooth

An introduction to kernel density estimation on Euclidean and more general metric spaces

Adam Furlong

Thesis presented in partial fulfillment of the requirements of the degree of Master of Science in Theoretical Physics and Mathematics to the Maynooth University Department of Mathematics and Statistics.

Supervisor: Dr. Galatia Cleanthous

Department Head: Prof. Stephen Buckley

August 11, 2025

Abstract

Kernel density estimation, a method to produce estimators of an unknown probability density f, is introduced over the real numbers. Some classical upper bounds on the error of these estimators are derived, assuming f lies in some regularity space. Recent work in the field has developed kernel density estimation over a broad class of metric spaces. We consider a measure metric space satisfying a volume doubling condition, which admits a non-negative self-adjoint operator whose heat kernels enjoy Gaussian regularity. This framework includes many natural spaces such as Euclidean space, spheres, balls, and a wide class of Riemannian manifolds. In this setting, analagous upper bound to those on Euclidean space are derived.

Contents

1 Introduction

In the modern world, science has become largely based on statistical methods, due to the endless stream of data provided by ever-improving technology. In astronomy, new telescopes and techniques are detecting a growing number of astrophysical phenomena, such as new galaxies, supernovae and black hole mergers. In earth science, both ground-based and orbital sensors monitor global events such as earthquakes and weather patterns. Medicine uses scans, such as magnetic resonance imaging and x-rays, to study the interior of our bodies.

Many datasets can be understood as many independent realisations of a random variable X, which is distributed according to an unknown probability density f over some geometric space $\mathcal{M}$. The distances to newly discovered galaxies are distributed over the positive reals. Other data, such as the direction to a detected supernova and the locations of earthquakes, are distributed over the sphere. In other fields such as signal processing and medicine, the data may be interpreted over more obscure geometric spaces.

This leads to the problem of *density estimation*. Given a random sample $(X_1, \cdots, X_n)$ of a random variable with pdf f, we wish to produce an estimator of f, that is a measurable function $\hat{f} : \mathcal{M}^n \times \mathcal{M} \to \mathbb{R}$, which depends on the data and the variable upon which f depends.

One widely used method for tackling this problem is *kernel density estimation*. This was suggested first by Rosenblatt (1956) [13] and developed further by Parzen (1962) [12]. One of the first works in attaining optimal convergence rates of these estimators was done by Bretagnolle and Huber (1979) [1], under the assumption that $f : \mathbb{R} \to \mathbb{R}$ lies in a Sobolev space. Work over more general regularity spaces has also been performed, for example Besov spaces over $\mathbb{R}^n$ in Kerkyacharian and Picard (1992) [10]. A thorough introduction to the subject can be found in the now standard textbook Tsybakov (2009) [14].

However, many data sets are not distributed over Euclidean space, and each time a new setting is introduced, the entire machinery of kernel density estimation must be built again for that specific space. The field would benefit greatly from unifying the approach over a broad class of geometric spaces. Work in this direction has been undertaken by Coulhon et al. (2012) [5] and Kerkyacharian et al. (2015) [11], allowing us to consider kernel density estimation over many metric measure spaces equipped with a Laplacian-like operator. This modern framework hosts many examples which were of established interest, such as Euclidean space, spheres, balls, and intervals, and includes a great many more, such as all Riemannian manifolds of non-negative Ricci curvature, equipped with their Laplace or Laplace-Beltrami operator.

Building on this work, upper bounds on the convergence rates of these kernel density estimators over Hölder and Sobolev regularity spaces have been derived in Cleanthous et al. (2022) [3] and Cleanthous et al. (2025) [4] respectively. These rates match those known to be optimal over Euclidean space and the sphere. One of the goals of the thesis is to build up to these results.

We now present a roadmap for this thesis. The next two sections will follow the outline of the first chapter of [14]. In Section 2, we introduce the procedure of kernel density estimation on $\mathbb{R}$. Kernels, bandwidths and kernel density estimators are discussed, as well as the risks used to measure the accuracy of the estimators. Section 3 is dedicated to deriving some classical results on the convergence rates of kernel density estimators on $\mathbb{R}$, specifically when f lies in a Hölder, Nikol'skii or Sobolev regularity space.

We then move to the modern setting, preparing to perform kernel density estimation on a wide range of geometric spaces. In Section 4, the geometric aspect is explored: we assume a metric measure space $(\mathcal{M}, \rho, \mu)$ which has some minimal geometric structure, such as a volume-doubling condition. Section 5 is dedicated to outlining a minimal background in the powerful spectral theory necessary. There, we assume the space admits an essentially self-adjoint operator L, to be thought of as similar to the Laplacian of $\mathbb{R}^n$, with Gaussian-like heat kernels. At the end of these sections, we discuss some examples of spaces satisfying the assumptions.

The background of the modern setting builds up to Section 6, where we will then outline recent developments on such spaces from [3] and [4]. This will rely on the Assumptions I and II from Sections 4 and 5 respectively, as well as machinery built in [5] and [11]. Upper bounds analogous to each of those presented in Section 3 are obtained, and so these sections share the same structure.

Notation: We denote by $\mathbb{N}$, $\mathbb{R}$, $\mathbb{R}_+$ the sets of positive integers, reals and non-negative reals respectively. If $\tau \in \mathbb{N}$, the class of differentiable functions on $\mathbb{R}_+$ with continuous derivatives up to order τ will be stated as $C^\tau(\mathbb{R}_+)$. For $s > 0$, we will denote by $\lfloor s \rfloor$ the greatest integer strictly less than s.

2 Kernel density estimation

Let $(X_1, \cdots, X_n)$ be independent random variables identically distributed on $\mathbb{R}$ according to an unknown probability density function (pdf) $f : \mathbb{R} \to [0, \infty)$. Our goal is to construct some estimator $\hat{f}_n$ of f, a measurable function from $\mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$, based on the n observations. If we were to assume f could be described by finitely many parameters, such as a Gaussian being described by its mean and variance, this would simplify to choosing those parameters to best fit the observation. Here, we do not assume that f takes such a form, but rather that f belongs to some vast regularity space such as a Sobolev space. Kernel density estimation is one approach to this problem without relying on a parameterisation.

2.1 Kernels and bandwidths

In Figure 1, we plot an example probability density f and some data sampled from it. We see that the areas where the data is concentrated typically correspond to peaks of f, and hope that these concentrations will help us to estimate the pdf.

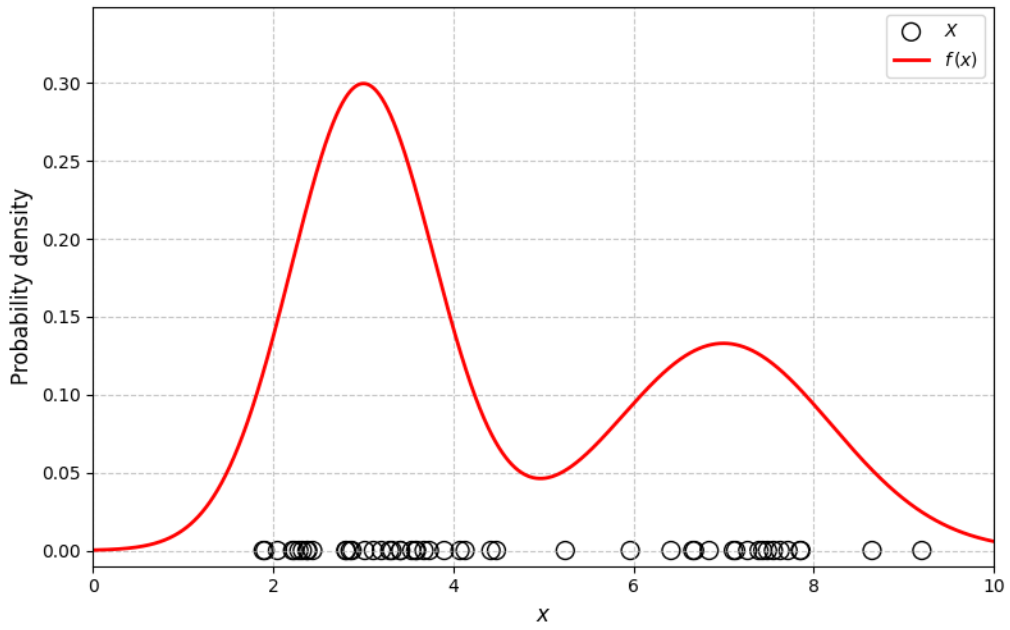

Figure 1: We use $f = 0.6\mathcal{N}(3, 0.8) + 0.4\mathcal{N}(7, 1.2)$ as an example probability density function, where $\mathcal{N}(\mu, \sigma)$ is a normal distribution of mean μ and standard deviation σ. The black circles are $n = 50$ points sampled from f, referred to as the data X. These are seen to be most concentrated around $x = 3$ and $x = 7$, corresponding to the peaks of f.

Given a probability density f, the corresponding cumulative distribution function (cdf) is defined as

$$F(x) := \int_{-\infty}^{x} f(t)\,\mathrm{d}t.$$

This measures how much of the mass of the pdf lies to the left of each $x \in \mathbb{R}$. Only having access to the data, we can estimate the cdf as the fraction of the observations X_i that lie to the left of x. Making use of the indicator function $I(\cdot)$, which is one when its argument is true and vanishes otherwise, this estimator can be written as

$$\hat{F}_n(x) := \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x).$$

By the strong law of large numbers, we have $\hat{F}_n(x) \xrightarrow{n\to\infty} F(x)$ almost surely for every $x \in \mathbb{R}$, so this is a consistent estimator of the cdf. In Figure 2 we plot a cdf F and its estimator $\hat{F}_n$.

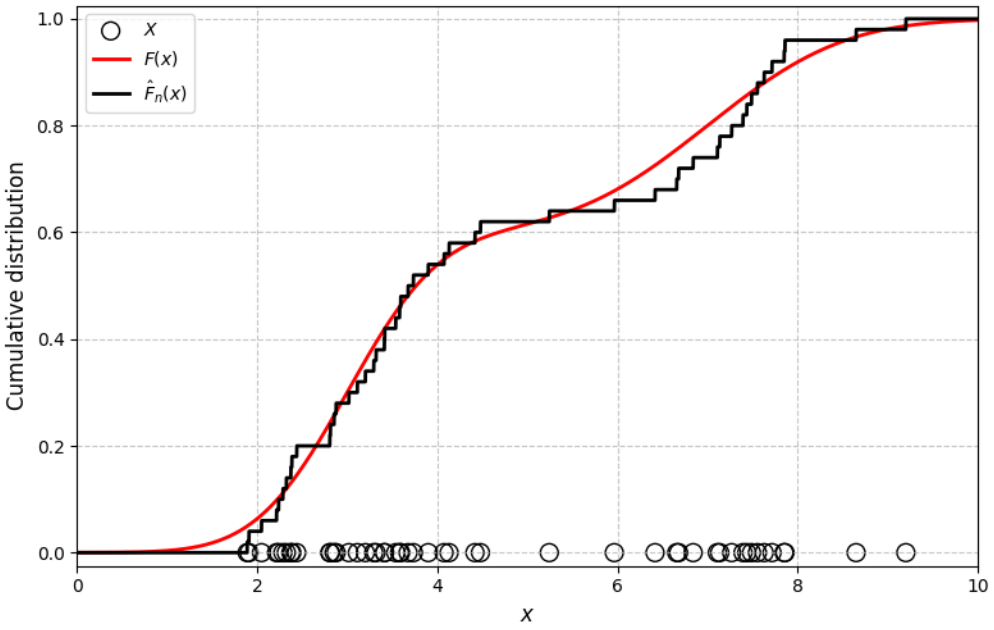

Figure 2: Using the same example pdf f and data X from Figure 1, we plot the cdf F and its estimator $\hat{F}_n$. The estimator is a step function, increasing by $1/n$ at each point X_i in the dataset.

Now we use this to estimate the pdf. Clearly $f(x) = F'(x)$, by the Fundamental Theorem of Calculus. So, one obvious approach is to estimate f by the derivative of the estimator $\hat{F}_n$. However, $\hat{F}_n$ is a step-function; the derivative is zero everywhere it exists. In its place, we use a symmetric discrete derivative. Choosing some $h > 0$, which is called the **bandwidth**, the Rosenblatt estimator is defined as

$$\hat{f}^R_{n,h}(x) := \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h}.$$

This estimator is plotted in Figure 3. The choice of bandwidth is important and we will return to it.

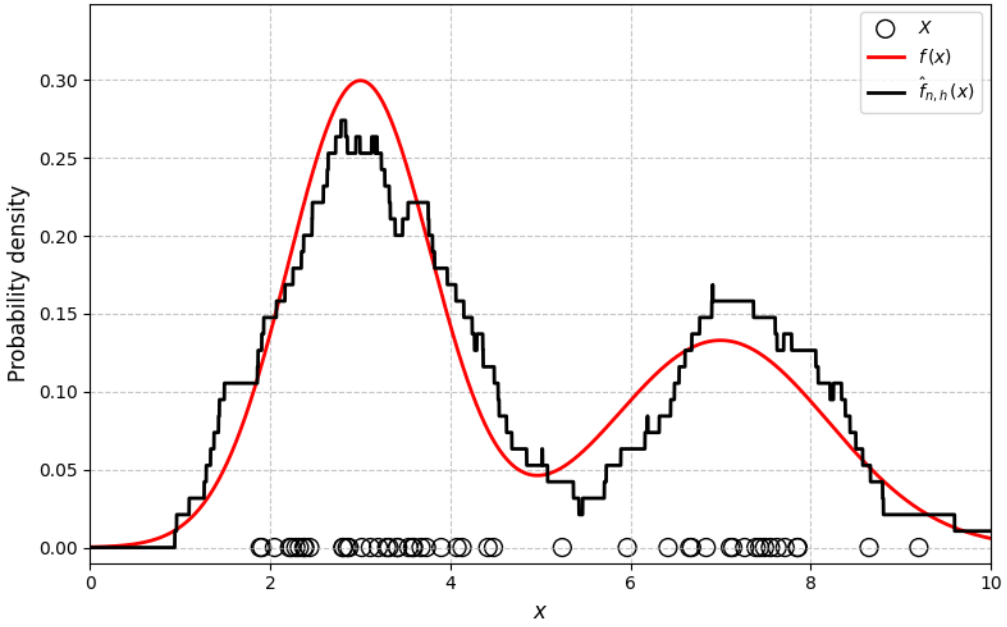

Figure 3: Using the example pdf f and data X from Figure 1, we plot the Rosenblatt estimator $\hat{f}_{n,h}$ against f. The bandwidth used is $h = 0.95$. The estimator is seen to mimic the structure of the pdf, exhibiting two distinct peaks.

Let us now rewrite this estimator to gain some intuition for how it works. Notice first that

$$I(X_i \leq x+h) - I(X_i \leq x-h) = I(x-h \leq X_i \leq x+h) = I\left(\frac{|X_i - x|}{h} \leq 1\right).$$

This allows us to express the estimator as

$$\hat{f}^R_{n,h}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x),$$

where we have defined $K(u) := \frac{1}{2}I(|u| \leq 1)$ and $K_h(u) := h^{-1}K(u/h)$. We call K the rectangular kernel, and it is plotted in Figure 4. K_h can be understood as an adaptation of K which is scaled horizontally (interior $1/h$) and vertically (exterior $1/h$), with $\int K_h(u)\,\mathrm{d}u = \int K(u)\,\mathrm{d}u = 1$. The Rosenblatt estimator can then be understood as placing a copy of the function K_h centred at each observation X_i, and then taking the average of these n functions. Examining the plot of K, we may interpret this as attributing a region of high probability around each data-point X_i. Hopefully this seems an intuitive approach to estimating the probability density.

Of course, there are many ways to distribute this probability. Any function $K : \mathbb{R} \to \mathbb{R}$ such that $\int K(u)\,\mathrm{d}u = 1$ can be used, and is called a **kernel**. Then, for some chice of bandwidth $h > 0$, we again denote $K_h(u) = h^{-1}K(u/h)$ and define the associated **kernel density estimator** (kde) as

$$\hat{f}_{n,h}(x) := \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x).$$

This is often referred to as a Parzen-Rosenblatt estimator, named after Rosenblatt who suggested the method in 1956 [13], and Parzen who developed it further in 1962 [12]. The word kernel comes from German, meaning "core" or "the most important part". This name is fitting as the kernel K contains all the information of the kde.

It can be noticed by the substitution $u = (X_i - x)/h$ that $\int \hat{f}_{n,h}(x)\,\mathrm{d}x = \int K_h(X_i - x)\,\mathrm{d}x = \int K(u)\,\mathrm{d}u$. Thus the benefit of K integrating to unity is that $\hat{f}_{n,h}$ also does. Then, if the kernel is chosen to be non-negative, the kde is also a probability density. However, we do not require this to be the case. This is discussed further in Section 2.2, along with further assumptions on the kernels. Typically, the kernel is chosen to be symmetric and localised about the origin. Some examples of commonly used kernels are:

1. the rectangular kernel, $K(u) = \frac{1}{2}I(|u| \leq 1)$,
2. the triangular kernel, $K(u) = (|u| - 1)I(|u| \leq 1)$,
3. the Epanechnikov kernel, $K(u) = \frac{3}{4}(1 - u^2)I(|u| \leq 1)$,
4. the biweight kernel, $K(u) = \frac{15}{16}(1 - u^2)^2 I(|u| \leq 1)$,
5. the Gaussian kernel kernel, $K(u) = \frac{1}{\sqrt{2\pi}}\exp(-u^2/2)$,
6. and fourth-order kernels, such as $K(u) = \frac{3}{8}(3 - 5u^2)I(|u| \leq 1)$.

These kernels are plotted in Figure 4. All but the gaussian are compactly supported on $[-1, 1]$, and all but the fourth order kernel are non-negative. The Epanechnikov kernel is a cutoff parabola, and the biweight is similar but differentiable at -1 and 1.

Let us now return to the concept of the bandwidth. It can be seen that K_h becomes more localised as h decreases. For example, with the Guassian kernel K, the bandwidth is precisely the standard deviation of K_h. Then, as h decreases, we expect K_h to become sharply peaked about the origin. The estimator $\hat{f}_{n,h}$ will inherit these peaks about each observation X_i. On the other hand, a large value of h will cause the various copies of K_h to overlap and blur the estimator. These effects are illustrated in Figure 5.

It is up to us to choose some h within these extremes. The best value for h may depend on n. That is, we will need the bandwidth to not be a value, but a sequence $h = h_n$. The size and rate of decay of this sequence is to be chosen by us to minimise the error of the kernel density estimator.

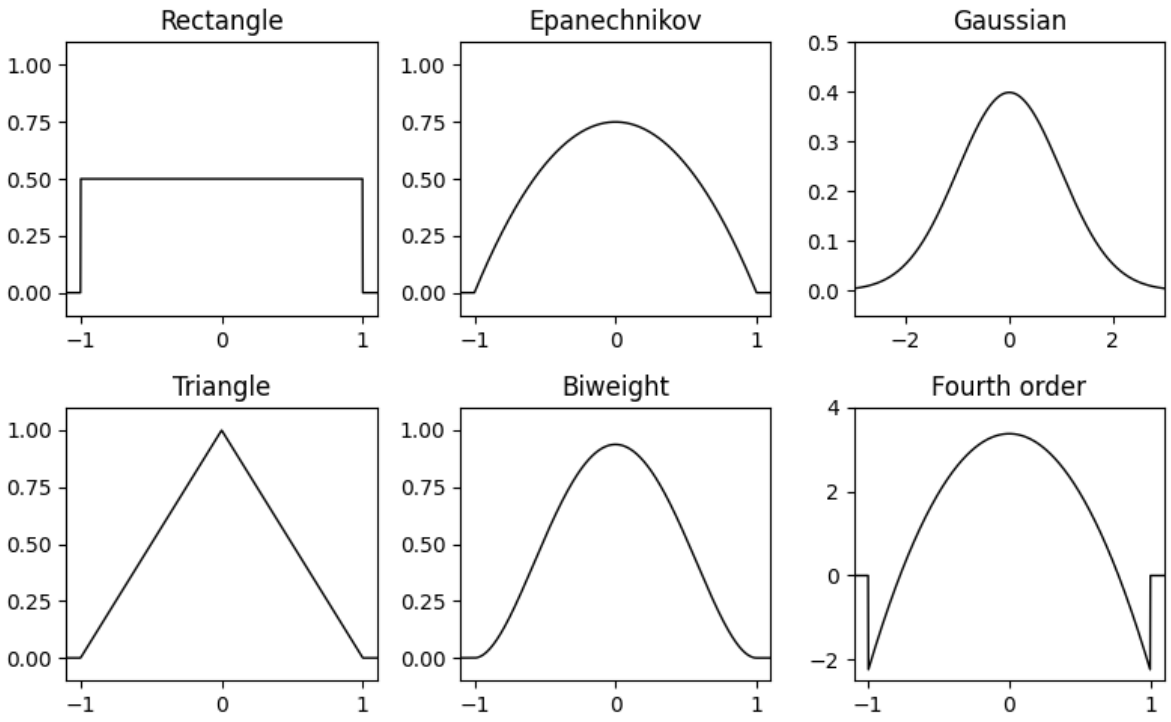

Figure 4: Plots of some common kernels $K(u)$ over u.

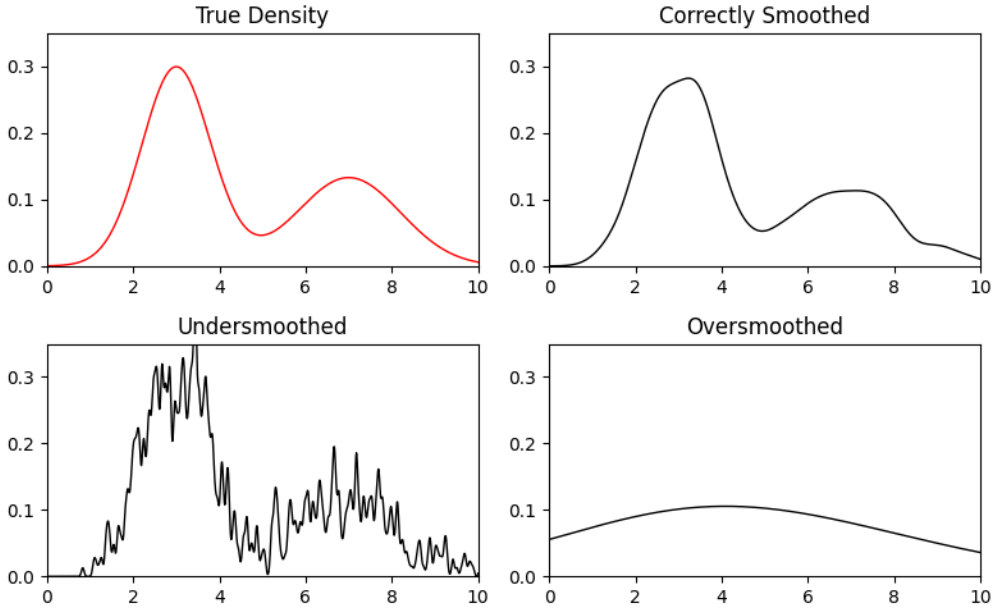

Figure 5: $n = 1000$ points are sampled from the probability density f from Figure 1. Using the Gaussian kernel, we plot the kernel density estimator $\hat{f}_{n,h}(x)$ over $x \in [0, 10]$ for 3 values of the bandwidth h. *Top left:* The pdf $f(x)$ is plotted for reference. *Top right:* $h = 0.3$. Displays two distinct peaks, in good agreement with the true density. *Bottom left:* $h = 0.03$. Noisy and overfit to the data, peaking sharply around observations. *Bottom right:* $h = 3$. Too spread out, and blind to the local structure of f.

2.2 Risks

By choosing different kernels K and bandwidths h, a wide variety of kernel density estimators $\hat{f}_{n,h}$ can be constructed. Some are clearly better than others at estimating the probability density f. In this section we will collect some ways to quantify the accuracy of an estimator $\hat{f}$. We drop the subscripts n, h to avoid clutter, and because these are true for any estimators $\hat{f}$ of f.

Before proceeding to define these risks, we lay down some information on integration. In this section and the next, all integrals are performed with respect to the Lebesgue measure λ, and we write it as usual $\int g(x)\,\mathrm{d}x$. Integrals without a specified domain are understood to be over the entirety of $\mathbb{R}$.

Again we let X_i, $1 \leq i \leq n$, be iid random variables sampled from a probability density function f. Suppose $g : \mathbb{R} \to \mathbb{R}$ is a function that may depend on each X_i. The **expected value** of g at $x \in \mathbb{R}$, written $\mathbb{E}[g(x)]$, is defined as

$$\mathbb{E}\left[g(x)\right] := \int \cdots \int g(x, x_1, \cdots, x_n) \left(\prod_{i=1}^{n} f(x_i)\,\mathrm{d}x_i \right). \tag{2.1}$$

This has all the properties we know of the expectation value: it is linear, if g is a function that is independent of each X_i then it obeys $\mathbb{E}[g(x)] = g(x)$, and in particular $\mathbb{E}\left[\mathbb{E}\left[g(x)\right]\right] = \mathbb{E}\left[g(x)\right]$ for every g.

We will make use of a very important class of functions, the L^p spaces. To define these, we first need the p-norms $\|\cdot\|_p$. Let $g : \mathbb{R} \to \mathbb{C}$ be a measurable function. For $1 \leq p < \infty$ the p-norms are defined as

$$\|g\|_p := \left(\int |g(x)|^p \,\mathrm{d}x \right)^{1/p}.$$

In the case $p = \infty$, it is defined as the least upper bound of $|g|$,

$$\|g\|_\infty := \inf \left\{ M \in \mathbb{R}_+ \mid |g(x)| \leq M \text{ for } \lambda\text{-almost every } x \in \mathbb{R} \right\}.$$

Then, for $1 \leq p \leq \infty$ the L^p spaces are defined as

$$L^p = L^p(\mathbb{R}, \lambda) := \left\{ g : \mathbb{R} \to \mathbb{C} \mid g \text{ measurable and } \|g\|_p < \infty \right\} / \sim,$$

where $\sim$ is the equivalence relation $h \sim g$ when $h(x) = g(x)$ for λ-almost every $x \in \mathbb{R}$. $\|\cdot\|_p$ is a norm on L^p, called the p-norm of g, and with it L^p is a Banach space. We can now define our risks.

Definition 2.1. *Suppose $\hat{f}$ is an estimator of f. Then we define*

(i) the Mean Squared Error

$$\mathrm{MSE}(x) := \mathbb{E}\left[(\hat{f}(x) - f(x))^2 \right]; \tag{2.2}$$

(ii) the Mean Integrated Square Error

$$\mathrm{MISE} := \mathbb{E} \int (\hat{f}(x) - f(x))^2 \,\mathrm{d}x; \tag{2.3}$$

(iii) and, for $1 \leq p < \infty$, the L^p Risk

$$\mathcal{R}_p := \mathbb{E} \left\| \hat{f} - f \right\|_p^p. \tag{2.4}$$

Look for a moment at the MISE, and we can find the relationship between these risks. First, notice the integrand is positive. Then, Tonelli's theorem tells us we can swap the order of integration,

$$\mathrm{MISE} = \int \mathbb{E}\left[(\hat{f}(x) - f(x))^2 \right] \mathrm{d}x = \int \mathrm{MSE}(x)\,\mathrm{d}x. \tag{2.5}$$

Because they allow us to swap integrals as above, Tonelli's theorem will be among our most frequently used tools. Secondly, we can see that the L^p risk is just the generalisation of the Mean Integrated Squared Error to other values of p,

$$\mathrm{MISE} = \mathbb{E} \int (\hat{f}(x) - f(x))^2 \,\mathrm{d}x = \mathbb{E} \left\| \hat{f} - f \right\|_2^2 = \mathcal{R}_2. \tag{2.6}$$

The Mean Squared Error is a pointwise measurement of the error, and is used to study the Hölder spaces. On the other hand, the L^p risks are global measurements, and will be employed to study subsets of the L^p space, specifically the Nikol'skii and Sobolev spaces. However, working on all L^p spaces proves cumbersome, especially in the range $1 \leq p < 2$. We shall focus on the special case $p = 2$. There, the L^2 risk will be used, which we saw is the Mean Integrated Squared Error.

We are now ready to outline our goal. Let $\mathcal{R}$ be one of the risks defined above. As n grows to infinity, we would expect our estimator $\hat{f}_{n,h}$ to better approximate f, and we hope the risk $\mathcal{R}$ would approach zero. Of course, without having access to f, we cannot compute the $\mathcal{R}$, though we can bound it. In particular, we hope there are constants $R > 0$, called the **convergence rate**, and $c > 0$ independent of n such that $\mathcal{R} \leq cn^{-R}$ for every $n \in \mathbb{N}$.

In our problem, we suppose that f lies in some regularity space $\mathbb{F}$. It is to be expected that some choices of kernel K and bandwidths h lead to faster convergence. Our goal is to find some conditions on the kernel and bandwidth to obtain some convergence rate *over the regularity space* $\mathbb{F}$. That is, to maximise the $R > 0$ such that there exists a constant $c > 0$, that may depend on K, h and the nature of $\mathbb{F}$, such that for every $n \in \mathbb{N}$

$$\sup_{f \in \mathbb{F}} \mathcal{R} \leq cn^{-R}.$$

This allows us to bound the risk without knowing what f is, only knowing what space it belongs to. In this work, we are not at all interested in the value of c, and will not take much care to optimise its value. We are only interested in attaining as fast a rate R as possible.

2.3 Bias and variance

We have established that our aim is to bound the risks defined in the previous section. In this section we define the bias and variance of the kernel density estimator, and use them to decompose the MSE and MISE into two terms. This is the most common method to estimate these errors. Unfortunately, the L^p risks do not have such a nice decomposition, and will be dealt with separately when the time comes.

Definition 2.2. *Suppose $\hat{f}$ is an estimator of f. Then we define*

(i) the ***bias*** $b(x)$

$$b(x) := \mathbb{E}\left[\hat{f}(x)\right] - f(x),$$

(ii) and the ***variance*** $\sigma^2(x)$

$$\sigma^2(x) := \mathbb{E}\left[\left(\hat{f}(x) - \mathbb{E}\left[\hat{f}(x)\right]\right)^2\right].$$

Typically in kernel density estimation, as the bandwidth h decreases, the bias decreases while the variance increase. This is known as the *bias-variance trade-off*. Estimators with large bias, variance are then oversmooothed, undersmoothed respectively. We shall see that the bias depends heavily on the regularity enjoyed by f, and the variance not so much.

The Mean Squared Error and Mean Integrated Squared Error have very well known decompositions in terms of the bias and variance. Because of this, every proof on bounds of the MSE and MISE will be split into two parts: estimating the bias, and estimating the variance.

Proposition 2.3. *Suppose $\hat{f}$ is an estimator of f. Then*

$$\text{MSE} = \sigma^2 + b^2, \text{ and} \tag{2.7}$$

$$\text{MISE} = \int \sigma^2(x)\,\mathrm{d}x + \int b^2(x)\,\mathrm{d}x. \tag{2.8}$$

Proof. Insert $\mathbb{E}\hat{f} - \mathbb{E}\hat{f}$ into $(\hat{f} - f)^2$ to split it into three terms. Note that $\mathbb{E}\hat{f} - f$ is independent of each X_i, and $\mathbb{E}\left[\hat{f} - \mathbb{E}\hat{f}\right] = 0$, so when we take the expectation, the central term vanishes, leaving

$$\text{MSE} = \mathbb{E}\left[(\hat{f} - f)^2\right] = \mathbb{E}\left[\left(\hat{f} - \mathbb{E}\hat{f}\right)^2\right] + 2\mathbb{E}\left[\hat{f} - \mathbb{E}\hat{f}\right]\left(\mathbb{E}\hat{f} - f\right) + \left(\mathbb{E}\hat{f} - f\right)^2 = \sigma^2 + b^2.$$

The expression for the Mean Integrated Squared Error follows from this and (2.5). □

3 Classical results on $\mathbb{R}$

The purpose of this section is to derive upper bounds on the Mean Squared Error and Mean Integrated Squared Error under certain regularity assumptions. Specifically, we follow Chapter 1 of [14], aiming to prove Theorems 1.1, 1.2 and 1.3. Some results under the L^p risks will also be stated.

3.1 Regularity spaces of interest

We have discussed that we do not wish to assume the probability density f has any particular form, but rather we assume f is in some sense "smooth". In our study, we will focus on three examples of regularity spaces: Hölder, Nikol'skii and Sobolev. Each has a parameter $s > 0$. We take $\ell = \lfloor s \rfloor$ to be the greatest integer strictly less than s. In particular, if $s \in \mathbb{N}$, then $\ell = \lfloor s \rfloor = s - 1$.

Definition 3.1. *Let $s > 0$, $\ell = \lfloor s \rfloor$, and $1 \leq p < \infty$. An ℓ-times differentiable function $f : \mathbb{R} \to \mathbb{R}$ belongs to*

(i) the Hölder space, $\mathcal{H}^s$, if

$$\|f\|_{\mathcal{H}^s} := \|f\|_\infty + \sup_{x \neq y} \frac{\left|f^{(\ell)}(y) - f^{(\ell)}(x)\right|}{|y - x|^{s-\ell}} < \infty; \tag{3.1}$$

(ii) the Nikol'skii space, $\mathcal{N}_p^s$, if

$$\|f\|_{\mathcal{N}_p^s} := \|f\|_p + \sup_{t \in \mathbb{R}} \frac{\left[\int \left|f^{(\ell)}(x+t) - f^{(\ell)}(x)\right|^p \, \mathrm{d}x\right]^{1/p}}{|t|^{s-\ell}} < \infty; \tag{3.2}$$

(iii) and for $s \in \mathbb{N}$, the Sobolev space, $\mathcal{W}_p^s$, if $f^{(\ell)}$ is absolutely continuous and

$$\|f\|_{\mathcal{W}_p^s} := \|f\|_p + \left\|f^{(s)}\right\|_p < \infty. \tag{3.3}$$

The Hölder condition is a mostly a local one, affecting the deviation of the function in a neighbourhood around a point, though it does require the function to be globally bounded. In the case that s is an integer, the finiteness of the second term is a Lipschitz condition on $f^{(\ell)}$. Any smooth probability density, such as our example f from Figure 1, is a member of $\mathcal{H}^s$ for every $s > 0$.

The Nikol'skii and Sobolev conditions are global, concerning integrals over the whole space. In the case that s is an integer, we have $\mathcal{W}_p^s \subset \mathcal{N}_p^s$. This fact is proved in Lemma 3.12, and will be used to prove the bounds on the Sobolev space once they are proved them for the Nikol'skii case. Such an inclusion is not as obvious when we move to more general geometric spaces, where we will have to treat these spaces separately.

Now let us speak of the above three cases simultaneously. To do this, let $\mathbb{F}$ among $\mathcal{H}^s, \mathcal{W}_p^s, \mathcal{N}_p^s$. In each case, $\|\cdot\|_{\mathbb{F}}$ is a norm on $\mathbb{F}$. In fact, $(\mathbb{F}, \|\cdot\|_{\mathbb{F}})$ is a Banach space. We will be interested in members of these spaces that are probability distributions with a bounded norm, so we define

$$\mathcal{P}_m(\mathbb{F}) := \left\{ f \in \mathbb{F} \;\middle|\; f \geq 0, \; \int f(x) \, \mathrm{d}x = 1, \; \|f\|_{\mathbb{F}} \leq m \right\}$$

for each $m > 0$. It is over these balls that we take wish to take the supremum, i.e. if $\mathcal{R}$ is the risk, we wish to find $R > 0$ such that for some constant $c > 0$, which may depend on $\mathbb{F}$ and m,

$$\sup_{f \in \mathcal{P}_m(\mathbb{F})} \mathcal{R} \leq c n^{-R}$$

for every $n \in \mathbb{N}$. In any of our results of this form, it is assumed that $\mathcal{P}_m(\mathbb{F})$ is non-empty. This is not always the case; $\mathcal{P}_m(\mathcal{W}_1^s)$ is empty for every $m < 1$, as the first term is $\|f\|_1 = 1$.

Before proceeding, note that the definitions above are not those used in [14], which will be explained now. In each case, the second term of $\|\cdot\|_{\mathbb{F}}$, which we will denote by $\|\cdot\|_{\dot{\mathbb{F}}}$, is a semi-norm on $\mathbb{F}$. It can

be shown that if f is a probability density and $\|\cdot\|_{\dot{\mathbb{F}}} < \infty$, then also $\|\cdot\|_{\mathbb{F}} < \infty$. For example, in the proof of Theorem 1.1 in [14], it is proven that for every $s > 0$ there are constants $a, b > 0$ such that $\|f\|_\infty \le a \|f\|_{\dot{\mathcal{H}}^s} + b$. That is to say, the subsets of $\mathbb{F}$ and $\dot{\mathbb{F}}$ which are probability densities coincide. As we are only interested in probability densities f, this change does not impose any additional constraints. Introducing this first term makes it cleaner to bound the variance. The definitions given above are the inhomogenous versions of the spaces, whereas using only the seminorm gives the homogenous spaces $\dot{\mathbb{F}}$. We will not mention $\|\cdot\|_{\dot{\mathbb{F}}}$ and $\dot{\mathbb{F}}$ will not again, just know that the bias terms are bounded in terms of $\|\cdot\|_{\dot{\mathbb{F}}}$, and not $\|\cdot\|_{\mathbb{F}}$ as stated.

3.2 Conditions on the kernel

One of the questions motivating our study is this: what properties of the kernel K lead to good convergence rates? The answer is largely foreshadowed by the following definition, which will be justified by its use in bounding the bias (see Lemma 3.4).

Definition 3.2. *Let $s > 0$ and $\ell = \lfloor s \rfloor$. We say that $K : \mathbb{R} \to \mathbb{R}$ is a **kernel of order s** if for each integer $1 \le j \le \ell$ the following integrals exist and*

$$\int K(u)\,\mathrm{d}u = 1, \qquad \int u^j K(u)\,\mathrm{d}u = 0, \qquad C_1(s,K) := \frac{1}{\ell!}\int |u|^s\,|K(u)|\,\mathrm{d}u < \infty \tag{3.4}$$

We shall see that kernels of higher order attain faster convergence rates, assuming f is sufficiently smooth. It can be shown that if $0 < t < s$ and K is bounded kernel of order s, then K is also a kernel of order t. In situations where we have an integral of the form $\int K(u)g(u)\,\mathrm{d}u$, and g is many times differentiable, assuming that K is a kernel of order s will allow us to kill all terms up to order ℓ in the Taylor expansion of g.

The first of the above conditions was already assumed to ensure the estimator integrates to one. The last condition corresponds to a sufficiently fast decay of K. It is the vanishing moments condition that may be difficult to satisfy. If K is even, that is $K(-u) = K(u)$, then all of the odd moments vanish. It can then be seen that every kernel in Figure 4 is of order 2.

The even moments require some care. We now follow the discussion in section 1.2.2 of [14] to show that such kernels may be constructed from orthonormal bases of polynomials.

Proposition 3.3. *Let $\{\phi_m\}_{m=0}^\infty$ be a family of polynomials, where ϕ_m is a polynomial of degree m, that are orthonormal with respect to a positive weight function $w : \mathbb{R} \to \mathbb{R}_+$. That is,*

$$\int w(u)\,\mathrm{d}u = 1 \qquad \text{and} \qquad \int \phi_m(u)\phi_k(u)w(u)\,\mathrm{d}u = \delta_{mk}$$

for all non-negative integers m, k. Let $s > 0$ and $\ell = \lfloor s \rfloor$. Then the following is a kernel of order s:

$$K(u) = \sum_{m=0}^{\ell} \phi_m(0)\phi_m(u)w(u).$$

Proof. First notice that $\phi_0(u) = \phi_0(0) \neq 0$ for every $u \in \mathbb{R}$. Then, by the orthonormality condition,

$$\int K(u)\,\mathrm{d}u = \int \phi_0(u)\phi_m(u)w(u)\,\mathrm{d}u + \sum_{m=1}^{\ell} \frac{\phi_m(0)}{\phi_0(0)} \int \phi_0(u)\phi_m(u)w(u)\,\mathrm{d}u = 1.$$

Next, let $1 \le j \le \ell$. As ϕ_n is a polynomial of degree n, there exist coefficients $c_{qj} \in \mathbb{R}$ such that $u^j = \sum_{q=0}^{j} c_{qj}\phi_j(u)$ for every $u \in \mathbb{R}$. This can be used to show the necessary moments vanish:

$$\int u^j K(u)\,\mathrm{d}u = \sum_{q=0}^{j}\sum_{m=0}^{\ell} c_{qj}\phi_m(0) \int \phi_q(u)\phi_m(u)w(u)\,\mathrm{d}u = \sum_{q=0}^{j} c_{qj}\phi_q(0) = 0^j = 0.$$

Finally, it must be checked that $C_1(s,K) < \infty$, but this is guaranteed by the necessarily fast decay of the weight w. □

There are many such families of polynomials, for example the Legendre polynomials with weight $w(u) = \frac{1}{2}I(|u| \leq 1)$, and the Hermite polynomials with weight $w(u) = \exp(-x^2/2)/\sqrt{2\pi}$. Notice that these weights correspond to the rectangular and Gaussian kernels respectively. This construction then has another interpretation: we take a kernel of lower order, and scale it by some polynomial, choosing the coefficients to cause the higher moments to vanish. The example of a fourth order kernel in Figure 4 is made by scaling the rectangular kernel by a quadratic polynomial. In Figure 6, we plot two more examples of order 4 kernels, based on the Gaussian and Epanechnikov kernels.

A kernel K constructed in this way inherits some nice properties from the weight function w: K is bounded if w is, K shares the support of w, K possesses as many derivatives as w, and K is even if w is. The last is true as the family of polynomials obeys $\phi_m(u) = (-1)^m\phi_m(u)$ when w is even. In particular, $\phi_m(0) = 0$ for every odd m, and the kernel is then only constructed from every second polynomial in the family. This is the case for the above polynomial families, and below kernels.

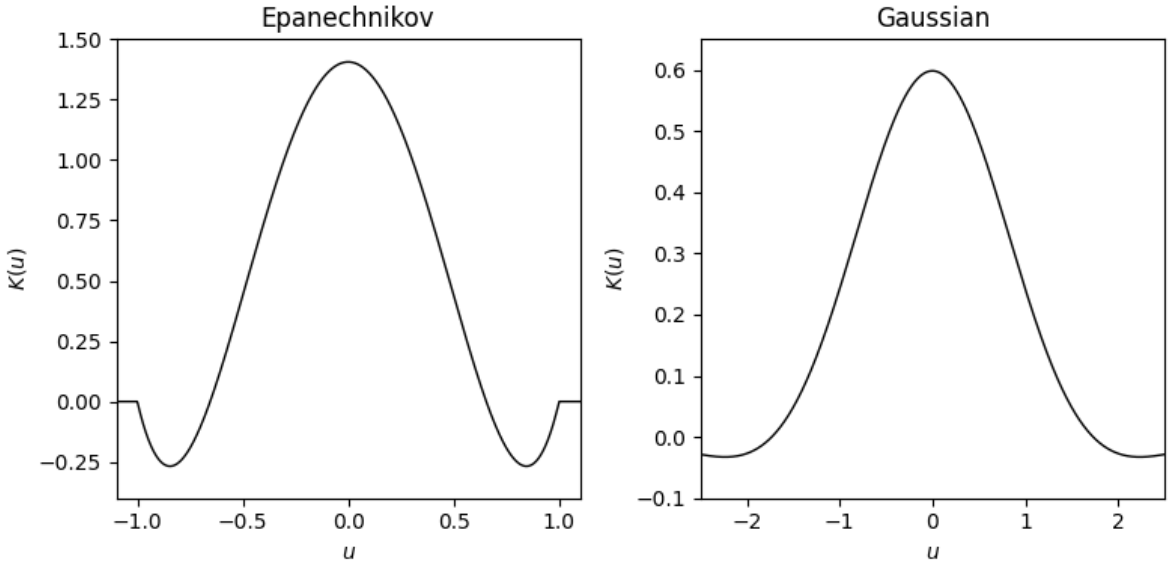

Figure 6: Examples of fourth order kernels are plotted. *Left:* $K(u) = \frac{15}{8}\left(1 - \frac{7}{3}u^2\right)K_E(u)$, where K_E is the Epanechnikov kernel. *Right:* $K(u) = \frac{1}{2}(3 - u^2)K_G(u)$, where K_G is the Gaussian kernel. Both kernels display taller central lobes than their order 2 counterparts, but also possess negative lobes further from the origin.

It can be seen that kernels of order $s > 2$ must be negative on a set of positive Lebesgue measure. This is because in order for $\int u^2K(u)\,\mathrm{d}u$ to vanish and K be non-negative almost everywhere, K must vanish almost everywhere, and so cannot integrate to 1. Thus, the kernel density estimator $\hat{f}_{n,h}$ is not guaranteed to be everywhere non-negative. For this reason, when using a kernel of order $s > 2$, $\hat{f}_{n,h}$ must not be interpreted as a probability density, only as a function that approximates the probability density f. This sacrifice is one we are willing to make — our interest is in the rate of convergence.

However, many authors take issue with the use of estimators which may be negative, and some insist that the estimator should itself be a probability density (see for example [6]). The simplest way to guarantee this is to use only kernels of second order, but this will cap the convergence rate of the estimator. Another solution is to use higher order kernels and modify the estimator to force it to be non-negative, for example $\hat{f}^*_{n,h}(x)$ either $\max(0, \hat{f}_{n,h}(x))$ or $|\hat{f}_{n,h}(x)|$. Clearly $|\hat{f}^*_{n,h}(x) - f(x)| \leq |\hat{f}_{n,h}(x) - f(x)|$, and so all the risks mentioned (MSE, MISE, $\mathcal{R}_p$) decrease by using this switch. In particular, every bound on the error and convergence rate to come will hold true for $\hat{f}^*_{n,h}$. However, it may be the case $\int \hat{f}^*_{n,h}(x)\,\mathrm{d}x > 1$, and so $\hat{f}^*_{n,h}$ is also not a probability density, so must be renormalised, typically by numerical means. Under certain conditions, such an estimator may still attain the faster convergence rates (see [9] for some numerical examples).

Throughout this section, we will be dealing with integrals of the kernel. For aesthetic reasons, we make use of the linear substitution $u = (y - x)/h$. Here we list some that will be used later. The first is used to bound the variance,

$$\mathbb{E}\left[K_h(X_i - x)^2\right] = \int K_h(y - x)^2 f(y)\,\mathrm{d}y = \frac{1}{h}\int K(u)^2 f(x + uh)\,\mathrm{d}u. \tag{3.5}$$

The next rewrites the expected value of the estimator, which is used when bounding the bias,

$$\mathbb{E}\left[\hat{f}_{n,h}(x)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[K_h(X_i - x)\right] = \int K(u) f(x + uh)\,\mathrm{d}u. \tag{3.6}$$

Of course, the reverse substitution, $y = x + uh$, tells us that the following integral is still unity,

$$\int f(x + uh)\,\mathrm{d}u = \int f(y)\,\mathrm{d}y = 1. \tag{3.7}$$

3.3 Decomposing the bias

The bias term relies very heavily on the regularity enjoyed by f. As such, it will be mainly dealt with separately in each case. However, our proofs in this chapter will rely on the same method: using kernels of a certain order and Taylor's theorem to express the bias in terms of derivatives of f. This method is expressed in the form of the next lemma.

Lemma 3.4. *Suppose $s > 0$, $\ell = \lfloor s \rfloor$, f be ℓ times differentiable, and K a kernel of order s. Then for every $h > 0$ and $x \in \mathbb{R}$ the bias may be written as either*

$$b(x) = \int K(u)\frac{(uh)^\ell}{\ell!} f^{(\ell)}(x + zuh)\,\mathrm{d}u \qquad \text{where } 0 \le z \le 1 \text{ may depend on } x, u, h, \tag{3.8}$$

$$\text{or} \quad b(x) = \int K(u)\frac{(uh)^\ell}{(\ell-1)!}\int_0^1 (1 - z)^{\ell-1} f^{(\ell)}(x + zuh)\,\mathrm{d}z\,\mathrm{d}u. \tag{3.9}$$

Proof. As f is ℓ times differentiable, the Taylor expansion of $f(x + t)$ about x can be written as

$$f(x + t) = f(x) + \sum_{j=1}^{\ell-1}\frac{f^{(j)}(x)}{j!}t^j + R_\ell(x, t)$$

where R_ℓ is the remainder term, which may be written as either

$$R_\ell(x, t) = \frac{t^\ell}{\ell!} f^{(\ell)}(x + zt) \qquad \text{for some } 0 \le z \le 1 \text{ that may depend on } x, t, \text{ or}$$
$$R_\ell(x, t) = \frac{t^\ell}{(\ell-1)!}\int_0^1 (1 - z)^{\ell-1} f^{(\ell)}(x + zt)\,\mathrm{d}z.$$

This Taylor expansion with $t = uh$ is used together with (3.6) to rewrite the expectation of the estimator. Since K is a kernel of order s, the first integral is unity and the intermediate integrals vanish as $1 \le j < \ell$,

$$\begin{aligned}
\mathbb{E}\left[\hat{f}_{n,h}(x)\right] &= \int K(u) f(x + uh)\,\mathrm{d}u \\
&= f(x)\int K(u)\,\mathrm{d}u + \sum_{j=1}^{\ell-1}\frac{f^{(j)}(x)}{j!}h^j\int u^j K(u)\,\mathrm{d}u + \int K(u) R_\ell(x, uh)\,\mathrm{d}u \\
&= f(x) + \int K(u) R_\ell(x, uh)\,\mathrm{d}u.
\end{aligned}$$

Then, the bias only depends on the remainder term

$$b(x) = \mathbb{E}\left[\hat{f}_{n,h}(x)\right] - f(x) = \int K(u) R_\ell(x, uh)\,\mathrm{d}u.$$

The first and second results follow by using the first and second remainder terms respectively.

□

3.4 Estimating the variance

The quantity σ^2 can usually be bounded with minimal assumptions on the regularity of f. Firstly, it can be bounded by a certain integral of the kernel as follows.

Lemma 3.5. *Suppose K is a kernel. Then for any $h > 0, n \in \mathbb{N}$ and $x \in \mathbb{R}$,*

$$\sigma^2(x) \leq \frac{1}{nh} \int K(u)^2 f(x+uh)\,\mathrm{d}u. \tag{3.10}$$

Proof. We introduce the random variables

$$Y_i(x) := K_h(X_i - x) - \mathbb{E}\left[K_h(X_i - x)\right], \qquad \text{for each } 1 \leq i \leq n. \tag{3.11}$$

These have many helpful properties, which we show using the linearity of $\mathbb{E}$. Firstly, the sum of Y_i can be expressed in terms of $\hat{f}_{n,h}$,

$$\frac{1}{n}\sum_{i=1}^{n} Y_i(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x) - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)\right] = \hat{f}_{n,h}(x) - \mathbb{E}\left[\hat{f}_{n,h}(x)\right].$$

Secondly, note that Y_i are independent and identically distributed with expected value 0. Then, expected values of cross-terms vanish,

$$i \neq j \implies \mathbb{E}\left[Y_i(x)Y_j(x)\right] = \mathbb{E}\left[Y_i(x)\right]\mathbb{E}\left[Y_j(x)\right] = 0.$$

Using this, notice only $\mathbb{E}\left[Y_i^2(x)\right]$ is needed,

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} Y_i(x)\right)^2\right] = \mathbb{E}\left[\sum_{i=1}^{n} Y_i(x)^2 + 2\sum_{i\neq j} Y_i(x)Y_j(x)\right] = \sum_{i=1}^{n} \mathbb{E}\left[Y_i(x)^2\right].$$

The last thing to notice is that (3.5) can be used to bound the expected value of $Y_i(x)^2$,

$$\mathbb{E}\left[Y_i^2(x)\right] = \mathbb{E}\left[K_h(X_i - x)^2\right] - \mathbb{E}\left[K_h(X_i - x)\right]^2 \leq \mathbb{E}\left[K_h(X_i - x)^2\right] = \frac{1}{h}\int K(u)^2 f(x+uh)\,\mathrm{d}u. \tag{3.12}$$

The result follows by stringing these facts together:

$$\begin{aligned}\sigma^2(x) &= \mathbb{E}\left[\left(\hat{f}_{n,h}(x) - \mathbb{E}\left[\hat{f}_{n,h}(x)\right]\right)^2\right] = \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} Y_i(x)\right)^2\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{E}\left[Y_i(x)^2\right] \leq \frac{1}{nh}\int K(u)^2 f(x+uh)\,\mathrm{d}u.\end{aligned}$$

□

We see that if we could estimate the integral, the variance would be bounded in terms of $1/nh$. Next, in order to construct bounds on the MSE and MISE, we estimate $\sigma^2(x)$ and $\int \sigma^2(x)\,\mathrm{d}x$. In the first case, we will need to assume that f is essentially bounded.

Proposition 3.6. *Suppose $K \in L^2$. Then for every $0 < h$ and $n \in \mathbb{N}$, the following estimates hold:*

(i) For any $x \in \mathbb{R}$ and $f \in L^\infty$,

$$\sigma^2(x) \leq \|f\|_\infty \|K\|_2^2 \frac{1}{nh}. \tag{3.13}$$

(ii) For any probability density f,

$$\int \sigma^2(x)\,\mathrm{d}x \leq \|K\|_2^2 \frac{1}{nh}. \tag{3.14}$$

Proof. To prove (i), start with (3.10). The constant is derived using $f \leq \|f\|_\infty$ almost everywhere,

$$\int K(u)^2 f(x+uh)\,\mathrm{d}u \leq \|f\|_\infty \int K(u)^2\,\mathrm{d}u = \|f\|_\infty \|K\|_2^2 .$$

To prove (ii), integrate (3.10) over $x \in \mathbb{R}$. The order of integration can be swapped by Tonelli's theorem since the integrand is non-negative. The inner integral is unity by (3.7), so the constant is given by

$$\int\int K(u)^2 f(x+uh)\,\mathrm{d}u\,\mathrm{d}x = \int K(u)^2 \int f(x+uh)\,\mathrm{d}x\,\mathrm{d}u = \int K(u)^2\,\mathrm{d}u = \|K\|_2^2 .$$

□

These estimates indicate it is sufficient to have $nh \xrightarrow{n\to\infty} \infty$ in order for the variance terms to decay to 0. This happens for example when $h = cn^{-q}$ for $0 < q < 1$. Thus we want the bandwidth to limit to zero, in order for the bias terms to decay, but not too quickly.

3.5 Hölder spaces

In this section we aim to prove Theorem 1.1 from [14], which bounds the Mean Squared Error over the Hölder space $\mathcal{P}_m(\mathcal{H}^s)$ for $s > 0, m > 0$. This will be the blueprint for similar proofs in the future.

Recall (2.7) from Proposition 2.3, which states MSE $= \sigma^2 + b^2$. We now try to bound the terms of the right-hand side. Assuming $K \in L^2$, Proposition 3.6 with $\|f\|_\infty \leq \|f\|_{\mathcal{H}^s} \leq m$ gives $\sigma^2 \leq \|K\|_2^2\, m/nh$. To bound the bias term, the regularity of f must be used.

Proposition 3.7. *Suppose $s > 0$, $f \in \mathcal{H}^s$ and K is a kernel of order s. Then for every $h > 0, n \in \mathbb{N}$*

$$|b(x)| \leq C_1(s,K)\,\|f\|_{\mathcal{H}^s}\, h^s. \tag{3.15}$$

Proof. Recall (3.8) from Lemma 3.4. Since K is a kernel of order s, we have $\int u^\ell K(u) f^{(\ell)}(x)\,\mathrm{d}u = 0$. Smuggling this term into the integral, we express the bias as

$$b(x) = \int K(u) \frac{(uh)^\ell}{\ell!} (f^{(\ell)}(x+zuh) - f^{(\ell)}(x))\,\mathrm{d}u,$$

where $0 \leq z \leq 1$ may depend on x, u, h. We now use $f \in \mathcal{H}^s$ and $|z| \leq 1$ to write

$$\left| f^{(\ell)}(x+zuh) - f^{(\ell)}(x) \right| \leq \|f\|_{\mathcal{H}^s} |zuh|^{s-\ell} \leq \|f\|_{\mathcal{H}^s} |uh|^{s-\ell} .$$

Together these give the bound

$$\begin{aligned} |b(x)| &\leq \frac{1}{\ell!} \int |K(u)|\,|uh|^\ell \left| f^{(\ell)}(x+zuh) - f^{(\ell)}(x) \right| \mathrm{d}u \\ &\leq \frac{1}{\ell!} \int |K(u)|\,|uh|^\ell\, \|f\|_{\mathcal{H}^s}\, |uh|^{s-\ell}\,\mathrm{d}u \\ &= h^s \frac{\|f\|_{\mathcal{H}^s}}{l!} \int |u|^s\,|K(u)|\,\mathrm{d}u. \end{aligned}$$

□

This result with $\|f\|_{\mathcal{H}^s} \leq m$ gives $b^2 \leq (C_1(s,K)m)^2 h^{2s}$. Notice that the variance decreases with bandwidth, while the bias increases. The is again the *bias-variance tradeoff* mentioned in Section **??**: minimising one will cause the other to grow very large. We must choose some h that minimises their sum, the Mean Squared Error. Assuming the hypotheses of both propositions are met, together they give the estimate

$$\mathrm{MSE}(x) \leq m \|K\|_2^2 \frac{1}{nh} + (C_1(s,K)m)^2 h^{2s},$$

for every $f \in \mathcal{P}_m(\mathcal{H}^s)$ and $x \in \mathbb{R}$. Our aim is to choose a sequence of bandwidths $h = h_n$ to achieve a tight bound on the Mean Squared Error. An obvious path is to choose the h that minimises the above estimate. Some simple calculus shows this happens at

$$h_n^* = \left(\frac{\|K\|_2^2}{2msC_1(s,K)^2} \right)^{1/(2s+1)} n^{-1/(2s+1)}.$$

As mentioned before, the constant is not of interest, only the rate. Choosing $h = h_n = n^{-1/(2s+1)}$ gives

$$(nh)^{-1} = h^{2s} = n^{-2s/(2s+1)}. \tag{3.16}$$

This in turn gives $\mathrm{MSE} \leq Cn^{-2s/(2s+1)}$, where $C = C(s,m,K) = m\|K\|_2^2 + (C_1(s,K)m)^2$. Notice this bound does not depend on f, only that it lies within $\mathcal{P}_m(\mathcal{H}^s)$. This discussion can be distilled into the following result.

Theorem 3.8. *Suppose $s > 0$, $\mathbb{F} = \mathcal{H}^s$ and $K \in L^2$ a kernel of order s. Choose $h = h_n = n^{-1/(2s+1)}$. Then for every $m > 0$ there exists a constant $C = C(s,K,m) > 0$ such that for every $n \in \mathbb{N}$ and $x \in \mathbb{R}$, the corresponding kernel density estimator $\hat{f}_{n,h}$ satisfies*

$$\sup_{f \in \mathcal{P}_m(\mathbb{F})} \mathrm{MSE}(x) \leq Cn^{-2s/(2s+1)}. \tag{3.17}$$

This is what we were looking for; the rate R in this setting is at least $2s/(2s+1)$. Theoretically, as s grows to infinity, i.e. as we look at smoother families of functions, the rate of convergence R approaches 1. This is the typical rate of convergence when using the Mean Squared Error for parametric problems, such as estimating the mean of a distribution with finite variance.

The rate of convergence attained here is in fact the optimal one, as will be the case for each theorem in this section. We will not prove this here, as it requires an altogether different approach which can be found in section 2 of [14].

3.6 Nikol'skii spaces

We now move on to our second regularity space. The Nikol'skii condition (3.2) is a global one, concerning an integral over the entire space. For this reason we use the Mean Integrated Squared Error. The goal of this section is to prove Theorem 1.2 from [14].

Theorem 3.9. *Suppose $s > 0, \mathbb{F} = \mathcal{N}_2^s$ and $K \in L^2$ a kernel of order s. Choose $h = h_n = n^{-1/(2s+1)}$. Then for every $m > 0$ there exists a constant $C = C(s,K,m) > 0$ such that for every $n \in \mathbb{N}$, the corresponding kernel density estimator $\hat{f}_{n,h}$ satisfies*

$$\sup_{f \in \mathcal{P}_m(\mathbb{F})} \mathrm{MISE} \leq Cn^{-2s/(2s+1)}. \tag{3.18}$$

The proof of the above claim will follow the same steps as the previous Theorem. We must first find a bound for the bias assuming that f lies in a Nikol'skii space. To prepare for the cases $p \neq 2$ that we mention later, we bound the bias for every $1 \leq p < \infty$.

Proposition 3.10. *Suppose $s > 0$, $1 \le p < \infty$, $f \in \mathcal{N}_p^s$ and K is a kernel of order s. Then for every $h > 0$ and $n \in \mathbb{N}$, the bias satisifes*

$$\int |b(x)|^p \,\mathrm{d}x \le \left(C_1(s,K)\,\|f\|_{\mathcal{N}_p^s}\right)^p h^{sp}$$

Proof. Recall (3.8) from Lemma 3.4. Since K is a kernel of order s, we have $\int u^\ell K(u) f^{(\ell)}(x)\,\mathrm{d}u = 0$. Moreover, because $\int_0^1 (1-z)^{(\ell-1)}\,\mathrm{d}z = 1/\ell$, we can smuggle a term $-f^{(\ell)}(x)$ into the inner integral to attain

$$b(x) = \int K(u) \frac{(uh)^\ell}{(\ell-1)!} \int_0^1 (1-z)^{\ell-1} (f^{(\ell)}(x+zuh) - f^{(\ell)}(x))\,\mathrm{d}z\,\mathrm{d}u.$$

Now take the absolute value, use the triangle inequality, raise to the power p and integrate over $x \in \mathbb{R}$. The generalised Minkowski inequality, see for example Proposition 6.19 from [8], is used to swap the integrals over x and u,

$$\begin{aligned}\int |b(x)|^p \,\mathrm{d}x &\le \int \left(\int |K(u)| \frac{|uh|^\ell}{(\ell-1)!} \int_0^1 (1-z)^{\ell-1} \left| f^{(\ell)}(x+zuh) - f^{(\ell)}(x) \right| \mathrm{d}z\,\mathrm{d}u \right)^p \mathrm{d}x \\ &\le \left(\int |K(u)| \frac{|uh|^\ell}{(\ell-1)!} \left[\int \left(\int_0^1 (1-z)^{\ell-1} \left| f^{(\ell)}(x+zuh) - f^{(\ell)}(x) \right| \mathrm{d}z \right)^p \mathrm{d}x \right]^{1/p} \mathrm{d}u \right)^p.\end{aligned}$$

Focus now on the inner two integrals. Again using the generalised Minkowski inequality, we swap the integrals over z and x. Because $f \in \mathcal{N}_p^s$ and $|z| \le 1$, the factor on the right is bounded by $\|f\|_{\mathcal{N}_p^s} |uh|^{s-\ell}$. As noted before, $\int_0^1 (1-z)^{\ell-1}\,\mathrm{d}z = 1/\ell$, so we get the bound

$$\begin{aligned} I := &\left[\int \left(\int_0^1 (1-z)^{\ell-1} \left| f^{(\ell)}(x+zuh) - f^{(\ell)}(x) \right| \mathrm{d}z \right)^p \mathrm{d}x \right]^{1/p} \\ &\le \int_0^1 (1-z)^{\ell-1} \left(\int \left| f^{(\ell)}(x+zuh) - f^{(\ell)}(x) \right|^p \mathrm{d}x \right)^{1/p} \mathrm{d}z \\ &\le \int_0^1 (1-z)^{\ell-1} \|f\|_{\mathcal{N}_p^s} |uh|^{s-\ell}\,\mathrm{d}z \le \|f\|_{\mathcal{N}_p^s} \frac{|uh|^{s-\ell}}{\ell}. \end{aligned}$$

Finally we simplify the result, and recognise $C_1(s,K)$,

$$\int |b(x)|^p \,\mathrm{d}x \le \left(\int |K(u)| \frac{|uh|^\ell}{(\ell-1)!} \|f\|_{\mathcal{N}_p^s} \frac{|uh|^{s-\ell}}{\ell}\,\mathrm{d}u \right)^p = h^{sp} \left(\frac{\|f\|_{\mathcal{N}_p^s}}{\ell!} \int |u|^s |K(u)|\,\mathrm{d}u \right)^p.$$

□

We are now ready to conclude the proof of the above theorem. Recall (2.8) from Proposition 2.3, which states MISE $= \int \sigma^2(x)\,\mathrm{d}x + \int b^2(x)\,\mathrm{d}x$. The first term is bounded by Proposition 3.6. Use Proposition 3.10 with $p = 2$ and $\|f\|_{\mathcal{N}_2^s} \le m$ to bound the second term. The theorem is then proved with

$$C(s,K,m) = \|K\|_2^2 + (C_1(s,K)m)^2\,. \tag{3.19}$$

3.7 Sobolev spaces

Our last regularity space is the Sobolev space. Here we prove a version of Theorem 1.3 from [14].

Theorem 3.11. *Suppose $s \in \mathbb{N}$, $\mathbb{F} = \mathcal{W}_2^s$ and $K \in L^2$ a kernel of order s. Choose $h = h_n = n^{-1/(2s+1)}$. Then for every $m > 0$ there exists a constant $C = C(s, K, m) > 0$ such that for every $n \in \mathbb{N}$ the corresponding kernel density estimator $\hat{f}_{n,h}$ satisfies*

$$\sup_{f \in \mathcal{P}_m(\mathbb{F})} \text{MISE} \leq C n^{-2s/(2s+1)}. \tag{3.20}$$

Notice that this is precisely Theorem 3.9 with s an integer and $\mathcal{W}_2^s$ replacing $\mathcal{N}_2^s$. The result is rather immediate from that result, and the observation that $\mathcal{W}_p^s \subset \mathcal{N}_p^s$ for every $s \in \mathbb{N}$ and $p \geq 1$.

Lemma 3.12. *Suppose $s \in \mathbb{N}$, $1 \leq p < \infty$ and $f \in \mathcal{W}_p^s$. Then, $\|f\|_{\mathcal{N}_p^s} \leq \|f\|_{\mathcal{W}_p^s}$.*

Proof. Since $f \in \mathcal{W}_p^s$, we have that $f^{(\ell)}$ exists and is absolutely continuous. Fix $t \in \mathbb{R}$. Use a Taylor expansion on $f^{(\ell)}$ to write

$$f^{(\ell)}(x+t) - f^{(\ell)}(x) = t \int_0^1 f^{(\ell+1)}(x+zt)\,\mathrm{d}z,$$

where we recognise $\ell + 1 = s$. Now raise the absolute value to the power p and integrate over $x \in \mathbb{R}$. Use the generalised Minkowski inequality to swap the order of integration. The inner integral can be recognised to be $\left\|f^{(s)}\right\|_p$ using a linear substitution,

$$\begin{aligned}
|t|^{-p} \int \left|f^{(\ell)}(x+t) - f^{(\ell)}(x)\right|^p \mathrm{d}x &\leq \int \left(\int_0^1 \left|f^{(s)}(x+zt)\right| \mathrm{d}z\right)^p \mathrm{d}x \\
&\leq \left(\int_0^1 \left[\int \left|f^{(s)}(x+zt)\right|^p \mathrm{d}x\right]^{1/p} \mathrm{d}z\right)^p \\
&\leq \left(\int_0^1 \left\|f^{(s)}\right\|_p \mathrm{d}z\right)^p = \left\|f^{(s)}\right\|_p^p.
\end{aligned}$$

We conclude by taking the p^{th} root of this, recalling $1 = s - \ell$ and taking the supremum over $t \in \mathbb{R}$,

$$\|f\|_{\mathcal{N}_p^s} = \|f\|_p + \sup_{t \in \mathbb{R}} \frac{\left[\int \left|f^{(\ell)}(x+t) - f^{(\ell)}(x)\right|^p \mathrm{d}x\right]^{1/p}}{|t|^{s-\ell}} \leq \|f\|_p + \left\|f^{(s)}\right\|_p = \|f\|_{\mathcal{W}_p^s}.$$

□

In particular, this Lemma states that $\mathcal{P}_m(\mathcal{W}_2^s) \subset \mathcal{P}_m(\mathcal{N}_2^s)$ for every $s \in \mathbb{N}$ and $m > 0$, so Theorem 3.11 is just a special case of 3.9. The inclusions of one regularity space within another are very common, but they may not hold in the more generalised spaces we see later.

3.8 L^p risks and other results

We bring this section to a close by discussing some results which we will not cover in detail. Firstly, there similar results to Theorems 3.9 and 3.11 for all values of $p \in [1, \infty)$, using the L^p risk $\mathcal{R}_p$. This was done originally in [1] for Sobolev spaces. The rate of convergence acquired is $R = sp/(2s+1)$, and is proved to be optimal. This result requires additional assumptions, such as K being bounded. Also, in the range $1 \leq p < 2$, the supremum is taken over the space of probability densities f in $\mathbb{F}$ having $\|f\|_{\mathbb{F}} \leq m$ and compact support in some interval $(x_0 - r, x_0 + r)$, written $\mathcal{P}_m(\mathbb{F}, x_0, r)$. One way to approach this problem is to decompose the L^p risk into two terms.

Proposition 3.13. *Suppose $\hat{f}$ is an estimator of f and $1 \leq p < \infty$. Define the* ***stochastic error*** $S := \mathbb{E}\left\|\hat{f}(x) - \mathbb{E}\left[\hat{f}(x)\right]\right\|_p^p$ *and the* ***approximation error*** $B := \left\|\mathbb{E}\left[\hat{f}\right] - f\right\|_p^p$. *Then:*

$$\mathcal{R}_p \leq 2^{p-1}\left(S + B\right). \tag{3.21}$$

Proof. Start with this application of the triangle inequality of the p-norm,

$$\left\|\hat{f} - f\right\|_p \leq \left\|\hat{f} - \mathbb{E}\left[\hat{f}\right]\right\|_p + \left\|\mathbb{E}\left[\hat{f}\right] - f\right\|_p.$$

Then use the power mean inequality,

$$\left\|\hat{f} - f\right\|_p^p \leq 2^{p-1}\left(\left\|\hat{f} - \mathbb{E}\left[\hat{f}\right]\right\|_p^p + \left\|\mathbb{E}\left[\hat{f}\right] - f\right\|_p^p\right).$$

To get the result, take the expectation, noting that the second term is independent of each X_i,

$$\mathcal{R}_p = \mathbb{E}\left\|\hat{f} - f\right\|_p^p \leq 2^{p-1}\left(\mathbb{E}\left\|\hat{f} - \mathbb{E}\left[\hat{f}\right]\right\|_p^p + \left\|\mathbb{E}\left[\hat{f}\right] - f\right\|_p^p\right) = 2^{p-1}(S+B).$$

□

We can see that the approximation error can be expressed simply in terms of the bias,

$$B = \|b\|_p^p = \int |b(x)|^p \, \mathrm{d}x.$$

Estimating this term then depends on the regularity space $\mathbb{F}$. In the case of $f \in \mathcal{N}_p^s$, $f \in \mathcal{W}_p^s$, we can do this using Proposition 3.10, Lemma 3.12. Bounding the L^p risks then only requires estimating the stochastic error. When $p = 2$, we can use Tonelli's theorem to write express S in terms of σ^2,

$$S = \mathbb{E}\int \left(\hat{f}(x) - \mathbb{E}\left[\hat{f}(x)\right]\right)^2 \mathrm{d}x = \int \mathbb{E}\left[\left(\hat{f}(x) - \mathbb{E}\left[\hat{f}(x)\right]\right)^2\right] \mathrm{d}x = \int \sigma^2(x) \, \mathrm{d}x.$$

We then think of the stochastic error as being similar to an integral of the variance. Indeed, the stochastic error will also not depend much on the regularity of f. Unfortunately, bounding this term is more difficult than the standard variance term.

Secondly, is possible to study kernel density estimation on $\mathbb{R}^d$. For example, by using some kernel $K : \mathbb{R}^n \to \mathbb{R}$ which integrates to one. $K_h(u)$ is defined as $h^{-d}K(u/h)$ to ensure that K_h and the kernel density estimator integrate to one also. There, the variance estimate becomes $1/nh^d$, and the convergence rates become $2s/(2s+d)$ and $sp/(2s+d)$. This worsening with dimension is known as the *curse of dimensionality*.

Lastly, notice that we used the same bandwidths and attained the same errors for each of our spaces. Each of the spaces we studied are special cases of the Besov spaces (see [15]). Kernel density estimation over Besov spaces on $\mathbb{R}^n$ is studied in [10].

4 Spaces of homogenous type

In this section we present the modern geometric setting. We work on a metric measure space $(\mathcal{M}, \rho, \mu)$. This is the minimal structure over which kernel density estimation makes sense; we can assign a distance to pairs of points, and we can integrate.

The metric gives the notion of "closeness" and we can define the open ball of radius r about a point $x \in \mathcal{M}$ as

$$B(x, r) := \{\, y \in \mathcal{M} \mid \rho(x, y) < r \,\}.$$

The open balls are ubiquitous in what follows, so we assume they are measurable and label their volumes

$$V(x, r) := \mu(B(x, r)).$$

We currently have no control over the change in $V(x, r)$ as the centre x is moved or the radius r increases, as we would in the case of $\mathbb{R}$. In order to give us some bounds on the growth, we assume some geometric constraints.

Assumption I

We assume that $(\mathcal{M}, \rho, \mu)$ is a measure metric space such that μ is a positive Radon measure, $(\mathcal{M}, \rho)$ is locally compact, and the following two conditions are satisfied:

1. The volume doubling condition,
   $\exists c_0 > 1 : \qquad 0 < V(x, 2r) \leq c_0 V(x, r) < \infty \qquad \forall x \in \mathcal{M}, \forall r > 0.$
2. The non-collapsing condition,
   $\exists c_1 > 0 : \qquad V(x, 1) > c_1 \qquad \forall x \in \mathcal{M}.$

The non-collapsing condition will not be used until the end of this section, where it will be used to bound the volumes of balls with radii $r \leq 1$ in terms of their radius. It is also an added assumption only when $\mu(\mathcal{M}) = \infty$. That is to say, if a space satisfies the volume doubling condition and $\mu(\mathcal{M}) < \infty$, then the second condition holds true, (see Proposition 2.1 of [5]).

We focus now on the volume doubling condition. This tells us that each ball has positive and finite volume, and it allows us to bounds the growth of balls as the radius increases. This in turn allows the estimation of some simple integrals (see Lemma 4.4). A space satisfying the volume doubling condition is said to be a **space of homogenous type**.

4.1 A notion of dimension

We can see the volume doubling condition allows us to bound the volume as the radius increases by any factor, and this will give us a notion of dimension in our framework. For the current discussion, define $d_0 := \log_2 c_0$.

Proposition 4.1. *Let $\lambda > 1$, $r > 0$ and $x \in \mathcal{M}$. Then*

$$V(x, \lambda r) \leq c_0 \lambda^{d_0} V(x, r)$$

Proof. It can be proved by induction that for each $k \in \mathbb{N}$

$$V(x, 2^k r) \leq c_0^k V(x, r) \qquad \text{for every } x \in \mathcal{M} \text{ and } r > 0.$$

Since $\lambda > 1$, there is some $n \in \mathbb{N}_0$ such that $2^n \leq \lambda < 2^{n+1}$. Then

$$c_0^n \leq (2^{d_0})^{\log_2(\lambda)} = \lambda^{d_0}$$

Since $\lambda < 2^{n+1}$, we have $B(x, \lambda r) \subset B(x, 2^{n+1} r)$, and so we have

$$V(x, \lambda r) \leq V(x, 2^{n+1} r) \leq c_0^{n+1} V(x, r) \leq c_0 \lambda^{d_0} V(x, r)$$

□

Thus we see for $d' = d_0 = \log_2 c_0$, the following is true

$$\exists c_0' \geq 1 : \qquad V(x, \lambda r) \leq c_0' \lambda^{d'} V(x, r) \qquad \forall x \in \mathcal{M}, r > 0 \tag{4.1}$$

In a sense then, d_0 acts like a dimension — remember that on $\mathbb{R}^n$, $V(x, \lambda r) = \lambda^n V(x, r)$. However, this could be true for some $d' < d_0$. In such a case, it is unnatural to refer to d_0 as the dimension of the space. This motivates the following definition.

Definition 4.2. *Let $(\mathcal{M}, \rho, \mu)$ be a space satsifying the volume doubling condition. The minimal value of d' such that (4.1) holds, should it exist, is said to be the* ***homogenous dimension*** *of the space.*

Later we will require that some quantities are larger than d (see Lemma 4.4). So, it will be less restrictive to use as small a value of d. In what follows, we use d to mean the homogenous dimension should it exist, otherwise we will use any $d \leq d_0$ obeying (4.1), with c_0' the corresponding constant.

Using this notion of a dimension, we can relate volumes of balls of different centres.

Proposition 4.3. *Let $x, y \in \mathcal{M}, r > 0$. Then*

$$V(x, r) \leq c_0' \left(1 + \frac{\rho(x, y)}{r}\right)^d V(y, r). \tag{4.2}$$

Proof. It can be proven using the triangle inequality that $B(x, r) \subseteq B(y, r + \rho(x, y))$. Then, noting that $1 + r^{-1}\rho(x, y) > 1$ and applying (4.1), we get

$$V(x, r) \leq V\left(y, r + \rho(x, y)\right) = V\left(y, \left(1 + \frac{\rho(x, y)}{r}\right) r\right) \leq c_0' \left(1 + \frac{\rho(x, y)}{r}\right)^d V(y, r)$$

□

4.2 Useful integral estimates

We turn our attention to an important quantity. For $\delta > 0, \tau > 0$ and $x, y \in \mathcal{M}$, we define

$$D_{\delta,\tau}(x, y) := \frac{1}{\sqrt{V(x, \delta) V(y, \delta)}} \left(1 + \frac{\rho(x, y)}{\delta}\right)^{-\tau}. \tag{4.3}$$

This symmetric function is a very helpful tool in the future, and will be used to localise the kernels (see Theorem 6.2). As seen in Section 1, we will deal with integrals of the kernel and its square. The remainder of this chapter is then dedicated to estimating the integrals of $D_{\delta,\tau}$ and $D_{\delta,\tau}^2$.

The first step is the following estimate, which comes from applying (4.2) to bound $V(y, \delta)^{-1/2}$,

$$D_{\delta,\tau}(x, y) \leq \sqrt{c_0'} V(x, \delta)^{-1} \left(1 + \frac{\rho(x, y)}{\delta}\right)^{-\tau + d/2}. \tag{4.4}$$

We then see that estimating the integrals of $D_{\delta,\tau}(x, y)$ and $D_{\delta,\tau}(x, y)^2$ over y will rely on integrating the factor $1 + \delta^{-1}\rho(x, y)$. If we raise it to a negative power $-\tau$, it is a generalisation of the function $\mathbb{R}^n \to \mathbb{R}_+ : x \mapsto (1 + \delta^{-1}\|x\|_2)^{-\tau}$, and captures the idea of polynomial decay on our space $\mathcal{M}$. This real version is known to be integrable when $\tau > n$, which can be shown by using polar coordinates. We then suspect $(1 + \delta^{-1}\rho(x, y))^{-\tau}$ is integrable when $\tau > d$, which we shortly prove. This gives us the opportunity to illustrate how integrals may be estimated despite the little geometric structure.

Before continuing, we make another quick note on integration. From now on, integrals will be performed with respect to the measure μ. They will be written as $\int g \, \mathrm{d}\mu$ for single variable functions, or by $\int g(x, y) \, \mathrm{d}\mu(y)$ for multi-variable functions to distinguish which variable is integrated over. An integral with an unspecified domain is understood to be over the entire space $\mathcal{M}$. The expectation $\mathbb{E}$ and $L^p = L^p(\mathcal{M}, \mu)$ are defined similarly to before, but where the integration is over $\mathcal{M}$ with respect to μ, and $\sim$ is defined by $f \sim g$ when $f = g$ μ-almost everywhere. The properties mentioned before still hold, for example $\mathbb{E}$ is linear, and $(L^p, \|\cdot\|_p)$ is a complete normed vector space. Also, the decompositions from Proposition 2.3 still hold: $\mathrm{MSE} = \sigma^2 + b^2$ and $\mathrm{MISE} = \int \sigma^2 \, \mathrm{d}\mu + \int b^2 \, \mathrm{d}\mu$.

Lemma 4.4. *Suppose $\tau > d$. Then there exists a constant $C_2(\tau) > 0$ such that*

$$I_{\delta,\tau}(x) := \int \left(1 + \frac{\rho(x,y)}{\delta}\right)^{-\tau} \mathrm{d}\mu(y) \leq C_2(\tau) V(x,\delta) \tag{4.5}$$

for every $\delta > 0$ and $x \in \mathcal{M}$.

Proof. We use the idea of dyadic decomposition. Fixing some $x \in \mathbb{R}$ and $\delta > 0$, split the metric space into the nested annuli

$$\mathcal{M} = B(x,\delta) \cup \bigcup_{\nu=1}^{\infty} M_\nu,$$

where $M_\nu := B(x, 2^\nu \delta) \setminus B(x, 2^{\nu-1}\delta)$ for $\nu \in \mathbb{N}$. Then the integral can be expressed as

$$I_{\delta,\tau}(x) = \int_{B(x,\delta)} \left(1 + \frac{\rho(x,y)}{\delta}\right)^{-\tau} \mathrm{d}\mu(y) + \sum_{\nu=1}^{\infty} \int_{M_\nu} \left(1 + \frac{\rho(x,y)}{\delta}\right)^{-\tau} \mathrm{d}\mu(y).$$

On $B(x,\delta)$, $\rho(x,y) \geq 0$ and so $(1 + \delta^{-1}\rho(x,y))^{-\tau} \leq 1$. The first term can be estimated by

$$\int_{B(x,\delta)} \left(1 + \frac{\rho(x,y)}{\delta}\right)^{-\tau} \mathrm{d}\mu(y) \leq V(x,\delta).$$

For any $\nu \in \mathbb{N}$, we have by definition $M_\nu \subset B(x, 2^\nu \delta)$. This gives us the estimate

$$\mu(M_\nu) \leq V(x, 2^\nu \delta) \leq c_0' 2^{\nu d} V(x,\delta).$$

On M_ν, $\rho(x,y) \geq 2^{\nu-1}\delta$ and so $(1 + \delta^{-1}\rho(x,y))^{-\tau} \leq 2^{-\tau(\nu-1)}$. The ν^{th} term is then bounded by

$$\int_{M_\nu} \left(1 + \frac{\rho(x,y)}{\delta}\right)^{-\tau} \mathrm{d}\mu(y) \leq 2^{-\tau(\nu-1)} \mu(M_\nu) \leq c_0' 2^d 2^{(\nu-1)(d-\tau)} V(x,\delta).$$

Using these estimates, the integral satisfies

$$I_{\delta,\tau}(x) \leq \left[1 + c_0' 2^d \sum_{\nu=1}^{\infty} 2^{(\nu-1)(d-\tau)}\right] V(x,\delta).$$

Since $\tau > d$, this geometric series converges to $(1 - 2^{d-\tau})^{-1}$ and so we have proved (4.5) with

$$C_2(\tau) = 1 + \frac{c_0'}{2^{-d} - 2^{-\tau}}. \tag{4.6}$$

□

There are a few ways this could be bounded, but all of them take the form $I_{\delta,\tau}(x) \leq cV(x,\delta)$ for some constant $c = c(\tau) > 0$. This is illustrated by the corresponding lower bound

$$I_{\delta,\tau}(x) \geq \int_{B(x,r)} \left(1 + \frac{\rho(x,y)}{\delta}\right)^{-\tau} \mathrm{d}\mu(y) \geq 2^{-\tau} V(x,\delta).$$

In this sense we say the integral estimate above is sharp. This Lemma allows us to estimate the integral of $D_{\delta,\tau}$ as follows.

Proposition 4.5. *Suppose $\tau > 3d/2$. Then there exists a constant $c = c(\tau) > 0$ such that*

$$\int D_{\delta,\tau}(x,y)\, \mathrm{d}\mu(y) \leq c$$

for any $x \in \mathcal{M}$ and $\delta > 0$.

Proof. First use (4.4) to bound $V(y,\delta)^{-1/2}$. Pull the constants from the integral to get

$$\int D_{\delta,\tau}(x,y)\,\mathrm{d}\mu(y) \le \sqrt{c_0'}V(x,\delta)^{-1}\int\left(1+\frac{\rho(x,y)}{\delta}\right)^{-\tau+d/2}\mathrm{d}\mu(y)$$

We recognise the integral as $I_{\delta,\tau-d/2}(x)$. Since $\tau > 3d/2$, we can use Lemma 4.4 to bound it by $C_2(\tau-d/2)V(x,\delta)$. This proves the claim with $c=\sqrt{c_0'}C_2(\tau-d/2)$. □

This integral estimate is used so often that the condition $\tau > 3d/2$ permeates through many of the later results. In this proof, we had only one factor of $V(x,\delta)^{-1}$, which cancelled against the $V(x,\delta)$ from Lemma 4.4. If we attempt to bound the integral of $D_{\delta,\tau}(x,y)^p$ for some $p>1$, there would be factors of the volume remaining. To deal with this, we finally turn to the non-collapsing condition from Assumption I.

Proposition 4.6. *Suppose $0<\delta\le 1$. Then $V(x,\delta)\ge\frac{c_1}{c_0'}\delta^d$ for every $x\in\mathcal{M}$.*

Proof. The result comes from the non-collapsing condition, $\delta\le 1$ and the volume growth condition,

$$c_1 \le V(x,1) = V(x,\delta^{-1}\delta) \le c_0'\delta^{-d}V(x,\delta).$$

□

Another statement of the above is

$$V(x,\delta)^{-1}\le\frac{c_0'}{c_1}\delta^{-d}\qquad\text{for every } x\in\mathcal{M}, 0<\delta\le 1. \tag{4.7}$$

This is used to pass bounds in terms of $V(x,\delta)^{-1}$ into bounds in terms of δ^{-d}, which helps bound the variance terms in Section 5. We are now ready for the other integral estimate.

Proposition 4.7. *Suppose $\tau > d/2$. Then there is a constant $c=c(\tau)>0$ such that*

$$\int D_{\delta,\tau}(x,y)^2\,\mathrm{d}\mu(y)\le c\delta^{-d}$$

for any $x\in\mathcal{M}$ and $0<\delta\le 1$.

Proof. First use (4.7) to bound $V(y,\delta)^{-1}$. Pull the constants from the integral to get

$$\int D_{\delta,\tau}(x,y)^2\,\mathrm{d}\mu(y)\le\frac{c_0'}{c_1}\delta^{-d}V(x,\delta)^{-1}\int\left(1+\frac{\rho(x,y)}{\delta}\right)^{-2\tau}\mathrm{d}\mu(y).$$

We recognise the integral as $I_{\delta,2\tau}(x)$. Since $\tau>d/2$, we use Lemma 4.4 to bound it by $C_2(2\tau)V(x,\delta)$. This proves the claim with $c=\frac{c_0'}{c_1}C_2(2\tau)$.

□

A similar proof shows that we may estimate the integral of $D_{\delta,\tau}^p$ by $c\delta^{-d(p-1)}$ if $\tau>d(2/p-1/2)$ when $1\le p\le 2$, and if $\tau>d/p$ when $p\ge 2$. However, our main tool to make kernels, Theorem 6.2, will require us to always have $\tau>d$.

4.3 Examples

We now provide some examples of spaces satisfying Assumption I.

1. Euclidean $\mathbb{R}^n$ under the Lebesgue measure. This is of course the space that inspires our studies, and has homogenous dimension $d=n$.
2. The sphere $\mathbb{S}^n$ embedded in Euclidean $\mathbb{R}^{n+1}$. From this embedding, the sphere inherits its measure (the n-dimensional Haussdorf measure restricted to $\mathbb{S}^n$) and spherical metric $\rho(x,y)=\arccos(\langle x,y\rangle)$, where $\langle\cdot,\cdot\rangle$ is the inner product on $\mathbb{R}^{n+1}$. This is an example of a space of finite volume and diameter, and has homogenous dimension $d=n$.

3. Riemannian manifolds M^n of non-negative Ricci curvature. The measure and metric are those inherited naturally from $\mathbb{R}^n$. The fact that such manifolds obey Assumption I follows from the Bishop-Gromov inequality, which also informs us that the homogenous dimension of these spaces is $d = n$.

In each case above we see that the homogenous dimension coincides with the standard geometric dimension. However, this does not have to be the case. It is possible to take a space we normally think of having dimension n, and equip it with a new measure and metric such that the resulting space satsifies Assumption I, but for a homogenous dimension $d \neq n$. Also, d may take any value greater than 0, not restricted to the integers. Two examples of such spaces, the weighted ball and the interval $[-1, 1]$, are explored in [3].

Alfhors regularity

Many examples of spaces satisfying Assumption I, such as $\mathbb{R}^n, \mathbb{S}^n$ and many Riemannian manifolds, have far more geometric structure than strictly necessary, enjoying a property known as Alfhors regularity. We take a moment to see how this added structure makes our work easier.

A metric measure space $(\mathcal{M}, \rho, \mu)$ is said to be **Alfhors regular** if there exist constants $a, b, d > 0$ such that

$$ar^d \leq V(x, r) \leq br^d$$

for every $x \in \mathcal{M}$ and $0 < r \leq \operatorname{diam}(\mathcal{M})$. This is a stronger geometric assumption than our own, and both 1. and 2. of Assumption I follow from it. Such a space has homogenous dimension d.

The improved control of the volumes of balls granted by this property help us to improve some bounds. Firstly, in place of 4.3, we redefine

$$D_{\tau,\delta}(x, y) := \delta^{-d} \left(1 + \frac{\rho(x, y)}{\delta}\right)^{-\tau}.$$

The constant in the integral estimate Lemma 4.4 may be improved, but that is not important. We can use that estimate to show that for $p > 0$ and $\tau > 1/p$, we have

$$\int D^p_{\tau,\delta}(x, y) \leq \delta^{-dp} I_{p\tau,\delta} \leq C_2(p\tau)\delta^{d(1-p)},$$

without assuming that $\delta \leq 1$. This estimate simplifies the requirement for $\tau > 3d/2$ from Proposition 4.5. In fact, any appearance of that requirement can be replaced by $\tau > d$.

5 Spectral theory

5.1 The Laplacian and heat kernels

One of the most famous partial differential equations on $\mathbb{R}$ is the heat equation:

$$\partial_t u = \Delta u$$

where $u = u(x, t)$, $\Delta = \nabla^2 = \nabla \cdot \nabla = \partial_x^2$ is the Laplacian operator on $\mathbb{R}$, the initial conditions $u(x, 0)$ are known and Dirichlet boundary conditions, $\lim_{|x|\to\infty} u(x, t) = 0$ for every $t \geq 0$, are imposed. The solution for $t > 0$ is known to be (see for example [7])

$$u(x, t) = \int p_t(x, y)u(y, 0)\,\mathrm{d}y,$$

where p_t is known as the *fundamental solution* of the heat equation, or the **heat kernel** of the Laplacian,

$$p_t(x, y) := \frac{1}{(4\pi t)^{1/2}} \exp\left(-\frac{(x - y)^2}{4t}\right).$$

We note some important properties by the heat kernel, which we notice is a Gaussian distribution. Firstly, $p_t(x,y) = p_t(y,x)$ and so we say it is symmetric. Secondly, it is smooth in each of x, y, t. As a consequence of the smoothness in y, it is α-Hölder continuous in y for every $\alpha \in (0,1)$. That is to say, for any fixed $t > 0$ and $\alpha \in (0,1)$, there is a constant C such that

$$|p_t(x,y) - p_t(x,y')| \leq C|y - y'|^\alpha.$$

Thirdly, it obeys a Markov property

$$\int p_t(x,y)\,\mathrm{d}y = 1 \text{ for every } x \in \mathbb{R}, t > 0.$$

5.2 Functional calculus

We assume now that $L^2(\mathcal{M}, \mu)$ admits some non-negative essentially self-adjoint operator L, which maps real-valued functions to real-valued functions. This is to be thought of as the Laplacian of $\mathbb{R}$.

In the same way that we can define functions of finite-dimensional operators, by diagonalising its matrix and acting the function on the diagonal elements (its eigenvalues), we can define functions of the operator L using its eignevalues. This is done using the powerful spectral theorem (see for example [16]).

By the spectral theorem, for any Borel measurable function $g : \mathbb{R}_+ \to \mathbb{R}$, we may define $g(L)$ as

$$g(L) := \int_0^\infty g(\lambda)\,\mathrm{d}E_\lambda,$$

where $E_\lambda, \lambda \geq 0$ is the spectral decomposition of L. The operator $g(L)$ is also essentially self-adjoint operator L and maps real-valued functions to real-valued functions, and is called the *spectral multiplier associated with g and L*. If additionally g is non-negative, then $g(L)$ is non-negative, and if g is bounded, then $g(L)$ is bounded. For example, for any $s > 0$ we may define

$$L^s = \int_0^\infty \lambda^s\,\mathrm{d}E_\lambda.$$

In particular we can define $\sqrt{L}$, and it is also non-negative, essentially self-adjoint operator L and maps real-valued functions to real-valued functions. We will make use of $\sqrt{L}$ later.

A second example is the function $\lambda \to e^{-t\lambda}$. Using it, we can define the operators

$$P_t := e^{-tL} = \int_0^\infty e^{-t\lambda}\,\mathrm{d}E_\lambda.$$

These operators form a commutative semigroup under composition

$$P_t \circ P_s = P_{t+s},$$

called the *associated semigroup of L*.

Suppose G is an operator and there is a measurable function $\mathcal{G} : \mathcal{M} \times \mathcal{M} \to \mathbb{C}$ such that

$$(Gf)(x) = \int \mathcal{G}(x,y) f(y)\,\mathrm{d}\mu(y), \text{ for every } f \in \mathrm{Dom}(G), x \in \mathrm{Dom}(f).$$

Then G is called an *integral operator*, and $\mathcal{G}$ is called its *kernel*. As in the previous use of the word kernel, this function contains all the information of the larger object G. These are precisely the kernels we will use for kernel density estimation in the next section. In the case that G is self-adjoint and maps real-valued functions to real-valued functions, then its kernel $\mathcal{G}$ is real-valued and symmetric.

For example, if $P_t, t > 0$ is an integral operator, then we call its kernel the *heat kernel* $p_t(x,y)$. These contain a wealth of information about the operator L, and the ambient space $\mathcal{M}$. In the case of Euclidean $\mathbb{R}^n$ under the Lebesgue measure and $L = -\Delta$, then these are precisely the heat kernels from the heat equation.

5.3 Kernel density estimation

We are now ready to present the second half of the assumptions on our space, which we will use to create kernel density estimators. The key idea is that we assume the space has some Laplacian-like operator whose heat kernels exist and enjoy similar properties to those of the Laplacian. These assumptions on the heat kernels are largely for technical reasons behind the proof of Theorem 6.2.

Assumption II

We assume that there exists an essentially self-adjoint non-negative operator L densely defined on $L^2(\mathcal{M}, \mu)$ which maps real-valued functions to real-valued functions, such that the associated semigroup $P_t = e^{-tL}, t > 0$, consists of integral operators with heat kernel $p_t(x, y)$ obeying the following:

(i) *Gaussian localisation*: There exists constants $c_2, c_3 > 0$ such that

$$|p_t(x,y)| \leq \frac{c_2 \exp\left(-c_3 \rho(x,y)^2/t\right)}{\sqrt{V(x,\sqrt{t})V(y,\sqrt{t})}} \qquad \text{for every } x, y \in \mathcal{M} \text{ and } t > 0. \tag{5.1}$$

(ii) *Hölder continuity*: There exists a constant $\alpha > 0$ such that

$$|p_t(x,y) - p_t(x,y')| \leq \left(\frac{\rho(y,y')}{t}\right)^{\alpha} \frac{c_2 \exp\left(-c_3 \rho(x,y)^2/t\right)}{\sqrt{V(x,\sqrt{t})V(y,\sqrt{t})}} \tag{5.2}$$

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

(iii) *Markov property*:

$$\int p_t(x,y)\, \mathrm{d}\mu(y) = 1 \qquad \text{for every } x \in \mathcal{M} \text{ and } t > 0. \tag{5.3}$$

Definition 5.1. *Let $k : \mathbb{R}_+ \to \mathbb{R}$ be Borel measurable and bounded. Then k is referred to as a **symbol**. Let $h > 0$. The associated **spectral multiplier** is defined as*

$$K_h := k(h\sqrt{L}) = \int_0^\infty k(h\lambda)\, \mathrm{d}F_\lambda,$$

where $F_\lambda, \lambda \geq 0$ is the spectral decomposition associated with $\sqrt{L}$. If the associated spectral mulitplier K_h happens to be an integral operator, that is if there is a symmetric measurable function $\mathcal{K}_h : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ such that

$$(K_h f)(x) = \int \mathcal{K}_h(x,y) f(y)\, \mathrm{d}\mu(y)$$

*for every $f \in Dom(K), x \in Dom(f)$, then $\mathcal{K}_h$ is called the **kernel** of K_h.*

To avoid confusion between these objects, we denote the symbols by lowercase letters such as k, g, the corresponding spectral multipliers by the upper case letters K, G and their kernels by the calligraphic letters $\mathcal{K}, \mathcal{G}$. We are now ready to set up kernel density estimation.

Definition 5.2. *Let $X_1, \cdots, X_n$ be indepedent random variables identically distributed on $\mathcal{M}$ by their shared probability density f. Let k be a symbol and $h > 0$ such that the associated spectral multiplier K_h is an integral operator with kernel $\mathcal{K}_h$ that satisifies*

$$\int \mathcal{K}_h(x,y)\, \mathrm{d}\mu(y) = 1, \text{ for every } x \in \mathcal{M}.$$

*Then we define the associated **kernel density estimator** as*

$$\hat{f}_{n,h}(x) := \frac{1}{n} \sum_{i=1}^{n} \mathcal{K}_h(X_i, x).$$

As before, we are curious what conditions on the symbol ensure good convergence rates for the estimator $\hat{f}_{n,h}$. In the new context, we have a second question: what conditions are needed for the kernel to exist and to satisfy that Markov property, for the range of values of the bandwidth h we might use?

5.4 Examples

The inspiration of our studies is of course Euclidean $\mathbb{R}^n$ equipped with the Laplacian $\Delta = \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}$. We must take care to use $L = -\Delta$ to ensure non-negativity. The sphere, and indeed any Riemannian manifold with non-negative Ricci curvature, may be equipped with a Laplace-Beltrami operator. These form the majority of nice examples of spaces under our framework.

Another example is to equip a space we think of as naturally possessing one operator, with a different choice of operator. For example $\mathbb{R}^n$ may equipped with an elliptic operator, such as an anisotropic Laplacian. Also, the examples of the weighted ball and interval with their new metric and measure can be equipped with a corresponding operator (see [3]).

6 Recent results on spaces of homogenous type

6.1 Regularity spaces of interest

The first step in this process is again to define the function spaces in our study: Hölder, Nikol'skii and Sobolev. We do this by translating each concept from the original definitions to our new setting. Every instance of $|x-y|$, the metric on $\mathbb{R}$, should be replaced with $\rho(x,y)$, the metric on $\mathcal{M}$. The p-norms are now defined with respect to the measure μ over $\mathcal{M}$.

However, it is not so obvious how to define the derivative of f in this context. The space does not necessarily have a translation, so certainly no immediate definition of a derivative. Remember though that we have equipped this space with a Laplacian-like operator L, and that the Laplacian on $\mathbb{R}$ is $\Delta = \mathrm{d}^2/\,\mathrm{d}x^2$, the second derivative. L then gives us a notion of a second derivative over these metric spaces. Loosely speaking, we think of $L^{1/2}$ as a differential operator, and redefine the function spaces by replacing the derivatives $f^{(\ell)}$ by $L^{\ell/2}$. In the case of $\mathbb{R}$, this returns the derivatives when ℓ is even, and captures a similar idea when ℓ is odd.

The last obstacle is the discrete difference $g(x+t) - g(x)$ in the Nikol'skii space. Again, there is no concept of a translation by t, so we define it instead as the average of the quantity over a ball of radius t. In the case of $\mathbb{R}$, this is an equivalent definition.

Definition 6.1. *Let $s > 0$ and $1 \le p < \infty$. A function $f : \mathcal{M} \to \mathbb{R}$ belongs to*

(i) the Hölder space, $\mathcal{H}^s$, if

$$\|f\|_{\mathcal{H}^s} := \|f\|_\infty + \sup_{x \neq y} \frac{\left|L^{\ell/2} f(x) - L^{\ell/2} f(y)\right|}{\rho(x,y)^{s-\ell}} < \infty;$$

(ii) the Nikol'skii space, $\mathcal{N}_p^s$, if

$$\|f\|_{\mathcal{N}_p^s} := \|f\|_p + \sup_{t>0} t^{\ell-s} \left[\int V(x,t)^{-1} \int_{B(x,t)} \left| L^{\ell/2} f(y) - L^{\ell/2} f(x) \right|^p \,\mathrm{d}\mu(y)\,\mathrm{d}\mu(x) \right]^{1/p} < \infty;$$

(iii) and for $s \in \mathbb{N}$, the Sobolev space, $\mathcal{W}_p^s$, if

$$\|f\|_{\mathcal{W}_p^s} := \|f\|_p + \left\| L^{s/2} f \right\|_p < \infty.$$

The above spaces capture the same spirit of the classical functions spaces over $\mathbb{R}$. Once again, it is over the subsets $\mathcal{P}_m(\mathbb{F})$ of these spaces that are probability distributions and the norms are bounded by $m > 0$, which are assumed to be non-empty, that we will study kernel density estimation.

6.2 Conditions on the symbol

As in section 3.2, we anticipate that we will need some restrictions on our kernels. In the previous section, it was discussed that we start with a symbol k, we can create spectral multipliers $K_h := k(h\sqrt{L})$, and if K_h happens to be an integral operator, its kernel $\mathcal{K}_h$ is precisely the kernel we use for our kernel density estimation. This time then, we have two questions: what conditions must the symbol obey so that (i) K_h is an integral operator, and (ii) the kernel density estimator achieves 'good' convergence rates?

Question (i) is answered by the following result, which was developed as Theorem 3.4 in [5] and [11]. These papers use d_0 from the doubling condition as their notion of dimension, and so we write Theorem 2.1 of [4] which matches our notation.

Theorem 6.2. *Suppose $k \in C^\tau(\mathbb{R}_+)$ for $\tau > d$ such that*

$$k^{(2\nu+1)}(0) = 0 \text{ for every } \nu \geq 0 \text{ where } 1 \leq 2\nu + 1 \leq \tau, \tag{6.1}$$

and for some $r > \tau + d$ there exists a constant $C > 0$ such that

$$|k^{(\nu)}(\lambda)| \leq C(1+\lambda)^{-r} \text{ for every } \lambda \geq 0 \text{ and } 0 \leq \nu \leq \tau. \tag{6.2}$$

Then for every $h > 0$, K_h is an integral operator, and its kernel $\mathcal{K}_h(x, y)$ satisfies

$$\int_{\mathcal{M}} \mathcal{K}_h(x, y)\, \mathrm{d}\mu(y) = k(0) \text{ for every } x \in \mathcal{M} \tag{6.3}$$

and furthermore enjoys the decay

$$|\mathcal{K}_h(x, y)| \leq cCD_{h,\tau}(x, y) \tag{6.4}$$

where $c > 0$ is a constant depending on τ and the geometric constants of the setting.

Using a kernel satisfying the hypothesis and $k(0) = 1$, this theorem gives us exactly the function $\mathcal{K}_h$ we need to define our kernel density estimators as in Definition 5.2. It also provides a localisation estimate 6.4 in terms of the functions $D_{\delta,\tau}$ we studied in Section 4.2. It is therefore one of our most important tools, and all of the results to follow will rely on it. To shorten the statements of these results, we use the following terminology.

Definition 6.3. *Let k satisfy the hypothesis of Theorem 6.2. Then k is a **symbol of order** τ. Furthermore, k is a **strong symbol of order** τ if it also satisifes*

$$k^{(\nu)}(0) = 0 \text{ for every } 1 \leq \nu \leq \tau. \tag{6.5}$$

Clearly a strong symbol of order τ is a symbol of order τ, which is a symbol. Before proceeding, let us consider the conditions placed on the symbols k by the hypothesis of Theorem 6.2. The first condition is equivalent to the symbol having an even extension in $C^\tau(\mathbb{R})$. The vanishing derivatives here and in (6.5) correspond to vanishing moments of the Fourier transform of k, which is a common tool in studying kernel density estimators on $\mathbb{R}^n$. The second condition is met if the function is compactly supported, or decays faster than all powers of λ.

We should again be noted that such symbols exist. It can be checked that the map $\lambda \mapsto (1+\lambda^{\tau+1})^{-1}$ is a strong symbol of order τ. The map $\lambda \mapsto e^{-\lambda^2}$ is a symbol of order τ for every $\tau \in \mathbb{N}$, and its associated kernel $\mathcal{K}_h = p_h$ is the heat kernel of L.

6.3 Decomposing the bias

The goal of this subsection is to introduce an analogue of Lemma 3.4, which wrote the bias term in terms of a derivative of f using a Taylor expansion. Here, we bound the bias in terms a power of the operator $\sqrt{L}$ acting on f, a method found in [3] and [4], which relies on machinery built in [5].

First note that the expectation value of the kernel $\mathcal{K}_h$ is K_h applied to f,

$$\mathbb{E}\left[\mathcal{K}_h(X_i,x)\right] = \int_{\mathcal{M}} \mathcal{K}_h(X_i,x) f(X_i)\,\mathrm{d}\mu(X_i) = \int_{\mathcal{M}} \mathcal{K}_h(x,y) f(y)\,\mathrm{d}\mu(y) = K_h[f](x),$$

and so the bias term can be written in terms of K_h, f and the identity operator I,

$$b(x) = \mathbb{E}\left[\hat{f}_{n,h}(x)\right] - f(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathcal{K}_h(X_i,x)\right] - f(x) = (K_h - I)f(x). \tag{6.6}$$

If we let $h > 0$ and i be the integer with $2^{-i} \le h < 2^{-i+1}$, then the following series can be bounded in terms of h

$$\sum_{j=i}^{\infty} 2^{-j\ell} 2^{-j(s-\ell)} = \sum_{j=i}^{\infty} 2^{-js} = 2^{-(i-1)s} = 2^s 2^{-is} \le 2^s h^s \tag{6.7}$$

The following decomposition will eventually be used with $q = \ell$ or $q = s$ to take advantage of the above geometric series.

Lemma 6.4. *Suppose $q \in \mathbb{N}$ and k a strong symbol of order $\tau > d + q$ with $k(0) = 1$. Choose some $0 < h \le 1$ and let i be the integer with $2^{-i} \le h < 2^{-i+1}$. Then there exists a constant $c = c(\tau, q) > 0$ such that both*

$$|b(x)| \le c \sum_{j=i}^{\infty} 2^{-jq} \int_{\mathcal{M}} D_{2^{-j},\tau-q}(x,y) \left| L^{q/2} f(y) \right| \mathrm{d}\mu(y), \text{ and} \tag{6.8}$$

$$|b(x)| \le c \sum_{j=i}^{\infty} 2^{-jq} \int_{\mathcal{M}} D_{2^{-j},\tau-q}(x,y) \left| L^{q/2} f(y) - L^{q/2} f(x) \right| \mathrm{d}\mu(y), \tag{6.9}$$

for every $x \in \mathcal{M}$ and probability density f.

Proof. We first create a decomposition of f. Choose some symbol $\psi \in C^\infty(\mathbb{R}_+)$ with

$$\operatorname{supp}\psi \subset [0,2], \qquad \psi(\lambda) = 1 \quad \forall \lambda \in [0,1], \qquad 0 \le \psi(\lambda) \le 1 \quad \forall \lambda \in [0,2].$$

Then define $\phi(\lambda) := \psi(\lambda) - \psi(2\lambda)$. Clearly $\phi \in C^\infty(\mathbb{R}_+)$ and $\operatorname{supp}\phi \subset [2^{-1}, 2]$. Because of telescoping, for every $\lambda \in \mathbb{R}_+$ we have

$$\begin{aligned}
\sum_{j=i+1}^{\infty} \phi(2^{-j}\lambda) &= \lim_{J\to\infty} \sum_{j=i+1}^{J} \left[\psi(2^{-j}\lambda) - \psi(2^{-(j-1)}\lambda)\right] \\
&= \lim_{J\to\infty} \left[\psi(2^{-J}\lambda) - \psi(2^{-i}\lambda)\right] \\
&= \psi(0) - \psi(2^{-i}\lambda).
\end{aligned}$$

Remembering that $\psi(0) = 1$, we have

$$\psi(2^{-i}\lambda) + \sum_{j=i+1}^{\infty} \phi(2^{-j}\lambda) = 1.$$

Then by Corollary 3.9 of [5]

$$\Psi_{2^{-i}} f + \sum_{j=i+1}^{\infty} \Phi_{2^{-j}} f = f, \tag{6.10}$$

where the convergence is strong in L^1, and as usual by the capital letters we denote the spectral multipliers $\Psi_{2^{-i}} = \psi(2^{-i}\sqrt{L})$ and $\Phi_{2^{-j}} = \phi(2^{-j}\sqrt{L})$. Using this and (6.6), we can express the bias as

$$b(x) = (K_h - I)\Psi_{2^{-i}} f + \sum_{j=i+1}^{\infty} (K_h - I)\Phi_{2^{-j}} f. \tag{6.11}$$

We now introduce the symbols

$$g^i(\lambda) := \frac{(k(h2^i\lambda) - 1)\psi(\lambda)}{\lambda^q} \quad \text{and} \quad g^j(\lambda) := \frac{(k(h2^j\lambda) - 1)\phi(\lambda)}{\lambda^q} \quad \text{for } j > i. \tag{6.12}$$

I claim that each of these are symbols of order $\tau - q$, that they vanish at the origin and there exists a constant $C' = C'(\tau, q) > 0$ independent of j such that

$$\left|(g^j)^{(\nu)}(\lambda)\right| \leq C' \quad \text{for every } \lambda \geq 0,\ 0 \leq \nu \leq \tau - q \text{ and } j \geq i. \tag{6.13}$$

Before proceeding, we prove these claims.

Clearly for $j > i$ we have $\operatorname{supp} g^j \subset \operatorname{supp} \phi \subset [2^{-1}, 2]$. This compact support away from the origin prevents the λ^{-q} term from blowing up, guaranteeing $g^j \in C^\tau(\mathbb{R}_+)$, and so g^j are strong symbols of order τ,. They are certainly symbols of order $\tau - q$.

The case of g^i is more delicate as $\operatorname{supp} g^i \subset \operatorname{supp} \psi \subset [0, 2]$; we must be careful at the origin due to the denominator. Use q applications of L'Hopital's Rule,

$$g^i(0) = \lim_{\lambda \to 0^+} \frac{(k(h2^i\lambda) - 1)\psi(\lambda)}{\lambda^q} = \lim_{\lambda \to 0^+} \frac{(k(h2^i\lambda) - 1)}{\lambda^q} = c \lim_{\lambda \to 0^+} k^{(q)}(h2^i\lambda) = 0$$

Notice that in these limits, the $\psi(\lambda)$ can be set to unity as $\psi(\lambda) = 1$ on $[0, 1]$. The result is zero since k is a strong symbol of order $\tau \geq q$.

We now use the same approach to show the derivatives vanish at the origin. Let $1 \leq \nu \leq \tau - q$. Using the same reasoning, any term containing a derivative of ψ can be set to zero as all derivatives of ψ are zero on $[0, 1]$. Thus, when computing the derivatives of g^i at zero using the product rule, the only terms we must compute are the ones where n of the derivatives act on the denominator, and $\nu - n$ derivatives act on $k(h2^i\lambda) - 1$. The n^{th} term may then be computed using $n + q$ applications of L'Hopital's Rule.

$$\begin{aligned}
(g^i)^{(\nu)}(0) &= \sum_{n=0}^{\nu} \lim_{\lambda \to 0^+} \left(\frac{\mathrm{d}^{\nu-n}}{\mathrm{d}\lambda^{\nu-n}}(k(h2^i\lambda) - 1)\right)\left(\frac{\mathrm{d}^n}{\mathrm{d}\lambda^n}\lambda^{-q}\right) \\
&= c \lim_{\lambda \to 0^+} (k(h2^i\lambda) - 1)\lambda^{-q-\nu} + \sum_{n=0}^{\nu-1} c \lim_{\lambda \to 0^+} k^{(\nu-n)}(h2^i\lambda)\lambda^{-q-n} \\
&= \sum_{n=0}^{\nu} c \lim_{\lambda \to 0^+} k^{(n+q)}(h2^i\lambda) = 0
\end{aligned}$$

The n^{th} term vanishes as long as $n + q \leq \tau$, which is guaranteed by $n \leq \nu \leq \tau - q$.

And so for every $1 \leq \nu \leq \tau - q$ we have $(g^i)^{(\nu)}(0) = 0$. Combined with the fact that g^i is compactly supported, we may conclude that g^i is a strong symbol of order $\tau - q$. This also tells us that the derivatives are bounded by some constant C' that depends on α, τ, i. Notice though that all dependence on i comes from factors of $h2^i$ in the constant, which come from the chain rule. This dependence can be removed by noticing $h2^i < 2$.

Finally, we must show that there is some constant bounding all the derivatives of g^j independent of j. Notice that k is a symbol and so is bounded. Notice also that $\lambda \mapsto \phi(\lambda)\lambda^{-q}$ is $C^\infty(\mathbb{R}_+)$ and is supported within $\operatorname{supp} \phi \subset [2^{-1}, 2]$, and so each of its derivatives up to order $\tau - q$ can be bounded by some c depending only on τ and q. We can express the derivative of order $\nu > 0$ in terms of these, by acting n derivatives on $(k(h2^j\lambda) - 1)$ and $\nu - n$ on $\phi(\lambda)\lambda^{-q}$.

$$\left|(g^j)^{(\nu)}(\lambda)\right| \leq \sum_{n=0}^{\nu} \left|\frac{\mathrm{d}^n}{\mathrm{d}\lambda^n}(k(h2^j\lambda) - 1)\right| \left|\frac{\mathrm{d}^{\nu-n}}{\mathrm{d}\lambda^{\nu-n}}\phi(\lambda)\lambda^{-q}\right| \leq c'\nu \sup_{0 \leq n \leq \nu} \sup_{\lambda \in [2^{-1}, 2]} \left|\frac{\mathrm{d}^n}{\mathrm{d}\lambda^n}(k(h2^j\lambda) - 1)\right|$$

For $n = 0$, $(k(h2^j\lambda) - 1)$ is bounded by $\|k\|_\infty + 1$. For $n > 0$, we use that k is a symbol of order $\tau > n$ to write a decay by a rate $r > \tau + d$. For every $\lambda > 0$, we have

$$\left|k^{(n)}(\lambda)\right| \le C(1+\lambda)^{-r} \le C(1+\lambda)^{-(\tau+d)} \le C\lambda^{-(\tau+d)},$$

where C is the constant from k being a symbol of order τ. This allows us to bound the derivatives

$$\left|\frac{\mathrm{d}^n}{\mathrm{d}\lambda^n}(k(h2^j\lambda) - 1)\right| = (h2^j)^n \left|k^{(n)}(h2^i\lambda)\right| \le (h2^j)^n C(1+(h2^j\lambda))^{-(\tau+d)} \le C(h2^j)^{n-r}\lambda^{-(\tau+d)} \le C2^{\tau+d},$$

where in the last inequality we used that $\lambda^{-(\tau+d)} \le 2^{\tau+d}$ on the domain $[2^{-1}, 2]$, and $(h2^j)^{n-r} < 1$ since $r > \tau + d > \nu \ge n$ and $h2^j > 1$. This proves (6.13) where the constant depends only on α, τ but not on j. This concludes the proof of the claims about the symbols g^j.

Then, by Theorem 6.2, we have that for $j \ge i$, the spectral mulitpliers $G^j_{2^{-j}}$ are integral operators, and their kernels $\mathcal{G}^j_{2^{-j}}$ satisfy

$$\int_{\mathcal{M}} \mathcal{G}^j_{2^{-j}}(x,y)\,\mathrm{d}\mu(y) = g^j(0) = 0 \tag{6.14}$$

and enjoy the decay

$$\left|\mathcal{G}^j_{2^{-j}}(x,y)\right| \le c' D_{2^{-j},\tau-q}(x,y) \tag{6.15}$$

where $c' = cC$ and c depends on τ and the geometric constants. Thus c' depends on τ, q but not on j.

We notice that

$$(k(h\lambda)-1)\psi(2^{-i}\lambda) = (2^{-i}\lambda)^q g^i(2^{-i}\lambda) \quad \text{and} \quad (k(h\lambda)-1)\phi(2^{-j}\lambda) = (2^{-j}\lambda)^q g^j(2^{-j}\lambda), \quad j > i,$$

and so, since each factor of λ gives an operator $L^{1/2}$,

$$(K_h - I)\Psi_{2^{-i}} = 2^{-iq} G^i_{2^{-i}} L^{q/2} \quad \text{and} \quad (K_h - I)\Phi_{2^{-j}} = 2^{-jq} G^j_{2^{-j}} L^{q/2}, \quad j > i.$$

This allows us to rewrite (6.11) as

$$b(x) = \sum_{j=i}^{\infty} 2^{-jq} G^j_{2^{-j}} L^{q/2} f(x)$$

We can write these in terms of the integral operators.

$$b(x) = \sum_{j=i}^{\infty} 2^{-jq} \int_{\mathcal{M}} \mathcal{G}^j_{2^{-j}}(x,y) L^{q/2} f(y)\,\mathrm{d}\mu(y) \tag{6.16}$$

We can bound the absolute value using (6.15) to get

$$\begin{aligned} |b(x)| &\le \sum_{j=i}^{\infty} 2^{-jq} \int_{\mathcal{M}} \left|\mathcal{G}^j_{2^{-j}}(x,y)\right| \left|L^{q/2} f(y)\right| \,\mathrm{d}\mu(y) \\ &\le c' \sum_{j=i}^{\infty} 2^{-jq} \int_{\mathcal{M}} D_{2^{-j},\tau-q}(x,y) \left|L^{q/2} f(y)\right| \,\mathrm{d}\mu(y), \end{aligned}$$

which is the first form. To get the second form, we go back to (6.16). Because of (6.14), we have the option to smuggle in a term of $-L^{q/2}f(x)$ within the integral,

$$b(x) = \sum_{j=i}^{\infty} 2^{-jq} \int_{\mathcal{M}} \mathcal{G}^j_{2^{-j}}(x,y)(L^{q/2}f(y) - L^{q/2}f(x))\,\mathrm{d}\mu(y).$$

Then use the same bounding argument as above.

□

6.4 Estimating the variance

Analagous results from the classical case hold by very similar arguments.

Lemma 6.5. *Suppose k is a symbol of order $\tau > d$. Then for any $h > 0, n \in \mathbb{N}$*

$$\sigma^2(x) \leq \frac{1}{n} \int \mathcal{K}_h(x,y)^2 f(y) \, \mathrm{d}\mu(y). \tag{6.17}$$

Proof. Proceed exactly as in the proof of Lemma 3.5, with two changes. Theorem 6.2 guarantees the existence of $\mathcal{K}_h$, so this time define the random variables Y_i as

$$Y_i(x) := \mathcal{K}_h(X_i, x) - \mathbb{E}\left[\mathcal{K}_h(X_i, x)\right],$$

and instead of (3.12), use

$$\mathbb{E}\left[Y_i^2(x)\right] \leq \mathbb{E}\left[\mathcal{K}_h(X_i,x)^2\right] = \int \mathcal{K}_h(x,y)^2 f(y) \, \mathrm{d}y.$$

□

As before, this Lemma gives us the ability to bound $\sigma^2(x)$ and its integral, which in turn will be used to bound the Mean Squared Error and Mean Integrated Squared Error.

Proposition 6.6. *Suppose k be a symbol of order $\tau > d$. Then there exists a constant $c = c(\tau) > 0$ such that for every $0 < h \leq 1$ and $n \in \mathbb{N}$, the following estimates hold:*

(i) For any $x \in \mathcal{M}$ and $f \in L^\infty$,

$$\sigma^2(x) \leq c \left\|f\right\|_\infty \frac{1}{nh^d}. \tag{6.18}$$

(ii) For any probability density f,

$$\int \sigma^2 \, \mathrm{d}\mu \leq c \frac{1}{nh^d}. \tag{6.19}$$

Proof. First we claim there is some $c = c(\tau) > 0$ such that

$$\int \mathcal{K}_h(x,y)^2 \, \mathrm{d}\mu(y) \leq ch^{-d}.$$

This follows from the localisation estimate (6.4); there exists some $c' = c'(\tau) > 0$ such that $|\mathcal{K}_y(x,y)| \leq c' D_{h,\tau}(x,y)$ for every $x, y \in \mathcal{M}$, and Proposition 4.7; there exists some $c'' = c''(\tau) > 0$ such that $\int D_{h,\tau}(x,y)^2 \, \mathrm{d}\mu(y) \leq c'' h^{-d}$ for every $x \in \mathcal{M}$. Now use 6.17 of the above Lemma. To prove (i), use $f \leq \left\|f\right\|_\infty$ almost everywhere,

$$\int \mathcal{K}_h(x,y)^2 f(y) \, \mathrm{d}\mu(y) \leq \left\|f\right\|_\infty \int \mathcal{K}_h(x,y)^2 \, \mathrm{d}\mu(y) \leq c \left\|f\right\|_\infty h^{-d}$$

To prove (ii), use Tonelli's theorem to swap the order of integration as the integrand is non-negative. The remaining integral is unity as f is a probability density,

$$\int \int \mathcal{K}_h(x,y)^2 f(y) \, \mathrm{d}y \, \mathrm{d}x = \int f(y) \int \mathcal{K}_h(x,y)^2 \, \mathrm{d}x \, \mathrm{d}y \leq ch^{-d} \int f(y) \, \mathrm{d}y = ch^{-d}.$$

□

6.5 Hölder Spaces

We are now ready to present a version the main result of [4], a generalisation of our Theorem 3.8.

Theorem 6.7. *Suppose $s > 0$, $\mathbb{F} = \mathcal{H}^s$ and k a strong symbol of order $\tau > 3d/2 + s$ with $k(0) = 1$. Choose $h = h_n = n^{-1/(2s+d)}$. Then for every $m > 0$ there exists a constant $C = C(s, \tau, m) > 0$ such that for every $n \in \mathbb{N}$ and $x \in \mathcal{M}$, the corresponding kernel density estimator $\hat{f}_{n,h}$ satisfies*

$$\sup_{f \in \mathcal{P}_m(\mathbb{F})} \mathrm{MSE}(x) \leq C n^{-2s/(2s+d)}. \tag{6.20}$$

Notice the convergence rate $R = 2s/(2s + d)$. This matches the rate we achieved on the one-dimensional Euclidean space $\mathbb{R}$. So, we have answered the question "how fast are the convergence rates on these metric spaces?" with the answer "at least as fast as on $\mathbb{R}^d$", which may seem surprising. We will find this is also the case for the remaining regularity spaces.

This theorem is proved similarly to Theorem 3.8: use (i) of Proposition 6.6 to bound the variance term, and the following to bound the bias term.

Proposition 6.8. *Suppose $s > 0$, $f \in \mathcal{P}(\mathcal{H}^s)$ and k a strong symbol of order $\tau > 3d/2+s$ with $k(0) = 1$. Then there is a constant $c = c(\tau, s) > 0$ such that for every $h > 0$ and $x \in \mathcal{M}$*

$$|b(x)| \leq c \, \|f\|_{\mathcal{H}^s} \, h^s. \tag{6.21}$$

Proof. Firstly, in this proof we use the shorthand

$$J_\delta(x) := \int_{\mathcal{M}} D_{\delta,\tau-\ell}(x, y) \left| L^{\ell/2} f(y) - L^{\ell/2} f(x) \right| \mathrm{d}\mu(y).$$

We invoke Lemma 6.4 with $q = \ell = \lfloor s \rfloor$, and use the above to write

$$|b(x)| \leq c' \sum_{j=i}^{\infty} 2^{-j\ell} \int_{\mathcal{M}} D_{2^{-j},\tau-\ell}(x, y) \left| L^{\ell/2} f(y) - L^{\ell/2} f(x) \right| \mathrm{d}\mu(y) = c' \sum_{j=i}^{\infty} 2^{-j\ell} J_{2^{-j}}(x), \tag{6.22}$$

where c' depends on τ and s. Notice it is sufficient to prove that

$$\exists C = C(\tau, s) > 0 : J_\delta(x) \leq C \, \|f\|_{\mathcal{H}^s} \, \delta^{s-\ell} \quad \forall \, \delta > 0. \tag{6.23}$$

The result would follow quickly from this, (6.22), (6.7),

$$|b(x)| \leq c' \sum_{j=i}^{\infty} 2^{-j\ell} J_{2^{-j}} \leq c' C \, \|f\|_{\mathcal{H}^s_p} \sum_{j=i}^{\infty} 2^{-j\ell} (2^{-j})^{s-\ell} \leq c' C 2^s \, \|f\|_{\mathcal{H}^s_p} \, h^s.$$

We proceed to prove (6.23). We use the fact that $f \in \mathcal{H}^s$ to write

$$\left| L^{\ell/2} f(y) - L^{\ell/2} f(x) \right| \leq \|f\|_{\mathcal{H}^s} \, \rho(x, y)^{s-\ell}.$$

Note that since $\rho(x, y) \leq \delta(1 + \delta^{-1}\rho(x, y))$, we have

$$\rho(x, y)^{s-\ell} \leq \delta^{s-\ell} (1 + \delta^{-1} \rho(x, y))^{s-\ell}.$$

Noticing that $D_{\delta,\tau-\ell}(x, y)(1 + \delta^{-1}\rho(x, y))^{s-\ell} = D_{\delta,\tau-s}(x, y)$, we get

$$\begin{aligned} J_\delta(x) &= \int_{\mathcal{M}} D_{\delta,\tau-\ell}(x, y) \left| L^{\ell/2} f(y) - L^{\ell/2} f(x) \right| \mathrm{d}\mu(y) \\ &\leq \|f\|_{\mathcal{H}^s} \, \delta^{s-\ell} \int_{\mathcal{M}} D_{\delta,\tau-s}(x, y) \, \mathrm{d}\mu(y). \end{aligned}$$

Since $\tau - s > 3d/2$, we can bound the integral using Proposition (4.5). This proves the the claim (6.23) with $C(\tau, s) = C(\tau, s) = \sqrt{c_0'} C_2(\tau - s - d/2)$.

□

6.6 Nikol'skii Spaces

Theorem 3.9 can be generalised similarly with the convergence rate $2s/(2s+d)$.

Theorem 6.9. *Suppose $s > 0$, $\mathbb{F} = \mathcal{N}_2^s$ and k a strong symbol of order $\tau > 3d/2 + s$ with $k(0) = 1$. Choose $h = h_n = n^{-1/(2s+d)}$. Then for every $m > 0$ there exists a constant $C = C(s, \tau, m) > 0$ such that for every $n \in \mathbb{N}$, the corresponding kernel density estimator $\hat{f}_{n,h}$ satisfies*

$$\sup_{f \in \mathcal{P}_m(\mathbb{F})} \mathrm{MISE} \leq C n^{-2s/(2s+d)}. \tag{6.24}$$

The variance term is bounded by (ii) of Proposition 6.6, and it remains only to bound the bias. This can be done muchj the same as we did in the Hölder case, though less cleanly.

Proposition 6.10. *Suppose $s > 0$, $1 \leq p < \infty$, $f \in \mathcal{N}_p^s$ and k a strong symbol of order $\tau > 3d/2 + s$ with $k(0) = 1$. Then there is a constant $c = c(\tau, s) > 0$ such that for every $h > 0$*

$$\|b\|_p \leq c \|f\|_{\mathcal{N}_p^s} h^s. \tag{6.25}$$

Proof. Again, we use the shorthand

$$J_\delta(x) := \int_{\mathcal{M}} D_{\delta, \tau-\ell}(x, y) \left| L^{\ell/2} f(y) - L^{\ell/2} f(x) \right| \mathrm{d}\mu(y).$$

We invoke Lemma 6.4 with $q = \ell = \lfloor s \rfloor$, and use the above to write

$$|b(x)| \leq c' \sum_{j=i}^{\infty} 2^{-j\ell} \int_{\mathcal{M}} D_{2^{-j}, \tau-\ell}(x, y) \left| L^{\ell/2} f(y) - L^{\ell/2} f(x) \right| \mathrm{d}\mu(y) = c' \sum_{j=i}^{\infty} 2^{-j\ell} J_{2^{-j}}(x), \tag{6.26}$$

where $c' = c'(\tau, s) > 0$ is a constant. Notice it is sufficient to prove the following:

$$\exists C = C(\tau, s) > 0 : \|J_\delta\|_p \leq C \|f\|_{\mathcal{N}_p^s} \delta^{s-\ell} \quad \forall \delta > 0. \tag{6.27}$$

The result follows quickly from this, (6.26), the p-norm triangle inequality and (6.7),

$$\|b\|_p \leq c' \sum_{j=i}^{\infty} 2^{-j\ell} \|J_{2^{-j}}\|_p \leq c' C \|f\|_{\mathcal{N}_p^s} \sum_{j=i}^{\infty} 2^{-j\ell} (2^{-j})^{s-\ell} \leq c' C 2^s \|f\|_{\mathcal{N}_p^s} h^s.$$

We proceed to prove (6.27). Again use the dyadic decomposition seen in Lemma 4.4,

$$\mathcal{M} = B(x, \delta) \cup \bigcup_{\nu=1}^{\infty} M_\nu,$$

where $M_\nu := B(x, 2^\nu \delta) \setminus B(x, 2^{\nu-1}\delta)$ for $\nu \in \mathbb{N}$. We bound the integrand on each piece, using $\sigma = \tau - \ell$ for brevity. Recall (4.4)

$$D_{\delta,\sigma}(x, y) \leq \sqrt{c_0'} V(x, \delta)^{-1} (1 + \delta^{-1} \rho(x, y))^{-\sigma + d/2}.$$

On $B(x, \delta)$, $\rho(x, y) \geq 0$ gives $(1 + \delta^{-1}\rho(x, y))^{-\sigma+d/2} \leq 1$. On M_ν, $\rho(x, y) \geq 2^{\nu-1}\delta$ and so

$$(1 + \delta^{-1}\rho(x, y))^{-\sigma+d/2} \leq 2^{(\nu-1)(-\sigma+d/2)}.$$

Then use the volume growth condition to write

$$V(x, \delta)^{-1} = \frac{V(x, 2^\nu \delta)}{V(x, \delta)} V(x, 2^\nu \delta)^{-1} \leq c_0' 2^{\nu d} V(x, 2^\nu \delta)^{-1}.$$

By combining these, we have for $y \in M_\nu$

$$D_{\delta,\sigma}(x, y) \leq 2^d \sqrt{c_0'}^3 2^{(\nu-1)(-\sigma+3d/2)} V(x, 2^\nu \delta)^{-1}.$$

Together these allow us to bound the integral by

$$\begin{aligned} J_\delta(x) &= \int_{\mathcal{M}} D_{\delta,\sigma}(x,y)F(x,y)\,\mathrm{d}\mu(y) \\ &= \int_{B(x,\delta)} D_{\delta,\sigma}(x,y)F(x,y)\,\mathrm{d}\mu(y) + \sum_{\nu=1}^{\infty}\int_{M_\nu} D_{\delta,\sigma}(x,y)F(x,y)\,\mathrm{d}\mu(y) \\ &\le \sqrt{c_0'}V(x,\delta)^{-1}\int_{B(x,\delta)} F(x,y)\,\mathrm{d}\mu(y) \\ &+ 2^d\sqrt{c_0'}^3\sum_{\nu=1}^{\infty} 2^{(\nu-1)(-\sigma+3d/2)}V(x,2^\nu\delta)^{-1}\int_{M_\nu} F(x,y)\,\mathrm{d}\mu(y). \end{aligned}$$

The integrand is positive and $M_\nu \subset B(x,2^\nu\delta)$, so we can change the integration domain from M_ν to $B(x,2^\nu\delta)$. Now we use the triangle inequality for the p-norm to write

$$\begin{aligned} \|J_\delta\|_p &\le \sqrt{c_0'}\left\|V(\cdot,\delta)^{-1}\int_{B(\cdot,\delta)} F(\cdot,y)\,\mathrm{d}\mu(y)\right\|_p \\ &+ 2^d\sqrt{c_0'}^3\sum_{\nu=1}^{\infty} 2^{(\nu-1)(-\sigma+3d/2)}\left\|V(\cdot,2^\nu\delta)^{-1}\int_{B(\cdot,2^\nu\delta)} F(\cdot,y)\,\mathrm{d}\mu(y)\right\|_p. \end{aligned}$$

Focus on this p-norm, for some $t>0$ replacing $2^\nu\delta$. Because $\mu/V(x,t)$ is a probability measure on $B(x,t)$ and $x\mapsto x^p$ is convex on $\mathbb{R}_+$ for $p\ge 1$, we can use Jensen's inequality to move the exponent of p inside the integral. It can be bounded using $f\in\mathcal{N}_p^s$,

$$\begin{aligned} \left\|V(\cdot,t)^{-1}\int_{B(\cdot,t)} F(\cdot,y)\,\mathrm{d}\mu(y)\right\|_p &= \left[\int\left(V(x,t)^{-1}\int_{B(x,t)} F(x,y)\,\mathrm{d}\mu(y)\right)^p\right]^{1/p} \\ &\le \left[\int V(x,t)^{-1}\int_{B(x,t)} F(x,y)^p\,\mathrm{d}\mu(y)\right]^{1/p} \\ &\le \|f\|_{\mathcal{N}_p^s}\,|t|^{s-\ell}. \end{aligned}$$

Then, we may rewrite the above as

$$\begin{aligned} \|J_\delta\|_p &\le \sqrt{c_0'}\,\|f\|_{\mathcal{N}_p^s}\,\delta^{s-\ell} + 2^d\sqrt{c_0'}^3\sum_{\nu=1}^{\infty} 2^{(\nu-1)(-\sigma+3d/2)}\,\|f\|_{\mathcal{N}_p^s}\,(2^\nu\delta)^{s-\ell} \\ &= \left[\sqrt{c_0'} + 2^{d+s-\ell}\sqrt{c_0'}^3\sum_{\nu=0}^{\infty} 2^{\nu(-\sigma+3d/2+s-\ell)}\right]\|f\|_{\mathcal{N}_p^s}\,\delta^{s-\ell}. \end{aligned}$$

Since $\sigma=\tau-\ell>3d/2+s-\ell$, this geometric series converges and so we have proved (6.27) with

$$C(\tau,s) = \sqrt{c_0'} + 2^{d+s-\ell}\sqrt{c_0'}^3\frac{2^\sigma}{2^\sigma - 2^{3d/2+s-\ell}} = \sqrt{c_0'}\left[1 + c_0'\frac{2^{\tau+d+s-\ell}}{2^\tau - 2^{3d/2+s}}\right]. \tag{6.28}$$

□

6.7 Sobolev Spaces

Once again our final regularity space is the Sobolev space. We now present the $p=2$ case of Theorem 3 of [3]. This generalises our Theorem 3.11, with the same convergence rate as before.

Theorem 6.11. *Suppose $s\in\mathbb{N}$, $\mathbb{F}=\mathcal{W}_2^s$ and k a strong symbol of order $\tau>3d/2+s$ with $k(0)=1$. Choose $h=h_n=n^{-1/(2s+d)}$. Then for every $m>0$ there exists a constant $C=C(s,\tau,m)>0$ such that for every $n\in\mathbb{N}$, the corresponding kernel density estimator $\hat{f}_{n,h}$ satisfies*

$$\sup_{f\in\mathcal{P}_m(\mathbb{F})}\mathrm{MISE}\le Cn^{-2s/(2s+d)}. \tag{6.29}$$

This time around, we do not have an obvious inclusion of $\mathcal{W}_p^s$ within $\mathcal{N}_p^s$, so the bias term must be bounded specifically for the Sobolev case.

Proposition 6.12. *Suppose $s \in \mathbb{N}$, $1 \leq p < \infty$, $f \in \mathcal{W}_p^s$ and k a strong symbol of order $\tau > 3d/2 + s$ with $k(0) = 1$. Then there is a constant $c = c(\tau, s) > 0$ such that for every $h > 0$*

$$\|b\|_p \leq c \|f\|_{\mathcal{W}_p^s} h^s. \tag{6.30}$$

Proof. We invoke Lemma 6.4 with $q = s$ to write

$$|b(x)| \leq c' \sum_{j=i}^{\infty} 2^{-js} \int_{\mathcal{M}} D_{2^{-j},\tau-s}(x,y) \left| L^{s/2} f(y) \right| \mathrm{d}\mu(y) = c' \sum_{j=i}^{\infty} 2^{-js} H \left| L^{s/2} f \right| (x) \tag{6.31}$$

where we have defined H to be the integral operator with kernel $\mathcal{H}(x,y) = D_{2^{-j},\tau-s}(x,y)$. Since $\tau - s > 3d/2$ and the function is symmetric, Lemma 4.4 states that both $\|\mathcal{H}(x,\cdot)\|_1 \leq C$ and $\|\mathcal{H}(\cdot,y)\|_1 \leq C$, where $C = C(\tau,s) = \sqrt{c_0'} C_2(\tau - s - d/2)$. A very well known result of integral operators, see for example Theorem 6.36 of [8], is that then $\|Hg\|_p \leq C \|g\|_p$ for every g in the domain of H. In particular, we use this on $\left| L^{s/2} f \right|$

$$\left\| H \left| L^{s/2} f \right| \right\|_p \leq C \left\| L^{s/2} f \right\|_p \leq C \|f\|_{\mathcal{W}_p^s}. \tag{6.32}$$

The result is attained by combining this with (6.31), the triangle inequality of the p-norm and (6.7).

$$\|b\|_p \leq c \sum_{j=i}^{\infty} 2^{-js} \left\| H \left| L^{s/2} f \right| \right\|_p \leq c' C \|f\|_{\mathcal{W}_p^s} \sum_{j=i}^{\infty} 2^{-js} \leq c' C 2^s \|f\|_{\mathcal{W}_p^s} h^s$$

□

6.8 L^p risks and other results

We bring the thesis to a close by again mentioning some results we will not prove. Theorems 6.9 and 6.11 have generalisations to $p \neq 2$ under the L^p risks.

Theorem 6.13. *Suppose $s > 0$, $1 \leq p < \infty$, $m > 0$, $\mathbb{F} = \mathcal{N}_p^s$ and k a strong symbol of order $\tau > 3d/2 + s$ with $k(0) = 1$. Choose $h = h_n = n^{-1/(2s+d)}$. Then the kernel density estimator $\hat{f}_{n,h}$ obeys the following:*

(i) Let $p \geq 2$. Then there is a constant $c = c(s,p,\tau,m) > 0$ such that for every $n \in \mathbb{N}$

$$\sup_{f \in \mathcal{P}_m(\mathbb{F})} \mathcal{R}_p \leq c n^{-sp/(2s+d)}.$$

(ii) Let $1 \leq p < 2$, $\tau > d(1 + 1/p)$, $x_0 \in \mathcal{M}$, $R > 0$. Then there is a constant $c = c(s,p,\tau,m,x_0,R) > 0$ such that for every $n \in \mathbb{N}$

$$\sup_{f \in \mathcal{P}_m(\mathbb{F},x_0,R)} \mathcal{R}_p \leq c n^{-sp/(2s+d)}.$$

The above statement is also true if we insert $s \in \mathbb{N}$ and replace $\mathbb{F} = \mathcal{W}_p^s$. As in Proposition 3.13, the L^p risk can be bounded by two terms. The approximation error $B = \|b\|_p^p$ is estimated using Proposition 6.10 or Proposition 6.12 as appropriate. The remaining stochastic term is estimated by Theorem 2 in [3]. Notice that the requirement of bounded support for $p < 2$ is also present here, and again the rate we achieve aligns with the classical case.

In [3], the Sobolev spaces are defined for any $s > 0$, and the results we have provided are derived under slightly weaker assumptions on the symbol k. Proposition 6.4 would then requires more care for non-integer $q = s$.

The more general class of Besov spaces can also be defined over these metric spaces. These are studied in detail in [11] and [2], attaining analagous upper bounds to those on $\mathbb{R}$ — once again, kernel density estimation on these metric spaces converges as fast as on Euclidean space.

References

[1] Bretagnolle, J., and Huber, C. (1979), 'Estimation des densités: risque minimax' (French) *Z. Wahrsch. Verw. Gebiete*, 47 (2), 119-137.

[2] Cleanthous, G., Georgiadis, A. G., Kerkyacharian, G., Petrushev, P., and Picard, D. (2020), 'Kernel and wavelet density estimators on manifolds or more general metric spaces'. *Bernoulli*, 26 (3), 1832-1862.

[3] Cleanthous, G., Georgiadis, A.G., Porcu, E. (2022), 'Oracle inequalities and upper bounds for kernel density estimators on manifolds and more general metric spaces'. *Journal of Nonparametric Statistics*, 34 (4), 734-757.

[4] Cleanthous, G., Georgiadis, A.G., White, P.A. (2025) 'Pointwise density estimation on metric spaces and applications in seismology', *Metrika 88*, 119–148.

[5] Coulhon, T., Kerkyacharian, G., and Petrushev, P. (2012), 'Heat Kernel Generated Frames in the Setting of Dirichlet Spaces'. *Journal of Fourier Analysis and Applications*, 18 (5), 995–1066.

[6] Devroye, L., and Györfi, L., *Nonparametric Density Estimation: The L_1 View*, New York: Wiley, 1985.

[7] Evans, L. C. (2010), 'Partial Differential Equations', *American Mathematical Society*, 2nd ed.

[8] Folland, G.B. (1999), *Real analysis: Modern techniques and their applications,* Pure and Applied Mathematics, New York: Wiley.

[9] Hall, P., and Murison, R. D., 'Correcting for Negativity in High-Order Kernel Density Estimation', *Journal of Multivariate Analysis*, 47(1), 1993, 103–121.

[10] Kerkyacharian, G., and Picard, D. (1992), 'Density estimation in Besov spaces', *Statistics & Probability Letters*, 13, 15–24.

[11] Kerkyacharian, G., and Petrushev, P. (2015), 'Heat kernel based decomposition of spaces of distributions in the framework of Dirichlet spaces', *Transactions of the American Mathematical Society*, 367, 121–189.

[12] Parzen, E. (1962), 'On the estimation of a probability density function and mode'. *Annals of Mathematical Statistics*, 33, 1065-1076.

[13] Rosenblatt, M. (1956), 'Remarks on some nonparametric estimates of a density function', *Annals of Mathematical Statistics*, 27, 832-837.

[14] Tsybakov, A.B. (V. Zaiats, trans.) (2009), *Introduction to nonparametric estimation,*, Springer Series in Statistics, New York: Springer

[15] Triebel, H., Theory of function spaces, Monographs in Math. Vol. 78, Birkhäuser, Verlag, Basel, 1983.

[16] Yoshida, K. (1978), *Functional analysis*, Berlin: Springer-Verlag.