Bayesian Nonparametric Estimation under the Manifold Hypothesis

Paul Rosa

Department of Statistics & Jesus College

University of Oxford

A thesis presented for the degree of

Doctor of Philosophy

Statement of originality

I hereby declare that except where specific reference is made to the work of others either in the text or the statements of authorship forms at the end of Chapters 23 & 4, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university.

Paul Rosa, Oxford, 17/01/2025.

Acknowledgements

My first thoughts obviously go towards my supervisor Judith, without whom completing this DPhil journey would not have been possible. Her patience and friendliness allowed me to grow into a better researcher, and she has always welcomed and valued my ideas. For this I would like to thank her deeply. Another big thank you goes to my co-authors Clément, Slava & Alex from which I learned a lot thanks to our collaboration, either in geometry, functional analysis or Gaussian processes; I hope our word lines meet again. Another thank you goes to Émile, Dan & Michael, who definitely made me feel welcome in office 1.14 (and later for Émile in the East End), and also to Shahine, Tyler and my other friends from the stats department. Another thought to my Oxford (and later for some of them, London) friends Rein, Simone, Billie, Simon, Lila & Camille with which I had some of my fondest memories here. Also to the Jesus people : Kyu Min, Dani, Lucas, Francis, Chris, Clara, Lou, Lisa, Fynn, Luka, Sophie, Hadley & others with whom I had the best time, either at the wine & cheese nights, bops and pubs or simply eating soup on Saturdays at Ship Street 13. To my Ker Lann friends, Julien, JB, Mathias, Paul, Julian, Hermes who definitely made my mathematical journey a fun one, either at the Opera, l'Astrolabe, l'Etoile or the Sur Mesure. Also to my Crous friends Ines & Eliot that I would like to thank for these two wonderful years at Les Arceaux, I wish we could do it again ! And to the Montpellier gang : Maxime, Thibault, Thomas, Stéphane & Loic, the Peyrou times may be over but we're definitely not done having fun, I can't wait for our next reunion. Of course to my old friend Yoann, being your mate for more than 16 years has definitely been a gift, and coming to La Croix-Rousse is always a highlight of any trip to France. This thesis would also not have been possible without my past teachers, especially in sixth form and classes prépa, and for this I deeply thank Mme.Massonnat, Mr.Marino & Mme.Chireux for their support. A special thought goes to Michou, whose particular interest in involving the heat kernel in every possible mathematical situations has definitely influenced in my thinking. Finally I would like to thank my family. To my grandparents Mamie Thérézou, Papi Moustache, Yaya & Barthélémy, I know you're there somewhere and I hope we can meet again so that I can properly say how lucky I am to be your grandson. To my parents

of course, whose immense support since day one has made me extremely grateful. I would not have been able to do it without you. To my siblings Camille & Jean, who will always be two of my best friends, and Prune to whom I hope we gave the best life she could have had. Finally to Gwen, whose presence has definitely sweetened my life, whether baking in the Provence sun or freezing in icy Port Meadows. Thanks for sticking with me even throughout the hardest times.

Bayesian Nonparametric Estimation under the Manifold Hypothesis

Paul Rosa

Abstract

Developing methodologies and theoretical understanding of statistical methods at large scales is a pressing matter given the recent growth in the high volume of data available, both in terms of sample sizes (large n) and number of predictors (large p). While the first regime leads mainly to computational issues, the second one also comes with specific statistical questions, as a consequence of the curse of dimensionality. In this thesis, we investigate how the manifold hypothesis, the idea that high dimensional predictors often possesses unknown low dimensional structures, can improve statistical convergence results for Bayesian nonparametric estimation procedures in this context.

In the introduction of this thesis, we will introduce the mathematical tools used in the subsequent chapters to tackle the aforementioned questions. We begin by reviewing in Section 1.1 basic concepts of differential and Riemannian geometry, focusing on submanifolds of Euclidean spaces. We will then in Section 1.2 present fundamental results of spectral theory of compact submanifolds and finite graphs, and show how random graphs can provide meaningful approximations of unknown submanifolds. The proof techniques specific to Bayesian nonparametric statistics will be presented in Section 1.3, where we will focus on the existing methods to derive posterior contraction rates. We will then review in Section 1.4 the state of knowledge related to the questions in nonparametric estimation theory studied in Chapters 2 3 & 4.

As an integrated thesis, Chapters 2 3 & 4 contain published or unpublished but submitted research papers. In Chapters 2 & 3 we will be investigating the problem of nonparametric regression with covariates supported on unknown submanifolds, where we will obtain posterior contraction rates depending only on the intrinsic dimensionality of the data for a class

of methodologies where limited results were previously known. We will then be focusing on the problem of density estimation for probability distributions supported near unknown submanifolds, by designing a new class of nonparametric Gaussian mixture models and deriving corresponding posterior contraction rates, together with the implementations of the methods. Finally, Chapter 5 ends with a discussion.

Contents

Chapter 1

Introduction

From the frequentist point of view, the goal of inferential statistics is to deduce information about a parameter f_0 (an element of a measurable space $(\mathcal{F}, \mathcal{B})$) based only on observations $\mathbb{Y}^n$ (element of a measurable space $\left(\mathcal{Y}^{(n)}, \mathcal{A}^{(n)}\right)$ distributed according to some probability distribution $P_{f_0}^{(n)}$. Here n represents an amount of "information" that is available based on $\mathbb{Y}^n$, which is typically a sample size (as in density estimation or regression) but can be something else such as the noise precision in the white noise model, see e.g. Giné & Nickl [56]. In the context of nonparametric estimation theory the parameter set $\mathcal{F}$ is a subset of a high (i.e. whose dimension increases with n) or infinite dimensional vector space, and we are interested in producing either

- **estimators** $\hat{f}$, i.e. measurable maps $\hat{f} : \mathcal{Y}^{(n)} \to \mathcal{F}$,
- or **Bayesian procedures**, i.e. posterior distributions $\Pi\left[\cdot|\mathbb{Y}^n\right]$ defined as the conditional distributions of the parameter $f \in \mathcal{F}$ given the observations $\mathbb{Y}^n$ in the joint prior model $f \sim \Pi, \mathbb{Y}^n | f \sim P_f^{(n)}$, for some probability distribution Π on $\mathcal{F}$.

Notice that the posterior distribution is not necessarily well defined without further assumptions on the measurable spaces and the prior distribution. Sufficient conditions are discussed in [55], Chapter 1 and we recall them in Section 1.3.

In this thesis, we will be focusing on Bayesian procedures (at the exception of our result regarding the PCR-LE estimator in [92], Chapter 3) and analysing their performances from the frequentist point of view, which can be referred to as the “frequentist Bayes” paradigm, see Rousseau [94], Castillo [24] or Ghosal & Van der Vaart [55]. As for the Cramer-Rao lower bound in the parametric case, fundamental limitations to the performances of statistical estimations exist in the nonparametric setting, and can typically be formulated in the **minimax** framework, which we now introduce. Let us suppose that we are interested in estimating an element f_0 of some set $B \subset \mathcal{F}$: typically $(\mathcal{F}, \|\cdot\|)$ is a normed function space and B is a ball centered at 0 (whose radius may be unknown). In essence this means that we are making a smoothness assumption on f_0; and often the subset B is a Hölder, Sobolev or Besov ball. Then it is possible to show lower bounds of the type

$$\exists C > 0, \quad \inf_{\hat{f}} \sup_{f \in B} \mathbb{E}_{\mathbb{Y}^n \sim P_f^{(n)}} \left[\rho \left(\hat{f}, f \right) \right] \geq C \varepsilon_n$$

for various distances (or semi distances) ρ over the space of probability distributions on $\mathcal{Y}^{(n)}$ and some sequence ε_n of positive real numbers. Here the infimum is taken over all possible estimators $\hat{f}$. In other words, for any estimator $\hat{f}$ there will always exist an element f in the set B such that $\mathbb{E}_{\mathbb{Y}^n \sim p_f^{(n)}} \left[\rho \left(\hat{f} - f \right) \right] \geq C \varepsilon_n$. If it exists, an estimator $\hat{f}$ achieving the lower bound (perhaps up to a constant), i.e. satisfying

$$\sup_{f \in B} \mathbb{E}_{\mathbb{Y}^n \sim P_f^{(n)}} \left[\rho \left(\hat{f}, f \right) \right] \leq C' \varepsilon_n$$

for some $C' \geq C$ is then called minimax optimal (or just minimax) over the set B, and the sequence ε_n is called the minimax rate. It is often the case to encounter estimators $\hat{f}$ that are minimax up to a logarithmic factor, i.e. such that for some $C', \kappa > 0$

$$\sup_{f \in B} \mathbb{E}_{\mathbb{Y}^n \sim P_f^{(n)}} \left[\rho \left(\hat{f}, f \right) \right] \leq C' (\ln n)^\kappa \varepsilon_n$$

From the asymptotic point of view the extra $(\ln n)^\kappa$ factor is often considered negligible, and the estimator is called near minimax optimal. We refer to [110] for an introduction to the theory of minimax estimation.

When the parameter of interest f_0 is contained in a collection of the form $\bigcup_{\lambda\in\Lambda} B_\lambda$ and λ is unknown (a typical situation when λ is a smoothness parameter and $B_\lambda \subset \mathcal{F}_\lambda$ is a ball in a normed function space $\mathcal{F}_\lambda$), then it is often possible to design so called (near) minimax adaptive estimators, that is estimators $\hat{f}$ satisfying

$$\sup_{\substack{\lambda\in\Lambda \\ f\in B_\lambda}} v_n(\lambda)^{-1} \mathbb{E}_{\mathbb{Y}^n\sim P_f^{(n)}} \left[\rho\left(\hat{f}, f\right)\right] \leq C,$$

where for $\varepsilon_n(\lambda)$ the minimax rate of B_λ we have $v_n(\lambda) = \varepsilon_n(\lambda) \ln^{c_\lambda} n$, for some constant $C, c_\lambda > 0$.

Goldenslugher and Lepski [59] have investigated the problem of adaptive minimax estimation of a density on $\mathbb{R}^D, \quad D \geq 1$ based on an *i.i.d.* sample of size $n \geq 1$. Under the L^1-loss $\rho(f,g) = \int_{\mathbb{R}^D} |f-g|$, with the smoothness spaces $\mathcal{F}_\lambda$ being the *anisotropic Besov spaces* $B^{\bar{\beta}}_{\infty\infty}\left(\mathbb{R}^D\right), \quad \bar{\beta} \in (0,\infty)^D$ and choosing balls in $B^{\bar{\beta}}_{\infty\infty}\left(\mathbb{R}^D\right)$ together with a tail condition as models for B_λ, they show that the minimax rate for the corresponding class is lower bounded by a multiple of $(n/\ln n)^{-\frac{\beta}{2\beta+D}}$, where $\beta^{-1} = \frac{1}{D}\sum_{j=1}^D \beta_j^{-1}$ is the harmonic mean of the smoothness along the coordinates axis $\beta_j' s$. Moreover, they construct an estimator $\hat{f}$ achieving this lower bound adaptively, in the sense that the estimator does not require the preliminary knowledge of $\bar{\beta}$. A similar behaviour arises in the nonparametric regression problem on $[0,1]^D$, where one observes $Y_i = f_0(X_i) + \varepsilon_i$ for some $\beta-$Hölder function $f_0 : [0,1]^D \to \mathbb{R}$ and (possibly random covariates) $X_i \in [0,1]^D$, see Tsybakov [110] and the references therein. In both cases, although it may introduce additional $\ln n$ factors, considering adaptive estimators is not responsible for the exponent $\frac{\beta}{2\beta+D}$ as it is already the one appearing in the non adaptive lower bounds for the rates.

Even if for a given statistical problem the minimax rates can exhibit very different behaviour depending on the semi metric ρ and the parameter classes $\mathcal{F}_\lambda$, a general and intuitive rule of thumb is always satisfied : the minimax rates (adaptive or not)

1. get smaller when the regularity of the functional parameter increases,

2. get larger when the dimension of the input space increases.

The second item is an instance of the so-called **curse of dimensionality**, the idea that statistical problems grow in complexity with the dimension of the data space. In the setting of minimax estimation, it is characterised by an exponential dependence of the minimax rates in the dimension of the input space. For instance, if we want to estimate a $\mathcal{C}^1$ function over $[0,1]^D$ in the nonparametric regression problem then, according to Stone [103] and under technical assumptions the corresponding minimax rate over a Lipschitz ball (for instance with respect to the L^2-norm) is $n^{-\frac{1}{D+2}}$: in order to maintain an error $\mathcal{O}(\varepsilon)$, we need n to be at least of the order of $\varepsilon^{-(D+2)}$. This exponential dependence of the required sample size in the dimension D of the input space is prohibitive and cannot be tempered without further assumptions.

In the Bayesian framework, if the prior and the likelihood are conjugate, it is possible to derive rates ε_n satisfying

$$\int\int \rho(f,f_0)\Pi\left[df|\mathbb{Y}^n\right] P_{f_0}^{(n)}(d\mathbb{Y}^n) \leq C\varepsilon_n \tag{1.1}$$

or

$$\mathbb{E}_{\mathbb{Y}^n\sim P_{f_0}^{(n)}}\Pi\left[\rho(f,f_0)>C\varepsilon_n|\mathbb{Y}^n\right] = \int\int_{\rho(f,f_0)>C\varepsilon_n} \Pi\left[df|\mathbb{Y}^n\right] P_{f_0}^{(n)}(d\mathbb{Y}^n) \xrightarrow[n\to\infty]{} 0 \tag{1.2}$$

by using the explicit form of the posterior distribution (see for instance Section 1.7 in Castillo [24] for the example of Gaussian priors in the white noise model). Such estimates are called posterior contraction rates, in expectation 1.1 or probability 1.2. In non-conjugate situations this is much more challenging, and alternative arguments must be used. This is the object of the **prior mass and testing approach**, popularised notably by Ghosal, Ghosh & Van der Vaart [52] and Ghosal & Van der Vaart [53], which by Baye's formula allows to derive posterior contraction rates using only assumptions on $P_{f_0}^{(n)}$, the semi metric ρ and the prior Π and that we will present in Section 1.3.

We present the necessary geometry and functional analysis background in Sections 1.1

& 1.2. In Section 1.4 we introduce the idea of the **manifold hypothesis** (a framework allowing to reduce the influence of the curse of dimensionality in the presence of implicit low-dimensional structures in the data) and the associated relevant and recent literature and a brief summary of our contributions. Chapters 2 3 & 4 contain a detailed account of our contributions (in the form of two published papers and a submitted one). Chapter 5 ends with a discussion.

1.1 Elements of differential geometry of submanifolds

In this section we give a brief introduction to differential and Riemannian geometry. We focus on submanifolds of $\mathbb{R}^D$ as our work in Chapters 23 & 4 does not involve general, non-embedded manifolds. Section 1.1.1 covers basic definitions of manifold and tangent spaces while Section 1.1.2 introduces some concepts of Riemannian geometry, following Robbin & Salamon [90].

1.1.1 Submanifolds

In this section, we introduce the notion of submanifolds and tangent spaces that we will use in Chapters 2, 3 and 4. From now on, unless stated otherwise the term "smooth" refers to infinitely differentiable functions and "diffeomorphism" refers to bijective smooth maps whose inverse is also smooth.

Roughly speaking, a subset $\mathcal{M}$ of $\mathbb{R}^D, D \geq 1$ is a $d-$dimensional **submanifold** if it locally "looks like" $\mathbb{R}^d$. This can be formalised as follows:

Definition 1. *Submanifolds of $\mathbb{R}^D$, [90] Definition 2.1.3*
*A subset $\mathcal{M}$ of $\mathbb{R}^D$ is called a $d-$dimensional submanifold if for any $p \in \mathcal{M}$ there exists an open neighbourhood U of p in $\mathbb{R}^D$ and a diffeomorphism $\varphi : U \cap \mathcal{M} \mapsto \Omega = \varphi\left(U \cap \mathcal{M}\right) \subset \mathbb{R}^d$ (see Remark 1 below). The diffeomorphism φ is called a **coordinate chart** and its inverse $\psi = \varphi^{-1} : \Omega \to U \cap \mathcal{M}$ a **local parametrisation** of $U \cap \mathcal{M}$.*

remark 1. *In Definition 1, we say that a map $\varphi : X \subset \mathbb{R}^D \to Y \subset \mathbb{R}^d$ is smooth (resp. $\mathcal{C}^k$, differentiable) if for any point $p \in X$ we can find an open neighbourhood U of p in $\mathbb{R}^D$*

and a smooth (resp. $\mathcal{C}^k$, differentiable) map $F : U \to \mathbb{R}^d$ such that $F_{|X} = \varphi$. We say that $\varphi : X \to Y$ is a diffeomorphism (resp. $\mathcal{C}^k-$diffeomorphism) if it is bijective and both $\varphi : X \to Y$ and $\varphi^{-1} : Y \to X$ are smooth (resp. $\mathcal{C}^k$).

An **atlas** of $\mathcal{M}$ will then be a family $(U_j, \varphi_j)_{j\in I}$ of coordinate charts with associated neighbourhoods satisfying $\mathcal{M} = \cup_{j\in I} U_j$.

We will always implicitely equip submanifolds of $\mathbb{R}^D$ with the induced subset topology so that the set $U \cap \mathcal{M}$ of Definition 1 is a generic open neighbourhood of $p \in \mathcal{M}$. It turns out that submanifolds of $\mathbb{R}^D$ can be described equivalently in various ways :

Theorem 1. *Equivalent descriptions of submanifolds, [90] Theorem 2.1.10*
Let $\mathcal{M} \subset \mathbb{R}^D$ and $p_0 \in \mathcal{M}$. Then the following are equivalent :

i) There exists an open neighbourhood U_0 of $p_0 \in \mathcal{M}$ and a diffeomorphism $\varphi_0 : U_0 \to \Omega_0 \subset \mathbb{R}^d$ where Ω_0 is open

ii) There exists open sets U, Ω of $\mathbb{R}^D$ and a diffeomorphism $\varphi : U \to \Omega$ such that $p_0 \in U$ and

$$\varphi\left(U \cap \mathcal{M}\right) = \Omega \cap \left(\mathbb{R}^d \times \{0\}^{D-d}\right)$$

iii) There exists an open set $U \subset \mathbb{R}^D$ and a smooth map $f : U \to \mathbb{R}^{D-d}$ such that $p_0 \in U$, the differential $df_p : \mathbb{R}^D \to \mathbb{R}^{D-d}$ is surjective for any $p \in U \cap \mathcal{M}$ and

$$U \cap \mathcal{M} = f^{-1}\left(\{0\}^{D-d}\right) = \{p \in U : f(p) = 0\}$$

Moreover if $i)$ holds then the diffeomorphism φ in $ii)$ can be chosen such that

$$\varphi(p) = (\varphi_0(p), 0)\,.$$

remark 2. *Theorem 1 is stated for smooth (i.e. infinitely differentiable) coordinates charts φ_0, "lifting" maps φ or constraints f. Its proof relies mainly on the inverse function theorem which also allows for non infinitely differentiable functions. Thus, the statement of Theorem*

1 still holds if the smoothness property is replaced more generally by a $\mathcal{C}^k-$assumption on the diffeomorphisms and smooth maps, $1 \leq k \leq +\infty$.

As we can see from Theorem 1, a submanifold $\mathcal{M} \subset \mathbb{R}^D$ initially described by coordinate charts in Definition 1 can also be described "lifting" maps (item *ii*)) or by an equation (item *iii*)), but this only holds locally.

We now define the **tangent and normal spaces**. From now on, for a curve $\gamma : t \in I \mapsto \gamma_t$ defined on an interval $I \subset \mathbb{R}$, we write $\dot{\gamma}_t$ (resp. $\ddot{\gamma}_t$) for its first (resp. second) derivative at t.

Definition 2. *Tangent & normal spaces*
$v \in \mathbb{R}^D$ is said to be tangent to the submanifold $\mathcal{M} \subset \mathbb{R}^D$ at $p \in \mathcal{M}$ if there exists a smooth curve $\gamma : \mathbb{R} \to \mathcal{M}$ such that $\dot{\gamma}_0 = v$ and $\gamma_0 = p$. The set $T_p\mathcal{M}$ of such v's is called the tangent space of $\mathcal{M}$ at p. The orthogonal complement $(T_p\mathcal{M})^\perp = \left\{u \in \mathbb{R}^D : \langle u|v\rangle = 0 \quad \forall v \in T_p\mathcal{M}\right\}$ of $T_p\mathcal{M}$ in $\mathbb{R}^D$ is called the normal space of $\mathcal{M}$ at p.

By definition, normal spaces are linear subspaces of $\mathbb{R}^D$. However regarding the tangent spaces Definition 2 only defines a set $T_p\mathcal{M} \subset \mathbb{R}^D$ which has no guarantee a priori to be an affine or linear subspace of $\mathbb{R}^D$. Fortunately, given our definition of submanifolds 1 and their equivalent characterisations 1 we can actually describe the tangent space $T_p\mathcal{M}$ at p as a linear subspace of $\mathbb{R}^D$, thus justifying the terminology. From now on, for any differentiable map $\varphi : U \to V$ between subsets of U, V of $\mathbb{R}^{D_1}, \mathbb{R}^{D_2}, \quad p \in \mathbb{R}^{D_1}$ and linear map $A : \mathbb{R}^{D_1} \to \mathbb{R}^{D_2}$ we write $d\varphi_p$ for the differential of φ at p and $\ker(A) = A^{-1}(\{0\})$ (resp. $\mathrm{Im}(A) = A\left(\mathbb{R}^{D_1}\right)$) for the kernel of A (resp. its range).

Theorem 2. *Description of the tangent spaces*
Let $\mathcal{M} \subset \mathbb{R}^D$ be a submanifold of dimension d and $p \in \mathcal{M}$. Then the following holds :

i) If $U_0 \subset \mathcal{M}$ is open, $p \in U_0$ and $\varphi_0 : U_0 \to \Omega_0$ is a diffeomorphism onto an open subset $\Omega_0 \subset \mathbb{R}^d$, then with $x_0 = \varphi_0(p), \psi_0 = \varphi_0^{-1} : \Omega_0 \to U_0$ we have

$$T_p\mathcal{M} = \mathrm{Im}\left(d\psi_{x_0}\right)$$

ii) Let $U, \Omega \subset \mathbb{R}^D$ be open sets and $\varphi : U \to \Omega$ be a diffeomorphism such that $p \in U$ and $\varphi(U \cap \mathcal{M}) = \Omega \cap \left(\mathbb{R}^d \times \{0\}^{D-d}\right)$. Then we have

$$T_p\mathcal{M} = d\varphi_p^{-1}\left(\mathbb{R}^d \times \{0\}^{D-d}\right)$$

iii) Let $U \subset \mathbb{R}^D$ be an open neighbourhood of p and $f : U \to \mathbb{R}^{D-d}$ be a smooth map such that $0 \in \mathbb{R}^{D-d}$ is a regular value of f and $U \cap \mathcal{M} = f^{-1}(\{0\})$. Then

$$T_p\mathcal{M} = \ker df_p$$

iv) $T_p\mathcal{M}$ is a linear subspace of $\mathbb{R}^D$.

remark 3. *Regarding item iii) : a regular value of a differentiable map $f : U \to \mathbb{R}^{D-d}$ is an element c of $\mathbb{R}^{D-d}$ such that $df_p : \mathbb{R}^D \to \mathbb{R}^{D-d}$ is surjective whenever $f(p) = c$.*

remark 4. *Theorem 2 implies that the tangent spaces $T_p\mathcal{M}, p \in \mathcal{M}$ are all $d-$ dimensional linear subspaces of $\mathbb{R}^D$. Thus we do not necessarily have $p \in T_p\mathcal{M}$; a tangent subspace that intersects $\mathcal{M}$ at p has to be affine in general and is described by $p + T_p\mathcal{M} = \{x + y : y \in T_p\mathcal{M}\}$.*

remark 5. *If $\mathcal{M} \subset \mathbb{R}^D$ is a $d-$dimensional submanifold and $p \in \mathcal{M}$ is such that the projection from $T_p\mathcal{M}$ to $\mathbb{R}^d \times \{0\}^{D-d}$ is invertible then there exists an open set $\Omega \subset \mathbb{R}^d$ and a smooth map $h : \Omega \to \mathbb{R}^{D-d}$ such that the set $\{(x, h(x)) : x \in \Omega\} \subset \mathbb{R}^D$ is actually an open neighbourhood of p in $\mathcal{M}$. Of course the projection is in general not guaranteed to be invertible, but we can always find an orthogonal transformation O of $\mathbb{R}^D$, an open neighbourhood $U_0 \subset \mathbb{R}^D$ of $O \cdot p$ and a smooth map $h : U_0 \to \mathbb{R}^d$ such that the set*

$$\left\{O^{-1} \cdot (x, h(x)) : x \in U_0\right\}$$

is an open neighbourhood of p in $\mathcal{M}$. This shows that we can always describe locally a submanifold $\mathcal{M}$ at a point $p \in \mathcal{M}$ as the graph of a function defined on an open set of $\mathbb{R}^d \simeq T_p\mathcal{M}$.

With the concept of tangent space at hand we can now introduce the notion of **basis vec-**

tors : given a $d-$dimensional submanifold $\mathcal{M} \subset \mathbb{R}^D$, a point $p \in \mathcal{M}$, an open neighbourhood U of p in $\mathbb{R}^D$ and a coordinate chart $\varphi : U \cap \mathcal{M} \to \Omega \subset \mathbb{R}^d$ with inverse ψ, we define the coordinate basis vectors at p as the elements

$$\partial_i = \frac{d}{dt}_{t=0} \psi\left(\varphi(p) + te_i\right) = \frac{\partial \psi}{\partial x_i}, \quad 1 \leq i \leq d$$

of $T_p\mathcal{M}$, where $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ are the canonical basis vectors of $\mathbb{R}^d$. Since φ is a diffeomorphism the vectors $(\partial_i)_{i=1}^d$ are linearly independent and form a basis of the tangent space $T_p\mathcal{M}$. These basis vectors are a natural generalisation of the standard basis of $\mathbb{R}^d$ to submanifolds once a set of local coordinates has been chosen.

As we have seen in this section, a submanifold $\mathcal{M} \subset \mathbb{R}^D$ can be described locally by $d-$dimensional coordinates. However not all coordinate systems are made equal : even in one dimension the circle

$$S^1 = \left\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\right\}$$

can be described around $p_0 = (1, 0)$ by the vector $e_\alpha = \{\cos(\alpha), \sin(\alpha)\}, \alpha \in I = [-\pi/2, \pi/2]$. But if $t \in I \mapsto \alpha(t) \in I$ is a smooth increasing map from I to itself with $\alpha(-\pi/2) = -\pi/2, \alpha(\pi/2) = \pi/2$ then $e'_t = e_{\alpha(t)}$ is also a valid parametrisation of S^1 around x_0 even though it can heavily distort distances when $\alpha(t)$ is nonlinear. As we will see in the next section, we can always find well behaved (in some sense) parametrisations of a submanifold $\mathcal{M}$ around each of its point, but for this we will first need to define the concept of **geodesics**.

1.1.2 Riemannian structure of submanifolds

Intuitively, a differentiable curve $\gamma : \mathbb{R} \to \mathcal{M}$ should be "straight" if a "bug" living on $\mathcal{M}$ and sitting at γ_t should not perceive any change of direction of the vector $\dot{\gamma}_t$. For instance, if

$$\mathcal{M} = S^2 := \left\{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1\right\}$$

and

$$\gamma_t = \left(\cos(t)\cos(\alpha), \sin(t)\cos(\alpha), \sin(\alpha)\right), \quad \alpha \in]-\pi/2, \pi/2[,$$

then

$$\dot{\gamma}_t = \cos(\alpha)\,(-\sin(t), \cos(t), 0)\,, \quad \ddot{\gamma}_t = -\cos(\alpha)\,(\cos(t), \sin(t), 0)\,.$$

Thus the acceleration vector $\ddot{\gamma}_t$ is never equal to 0. But from the perspective of the bug living on $\mathcal{M}$ and sitting at γ_t, only the tangential component of $\ddot{\gamma}_t$ is perceived. The latter is equal to

$$\Big(\big(\cos^2(\alpha) - 1\big)\cos(\alpha)\cos(t), \big(\cos^2(\alpha) - 1\big)\cos(\alpha)\sin(t), \cos^2(\alpha)\sin(\alpha)\Big)$$

and is therefore zero if and only if $\alpha = 0$: this explains the intuition that the "straight lines" of $\mathcal{M}$ should be the great circles.

In general for a submanifold $\mathcal{M} \subset \mathbb{R}^D$ and a curve γ_t on $\mathcal{M}$ we define $\nabla_{\dot{\gamma}_t}\dot{\gamma}_t$ as the orthogonal projection of $\ddot{\gamma}_t$ onto $T_{\gamma_t}\mathcal{M}$. $\nabla_{\dot{\gamma}_t}\dot{\gamma}_t$ is called the **covariant derivative** of $\dot{\gamma}_t$ along the curve γ_t. We say that a curve γ is a geodesic if and only if $\nabla_{\dot{\gamma}}\dot{\gamma} = 0$: this is a natural generalisation of the concept of straight lines to submanifolds. Notice that geodesics automatically have constant velocity :

$$\frac{d}{dt}\,\|\dot{\gamma}_t\|^2 = 2\langle\dot{\gamma}_t|\ddot{\gamma}_t\rangle \underbrace{=}_{\dot{\gamma}_t \in T_{\gamma_t}\mathcal{M}} 2\langle\dot{\gamma}_t|\nabla_{\dot{\gamma}_t}\dot{\gamma}_t\rangle = 2\langle\dot{\gamma}_t|0\rangle = 0.$$

It is interesting to write out the equation $\nabla_{\dot{\gamma}}\dot{\gamma} = 0$ in local coordinates; and for this we need to introduce the **metric tensor**. If $\mathcal{M}$ is a $d-$dimensional submanifold of $\mathbb{R}^D$ and $p \in \mathcal{M}$ we can define an inner product structure on $T_p\mathcal{M}$ by restriction of the ambient one :

$$\forall u, v \in T_p\mathcal{M}, \quad \langle u|v\rangle_{T_p\mathcal{M}} := \langle u|v\rangle_{\mathbb{R}^D}$$

where $\langle\cdot|\cdot\rangle_{\mathbb{R}^D}$ is the Euclidean inner product in $\mathbb{R}^D$. From now on unless otherwise stated $\langle\cdot|\cdot\rangle$ will refer to the inner product in $\mathbb{R}^D$, and thus on the tangent spaces by restriction.

We define the metric tensor in coordinates $x = \varphi(p)$ (where φ is a diffeomorphism onto a subset of $\mathbb{R}^d$) as the Gram matrix of size $d \times d$ with components

$$g_{ij}(x) = \langle\partial_i|\partial_j\rangle, \quad 1 \leq i, j \leq d$$

where the basis vectors are implicitely computed at p. We will write g^{ij} for the components of the matrix inverse of the metric tensor. With these definitions at hand we can now describe the geodesics explicitely.

Proposition 2.1. *Geodesic equation*

If $x = \varphi(p)$ are local coordinates then a curve $\gamma : \mathbb{R} \to \mathcal{M}$ is a geodesic if and only if it satisfies the equation

$$\ddot{x}_t^k + \Gamma_{ij}^k (x_t)\, \dot{x}_t^i \dot{x}_t^j = 0, \quad k = 1, \ldots, d \tag{1.3}$$

where $\left(\dot{x}_t^i\right)_{i=1}^d$ (resp. $\left(\ddot{x}_t\right)_{i=1}^d$) are the coordinates of $\dot{x}_t$ (resp. $\ddot{x}_t$) and

$$\Gamma_{ij}^k (x_t) = \frac{1}{2} \sum_l g^{kl} \left(\frac{\partial}{\partial x^i} g_{lj} + \frac{\partial}{\partial x^j} g_{il} - \frac{\partial}{\partial x^l} g_{ij} \right) = \sum_l g^{kl} \langle \frac{\partial^2 \varphi^{-1}}{\partial x^i \partial x^j} | \partial_l \rangle.$$

Proof. If $x^i = \varphi^i(p)$ are local coordinates then with $\psi = \varphi^{-1}$ we have $\gamma_t = \psi(x_t)$ and

$$\dot{\gamma}_t = \sum_i \dot{x}_t^i \frac{\partial \psi}{\partial x^i} (x_t),$$

which implies

$$\ddot{\gamma}_t = \sum_i \ddot{x}_t^i \frac{\partial \psi}{\partial x^i} (x_t) + \sum_{i,j} \dot{x}_t^i \dot{x}_t^j \frac{\partial^2 \psi}{\partial x^i \partial x^j} (x_t).$$

Thus γ is a geodesic if and only if $\langle \ddot{\gamma}_t | \partial_l \rangle = \langle \ddot{\gamma}_t | \frac{\partial \psi}{\partial x^l} \rangle = 0$ for any $l = 1, \ldots, d$, i.e. if and only if

$$\sum_i \ddot{x}^i g_{il} + \sum_{i,j} \dot{x}^i \dot{x}^j \langle \frac{\partial^2 \psi}{\partial x^i \partial x^j} | \partial_l \rangle = 0, \quad l = 1, \ldots, d.$$

Mutiplying each side by g^{kl} (the inverse of the metric tensor) and summing over $l = 1, \ldots, d$ we see that this is equivalent to

$$\ddot{x}^k + \sum_{i,j} \dot{x}^i \dot{x}^j \sum_l g^{kl} \langle \frac{\partial^2 \psi}{\partial x^i \partial x^j} | \partial_l \rangle = 0, \quad k = 1, \ldots, d$$

Finally, noticing that

$$
\begin{aligned}
&\frac{\partial}{\partial x^i} g_{lj} + \frac{\partial}{\partial x^j} g_{il} - \frac{\partial}{\partial x^l} g_{ij} \\
&= \langle \frac{\partial^2 \psi}{\partial x^i \partial x^l} | \partial_j \rangle + \langle \partial_l | \frac{\partial^2 \psi}{\partial x^i \partial x^j} \rangle + \langle \frac{\partial^2 \psi}{\partial x^j \partial x^i} | \partial_l \rangle + \langle \partial_i | \frac{\partial^2 \psi}{\partial x^j \partial x^l} \rangle \\
&- \langle \frac{\partial^2 \psi}{\partial x^l \partial x^i} | \partial_j \rangle - \langle \partial_i | \frac{\partial^2 \psi}{\partial x^l \partial x^j} \rangle \\
&= 2 \langle \frac{\partial^2 \psi}{\partial x^i \partial x^j} | \partial_l \rangle
\end{aligned}
$$

gives the result. □

The quantities Γ_{ij}^k are called **Christoffel symbols of the second kind**. They depend both on the submanifold $\mathcal{M}$ and the chosen coordinate system x^i; for instance if $\mathcal{M}$ is a $2-$dimensional linear subspace and x^i are Cartesian coordinates then $\Gamma_{ij}^k = 0$, but this is no longer true in polar coordinates $y = (r, \theta)$. Indeed, in these coordinates we have

$$
g = \begin{bmatrix} 1 & 0 \\ 0 & r^2 \end{bmatrix},
$$

and therefore

$$
\Gamma_{22}^1 = \frac{1}{2} g^{1l} \sum_l \left\{ \frac{\partial}{\partial y^2} g_{l2} + \frac{\partial}{\partial y^2} g_{2l} - \frac{\partial}{\partial y^l} g_{22} \right\} = -\frac{1}{2} \frac{\partial}{\partial r} r^2 = -r \neq 0.
$$

Geodesics however do not depend on the chosen chart : if Equation 1.3 is satisfied in a coordinate system then it automatically holds in any other one, simply because Equation 1.3 is equivalent to $\nabla_{\dot{\gamma}} \dot{\gamma} = 0$, which did not require the construction of a coordinate system in the first place. In this sense, geodesics are an invariant notion.

The geodesic equation 1.3 is a system of non linear ordinary differential equations, thus the existence of a maximal solution to the Cauchy problem defined by 1.3 with initial conditions $\gamma_0 = p \in \mathcal{M}, \dot{\gamma}_0 = v \in T_p\mathcal{M}$ is only guaranteed locally in general (that is, on some interval $]-\delta, \delta[$ where $\delta > 0$ that depends on p). For some initial conditions $\gamma_0 = p \in \mathcal{M}, \dot{\gamma}_0 = v \in T_p\mathcal{M}$, let $I_{p,v}$ be the maximal interval of existence of a solution to the geodesic equation 1.3

and V_p the subset of the tangent space defined as

$$V_p = \{v \in T_p\mathcal{M} : 1 \in I_{p,v}\}.$$

In other words, $I_{p,v}$ is the set of initial velocities v for which we can shoot a geodesic γ satisfying $\gamma_0 = p, \dot{\gamma}_0 = v$ that still exists up to time 1. The **exponential map** at p is then defined by

$$\forall v \in V_p, \quad \exp_p(v) = \gamma_1$$

(this is unambiguous as two geodesics with the same initial conditions must coincide by standard ODE theory) and the **injectivity radius** (at p) $\mathrm{inj}_{\mathcal{M}}(p)$ as the supremum of the $r > 0$ such that $B_r(p) := \{v \in T_p\mathcal{M} : \|v\| < r\} \subset V_p$ and that $\exp_p : B_r(p) \to \exp_p(B_r(p))$ is a diffeomorphism. In [90] Sections 4.3 & 7.2 is it shown that $\mathrm{inj}_{\mathcal{M}}$ depends continuously on p and is positive at each p, so that in particular the global injectivity radius

$$\mathrm{inj}_{\mathcal{M}} := \inf_{p \in \mathcal{M}} \mathrm{inj}_{\mathcal{M}}(p)$$

is always positive for compact submanifolds. The exponential map at $p \in \mathcal{M}$ allows us to define a set of local coordinates ω

$$\omega = \exp_p^{-1}(q) \in T_p\mathcal{M}, \quad q \in B_{\mathrm{inj}_{\mathcal{M}}(p)}(p),$$

which are called **geodesic normal coordinates** (or simply normal coordinates) at p. Since ω is an element of $T_p\mathcal{M}$, these coordinates do not define a coordinate chart per se in the sense of Definition 1. However ω can be lifted back to $\mathbb{R}^d$ via a choice of orthonormal basis of $T_p\mathcal{M}$ yielding an inverse coordinate chart $\bar{\exp}_p : B_{\mathbb{R}^d}(0, \mathrm{inj}_{\mathcal{M}}(p)) \to \mathcal{M}$, and we can therefore define the metric tensor components g_{ij}. It turns out that the Riemannian structure in normal coordinates (with an orthonormal choice of basis) is remarkably simplified : $\mathcal{M}$ looks like a Euclidean space up to the second order when seen through Riemannian normal coordinates.

Proposition 2.2. *Let $\mathcal{M} \subset \mathbb{R}^D$ be a $d-$dimensional submanifold, $p \in \mathcal{M}$ and $\omega = (\omega^i)_{i=1}^d$ a* *system of normal coordinates at p (with an orthonormal choice of basis on $T_p\mathcal{M}$). Then*

i) The Christoffel symbols satisfy $\Gamma^k_{ij}(0) = 0$,

ii) The components of the metric tensor satisfy $g_{ij}(\omega) = \delta_{ij} + \mathcal{O}\left(\|\omega\|^2\right)$.

Proof. For any $\omega \in \mathbb{R}^d$ let ω_t be the curve with coordinates $\omega^i_t = t\omega^i, i = 1, \ldots, d$. Then ω_t is a geodesic passing through p at 0 with initial velocity ω.

i) Since $\ddot{\omega}_t = 0$ the geodesic equation 1.3 (where here $\phi^{-1} = \mathrm{e\bar{x}p}_p$) implies

$$\forall k = 1, \ldots, d, \quad \sum_{i,j} \Gamma^k_{ij}(\omega_t)\omega^i\omega^j = 0$$

In particular for $t = 0$ we get

$$\forall \omega \in \mathbb{R}^d, \quad \forall k = 1, \ldots, d, \quad \sum_{i,j} \Gamma^k_{ij}(0)\omega^i\omega^j = 0$$

Since $\Gamma^k_{ij}(0)$ is symmetric in its lower indices, this implies that $\Gamma^k_{ij}(0) = 0$.

ii) The time derivative at $t = 0$ of the components g_{ij} is equal to

$$\begin{aligned}
&\frac{\partial}{\partial t}\langle \frac{\partial \mathrm{e\bar{x}p}_p}{\partial x^i}(\omega_t) | \frac{\partial \mathrm{e\bar{x}p}_p}{\partial x^j}(\omega_t)\rangle \\
&= \sum_l \langle \frac{\partial^2 \mathrm{e\bar{x}p}_p}{\partial x^i \partial x^i}(\omega_t)\omega^l | \frac{\partial \mathrm{e\bar{x}p}_p}{\partial x^j}(\omega_t)\rangle + \langle \frac{\partial \mathrm{e\bar{x}p}_p}{\partial x^i}(\omega_t) | \frac{\partial^2 \mathrm{e\bar{x}p}_p}{\partial x^j \partial x^l}(\omega_t)\omega^l\rangle
\end{aligned}$$

But since by $i)$ we have

$$0 = \Gamma^k_{ij} = \sum_l g^{kl}\langle \frac{\partial^2 \mathrm{e\bar{x}p}_p}{\partial x^i \partial x^j} | \frac{\partial \mathrm{e\bar{x}p}_p}{\partial x^l}\rangle$$

and since g^{ij} is invertible this implies in particular

$$\frac{\partial}{\partial t}\langle \frac{\partial \mathrm{e\bar{x}p}_p}{\partial x^i}(\omega_t) | \frac{\partial \mathrm{e\bar{x}p}_p}{\partial x^j}(\omega_t)\rangle = 0$$

The result follows now by a Taylor expansion.

$\square$

Geodesic normal coordinates are therefore an interesting choice of coordinates as they do not introduce distortion in the metric tensor up to the second order in $\|\omega\|$. Sometimes other parametrisations can be preferred for regularity issues (see e.g. [39] or [2]), but in general the nice properties of the exponential map make it easier to work with.

Proposition 2.2 shows that for any $p \in \mathcal{M}$ and in normal coordinates, the quantity

$$J_p(\omega) = \sqrt{|g(\omega)|}$$

(square root of the determinant of the metric tensor) satisfies $J_p(\omega) = 1 + \mathcal{O}\left(\|\omega\|^2\right)$. This is extremely useful in computations involving integration on submanifolds, as we shall see in Chapter 3. Indeed, it turns out that a construction of an integration theory mirroring the one of Euclidean spaces exist on Riemannian manifolds, and J can be seen to be the density in local coordinates of the analog μ of the Lebesgue measure in this context, called the **Riemannian volume element**. In particular if $U \subset \mathcal{M}$ is open, $\phi : U \to \mathbb{R}^d$ is a coordinate chart (not necessarily normal) then for any smooth $f : \mathcal{M} \to \mathbb{R}$ with support in U, the integral of f with respect to μ is defined as

$$\int f d\mu := \int_{\varphi(U)} \left(f \circ \phi^{-1}\right)(\omega) \sqrt{|g(\omega)|} d\omega$$

This definition is easily seen to be independent of the chart φ. If f is not supported in a single chart, then a **partition of the unity** (see e.g. [90] Section 2.9.2) must be used in order to "glue" the charts together. Formally, a partition of the unity $(\chi_j)_{j\in I}$ subordinated to a family of coordinate charts $(\varphi_j)_{j\in I}, \varphi_j : U_j \cap \mathcal{M} \to \mathbb{R}^d, \quad \mathcal{M} = \cup_{j\in I} U_j$ is a family of smooth functions $\chi_j : \mathcal{M} \to [0,1]$ satisfying $\operatorname{supp}(\chi_j) \subset U_j$ and $\sum_{j\in I} \chi_j = 1$ (we can always assume that the sum is locally finite by paracompactness of $\mathcal{M}$, see [90] Section 2.9). We refer to [63] for more details about integration on Riemannian manifolds.

The concept of geodesics is also related to the one of **geodesic distance**, which is intuitively the length of a shortest path connecting two point on a given submanifold. For a curve

$\gamma : I \to \mathcal{M}$ over an open interval $I = [a, b] \subset \mathbb{R}, -\infty < a < b < +\infty$, we define its length $L(\gamma)$ as

$$L(\gamma) = \int_a^b \|\dot{\gamma}_t\| \, dt$$

The concept of length of curves then leads to a well defined metric space structure on $\mathcal{M}$:

Proposition 2.3. *[90] Lemma 4.2.3*

Let $\Omega_{p,q}$ is the set of smooth curves $\gamma : [0,1] \to \mathcal{M}$ satisfying $\gamma_0 = p, \gamma_1 = q$. Then the bivariate function d

$$d : (p, q) \mapsto d(p, q) = \inf_{\Omega_{p,q}} L(\gamma)$$

defines a metric on $\mathcal{M}$ called the geodesic distance.

remark 6. *By the triangle inequality, for any $p, q \in \mathcal{M}$ and curve $\gamma \in \Omega_{p,q}$ we have*

$$L(\gamma) = \int_0^1 \|\dot{\gamma}\| \geq \left\| \int_0^1 \dot{\gamma} \right\| = \|p - q\| \, .$$

Thus the geodesic distance is always greater than the Euclidean distance. The inequality cannot be reversed in general but Lemma 3 in [18] shows that if

$$r_0 := \frac{1}{\sup_{\gamma,t} \|\ddot{\gamma}_t\|}$$

(where the supremum is over every possible geodesic on $\mathcal{M}$ and time t), then for any geodesic in $\mathcal{M}$ from p to q with length less than or equal to πr_0 we have

$$2r_0 \sin\left(L(\gamma)/2r_0\right) \leq \|p - q\| \leq L(\gamma)$$

It turns out that the geodesics are related with the geodesic distance (hence the name), however the link is subtle. Indeed for instance on the circle $S^1 = \left\{x \in \mathbb{R}^2 : \|x\| = 1\right\}$ if we consider $x_0 = (1, 0)$ and $x_1 = \{\cos(\alpha), \sin(\alpha)\}$ for a small $\alpha > 0$, then the curve

$$\gamma_t = (\cos(t\alpha), \sin(t\alpha)), \quad t \in [0, 1]$$

has length α (which implies in particular that $d(x_0, x_1) \leq \alpha$), while the curve

$$\gamma_t = (\cos(-t\,(2\pi - \alpha)), \sin(-t\,(2\pi - \alpha)))\,, \quad t \in [0,1]$$

can be seen to be a geodesic with length $2\pi - \alpha > \alpha \geq d(x_0, x_1)$. This implies that in general geodesics are not automatically length minimising. However, for any point p and close enough point q there always exist a unique length minimising curve from p to q (up to time reparametrisation), and it is given by a geodesic.

Theorem 3. *Existence of minimal geodesics, [90] Theorem 4.4.4*
If $\mathcal{M} \subset \mathbb{R}^D$ is a submanifold, $p \in \mathcal{M}$, $0 < r < \mathrm{inj}_{\mathcal{M}}(p)$ and $v \in B_r(p) = \{u \in T_p\mathcal{M} : \|u\| < r\}$ then

$$d(p,q) = \|v\| \text{ where } q = \exp_p(v).$$

Moreover a curve $\gamma \in \Omega_{p,q}$ has length $d(p,q)$ if and only if there exists a smooth $\beta : [0,1] \to [0,1]$ such that $\beta(0) = 0, \quad \beta(1) = 1, \quad \dot{\beta} \geq 0$ and $\gamma_t = \exp_p(\beta(t)v)\,, \quad 0 \leq t \leq 1.$

We finish this section by introducing the **reach** of a submanifold. For a subset $K \subset \mathbb{R}^D$, we define its distance function as $d(a, K) = \inf_{p \in K} \|a - p\|\,, a \in \mathbb{R}^D$.

Definition 3. *Reach*
For $\mathcal{M}$ a submanifold of $\mathbb{R}^D$ define the medial axis of $\mathcal{M}$

$$\mathrm{Med}(\mathcal{M}) = \left\{a \in \mathbb{R}^D : \exists p \neq q \in \mathcal{M}, \quad \|a - p\| = \|a - q\| = d(a, \mathcal{M})\right\}$$

Then the reach τ of $\mathcal{M}$ is defined by

$$\tau = \inf_{p \in \mathcal{M}} d(p, \mathrm{Med}(\mathcal{M}))$$

Thus, for any closed submanifold $\mathcal{M}$ and point a at distance strictly less than τ from $\mathcal{M}$, the minimiser p^* of $p \mapsto d(p,a)$ over $\mathcal{M}$ exists (since $\mathcal{M}$ is closed) and is unique (since $d(a,\mathcal{M}) < \tau$). Since for any smooth curve γ_t satisfying $\gamma_0 = p^*$ we must have $\langle \dot{\gamma}_0 | p^* - a \rangle = 0$

(because 0 is a minimum of $t \mapsto \|\gamma_t - a\|^2$), this shows that $p^* - a \in N_{p^*}\mathcal{M}$. This suggests that the tubular neighbourhood

$$\mathcal{M}^\tau := \left\{a \in \mathbb{R}^D : d\,(a, \mathcal{M}) < \tau\right\}$$

can be parametrised locally by $p^* \in \mathcal{M}$ and the normal displacement $a - p^* \in N_{p^*}\mathcal{M}$. This is true by combining Proposition 2.14 in [72] with the construction of **Fermi coordinates** of [97] (Definition 10.1.3; after remarking that their assumptions are fulfilled when one assumes $\tau > 0$, thanks to Proposition 6.1 in [82] and the fact that $\mathrm{inj}_{\mathcal{M}} \geq \pi\tau$, as explained in the proof of Lemma A.1 in [2]). We will use Fermi coordinates in Chapter 4, with a precise control on the derivatives of the maps involved. More generally, a lower bound on the reach is often necessary in order to estimate either geometric properties of $\mathcal{M}$ or functions defined on it in statistical contexts (see for instance [16] for an impossibility result in density estimation or [2] for one in tangent space and extrinsic curvature estimation).

This concludes our introduction to the geometry of submanifolds of $\mathbb{R}^D$. We have not tried to be exhaustive, but only to provide the reader with the necessary preliminary background that will be used subsequently in Chapters 2 3 & 4. In particular we have not introduced the second fundamental form or the general concept of linear connection and covariant derivative, which are essential in differential geometry even beyond the Riemannian setting. Furthermore, we have purposefully avoided to present general manifolds (i.e. well behaved topological spaces which locally look like $\mathbb{R}^d$ while not being embedded in $\mathbb{R}^D, D \geq d$) as our work in Chapters 2 3 & 4 only focuses on submanifolds of $\mathbb{R}^D$. General non embedded manifolds are fundamental to numerous areas of mathematics and physics, and all of the results presented above in Sections 1.1.1 & 1.1.2 (at the exception of the reach and the normal spaces, although these can be defined for submanifolds of general manifolds) can be generalised to them with a more abstract construction of the tangent spaces. In particular the geodesics equation 1.3 can be generalised by defining the Christoffel symbols using only the derivatives of the metric tensor, yielding the so-called Levi-Civita connection. We refer the reader to [90, 72, 48, 63] or [97] for a more thorough introduction to manifolds.

1.2 Sobolev spaces and heat diffusion

In this section we introduce the necessary spectral theory background that we will use in Chapters 2 & 3, in particular the theory of heat diffusion and the construction of heat kernels. The content presented in Section 1.2.1 focuses on the case of compact manifolds and will be used in Chapter 2. Section 1.2.1 also presents the basic properties of the heat kernel on compact manifolds, which provides intuition regarding the content of Section 1.2.2 where the spectral theory of finite graphs is studied. Finally, Section 1.2.2 ends with heuristic arguments which allow to view random geometric graphs as approximations of submanifolds; this point of view will crucially be used in Chapter 3.

1.2.1 Compact submanifolds

Let $\mathcal{M} \subset \mathbb{R}^D$ be a submanifold as defined in Section 1.1. We are going to describe the generalisations of the **gradient** and **Laplacian** operators of classical multivariate calculus to spaces of functions over $\mathcal{M}$, as well as subsequent functional analytic properties. We start by defining the concept of **differential** of functions on manifolds.

Definition 4. *Differential*
If $\mathcal{M} \subset \mathbb{R}^D$ be a submanifold of $\mathbb{R}^D, U \subset \mathbb{R}^D$ open, $\varphi : U \cap \mathcal{M} \to \mathbb{R}^d$ a coordinate chart in the sense of Definition 1 and $f : \mathcal{M} \to \mathbb{R}$ differentiable (see Remark 1). We define the differential of f at $p \in U \cap \mathcal{M}$ as the linear map $df_p : T_p\mathcal{M} \to \mathbb{R}$ defined by

$$\forall v \in T_p\mathcal{M}, \quad df_p.v = \frac{d}{dt}_{t\to 0} f\left(\gamma_t\right)$$

where γ_t is any differentiable curve in $\mathcal{M}$ satisfying $\gamma_0 = p, \dot{\gamma}_0 = v$.

remark 7. *Using the chain rule, Definition 4 is easily seen to be independent of the choice of the curve γ. Moreover using the coordinate basis vectors ∂_l of the chart φ we have*

$$df_p.\partial_l = \frac{\partial\left(f \circ \varphi^{-1}\right)}{\partial x^l}(x),$$

where $x = \varphi(p)$.

Definition 4 thus extends the usual concept of differential as a linear form on the tangent space. We now define the gradient of a function by Riesz representation of the differential through the inner product structure.

Definition 5. *Gradient*
If $f : \mathcal{M} \to \mathbb{R}$ is differentiable we define the gradient of f at $p \in \mathcal{M}$ as the unique element $(\nabla f)_p \in T_p\mathcal{M}$ satisfying

$$\forall v \in T_p\mathcal{M}, \quad df_p.v = \langle (\nabla f)_p | v \rangle.$$

remark 8. *The element $(\nabla f)_p$ in Definition 5 is unambiguously defined by definiteness of the inner product. Written in a local chart we have*

$$(\nabla f)_p = \sum_{l=1}^{d} \frac{\partial \left(f \circ \varphi^{-1}\right)}{\partial x^l} \partial_l.$$

Is is immediate to check that the coordinates $\frac{\partial(f \circ \varphi^{-1})}{\partial x^l}$ of ∇f in the local coordinate basis $(\partial_l)_{l=1}^d$ are $\mathcal{C}^{k-1}$ in x if f is $\mathcal{C}^k$ and $\mathcal{M}$ is smooth : in this sense $\nabla f : \mathcal{M} \to T\mathcal{M} := \cup_{p \in \mathcal{M}} T_p\mathcal{M}$ is a $\mathcal{C}^{k-1}$ vector field, i.e. a $\mathcal{C}^{k-1}$ assignment of a tangent vector at each point of $\mathcal{M}$. Here, because $\mathcal{M}$ is a submanifold of $\mathbb{R}^D$, every tangent space is a linear subspace of $\mathbb{R}^D$, and therefore this notion of smooth vector field is well defined. In the intrinsic setting (i.e. when $\mathcal{M}$ is a general manifold and not a submanifold of $\mathbb{R}^D$), a rigorous definition of vector fields can be made by first describing a manifold structure on $T\mathcal{M} := \cup_{p \in \mathcal{M}} T_p\mathcal{M}$ (the tangent bundle, i.e. the collection of tangent spaces) and defining the notion of $\mathcal{C}^k$ maps and their differentials between two general manifolds, see e.g. [90] Section 2.8.2 and Definition 2.8.18 (this is stated for smooth maps, but can be readily generalised to $\mathcal{C}^k$ maps).

As in the Euclidean case, the gradient represents the direction of the greatest change of f over $\mathcal{M}$. We can also define the Laplacian of a $\mathcal{C}^2$ function f over a submanifold $\mathcal{M}$.

Definition 6. *Laplace-Beltrami operator*
If f is a $\mathcal{C}^2$ function on $\mathcal{M}$ we define the Laplacian of f as the function Δf written in local

coordinates as

$$\Delta f(p) = -|g|^{-1/2}\sum_{i=1}^{d}\frac{\partial}{\partial x^i}\left(|g|^{1/2}\sum_{j=1}^{d}g^{ij}\frac{\partial\left(f\circ\varphi^{-1}\right)}{\partial x^j}\right)$$

where $|g|$ is the determinant of the metric tensor. The operator Δ is called the ***Laplace-Beltrami operator*** *of $\mathcal{M}$.*

remark 9. *Using the chain rule, the Laplace-Beltrami operator as defined in Definition 6 can be seen to be independent of the coordinate chart. Notice that if $\mathcal{M} = \mathbb{R}^d$ then using the Cartesian chart and the associated coordinates $(x_i)_{i=1}^d$ we actually find $\Delta f = -\sum_{i=1}^d \frac{\partial^2 f}{\partial x_i^2}$, thus Definition 6 generalises the usual Laplacian operator (up to a minus sign).*

Let's consider the space $L^2 = L^2(\mathcal{M},\mu)$ of measurable and square integrable functions with respect to the Riemannian volume μ (see Section 1.1). Then using the manifold version of Green's identity (see [19] Theorem 9.21) we get, for any $f, g \in \mathcal{C}^\infty(\mathcal{M})$ (the space of smooth functions on $\mathcal{M}$)

$$\langle\Delta f|g\rangle_{L^2} = \int\langle\nabla f|\nabla g\rangle_{L^2}d\mu,$$

so that $(\Delta, \mathcal{C}^\infty(\mathcal{M}))$ is a symmetric unbounded operator on L^2 (see [19] for an introduction to unbounded operators and spectral theory). Moreover via $f = g$ we have

$$\langle\Delta f|f\rangle_{L^2} = \int\|\nabla f\|^2 d\mu \geq 0,$$

and therefore $(\Delta, \mathcal{C}^\infty(\mathcal{M}))$ is also a positive operator (which explains the choice of a minus sign in Definition 6).

It turns out that Δ can be extended to a non-negative **self adjoint operator** on an appropriate domain. We first introduce the **Sobolev spaces** on $\mathcal{M}$. From now on we assume that $\mathcal{M}$ is compact.

Definition 7. *Sobolev spaces*

If $(U_j,\varphi_j)_{j\in I}$ is an atlas of $\mathcal{M}$ (we may assume I finite by compactness of $\mathcal{M}$) and $(\chi_j)_{j\in I}$ is an associated partition of the unity, for $s > 0$ we define the Sobolev space $H^s(\mathcal{M})$ as the *space of distributions f over $\mathcal{M}$ satisfying*

$$\|f\|^2_{H^s(\mathcal{M})} := \sum_{j\in I} \left\|(\chi_j f)\circ\varphi_j^{-1}\right\|^2_{H^s\left(\mathbb{R}^d\right)} < +\infty$$

remark 10. *Definition 7 is unambiguous since, for another atlas (V_i, ψ_i) with associated partition of the unity ρ_i, using the diffeomorphisms and multiplication theorems of Sections 4.2 & 4.3 of [109] and the smoothness of $\chi_j, \rho_i, \varphi_j, \psi_i$ we have for any f*

$$\begin{aligned}
\left\|(\chi_j f)\circ\varphi_j^{-1}\right\|_{H^s\left(\mathbb{R}^d\right)} &= \left\|\sum_i (\rho_i\chi_j f)\circ\psi_i^{-1}\circ\psi_i\circ\varphi_j^{-1}\right\|_{H^s\left(\mathbb{R}^d\right)} \\
&\le \sum_i \left\|(\rho_i\chi_j f)\circ\psi_i^{-1}\circ\psi_i\circ\varphi_j^{-1}\right\|_{H^s\left(\mathbb{R}^d\right)} \\
&= \sum_i \left\|\left(\chi_j\circ\varphi_j^{-1}\right)\cdot\left((\rho_i f)\circ\psi_i^{-1}\right)\circ\left(\psi_i\circ\varphi_j^{-1}\right)\right\|_{H^s\left(\mathbb{R}^d\right)} \\
&\le C\sum_i \left\|(\rho_i f)\circ\psi_i^{-1}\right\|_{H^s\left(\mathbb{R}^d\right)}
\end{aligned}$$

for some constant $C > 0$ that depends only on the atlases and s, d.

Then, as a consequence of Theorem 9.24 in [19], Δ can be extended to a non-negative self-adjoint operator on the domain $\mathcal{D}(\Delta) \subset L^2$ consisting of functions $u \in H^1(\mathcal{M})$ with weak Laplacian $\Delta u \in L^2$ (i.e. such that $\phi \in \mathcal{C}^\infty(\mathcal{M}) \mapsto \int u\Delta\phi$ extends to a bounded linear functional in L^2). Moreover the operator Δ can be shown to have the following **spectral decomposition** :

Theorem 4. *Spectral decomposition of the Laplace-Beltrami operator*
There exists an L^2-orthonormal basis $(u_k)_{k\ge 0}$ of L^2 satsfying the following three conditions :

i) each u_k is an eigenfunction of Δ associated with an eigenvalue $\lambda_k \ge 0$,

ii) for $k \ge 0$ the eigenvalues satisfy $\lambda_k \le \lambda_{k+1}$ and $\lim_{k\ge\infty}\lambda_k = +\infty$,

iii) each u_k is an element of $\mathcal{C}^\infty(\mathcal{M})$.

Proof. The result is essentially a corollary of the compactness of $(I+\Delta)^{-1} : L^2 \to L^2$ and the

Hilbert-Schmidt theorem ([19] Theorem 4.21). Details of the proof can be found in Section 9.4 of [19]. □

Theorem 4 (sometimes refered to as a Sturm-Liouville decomposition) has far-reaching implications. In particular it allows to develop a theory of Sobolev spaces over $\mathcal{M}$ entirely in terms of the spectral decomposition of Δ (in analogy with the characterisation of Sobolev spaces on $\mathbb{R}^d$ in terms of Fourier transform), a principle that we will use in Chapter 2. More precisely, the following theorem (which is a compilation of classical results from the functional analysis litterature) shows that $f \in H^s(\mathcal{M})$ in the sense of Definition 7 if and only if its coefficients $(\langle u_k|f\rangle)_{k\geq 0}$ in the Laplace-Beltrami eigenbasis $(u_k)_{k\geq 0}$ satisfy a summability condition.

Theorem 5. *For $s > 0$, the following statements are equivalent :*

i) $f \in H^s(\mathcal{M})$,

ii) $\|f\|^2_{H^{s,2}(\mathcal{M})} = \sum_{k\geq 0} (1+\lambda_k)^s |\langle u_k|f\rangle|^2 < +\infty$,

iii) $\|f\|^2_{H^{s,3}(\mathcal{M})} = \sum_{k\geq 0} (1+\lambda_k^s) |\langle u_k|f\rangle|^2 < +\infty$.

Moreover, there exists constants $c_1, c_2, c_3 > 0$ such that for any $f \in H^s(\mathcal{M})$,

$$\|f\|_{H^s(\mathcal{M})} \leq c_1 \|f\|_{H^{s,2}(\mathcal{M})} \leq c_2 \|f\|_{H^{s,3}(\mathcal{M})} \leq c_3 \|f\|_{H^s(\mathcal{M})}.$$

Proof. See [37] Section 3 & 5. □

Functions in the Sobolev spaces $H^s(\mathcal{M})$ are thus as smooth as usual Sobolev functions on $\mathbb{R}^d$, once seen through charts. Furthermore, the well known Sobolev embedding theorem (see [108] Section 2.7) implies that if $s > d/2$ then we actually have a continuous injection $H^s(\mathbb{R}^d) \hookrightarrow \mathcal{C}^{s-d/2}(\mathbb{R}^d)$ where $\mathcal{C}^{s-d/2}(\mathbb{R}^d)$ is the Hölder space (see e.g. [108] Section 2.2.2 or [109] Section 1.2.2). This implies the following result :

Theorem 6. *Sobolev embedding theorem*

With $\mathcal{C}^s(\mathcal{M})$ the space of functions $f : \mathcal{M} \to \mathbb{R}$ satisfying

$$\|f\|_{\mathcal{C}^s(\mathcal{M})} := \sum_{j\in I} \left\| (\chi_j f) \circ \varphi_j^{-1} \right\|_{\mathcal{C}^s(\mathbb{R}^d)} < +\infty$$

where $(U_j, \varphi_j)_{j\in I}$ is an arbitrary atlas and $(\chi_j)_{j\in I}$ an associated partition of the unity, the definition of $\mathcal{C}^s(\mathcal{M})$ is independent of the choice of $(U_j, \varphi_j, \chi_j)_{j\in I}$ and we have the continuous embedding

$$\forall s > d/2, \quad H^s(\mathcal{M}) \hookrightarrow \mathcal{C}^{s-d/2}(\mathcal{M})$$

Proof. The independence of the choice of $(U_j, \varphi_j, \chi_j)_{j\in I}$ goes as in Remark 10, except in this case it is even simpler since we do not have to use a diffeomorphism or multiplication theorem (as here the smoothness is measured in the classical Hölder sense), and the embedding is then a direct consequence of the Euclidean Sobolev embedding theorem together with Definition 7. $\square$

In particular, Theorem 6 implies the following fundamental property of the spaces $H^{s+d/2}(\mathcal{M}), s > 0$:

Corollary 6.1. *The spaces $H^{s+d/2}(\mathcal{M}), s > 0$ (which contain only classical functions) have the **reproducing property** : the evaluation maps*

$$\delta_p : f \mapsto f(p), \quad p \in \mathcal{M}$$

are norm continuous. As a consequence for each $s > 0$ and $p \in \mathcal{M}$ we can find a Riesz representation of δ_p by an element $k_s(p, \cdot) \in H^{s+d/2}(\mathcal{M})$ satisfying

$$\forall f \in H^{s+d/2}(\mathcal{M}), \quad f(p) = \langle k_s(p, \cdot) | f \rangle_{H^{s+d/2}(\mathcal{M})}$$

Moreover k_s is jointly continuous and

$$k_s(p, q) = \sum_{k\geq 0} (1 + \lambda_k)^{-(s+d/2)} u_k(p) u_k(q)$$

where the convergence is absolute and uniform on $\mathcal{M} \times \mathcal{M}$.

Proof. The existence of a kernel k_s follows by Theorem 6 and Riesz representation. For the continuity property we use the reproducing property to find

$$k_s(p,q) = \langle k_s(p,\cdot)|k_s(q,\cdot)\rangle_{H^{s+d/2}(\mathcal{M})},$$

and therefore it suffices to show that $p \mapsto k_s(p,\cdot) \in H^{s+d/2}(\mathcal{M})$ is norm continuous. For this we use the dual characterisation of norms in Hilbert spaces :

$$\begin{aligned}\|k_s(p,\cdot) - k_s(q,\cdot)\|_{H^{s+d/2}(\mathcal{M})} &= \sup_{\|f\|_{H^{s+d/2}(\mathcal{M})}\leq 1} |\langle k_s(p,\cdot) - k_s(q,\cdot)|f\rangle| \\ &= \sup_{\|f\|_{H^{s+d/2}(\mathcal{M})}\leq 1} |f(p) - f(q)|\,.\end{aligned}$$

The Sobolev embedding theorem 6 now implies that for some $C > 0$ this can be subsequently bounded by

$$\sup_{\|f\|_{\mathcal{C}^s(\mathcal{M})}\leq C} |f(p) - f(q)| \leq Cd(p,q)^{s\wedge 1},$$

where $d(p,q)$ is the geodesic distance on $\mathcal{M}$. This shows continuity of $p \mapsto k_s(p,\cdot) \in H^{s+d/2}(\mathcal{M})$ (actually even $s \wedge 1-$Hölder continuity). The absolute and uniform convergence is now a consequence of Mercer's theorem, see e.g. [32] Theorem 4.49. □

remark 11. *In general a Hilbert space of functions satisfying the reproducing property is called a* ***reproducing kernel Hilbert space (RKHS)****.*

The bivariate function $k_s : \mathcal{M}\times\mathcal{M} \to \mathbb{R}$ is called the **reproducing kernel** of $H^{s+d/2}(\mathcal{M})$. It is positive semi-definite in the sense that p_t is symmetric in its arguments and that for any $L \geq 1, \quad \alpha \in \mathbb{R}^L$ and $p_1, \ldots, p_L \in \mathcal{M}$ we have

$$\sum_{i,j=1}^{L} \alpha_i\alpha_j k_s(p_i,p_j) = \left\|\sum_{i=1}^{L} \alpha_i k_s(p_i,\cdot)\right\|_{H^{s+d/2}(\mathcal{M})} \geq 0.$$

We will use Corollary 6.1 in Chapter 2 in order to construct and analyse the behaviour of

Matérn Gaussian processes on compact manifolds.

We finish this section by saying a few words on the heat semigroup. The family $(P_t)_{t\geq 0}$ of bounded linear operators defined as $P_0 = I_{L^2}$ and for $t > 0$,

$$\forall k \geq 0, (P_t f) := \sum_{k\geq 0} e^{-t\lambda_k} \langle u_k | f \rangle_{L^2} u_k$$

(the sum is absolutely convergent in L^2) satisfies $P_t \circ P_s = P_{t+s}$ for any $t, s > 0$ and can be seen to satisfy $\|P_t f\|_{L^2} \leq \|f\|_{L^2}$ and $\lim_{t\to 0} \|P_t f - f\|_{L^2} = 0$ for any $f \in L^2$: in this sense $(P_t)_{t\geq 0}$ is a contraction semigroup of operators (see e.g. [20] Section 7 and the references therein). Let $\mathcal{H}^t$ by the Hilbert space defined as the subspace of L^2 by the f satisfying

$$\|f\|_{\mathcal{H}^t}^2 = \sum_{k\geq 0} e^{\lambda_k t} \left| \langle u_k | f \rangle_{L^2} \right|^2 < +\infty.$$

It is shown in [37] that for any $s, t > 0$ we have the following continuous embeddings

$$\mathcal{H}^t \hookrightarrow H^s(\mathcal{M}) \text{ and } \mathcal{H}^t \hookrightarrow \mathcal{C}^s(\mathcal{M}).$$

In particular, arguing as in Corollary 6.1 we find :

Corollary 6.2. *The spaces $\mathcal{H}^t, t > 0$ have the reproducing property. The reproducing kernel p_t of $\mathcal{H}^t$ is jointly continuous and is given by the formula*

$$\forall x, y \in \mathcal{M}, \quad p_t(x, y) = \sum_{k\geq 0} e^{-t\lambda_k} u_k(x) u_k(y),$$

where the convergence is absolute and uniform.

It turns out that the kernel p_t is closely related to the theory of heat diffusion on Riemannian manifolds : it is actually equal to the so-called **heat kernel** of $\mathcal{M}$.

Theorem 7. *Heat kernel on a compact manifold*

The kernel

$$p : \left(\begin{array}{ccc} (0,\infty)\times\mathcal{M}\times\mathcal{M} & \to & \mathbb{R} \\ (t,x,y) & \mapsto & p_t(x,y) \end{array} \right)$$

is of class $\mathcal{C}^\infty$ in the three variables (t,x,y) and satisfies

i) for any $x \in \mathcal{M}$, $\quad p_t(x,\cdot)$ is a probability density on $\mathcal{M}$ with respect to μ,

ii) for every $f \in L^2$, $\quad P_t f = \int p_t(x,\cdot) f(x)\mu(dx)$,

*iii) for any $y \in \mathcal{M}$ the function $(t,x) \mapsto p_t(x,y)$ is smooth and satisfies the **heat equation** on $(0,\infty)\times\mathcal{M}$:*

$$\frac{\partial}{\partial t} p_t(x,y) = -\Delta p_t(\cdot,y),$$

where the Laplacian acts on the variable x only.

Proof. It is shown in [63] Theorem 10.13 that the reproducing kernel p_t of $\mathcal{H}^t$ is indeed the heat kernel of $\mathcal{M}$ since they both have the same expression

$$\sum_{k\geq 0} e^{-t\lambda_k} u_k(x) u_k(y).$$

Then items *ii*) & *iii*) are given in [63] Theorems 7.13 and 10.13. Item *i*) is given in [63] Section 8.4. □

Item *iii*) of Theorem 7 together with $P_t f \to f$ as $t \to 0$ says that the heat kernel is the fundamental solution to the heat equation on $\mathcal{M}$. In this sense and thanks to item *i*) the heat kernel is a natural generalisation of the Gaussian kernel on $\mathbb{R}^d$ to compact manifolds (for non compact manifolds, the heat kernel may fail to be a probability density in general, see [63] Section 8.4). In light of the analogies between the heat kernel and the Gaussian kernel, it is then interesting to wonder whether or not there exists $\sigma > 0$ such that

$$p_t(x,y) \sim \left(2\pi\sigma\sqrt{t}\right)^{-d} \exp\left(-\frac{d(x,y)^2}{\sigma^2 t}\right)$$

in some sense. While there is not equality in general, an asymptotic expansion of the form

$$\forall N \geq 1, \quad p_t(x,y) = \frac{\exp\left(-\frac{d(x,y)^2}{4\pi t}\right)}{(4t)^{d/2}} \sum_{l=0}^{N} \theta_l(x,y) t^l + o_{t\to 0}\left(t^N\right) \tag{1.4}$$

is proved for compact connected oriented manifolds in [93] Chapter 3, for some functions θ_l (see also [7] Chapter 5). As explained in [31] Chapter 6, equation 1.4 implies the following corollary :

Corollary 7.1. *Weyl's law*
There exists a constant c that depends only on $\mathcal{M}$ such that $\lambda_k \sim ck^{2/d}$.

We will use this property in Chapter 2. A weaker form of Equation 1.4 are the so-called **heat kernel bounds**, i.e. inequalities of the form

$$\forall x, y \in \mathcal{M}, \quad A_1 \frac{\exp\left(-\frac{d(x,y)^2}{c_1 t}\right)}{(4t)^{d/2}} \leq p_t(x,y) \leq A_2 \frac{\exp\left(-\frac{d(x,y)^2}{c_2 t}\right)}{(4t)^{d/2}} \tag{1.5}$$

for some $A_i, c_i > 0$. Inequalities of the type 1.5 have been shown to hold for compact manifolds and even beyond this setting such as on **graphs** or fractals (where the notion of heat kernel can be generalised, see e.g.[62]).

In the next subsection, we present the spectral theory of finite graphs. The exposition will be substantially easier as it only requires finite dimensional linear algebra arguments.

1.2.2 Graphs

A (finite, undirected) graph $G = (V, E)$ is a tuple where V is a finite set of **vertices** and E is a set of **edges**, formally defined as subset of $\{0,1\}^V$. We write $x \sim y \iff \{x,y\} \in E$ the neighbouring relation on V, $D = \operatorname{diag}(\mu)$ for the **degrees** matrix $D_{xy} = \delta_x^y \mu_x, \mu_x = \#\{y : x \sim y\}$ and A for the **adjacency matrix** $A_{xy} = \mu_{xy} = \mathbb{1}_{x\sim y}$.

Definition 8. *Graph Laplacian*

*We define the **graph Laplacian** as $L = I - D^{-1}A$. Equivalently*

$$\forall f \in \mathbb{R}^V, (Lf)(x) = \frac{1}{\mu_x} \sum_{y \sim x} (f(x) - f(y)).$$

Other definitions of the graph Laplacian can be made, see for instance [49].

Let ν be the normalized degree measure $\nu = \frac{\mu}{\mu(V)}$ and the associated inner product $\langle \cdot | \cdot \rangle_{L^2(\nu)}$ defined by

$$\forall f, g \in L^2(\nu), \langle f|g \rangle_{L^2(\nu)} = \sum_{y \in V} f(y) g(y) \nu_y.$$

We then have the following identity :

Proposition 7.1. *For every $f, g \in \mathbb{R}^V$:*

$$\langle f|Lg \rangle_{L^2(\nu)} = \frac{1}{2\mu(V)} \sum_{x \sim y} (f(x) - f(y))(g(x) - g(y))$$

As a consequence, the graph Laplacian L is a nonnegative self-adjoint operator from $L^2(\nu)$ to itself.

Proof. We have

$$\begin{aligned}
\langle f|Lg \rangle_{L^2(\nu)} &= \sum_{x \in V} f(x) \left\{ \frac{1}{\mu_x} \sum_{y \sim x} g(x) - g(y) \right\} \nu_x \\
&= \frac{1}{\mu(V)} \sum_{x \sim y} f(x)(g(x) - g(y)) \\
&= \frac{1}{\mu(V)} \sum_{x \sim y} (f(x) - f(y))(g(x) - g(y)) - \langle f|Lg \rangle_{L^2(\nu)},
\end{aligned}$$

and thus the result. $\square$

remark 12. *If $V = h\mathbb{Z}^d \cap \left\{ x \in \mathbb{R}^d : \max_{1 \leq j \leq d} |x_j| < 1 \right\}$ for some $h > 0$, then by a second* *order Taylor expansion, if $f : \mathbb{R}^d \to \mathbb{R}$ is $\mathcal{C}^2$*

$$(Lf)(0) = \frac{1}{\mu_0}\sum_{y\sim 0} f(0) - f(y) \approx -\frac{1}{2d}\sum_{i=1}^{d} h^2 \frac{\partial^2 f}{\partial x_i^2}(0) = \frac{h^2}{2d}(\Delta f)(0),$$

where Δ is the standard mutivariate Laplacian operator $-\sum_{i=1}^d \frac{\partial^2}{\partial x_i^2}$. The same result holds at other points in V and explains the intuition behind Definition 8.

By self-adjointness and the (finite dimensional) spectral theorem there exists an $L^2(\nu)-$orthonormal eigenbasis $(u_j)_{j=1}^N, N = \#V$ for L, associated with eigenvalues $0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{N-1}$ ($\lambda_0 = 0$ since the function $\mathbb{1} \in \mathbb{R}^V$ identically equal to 1 satisfies $L\mathbb{1} = 0$). In analogy with the situation on compact manifolds, for $t > 0$ we can then define the **heat kernel of** G associated with the operator L as the bivariate function

$$p_t : (x, y) \in V \times V \mapsto p_t(x, y) = \sum_{j=0}^{N-1} e^{-t\lambda_j} u_j(x)u_j(y).$$

It can be seen to satisfy the heat equation on the graph

$$\forall x, y \in V, \quad \frac{\partial}{\partial t} p_t(x, y) = -(Lp_t(\cdot, y))(x),$$

and $p_0(x, y) = \delta_{xy} = \mathbb{1}_{x=y}$. Is is moreover the kernel of the linear operator e^{-tL} (defined by matrix exponentiation) with respect to the measure ν in the sense that for any $f \in \mathbb{R}^V, t > 0$ and $x \in V$ we have

$$\left(e^{-tL} f\right)(x) = \sum_{y\in V} p_t(x, y)f(y)\nu_y.$$

In other words $p_t(x, y) = \frac{e^{-tL}(x,y)}{\nu_y}$. Furthermore, e^{-tL} can be seen to be a stochastic matrix (as it can be checked that $-L$ is itself a $Q-$matrix, see [83] Section 2.1), i.e. $p_t(x, \cdot)$ is a probability density on V with respect to the measure ν.

Thus the situation is very similar to the one of compact manifolds with V (resp. L, ν) replacing V (resp. Δ, μ), only that no functional analytic theory is needed as the graph is assumed to be finite.

More than just a formal analogy, graphs can actually be used in a concrete way to approximate in some sense a given manifold $\mathcal{M}$: if $V = \{x_1, \ldots, x_N\}$ is a set of points on a submanifold $\mathcal{M}$ of $\mathbb{R}^D$, and the edges are defined as $x_i \sim x_j \iff \|x_i - x_j\|_{\mathbb{R}^D} < h$ (for some choice of **connectivity parameter** $h > 0$), then the resulting graph G can be thought of as an approximation of the submanifold $\mathcal{M}$. Indeed, if for instance

$$\mathcal{M} = \{e_\theta := (\cos(\theta), \sin(\theta)) : \theta \in \mathbb{R}\}$$

and $x_j = e_{2j\pi/N}$ for $0 \leq j \leq N-1$ then for $\|x_0 - x_1\|_{\mathbb{R}^D} < h < \|x_0 - x_2\|_{\mathbb{R}^D}$ the resulting graph G is a polygonal approximation of $\mathcal{M}$, i.e.

$$E = \{\{x_i, x_i\}, \{x_i, x_{i+1}\} : 0 \leq i \leq N-2\} \cup \{\{x_{N-1}, x_{N-1}\}, \{x_{N-1}, x_0\}\}.$$

The presence of the self-loops $\{x_i, x_i\}, \quad 1 \leq i \leq N-1$ is only a formal artefact and typically does not introduce issues (for instance, it can be immediately checked that the graph Laplacian L remains unchanged when these are present or not). Obviously the choice of h is crucial, as h too small (resp. too large) will force G to have no edges other than the self-loops (resp. to contain all possible edges $\{x_i, x_j\}, \quad 1 \leq i, j \leq N-1$).

A graph G constructed as above where the vertices are drawn *i.i.d.* according to some probability distribution μ_0 supported on $\mathcal{M}$ is called a **random geometric graph** (on $\mathcal{M}$), and the point of this construction is that no knowledge of $\mathcal{M}$ is needed to construct G, but only a set of vertices that we know to be contained in $\mathcal{M}$ a priori. This is particularly useful for statistical applications when $\mathcal{M}$ is unknown or does not admit a closed-form description, as we will see for instance in Chapter 3 (see also [43, 45, 61] or [96]). Note that we focus here on random geometric graphs, but other choice of random graph models can also induce useful graph approximations to submanifold, see for instance the case of $k-$nearest neighbours graphs [22].

Building on the intuition that random geometric graphs can accurately approximate submanifolds (for an appropriate choice of connectivity parameter $h > 0$), it is interesting to wonder whether or not their graph Laplacians can be compared to the Laplace-Beltrami operator of the submanifold in question. It actually does not hold under some assumptions : as showed in [64] & [49], under technical conditions on μ_0 and $\mathcal{M}$ the graph Laplacian does converge in some sense to a "limiting" differential operator on $\mathcal{M}$, but it is not equal to the Laplace-Beltrami operator, unless the probability distribution μ_0 is uniform, i.e. proportional to the Riemannian volume (see also [42] where a similar phenomenon for the heat kernels is shown). In order to recover the Laplace-Beltrami operator it is necessary to introduce further normalisations in the construction of the random geometric graph, therefore in general care must be taken when approximating a manifold with a graph.

Sometimes, establishing graph heat kernel bounds (or weak versions thereof) similar as Equation 1.5 in the continuous setting is useful for the problem at hand : in particular in Chapter 3, building on [58] and using the proof techniques of [9] we will show how to obtain a (weak version of) Weyl's type asymptotic 7.1 for the graph Laplacian eigenvalues of a random geometric graph approximating an unknown submanifold (with high probability), which will be subsequently used to bound the $L^{\infty}(\nu)$ norm of functions over V by a multiple of their $L^{2}(\nu)$ norms in this context.

This concludes our introduction to the spectral theory of compact manifolds and graphs. More details about the properties of the Laplacian operators and heat kernels on manifolds can be found in [63, 19] or [93] in the manifold case, as well as [9] for graphs. We have also only presented the properties of Sobolev spaces on manifolds that we will use in Chapter 2; more results (including properties of function spaces on $\mathbb{R}^d$) can be found in [109, 97] or [56] (where in the last reference relationships with wavelets theory are presented). Moreover, heat kernel bounds in the general setting of Dirichlet spaces (abstract spaces that in particular generalise manifolds and graphs) have been shown in [35] and [67] to lead to characterisations of some classes of functional spaces in terms of the heat kernel, together with the construction of frames for these functional spaces. This has subsequently been used in statistical applications, see e.g. [26, 33].

1.3 Bayesian asymptotics

In this thesis, our primary aim will be to derive **asymptotic frequentist properties** for some classes of Bayesian procedures. The original works of Bayes [12] and Laplace [40] introduced the foundations of what is now commonly referred to as the Bayesian point of view on probability theory and its subsequent use as a statistical estimation procedure. As mentioned previously, deriving frequentist properties of Bayesian procedures in non-conjugate situations is challenging since the posterior distribution lacks a closed-form expression. We will detail in Section 1.3.1 general theorems and approaches before considering a few examples in Section 1.3.2.

1.3.1 Posterior consistency and contraction rates

The first modern investigations on this matter include the contributions of [71, 99, 46, 47, 11]. In [99], Lorraine Schwartz laid the foundations of what will become later the prior mass and testing technique by investigating the behaviour of posterior distributions over densities in the *i.i.d.* sampling model.

Let us precise our assumptions : we consider a measured space $(\mathcal{X}, \mu)$ and a set $\mathcal{F}$ of probability densities with respect to μ (i.e. we consider a *dominated model*) equipped with a $\sigma-$field $\mathcal{T}$. We now formulate a frequentist assumption, namely that there exists a density f_0 (not necessarily an element of $\mathcal{F}$) with respect to μ generating the sample $\mathbb{X}^n := (X_i)_{i=1}^n$:

$$X_i \overset{i.i.d.}{\sim} f_0, \quad i = 1, \dots, n. \tag{1.6}$$

We write P_0 (resp. P_0^∞) for the distribution of X_1 (resp. $\mathbb{X}^\infty := (X_i)_{i \geq 1}$). Then, for a prior Π on $\mathcal{F}$ (i.e. a probability distribution on $(\mathcal{F}, \mathcal{T})$) and under appropriate measurability conditions (for instance it suffices for μ to be $\sigma-$finite and $(f, x) \mapsto f(x)$ to be jointly measurable, see the discussion in Chapter 1 of [55]), **Bayes' formula**

$$\Pi\left[df | \mathbb{X}^n\right] := \frac{\Pi_{i=1}^n f\left(X_i\right) \Pi\left[df\right]}{\int \Pi_{i=1}^n f\left(X_i\right) \Pi\left[df\right]} \tag{1.7}$$

defines a Markov kernel over $(\mathcal{T}, \mathcal{X}^n)$ that we call the posterior distribution of f given the sample $\mathbb{X}^n$. Some care must be taken here as the posterior distribution is defined only up to a null set of the marginal density $\int \Pi_{i=1}^n f(X_i) \Pi[df]$ of $\mathbb{X}^n$ with respect to μ, and under the frequentist assumption 1.6 the sample may end up outside of its support. Thus in general when combining a Bayesian approach with a frequentist assumption, we may ask for a condition of the form

$$f_0 \cdot \mu << \int f \Pi[df] \cdot \mu$$

in order to ensure that the posterior distributions involved in the analysis are all well defined. This condition will be satisfied in all examples studied in this thesis.

The goal is now to investigate whether or not the (random) probability distribution $\Pi[\cdot|\mathbb{X}^n]$ possesses interesting properties as $n \to \infty$. More precisely it is of particular interest to understand under which conditions it will eventually converge in some sense to the Dirac measure δ_{f_0}, which is the focus of the following theorem.

Theorem 8. *Schwartz' theorem*

Define the Kullback-Leibler support of Π

$$KL(\Pi) = \left\{ f_0 : \forall \varepsilon > 0, \quad \Pi\left(\left\{ f \in \mathcal{F} : \int_{\mathcal{X}} f_0 \ln \frac{f_0}{f} d\mu < \varepsilon \right\}\right) > 0 \right\}.$$

Then if $\mathcal{V}$ is any topology on $\mathcal{F}$ such that for each open neighbourhood V of f_0 there exists a test ϕ_n satisfying

$$P_0\phi_n = \int \phi_n f_0 = o(1) \text{ and } \sup_{f \in V^c} P_f(1-\phi_n) = \sup_{f \in V^c} \int (1-\phi_n) f = o(1),$$

then for any neighbourhood U of f_0 we have

$$\Pi[U^c|\mathbb{X}^n] \xrightarrow[n\to\infty]{P_0^\infty - a.s} 0$$

If in addition $\mathcal{V}$ is first countable then the almost sure convergence holds simultaneously for any neighbourhood of f_0.

Proof. See [55], Chapter 6. It is only stated for a first countable topology $\mathcal{V}$, but the almost sure convergence $\Pi\left[U^c|\mathbb{X}^n\right] = o_{P_0^\infty}(1)$ still holds U by U otherwise. □

Thus, under the testing condition, if the prior assigns enough mass around f_0 in the Kullback-Leibler sense, every neighbourhood of f_0 will eventually receive all the mass of the posterior distribution. Importantly the testing condition always holds for the weak topology, which also happens to be first countable (see [55] Chapter 6), so that the posterior distribution is always weakly consistent at every f_0 in the Kullback-Leibler support of Π. Consistency with respect to stronger topologies (such as the ones generated by the L^1 or Hellinger distances) has also been investigated for instance in [51] and [10], where it is shown that controlling the Hellinger or L^1 metric entropy of a submodel with sufficiently large prior probability is sufficient to extend the approach of Schwartz to this context.

As interesting as consistency can be, this notion is however too weak to be related to the minimax framework. In particular, a consistency result with respect to a topology (even a strong one) is not enough in order to derive posterior contraction rates in the sense of Equation 1.2. In the seminal work [52], Ghosal, Ghosh and Van der Vaart refined the approach of Schwartz in order to replace the neighbourhoods U of Theorem 8 by the complement of a ball $B_\rho\left(f_0, \varepsilon_n\right) := \{f \in \mathcal{F} : \rho\left(f_0, f\right) < \varepsilon_n\}$, where here ρ is either the Hellinger or L^1 distance on $\mathcal{F}$ and $\varepsilon_n \to 0$. For this we first define the Kullback-Leibler neighbourhood at f_0

$$\forall \varepsilon > 0, \quad B_{KL}\left(f_0, \varepsilon\right) := \left\{f : KL(f_0, f) \leq \varepsilon^2, \quad V(f_0, f) \leq \varepsilon^2\right\},$$

where $KL(f_0, f) := \int f_0 \ln \frac{f_0}{f} \leq \varepsilon^2$ is the Kullback-Leibler divergence and $V_2(f_0, f) := \int f_0 \ln^2 \frac{f_0}{f} \leq \varepsilon^2$ its second-order version.

Theorem 9. *Rate theorem - density estimation*
In the i.i.d. sampling model $\mathbb{X}^n = \left(X_1, \ldots, X_n\right) \overset{i.i.d.}{\sim} f_0, f_0 \in \mathcal{F}$, let ρ be either the Hellinger or L^1 metric on $\mathcal{F}$ and for a subset $\mathcal{F}_n \subset \mathcal{F}$ let $\mathcal{N}\left(\varepsilon, \mathcal{F}_n, \rho\right)$ be the $\varepsilon-$packing number of $\mathcal{F}_n$ with respect to ρ, that is the maximal number of points in $\mathcal{F}_n$ with pairwise distance at least ε. Assume that for some sequence $\varepsilon_n \to 0, n\varepsilon_n^2 \to \infty$ and constant $C > 0$ we have

1) $\Pi\left[\mathcal{F}_n^c\right] \leq e^{-(C+4)n\varepsilon_n^2}$,

2) $\ln \mathcal{N}\left(\varepsilon_n, \mathcal{F}_n, \rho\right) \leq n\varepsilon_n^2$,

3) $\Pi\left(B_{KL}\left(f_0, \varepsilon_n\right)\right) \geq e^{-Cn\varepsilon_n^2}$.

Then for sufficiently large $M > 0$ we have

$$\Pi\left[\{f : \rho(f_0, f) > M\varepsilon_n\} | \mathbb{X}^n\right] \xrightarrow[n\to\infty]{P_0^\infty} 0.$$

Proof. See the original paper [52], or Chapter 8 in [55]. □

As for the strong consistency results, the assumptions of Theorem 9 require the prior to put most of its mass on sets with suitably bounded L^1 or Hellinger entropy, and to assign enough mass around f_0 in the Kullback-Leibler sense.

Actually, Theorem 9 can be generalised in the following way to handle more general semi metrics and statistical models, using essentially the same proof. Let $\left(P_{f^{(n)}}\right)_{f^{(n)}\in\mathcal{F}^{(n)}}$ be a dominated statistical model, with associated densities $f^{(n)} \in \mathcal{F}^{(n)}$. Note that $f^{(n)}$ is not necessarily a tensor product of a density f. For instance in the case of the Gaussian nonparametric regression model it is a Gaussian likelihood. As in the *i.i.d.* setting, we define the Kullback-Leibler neighbourhoods at $f_0^{(n)}$

$$B_{KL}\left(f_0^{(n)}, \varepsilon\right) := \left\{f^{(n)} \in \mathcal{F}^{(n)} : \int f_0^{(n)} \ln \frac{f_0^{(n)}}{f^{(n)}} < n\varepsilon^2, \quad \int f_0^{(n)} \ln^2 \frac{f_0^{(n)}}{f^{(n)}} < n\varepsilon^2\right\}.$$

Then we have the following fundamental result :

Theorem 10. *Rate theorem - generic version*
For observations $\mathbb{X}^n \sim P_{f_0}^{(n)}, \quad f_0^{(n)} \in \mathcal{F}^{(n)}$, consider ρ_n a semi metric on $\mathcal{F}^{(n)}$ satisfying the following : for any $\varepsilon > 0, \quad f_1^{(n)} \in \mathcal{F}^{(n)}$ such that $\rho_n(f_1^{(n)}, f_0^{(n)}) > \varepsilon$ and $n \geq 1$ there exists

tests ϕ_n (i.e. measurable function of $\mathbb{X}^n$ with values in $[0,1]$) such that, for some $A, K, \xi > 0$

$$\mathbb{E}_{\mathbb{X}^n \sim f_0^{(n)}} \phi_n \leq A e^{-Kn\varepsilon^2}, \qquad \sup_{\rho_n(f^{(n)}, f_1^{(n)}) < \xi\varepsilon} \mathbb{E}_{\mathbb{X}^n \sim f^{(n)}} (1 - \phi_n) \leq A e^{-Kn\varepsilon^2}.$$

Assume further that there exists a measurable subset $\mathcal{F}_n \subset \mathcal{F}^{(n)}$, positives sequences $\bar{\varepsilon}_n, \varepsilon_n, n\bar{\varepsilon}_n^2 \to \infty, n\varepsilon_n^2 \to \infty$ and a constant $C > 0$ such that, for a prior probability distribution Π on $\mathcal{F}^{(n)}$

1) $\Pi\left[\mathcal{F}_n^c\right] \leq e^{-(C+4)n\varepsilon_n^2}$,

2) $\ln \mathcal{N}\left(\bar{\varepsilon}_n, \mathcal{F}_n, \rho_n\right) \leq n\bar{\varepsilon}_n^2$,

3) $\Pi\left[B_{KL}\left(f_0^{(n)}, \varepsilon_n\right)\right] \geq e^{-Cn\varepsilon_n^2}$

Then with $\varepsilon_n^* := \varepsilon_n \vee \bar{\varepsilon}_n$ there exists $M > 0$ such that

$$\Pi\left[\rho_n(f_0^{(n)}, f^{(n)}) > M\varepsilon_n^* | \mathbb{X}^n\right] \xrightarrow[n\to\infty]{P_0^\infty} 0.$$

Proof. See the original paper [53], the review [94] or [55], Chapter 8. The statements in [53] are a bit more complicated but our slightly modified formulation can be shown to hold using the same proof. □

remark 13. *Theorems 9 & 10 still hold if Π depends on n but still satisfies the same set of assumptions with fixed constants C, A, K, ξ.*

remark 14. *The testing condition in Theorem 10 can be seen to hold in various statistical models, see the examples in [55], Chapter 9. The semi metrics ρ_n involved in this testing condition are often referred to as statistical or testing distances and cannot be arbitrary, see for instance the discussion in [94, 65].*

Theorem 10 can be refined in various ways (see [55] Chapter 8), but at its core remains the standard approach in order to generically derive posterior contraction rates. This proof strategy can be referred to as the **prior mass and testing approach**, given the type of conditions required in Theorems 9 & 10. In a lot of cases, the posterior contractions rates obtained by the application of theorems 9 & 10 (or their variants) match (near) minimax optimal rates of estimation for appropriate choices of priors. We present in Section **??** a concrete application of Theorem 10 to the white noise model with a non-conjugate prior.

An interesting particular case where the application of Theorem 10 can be elegantly simplified is the one of **Gaussian process priors**. Following Van der Vaart & Van zanten [113, 111], a random variable $f \sim \Pi$ taking values in a separable Banach space $(\mathbb{B}, \|\cdot\|)$ is a (zero mean) **Gaussian random element** if the random variable $b^*(f)$ has mean zero and is normally distributed in $\mathbb{R}$ for any element $b^* \in \mathbb{B}^*$ (the topological dual space of $\mathbb{B}$, i.e. the set of continuous linear functional over $\mathbb{B}$). The **RKHS** (reproducing kernel Hilbert space) $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ of f is then defined to be the completion of the range of the map $S : \mathbb{B}^* \to \mathbb{B}$ defined by $Sb^* = \Pi[b^*(f)f]$ (where the last expectation with respect to the measure Π is understood as the *Pettis integral* of $b^*(f)f$, i.e. the only element $Sb^* = \omega \in \mathbb{B}$ satisfying $b_1^*(\omega) = \Pi[b^*(f)b_1^*(f)]$ for any $b_1^* \in \mathbb{B}^*$, see Section 2.2 in [113]) for the inner product

$$\langle Sb_1^* | Sb_2^* \rangle_{\mathbb{H}} := \Pi[b_1^*(f)b_2^*(f)].$$

remark 15. *In Remark 11, we have defined RKHSs as Hilbert spaces of functions having the reproducing property, which does not appear to be directly related to the notion of RKHS of Gaussian random elements. However, when the Banach space* $(\mathbb{B}, \|\cdot\|)$ *is the space of continuous functions over a compact set* $\mathcal{M}$ *equipped with the supremum norm (for instance, the space of continuous functions over a compact manifold* $\mathcal{M}$*), then the RKHS of the Gaussian random element can be identified with its stochastic process RKHS, defined as the completion of the space of linear combinations of the functions* $K(x, \cdot) := \mathrm{Cov}(f(x), f(\cdot)), \quad x \in \mathcal{M}$ *for the inner product*

$$\langle K(x, \cdot) | K(y, \cdot) \rangle = K(x, y),$$

and in this case $S\delta_x = K(x, \cdot)$*, where* δ_x *is the evaluation map* $\delta_x : b \in \mathbb{B} \mapsto b(x)$*, see [113] for a proof and a more precise statement. Moreover, for a given RKHS* $\mathbb{H}$ *of functions over* *a compact set $\mathcal{M}$ in the sense of remark 11 with associated kernel K, Theorem 12.1.3 in Dudley [41] implies the existence of a zero mean Gaussian process f over $\mathcal{M}$ (in the usual sense, i.e. a stochastic process with Gaussian finite dimensional marginals) having K as its covariance kernel. By the Moore-Aronszajn theorem [6], $\mathbb{H}$ must coincide with the stochastic process RKHS of f. Therefore, if f has values in the Banach space of continuous functions over $\mathcal{M}$, all 3 notions of RKHSs must coincide in this case.*

Although potentially less familiar to the machine learning or Gaussian process audience, it turns out that working with the notion of Gaussian random element is particularly convenient when investigating posterior contraction rates as we will see in the next theorem, provided in [111] or [55], Theorem 11.20.

Theorem 11. *Gaussian rate theorem*

Let $f \sim \Pi$ be a Gaussian random element in a separable Banach space $(\mathbb{B}, \|\cdot\|)$ with RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$. If $f_0 \in \mathbb{B}$ and $\varepsilon_n \to 0, n\varepsilon_n^2 \to \infty$ are such that

$$\phi_{f_0}(\varepsilon_n) := -\ln \Pi\left[\|f\| < \varepsilon_n\right] + \frac{1}{2} \inf_{h \in \mathbb{H}: \|f_0 - h\| < \varepsilon_n} \|h\|_{\mathbb{H}}^2 \leq n\varepsilon_n^2 \tag{1.8}$$

then for any $C > 1$ such that $Cn\varepsilon_n^2 > \ln 2$ there exists a measurable subset $B_n \subset \mathbb{B}$ such that

1) $\ln \mathcal{N}(3\varepsilon_n, B_n, \|\cdot\|) \leq 6Cn\varepsilon_n^2$,

2) $\Pi\left[f \notin B_n\right] \leq e^{-Cn\varepsilon_n^2}$,

3) $\Pi\left[\|f - f_0\| < 2\varepsilon_n\right] \geq e^{-n\varepsilon_n^2}$.

remark 16. *Using Borell's inequality (Proposition 11.17 in [55]), the set B_n can be constructed as $B_n = \varepsilon_n \mathbb{B}_1 + M_n \mathbb{H}_1$ where $\mathbb{B}_1 := \{f \in \mathbb{B} : \|f\| \leq 1\}$ and $\mathbb{H}_1 := \{h \in \mathbb{H} : \|h\|_{\mathbb{H}} \leq 1\}$ are the unit balls of $\mathbb{B}$ and $\mathbb{H}$ and $M_n = -2\Phi^{-1}\left(e^{-Cn\varepsilon_n^2}\right)$ where Φ is the standard Gaussian cumulative distribution function, see the proof of Theorem 11.20 in [55]. It is known (see [55] Lemma K.6) that $\sqrt{Cn\varepsilon_n^2} \leq M_n \leq \sqrt{10Cn\varepsilon_n^2}$.*

remark 17. *The function ϕ_{f_0} in Theorem 11 is usually referred to as the **concentration function** (at f_0) and can be actually be shown (see Proposition 11.19 in [55]) to satisfy for* *any* $\varepsilon > 0$

$$\phi_{f_0}(\varepsilon) \leq -\ln \Pi\left[\|f - f_0\| < \varepsilon\right] \leq \phi_{f_0}(\varepsilon/2),$$

so that the condition in Theorem 11 is actually precisely a lower bound on the amount of prior mass around f_0*, measured in terms of the norm* $\|\cdot\|$ *of* $\mathbb{B}$.

As presented, Theorem 11 is not directly useful in order to apply Theorem 10. However, when the norm $\|\cdot\|$ upper bounds a semi metric ρ_n satisfying the testing condition and the Kullback-Leibler divergences can be appropriately controlled by $\|\cdot\|$ then Theorem 11 directly implies a posterior contraction rate theorem in terms of the semi metric ρ_n, as we will see in the second part of the next section 1.3.2 and in Chapters 2 & 3.

1.3.2 Examples

White noise model

Here we present an application of Theorem 10 to the white noise model under a potentially non-conjugate prior. In this model, we assume observations $\mathbb{Y}^n := (y_j)_{j\geq 1}$ generated as

$$y_j = f_{0j} + \frac{1}{\sqrt{n}}\varepsilon_j, \quad \varepsilon_j \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \quad j \geq 1$$

for some $n \geq 1$, $f_0 \in \ell^2 := \ell^2(\mathbb{N}^*)$, which can be viewed equivalently as the coefficients of a solution Y to the stochastic differential equation

$$dY_t = f_0(t)dt + \frac{1}{\sqrt{n}}dB_t, \quad t \in [0,1]$$

in an arbitrary orthonormal basis of $L^2 := L^2([0,1])$, where here $f_0 \in L^2$ $(B_t)_{t\in[0,1]}$ is a standard Brownian motion. We refer the reader to Section 1.2 of [56] for a more precise introduction to the white noise model and its connections with nonparametric regression.

We assume a smoothness condition on the vector $f_0 \in \ell^2$: for some $M, \beta > 0$ and for every $J \geq 1$

$$\sum_{j>J} f_{0j}^2 \leq M^2 J^{-2\beta}. \tag{1.9}$$

Assumption 1.9 holds for instance if $f_0(t)$ is periodic and is an element of the Besov space $B^{\beta}_{2\infty}$ over the unit circle S^1 (in particular this is also the case if $f_0(t)$ is in the Sobolev space H^{β} of order β over S^1 as $H^{\beta} = B^{\beta}_{22} \subset B^{\beta}_{2\infty}$, see the discussion in the Appendix of Chapter 2). We refer the reader to [26, 37] and [35] for more details about Besov and Sobolev spaces on manifolds.

We choose an index $J \in \{1, \dots, n\}$ (that may depend on n) and a prior Π over f of the form

$$f \sim \Pi \iff \begin{cases} f_j \overset{i.i.d.}{\sim} K \text{ if } j \leq J, \\ \\ f_j = 0, \text{ if } j > J, \end{cases}$$

where K is a positive and continuous density on $\mathbb{R}$ satisfying

$$\forall a \geq a_0, \quad \int_{|x|>a} K(x)dx \leq c_1 e^{-c_2 a^{c_3}},$$

for some $a_0, c_1, c_2, c_3 > 0$. For instance a standard Gaussian density is allowed (and in that case it can be checked that the resulting posterior distribution is explicit by conjugacy), but more general densities are also possible (and in this case conjugacy no longer holds).

We are going to apply Theorem 10 in order to deduce an L^2-posterior contraction at f_0 with rate

$$\varepsilon_n = \sqrt{\frac{J \ln n}{n}} + J^{-\beta}.$$

The distributions $(Q_f)_{f\in\ell^2}$ of the observations

$$y_j = f_f + \frac{1}{\sqrt{n}}\varepsilon_j, \quad j \geq 1$$

can be seen to be tight Gaussian Borel probability measures in an appropriate separable Hilbert space and are all absolutely continuous with respect to Q_0, with explicit likelihood ratios (see Theorem 6.1.6 in [56]). We thus apply Theorem 10 by taking $P_f^{(n)} = \frac{dQ_f}{dQ_0}$: as explained in Section 8.3.4 of [55], the L^2-metric $(f, g) \mapsto \|f - g\|_2$ satisfies the testing assumption of Theorem 10 in this context, and the Kullback-Leibler neighbourhood $B_{KL}(Q_{f_0}, \varepsilon_n)$

contains an L^2-ball at f_0 with radius a constant multiple of ε_n. Using

$$\begin{aligned}\|f - f_0\|_2^2 &= \sum_{j=1}^{J} (f_j - f_{0j})^2 + \sum_{j>J} f_{0j}^2 \\ &\leq M^2 J^{-2\beta} + J \max_{1\leq j\leq J} |f_j - f_{0j}|^2 \\ &\leq \varepsilon_n + J \max_{1\leq j\leq J} |f_j - f_{0j}|^2 ,\end{aligned}$$

if $\max_{1\leq j\leq J} |f_j - f_{0j}| \leq \frac{\varepsilon_n}{\sqrt{J}}$ we get $f \in B_{KL}(Q_{f_0}, c\varepsilon_n)$ for some constant $c > 0$. The prior probability of the latter event is lower bounded by

$$\begin{aligned}\Pi\left[\forall 1 \leq j \leq J, \quad |f_j - f_{0j}| \leq \varepsilon_n/\sqrt{J}\right] &= \Pi_{j=1}^{J} \int_{I_j} K(x)dx \\ &\geq \Pi_{j=1}^{J} \inf_{I_j} K,\end{aligned}$$

where here $I_j = \left[f_{0j} - \varepsilon_n/\sqrt{J}, f_{0j} + \varepsilon_n/\sqrt{J}\right]$. Since $\sup_{j\geq 1} |f_0 j| + \varepsilon_n/\sqrt{J}$ is bounded by a constant independent of n (because $f_0 \in \ell^2$) and K is lower bounded on every compact interval, we get for some $c, c' > 0$

$$\Pi\left[B_{KL}(f_0, c\varepsilon)\right] \geq \Pi\left[\forall 1 \leq j \leq J, \quad |f_j - f_{0j}| \leq \varepsilon_n/\sqrt{J}\right] \geq \exp\left(-c' J \ln n\right).$$

Here we have also used the fact that $J \leq n$, so that $\frac{\varepsilon_n}{\sqrt{n}}$ is always greater than some inverse power of n. Moreover, by the union bound and our assumption on K, for any $a \geq a_0$

$$\Pi\left[\exists 1 \leq j \leq J, \quad |f_j| > a\right] \leq J \int_{|x|>a} K(x)dx \leq c_1 J e^{-c_2 a^{c_3}},$$

so that for any $C > 0$ there exists C' such that

$$\Pi\left[f \notin \mathcal{F}_{C'}\right] \leq e^{-Cn\varepsilon_n^2},$$

where $\mathcal{F}_{C'}$ is set of $f \in \ell^2$ such that

$$\begin{cases} |f_j| \leq C' \left(n\varepsilon_n^2\right)^{1/c_3} & \text{for every } 1 \leq j \leq J \\ f_j = 0 & \text{otherwise} \end{cases}.$$

Being an $\ell^\infty-$ball in a $J-$dimensional space of radius at most polynomial in n, standard packing arguments show that for some $C'' > 0$

$$\ln \mathcal{N}\left(\varepsilon_n, \mathcal{F}_{C'}, \|\cdot\|_2\right) \leq C'' J \ln n.$$

Since $J \ln n \leq n\varepsilon_n^2$, we deduce by an application of Theorem 10 the existence of $C > 0$ such that

$$Q_{f_0} \Pi\left[\|f - f_0\|_2 > C\varepsilon_n | \mathbb{Y}^n\right] \xrightarrow[n \to \infty]{} 0.$$

Notice that in that case, choosing $J = \lfloor (n/\ln n)^{\frac{1}{2\beta+1}} \rfloor$ to balance both terms in ε_n leads to a posterior contraction rate of order $(n/\ln n)^{-\frac{\beta}{2\beta+1}}$, which is the typical minimax rate of estimation of a $\beta-$Sobolev function on $[0,1]$ in the nonparametric regression model with random design, up to a logarithmic factor.

This proof of posterior contraction in the white noise model is similar at a really high level to the one presented in Chapter 3 in the nonparametric regression model over random geometric graphs, although the smoothness assumption there is much more subtle and requires to prove involved approximations arguments. In the following subsections we will present other classes of priors that we will use in Chapters 4 & 2 and for which the theory of posterior contraction has been well studied.

Nonparametric mixtures

A popular class of priors for the density estimation problem based on an *i.i.d.* sample is the one of **mixture models**, defined as the probability distribution Π of a density f_P of the form

$$f_P = \int_\Theta K_\theta P(d\theta),$$

where Π_1 is a typically discrete probability distribution on Θ such as a Dirichlet Process (see [44] and Chapter 4 in [55]), K_θ is a probability distribution on a set $\mathcal{X}$ for each $\theta \in \Theta$ and therefore the resulting function $f_P \sim \Pi$ is a random probability measure (assuming suitable measurability conditions). A fundamental example is the case where $\mathcal{X} = \mathbb{R}^d, \Theta = \mathbb{R}^d \times \mathcal{S}_d^{++}(\mathbb{R})$ (the open cone of positive definite symmetric matrices of size $d \times d$) and $K_{\mu,\Sigma} = \mathcal{N}(\mu, \Sigma)$, yielding the class of Gaussian mixture models, whose consistency and contraction rate properties have been investigated in [51, 54, 70, 100, 23].

In order to derive contraction rates, [54, 70, 100, 23] used a refined version of the rate Theorem 9. In this context, verifying the assumptions of the rate theorem (in particular showing an appropriate lower bound on the prior probability of the Kullback-Leibler neighbourhoods) requires to show preliminary results on the quality of approximation of the true density of the observations by mixtures of Gaussians. In Chapter 4, we will apply the same proof strategy to analyse the posterior contraction properties of a new class of Gaussian mixture models under geometric assumptions. From the practical side, using the likelihood arising from the *i.i.d.* sampling model, one can then compute an approximate posterior sample using MCMC, see for instance [77] in the case of Dirichlet Process mixture models.

Gaussian processes

As introduced in the last section, another popular class of priors is the one of Gaussian processes. For instance, in the nonparametric regression model

$$Y_i = f(X_i) + \varepsilon_i, \quad 1 \leq i \leq n \tag{1.10}$$

where ε_i are *i.i.d.* $\mathcal{N}\left(0, \sigma^2\right)$ for some $\sigma > 0$, X_i covariates that are elements of a compact set $\mathcal{X}$ (that can themselves be random) and $f : \mathcal{X} \to \mathbb{R}$, when putting a prior on f and under the frequentist assumption of existence of a "true" $f_0 : \mathcal{M} \to \mathbb{R}$ generating the observations $(X_i, Y_i)_{i=1}^n$ via the model 1.10, it is possible to apply Theorem 10 with ρ_n being the **empirical**

L^2 distance

$$\rho_n(f,g) := \left(\frac{1}{n}\sum_{i=1}^{n}\left(f(X_i) - g(X_i)\right)^2\right)^{1/2}$$

by considering as prior over f a Gaussian random element taking values in $\mathcal{C}(\mathcal{X})$, the space of continuous functions over $\mathcal{X}$ (equipped with the supremum norm $\|\cdot\|_\infty$). Indeed, the semi metric ρ_n satisfies the testing condition (by constructing a likelihood ratio test, see e.g. Lemma 8.27 in [55]) and is always trivially upper bounded by $\|\cdot\|_\infty$. Moreover, the Kullback-Leibler neighbourhoods at f_0 of radius ε_n can be seen in this context to contain a $\|\cdot\|_\infty -$ball at f_0 of radius a constant multiple of ε_n (see Lemma 2.7 in [55]), and therefore an application of Theorem 11 with $(\mathbb{B}, \|\cdot\|) = (\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty)$ directly yields ρ_n-posterior contraction at rate (a multiple of) ε_n, whenever ε_n satisfies the rate equation 1.8 (see [55] Chapter 11 or [26] for detailed and concrete examples).

Furthermore, when the covariates X_i are themselves *i.i.d.* sampled according to some probability distribution μ_0 on $\mathcal{X}$, under additional technical assumptions and with a bit of extra work it is possible to deduce a posterior contraction rate with respect to the "global" $L^2(\mu_0)$ loss

$$\|f-g\|_{L^2(\mu_0)} = \left(\int_{\mathcal{X}}\left(f(x) - g(x)\right)^2 \mu_0(dx)\right)^{1/2},$$

from that with respect to ρ_n, see for instance [112] in the case $\mathcal{X} = [0,1]^d$ where a variation of this proof technique is used to get a posterior contraction rate in expectation in the sense of 1.1. We will apply this set of proof techniques in Chapter 2 when analysing the posterior contraction properties of some classes of Gaussian processes in geometrical settings. For the nonparametric regression problem with Gaussian errors and Gaussian priors the posterior is also a Gaussian process with explicit mean and covariance by conjugacy, and therefore posterior sampling is straightforward up to the inversion of a high dimensional matrix which can be challenging, see for instance [79, 102] and the discussions therein. It should also be noted that the nonparametric regression problem is only a first example, and it is possible to use Gaussian process priors in other contexts such as density estimation, classification or inverse problems (even though conjugacy may not hold and posterior sampling requires generic MCMC techniques), see e.g. [55] Chapter 11 and [78].

Random basis expansions

Still in the regression model 1.10, other Gaussian priors can be of interest. For instance when $(u_j)_{j\geq 1}$ is a (typically orthonormal, according to some inner product) family of functions over $\mathcal{X}$, then a possible prior on f_0 can be constructed as the probability distribution of the random function

$$f = \sum_{j\geq 1} \sigma_j Z_j u_j \tag{1.11}$$

(assuming almost sure convergence), where $\sigma_j \geq 0$ are scaling coefficients and Z_j are *i.i.d.* standard Normal random variables. We will use this type of priors in Chapter 2, with the family $(u_j)_{j\geq 1}$ being an eigenbasis of the Laplace-Beltrami operator on a compact Riemannian manifold, see Section 1.1. Such a construction can be generalised to give other non Gaussian priors by considering $Z_j \overset{i.i.d.}{\sim} \Psi$ for some other base distribution Ψ, as done in Section 1.3.2. Such priors as commonly referred to as **random basis expansions**. Even though the resulting f is no longer a Gaussian random element and therefore Theorem 11 does not apply as above, it is still possible to use Theorem 10 directly (with ρ_n still being the empirical L^2 distance). This is the approach we take in Chapter 3 where we study the case of $\mathcal{X}$ being a random geometric graph.

Theorems 9 10 & 11 and their variants form the core part of the proof techniques used to prove posterior contraction rates in generic, non-conjugate settings. As the focus of this thesis is on posterior contraction rates (i.e. rates of estimation), we have not tried to present the theory of Bayesian uncertainty quantification, the problem to determine whether or not Bayesian credible sets (subsets of the model parameters with high posterior probability) also have high frequentist coverage (i.e. are also confidence intervals). The investigation of this problem typically relies on a preliminary proof of the (either nonparametric or semiparametric) Bernstein-von Mises theorem, see [28, 27, 29, 88, 55, 24].

We will now introduce in the next section the manifold hypothesis, a geometrical framework that aims to provide insights on the curse of dimensionality in high dimensional spaces, and the different associated methods to temper its effects. We will also review the recent relevant literature to the research questions tackled in this thesis, as well as our contributions using nonparametric Bayesian techniques.

1.4 The manifold hypothesis

As we argued at the start of Chapter 1, minimax lower bounds for the estimation of functional parameters imply an unavoidable exponentially bad dependence of the risk on the inputs domain dimension. Since modern large-scale statistical problems often include large dimensionality, understanding to which extent the data presents intrinsic lower dimensional structures can be of interest. We will go over the existing literature on this question in Section 1.4.1 before presenting our contributions in Section 1.4.2.

1.4.1 Literature review

There are good empirical evidence confirming that such low dimensional structures exist in real world applications, with examples found in texts, sounds and images data sets, see for instance [13, 69, 68, 116] & [21]. Intuitively, this can be understood as stemming from physical constraints, see for instance Section 5.11.3 in [60] where it is argued that most points in the space of images are actually noise, and that only a subset of this space of much lower dimension contains the relevant images for practical applications.

Leveraging the tools we introduced in Section 1.1, a precise mathematical formulation of such a phenomenon is then clear : in high dimensional statistical problems (where the variables are elements of $\mathbb{R}^D, \quad D \geq 1$ large), the data sets tend to be supported on low dimensional submanifolds (of $\mathbb{R}^D$ in this context). This is what is often referred to as the manifold hypothesis. A more flexible and realistic formulation requires the data to be supported on a small neighbourhood of a low dimensional manifold, a model which allows possibly small observation noise in high dimension, see for instance [30] where the case of additive Gaussian noise is considered. Such noise models (additive or not) are still referred to as part of the manifold hypothesis framework, even though they are technically a small relaxation of it. Other possible extensions allow the data to be supported on potentially irregular subsets of $\mathbb{R}^D$ (i.e. not necessarily submanifolds) with low intrinsic dimension (where the notion of dimension is now understood in the Minkowski sense), see for instance [115, 25] or union of manifolds [21, 1].

Numerous different questions can then be of interest when working under the manifold hypothesis. To start with, one can wonder to which extent it is possible to estimate the probability distribution of a random *i.i.d.* sample $X_1, \ldots, X_n$ in $\mathbb{R}^D$ if its probability distribution is supported on a (typically unknown) low dimensional subset $\mathcal{M} \subset \mathbb{R}^D$ (or a small neighbourhood of it). The deep learning literature has a lot to offer from this perspective, with numerous recent methodological and theoretical advances with generative adversarial networks, energy-based models, normalising flows, variational auto-encoders or diffusion models [4, 5, 66, 74, 73, 68, 75, 36, 101, 105, 8, 86]. In [105] the resulting estimator is shown to achieve minimax rates of estimation of the underlying probability distribution under the $1-$Wasserstein distance. In the nonparametric statistics literature, the behaviour of kernel density estimators has also been studied theoretically under the manifold hypothesis, for the Wasserstein [39] or pointwise losses [16, 84], with minimax rates of convergence. In particular, the rates obtained there are only impacted by the dimensionality of the submanifold $\mathcal{M}$ and not the one of $\mathbb{R}^D$. Estimators based on mixtures models have also been investigated empirically in [114] and in [76] in the Bayesian context, where a new class of kernels (the "Fisher-Gaussian" kernels) more flexible than the Gaussian ones is presented, but where however no posterior contraction rates are derived. Compared with the usual Gaussian kernels, the Fisher-Gaussian kernels are shown empirically in [76] to be able to estimate more accurately probability distributions supported on small neighbourhoods of submanifolds thanks to an additional curvature parameter.

Instead of estimating a probability distribution, recovering the geometrical structure of the submanifold in question can also be of interest. In particular, the estimation of the tangent spaces, the extrinsic curvature or the submanifold itself (as a subset of $\mathbb{R}^D$, under the Hausdorff loss) is for instance studied in a minimax framework in [50, 2, 38, 87, 8].

Finding a meaningful low dimensional representation of the data can also be useful, a problem commonly referred to as (non linear) dimension reduction, which can be understood as estimating a set of coordinates charts. To this end various methods have been suggested such as Principal Component Analysis [85], its kernelised version [98], Locally Linear Embeddings [95] or graph based methods (which implicitely approximate the latent submanifold by a random graph, a technique we have already discussed in Section 1.2) such as Isomap [107], Laplacian Eigenmaps [14, 15] or diffusion maps [34].

Finally, the problem of studying the nonparametric regression problem 1.10 in the case where the covariates X_i are supported on an unknown low dimensional subset has also received much attention, with estimation procedures based for instance on Bayesian neural networks [25], the restriction of ambient Gaussian process priors (i.e. defined globally on $\mathbb{R}^D$) [117, 106] or graph based methods [96, 61, 45, 81, 43] (based either on least square estimators or Bayesian series priors of the form 1.11 in the graph spectral domain). While optimal rates of convergence are obtained in quite general settings in [25, 117, 106], for graph based methods the situation is not as clear, with the only known results either requiring the smoothness of the true regression function to belong to $\{1, 2, 3\}$ [61] or to have access to a large amount of additional unlabeled covariates $\{x_{n+1}, \dots, x_N\}, \quad N > n$ [96].

1.4.2 Contributions

We now present the three main chapters of this thesis, the first two focusing on nonparametric regression with covariates located on an unknown submanifold, and the last one on density estimation in the vicinity of an unknown submanifold.

Chapter 2 : Posterior Contraction Rates for Matérn Gaussian Processes on Riemannian Manifolds [91]

This is a joint work with Viacheslav Borovitskiy, Alexander Terenin and Judith Rousseau, and has been published in the *Advances in Neural Information Processing Systems 36* (NeurIPS 2023, spotlight poster). In this paper we tackle the nonparametric regression problem 1.10 when the covariates are located on a possibly unknown submanifold $\mathcal{M}$ of $\mathbb{R}^D, \quad D \geq 1$ using a Bayesian nonparametric approach based on Gaussian process priors (as presented in Section 1.3.2). We consider two main classes of Gaussian processes : the intrinsic ones, which are those defined by random series expansion as in 1.11 and where the basis functions are the eigenfunctions of the Laplace-Beltrami operator (see Section 1.2), and the extrinsic ones, those defined (implicitely) by the restriction of a ambient Gaussian process (i.e. : defined everywhere on $\mathbb{R}^D$) to $\mathcal{M}$. We focus on the class of Mátern Gaussian processes and we adapt the proof strategy of [112] to the manifold case. Surprisingly, our results indicate that both priors (intrinsic and extrinsic) achieve the optimal rate of convergence. Our proof crucially relies on an application of a trace and extension theorem between manifold and Euclidean Sobolev spaces, showing in our context that both processes have norm equivalent RKHS's. Compared with [117] (which focuses on the restriction of ambient square exponential processes), our proof is not limited to smoothness less than 2. It should be noted that [117] has been extended later to arbitrary smoothness and to the adaptive case in [106].

Chapter 3 : Nonparametric regression on random geometric graphs sampled from submanifolds [92]

This is a joint work with Judith Rousseau and has been submitted to the *Journal of Machine Learning Research*. It is a logical extension of the work previously done in [91], where we still consider the same nonparametric regression problem 1.10 (with covariates X_i located on an unknown $d-$dimensional submanifold $\mathcal{M}$ of $\mathbb{R}^D, \quad D \geq 1$) but we approximate the intrinsic Gaussian process defined by random basis expansion in the Laplace-Beltrami eigenbasis by its analog on a random graph approximating $\mathcal{M}$ (see Section 1.2). This technique therefore allows to use an approximation of the intrinsic processes even when the manifold in question is unknown, as the prior process is this time defined by a random basis expansion in the eigenbasis of the graph Laplacian. Actually, as we discussed in Section 1.3.2, there is no reason to limit ourselves to Gaussian processes, and we present a very general result valid for a large class of priors on the coefficients in the basis expansion with exponentially small tails. Assuming $\beta-$Hölder smoothness of the underlying regression function, we present priors achieving (near) minimax optimal posterior contraction rates in the adaptive or non-adaptive case for any $\beta > 0$ ($\beta > d/2$ when the global L^2 loss over the graph is considered) without requiring any additional unlabeled covariates, thus significantly extending the results in [61, 96]. Our proof is based on a novel approximation argument in the graph Laplacian spectral domain. As a byproduct of our approximation argument we prove that the PCR-LE (least square) estimator proposed in [61] also achieves the (near) minimax rate under the same conditions.

Chapter 4 : Estimating a density near an unknown manifold: a Bayesian non-parametric approach [16]

This is a joint work with Clément Berenfeld and Judith Rousseau, and has been published in *The Annals of Statistics.*

This was the first problem investigated during the DPhil. In this paper, we investigate the problem of estimating a density f_0 supported on a small tubular neighbourhood (see Section 1.1 of Chapter 1) of an unknown $d-$dimensional submanifold $\mathcal{M}$ of $\mathbb{R}^D$, based on an *i.i.d.* sample from f_0. To this end we design a new class of Bayesian nonparametric priors for the density based on location-scale mixtures of Gaussian kernels, where the prior for the local covariance matrices is carefully designed by putting separate priors on their directions and eigenvalues. After defining a new type of Hölder regularity assumptions on the density f_0 where the smoothness is measured along a moving coordinate frame in the tubular neighbourhood (instead of the coordinates axis of $\mathbb{R}^D$), we show (near) minimax optimal posterior contraction rates for the resulting procedure with respect to the Hellinger distance provided that the width of the tubular neighbourhood is not too small. The density f_0 is shown to satisfy this Hölder regularity assumption in the case of isotropic and orthogonal additive noise models, and the resulting posterior contraction rates are in this setting essentially only impacted by the dimensionality of $\mathcal{M}$ and not the one of $\mathbb{R}^D$, provided that the density of the noise is infinitely smooth. Moreover, we present a Gibbs sampling and a MAP (maximum a posteriori) algorithm showing that the resulting Bayesian procedures significantly outperform their more standard counterparts based on inverse Wishart prior for the covariances.

Bibliography

[1] Eddie Aamari and Clément Berenfeld. A theory of stratification learning, May 2024.

[2] Eddie Aamari and Clément Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. *The Annals of Statistics*, 47(1):177–204, February 2019.

[3] Sergios Agapiou, Masoumeh Dashti, and Tapio Helin. Rates of contraction of posterior distributions based on p-exponential priors. *Bernoulli*, 27(3):1616–1642, May 2021.

[4] Martin Arjovsky and Leon Bottou. Towards Principled Methods for Training Generative Adversarial Networks. In *International Conference on Learning Representations*, 2017.

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 214–223. PMLR, July 2017.

[6] N. Aronszajn. Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[7] Ivan Avramidi. *Heat Kernel Method and Its Applications*. Springer International Publishing, Cham, 2015.

[8] Iskander Azangulov, George Deligiannidis, and Judith Rousseau. Convergence of Diffusion Models Under the Manifold Hypothesis in High-Dimensions, September 2024.

[9] Martin T. Barlow. *Random Walks and Heat Kernels on Graphs*. London Mathematical Society Lecture Note Series. Cambridge University Press, 2017.

[10] Andrew Barron, Mark J. Schervish, and Larry Wasserman. The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27(2):536–561, April 1999.

[11] Andrew R. Barron. The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Technical Report, University of Illinois at Urbana-Champaign, 1988.

[12] Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 53, 1763.

[13] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

[14] Mikhail Belkin and Partha Niyogi. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

[15] Mikhail Belkin and Partha Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, 15(6):1373–1396, June 2003.

[16] Clément Berenfeld and Marc Hoffmann. Density estimation on an unknown submanifold. *Electronic Journal of Statistics*, 15(1):2179–2223, January 2021.

[17] Clément Berenfeld, Paul Rosa, and Judith Rousseau. Estimating a density near an unknown manifold: A Bayesian nonparametric approach. *The Annals of Statistics*, 52(5):2081–2111, October 2024.

[18] Mira Bernstein, Vin de Silva, John C Langford, and Joshua B Tenenbaum. Graph approximations to geodesics on embedded manifolds.

[19] David Borthwick. *Spectral Theory: Basic Concepts and Applications | SpringerLink*, volume 284 of *Graduate Texts in Mathematics*. 2020.

[20] Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, New York, NY, 2011.

[21] Bradley C.A. Brown, Anthony L. Caterini, Brendan Leigh Ross, Jesse C. Cresswell, and Gabriel Loaiza-Ganem. Verifying the union of manifolds hypothesis for image data. In *International Conference on Learning Representations*, 2023.

[22] Jeff Calder and Nicolas Garcia Trillos. Improved spectral convergence rates for graph Laplacians on epsilon-graphs and k-NN graphs, June 2020.

[23] Antonio Canale and Pierpaolo De Blasi. Posterior asymptotics of nonparametric location-scale mixtures for multivariate density estimation. *Bernoulli*, 23(1):379–404, February 2017.

[24] Ismaël Castillo. *Bayesian Nonparametric Statistics: École d'Été de Probabilités de Saint-Flour LI - 2023*, volume 2358 of *Lecture Notes in Mathematics*. Springer Nature Switzerland, Cham, 2024.

[25] Ismaël Castillo and Paul Egels. Posterior and variational inference for deep neural networks with heavy-tailed weights, June 2024.

[26] Ismaël Castillo, Gérard Kerkyacharian, and Dominique Picard. Thomas Bayes' walk on manifolds. *Probability Theory and Related Fields*, 158(3):665–710, April 2014.

[27] Ismaël Castillo and Richard Nickl. Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *The Annals of Statistics*, 41(4):1999–2028, August 2013.

[28] Ismaël Castillo and Richard Nickl. On the Bernstein-Von Mises Phenomenon for Nonparametric Bayes Procedures. *The Annals of Statistics*, 42(5):1941–1969, 2014.

[29] Ismaël Castillo and Judith Rousseau. A Bernstein-Von Mises Theorem for Smooth Functionals in Semiparametric Models. *The Annals of Statistics*, 43(6):2353–2383, 2015.

[30] Minwoo Chae, Dongha Kim, Yongdai Kim, and Lizhen Lin. A likelihood approach to nonparametric estimation of a singular distribution using deep generative models. *J. Mach. Learn. Res.*, 24(1):77:3414–77:3455, March 2024.

[31] Isaac Chavel. *Eigenvalues in Riemannian Geometry*, volume 115 of *Pure and Applied Mathematics*. Academic Press.

[32] Andreas Christmann and Ingo Steinwart. *Support Vector Machines*. Information Science and Statistics. Springer, New York, NY, 2008.

[33] Galatia Cleanthous, Athanasios G. Georgiadis, Gerard Kerkyacharian, Pencho Petrushev, and Dominique Picard. Kernel and wavelet density estimators on manifolds and more general metric spaces. *Bernoulli*, 26(3):1832–1862, August 2020.

[34] Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, July 2006.

[35] Thierry Coulhon, Gerard Kerkyacharian, and Pencho Petrushev. Heat kernel generated frames in the setting of Dirichlet spaces. *Journal of Fourier Analysis and Applications*, 18:995–1066, 2012.

[36] Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.

[37] Ernesto De Vito, Nicole Mucke, and Lorenzo Rosasco. Reproducing kernel Hilbert spaces on manifolds: Sobolev and diffusion spaces. *Analysis and Applications*, 19(03):363–396, May 2021.

[38] Vincent Divol. Minimax adaptive estimation in manifold inference. *Electronic Journal of Statistics*, 15(2):5888–5932, January 2021.

[39] Vincent Divol. Measure estimation on manifolds: An optimal transport approach. *Probability Theory and Related Fields*, 183(1):581–647, June 2022.

[40] Pierre-Simon de (1749-1827) Auteur du texte Laplace. *Théorie Analytique Des Probabilités ; Par M. Le Comte Laplace,...* 1812.

[41] R. M. Dudley. *Real Analysis and Probability.* Cambridge University Press, 2002.

[42] David B. Dunson, Hau-Tieng Wu, and Nan Wu. Spectral Convergence of Graph Laplacian and Heat Kernel Reconstruction in \$L^\infty\$ from Random Samples. *Applied and Computational Harmonic Analysis*, 55:282–336, 2021.

[43] David B. Dunson, Hau-Tieng Wu, and Nan Wu. Graph based Gaussian processes on restricted domains. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(2):414–439, 2022.

[44] Thomas S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230, March 1973.

[45] Bernardo Fichera, Viacheslav Borovitskiy, Andreas Krause, and Aude Billard. Implicit Manifold Gaussian Process Regression. In *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023.

[46] David A. Freedman. On the Asymptotic Behavior of Bayes' Estimates in the Discrete Case. *The Annals of Mathematical Statistics*, 34(4):1386–1403, December 1963.

[47] David A. Freedman. On the Asymptotic Behavior of Bayes Estimates in the Discrete Case II. *The Annals of Mathematical Statistics*, 36(2):454–456, April 1965.

[48] Sylvestre Gallot, Dominique Hulin, and Jacques Lafontaine. *Riemannian Geometry.* Universitext. Springer, Berlin, Heidelberg, 2004.

[49] Nicolás García Trillos, Moritz Gerlach, Matthias Hein, and Dejan Slepčev. Error Estimates for Spectral Convergence of the Graph Laplacian on Random Geometric Graphs Toward the Laplace–Beltrami Operator. *Foundations of Computational Mathematics*, 20(4):827–887, August 2020.

[50] Christopher Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Minimax Manifold Estimation. *Journal of Machine Learning Research*, 13(43):1263–1291, 2012.

[51] Subhashis Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, 1999.

[52] Subhashis Ghosal, Jayanta K. Ghosh, and Aad W. van der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531, April 2000.

[53] Subhashis Ghosal and Aad van der Vaart. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223, February 2007.

[54] Subhashis Ghosal and Aad van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics*, 35(2):697–723, April 2007.

[55] Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2017.

[56] Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2015.

[57] Matteo Giordano. Besov-Laplace priors in density estimation: Optimal posterior contraction rates and adaptation. *Electronic Journal of Statistics*, 17(2):2210–2249, January 2023.

[58] Franziska Göbel and Gilles Blanchard. Volume Doubling Condition and a Local Poincar\'e Inequality on Unweighted Random Geometric Graphs, March 2020.

[59] A. Goldenshluger and O. Lepski. On adaptive minimax density estimation on $$R^d$$. *Probability Theory and Related Fields*, 159(3):479–543, August 2014.

[60] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[61] Alden Green, Sivaraman Balakrishnan, and Ryan Tibshirani. Minimax Optimal Regression over Sobolev Spaces via Laplacian Regularization on Neighborhood Graphs. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 2602–2610. PMLR, March 2021.

[62] Alexander Grigor'yan. Heat Kernels on Manifolds, Graphs and Fractals. In Carles Casacuberta, Rosa Maria Miró-Roig, Joan Verdera, and Sebastià Xambó-Descamps, editors, *European Congress of Mathematics*, pages 393–406, Basel, 2001. Birkhäuser.

[63] Alexander Grigor'yan. *Heat Kernel and Analysis on Manifolds*. AMS/IP Studies in Advanced Mathematics. AMS/International Press of Boston, 2013.

[64] Matthias Hein, Jean-Yves Audibert, and Ulrike von Luxburg. Graph Laplacians and their Convergence on Random Neighborhood Graphs. *Journal of Machine Learning Research*, 8(48):1325–1368, 2007.

[65] Marc Hoffmann, Judith Rousseau, and Johannes Schmidt-Hieber. On adaptive posterior concentration rates. *The Annals of Statistics*, 43(5):2259–2295, October 2015.

[66] Christian Horvat and Jean-Pascal Pfister. Density estimation on low-dimensional manifolds: An inflation-deflation approach. *Journal of Machine Learning Research*, 24(61):1–37, 2023.

[67] Gerard Kerkyacharian and Pencho Petrushev. Heat kernel based decomposition of spaces of distributions in the framework of Dirichlet spaces. *Transactions of the American Mathematical Society*, 367(1):121–189, 2015.

[68] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes, December 2022.

[69] W. Klein, R. Plomp, and L. C. Pols. Vowel spectra, vowel spaces, and vowel identification. *The Journal of the Acoustical Society of America*, 48(4):999–1009, October 1970.

[70] Willem Kruijer, Judith Rousseau, and Aad van der Vaart. Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4(none):1225–1257, January 2010.

[71] Lucien Le Cam. *On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes' Estimates.* PhD thesis, UC Berkeley, 1952.

[72] John M. Lee. *Introduction to Riemannian Manifolds*, volume 176 of *Graduate Texts in Mathematics*. Springer International Publishing, Cham, 2018.

[73] Brendan Leigh Ross, Gabriel Loaiza-Ganem, Anthony L. Caterini, and Jesse C. Cresswell. Neural Implicit Manifold Learning for Topology-Aware Density Estimation. *Transactions on Machine Learning Research*, 2023.

[74] Gabriel Loaiza-Ganem, Brendan Leigh Ross, Jesse C. Cresswell, and Anthony L. Caterini. Diagnosing and Fixing Manifold Overfitting in Deep Generative Models. *Transactions on Machine Learning Research*, 2022.

[75] Gabriel Loaiza-Ganem, Brendan Leigh Ross, Rasa Hosseinzadeh, Anthony L. Caterini, and Jesse C. Cresswell. Deep Generative Models through the Lens of the Manifold Hypothesis: A Survey and New Connections. *Transactions on Machine Learning Research*, 2024.

[76] Minerva Mukhopadhyay, Didong Li, and David B. Dunson. Estimating densities with non-linear support by using Fisher-Gaussian kernels. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 82(5):1249–1271, December 2020.

[77] Radford M. Neal. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

[78] Richard Nickl. *Bayesian Non-linear Statistical Inverse Problems.* Zurich Lectures in Advanced Mathematics. EMS Press, 2023.

[79] Dennis Nieman, Botond Szabo, and Harry van Zanten. Contraction rates for sparse variational approximations in Gaussian process regression. *Journal of Machine Learning Research*, 23(205):1–26, 2022.

[80] Jonathan Niles-Weed and Quentin Berthet. Minimax estimation of smooth densities in Wasserstein distance. *The Annals of Statistics*, 50(3):1519–1540, June 2022.

[81] Mu Niu, Zhenwen Dai, Pokman Cheung, and Yizhu Wang. Intrinsic Gaussian Process on Unknown Manifolds with Probabilistic Metrics. *Journal of Machine Learning Research*, 24(104):1–42, 2023.

[82] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the Homology of Submanifolds with High Confidence from Random Samples. *Discrete & Computational Geometry*, 39(1):419–441, March 2008.

[83] J. R. Norris. *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1997.

[84] Arkadas Ozakin and Alexander Gray. Submanifold density estimation. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

[85] Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space. November 1901.

[86] Peter Potaptchik, Iskander Azangulov, and George Deligiannidis. Linear Convergence of Diffusion Models Under the Manifold Hypothesis, October 2024.

[87] Nikita Puchkin and Vladimir Spokoiny. Structure-adaptive Manifold Estimation. *Journal of Machine Learning Research*, 23(40):1–62, 2022.

[88] Vincent Rivoirard and Judith Rousseau. Bernstein–von Mises theorem for linear functionals of the density. *The Annals of Statistics*, 40(3):1489–1523, June 2012.

[89] Vincent Rivoirard and Judith Rousseau. Posterior Concentration Rates for Infinite Dimensional Exponential Families. *Bayesian Analysis*, 7(2):311–334, June 2012.

[90] Joel W. Robbin and Dietmar A. Salamon. *Introduction to Differential Geometry.* Springer, Berlin, Heidelberg, 2022.

[91] Paul Rosa, Slava Borovitskiy, Alexander Terenin, and Judith Rousseau. Posterior Contraction Rates for Matérn Gaussian Processes on Riemannian Manifolds. *Advances in Neural Information Processing Systems*, 36:34087–34121, December 2023.

[92] Paul Rosa and Judith Rousseau. Nonparametric regression on random geometric graphs sampled from submanifolds, November 2024.

[93] Steven Rosenberg. *The Laplacian on a Riemannian Manifold: An Introduction to Analysis on Manifolds.* London Mathematical Society Student Texts. Cambridge University Press, Cambridge, 1997.

[94] Judith Rousseau. On the Frequentist Properties of Bayesian Nonparametric Methods. *Annual Review of Statistics and Its Application*, 3(Volume 3, 2016):211–231, June 2016.

[95] Sam T. Roweis and Lawrence K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, December 2000.

[96] Daniel Sanz-Alonso and Ruiyi Yang. Unlabeled Data Help in Graph-Based Semi-Supervised Learning: A Bayesian Nonparametrics Perspective. *Journal of Machine Learning Research*, 23(97):1–28, 2022.

[97] Cornelia Schneider. *Beyond Sobolev and Besov: Regularity of Solutions of PDEs and Their Traces in Function Spaces*, volume 2291 of *Lecture Notes in Mathematics.* Springer International Publishing, Cham, 2021.

[98] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Technical report.

[99] Lorraine Schwartz. On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 4(1):10–26, March 1965.

[100] Weining Shen, Surya T. Tokdar, and Subhashis Ghosal. Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3):623–640, 2013.

[101] Jan Pawel Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Diffusion Models Encode the Intrinsic Dimension of Data Manifolds. In *International Conference on Machine Learning*, 2024.

[102] Bernhard Stankewitz and Botond Szabo. Contraction rates for conjugate gradient and Lanczos approximate posteriors in Gaussian process regression, June 2024.

[103] Charles J. Stone. Optimal Global Rates of Convergence for Nonparametric Regression. *The Annals of Statistics*, 10(4):1040–1053, December 1982.

[104] Rong Tang and Yun Yang. Minimax rate of distribution estimation on unknown submanifolds under adversarial losses. *The Annals of Statistics*, 51(3):1282–1308, June 2023.

[105] Rong Tang and Yun Yang. Adaptivity of Diffusion Models to Manifold Structures. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pages 1648–1656. PMLR, April 2024.

[106] Tao Tang, Nan Wu, Xiuyuan Cheng, and David Dunson. Adaptive Bayesian Regression on Data with Low Intrinsic Dimensionality, September 2024.

[107] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, December 2000.

[108] Hans Triebel. *Theory of Function Spaces*. Springer, Basel, 1983.

[109] Hans Triebel. *Theory of Function Spaces II*. Springer, Basel, 1992.

[110] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, NY, 2009.

[111] A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463, June 2008.

[112] Aad van der Vaart and Harry van Zanten. Information Rates of Nonparametric Gaussian Process Methods. *Journal of Machine Learning Research*, 12(60):2095–2119, 2011.

[113] J. H. van Zanten and A. W. van der Vaart. Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, volume 3, pages 200–223. Institute of Mathematical Statistics, January 2008.

[114] Pascal Vincent and Yoshua Bengio. Manifold Parzen Windows. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002.

[115] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, November 2019.

[116] Kilian Q. Weinberger and Lawrence K. Saul. Unsupervised Learning of Image Manifolds by Semidefinite Programming. *International Journal of Computer Vision*, 70(1):77–90, October 2006.

[117] Yun Yang and David B. Dunson. Bayesian manifold regression. *The Annals of Statistics*, 44(2):876–905, April 2016.

Chapter 2

Posterior Contraction Rates for Matérn Gaussian Processes on Riemannian Manifolds

R., Borovitskiy, Terenin & Rousseau, *Posterior Contraction Rates for Matérn Gaussian Processes on Riemannian Manifolds*, Advances in Neural Information Processing Systems 36 (NeurIPS 2023)

Posterior Contraction Rates for Matérn Gaussian Processes on Riemannian Manifolds

Paul Rosa
University of Oxford

Viacheslav Borovitskiy
ETH Zürich

Alexander Terenin
University of Cambridge
and Cornell University

Judith Rousseau
University of Oxford

Abstract

Gaussian processes are used in many machine learning applications that rely on uncertainty quantification. Recently, computational tools for working with these models in geometric settings, such as when inputs lie on a Riemannian manifold, have been developed. This raises the question: can these intrinsic models be shown theoretically to lead to better performance, compared to simply embedding all relevant quantities into $\mathbb{R}^d$ and using the restriction of an ordinary Euclidean Gaussian process? To study this, we prove optimal contraction rates for intrinsic Matérn Gaussian processes defined on compact Riemannian manifolds. We also prove analogous rates for extrinsic processes using trace and extension theorems between manifold and ambient Sobolev spaces: somewhat surprisingly, the rates obtained turn out to coincide with those of the intrinsic processes, provided that their smoothness parameters are matched appropriately. We illustrate these rates empirically on a number of examples, which, mirroring prior work, show that intrinsic processes can achieve better performance in practice. Therefore, our work shows that finer-grained analyses are needed to distinguish between different levels of data-efficiency of geometric Gaussian processes, particularly in settings which involve small data set sizes and non-asymptotic behavior.

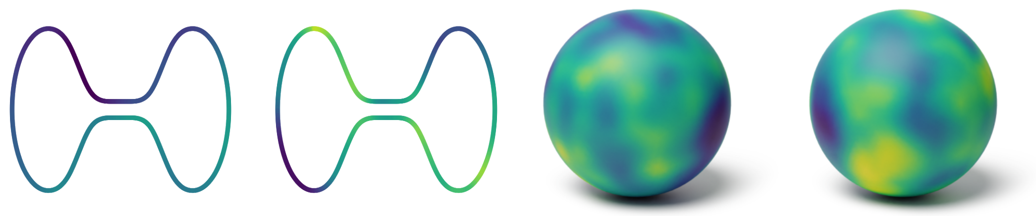
1 Introduction

Gaussian processes provide a powerful way to quantify uncertainty about unknown regression functions via the formulation of Bayesian learning. Motivated by applications in the physical and engineering sciences, a number of recent papers [11, 9, 10, 37] have studied how to extend this model class to spaces with geometric structure, in particular Riemannian manifolds including important special cases such as spheres and Grassmannians [4], hyperbolic spaces and spaces of positive definite matrices [5], as well as general manifolds approximated numerically by a mesh [11].

These Riemannian Gaussian process models are starting to be applied for statistical modeling, and decision-making settings such as Bayesian optimization. For example, in a robotics setting, Jaquier et al. [27] has shown that using Gaussian processes with the correct geometric structure allows one to learn quantities such as the orientation of a robotic arm with less data compared to baselines. The same model class has also been used by Coveney et al. [15] to perform Gaussian process regression on a manifold which models the geometry of a human heart for downstream applications in medicine.

Given these promising empirical results, it is important to understand whether these learning algorithms have good theoretical properties, as well as their limitations. Within the Bayesian framework, a natural way to quantify data-efficiency and generalization error is to posit a data-generating mech-

37th Conference on Neural Information Processing Systems (NeurIPS 2023).



(a) Intrinsic (b) Extrinsic (c) Intrinsic (d) Extrinsic

Figure 1: Samples from different Matérn Gaussian processes on different manifolds, namely a one-dimensional dumbbell-shaped manifold and a two-dimensional sphere. Notice that the values across the dumbbell's bottleneck can be very different for the intrinsic process in (a), despite being very close in the ambient Euclidean distance and in contrast to the situation for the extrinsic model in (b). On the other hand, there is little qualitative difference between (c) and (d), since the embedding produces a reasonably-good global approximation to geodesic distances on the sphere.

anism model and study if—and how fast—the posterior distribution concentrates around the true regression function as the number of observations goes to infinity.

Within the Riemannian setting, it is natural to compare *intrinsic* methods, which are formulated directly on the manifold of interest, with *extrinsic* ones, which require one to embed the manifold within a higher-dimensional Euclidean space. For example, the two-dimensional sphere can be embedded into the Euclidean space $\mathbb{R}^3$: intrinsic Gaussian processes model functions on the sphere while extrinsic ones model functions on $\mathbb{R}^3$, which are then restricted to the sphere. Are the former more efficient than the latter? Since embeddings—even isometric ones—at best only preserve distances locally, they can induce spurious dependencies, as points can be close in the ambient space but far away with respect to the intrinsic geodesic distance: this is illustrated in Figure 1. In cases where embeddings significantly alter distances, one can expect intrinsic models to perform better, and it is therefore interesting to quantify such differences.

In other settings, the manifold on which the data lies can be unknown, which makes using intrinsic methods directly no longer possible. There, one would like to understand how well extrinsic methods can be expected to perform. According to the *manifold hypothesis* [18], it is common for perceptual data such as text and images to concentrate on a lower-dimensional submanifold within, for instance, pixel space or sequence space. It is therefore also interesting to investigate how Gaussian process models—which, being kernel-based, are simpler than for instance deep neural networks—perform in such scenarios, at least in the asymptotic regime.

In this work, we develop geometric analogs of the Gaussian process posterior contraction theorems of van der Vaart and van Zanten [56]. More specifically, we derive posterior contraction rates for three main geometric model classes: (1) the intrinsic Riemannian Matérn Gaussian processes, (2) truncated versions of the intrinsic Riemannian Matérn Gaussian processes, which are used in practice to avoid infinite sums, and (3) the extrinsic Euclidean Matérn Gaussian processes under the assumption that the data lies on a compact Riemannian manifold. In all cases, we focus on IID randomly-sampled input points—commonly referred to as *random design* in the literature—and contraction in the sense of the $L^2(p_0)$ distance, defined in Section 2. We focus on *compact* Riemannian manifolds: this allows one to define Matérn Gaussian processes through their Karhunen–Loève expansions, which requires a discrete spectrum for the Laplace-Beltrami operator—see for instance Borovitskiy et al. [11] and Chavel [13], Chapter 1—and is a common setting in statistics [39].

Contributions. We show that all three classes of Gaussian processes lead to optimal procedures, in the minimax sense, as long as the smoothness parameter of the kernel is aligned with the regularity of the unknown function. While this result is natural—though non trivial—in the case of intrinsic Matérn processes, it is rather remarkable that it also holds for extrinsic ones. This means that in order to understand their differences better, finite-sample considerations are necessary. We therefore present experiments that compute the worst case errors numerically. These experiments highlight that intrinsic models are capable of achieving better performance in the small-data regime. We conclude with a discussion of why these results—which might at first seem counterintuitive—are very natural when viewed from an appropriate mathematical perspective: they suggest that optimality is perhaps best seen as a basic property or an important guarantee that any sensible model should satisfy.

2 Background

Gaussian process regression is a Bayesian approach to regression where the modeling assumptions are $y_i = f(x_i) + \varepsilon_i$, with $\varepsilon_i \sim \mathrm{N}(0, \sigma_\varepsilon^2)$, $x_i \in X$, and f is assigned a Gaussian process prior. A *Gaussian process* is a random function $f : X \to \mathbb{R}$ for which all finite-dimensional marginal distributions are multivariate Gaussian. The distribution of such a process is uniquely determined by its *mean function* $m(\cdot) = \mathbb{E}(f(\cdot))$ and *covariance kernel* $k(\cdot, \cdot') = \mathrm{Cov}(f(\cdot), f(\cdot'))$, hence we write $f \sim \mathrm{GP}(m, k)$.

For Gaussian process regression, the posterior distribution given the data is also a Gaussian process with probability kernel $\Pi(\cdot \mid \boldsymbol{x}, \boldsymbol{y}) = \mathrm{GP}(m_{\Pi(\cdot|\boldsymbol{x},\boldsymbol{y})}, k_{\Pi(\cdot|\boldsymbol{x},\boldsymbol{y})})$, see Rasmussen and Williams [41],

$$m_{\Pi(\cdot|\boldsymbol{x},\boldsymbol{y})}(\cdot) = \mathbf{K}_{(\cdot)\boldsymbol{x}}(\mathbf{K}_{\boldsymbol{x}\boldsymbol{x}} + \sigma_\varepsilon^2\mathbf{I})^{-1}\boldsymbol{y}, \tag{1}$$

$$k_{\Pi(\cdot|\boldsymbol{x},\boldsymbol{y})}(\cdot, \cdot') = \mathbf{K}_{(\cdot,\cdot')} - \mathbf{K}_{(\cdot)\boldsymbol{x}}(\mathbf{K}_{\boldsymbol{x}\boldsymbol{x}} + \sigma_\varepsilon^2\mathbf{I})^{-1}\mathbf{K}_{\boldsymbol{x}(\cdot')}. \tag{2}$$

These quantities describe how incorporating data updates the information contained within the Gaussian process. We will be interested studying the case where X is a Riemannian manifold, but first review the existing theory on the asymptotic behaviour of the posterior when $X = [0, 1]^d$.

2.1 Posterior Contraction Rates

Posterior contraction results describe how the posterior distribution concentrates around the true data generating process, as the number of observations increases, so that it eventually uncovers the true data-generating mechanism. The area of *posterior asymptotics* is concerned with understanding conditions under which this does or does not occur, with questions of *posterior contraction rates*—how fast such convergence occurs—being of key interest. At present, there is a well-developed literature on posterior contraction rates, see Ghosal and van der Vaart [20] for a review.

In the context of Gaussian process regression with *random design*, which is the focus of this paper, the true data generating process is assumed to be of the form

$$y_i \mid x_i \sim \mathrm{N}(f_0(x_i), \sigma_\varepsilon^2) \qquad\qquad x_i \sim p_0 \tag{3}$$

where $f_0 \in \mathcal{F} \subseteq \mathbb{R}^X$, a class of real-valued functions, and $\mathrm{N}(\mu, \sigma^2)$ denotes the Gaussian with moments μ, σ^2. Note that, in this particular variant, these equations exactly mirror those of the Gaussian process model's likelihood, including the use of the same noise variance σ_ε^2 in both cases: in this paper, we focus on the particular case where σ_ε is known in advance. This setting is restrictive, one can extend to an unknown $\sigma_\varepsilon > 0$ using techniques that are not specific to our geometric setting: for instance, the approach of [55] allows to handle an unknown σ_ε if one assumes an upper and lower bound on it and keep the same contraction rates. In practice, more general priors, including ones that do not assume an upper or lower bound on σ_ε, can be used, such as a conjugate one like in Banerjee [6]—these can also be analyzed to obtain contraction rates, albeit with additional considerations. The generalization error for prediction in such models is strongly related to the *weighted L^2 loss* given by

$$\|f - f_0\|_{L^2(p_0)} = \left(\int_X |f(x) - f_0(x)|^2 \,\mathrm{d}p_0(x)\right)^{1/2} \tag{4}$$

which is arguably the natural way of measuring discrepancy between f and f_0, given the fact that the covariates x_i are sampled from p_0. The posterior contraction rate is then defined as

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\, \mathbb{E}_{f\sim\Pi(\cdot|\boldsymbol{x},\boldsymbol{y})} \|f - f_0\|_{L^2(p_0)}^2 \tag{5}$$

where $\mathbb{E}_{f\sim\Pi(\cdot|\boldsymbol{x},\boldsymbol{y})}(\cdot)$ denotes expectation under the posterior distribution while $\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}(\cdot)$ denotes expectation under the true data generating process.[1] In the case of covariates distributed on $[0, 1]^d$, posterior contraction rates have been derived under Matérn Gaussian process priors [47] in van der Vaart and van Zanten [56], who showed the following result.

[1] Note that other notions of posterior contraction can be found in the literature, see Ghosal and van der Vaart [20] and Rousseau [42] for examples that are slightly weaker than the definition we work with.

Result 1 (Theorem 2 of van der Vaart and van Zanten [56]). *In the Bayesian regression model, let f be a mean-zero Matérn Gaussian process prior on $\mathbb{R}^d$ with amplitude σ_f^2, length scale κ, and smoothness $\nu > d/2$. Assume that the true data generating process is given by (3), where p_0 has a Lebesgue density on $X = [0,1]^d$ which is bounded from below and above by $0 < c_{p_0} < C_{p_0} < \infty$, respectively. Let $f_0 \in H^\beta \cap \mathcal{CH}^\beta$ with $\beta > d/2$, where H^β and $\mathcal{CH}^\beta$ the Sobolev and Hölder spaces, respectively. Then there exists a constant $C > 0$, which does not depend on n but does depend on d, σ_f^2, ν, κ, β, p_0, σ_ε^2, $\|f_0\|_{H^\beta(\mathcal{M})}$, and $\|f_0\|_{\mathcal{CH}^\beta(\mathcal{M})}$, such that*

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\, \mathbb{E}_{f\sim\Pi(\cdot|\boldsymbol{x},\boldsymbol{y})} \|f - f_0\|^2_{L^2(p_0)} \leq C n^{-\frac{2\min(\beta,\nu)}{2\nu+d}} \tag{6}$$

and, moreover, the posterior mean satisfies

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \|m_{\Pi(\cdot|\boldsymbol{x},\boldsymbol{y})} - f_0\|^2_{L^2(p_0)} \leq C n^{-\frac{2\min(\beta,\nu)}{2\nu+d}}. \tag{7}$$

Note that $m_{\Pi(\cdot|\boldsymbol{x},\boldsymbol{y})}$ is the Bayes estimator [52] of f associated to the weighted L^2 loss and that the second inequality above is a direct consequence of the first. Therefore the posterior contraction rate implies the same convergence rate for $m_{\Pi(\cdot|\boldsymbol{x},\boldsymbol{y})}$. The best rate is attained when $\beta = \nu$: that is, when true smoothness and prior smoothness match—which is known to be minimax optimal in the problem of estimating f_0: see Tsybakov [52]. In this paper, we extend this result to the manifold setting.

2.2 Related Work and Current State of Affairs

The formalization of posterior contraction rates of Bayesian procedures dates back to the work of Schwartz [46] and Le Cam [29], but has been extensively developed since the seminal paper of Ghosal et al. [19] for various sampling and prior models, see for instance [20, 42] for reviews. This includes, in particular, work on Gaussian process priors [54, 56, 57, 43, 49]. Most of the results in the literature, however, assume Euclidean data: as a consequence, contraction properties of Bayesian models under manifold assumptions are still poorly understood, with exception of some recent developments in both density estimation [7, 8, 60] and regression [63, 60].

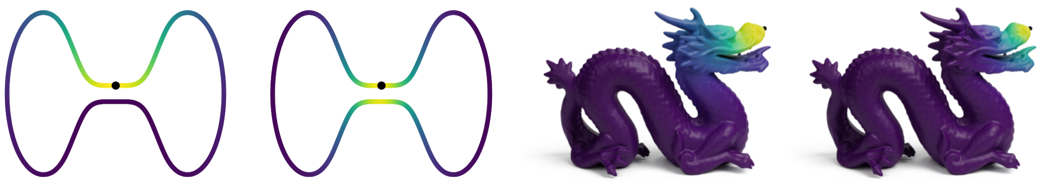
The results closest to ours are those of Yang and Dunson [63] and Castillo et al. [12]. In the former, the authors use an extrinsic length-scale-mixture of squared exponential Gaussian processes to achieve optimal contraction rates with respect to the weighted L^2 norm, using a completely different proof technique compared to us, and their results are restricted to f_0 having Hölder smoothness of order less than or equal to two. On the other hand Castillo et al. [12] consider, as an intrinsic process on the manifold, a hierarchical Gaussian process based on its heat kernel and provide posterior contraction rates. For the Matérn class, Li et al. [30] presents results which characterize the asymptotic behavior of kernel hyperparameters: our work complements these results by studying contraction of the Gaussian process itself toward the unknown ground-truth function. One can also study analogous discrete problems: Dunson et al. [17] and Sanz-Alonso and Yang [45] present posterior contraction rates for a specific graph Gaussian process model in a semi-supervised setting. In the next section, we present our results on Matérn processes, defined either by restriction of an ambient process or by an intrinsic construction, and discuss their implications.

3 Posterior Contraction Rates on Compact Riemannian Manifolds

We now study posterior contraction rates for Matérn Gaussian processes on manifolds, which are arguably the most-widely-used Gaussian process priors in both the Euclidean and Riemannian settings. We begin by more precisely describing our geometric setting before stating our key results and discussing their implications. From now on, we write $X = \mathcal{M}$, to emphasize that the covariate space is a manifold.

Assumption 2. *Assume that $\mathcal{M} \subset \mathbb{R}^D$ is a smooth, compact submanifold (without boundary) of dimension $d < D$ equipped with the standard Riemannian volume measure μ.*

We denote $|\mathcal{M}| = \int_{\mathcal{M}} d\mu(x)$ for volume of $\mathcal{M}$. With this geometric setting defined, we will need to describe regularity assumptions in terms of functional spaces on the manifold $\mathcal{M}$. We work with Hölder spaces $\mathcal{CH}^\gamma(\mathcal{M})$, defined using charts via the usual Euclidean Hölder spaces, the Sobolev spaces $H^s(\mathcal{M})$, and Besov spaces $B^s_{\infty,\infty}(\mathcal{M})$ which are one of the ways of generalizing



(a) Intrinsic (b) Extrinsic (c) Intrinsic (d) Extrinsic

Figure 2: Different Matérn kernels $k(\bullet, x)$ on different manifolds.

the Euclidean Hölder spaces of smooth functions to manifolds. We follow Coulhon et al. [14] and Castillo et al. [12], and define these spaces using the Laplace-Beltrami operator on $\mathcal{M}$ in Appendix A.

Recall that the data-generating process is given by (3), with f_0 as the true regression function and p_0 as the distribution of the covariates.

Assumption 3. *Assume that p_0 is absolutely continuous with respect to μ, and that its density, denoted by p_0, satisfies $c \leq p_0 \leq C$ for $0 < c, C < \infty$. Assume the regression function f_0 satisfies $f_0 \in H^\beta(\mathcal{M}) \cap B^\beta_{\infty,\infty}(\mathcal{M})$ for some $\beta > d/2$, and that $\sigma_\varepsilon^2 > 0$ is fixed and known.*

This setting can be extended to handle unknown variance σ_ε by putting a prior on σ_ε, following the strategy of Salomond [44] and Naulet and Barat [36]. Since we are focused primarily on the impact of the manifold, we do not pursue this here. With the setting fully defined, we proceed to develop posterior contraction results for different types of Matérn Gaussian process priors: intrinsic, intrinsic truncated and extrinsic.

3.1 Intrinsic Matérn Gaussian Processes

We now introduce the first geometric Gaussian process prior under study—the Riemannian Matérn kernel of Whittle [62], Lindgren et al. [33], and Borovitskiy et al. [11]. This process was originally defined using stochastic partial differential equations: here, we present it by its Karhunen–Loève expansion, to facilitate comparisons with its truncated analogs presented in Section 3.2.

Definition 4 (Intrinsic Matérn prior)**.** *Let $\nu > 0$, and let $(\lambda_j, f_j)_{j\geq 0}$ be the eigenvalues and orthonormal eigenfunctions of the Laplace–Beltrami operator on $\mathcal{M}$, in increasing order. Define the intrinsic* Riemannian Matérn Gaussian process *through its Karhunen–Loève expansion to be*

$$f(\cdot) = \frac{\sigma_f^2}{C_{\nu,\kappa}} \sum_{j=1}^{\infty} \left(\frac{2\nu}{\kappa^2} + \lambda_j\right)^{-\frac{\nu+d/2}{2}} \xi_j f_j(\cdot) \qquad\qquad \xi_j \sim \mathrm{N}(0,1) \tag{8}$$

where ν, κ, σ_f^2 are positive parameters and $C_{\nu,\kappa}$ is the normalization constant, chosen such that $\frac{1}{|\mathcal{M}|}\int_M \mathrm{Var}(f(x))\,\mathrm{d}\mu(x) = \sigma_f^2$, where Var *denotes the variance.*

The covariance kernels of these processes are visualized in Figure 2. With this prior, and the setting defined in Section 3, we are ready to present our first result: this model attains the desired optimal posterior contraction rate as soon as the regularity of the ground-truth function matches the regularity of the Gaussian process, as described by the parameter ν.

Theorem 5. *Let f be a Riemannian Matérn Gaussian process prior of Definition 4 with smoothness parameter $\nu > d/2$ and let f_0 satisfy Assumption 3. Then there is a $C > 0$ such that*

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\, \mathbb{E}_{f\sim\Pi(\cdot|\boldsymbol{x},\boldsymbol{y})} \|f - f_0\|^2_{L^2(p_0)} \leq C n^{-\frac{2\min(\beta,\nu)}{2\nu+d}}. \tag{9}$$

All proofs are given in Appendix B. Our proof follows the general approach of van der Vaart and van Zanten [56], by first proving a contraction rate with respect to the distance $n^{-1/2}\|f(\boldsymbol{x}) - f_0(\boldsymbol{x})\|_{\mathbb{R}^n}$ at input locations $\boldsymbol{x}$, and then extending the result to the true L^2-distance by applying a suitable concentration inequality. The first part is obtained by studying the *concentration function*, which is known to be the key quantity to control in order to derive contraction rates of Gaussian process priors—see Ghosal and van der Vaart [20] and van der Vaart and van Zanten [57] for an overview.

Given our regularity assumptions on f_0, the most difficult part lies in controlling the small-ball probabilities $\Pi\big[\|f\|_{\mathcal{C}(\mathcal{M})} < \varepsilon\big]$: we handle this by using results relating this quantity with the entropy of an RKHS unit ball with respect to the uniform norm. Since our process' RKHS is related to the Sobolev space $H^{\nu+d/2}(\mathcal{M})$ which admits a description in terms of charts, we apply results on the entropy of Sobolev balls in the Euclidean space to conclude the first part. Finally, to extend the rate to the true $L^2(p_0)$ norm, following van der Vaart and van Zanten [56], we prove a Hölder-type property for manifold Matérn processes, and apply Bernstein's inequality. Together, this gives the claim.

This result is good news for the intrinsic Matérn model: it tells us that asymptotically it incorporates the data as efficiently as possible at least in terms of posterior contraction rates, given that its regularity matches the regularity of f_0. An inspection of the proof shows that the constant $C > 0$ can be seen to depend on d,σ_f^2, ν, κ, β, p_0,σ_ε^2, $\|f_0\|_{H^\beta(\mathcal{M})}$, $\|f_0\|_{B^\beta_{\infty\infty}(\mathcal{M})}$, and $\|f_0\|_{\mathcal{CH}^\beta(\mathcal{M})}$. Theorem 5 extends to the case where the norm is raised to any power $q > 1$ rather than the second power, with the right-hand side raised to the same power: see Appendix B for details. We now consider variants of this prior that can be implemented in practice.

3.2 Truncated Matérn Gaussian Processes

The Riemannian Matérn prior's covariance kernel cannot in general be computed exactly, since Definition 4 involves an infinite sum. Arguably the simplest way to implement these processes numerically is to truncate the respective infinite series in the Karhunen–Loève expansion by taking the first J terms, which is also optimal in an $L^2(\mathcal{M})$-sense.

Note that the truncated prior is a randomly-weighted finite sum of Laplace–Beltrami eigenfunctions, which have different smoothness properties compared to the original prior: the truncated prior takes its values in $\mathcal{C}^\infty(\mathcal{M})$ since the eigenfunctions of $\mathcal{M}$ are smooth—see for instance De Vito et al. [16]. Nevertheless, if the truncation level is allowed to grow as the sample size increases, then the regularity of the process degenerates and one gets a function with essentially-finite regularity in the limit.

Truncated random basis expansions have been studied extensively in the Bayesian literature in the Euclidean setting—see for instance Arbel et al. [2] and Yoo et al. [64] or Ghosal and van der Vaart [20], Chapter 11 for examples with priors based on wavelet expansions. It is known that truncating the expansion at a high enough level usually allows one to retain optimality. Instead of truncating deterministically, it is also possible to put a prior on the truncation level and resort to MCMC computations which would then select the optimal number of basis functions adaptively, at the expense of a more computationally intensive method—this is done, for instance, in van der Meulen et al. [53] in the context of drift estimation for diffusion processes. Random truncation has been proven to lead in many contexts to adaptive posterior contraction rates, meaning that although the prior does not depend on the smoothness β of f_0, the posterior contraction rate—up to possible $\ln n$ terms—is of order $n^{-\beta/(2\beta+d)}$: see for instance Arbel et al. [2] and Rousseau and Szabo [43].

By analogy of the Euclidean case with its random Fourier feature approximations [40], we can call the truncated version of Definition 4 the *manifold Fourier feature* model, for which we now present our result.

Theorem 6. *Let f be a Riemannian Matérn Gaussian process prior on $\mathcal{M}$ with smoothness parameter $\nu > d/2$, modified to truncate the infinite sum to at least $J_n \geq cn^{\frac{d(\min(1,\nu/\beta))}{2\nu+d}}$ terms, and let f_0 satisfy Assumption 3. Then there is a $C > 0$ such that*

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\, \mathbb{E}_{f\sim\Pi(\cdot|\boldsymbol{x},\boldsymbol{y})} \|f - f_0\|^2_{L^2(p_0)} \leq Cn^{-\frac{2\min(\beta,\nu)}{2\nu+d}}. \tag{10}$$

The proof is essentially-the-same as the non-truncated Matérn, but involves tracking dependence of the inequalities on the truncation level J_n, which implicitly defines a sequence of priors rather than a single fixed prior.

This result is excellent news for the intrinsic models: it means that they inherit the optimality properties of the limiting one, even in the absence of the infinite sum—in spite of the fact that the corresponding finite-truncation prior places its probability on $\mathcal{C}^\infty(\mathcal{M})$. Again, the constant $C > 0$ can be seen to depend on d,σ_f^2, ν,κ,β,p_0,σ_ε^2, $\|f_0\|_{H^\beta(\mathcal{M})}$, $\|f_0\|_{B^\beta_{\infty\infty}(\mathcal{M})}$, and $\|f_0\|_{\mathcal{CH}^\beta(\mathcal{M})}$. This concludes our results for the intrinsic Riemannian Matérn priors. We now study what happens if, instead of working with a geometrically-formulated model, we simply embed everything into $\mathbb{R}^d$ and formulate our models there.

3.3 Extrinsic Matérn Gaussian Processes

The results of the preceding sections provide good reason to be excited about the intrinsic Riemannian Matérn prior: the rates it obtains match the usual minimax rates seen for the Euclidean Matérn prior and Euclidean data, provided that we match the smoothness ν with the regularity of f_0. Another possibility is to consider an extrinsic Gaussian process, that is, a Gaussian process defined over an ambient space. This has been considered by Yang and Dunson [63] for instance for the square-exponential process, in an adaptive setting where one does not assume that the regularity β of f_0 is explicitly known, but where $\beta \leq 2$. In this section we prove a non-adaptive analog of this result for the Matérn process.

Definition 7 (Extrinsic Matérn prior)**.** *Assume that the manifold $\mathcal{M}$ is isometrically embedded in the Euclidean space $\mathbb{R}^D$, such that we can regard $\mathcal{M}$ as a subset of $\mathbb{R}^D$. Consider the Gaussian process with zero mean and kernel given by restricting onto $\mathcal{M}$ the standard Euclidean Matérn kernel*

$$k_{\nu,\kappa,\sigma_f^2}(x, x') = \sigma_f^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{\|x - x'\|_{\mathbb{R}^D}}{\kappa}\right)^{\nu} K_\nu\left(\sqrt{2\nu}\frac{\|x - x'\|_{\mathbb{R}^D}}{\kappa}\right) \tag{11}$$

where $\sigma_f, \kappa, \nu > 0$ and K_ν is the modified Bessel function of the second kind [22].

Since the extrinsic Matérn process is defined in a completely agnostic way with respect to the manifold geometry, we would expect it to be less performant when $\mathcal{M}$ is known. However, it turns out that the extrinsic Matérn process converges at the same rate as the intrinsic one, as given in the following claim.

Theorem 8. *Let f be a mean-zero extrinsic Matérn Gaussian process prior with smoothness parameter $\nu > d/2$ on $\mathcal{M}$, and let f_0 satisfy Assumption 3. Then for some $C > 0$ we have*

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\, \mathbb{E}_{f\sim\Pi(\cdot|\boldsymbol{x},\boldsymbol{y})} \|f - f_0\|^2_{L^2(p_0)} \leq C n^{-\frac{2\min(\beta,\nu)}{2\nu+d}}. \tag{12}$$

Theorem 8 is a surprising result because the optimal rates in this setting only require the knowledge of the regularity β, but not the knowledge of the manifold or the intrinsic dimension. More precisely, the prior is not designed to be an adaptive prior, since it is a fixed Gaussian process, but it surprisingly adapts to the dimension of the manifold, and thus to the manifold.

The proof is also based on control of concentration functions. The main difference is that, although the ambient process has a well known RKHS—the Sobolev space $H^{s+D/2}(\mathbb{R}^D)$—the restricted process has a non-explicit RKHS, which necessitates further analysis. We tackle this issue by using results from Große and Schneider [24] relating manifold and ambient Sobolev spaces by linear bounded trace and extension operators, and from Yang and Dunson [63] describing a general link between the RKHS of an ambient process and its restriction. This allows us to show that the restricted process has an RKHS that is actually norm-equivalent to the Sobolev space $H^{\nu+d/2}(\mathcal{M})$, which allows us to conclude the result in the same manner as in the intrinsic case.

As consequence, our argument applies *mutatis mutandis* in any setting where suitable trace and extension theorems apply, with the Riemannian Matérn case corresponding to the usual Sobolev results. In particular, our arguments therefore apply directly to other processes possessing similar RKHSs, such as for instance various kernels defined on the sphere—see e.g. Wendland [61], Chapter 17 and Hubbert et al. [26]. The constant $C > 0$ can be seen to depend on d,D,σ_f^2, ν,κ,β,p_0,σ_ε^2, $\|f_0\|_{H^\beta(\mathcal{M})}$,$\|f_0\|_{B^\beta_{\infty\infty}(\mathcal{M})}$,$\|f_0\|_{\mathcal{CH}^\beta(\mathcal{M})}$—notice that here C depends implicitly on D because of the presence of trace and extension operator continuity constants. We now proceed to understand the significance of the overall results.

3.4 Summary of Results

As a consequence of our previous results, fixing a single common data generating distribution determined by p_0, f_0, under suitable conditions the intrinsic Matérn process, its truncated version, and the extrinsic Matérn process all possess the *same* posterior contraction rate with respect to the $L^2(p_0)$-norm, which depends on d, ν, and β, and is optimal if the regularities of f_0 and the prior match. These results imply the following immediate corollary, which follows by convexity of $\|\cdot\|^2_{L^2(p_0)}$ using Jensen's inequality.

Corollary 9. *Under the assumptions of Theorems 5, 6 and 8, it follows that, for some $C > 0$*

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \|m_{\Pi(\cdot|\boldsymbol{x},\boldsymbol{y})} - f_0\|^2_{L^2(p_0)} \leq C n^{-\frac{2\min(\beta,\nu)}{2\nu+d}} \tag{13}$$

where $m_{\Pi(\cdot|\boldsymbol{x},\boldsymbol{y})}$ is the posterior mean given a particular value of $(x_i, y_i)_{i=1}^n$.

When $\nu = \beta$, the optimality of the rates we present in the manifold setting can be easily inferred by lower bounding the L^2-risk of the posterior mean by the L^2-risk over a small subset of $\mathcal{M}$ and using charts, which translates the problem into the Euclidean framework for which the rate is known to be optimal—see for instance Tsybakov [52].

To contextualize this, observe that even in cases where the geometry of the manifold is non-flat, the asymptotic rates are unaffected by the choice of the prior's length scale κ—in either the intrinsic, or the extrinsic case—but only by the smoothness parameter ν. Indeed, the RKHS of the process is only determined—up to norm equivalence—by ν, which plays an important role in the proofs. This, and the fact that extrinsic processes attain the same rates, implies that the study of asymptotic posterior contraction rates *cannot detect geometry* in our setting, as was already hinted by Yang and Dunson [63]. Hence, in the geometric setting, optimal posterior contraction rates should be thought of more as a basic property that any reasonable model should satisfy. Differences in performance will be down to constant factors alone—but as we will see, these can be significant. To understand these differences, we turn to empirical analysis.

4 Experiments

From Theorems 5, 6 and 8, we know that intrinsic and extrinsic Gaussian processes exhibit the same posterior contraction rates in the asymptotic regime. Here, we study how these rates manifest themselves in practice, by examining how worst-case errors akin to those of Corollary 9 behave numerically. Specifically, we consider the pointwise worst-case error

$$v^{(\tau)}(t) = \sup_{\|f_0\|_{H^{\nu+d/2}} \leq 1} \mathbb{E}_{\varepsilon_i \sim \mathrm{N}(0,\sigma_\varepsilon^2)} \left| m^{(\tau)}_{\Pi(\cdot|\boldsymbol{x},\boldsymbol{y})}(t) - f_0(t) \right|^2 \tag{14}$$

where $m^{(\tau)}_{\Pi(\cdot|\boldsymbol{x},\boldsymbol{y})}$ is the posterior mean corresponding to the zero-mean Matérn Gaussian process prior with smoothness ν, length scale κ, amplitude σ_f^2, which is intrinsic if $\tau = \mathrm{i}$ or extrinsic if $\tau = \mathrm{e}$. We use a Gaussian likelihood with noise variance σ_ε^2 and observations $y_i = f_0(x_i) + \varepsilon_i$, and examine this quantity as a function of the evaluation location $t \in \mathcal{M}$. By allowing us to assess how error varies in different regions of the manifold, this provides us with a fine-grained picture of how posterior contraction behaves.

One can show that $v^{(\tau)}$ may be computed without numerically solving an infinite-dimensional optimization problem. Specifically, (14) can be calculated, in the respective intrinsic and extrinsic cases, using

$$v^{(\mathrm{i})}(t) = k^{(\mathrm{i})}(t,t) - \mathbf{K}^{(\mathrm{i})}_{t\mathbf{X}} \left(\mathbf{K}^{(\mathrm{i})}_{\mathbf{X}\mathbf{X}} + \sigma_\varepsilon^2 \mathbf{I} \right)^{-1} \mathbf{K}^{(\mathrm{i})}_{\mathbf{X}t} \tag{15}$$

$$v^{(\mathrm{e})}(t) \approx (\mathbf{K}^{(\mathrm{i})}_{t\,\mathbf{X}'} - \boldsymbol{\alpha}_t \mathbf{K}^{(\mathrm{i})}_{\mathbf{X}\,\mathbf{X}'})(\mathbf{K}^{(\mathrm{i})}_{\mathbf{X}'\mathbf{X}'})^{-1}(\mathbf{K}^{(\mathrm{i})}_{\mathbf{X}'\,t} - \mathbf{K}^{(\mathrm{i})}_{\mathbf{X}'\,\mathbf{X}} \boldsymbol{\alpha}_t^\top) + \sigma_\varepsilon^2 \boldsymbol{\alpha}_t \boldsymbol{\alpha}_t^\top \tag{16}$$

where, for the extrinsic case, $\boldsymbol{\alpha}_t = \mathbf{K}^{(\mathrm{i})}_{t\mathbf{X}}(\mathbf{K}^{(\mathrm{e})}_{\mathbf{X}\mathbf{X}} + \sigma_\varepsilon^2 \mathbf{I})^{-1}$, and $\mathbf{X}'$ is a set of points sampled uniformly from the manifold $\mathcal{M}$, the size of which determines approximation quality. The intrinsic expression is simply the posterior variance $k^{(\mathrm{i})}_{\Pi(\cdot|\boldsymbol{x},\boldsymbol{y})}(t,t)$, and its connection with worst-case error is a well-known folklore result mentioned somewhat implicitly in, for instance, Mutny and Krause [35]. The extrinsic expression is very-closely-related, and arises by numerically approximating a certain RKHS norm. A derivation of both is given in Appendix F. To assess the approximation error of this formula, we also consider an analog of (16) but instead defined for the intrinsic model, and compare it to (15): in all cases, the difference between the exact and approximate expression was found to be smaller than differences between models. By computing these expressions, we therefore obtain, up to numerics, the pointwise worst-case expected error in our regression model.

For $\mathcal{M}$ we consider three settings: a dumbbell-shaped manifold, a sphere, and the dragon manifold from the Stanford 3D scanning repository. In all cases, we perform computations by approximating

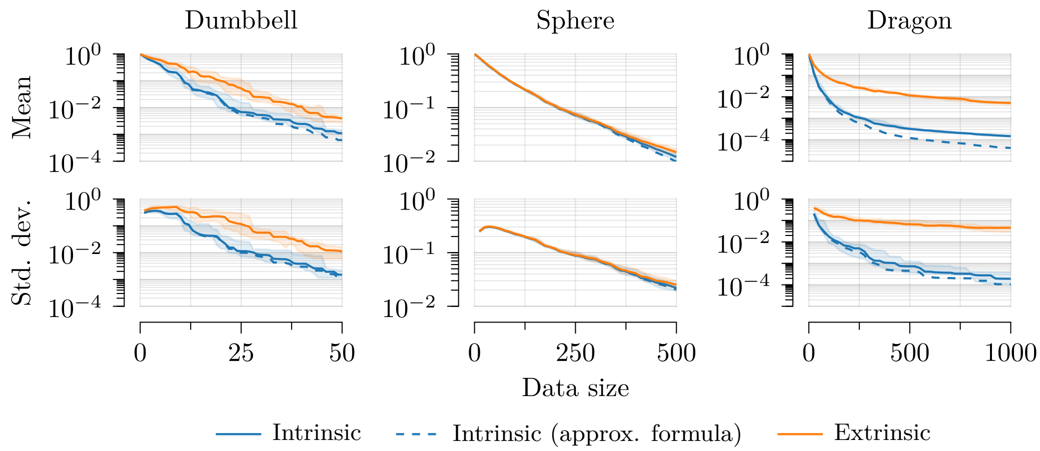


Figure 3: Worst-case error estimates for the intrinsic and extrinsic processes, on the *dumbbell*, *sphere*, and *dragon* manifolds (lower is better, y axis is in the logarithmic scale). We see that, on the dumbbell and dragon manifold, intrinsic models achieve lower expected errors than extrinsic models for the ranges considered (top), and that their expected error consistently varies less as a function of space (bottom). In contrast, on the sphere, both models achieve similar performance, with differences between models falling within the range of variability caused by different random number seeds. We also see that the difference between computing the pointwise worst-case error exactly and approximately, in the intrinsic case where computing this difference is possible, is small in all cases.

the manifold using a mesh, and implementing the truncated Karhunen–Loève expansion with $J = 500$ eigenpairs obtained from the mesh. We fix smoothness $\nu = \frac{5}{2}$, amplitude $\sigma_f^2 = 1$, and noise variance $\sigma_\varepsilon^2 = 0.0005$, for both the intrinsic and extrinsic Matérn Gaussian processes. Since the interpretation of the length scale parameter is manifold-specific, for the intrinsic Gaussian processes we set $\kappa = 200$ for the dumbbell, $\kappa = 0.25$ for the sphere, and $\kappa = 0.05$ for the dragon manifold. In all cases, this yielded functions that were neither close to being globally-constant, nor resembled noise. Each experiment was repeated 10 times to assess variability. Complete experimental details are given in Appendix G.[2]

The length scales κ are defined differently for intrinsic and extrinsic Matérn kernels: in particular, using the same length scale in both models can result in kernels behaving very differently. To alleviate this, for the extrinsic process, we set the length scale by maximizing the extrinsic process' marginal likelihood using the full dataset generated by the intrinsic process, except in the dumbbell's case where the full dataset is relatively small, and therefore a larger set of 500 points was used instead. This allows us to numerically match intrinsic and extrinsic length scales to ensure a reasonably-fair comparison.

Figure 3 shows the mean, and spatial standard deviation of $v_\tau(t)$, where by *spatial standard deviation* we mean the sample standard deviation computed with respect to locations in space, rather than with respect to different randomly sampled datasets. From this, we see that on the dumbbell and dragon manifold—whose geometry differs significantly from the respective ambient Euclidean spaces—intrinsic models obtain better mean performance. The standard deviation plot reveals that intrinsic models have errors that are less-variable across space. This means that extrinsic models exhibit higher errors in some regions rather than others—such as, for instance, regions where embedded Euclidean and Riemannian distances differ—whereas in intrinsic models the error decays in a more spatially-uniform manner.

In contrast, on the sphere, both models perform similarly. Moreover, both the mean and spatial standard deviation decrease at approximately the same rates, indicating that the extrinsic model's predictions are correct about-as-often as the intrinsic model's, as a function of space. This confirms the view that, since the sphere does not possess any bottleneck-like areas where embedded Euclidean distances are extremely different from their Riemannian analogs, it is significantly less affected by differences coming from embeddings.

[2]Code available at: HTTPS://GITHUB.COM/ATERENIN/GEOMETRIC_ASYMPTOTICS.

In total, our experiments confirm that there are manifolds on which geometric models can perform significantly better than non-geometric models. This phenomenon was also noticed in Dunson et al. [17], where a prior based on the eigendecomposition of a random geometric graph, which can be thought as an approximation of our intrinsic Matérn processes, is compared to a standard extrinsic Gaussian process. In our experiments, we see this through expected errors, mirroring prior results on Bayesian optimization performance. From our theoretical results, such differences cannot be captured through posterior contraction rates, and therefore would require sharper technical tools, such as non-asymptotic analysis, to quantify theoretically.

5 Conclusion

In this work, we studied the asymptotic behavior of Gaussian process regression with different classes of Matérn processes on Riemannian manifolds. By using various results on Sobolev spaces on manifolds we derived posterior contraction rates for intrinsic Matérn process defined via their Karhunen-Loeve decomposition in the Laplace–Beltrami eigenbasis, including processes arising from truncation of the respective sum which can be implemented in practice. Next, using trace and extension theorems which relate manifold and Euclidean Sobolev spaces, we derived similar contraction rates for the restriction of an ambient Matérn process in the case where the manifold is embedded in Euclidean space. These theoretical asymptotic results were supplemented by experiments on several examples, showing significant differences in performance between intrinsic and extrincic methods in the small sample size regime when the manifold's geometric structure differs from the ambient Euclidean space. Our work therefore shows that capturing such differences cannot be done through asymptotic contraction rates, motivating and paving the way for further work on non-asymptotic error analysis to capture empirically-observed differences between extrinsic and intrinsic models.

Acknowledgments

The authors are grateful to Mojmír Mutný and Prof. Andreas Krause for fruitful discussions concerning this work. PR and JR were supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 834175). VB was supported by an ETH Zürich Postdoctoral Fellowship. AT was supported by Cornell University, jointly via the Center for Data Science for Enterprise and Society, the College of Engineering, and the Ann S. Bowers College of Computing and Information Science.

References

[1] C. D. Aliprantis and K. C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 2006. Cited on page 29.

[2] J. Arbel, G. Gayraud, and J. Rousseau. Bayesian Optimal Adaptive Estimation Using a Sieve Prior. *Scandinavian Journal of Statistics*, 2013. Cited on page 6.

[3] T. Aubin. *Some nonlinear problems in Riemannian geometry*. Springer, 1998. Cited on page 14.

[4] I. Azangulov, A. Smolensky, A. Terenin, and V. Borovitskiy. Stationary Kernels and Gaussian Processes on Lie Groups and their Homogeneous Spaces I: the compact case. *arXiv:2208.14960*, 2022. Cited on page 1.

[5] I. Azangulov, A. Smolensky, A. Terenin, and V. Borovitskiy. Stationary Kernels and Gaussian Processes on Lie Groups and their Homogeneous Spaces II: non-compact symmetric spaces. *arXiv:2301.13088*, 2023. Cited on page 1.

[6] S. Banerjee. Modeling massive spatial datasets using a conjugate Bayesian linear modeling framework. *Spatial Statistics*, 2020. Cited on page 3.

[7] C. Berenfeld, P. Rosa, and J. Rousseau. Estimating a density near an unknown manifold: a Bayesian nonparametric approach. *arXiv:2205.15717*, 2022. Cited on page 4.

[8] A. Bhattacharya and D. B. Dunson. Nonparametric Bayesian density estimation on manifolds with applications to planar shapes. *Biometrika*, 2010. Cited on page 4.

[9] V. Borovitskiy, I. Azangulov, A. Terenin, P. Mostowsky, M. P. Deisenroth, and N. Durrande. Matérn Gaussian processes on graphs. In *International Conference on Artificial Intelligence and Statistics*, 2021. Cited on page 1.

[10] V. Borovitskiy, M. R. Karimi, V. R. Somnath, and A. Krause. Isotropic Gaussian Processes on Finite Spaces of Graphs. In *Artificial Intelligence and Statistics*, 2023. Cited on page 1.

[11] V. Borovitskiy, A. Terenin, P. Mostowsky, and M. P. Deisenroth. Matérn Gaussian processes on Riemannian manifolds. *Advances in Neural Information Processing Systems*, 2020. Cited on pages 1, 2, 5, 15, 23, 25.

[12] I. Castillo, G. Kerkyacharian, and D. Picard. Thomas Bayes' walk on manifolds. *Probability Theory and Related Fields*, 2013. Cited on pages 4, 5, 15, 32, 33.

[13] I. Chavel. *Eigenvalues in Riemannian Geometry*. Elsevier, 1984. Cited on pages 2, 14.

[14] T. Coulhon, G. Kerkyacharian, and P. Petrushev. Heat Kernel Generated frames in the setting of Dirichlet Spaces. *Journal of Fourier Analysis and Applications*, 2012. Cited on pages 5, 15, 32.

[15] S. Coveney, C. Corrado, C. H. Roney, D. O'Hare, S. E. Williams, M. D. O'Neill, S. A. Niederer, R. H. Clayton, J. E. Oakley, and R. D. Wilkinson. Gaussian Process Manifold Interpolation for Probabilistic Atrial Activation Maps and Uncertain Conduction Velocity. *Philosophical Transactions of the Royal Society A*, 2020. Cited on page 1.

[16] E. De Vito, N. Mücke, and L. Rosasco. Reproducing kernel Hilbert spaces on manifolds: Sobolev and Diffusion spaces. *Analysis and Applications*, 2021. Cited on pages 6, 14, 15, 17, 23, 25, 32.

[17] D. B. Dunson, H.-T. Wu, and N. Wu. Graph Based Gaussian Processes on Restricted Domains. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2021. Cited on pages 4, 10.

[18] C. Fefferman, S. Mitter, and H. Narayanan. Testing the Manifold Hypothesis. *Journal of the American Mathematical Society*, 2016. Cited on page 2.

[19] S. Ghosal, J. K. Ghosh, and A. van der Vaart. Convergence Rates of Posterior Distributions. *The Annals of Statistics*, 2000. Cited on page 4.

[20] S. Ghosal and A. van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2017. Cited on pages 3–6, 17, 21, 28, 31.

[21] E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2015. Cited on pages 14, 15, 32.

[22] I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, 7th edition, 2014. Cited on page 7.

[23] A. Grigoryan. *Heat Kernel and Analysis on Manifolds*. American Mathematical Society, 2009. Cited on page 14.

[24] N. Große and C. Schneider. Sobolev spaces on Riemannian manifolds with bounded geometry: General coordinates and traces. *Mathematische Nachrichten*, 2013. Cited on pages 7, 14, 24, 32.

[25] E. Hebey. *Sobolev spaces on Riemannian manifolds*, volume 1635. Springer, 1996. Cited on page 14.

[26] S. Hubbert, E. Porcu, C. J. Oates, and M. Girolami. Sobolev Spaces, Kernels and Discrepancies over Hyperspheres. *Transactions on Machine Learning Research*, 2023. Cited on page 7.

[27] N. Jaquier, V. Borovitskiy, A. Smolensky, A. Terenin, T. Asfour, and L. Rozo. Geometry-aware bayesian optimization in robotics using Riemannian Matérn kernels. In *Conference on Robot Learning*, 2022. Cited on page 1.

[28] M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv:1807.02582*, 2018. Cited on pages 24, 25.

[29] L. Le Cam. Convergence of Estimates Under Dimensionality Restrictions. *The Annals of Statistics*, 1973. Cited on page 4.

[30] D. Li, W. Tang, and S. Banerjee. Inference for Gaussian Processes with Matern Covariogram on Compact Riemannian Manifolds. *Journal of Machine Learning Research*, 2023. Cited on page 4.

[31] W. V. Li and W. Linde. Approximation, Metric Entropy and Small Ball Estimates for Gaussian Measures. *The Annals of Probability*, 1999. Cited on pages 32, 33.

[32] M. Lifshits. *Lectures on Gaussian processes*. Springer, 2012. Cited on page 17.

[33] F. Lindgren, H. Rue, and J. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2011. Cited on page 5.

[34] S. V. Lototsky and B. L. Rozovsky. *Stochastic partial differential equations*. Springer, 2017. Cited on page 26.

[35] M. Mutny and A. Krause. Experimental design for linear functionals in reproducing kernel hilbert spaces. *Advances in Neural Information Processing Systems*, 2022. Cited on page 8.

[36] Z. Naulet and É. Barat. Some Aspects of Symmetric Gamma Process Mixtures. *Bayesian Analysis*, 2018. Cited on page 5.

[37] M. Niu, P. Cheung, L. Lin, Z. Dai, N. D. Lawrence, and D. B. Dunson. Intrinsic Gaussian processes on complex constrained domains. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2018. Cited on page 1.

[38] T. Pinder and D. Dodd. GPJax: A Gaussian Process Framework in JAX. *Journal of Open Source Software*, 2022. Cited on page 35.

[39] E. Porcu, M. Bevilacqua, R. Schaback, and C. J. Oates. The Matérn Model: A Journey through Statistics, Numerical Analysis and Machine Learning. *arXiv:2303.02759*, 2023. Cited on page 2.

[40] A. Rahimi and B. Recht. Random Features for Large-Scale Kernel Machines. *Advances in Neural Information Processing Systems*, 2007. Cited on page 6.

[41] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. Cited on page 3.

[42] J. Rousseau. On the frequentist properties of Bayesian nonparametric methods. *Annual Review of Statistics and its Application*, 2016. Cited on pages 3, 4.

[43] J. Rousseau and B. Szabo. Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator. *The Annals of Statistics*, 2017. Cited on pages 4, 6.

[44] J.-B. Salomond. Testing Un-Separated Hypotheses by Estimating a Distance. *Bayesian Analysis*, 2018. Cited on page 5.

[45] D. Sanz-Alonso and R. Yang. Unlabeled Data Help in Graph-Based Semi-Supervised Learning: A Bayesian Nonparametrics Perspective. *Journal of Machine Learning Research*, 2022. Cited on page 4.

[46] L. Schwartz. On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 1965. Cited on page 4.

[47] M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, 1999. Cited on page 3.

[48] R. S. Strichartz. Analysis of the Laplacian on the complete Riemannian manifold. *Journal of functional analysis*, 52(1):48–79, 1983. Cited on page 14.

[49] W. Tang, L. Zhang, and S. Banerjee. On identifiability and consistency of the nugget in Gaussian spatial process models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2021. Cited on page 4.

[50] H. Triebel. *Theory of Function Spaces I*. Birkhäuser, 1983. Cited on pages 15, 27.

[51] H. Triebel. *Theory of Function Spaces II*. Birkhäuser, 1992. Cited on pages 14, 15.

[52] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2008. Cited on pages 4, 8.

[53] F. van der Meulen, M. Schauer, and H. van Zanten. Reversible jump MCMC for nonparametric drift estimation for diffusion processes. *Computational Statistics and Data Analysis*, 2014. Cited on page 6.

[54] A. van der Vaart and H. van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *The Annals of Statistics*, 2009. Cited on page 4.

[55] A. van der Vaart and H. van Zanten. Bayesian inference with rescaled Gaussian process priors. *Electronic Journal of Statistics*, 2007. Cited on page 3.

[56] A. van der Vaart and H. van Zanten. Information Rates of Nonparametric Gaussian Process Methods. *Journal of Machine Learning Research*, 2011. Cited on pages 2–6, 17–22, 30, 31.

[57] A. van der Vaart and H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, 2008. Cited on pages 4, 5, 17.

[58] A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer Verlag, 1996. Cited on page 22.

[59] H. van Zanten and A. van der Vaart. Reproducing kernel Hilbert spaces of Gaussian priors. *Institute of Mathematical Statistics Collections*, 2008. Cited on pages 17, 30, 32.

[60] X. Wang and G. Lerman. Nonparametric Bayesian regression on manifolds via Brownian motion. *arXiv:1507.06710*, 2015. Cited on page 4.

[61] H. Wendland. *Scattered Data Approximation*. Cambridge University Press, 2004. Cited on pages 7, 25.

[62] P. Whittle. Stochastic Processes in Several Dimensions. *Bulletin of the International Statistical Institute*, 1963. Cited on page 5.

[63] Y. Yang and D. B. Dunson. Bayesian manifold regression. *The Annals of Statistics*, 2016. Cited on pages 4, 7, 8, 24.

[64] W. W. Yoo, V. Rivoirard, and J. Rousseau. Adaptive supremum norm posterior contraction: Wavelet spike-and-slab and anisotropic Besov spaces. *arXiv:1708.01909*, 2017. Cited on page 6.

A Preliminaries

Here we describe preliminaries necessary for Appendices B to E. This includes some basic properties of the Laplace–Beltrami operator on compact manifolds, partitions of unity subordinate to atlases, function spaces such as Hölder, Sobolev and Besov spaces, general Gaussian random elements on Banach spaces, and a certain technical lemma. Hereinafter, the expression $a \lesssim b$ means $a \leq Cb$ for some constant $C > 0$ whose value is irrelevant to our claims.

A.1 Laplace–Beltrami Operator and Subordinate Partitions of Unity

Let $\mathcal{M}$ denote a compact connected Riemannian manifold without boundary of dimension $d \in \mathbb{Z}_{>0}$. The Laplace–Beltrami operator Δ on $\mathcal{M}$ is self-adjoint and positive semi-definite [48, Theorem 2.4]. Let $(L^2(\mathcal{M}), \langle\cdot,\cdot\rangle_{L^2(\mathcal{M})})$ denote the Hilbert space of square integrable (equivalence classes of) functions on $\mathcal{M}$ with respect to the standard Riemannian volume measure.

By standard theory [13, 23], there exists an orthonormal basis $\{f_j\}_{j=0}^{\infty}$ of $L^2(\mathcal{M})$ consisting of the eigenfunctions of Δ, namely $\Delta f_j = -\lambda_j f_j$, with $\lambda_j \geq 0$. We assume that the pairs (λ_j, f_j) are sorted such that $0 = \lambda_0 \leq \lambda_j \leq \lambda_{j+1}$. The growth of λ_j can be characterized as follows.

Result 10 (Weyl's Law)**.** *There exists a constant $C > 0$ such that for all j large enough we have*

$$C^{-1} j^{2/d} \leq \lambda_j \leq C j^{2/d}. \tag{17}$$

Proof. See Chavel [13], Chapter 1. □

Following De Vito et al. [16] and Große and Schneider [24] we fix a family $\mathcal{T} = (\mathcal{U}_l, \phi_l, \chi_l)_{l=1}^{L}$, where $L \in \mathbb{Z}_{>0}$, the local coordinates $\phi_l : \mathcal{U}_l \subset \mathcal{M} \to \mathcal{V}_l = \phi_l(\mathcal{U}_l) \subset \mathbb{R}^d$ are smooth diffeomorphisms, and the functions χ_l form a partition of unity subordinate to $\{\mathcal{U}_l\}_{l=1}^{L}$, that is $\chi_l \in \mathcal{C}^\infty(\mathcal{M})$, $\operatorname{supp}(\chi_l) \subset \mathcal{U}_l$, $0 \leq \chi_l \leq 1$ and $\sum_l \chi_l = 1$—here, we can choose L finite by compactness of $\mathcal{M}$. For convenience and without loss of generality, we assume that $\mathcal{V}_l = (0,1)^d$. With this, we can start defining function spaces on $\mathcal{M}$.

A.2 Hölder Spaces $\mathcal{CH}^\gamma$ and the spaces of continuous functions $\mathcal{C}$ and smooth functions $\mathcal{C}^\infty$

Consider an arbitrary domain $\mathcal{X} \subseteq \mathbb{R}^d$ or $\mathcal{X} \subseteq \mathcal{M}$. We denote the class of infinitely differentiable functions on $\mathcal{X}$ by $\mathcal{C}^\infty(\mathcal{X})$. Let $\mathcal{C}^k(\mathcal{X})$ denote the Banach space of $k \in \mathbb{Z}_{>0}$ times continuously differentiable functions on $\mathcal{X}$ with finite norm

$$\|f\|_{\mathcal{C}^k(\mathcal{X})} = \sup_{x\in\mathcal{X}} \left|\nabla^k f(x)\right| \tag{18}$$

where ∇^k is the kth covariant derivative, as in Hebey [25], Section 2.1 and Aubin [3], Definition 2.2. We also write $\mathcal{C}(\mathcal{X}) = \mathcal{C}^0(\mathcal{X})$ for the space of continuous functions on $\mathcal{X}$.

The Euclidean Hölder spaces $\mathcal{CH}^\gamma(\mathbb{R}^d)$, where $\gamma = k + \alpha$ with $k \in \mathbb{Z}_{\geq 0}$ and $0 < \alpha \leq 1$, are defined by[3]

$$\mathcal{CH}^\gamma(\mathbb{R}^d) = \left\{ f \in \mathcal{C}^k(\mathbb{R}^d) : \|f\|_{\mathcal{CH}^\gamma(\mathbb{R}^d)} < \infty \right\} \tag{19}$$

where

$$\|f\|_{\mathcal{CH}^\gamma(\mathbb{R}^d)} = \|f\|_{\mathcal{C}^k(\mathbb{R}^d)} + \sup_{\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^d,\, \boldsymbol{x}\neq\boldsymbol{y}} \frac{|f(\boldsymbol{x}) - f(\boldsymbol{y})|}{\|\boldsymbol{x} - \boldsymbol{y}\|_{\mathbb{R}^d}^{\alpha}}. \tag{20}$$

More information on these definitions may be found, for instance, in Giné and Nickl [21] and Triebel [51]. We now turn to the manifold versions of the Hölder spaces.

[3] Hölder spaces are often also denoted by $\mathcal{C}^\gamma$ as well, with $\gamma \in \mathbb{R}_{>0}$. Since, using this formulation, they do *not* coincide with $\mathcal{C}^k(\mathcal{X})$ when $k = \gamma \in \mathbb{Z}_{\geq 0}$, we use the notation $\mathcal{CH}^\gamma$ to avoid confusion.

Definition 11 (Hölder spaces)**.** *For all $\gamma > 0$ we define the Hölder space $\mathcal{CH}^\gamma(\mathcal{M})$ on the manifold $\mathcal{M}$ to be the space of all $f : \mathcal{M} \to \mathbb{R}$ satisfying*

$$\|f\|_{\mathcal{CH}^\gamma(\mathcal{M})} = \sum_{l=1}^{L} \|(\chi_l f) \circ \phi_l^{-1}\|_{\mathcal{CH}^\gamma(\mathbb{R}^d)} < \infty. \tag{21}$$

Since the charts ϕ_l are smooth, Definition 11 can be easily seen to be independent of the chosen atlas, with equivalence of norms.

A.3 Sobolev and Besov Spaces

We now introduce the manifold versions of the Sobolev and Besov spaces, whose definitions in the standard Euclidean case may be found, for instance, in Triebel [51]. For Sobolev spaces we use the Bessel-potential-based definition, following De Vito et al. [16].

Definition 12 (Sobolev spaces)**.** *For any $s > 0$ we define the Sobolev space $H^s(\mathcal{M})$ on the manifold $\mathcal{M}$ as the Hilbert space of functions $f \in L^2(\mathcal{M})$ such that $\|f\|^2_{H^s(\mathcal{M})} = \langle f, f\rangle_{H^s(\mathcal{M})} < \infty$ where*

$$\langle f, g\rangle_{H^s(\mathcal{M})} = \sum_{j=0}^{\infty} (1+\lambda_j)^s \langle f, f_j\rangle_{L^2(\mathcal{M})} \langle g, f_j\rangle_{L^2(\mathcal{M})}. \tag{22}$$

Remark 13. *It is easy to see that substituting $(1+\lambda_j)^s$ in (22) with $\beta(\alpha+\lambda_j)^s$ or with $(\alpha+\beta\lambda_j^s)$ for any $\alpha, \beta > 0$ results in the same set of functions and an equivalent norm. The former follows from Borovitskiy et al. [11], eq. (109). The latter follows from the Binomial Theorem.*

For Besov spaces we follow Coulhon et al. [14] and Castillo et al. [12] and define them in terms of approximations by low-frequency functions. We fix a function $\Phi \in \mathcal{C}^\infty(\mathbb{R}_{\geq 0}, \mathbb{R}_{\geq 0})$ such that $K = \operatorname{supp}(\Phi) \subset [0, 2]$ and $\Phi(x) = 1$ for $x \in [0, 1]$. We also define the functions $\Phi_j(x) = \Phi\big(2^{-j}x\big)$.

Coulhon et al. [14], Corollary 3.6 shows that the operators $\Phi_j(\sqrt{\Delta})$ defined by

$$\Phi_j\Big(\sqrt{\Delta}\Big) f = \sum_{j\geq 0} \Phi_j\Big(\sqrt{\lambda_j}\Big) \langle f_j, f\rangle f_j \tag{23}$$

are bounded in the space $L^p(\mathcal{M})$ for all $1 \leq p \leq \infty$.[4] Moreover, the same result also shows that we can express any $f \in L^p(\mathcal{M})$ as $f = \lim_{j\to\infty} \Phi_j(\sqrt{\Delta})f$ in $L^p(\mathcal{M})$. $\Phi_j(\sqrt{\Delta})f$ can intuitively be considered as a version of f filtered by a low-pass filter. The next definition introduces the Besov spaces $B^s_{p,q}(\mathcal{M})$, which are formulated in terms of quality-of-approximation by low-frequency functions.

Definition 14 (Besov spaces)**.** *For any $s > 0$ and $1 \leq p, q \leq \infty$ we define the Besov space $B^s_{p,q}(\mathcal{M})$ on the manifold $\mathcal{M}$ as the space of functions $f \in L^p(\mathcal{M})$ such that $\|f\|_{B^s_{p,q}(\mathcal{M})} < \infty$ where*

$$\|f\|_{B^s_{p,q}(\mathcal{M})} = \begin{cases} \|f\|_{L^p(\mathcal{M})} + \left(\sum_{j\geq 0}\left(2^{js}\|\Phi_j(\sqrt{\Delta})f - f\|_{L^p(\mathcal{M})}\right)^q\right)^{1/q} & \text{if } q < +\infty \\ \|f\|_{L^p(\mathcal{M})} + \sup_{j\geq 0} 2^{js}\|\Phi_j(\sqrt{\Delta})f - f\|_{L^p(\mathcal{M})} & \text{if } q = +\infty. \end{cases} \tag{24}$$

The classical Besov spaces $B^s_{2,2}$ coincide with the Sobolev spaces H^s on $\mathbb{R}^d$, in the sense that they define the same set of functions and equivalent norms—see for instance Giné and Nickl [21] section 4.3.6—and even on manifolds if one follows the construction of Triebel [51], pages 7.3–7.4 for Besov spaces. Since our definition of the Besov spaces is somewhat non-standard, we present a proof.

Proposition 15. *For all $s > 0$, $H^s(\mathcal{M}) = B^s_{2,2}(\mathcal{M})$ as sets and there exist two constants $C_1, C_2 > 0$ such that for all $f \in H^s(\mathcal{M}) = B^s_{2,2}(\mathcal{M})$ we have*

$$C_1\|f\|_{H^s(\mathcal{M})} \leq \|f\|_{B^s_{2,2}(\mathcal{M})} \leq C_2\|f\|_{H^s(\mathcal{M})}. \tag{25}$$

[4] The space $L^p(\mathcal{M})$ is the Banach space of functions (or rather their equivalence classes) that are integrable when raised to the power $p < \infty$ or essentially bounded for $p = \infty$. See for instance Triebel [50] for details.

Proof. It is enough to prove (25), the rest will follow automatically. The main technical tools used in the proof are Result 10 and summation by parts. First, we prove the upper bound. Denote the support set of Φ by $K = \operatorname{supp}(\Phi)$. Notice that

$$\left(\|f\|_{B^s_{2,2}(\mathcal{M})} - \|f\|_{L^2(\mathcal{M})}\right)^2 = \sum_{j\geq 0} 2^{2js} \left\|\Phi_j\left(\sqrt{\Delta}\right) f - f\right\|^2_{L^2(\mathcal{M})} \tag{26}$$

$$= \sum_{j\geq 0} 2^{2js} \sum_{l:\sqrt{\lambda_l}\notin 2^j K} \left|\langle f_l, f\rangle_{L^2(\mathcal{M})}\right|^2 \tag{27}$$

$$\leq \sum_{j\geq 0} 2^{2js} \sum_{l:\sqrt{\lambda_l}> 2^j} \left|\langle f_l, f\rangle_{L^2(\mathcal{M})}\right|^2. \tag{28}$$

The last inequality results from the fact that $[0,1]\subset K$. By Weyl's Law (Result 10), there exists a constant $C>0$ such that $\lambda_l \leq C l^{2/d}$. Without loss of generality, we can assume that $C = 2^{2r}$, $r\in\mathbb{Z}_{>0}$. Hence $\sqrt{\lambda_l} > 2^j$ implies $l > 2^{d(j-r)}$, thus we have

$$\left(\|f\|_{B^s_{2,2}(\mathcal{M})} - \|f\|_{L^2(\mathcal{M})}\right)^2 \leq \sum_{j\geq 0} 2^{2js} \sum_{l>2^{d(j-r)}} \left|\langle f_l, f\rangle_{L^2(\mathcal{M})}\right|^2 \tag{29}$$

$$= \sum_{0\leq j\leq r} 2^{2js} \sum_{l>2^{d(j-r)}} \left|\langle f_l, f\rangle_{L^2(\mathcal{M})}\right|^2 + \sum_{j>r} 2^{2js} \sum_{l>2^{d(j-r)}} \left|\langle f_l, f\rangle_{L^2(\mathcal{M})}\right|^2 \tag{30}$$

$$\leq r2^{2rs}\|f\|^2_{L^2(\mathcal{M})} + 2^{2rs} \sum_{j>0} 2^{2js} \sum_{l>2^{dj}} \left|\langle f_l, f\rangle_{L^2(\mathcal{M})}\right|^2. \tag{31}$$

Now let $R_j = \sum_{l>2^{dj}} \left|\langle f_l, f\rangle_{L^2(\mathcal{M})}\right|^2$ and $S_J = \sum_{j=1}^J 2^{2js} \leq \frac{2^{2s}}{2^{2s}-1} 2^{2Js}$, $S_0 = 0$. Write

$$\sum_{j>0} 2^{2js} \sum_{l>2^{dj}} \left|\langle f_l, f\rangle_{L^2(\mathcal{M})}\right|^2 = \sum_{j>0} (S_j - S_{j-1}) R_j \tag{32}$$

$$= \sum_{j>0} S_j (R_j - R_{j+1}) \tag{33}$$

$$= \sum_{j>0} S_j \sum_{2^{dj}<l\leq 2^{d(j+1)}} \left|\langle f_l, f\rangle_{L^2(\mathcal{M})}\right|^2 \tag{34}$$

$$\leq \frac{2^{2s}}{2^{2s}-1} \sum_{j>0} 2^{2js} \sum_{2^{dj}<l\leq 2^{d(j+1)}} \left|\langle f_l, f\rangle_{L^2(\mathcal{M})}\right|^2 \tag{35}$$

$$\leq \frac{2^{2s}}{2^{2s}-1} \sum_{j>0} \sum_{2^{dj}<l\leq 2^{d(j+1)}} l^{2s/d} \left|\langle f_l, f\rangle_{L^2(\mathcal{M})}\right|^2 \tag{36}$$

$$\leq \frac{c^s 2^{2s}}{2^{2s}-1} \sum_{j>0} \sum_{2^{dj}<l\leq 2^{d(j+1)}} \lambda_l^s \left|\langle f_l, f\rangle_{L^2(\mathcal{M})}\right|^2 \tag{37}$$

$$= \frac{c^s 2^{2s}}{2^{2s}-1} \sum_{l>2^d} \lambda_l^s \left|\langle f_l, f\rangle_{L^2(\mathcal{M})}\right|^2 \tag{38}$$

$$\leq \frac{c^s 2^{2s}}{2^{2s}-1} \sum_{l\geq 0} \lambda_l^s \left|\langle f_l, f\rangle_{L^2(\mathcal{M})}\right|^2 \tag{39}$$

where we have used Result 10 to get existence of a c such that $l^{2/d} \leq c\lambda_l$. This proves the upper bound. The proof for the lower bound is similar. □

Proposition 15 provides a characterization of the Sobolev spaces $H^s(\mathcal{M})$. There is yet another characterization of these spaces that will be useful later, in terms of charts. We present this characterization as part of the following result.

Theorem 16. *On the Sobolev space $H^s(\mathcal{M})$, the following norms are equivalent:*

$$\|f\|_{H^s(\mathcal{M})} = \left(\sum_{j=0}^{\infty}(1+\lambda_j)^s\langle f, f_j\rangle_{L^2(\mathcal{M})}^2\right)^{1/2} \tag{40}$$

$$\|f\|_{B^s_{2,2}(\mathcal{M})} = \|f\|_{L^2(\mathcal{M})} + \left(\sum_{j\geq 0}\left(2^{js}\|\Phi_j(\sqrt{\Delta})f - f\|_{L^2(\mathcal{M})}\right)^2\right)^{1/2} \tag{41}$$

$$\|f\|_{H^s_{\mathcal{T}}(\mathcal{M})} = \left(\sum_{l=1}^{L}\left\|(\chi_l f)\circ\phi_l^{-1}\right\|_{H^s(\mathbb{R}^d)}^2\right)^{1/2} \tag{42}$$

Proof. The equivalence between $\|\cdot\|_{H^s(\mathcal{M})}$ and $\|f\|_{B^s_{2,2}(\mathcal{M})}$ is given by Proposition 15. The proof of equivalence between $\|\cdot\|_{H^s(\mathcal{M})}$ and $\|f\|_{H^s_{\mathcal{T}}(\mathcal{M})}$ can be found in De Vito et al. [16]. $\square$

A.4 Gaussian Random Elements

Here we recall the definition of a Gaussian process as a Banach-space-valued random variable, following for instance van Zanten and van der Vaart [59].

Definition 17 (Gaussian random element)**.** *Let $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$ be a Banach space, and f be a Borel random variable with values in $\mathbb{B}$ almost surely. We say that f is a Gaussian random element if $b^*(f)$ is a univariate Gaussian random variable for every bounded linear functional $b^* : \mathbb{B} \to \mathbb{R}$.*

Random variables of this kind are also sometimes called *Gaussian in the sense of duality*. One should think of a Gaussian random element as a generalization of a Gaussian process, but which is better-behaved from a function-analytic point of view and in particular does not require the process to be an actual function—as opposed to, for instance, a measure or a distribution. Many connections between the usual Gaussian processes and Gaussian random elements exist, see Lifshits [32], Ghosal and van der Vaart [20], Appendix I, van der Vaart and van Zanten [57] for details. The following observation about Gaussian random elements will be useful later.

Lemma 18. *A Gaussian process f on the manifold $\mathcal{M}$ with almost surely continuous sample paths is a Gaussian random element in the Banach space $\left(\mathcal{C}(\mathcal{M}), \|\cdot\|_{\mathcal{C}(\mathcal{M})}\right)$ of continuous functions on $\mathcal{M}$.*

Proof. Since $\mathcal{C}(\mathcal{M})$ is separable, this follows from Lemma I.6 in Ghosal and van der Vaart [20]. $\square$

A.5 A Technical Lemma

In order to apply Bernstein's inequality when going from the error at input locations to the $L^2(p_0)$ error, we will use the following technical extrapolation lemma.

Lemma 19. *For any function $g : \mathcal{M} \to \mathbb{R}$, a number $\gamma \in \mathbb{R}_{>0} \setminus \mathbb{Z}_{>0}$ and a density $p_0 : \mathcal{M} \to \mathbb{R}_{>0}$ with $1 \lesssim p_0$, we have*

$$\|g\|_{L^\infty(\mathcal{M})} \lesssim \|g\|_{\mathcal{CH}^\gamma(\mathcal{M})}^{\frac{d}{2\gamma+d}}\|g\|_{L^2(p_0)}^{\frac{2\gamma}{2\gamma+d}}. \tag{43}$$

Proof. We use Lemma 15 from van der Vaart and van Zanten [56] and push it through charts. More precisely we have, using $B^\gamma_{\infty,\infty}\left([0,1]^D\right) = \mathcal{CH}^\gamma\left([0,1]^D\right)$ for $\gamma \notin \mathbb{Z}_{>0}$, that

$$\|g\|_{L^\infty(\mathcal{M})} \leq \sum_l \left\|(\chi_l g)\circ\phi_l^{-1}\right\|_{L^\infty(\mathcal{V}_l)} \tag{44}$$

$$\lesssim \max_l \left\|(\chi_l g)\circ\phi_l^{-1}\right\|_{\mathcal{CH}^\gamma(\mathcal{V}_l)}^{\frac{d}{2\gamma+d}}\left\|(\chi_l g)\circ\phi_l^{-1}\right\|_{L^2(\mathcal{V}_l)}^{\frac{2\gamma}{2\gamma+d}}. \tag{45}$$

By definition of the the manifold Hölder spaces, this gives

$$\|g\|_{L^\infty(\mathcal{M})} \lesssim \|g\|_{\mathcal{CH}^\gamma(\mathcal{M})}^{\frac{d}{2\gamma+d}} \max_l \left\|(\chi_l g)\circ\phi_l^{-1}\right\|_{L^2(\mathcal{V}_l)}^{\frac{2\gamma}{2\gamma+d}}. \tag{46}$$

Finally, since χ_l are bounded, the charts are smooth, and p_0 is lower bounded, we have

$$\left\|(\chi_l g) \circ \phi_l^{-1}\right\|_{L^2(\mathcal{V}_l)}^2 = \int_{\mathcal{V}_l} \left|(\chi_l g) \circ \phi_l^{-1}(y)\right|^2 dy \lesssim \int_{\mathcal{U}_l} g^2(x) p_0(x) \mu(dx) \lesssim \|g\|_{L^2(\mathcal{M})}^2 \tag{47}$$

which gives the result. $\square$

This concludes the necessary preliminaries. We now turn to the proofs of our main results.

B Proofs

Throughout this section we use the notation from Appendix A and results from Appendices C to E. We start by defining our notation for Gaussian likelihood and probability distribution of the sample.

Definition 20. *For every $\boldsymbol{x} \in \mathcal{M}^n$ and $f : \mathcal{M} \to \mathbb{R}$ we define $p_{\boldsymbol{x},\boldsymbol{y}}^{(f)}$ to be the joint distribution corresponding to the marginal $p_{\boldsymbol{x}} = p_0$ and conditional $p_{\boldsymbol{y}|\boldsymbol{x}}^{(f)} = \mathrm{N}(f(\boldsymbol{x}), \sigma_\varepsilon^2 \mathbf{I})$, where $f(\boldsymbol{x})$ is the vector with entries $f(x_i)$. Expectations with respect to $p_{\boldsymbol{x},\boldsymbol{y}}^{(f)}$, to $p_{\boldsymbol{x}}$ and to $p_{\boldsymbol{y}|\boldsymbol{x}}^{(f)}$ we denote by $\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}^{(f)}$, by $\mathbb{E}_{\boldsymbol{x}}$ and by $\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}^{(f)}$, respectively. Sometimes, when f is clear from the context, we omit the subscript ${}^{(f)}$.*

Using van der Vaart and van Zanten [56], Theorem 1, which is valid for any compact metric space and thus also for $\mathcal{M}$, we can deduce a posterior contraction rate at data input locactions, with respect to the *empirical L^2-norm*[5]

$$\|f\|_n = \left(\frac{1}{n}\sum_{i=1}^n f(x_i)^2\right)^{1/2} \tag{48}$$

by studying the *concentration functions* with respect to the uniform norm. This is the purpose of the following lemma, whose proof mainly follows [56], but crucially relies on two new components: (i) the *small ball asymptotics* for manifolds we study in Appendix E and (ii) the characterization of the RKHS corresponding to both intrinsic and extrinsic priors as manifold Sobolev spaces we obtain in Appendix C. We recall that the prior Π_n may depend on n if we consider a truncated intrinsic Matérn process.

Theorem 21. *Let Π_n denote the prior in either Theorem 5, Theorem 6 or Theorem 8 with smoothness parameter $\nu > d/2$. Let $\mathbb{H}_n$ denote the corresponding RKHS. Define the* CONCENTRATION FUNCTION *for $f_0 \in C(\mathcal{M})$ and $\varepsilon > 0$ by*

$$\varphi_{f_0}(\varepsilon) = -\ln \mathbb{P}_{f\sim\Pi_n}\left[\|f\|_{L^\infty(\mathcal{M})} < \varepsilon\right] + \inf_{f\in\mathbb{H}_n : \|f-f_0\|_{L^\infty(\mathcal{M})}<\varepsilon} \|f\|_{\mathbb{H}_n}^2. \tag{49}$$

Then if $f_0 \in H^\beta(\mathcal{M}) \cap B_{\infty,\infty}^\beta(\mathcal{M}), \beta > 0$ we have $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$ for ε_n a multiple of $n^{-\frac{\min(\nu,\beta)}{2\nu+d}}$.

Proof. The first term on the right-hand side of Equation (49) is bounded by $C\varepsilon^{-d/\nu}$ by Lemma 33. To bound the second term, we assume, without loss of generality,[6] that $\nu \geq \beta$. Consider an approximation $f = \Phi_j(\sqrt{\Delta}) f_0$ of f_0, where $c\varepsilon \leq 2^{-\beta j} \leq \varepsilon$ and $c > 0$ does not depend on j. Since we assume $f_0 \in B_{\infty,\infty}^\beta(\mathcal{M})$, by definition of $B_{\infty,\infty}^\beta(\mathcal{M})$ we have

$$\|f_0 - f\|_{L^\infty(\mathcal{M})} \leq \|f_0\|_{B_{\infty,\infty}^\beta(\mathcal{M})} 2^{-\beta j} \lesssim \varepsilon \tag{50}$$

where in the last inequality the $B_{\infty,\infty}^\beta(\mathcal{M})$-norm is the constant implied by notation $\lesssim$. We now show that

$$\|f\|_{\mathbb{H}_n}^2 \lesssim \varepsilon^{-\frac{2}{\beta}(\nu-\beta+d/2)}. \tag{51}$$

First notice that by Lemma 24 and Proposition 27, for any prior considered here we have $\mathbb{H}_n \subseteq H^{\nu+d/2}(\mathcal{M})$ and $\|\cdot\|_{\mathbb{H}_n} \lesssim \|\cdot\|_{H^{\nu+d/2}(\mathcal{M})}$ with a constant that does not depend on n. Hence using

[5]This is actually a seminorm, but we follow the rest of the literature in referring to it as a norm.

[6]If $\beta > \nu$ then $f_0 \in H^\beta(\mathcal{M}) \cap B_{\infty,\infty}^\beta \subseteq H^\nu(\mathcal{M}) \cap B_{\infty,\infty}^\nu(\mathcal{M})$ gives $\varepsilon_n \propto n^{-\frac{\nu}{2\nu+d}} = n^{-\frac{\min(\beta,\nu)}{2\nu+d}}$.

Result 10 (Weyl's Law) and properties of Φ, we have

$$\|f\|_{\mathbb{H}_n}^2 \lesssim \|f\|_{H^{\nu+d/2}(\mathcal{M})}^2 \tag{52}$$

$$= \sum_{l\geq 0}(1+\lambda_l)^{\nu+d/2}\Phi^2\left(2^{-j}\sqrt{\lambda_l}\right)\left|\langle f_l, f_0\rangle_{L^2(\mathcal{M})}\right|^2 \tag{53}$$

$$\leq \sum_{l:\sqrt{\lambda_l}\leq 2^{j+1}}(1+\lambda_l)^{\nu+d/2-\beta}(1+\lambda_l)^{\beta}\left|\langle f_l, f_0\rangle_{L^2(\mathcal{M})}\right|^2 \tag{54}$$

$$\leq 2^{(j+2)(\nu+d/2-\beta)}\sum_{l:\sqrt{\lambda_l}\leq 2^{j+1}}(1+\lambda_l)^{\beta}\left|\langle f_l, f_0\rangle_{L^2(\mathcal{M})}\right|^2 \tag{55}$$

$$\leq 2^{(j+2)(\nu+d/2-\beta)}\sum_{l\geq 0}(1+\lambda_l)^{\beta}\left|\langle f_l, f_0\rangle_{L^2(\mathcal{M})}\right|^2 \tag{56}$$

$$= 2^{(j+2)(\nu+d/2-\beta)}\|f_0\|_{H^\beta(\mathcal{M})}^2 \tag{57}$$

$$\lesssim \varepsilon^{-\frac{1}{\beta}(\nu+d/2-\beta)}\|f_0\|_{H^\beta(\mathcal{M})}^2 \lesssim \varepsilon^{-\frac{2}{\beta}(\nu+d/2-\beta)}\|f_0\|_{H^\beta(\mathcal{M})}^2. \tag{58}$$

Our assumption $\nu \geq \beta$ implies that

$$\frac{2}{\beta}(\nu-\beta+d/2) \geq \frac{d}{\beta} \geq \frac{d}{\nu}. \tag{59}$$

Hence, we have $\varepsilon^{-d/\nu} \leq \varepsilon^{-\frac{2}{\beta}(\nu-\beta+d/2)}$ which gives us $\varphi_{f_0}(\varepsilon) \lesssim \varepsilon^{-\frac{2}{\beta}(\nu-\beta+d/2)}$. It is then easy to check that $\varepsilon_n = Mn^{-\frac{\beta}{2\nu+d}}$ satisfies $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$ for $M > 0$ large enough. □

From this, we deduce an upper bound on the error in *the empirical L^2 norm* $\|\cdot\|_n$, that is, on the Euclidean distance between the posterior Gaussian process f and the ground truth function f_0 evaluated at data locations x_i.

Lemma 22. *Let Π_n denote the prior in either Theorem 5, Theorem 6 or Theorem 8 with smoothness parameter $\nu > d/2$. Fix $f_0 \in H^\beta(\mathcal{M}) \cap B^\beta_{\infty,\infty}(\mathcal{M})$ with $\beta > 0$. Then*

$$\mathbb{E}_{f\sim\Pi_n(\cdot|\boldsymbol{x},\boldsymbol{y})}\|f-f_0\|_n^2 \leq \varepsilon_n^2 \tag{60}$$

for $\varepsilon_n \propto n^{-\frac{\min(\nu,\beta)}{2\nu+d}}$ with constant depending on f_0, ν but not on $\boldsymbol{x}$.

Proof. By Theorem 21 for ε_n a multiple of $n^{-\frac{\min(\beta,\nu)}{2\nu+d}}$, we have $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$. By virtue of this, the proof of Theorem 1 and Proposition 11 of van der Vaart and van Zanten [56] imply the result. Indeed, the proof of Theorem 1 relies solely on the fact that $\varphi_{f_0}(\varepsilon_n/2) \leq n\varepsilon_n^2$ and an application of van der Vaart and van Zanten [56], Proposition 11. We have $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2 \leq n(2\varepsilon_n)^2$ and hence the condition is satisfied with ε_n replaced by $2\varepsilon_n$. □

We now turn to the proofs of our main results, Theorems 5, 6 and 8, which for convenience we restate below. The idea of these proofs is to extend the result of Lemma 22 from input locations to the whole manifold $\mathcal{M}$ using an appropriate concentration inequality. To this end, the proof closely follows the one of Theorem 2 in van der Vaart and van Zanten [56], but relies on Lemma 22 proved above and on the concentration inequality we prove in Appendix D along with some important sample differentiablity properties of the prior processes.

Theorem 5. *Let f be a Riemannian Matérn Gaussian process prior of Definition 4 with smoothness parameter $\nu > d/2$ and let f_0 satisfy Assumption 3. Then there is a $C > 0$ such that*

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\,\mathbb{E}_{f\sim\Pi(\cdot|\boldsymbol{x},\boldsymbol{y})}\|f-f_0\|_{L^2(p_0)}^2 \leq Cn^{-\frac{2\min(\beta,\nu)}{2\nu+d}}. \tag{9}$$

Theorem 6. *Let f be a Riemannian Matérn Gaussian process prior on $\mathcal{M}$ with smoothness parameter $\nu > d/2$, modified to truncate the infinite sum to at least $J_n \geq cn^{\frac{d(\min(1,\nu/\beta))}{2\nu+d}}$ terms, and let f_0 satisfy Assumption 3. Then there is a $C > 0$ such that*

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\,\mathbb{E}_{f\sim\Pi(\cdot|\boldsymbol{x},\boldsymbol{y})}\|f-f_0\|_{L^2(p_0)}^2 \leq Cn^{-\frac{2\min(\beta,\nu)}{2\nu+d}}. \tag{10}$$

Theorem 8. *Let f be a mean-zero extrinsic Matérn Gaussian process prior with smoothness parameter $\nu > d/2$ on $\mathcal{M}$, and let f_0 satisfy Assumption 3. Then for some $C > 0$ we have*

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\,\mathbb{E}_{f\sim\Pi(\cdot|\boldsymbol{x},\boldsymbol{y})}\|f-f_0\|^2_{L^2(p_0)} \le Cn^{-\frac{2\min(\beta,\nu)}{2\nu+d}}. \tag{12}$$

Proof of Theorems 5, 6 and 8. To ease notation, we denote the expectations under the true data generating process by $\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} = \mathbb{E}^{(f_0)}_{\boldsymbol{x},\boldsymbol{y}}$ and $\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}} = \mathbb{E}^{(f_0)}_{\boldsymbol{y}|\boldsymbol{x}}$, omitting the superscript $(\cdot)^{f_0}$. Take $\varepsilon_n \propto n^{-\frac{\min(\beta,\nu)}{2\nu+d}}$ satisfying $\varphi_{f_0}(\varepsilon_n/2) \le n\varepsilon_n^2$, noting that such a rate exists by Theorem 21. Then, for each n, there exists an element $f_n \in \mathbb{H}_n$, where $\mathbb{H}_n$ is the RKHS corresponding to Π_n, satisfying

$$\|f_n\|^2_{\mathbb{H}_n} \le n\varepsilon_n^2 \qquad\qquad \|f_n - f_0\|_{L^\infty(\mathcal{M})} \le \varepsilon_n/2. \tag{61}$$

Write

$$\varepsilon_n^{-2}\,\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\,\mathbb{E}_{f\sim\Pi_n(\cdot|\boldsymbol{x},\boldsymbol{y})}\|f-f_0\|^2_{L^2(p_0)} \lesssim \varepsilon_n^{-2}\,\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\,\mathbb{E}_{f\sim\Pi_n(\cdot|\boldsymbol{x},\boldsymbol{y})}\|f_n-f_0\|^2_{L^2(p_0)} \tag{62}$$

$$+\,\varepsilon_n^{-2}\,\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\,\mathbb{E}_{f\sim\Pi_n(\cdot|\boldsymbol{x},\boldsymbol{y})}\|f-f_n\|^2_{L^2(p_0)} \tag{63}$$

$$\lesssim 1 + \varepsilon_n^{-2}\,\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\,\mathbb{E}_{f\sim\Pi_n(\cdot|\boldsymbol{x},\boldsymbol{y})}\|f-f_n\|^2_{L^2(p_0)}. \tag{64}$$

Thus, we can work with $f - f_n$ instead of $f - f_0$. Define $\mathcal{B}(r) = \left\{\|f-f_n\|_{L^2(p_0)} > \varepsilon_n r\right\}$. Then

$$\varepsilon_n^{-2}\,\mathbb{E}_{f\sim\Pi_n(\cdot|\boldsymbol{x},\boldsymbol{y})}\|f-f_n\|^2_{L^2(p_0)}. = \int_0^\infty 2r\,\mathbb{P}_{f\sim\Pi_n(\cdot|\boldsymbol{x},\boldsymbol{y})}(\mathcal{B}(r))\,\mathrm{d}r. \tag{65}$$

Fix a γ such that $d/2 < \gamma < \nu, \gamma \notin \mathbb{Z}_{>0}$ and $s > 0, \tau > 0$ and define

$$\mathcal{B}^{(\mathrm{I})}(r) = \{2\|f-f_n\|_n > \varepsilon_n r\} \tag{66}$$

$$\mathcal{B}^{(\mathrm{II})}(r) = \left\{\|f\|_{\mathcal{CH}^\gamma(\mathcal{M})} > \tau\sqrt{n}\varepsilon_n r^s\right\} \tag{67}$$

$$\mathcal{B}^{(\mathrm{III})}(r) = \mathcal{B}(r)\setminus\left(\mathcal{B}^{(\mathrm{I})}(r)\cup\mathcal{B}^{(\mathrm{II})}(r)\right). \tag{68}$$

Then $\mathcal{B}(r) \subseteq \mathcal{B}^{(\mathrm{I})}(r)\cup\mathcal{B}^{(\mathrm{II})}(r)\cup\mathcal{B}^{(\mathrm{III})}(r)$, and thus for an indexed family of events $\mathcal{A}_r$ to be chosen later, we have

$$\varepsilon_n^{-2}\,\mathbb{E}_{f\sim\Pi_n(\cdot|\boldsymbol{x},\boldsymbol{y})}\|f-f_n\|^2_{L^2(p_0)} \lesssim \int_0^\infty r\,\mathbb{P}_{f\sim\Pi_n(\cdot|\boldsymbol{x},\boldsymbol{y})}\left(\mathcal{B}^{(\mathrm{I})}(r)\right)\mathrm{d}r + \int_0^\infty r\mathbb{1}_{\mathcal{A}_r^c}\,\mathrm{d}r \tag{69}$$

$$+\int_0^\infty r\mathbb{1}_{\mathcal{A}_r}\,\mathbb{P}_{f\sim\Pi_n(\cdot|\boldsymbol{x},\boldsymbol{y})}\left(\mathcal{B}^{(\mathrm{II})}(r)\right)\mathrm{d}r + \int_0^\infty r\mathbb{1}_{\mathcal{A}_r}\,\mathbb{P}_{f\sim\Pi_n(\cdot|\boldsymbol{x},\boldsymbol{y})}\left(\mathcal{B}^{(\mathrm{III})}(r)\right)\mathrm{d}r. \tag{70}$$

For the first term, by Lemma 22 applied conditionally on $\boldsymbol{x}$, for which we got a bound on the integrated empirical L^2-norm uniformly on the design points, we have

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\int_0^\infty r\,\mathbb{P}_{f\sim\Pi_n(\cdot|\boldsymbol{x},\boldsymbol{y})}\left(\mathcal{B}^{(\mathrm{I})}(r)\right)\mathrm{d}r \lesssim \varepsilon_n^{-2}\,\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\,\mathbb{E}_{f\sim\Pi_n(\cdot|\boldsymbol{x},\boldsymbol{y})}\|f-f_0\|_n^2 \lesssim \varepsilon_n^{-2}\varepsilon_n^2 = 1. \tag{71}$$

Moreover, by Lemma 14 of van der Vaart and van Zanten [56] applied with r in the notation of the reference being equal to $\sqrt{n}\varepsilon_n r^s$, for each $r > 0$, the event

$$\mathcal{A}_r(\boldsymbol{x}) = \left\{\boldsymbol{u}\in\mathbb{R}^n : \int\frac{p^{(f)}_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{u})}{p^{(f_0)}_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{u})}\,\mathrm{d}\Pi_n(f) \ge e^{-n\varepsilon_n^2 r^{2s}}\,\mathbb{P}_{f\sim\Pi_n}\left(\|f-f_0\|_{L^\infty(\mathcal{M})} < \varepsilon_n r^s\right)\right\} \tag{72}$$

$$\supseteq \left\{\boldsymbol{u}\in\mathbb{R}^n : \int\frac{p^{(f)}_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{u})}{p^{(f_0)}_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{u})}\,\mathrm{d}\Pi_n(f) \ge e^{-n\varepsilon_n^2 r^{2s}}\,\mathbb{P}_{f\sim\Pi_n}(\|f-f_0\|_n < \varepsilon_n r^s)\right\} \tag{73}$$

is such that

$$p^{(f_0)}_{\boldsymbol{y}|\boldsymbol{x}}[\mathcal{A}_r^c(\boldsymbol{x})] \le e^{-n\varepsilon_n^2 r^{2s}/8}. \tag{74}$$

It should be noted that the $\sqrt{n}$ factor disappears because of the discrepancy between the empirical L^2 norm $\|\cdot\|_n$ and the Euclidean norm used in van der Vaart and van Zanten [56]. By Fubini's Theorem and since $n\varepsilon_n^2 \geq n^{\frac{d}{2\nu+d}} \geq 1$, the second term is bounded by

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \int_0^\infty r \mathbb{1}_{\mathcal{A}_r^c(\boldsymbol{x})} \,\mathrm{d}r = \int_0^\infty r\, \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left[\mathbb{1}_{\mathcal{A}_r^c(\boldsymbol{x})}\right]\right] \mathrm{d}r \tag{75}$$

$$\leq \int_0^\infty r e^{-n\varepsilon_n^2 r^{2s}/8} \,\mathrm{d}r \tag{76}$$

$$\leq \int_0^\infty r e^{-r^{2s}/8} \,\mathrm{d}r = C_s < \infty. \tag{77}$$

It remains to bound the last two terms. By Bayes' Rule, we have the equality

$$\mathbb{P}_{f\sim\Pi_n(\cdot|\boldsymbol{x},\boldsymbol{y})}\left(\|f\|_{\mathcal{CH}^\gamma(\mathcal{M})} > \tau\sqrt{n}\varepsilon_n r^s\right) = \frac{\int_{\|f\|_{\mathcal{CH}^\gamma(\mathcal{M})} > \tau\sqrt{n}\varepsilon_n r^s} p^{(f)}_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y}) \,\mathrm{d}\Pi_n(f)}{\int p^{(f)}_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y}) \,\mathrm{d}\Pi_n(f)} \tag{78}$$

$$= \frac{\int_{\|f\|_{\mathcal{CH}^\gamma(\mathcal{M})} > \tau\sqrt{n}\varepsilon_n r^s} \frac{p^{(f)}_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y})}{p^{(f_0)}_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y})} \,\mathrm{d}\Pi_n(f)}{\int \frac{p^{(f)}_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y})}{p^{(f_0)}_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y})} \,\mathrm{d}\Pi_n(f)} \tag{79}$$

therefore for $\boldsymbol{y} \in \mathcal{A}_r(\boldsymbol{x})$, we have

$$\mathbb{P}_{f\sim\Pi_n(\cdot|\boldsymbol{x},\boldsymbol{y})}\left(\|f\|_{\mathcal{CH}^\gamma(\mathcal{M})} > \tau\sqrt{n}\varepsilon_n r^s\right) \tag{80}$$

$$\leq \frac{e^{n\varepsilon_n^2 r^{2s}}}{\mathbb{P}_{f\sim\Pi_n}\left(\|f - f_0\|_{L^\infty(\mathcal{M})} < \varepsilon_n r^s\right)} \int_{\|f\|_{\mathcal{CH}^\gamma(\mathcal{M})} > \tau\sqrt{n}\varepsilon_n r^s} \frac{p^{(f)}_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y})}{p^{(f_0)}_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y})} \,\mathrm{d}\Pi_n(f). \tag{81}$$

Hence taking expectations, using Tonelli's Theorem, and $\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \frac{p^{(f)}_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y})}{p^{(f_0)}_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y})} = 1$ gives

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left[\mathbb{1}_{\mathcal{A}_r(\boldsymbol{x})}\, \mathbb{P}_{f\sim\Pi_n(\cdot|\boldsymbol{x},\boldsymbol{y})}\left(\|f\|_{\mathcal{CH}^\gamma(\mathcal{M})} > \tau\sqrt{n}\varepsilon_n r^s\right)\right] \tag{82}$$

$$\leq \frac{e^{n\varepsilon_n^2 r^{2s}}}{\mathbb{P}_{f\sim\Pi_n}\left(\|f - f_0\|_{L^\infty(\mathcal{M})} < \varepsilon_n r^s\right)} \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \int_{\|f\|_{\mathcal{CH}^\gamma(\mathcal{M})} > \tau\sqrt{n}\varepsilon_n r^s} \frac{p^{(f)}_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y})}{p^{(f_0)}_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y})} \,\mathrm{d}\Pi_n(f) \tag{83}$$

$$= \frac{e^{n\varepsilon_n^2 r^{2s}}}{\mathbb{P}_{f\sim\Pi_n}\left(\|f - f_0\|_{L^\infty(\mathcal{M})} < \varepsilon_n r^s\right)} \int_{\|f\|_{\mathcal{CH}^\gamma(\mathcal{M})} > \tau\sqrt{n}\varepsilon_n r^s} \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \frac{p^{(f)}_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y})}{p^{(f_0)}_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y})} \,\mathrm{d}\Pi_n(f) \tag{84}$$

$$= \frac{e^{n\varepsilon_n^2 r^{2s}}}{\mathbb{P}_{f\sim\Pi_n}\left(\|f - f_0\|_{L^\infty(\mathcal{M})} < \varepsilon_n r^s\right)} \mathbb{P}_{f\sim\Pi_n}\left(\|f\|_{\mathcal{CH}^\gamma(\mathcal{M})} > \tau\sqrt{n}\varepsilon_n r^s\right). \tag{85}$$

Therefore, the third term can be bounded as

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \int_0^\infty r \mathbb{1}_{\mathcal{A}_r}\, \mathbb{P}_{f\sim\Pi_n(\cdot|\boldsymbol{x},\boldsymbol{y})}\left(\mathcal{B}^{(\mathrm{II})}(r)\right) \mathrm{d}r \tag{86}$$

$$\leq \int_0^\infty r \frac{e^{n\varepsilon_n^2 r^{2s}}}{\mathbb{P}_{f\sim\Pi_n}\left(\|f - f_0\|_{L^\infty(\mathcal{M})} < \varepsilon_n r^s\right)} \mathbb{P}_{f\sim\Pi_n}\left(\|f\|_{\mathcal{CH}^\gamma(\mathcal{M})} > \tau\sqrt{n}\varepsilon_n r^s\right) \mathrm{d}r. \tag{87}$$

Now, using Lemma 31, for a possibly small constant $c > 0$ independent of n, we have

$$\mathbb{P}_{f\sim\Pi_n(\cdot|\boldsymbol{x},\boldsymbol{y})}\left(\|f\|_{\mathcal{CH}^\gamma(\mathcal{M})} > \tau\sqrt{n}\varepsilon_n r^s\right) \leq e^{-c\tau^2 n\varepsilon_n^2 r^{2s}}. \tag{88}$$

Moreover, by using the bound on the concentration function in Theorem 21 and Ghosal and van der Vaart [20], Proposition 11.19, we can assume that

$$\mathbb{P}_{f\sim\Pi_n}\left[\|f - f_0\|_{L^\infty(\mathcal{M})} < \varepsilon_n r^s\right] \geq e^{-c^{-1} n\varepsilon_n^2 r^{2s}}. \tag{89}$$

Therefore, the third term is bounded by

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \int_0^\infty r \mathbb{1}_{\mathcal{A}_r(\boldsymbol{x})} \mathbb{P}_{f\sim\Pi_n(\cdot|\boldsymbol{x},\boldsymbol{y})}\left(\mathcal{B}^{(\mathrm{II})}(r)\right) \mathrm{d}r \leq \int_0^\infty r e^{-n(c\tau^2-1)\varepsilon_n^2 r^{2s}} e^{c^{-1}n\varepsilon_n^2 r^{2s}} \mathrm{d}r \tag{90}$$

$$\leq \int_0^\infty r e^{-r^{2s}} \mathrm{d}r < \infty \tag{91}$$

if $\tau^2 c > 2 + c^{-1}$. It remains to bound the last term. We have, by the same arguments as above, that

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \int_0^\infty r \mathbb{1}_{\mathcal{A}_r(\boldsymbol{x})} \mathbb{P}_{f\sim\Pi_n(\cdot|\boldsymbol{x},\boldsymbol{y})}\left(\mathcal{B}^{(\mathrm{III})}(r)\right) \mathrm{d}r \tag{92}$$

$$= \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \int_0^\infty r \mathbb{1}_{\mathcal{A}_r(\boldsymbol{x})} \tag{93}$$

$$\times \mathbb{P}_{f\sim\Pi_n(\cdot|\boldsymbol{x},\boldsymbol{y})}\left(\|f\|_{\mathcal{CH}^\gamma(\mathcal{M})} \leq \tau\sqrt{n}\varepsilon_n r^s, 2\|f-f_n\|_n \leq \varepsilon_n r \leq \|f-f_n\|_{L^2(p_0)}\right) \mathrm{d}r \tag{94}$$

$$\leq \int_0^\infty r \frac{e^{n\varepsilon_n^2 r^{2s}}}{\mathbb{P}_{f\sim\Pi_n}\left(\|f-f_0\|_{L^\infty(\mathcal{M})} < \varepsilon_n r^s\right)} \tag{95}$$

$$\times \mathbb{E}_{\boldsymbol{x}} \mathbb{P}_{f\sim\Pi_n}\left(\|f\|_{\mathcal{CH}^\gamma(\mathcal{M})} \leq \tau\sqrt{n}\varepsilon_n r^s, 2\|f-f_n\|_n \leq \varepsilon_n r \leq \|f-f_n\|_{L^2(p_0)}\right) \mathrm{d}r \tag{96}$$

$$\leq \int_0^\infty r e^{(c+1)n\varepsilon_n^2 r^{2s}} \tag{97}$$

$$\times \int_{\|f\|_{\mathcal{CH}^\gamma(\mathcal{M})} \leq \tau\sqrt{n}\varepsilon_n r^s, \varepsilon_n r \leq \|f-f_n\|_{L^2(p_0)}} \mathbb{E}_{\boldsymbol{x}} \mathbb{1}_{\|f-f_n\|_{L^2(p_0)} \geq 2\|f-f_n\|_n} \mathrm{d}\Pi_n(f) \mathrm{d}r. \tag{98}$$

As the squared empirical L^2-norm is a Monte Carlo approximation of the true L^2-norm, the probability in the integrand can be controlled via a concentration inequality. As in van der Vaart and van Zanten [56], we use Bernstein's inequality [58, Lemma 2.2.9]. For a collection $Y_1, \ldots Y_n$ of random variables such that $\mathbb{E}\, Y_i = 0$ and $Y_i \in [-M, M]$ almost surely for some constant $M > 0$, this inequality asserts

$$\mathbb{P}(|Y_1 + \ldots + Y_n| > x) \leq 2\exp\left(-\frac{1}{2}\frac{x^2}{v + Mx/3}\right) \tag{99}$$

where $v \geq \mathrm{Var}(Y_1 + \ldots + Y_n)$. We put $Y_i = \frac{1}{n}(f(x_i) - f_n(x_i))^2 - \frac{1}{n}\|f - f_n\|_{L^2(p_0)}^2$ where x_i are IID with $x_i \sim p_0$. It is easy to check that $\mathbb{E}\, Y_i = 0$ and $Y_i \in [-M, M]$ for $M = \frac{1}{n}\|f - f_n\|_{L^\infty(\mathcal{M})}$. Furthermore, $v = \frac{1}{n}\|f - f_n\|_{L^\infty(\mathcal{M})}^2 \|f - f_n\|_{L^2(\mathcal{M})}^2$ upper-bounds the variance of the respective sum, since

$$\mathrm{Var}(Y_1 + \ldots + Y_n) = \frac{1}{n}\mathrm{Var}_{x\sim p_0}(f(x) - f_n(x))^2 \leq \frac{1}{n}\mathbb{E}_{x\sim p_0}(f(x) - f_n(x))^4 \tag{100}$$

$$\leq \frac{\|f - f_n\|_{L^\infty(\mathcal{M})}^2}{n}\mathbb{E}_{x\sim p_0}(f(x) - f_n(x))^2 = v. \tag{101}$$

Using Bernstein's inequality with Y_i, M, v as above and with $x = \frac{3}{4}\|f - f_n\|_{L^2(p_0)}^2$, we have

$$\mathbb{E}_{\boldsymbol{x}} \mathbb{1}_{\|f-f_n\|_{L^2(p_0)} \geq 2\|f-f_n\|_n} = \mathbb{E}_{\boldsymbol{x}} \mathbb{1}_{\|f-f_n\|_{L^2(p_0)}^2 \geq 4\|f-f_n\|_n^2} \tag{102}$$

$$= \mathbb{E}_{\boldsymbol{x}} \mathbb{1}_{\|f-f_n\|_n^2 - \|f-f_n\|_{L^2(p_0)}^2 \leq -\frac{3}{4}\|f-f_n\|_{L^2(p_0)}^2} \tag{103}$$

$$\lesssim \exp\left(-\frac{1}{2}\frac{x^2}{v + Mx/3}\right) \tag{104}$$

$$= \exp\left(-\frac{\frac{9}{32}\|f-f_n\|_{L^2(p_0)}^4}{v + \frac{1}{n}\|f-f_n\|_{L^\infty(\mathcal{M})}^2 \frac{3}{4}\|f-f_n\|_{L^2(p_0)}^2/3}\right) \tag{105}$$

$$= \exp\left(-\frac{9n}{32}\frac{4}{5}\frac{\|f-f_n\|_{L^2(p_0)}^2}{\|f-f_n\|_{L^\infty(\mathcal{M})}^2}\right) \tag{106}$$

$$= \exp\left(-\frac{9n}{40}\frac{\|f-f_n\|_{L^2(p_0)}^2}{\|f-f_n\|_{L^\infty(\mathcal{M})}^2}\right). \tag{107}$$

Moreover, by Lemma 19, since $\gamma \notin \mathbb{Z}_{>0}$, we have

$$\|f - f_n\|_{L^\infty(\mathcal{M})} \lesssim \|f - f_n\|_{\mathcal{CH}^\gamma(\mathcal{M})}^{\frac{d}{2\gamma+d}} \|f - f_n\|_{L^2(p_0)}^{\frac{2\gamma}{2\gamma+d}}. \tag{108}$$

Using the Sobolev Embedding Theorem, given in De Vito et al. [16], Theorem 4, $\|f - f_n\|_{\mathcal{CH}^\gamma(\mathcal{M})} \lesssim \|f_n\|_{\mathbb{H}} + \|f\|_{\mathcal{CH}^\gamma(\mathcal{M})} \lesssim \tau\sqrt{n}\varepsilon_n r^s$ whenever $\|f\|_{\mathcal{CH}^\gamma(\mathcal{M})} \leq \tau\sqrt{n}\varepsilon_n r^s$. Therefore, for a constant $c > 0$, when $\|f\|_{\mathcal{CH}^\gamma(\mathcal{M})} \leq \tau\sqrt{n}\varepsilon_n r^s$ and $\varepsilon_n r \leq \|f - f_n\|_{L^2(p_0)}$, we have that

$$\mathbb{E}_{\boldsymbol{x}} \mathbb{1}_{\|f-f_n\|_{L^2(\mathcal{M})} \geq 2\|f-f_n\|_n} \lesssim \exp\left(-cn \frac{\|f - f_n\|_{L^2(p_0)}^2}{\|f - f_n\|_{\mathcal{CH}^\gamma(\mathcal{M})}^{\frac{2d}{2\gamma+d}} \|f - f_n\|_{L^2(p_0)}^{\frac{4\gamma}{2\gamma+d}}}\right) \tag{109}$$

$$\leq e^{-c\tau^{-\frac{2d}{2\gamma+d}} n^{\frac{2\gamma}{2\gamma+d}} r^{\frac{2d}{2\gamma+d}(1-s)}}. \tag{110}$$

Hence, we can bound the last term as

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \int_0^\infty r \mathbb{1}_{\mathcal{A}_r} \mathbb{P}_{f\sim\Pi_n(\cdot|\boldsymbol{x},\boldsymbol{y})}\left(\mathcal{B}^{(\mathrm{III})}(r)\right) \mathrm{d}r \tag{111}$$

$$\lesssim \int_0^\infty r e^{(c+1)n\varepsilon_n^2 r^{2s}} e^{-c\tau^{-\frac{2d}{2\gamma+d}} n^{\frac{2\gamma}{2\gamma+d}} r^{\frac{2d}{2\gamma+d}(1-s)}} \mathrm{d}r. \tag{112}$$

We have $n^{\frac{2\gamma}{2\gamma+d}} = n\left(n^{-\frac{d/2}{2\gamma+d}}\right)^2$. Since $\varepsilon_n \lesssim n^{-\frac{\min(\nu,\beta)}{2\nu+d}}$ and $\min(\nu,\beta) > d/2$, we have $n\varepsilon_n^2 \lesssim n^{\frac{2\gamma}{2\gamma+d}}$ for some $\gamma \in (d/2, \nu)$. Moreover, for this choice of γ and s small enough we have $\frac{2d}{2\gamma+d}(1 - s) \geq 2s$, which proves that for some constants $C, C', C'' > 0$ the fourth term is bounded by

$$C \int_0^\infty r e^{-C' r^{C''}} \mathrm{d}r < \infty. \tag{113}$$

This concludes the proof. □

Remark 23. *Following the proof, it is easy to see that $\|f - f_0\|_{L^2(p_0)}^2$ on the left-hand side of Equations (9), (10) and (12) can be replaced with $\|f - f_0\|_{L^2(p_0)}^q$ for any $q > 1$, changing the exponent $-\frac{2\min(\beta,\nu)}{2\nu+d}$ on the right-hand side to $-\frac{q\min(\beta,\nu)}{2\nu+d}$.*

C Characterizing Reproducing Kernel Hilbert Spaces of Matérn Kernels

We start by describing the reproducing kernel Hilbert spaces (RKHSs) of the (truncated) intrinsic Matérn processes, proving that they coincide with (certain subspaces of) the manifold Sobolev spaces. We follow the ideas of Borovitskiy et al. [11], where the same was shown somewhat implicitly. We consider the more-involved case of the extrinsic Matérn processes immediately after.

The next lemma describes the RKHS of the intrinsic Matérn processes, including truncated variants. This result is easy to obtain since we have defined them in terms of the Karhunen–Loève expansions.

Lemma 24. *Let $\mathbb{H}_J$ be the RKHS of the intrinsic Matérn Gaussian process with smoothness parameter ν truncated at the level $J \in \mathbb{Z}_{>0} \cup \{\infty\}$, and let $\{f_j\}_{j=0}^\infty$ be the orthonormal basis of Laplace–Beltrami eigenfunctions. The space $\mathbb{H}_J$ is norm-equivalent—with constants depending only on ν, κ and σ_f^2—to the set of functions $f = \sum_{j=1}^J b_j f_j$ with $b_j \in \mathbb{R}$, equipped with the inner product*

$$\left\langle \sum_{j=1}^J b_j f_j, \sum_{j=1}^J b'_j f_j \right\rangle_{\mathbb{H}_J} = \sum_{j=1}^J (1 + \lambda_j)^{\nu+d/2} b_j b'_j. \tag{114}$$

In particular, $\mathbb{H}_J \subset H^{\nu+d/2}(\mathcal{M})$ for all J, and for every $h \in \mathbb{H}_J$ we have $\|h\|_{\mathbb{H}_J} = \|h\|_{H^{\nu+d/2}(\mathcal{M})}$.

Proof. By direct computation, the covariance k of the (truncated) intrinsic Gaussian process is

$$k(x, x') = \frac{\sigma_f^2}{C_{\nu,\kappa}} \sum_{j=1}^J \left(\frac{2\nu}{\kappa^2} + \lambda_j\right)^{-(\nu+d/2)} f_j(x) f_j(x'). \tag{115}$$

Hence, the covariance operator $K : L^2(\mathcal{M}) \to L^2(\mathcal{M})$ defined by

$$(Kf)(x) = \int_{\mathcal{M}} k(x, x') f(x') \,\mathrm{d}x' \tag{116}$$

is diagonal in the basis $\{f_j\}_{j=1}^J$, with $K f_j = \frac{\sigma_f^2}{C_{\nu,\kappa}} \left(\frac{2\nu}{\kappa^2} + \lambda_j\right)^{-(\nu+d/2)} f_j$. Then, Kanagawa et al. [28], Theorem 4.2 implies that $\mathbb{H}_J$ consists of functions of form $f = \sum_{j=1}^J a_j f_j$ satisfying

$$\|f\|_{\mathbb{H}_J}^2 = \frac{\sigma_f^2}{C_{\nu,\kappa}} \sum_{j=1}^J \left(\frac{2\nu}{\kappa^2} + \lambda_j\right)^{\nu+d/2} |a_j|^2 < \infty. \tag{117}$$

Using the elementary inequality $\min\left(\frac{2\nu}{\kappa^2}, 1\right) \leq \frac{\frac{2\nu}{\kappa^2}+\lambda}{1+\lambda} \leq \max\left(\frac{2\nu}{\kappa^2}, 1\right)$, we find that this space is norm-equivalent to the space $H_J^{\nu+d/2}(\mathcal{M})$ of functions $f = \sum_{j=1}^J a_j f_j$ satisfying

$$\|f\|_{H_J^{\nu+d/2}(\mathcal{M})}^2 = \sum_{j=1}^J (1 + \lambda_j)^{\nu+d/2} |a_j|^2 < \infty \tag{118}$$

where the comparison constants $\sqrt{\frac{\sigma_f^2}{C_{\nu,\kappa}} \min\left(1, \frac{2\nu}{\kappa^2}\right)}$ and $\sqrt{\frac{\sigma_f^2}{C_{\nu,\kappa}} \max\left(1, \frac{2\nu}{\kappa^2}\right)}$ only depend on the parameters ν, κ, σ_f^2. $\square$

The next two theorems will be useful to characterize the RKHS of the extrinsic Matérn process on $\mathcal{M}$. We start by a lemma relating the RKHS of the restriction of a Gaussian process to the original one.

Lemma 25. *Assume that k is a kernel on $\mathbb{R}^d$, $f \sim \mathrm{GP}(0, k)$ with almost surely continuous sample paths and $\widetilde{\mathbb{H}}$ is the RKHS of k. If $\mathcal{M} \subseteq \mathbb{R}^d$ is a submanifold, then the RKHS $\mathbb{H}$ corresponding to the restricted process $f_{|\mathcal{M}}$ is the set of all restrictions $g_{|\mathcal{M}}$ of functions $g \in \widetilde{\mathbb{H}}$ equipped with the norm*

$$\|h\|_{\mathbb{H}} = \inf_{g \in \widetilde{\mathbb{H}},\ g_{|\mathcal{M}} = h} \|g\|_{\widetilde{\mathbb{H}}}. \tag{119}$$

Moreover there always exists an element $g \in \widetilde{\mathbb{H}}$ such that $g_{|\mathcal{M}} = f$ and $\|g\|_{\widetilde{\mathbb{H}}} = \|f\|_{\mathbb{H}}$.

Proof. Lemma 5.1 in Yang and Dunson [63]. $\square$

The last result will be used to characterize the RKHS of the extrinsic Matérn Gaussian processes using trace and extension operators. The second ingredient for this is the following.

Theorem 26. *If $s > \frac{D-d}{2}$ then the restriction operator extends to a bounded linear map $\mathrm{Tr} : H^s(\mathbb{R}^D) \to H^{s-\frac{D-d}{2}}(\mathcal{M})$. Moreover, for every $u > 0$ there exists a bounded right inverse $\mathrm{Ex} : H^u(\mathcal{M}) \to H^{u+\frac{D-d}{2}}(\mathbb{R}^D)$ such that $\mathrm{Tr} \circ \mathrm{Ex} = \mathrm{id}_{H^u(\mathcal{M})}$ where Tr corresponds to $s = u + \frac{D-d}{2}$.*

Proof. Theorem 4.10 in Große and Schneider [24]. $\square$

The last two results allow us to characterize the RKHS of the extrinsic Matérn process on $\mathcal{M}$.

Proposition 27. *The RKHS $\mathbb{H}$ of a restricted extrinsic Matérn process f with smoothness parameter ν on $\mathcal{M}$ is norm-equivalent to the Sobolev space $H^{\nu+d/2}(\mathcal{M})$.*

Proof. Using Lemma 25, the RKHS $\mathbb{H}$ can be characterized as the set of functions $f : \mathcal{M} \to \mathbb{R}$ that are the restrictions of some $g \in \widetilde{\mathbb{H}}$, where $\widetilde{\mathbb{H}}$ is the RKHS of the ambient Matérn process $\tilde{f}$, with

$$\|f\|_{\mathbb{H}} = \inf_{g \in \widetilde{\mathbb{H}},\ g_{|\mathcal{M}} = f} \|g\|_{\widetilde{\mathbb{H}}}. \tag{120}$$

Since $\widetilde{\mathbb{H}}$ is norm-equivalent to the Sobolev space[7] $H^{\nu+D/2}(\mathbb{R}^D)$—see for instance Kanagawa et al. [28]—by Theorem 26, for every $f \in \mathbb{H}$ we have

$$\|f\|_{\mathbb{H}} \lesssim \|\mathrm{Ex}(f)\|_{H^{\nu+D/2}(\mathbb{R}^D)} \lesssim \|f\|_{H^{\nu+D/2-\frac{D-d}{2}}(\mathcal{M})} = \|f\|_{H^{\nu+d/2}(\mathcal{M})}. \tag{121}$$

Similarly, for every $g \in \widetilde{\mathbb{H}}$ with $g_{|\mathcal{M}} = f$, we have

$$\|f\|_{H^{\nu+d/2}(\mathcal{M})} = \left\|g_{|\mathcal{M}}\right\|_{H^{\nu+d/2}(\mathcal{M})} \lesssim \|g\|_{H^{\nu+D/2}(\mathbb{R}^D)} \lesssim \|g\|_{\widetilde{\mathbb{H}}}. \tag{122}$$

Hence, taking the infimum, we obtain

$$\|f\|_{H^{\nu+d/2}(\mathcal{M})} \lesssim \inf_{g\in\widetilde{\mathbb{H}},\ g_{|\mathcal{M}}=f} \|g\|_{\widetilde{\mathbb{H}}} = \|f\|_{\mathbb{H}}. \tag{123}$$

The claim follows. □

D Concentration and Sample Path Regularity

In this section, we prove that intrinsic, truncated intrinsic and extrinsic Matérn processes are sub-Gaussian, uniformly with respect to the truncation parameter in the case of the truncated intrinsic Matérn process, and live in Hölder spaces with appropriate exponents. On our way to proving this, we characterize sample path regularity of (truncated) intrinsic Matérn processes, which is of independent interest. A simple way to do this would be to build on the results of Appendix C, using Driscoll's Theorem—given in Kanagawa et al. [28], Theorem 4.9—and the Sobolev Embedding Theorem—De Vito et al. [16], Theorem 4—but that would only give us that the sample paths are almost surely in $\mathcal{CH}^\gamma(\mathcal{M})$ for every $0 < \gamma < \nu - d/2, \gamma \notin \mathbb{N}$, whereas here we improve the range of index to $\gamma < \nu$.

Kolmogorov's continuity criterion is a standard tool to show that a given stochastic process has a Hölder continuous version: we re-prove it here because we will need a form of the result which gives explicit control of the Hölder norms, which is not usually included in the respective statement.

Lemma 28 (Kolmogorov's continuity criterion)**.** *If $g \sim \Pi$ is a zero-mean Gaussian process on $[0,1]^d$ with*

$$\mathbb{E}\,|g(x) - g(y)|^2 \leq C\|x-y\|^{2\rho}_{\mathbb{R}^d} \tag{124}$$

for some $0 < \rho \leq 1$ and $C > 0$, then there exists a version of g with samples paths in $\mathcal{CH}^\alpha\left([0,1]^d\right)$ for every $0 < \alpha < \rho$. Moreover for every $\alpha < \rho$ this version satisfies $\mathbb{E}\,\|g\|^2_{\mathcal{CH}^\alpha([0,1]^d)} \leq C'$ where $C' < +\infty$ depends only on C, ρ, α and d.

Proof. Take $x, y \in [0,1]^d$, $M > 0$ and $q \in \mathbb{N}$. Since the random variable $g(x) - g(y)$ is Gaussian we have

$$\mathbb{E}\,|g(x)-g(y)|^{2q} = \frac{(2q)!}{2^q q!}\left(\mathbb{E}\,|g(x)-g(y)|^2\right)^q \leq C_q\|x-y\|^{2\rho q}_{\mathbb{R}^d} \tag{125}$$

where we have defined $C_q = C^q\frac{(2q)!}{2^q q!}$. The reason for considering the the $2q$th power will become clear later in the proof. By Markov's inequality, for every $x, y \in [0,1]^d$ we have

$$\mathbb{P}(|g(x)-g(y)| > u) \leq C_q u^{-2q}\|x-y\|^{2q\rho}_{\mathbb{R}^d}. \tag{126}$$

Now, take $X = \cup_{k\geq 0} X_k, X_k = 2^{-k}\mathbb{Z}^d \cap [0,1]^d$. Then, the previous inequality applied to any adjacent $x, y \in X_k$, where we see X_k as a graph where two vertices are connected if they differ by 2^{-k}, and $u = M2^{-k\alpha}$, implies

$$\mathbb{P}\left(|g(x)-g(y)| > M2^{-k\alpha}\right) \leq C_q M^{-2q}2^{-2kq(\rho-\alpha)}. \tag{127}$$

[7]Note that this norm-equivalence is the only property of the Gaussian process we use in the proofs. Any other Gaussian process satisfying this, including ones different from the Matérn processes of Borovitskiy et al. [11], would also work. This is of potential interest since other Euclidean kernels, such as Wendland kernels [61], are known to possess RKHSs which are norm-equivalent to those of the Matérn kernel.

Denote $K = \sup_{x,y\in X \text{adjacent}} \frac{|g(x)-g(y)|}{\|x-y\|^{\alpha}}$. Summing over $k \geq 1$ and adjacent points in X—and there are at most $C2^{k(d+1)}$ of them where $C > 0$ is an absolute constant—gives us for $q > \frac{d+1}{2(\rho-\alpha)}$ that

$$\mathbb{P}(K > M) \leq \sum_{k\geq 0} \sum_{x,y\in X \text{adjacent}} \mathbb{P}\big(|g(x) - g(y)| > M2^{-k\alpha}\big) \tag{128}$$

$$\leq C \sum_{k\geq 0} 2^{k(d+1)} C_q M^{-2q} 2^{-2kq(\rho-\alpha)} = \frac{CC_q}{1 - 2^{2q(\rho-\alpha)-d-1}} M^{-2q}. \tag{129}$$

In particular, for all $q > \max\Big(1, \frac{d+1}{2(\rho-\alpha)}\Big)$, we have

$$\mathbb{E}(K^2) = 2\int_0^\infty M\,\mathbb{P}(K > M)\,\mathrm{d}M = 2\int_0^1 M\,\mathbb{P}(K > M)\,\mathrm{d}M + 2\int_1^\infty M\,\mathbb{P}(k > M)\,\mathrm{d}M \tag{130}$$

$$\leq 2 + 2\int_1^\infty M\,\mathbb{P}(K > M)\,\mathrm{d}M \leq C_{C,\alpha,\rho,d} \tag{131}$$

for some constant $C_{C,\alpha,\rho,d} < +\infty$. This means that K is finite almost surely. Since X is dense in $[0,1]^d$ and g is almost surely uniformly continuous on X by the classical version of the Kolmogorov's continuity criterion—see for instance Lototsky and Rozovsky [34]—g admits a unique continuous extension to $[0,1]^d$ on a probability one event $\mathcal{A}$. Let us define

$$\overline{g}(x) = \begin{cases} \lim\limits_{y\to x, y\in X} g(y) & \text{on } \mathcal{A}, \\ 0 & \text{otherwise,} \end{cases} \tag{132}$$

for $x \in [0,1]^d$. For any $x, y \in [0,1]^d$ and $x_n \to x$, $y_n \to y$, $x_n, y_n \in X$ we have on $\mathcal{A}$

$$|\overline{g}(x) - \overline{g}(y)| \leq \liminf_{n\to\infty} (|\overline{g}(x) - \overline{g}(x_n)| + |\overline{g}(x_n) - \overline{g}(y_n)| + |\overline{g}(y_n) - \overline{g}(y)|) \tag{133}$$

$$\leq \liminf_{n\to\infty} (|\overline{g}(x) - \overline{g}(x_n)| + K\|x_n - y_n\|^{\alpha}_{\mathbb{R}^d} + |\overline{g}(y_n) - \overline{g}(y)|) \tag{134}$$

$$= K\|x - y\|^{\alpha}_{\mathbb{R}^d}. \tag{135}$$

On the complement of $\mathcal{A}$, we have $\overline{g} = 0$. Hence, $\overline{g}$ is α-Hölder continuous on $[0,1]^d$ with the same (random) constant K, since

$$\mathbb{E}\,\|\overline{g}\|^2_{\mathcal{CH}^{\alpha}([0,1]^d)} = \mathbb{E}\left(\sup_{x,y\in X} \frac{|\overline{g}(x) - \overline{g}(y)|}{\|x-y\|^{\alpha}}\right)^2 \leq \mathbb{E}\,K^2 \leq C_{C,\alpha,\rho,d} < \infty. \tag{136}$$

Since $\overline{g}$ is a version of g, the claim follows. □

The next lemma applies our version of Kolmogorov's criterion, Lemma 28, to the intrinsic Matérn processes on $\mathcal{M}$ by considering charts. This allows us to show that the sample paths are almost surely in $\mathcal{CH}^{\gamma}(\mathcal{M})$ for every $\gamma < \nu$, which is both used in our arguments and also of independent interest. For the claims in Appendix B, we need to ensure that this property holds somewhat uniformly with respect to the truncation parameter, which is why we tracked the constants in our proof of Kolmogorov's criterion. As we will see, the main difficulty in the proof of the next result will be to tackle the case of regularity strictly larger than 1.

Lemma 29. *Let $f \sim \Pi_n$ be an intrinsic Matérn process with smoothness parameter $\nu > 0$ truncated at $J_n \in \mathbb{N} \cup \{\infty\}$. Then for every $\gamma < \nu$ we have*

$$\sup_n \mathbb{E}_{f\sim\Pi_n}\Big[\|f\|^2_{\mathcal{CH}^{\gamma}(\mathcal{M})}\Big] < \infty. \tag{137}$$

Proof. We start with the case $\nu \leq 1$. Take $1 \leq l \leq L$ and define $h_l = (\chi_l f) \circ \phi_l^{-1}$. Then h_l is a Gaussian process with covariance kernel $\widetilde{K}_l$ given by

$$\widetilde{K}_l(x,y) = (\chi_l \circ \phi_l^{-1})(x) K(\phi_l^{-1}(x), \phi_l^{-1}(y))(\chi_l \circ \phi_l^{-1})(y) \tag{138}$$

where $x, y \in \mathcal{V}_l$ and $K(x, y) = \operatorname{Cov}(f(x), f(y))$ is the covariance kernel of f. Let $\widetilde{\mathbb{H}}_l$ be the RKHS induced by $\widetilde{K}_l$. We seek to apply Lemma 28 to h_l. For all $x, y \in \mathcal{V}_l$, where we recall that we can assume that $\mathcal{V}_l = (0,1)^d$,, we have

$$\begin{aligned}
\mathbb{E}_{f\sim\Pi_n}|h_l(x) - h_l(y)|^2 &= \widetilde{K}_l(x,x) + \widetilde{K}_l(y,y) - 2\widetilde{K}_l(x,y) && (139)\\
&= \left\|\widetilde{K}_l(x,\cdot) - \widetilde{K}_l(y,\cdot)\right\|^2_{\widetilde{\mathbb{H}}_l} && (140)\\
&= \sup_{\|\varphi\|_{\widetilde{\mathbb{H}}_l}=1}\left|\left\langle \widetilde{K}_l(x,\cdot) - \widetilde{K}_l(y,\cdot), \varphi\right\rangle_{\widetilde{\mathbb{H}}_l}\right|^2 && (141)\\
&= \sup_{\|\varphi\|_{\widetilde{\mathbb{H}}_l}=1}|\varphi(x) - \varphi(y)|^2 && (142)\\
&\le \sup_{\|\varphi\|_{\widetilde{\mathbb{H}}_l}=1}\|\varphi\|^2_{\mathcal{CH}^\nu(\mathcal{V}_l)}\|x-y\|^{2\nu}_{\mathbb{R}^d} && (143)
\end{aligned}$$

where $\|\varphi\|^2_{\mathcal{CH}^\nu(\mathcal{V}_l)}$ can potentially be infinite or unbounded: we will show this is not the case. To do this it suffices to show that we have a continuous embedding $\widetilde{\mathbb{H}}_l \hookrightarrow \mathcal{CH}^\nu(\mathcal{V}_l)$, that is $\|\cdot\|_{\mathcal{CH}^\nu(\mathcal{V}_l)} \lesssim \|\cdot\|_{\widetilde{\mathbb{H}}_l}$. The RKHS $\widetilde{\mathbb{H}}_l$ is by definition the completion of

$$\left\{\sum_{i=1}^p \alpha_i \widetilde{K}_l(x_i,\cdot) : p \ge 1, \alpha_i \in \mathbb{R}, x_i \in \mathcal{V}_l\right\} \tag{144}$$

$$= \left\{\sum_{i=1}^p \alpha_i \big(\chi_l \circ \phi_l^{-1}\big)(x_i)\big(\chi_l \circ \phi_l^{-1}\big)(\cdot)K\big(\phi_l^{-1}(x_i), \phi_l^{-1}(\cdot)\big) : p \ge 1, \alpha_i \in \mathbb{R}, x_i \in \mathcal{V}_l\right\} \tag{145}$$

with respect to the topology induced by the RKHS norm

$$\left\|\sum_{i=1}^p \alpha_i \widetilde{K}_l(x_i,\cdot)\right\|^2_{\widetilde{\mathbb{H}}_l} = \sum_{i,j=1}^p \alpha_i\alpha_j \big(\chi_l \circ \phi_l^{-1}\big)(x_i)\big(\chi_l \circ \phi_l^{-1}\big)(x_j)K\big(\phi_l^{-1}(x_i), \phi_l^{-1}(x_j)\big). \tag{146}$$

Denote the RKHS of K by $\mathbb{H}$. By Theorem 16, and by the equality $\|\cdot\|_{\mathbb{H}} = \|\cdot\|_{H^{\nu+d/2}(\mathcal{M})}$ on $\mathbb{H}$ which follows by Lemma 24, we have

$$\begin{aligned}
&\left\|\sum_{i=1}^p \alpha_i \widetilde{K}_l(x_i,\cdot)\right\|^2_{H^{\nu+d/2}(\mathbb{R}^d)} && (147)\\
&\quad= \left\|\sum_{i=1}^p \alpha_i \big(\chi_l \circ \phi_l^{-1}\big)(x_i)\big(\chi_l \circ \phi_l^{-1}\big)(\cdot)K(\phi_l^{-1}(x_i), \phi_l^{-1}(\cdot))\right\|^2_{H^{\nu+d/2}(\mathbb{R}^d)} && (148)\\
&\quad\lesssim \left\|\sum_{i=1}^p \alpha_i \big(\chi_l \circ \phi_l^{-1}\big)(x_i)K(\phi_l^{-1}(x_i), \cdot)\right\|^2_{H^{\nu+d/2}(\mathcal{M})} && (149)\\
&\quad= \left\|\sum_{i=1}^p \alpha_i \big(\chi_l \circ \phi_l^{-1}\big)(x_i)K(\phi_l^{-1}(x_i), \cdot)\right\|^2_{\mathbb{H}} && (150)\\
&\quad= \sum_{i,j=1}^p \alpha_i\alpha_j \big(\chi_l \circ \phi_l^{-1}\big)(x_i)\big(\chi_l \circ \phi_l^{-1}\big)(x_j)K\big(\phi_l^{-1}(x_i), \phi_l^{-1}(x_j)\big) && (151)\\
&\quad= \left\|\sum_{i=1}^p \alpha_i \widetilde{K}_l(x_i,\cdot)\right\|^2_{\widetilde{\mathbb{H}}_l}. && (152)
\end{aligned}$$

Therefore, we have a continuous embedding $\widetilde{\mathbb{H}}_l \hookrightarrow H^{\nu+d/2}(\mathbb{R}^d)$ with $\|\cdot\|_{H^{\nu+d/2}(\mathbb{R}^d)} \lesssim \|\cdot\|_{\widetilde{\mathbb{H}}_l}$ on $\widetilde{\mathbb{H}}_l$. By the Sobolev Embedding Theorem in $\mathbb{R}^d$—see for instance Triebel [50], Section 2.7.1, Remark 2—we have $B_{2,2}^{\nu+d/2}(\mathbb{R}^d) = H^{\nu+d/2}(\mathbb{R}^d) \hookrightarrow \mathcal{CH}^{\nu}(\mathbb{R}^d)$, which implies $\widetilde{\mathbb{H}}_l \hookrightarrow \mathcal{CH}^{\nu}(\mathbb{R}^d)$ by composition. Thus, there exists a constant $C = C_\nu$ such that

$$\mathbb{E}_{f\sim\Pi_n} |h_l(x) - h_l(y)|^2 \leq C\|x-y\|_{\mathbb{R}^d}^{2\nu} \tag{153}$$

for $x, y \in \mathcal{V}_l$. Hence, by applying Lemma 28, there exists a version $\tilde{h}_l$ of h_l with almost surely α-Hölder continuous sample paths for every $\alpha < \nu$. Now consider $\tilde{f} = \sum_{l=1}^{L} \tilde{h}_l \circ \phi_l$. Then $\tilde{f}$ is a version of f. We proceed to bound $\mathbb{E}_{f\sim\Pi_n}\left[\|\tilde{f}\|_{\mathcal{CH}^\alpha(\mathcal{M})}^2\right]$. For any $1 \leq l, r \leq L$ write

$$\left|\tilde{h}_r(\phi_r(\phi_l^{-1}(x))) - \tilde{h}_r(\phi_r(\phi_l^{-1}(y)))\right| \leq K\left|\phi_r(\phi_l^{-1}(x)) - \phi_r(\phi_l^{-1}(y))\right|^\alpha \leq CK|x-y|^\alpha \tag{154}$$

where C is the Lipshitz constant of $\phi_r \circ \phi_l^{-1}$ which is well defined and finite because this composition is a diffeomorphism and K is a random constant with $\mathbb{E}_{f\sim\Pi_n} K^2 \leq C_{\alpha,\nu,d}$. Hence

$$\begin{aligned}
\mathbb{E}_{f\sim\Pi_n}\left\|\tilde{f}\right\|_{\mathcal{CH}^\alpha(\mathcal{M})}^2 &= \mathbb{E}_{f\sim\Pi_n} \sum_{l=1}^{L} \left\|\left(\chi_l \tilde{f}\right)\circ\phi_l^{-1}\right\|_{\mathcal{CH}^\alpha(\mathbb{R}^d)}^2 &(155)\\
&= \mathbb{E}_{f\sim\Pi_n} \sum_{l=1}^{L} \left\|\left(\chi_l \sum_{r=1}^{L} \tilde{h}_r\circ\phi_r\right)\circ\phi_l^{-1}\right\|_{\mathcal{CH}^\alpha(\mathbb{R}^d)}^2 &(156)\\
&\lesssim \sum_{l=1}^{L}\sum_{r=1}^{L} \mathbb{E}_{f\sim\Pi_n}\left\|\tilde{h}_r\circ\phi_r\circ\phi_l^{-1}\right\|_{\mathcal{CH}^\alpha(\mathbb{R}^d)}^2 \lesssim \mathbb{E}_{f\sim\Pi_n} K^2 \leq C_{\alpha,\nu,d}. &(157)
\end{aligned}$$

which gives the $\nu \leq 1$ case.

We now turn to the general case. The proof will be similar to the one of Ghosal and van der Vaart [20], Proposition I.3 although we need to control the Hölder norms and work through charts since our Gaussian processes are supported on manifolds. Assume for simplicity that $d = 1, 1 < \nu \leq 2$, otherwise it suffices to introduce coordinates and to proceed by induction on $\lfloor\nu\rfloor$. Let $l \in \{1, \ldots, L\}$, and as before define $\widetilde{K}_l(x, y) = \left(\chi_l\circ\phi_l^{-1}\right)(x)\left(\chi_l\circ\phi_l^{-1}\right)(y)K\left(\phi_l^{-1}(x), \phi_l^{-1}(y)\right)$ the covariance kernel of $h_l = (\chi_l f)\circ\phi_l^{-1}$ as well as $\widetilde{\mathbb{H}}_l$ its RKHS.

First, let us construct an $L^2(\Omega)$-derivative $\dot{h}_l$ of h_l, where $L^2(\Omega)$ is the space of random variables with finite variance with $\langle a, b\rangle_{L^2(\Omega)} = \mathbb{E}(ab)$. This derivative is a square integrable process on $\mathcal{V}_l$ such that

$$\mathbb{E}_{f\sim\Pi_n}\left|\frac{h_l(x+\delta) - h_l(x)}{h} - \dot{h}_l(x)\right|^2 \to 0 \tag{158}$$

as $h \to 0$, for all $x \in \mathcal{V}_l$. For this, we will first show that $\frac{\partial\widetilde{K}_l}{\partial x}(x, \cdot) \in \widetilde{\mathbb{H}}_l$ for every $x \in \mathcal{V}_l$—here $\frac{\partial\widetilde{K}_l}{\partial x}$ denotes the derivative of the function $\widetilde{K}_l(\cdot, \cdot')$ with respect to the first argument—and that

$$\left\|\frac{\partial\widetilde{K}_l}{\partial x}(x, \cdot) - \frac{\partial\widetilde{K}_l}{\partial x}(x', \cdot)\right\|_{\widetilde{\mathbb{H}}_l} \leq C_\nu|x-x'|^{\nu-1}. \tag{159}$$

We first show that $\frac{\widetilde{K}_l(x+\delta,\cdot)-\widetilde{K}_l(x,\cdot)}{h}$ is a Cauchy net[8] in $\widetilde{\mathbb{H}}_l$. We have

$$\left\|\frac{\widetilde{K}_l(x+\delta,\cdot)-\widetilde{K}_l(x,\cdot)}{h}-\frac{\widetilde{K}_l(x+\delta',\cdot)-\widetilde{K}_l(x,\cdot)}{h'}\right\|_{\widetilde{\mathbb{H}}_l} \tag{160}$$

$$=\sup_{\|\varphi\|_{\widetilde{\mathbb{H}}_l}=1}\left\langle\frac{\widetilde{K}_l(x+\delta,\cdot)-\widetilde{K}_l(x,\cdot)}{h}-\frac{\widetilde{K}_l(x+\delta',\cdot)-\widetilde{K}_l(x,\cdot)}{h'},\varphi\right\rangle_{\widetilde{\mathbb{H}}_l} \tag{161}$$

$$=\sup_{\|\varphi\|_{\widetilde{\mathbb{H}}_l}=1}\left(\frac{\varphi(x+\delta)-\varphi(x)}{h}-\frac{\varphi(x+\delta')-\varphi(x)}{h'}\right) \tag{162}$$

$$=\sup_{\|\varphi\|_{\widetilde{\mathbb{H}}_l}=1}\int_0^1\left[\varphi'(x+th)-\varphi'(x+th')\right]\mathrm{d}t \tag{163}$$

$$\leq\sup_{\|\varphi\|_{\widetilde{\mathbb{H}}_l}=1}\|\varphi'\|_{\mathcal{CH}^{\nu-1}(\mathcal{V}_l)}|h-h'|^{\nu-1} \tag{164}$$

$$\leq\sup_{\|\varphi\|_{\widetilde{\mathbb{H}}_l}=1}\|\varphi\|_{\mathcal{CH}^{\nu}(\mathcal{V}_l)}|h-h'|^{\nu-1} \tag{165}$$

where in (163) the derivative φ' exists because, exactly as in the case $\nu\leq 1$, we can show show that $\widetilde{\mathbb{H}}_l\hookrightarrow\mathcal{CH}^{\nu}\left(\mathbb{R}^d\right)$. This also implies that for a constant $C=C_\nu$ we have

$$\left\|\frac{\widetilde{K}_l(x+\delta,\cdot)-\widetilde{K}_l(x,\cdot)}{h}-\frac{\widetilde{K}_l(x+\delta',\cdot)-\widetilde{K}_l(x,\cdot)}{h'}\right\|_{\widetilde{\mathbb{H}}_l}\leq C|h-h'|^{\nu-1}. \tag{166}$$

As $|h-h'|^{\nu-1}\to 0$ when $h,h'\to 0$, because $\nu>1$, this proves that $\frac{\widetilde{K}_l(x+\delta,\cdot)-\widetilde{K}_l(x,\cdot)}{h}$ is a Cauchy net in $\widetilde{\mathbb{H}}_l$: by completeness of $\widetilde{\mathbb{H}}_l$ it converges in $\widetilde{\mathbb{H}}_l$ to a limit. Since by general properties of RKHSs, convergence in $\widetilde{\mathbb{H}}_l$ implies pointwise convergence, the limit satisfies

$$\lim_{h\to 0}\frac{\widetilde{K}_l(x+\delta,y)-\widetilde{K}_l(x,y)}{h}=\frac{\partial\widetilde{K}_l}{\partial x}(x,y). \tag{167}$$

Hence the partial derivative $\frac{\partial\widetilde{K}_l}{\partial x}(x,y)$ exists for all y and $\frac{\partial\widetilde{K}_l}{\partial x}(x,\cdot)\in\widetilde{\mathbb{H}}_l$. Moreover, by the isometry $L^2(\Omega)\ni h_l(x)\mapsto\mathbb{E}_{f\sim\Pi_n}h_l(x)h_l(\cdot)=\widetilde{K}_l(x,\cdot)\in\widetilde{\mathbb{H}}_l$, we deduce that h_l is actually $L^2(\Omega)$-differentiable, with an $L^2(\Omega)$-derivative denoted as $\dot{h}_l$, and that the derivative process $\dot{h}_l$ is Gaussian, as it is an $L^2(\Omega)$-limit of Gaussian random variables, satisfying $\mathbb{E}_{f\sim\Pi_n}\dot{h}_l(x)\dot{h}_l(y)=\left\langle\frac{\partial\widetilde{K}_l}{\partial x}(x,\cdot),\frac{\partial\widetilde{K}_l}{\partial x}(y,\cdot)\right\rangle_{\widetilde{\mathbb{H}}_l}$

Having established the existence of an $L^2(\Omega)$-derivative $\dot{h}_l$ of the process h_l, we would like now to show that $\dot{h}_l$ possesses a $(\gamma-1)$-regular version for every $\gamma<\nu$. For this, we would like to apply Lemma 28 to $\dot{h}_l$. Notice that, by isometry, for all $h>0$ we have

$$\mathbb{E}_{f\sim\Pi_n}\left|\dot{h}_l(x)-\dot{h}_l(y)\right|^2=\left\|\frac{\partial\widetilde{K}_l}{\partial x}(y,\cdot)-\frac{\partial\widetilde{K}_l}{\partial x}(x,\cdot)\right\|^2_{\widetilde{\mathbb{H}}_l} \tag{168}$$

$$\leq 3\left\|\frac{\widetilde{K}_l(y+\delta,\cdot)-\widetilde{K}_l(y,\cdot)}{h}-\frac{\partial\widetilde{K}_l}{\partial x}(y,\cdot)\right\|^2_{\widetilde{\mathbb{H}}_l} \tag{169}$$

$$+3\left\|\frac{\widetilde{K}_l(x+\delta,\cdot)-\widetilde{K}_l(x,\cdot)}{h}-\frac{\partial\widetilde{K}_l}{\partial x}(x,\cdot)\right\|^2_{\widetilde{\mathbb{H}}_l} \tag{170}$$

$$+3\left\|\frac{\widetilde{K}_l(x+\delta,\cdot)-\widetilde{K}_l(x,\cdot)}{h}-\frac{\widetilde{K}_l(y+\delta,\cdot)-\widetilde{K}_l(y,\cdot)}{h}\right\|^2_{\widetilde{\mathbb{H}}_l}. \tag{171}$$

[8] See for instance Aliprantis and Border [1] for a review of Cauchy nets.

Therefore, by the same arguments as above, we have

$$\left(\mathbb{E}_{f\sim\Pi_n}\left|\dot{h}_l(x)-\dot{h}_l(y)\right|^2\right)^{1/2} \lesssim \liminf_{h\to 0}\left\|\frac{\widetilde{K}_l(x+\delta,\cdot)-\widetilde{K}_l(x,\cdot)}{h}-\frac{\widetilde{K}_l(y+\delta,\cdot)-\widetilde{K}_l(y,\cdot)}{h}\right\|_{\widetilde{\mathbb{H}}_l} \tag{172}$$

$$\leq \liminf_{h\to 0}\sup_{\|\varphi\|_{\widetilde{\mathbb{H}}_l}=1}\int_0^1 |\varphi'(x+th)-\varphi'(y+th)|\,\mathrm{d}t \tag{173}$$

$$\leq \liminf_{h\to 0} C_\nu|x-y|^{\nu-1} = C_\nu|x-y|^{\nu-1} \tag{174}$$

where the transition from the second-to-last line to the last line is similar to (163)–(165).

Therefore, we can apply Lemma 28 to $\dot{h}_l$, and find a version $\tilde{h}_l'$ of $\dot{h}_l$ with sample paths in $\mathcal{CH}^{\alpha-1}(\mathcal{V}_l)$ almost surely for all $\alpha<\nu$, such that

$$\mathbb{E}_{f\sim\Pi_n}\left\|\tilde{h}_l'\right\|^2_{\mathcal{CH}^{\alpha-1}(\mathcal{V}_l)} \leq C_{\nu,\alpha} < +\infty \qquad \alpha<\nu. \tag{175}$$

This gives Hölder regularity of the respective derivatives: we now integrate these to obtain a Hölder-regular version of the process itself. Take any $c_l\in(0,1)$ and consider $\tilde{h}_l = h_l(c_l)+\int_{c_l}^{\cdot}\tilde{h}_l'(t)\,\mathrm{d}t$. Then since $\tilde{h}_l'$ is almost surely in $\mathcal{CH}^{\alpha-1}(\mathcal{V}_l)$, $\tilde{h}_l$ is has sample paths almost surely in $\mathcal{CH}^{\alpha}(\mathcal{V}_l)$. Moreover, it is easy to check using our previous results that $\tilde{h}_l$ has an $L^2(\Omega)$-derivative given by $\tilde{h}_l'$. This implies that $\tilde{h}_l$ is a version of h_l.

To conclude the argument, we construct $\tilde{f}$ from the obtained parts, by pulling $\tilde{h}_l$ back from the charts to the manifold. Consider $\tilde{f}=\sum_{l=1}^L \tilde{h}_l\circ\phi_l$. Arguing as in the case $\nu\leq 1$, we see that $\tilde{f}$ is a version of f with $\mathcal{CH}^\alpha(\mathcal{M})$ sample paths for every $\alpha<\nu$, and for every $\alpha<\nu$ we have $\mathbb{E}_{f\sim\Pi_n}\|\tilde{f}\|^2_{\mathcal{CH}^\alpha(\mathcal{M})}\leq C_{\alpha,\nu}<+\infty$. This gives the claim. □

With this, it is easy to prove that all Matérn Gaussian processes considered in this paper can be seen as Gaussian random elements in the Banach space $\left(\mathcal{C}(\mathcal{M}),\|\cdot\|_{\mathcal{C}(\mathcal{M})}\right)$ of continuous functions on $\mathcal{M}$. This allows us to use the same proof scheme as in van der Vaart and van Zanten [56] through the control of the *concentration functions* defined in Appendix B. It is also important that we work with Gaussian random elements in $\mathcal{C}(\mathcal{M})$—and not only with the classical notion of Gaussian process, as the concentration functions are defined using the *Gaussian random element RKHS* defined in van Zanten and van der Vaart [59], which can potentially be different from the classical RKHS. Fortunately, when the process is a Gaussian random element in $\mathcal{C}(\mathcal{M})$, van Zanten and van der Vaart [59], Theorem 2.1 implies that the two notions of RKHS coincide.

Corollary 30. *The intrinsic Matérn Gaussian processes of Definition 4, their truncated versions as in Theorem 6 as well as the extrinsic Matérn Gaussian processes of Definition 7 are Gaussian random elements in* $\left(\mathcal{C}(\mathcal{M}),\|\cdot\|_{\mathcal{C}(\mathcal{M})}\right)$.

Proof. By Lemma 18, it suffices to show that the processes have almost surely continuous sample paths. Since Euclidean Matérn Gaussian processes have continuous sample paths, this implies the same for their restrictions, the extrinsic Matérn Gaussian processes on $\mathcal{M}$. For the intrinsic Matérn process, we apply Lemma 29. □

Using Lemma 29 and known properties of Euclidean Matérn processes, we now show, in a sense, that all of the Matérn processes presented in this paper are sub-Gaussian, in a manner which holds uniformly with respect to the truncation parameter in the case of the truncated intrinsic Matérn process, and live in Hölder spaces with appropriate exponents. This result is used to control Hölder norms when going from the error at input locations to the $L^2(p_0)$-error. We use the notation Π_n to emphasize that the prior depends on n when we consider a truncated intrinsic Matérn process.

Lemma 31. *For $f\sim\Pi_n$ with Π_n the prior in either Definition 4, Theorem 6 or Definition 7, for every $\nu>0$ and $\gamma<\nu,\gamma\notin\mathbb{Z}_{>0}$, there exists a constant $\sigma(f)=\sigma_\gamma(f)$ independent of n we have for $x>0$ that*

$$\mathbb{P}\Big(\|f\|_{\mathcal{CH}^\gamma(\mathcal{M})}>(x+1)\sigma(f)\Big)\leq 2e^{-x^2/2}. \tag{176}$$

Proof. We start with the restriction f of an extrinsic Matérn process $\tilde{f}$ to $\mathcal{M}$, as in Definition 7. By van der Vaart and van Zanten [56], Section 3.1, for every $\gamma < \nu$ we have $\tilde{f} \in \mathcal{CH}^\gamma\big([0,1]^D\big)$ almost surely. By Ghosal and van der Vaart [20], Lemma I.7, for every $\gamma < \nu$, $\tilde{f}$ is a Gaussian random element in the Banach space $\mathcal{CH}^\gamma\big([0,1]^D\big)$. In particular, by the Borell–TIS inequality [20, Proposition I.8] we have for $x > 0$ that

$$\mathbb{P}\Big(\|\tilde{f}\|_{\mathcal{CH}^\gamma([0,1]^D)} > (x+1)\sigma\big(\tilde{f}\big)\Big) \leq 2e^{-x^2/2} \tag{177}$$

where $\sigma\big(\tilde{f}\big) = \Big(\mathbb{E}\,\|\tilde{f}\|^2_{\mathcal{CH}^\gamma([0,1]^D)}\Big)^{1/2} < \infty$. Since $\mathcal{M}$ is smooth, the restriction f also satisfies for $x > 0$ the expression

$$\mathbb{P}\Big(\|f\|_{\mathcal{CH}^\gamma(\mathcal{M})} > (x+1)\sigma(f)\Big) \leq 2e^{-x^2/2} \tag{178}$$

perhaps for a possibly larger constant $\sigma(f)$. Finally, the case of the intrinsic Matérn process $f \sim \Pi_n$ truncated at $J_n \in \mathbb{Z}_{>0} \cup \{\infty\}$ follows in the same manner, as we have shown in Lemma 29 that $\sup_{n\geq 1} \mathbb{E}_{f\sim\Pi_n} \|f\|^2_{\mathcal{CH}^\alpha(\mathcal{M})} \leq C_{\alpha,\nu}$. □

E Small Ball Asymptotics

Here, we bound the probability that a Matérn Gaussian process lies in the ε-ball with respect to the L^∞-norm for small ε. For a Banach space $\mathbb{B}$, an element $x \in \mathbb{B}$ and a number $r \in \mathbb{R}_{>0}$, let us denote the closed r-ball around x by $B^{\mathbb{B}}_r(x)$. We start by an upper bound on the metric entropy of Sobolev balls on $\mathcal{M}$ with respect to the uniform norm.

Lemma 32 (Entropy of Sobolev balls)**.** *For any $s > 0$ define the ε-covering number of $A \subseteq H^s(\mathcal{M})$ with respect to the norm $\|\cdot\|_{L^\infty(\mathcal{M})}$ by*

$$N\Big(\varepsilon, A, \|\cdot\|_{L^\infty(\mathcal{M})}\Big) = \operatorname*{arg\,min}_{J\in\mathbb{Z}_{>0}}\Bigg\{\exists h_1, .., h_J \in A : A \subset \bigcup_{j=1}^{J} B^{L^\infty(\mathcal{M})}_\varepsilon(h_j)\Bigg\}. \tag{179}$$

Then for any $s > d/2$, there exist $C, \varepsilon_0 > 0$ such that for every $\varepsilon \leq \varepsilon_0$

$$\ln N\Big(\varepsilon, B^{H^s(\mathcal{M})}_1(0), \|\cdot\|_{L^\infty(\mathcal{M})}\Big) \leq C\varepsilon^{-\frac{d}{s}}, \tag{180}$$

where the left-hand side of the inequality above, as a function of ε, is called the METRIC ENTROPY *of the Sobolev ball $B^{H^s(\mathcal{M})}_1(0)$ with respect to the uniform norm $\|\cdot\|_{L^\infty(\mathcal{M})}$.*

Proof. Using charts we will reduce the problem to the entropy of the unit ball of the Sobolev space $H^s\big([0,1]^d\big)$ for which the upper bound is known. Take $f \in B^{H^s(\mathcal{M})}_1(0)$ and consider approximations of f by g of the form

$$g = \sum_{l=1}^{L} \chi_l (g_l \circ \phi_l) \tag{181}$$

for some functions $g_l : \mathcal{V}_l \to \mathbb{R}$ where $\mathcal{V}_l \subseteq \mathbb{R}^d$. We have

$$\big\|f - g\big\|_{L^\infty(\mathcal{M})} = \big\|\sum_{l=1}^{L} \chi_l(g_l \circ \phi_l - f)\big\|_{L^\infty(\mathcal{M})} \leq \sum_{l=1}^{L}\big\|\chi_l(g_l \circ \phi_l - f)\big\|_{L^\infty(\mathcal{U}_l)} \tag{182}$$

$$\leq \sum_{l=1}^{L}\big\|g_l \circ \phi_l - f\big\|_{L^\infty(\mathcal{U}_l)} \leq \sum_{l=1}^{L}\big\|g_l - f \circ \phi_l^{-1}\big\|_{L^\infty(\mathcal{V}_l)} \tag{183}$$

$$\leq L \max_{1\leq l\leq L}\big\|g_l - f \circ \phi_l^{-1}\big\|_{L^\infty([0,1]^d)}. \tag{184}$$

This means that to approximate f by g uniformly on $\mathcal{M}$, we need to choose the functions g_l that approximate $f \circ \phi_l^{-1}$ well with respect to the uniform norm on $[0,1]^d$.

Next, we show that the functions $f \circ \phi_l^{-1}$ are contained in an Euclidean Sobolev ball of radius R, with R depending only on ν and the atlas. To do this we use Große and Schneider [24], Lemma 2.1 to write

$$\left\| f \circ \phi_l^{-1} \right\|_{H^s([0,1]^d)} = \left\| \sum_{l'=1}^{L} (\chi_{l'} f) \circ \phi_l^{-1} \right\|_{H^s([0,1]^d)} \leq \sum_{l'=1}^{L} \left\| (\chi_{l'} f) \circ \phi_l^{-1} \right\|_{H^s([0,1]^d)} \tag{185}$$

$$= \sum_{l'=1}^{L} \left\| (\chi_{l'} f) \circ \phi_{l'}^{-1} \circ \phi_{l'} \circ \phi_l^{-1} \right\|_{H^s([0,1]^d)} \tag{186}$$

$$\lesssim \sum_{l'=1}^{L} \left\| (\chi_{l'} f) \circ \phi_{l'}^{-1} \right\|_{H^s([0,1]^d)} \lesssim \|f\|_{H^s(\mathcal{M})}. \tag{187}$$

Note, importantly, that the remark just above Große and Schneider [24], Lemma 2.1 allows us to consider Besov spaces $B^s_{2,2}$ coinciding with the Sobolev spaces H^s instead of the Besov spaces $B^s_{2,\infty}$—to get from the second line to the third. Note also that the constant hidden behind the notation $\lesssim$ in the last line where we use Theorem 16 is the radius R. Without loss of generality we assume $R = 1$. By the Euclidean counterpart of the result we are proving [21, Theorem 4.3.36], we have

$$\ln N\left(\varepsilon, B_1^{H^s[0,1]^d}(0), \|\cdot\|_{L^\infty([0,1]^d)}\right) \lesssim \varepsilon^{-\frac{d}{s}}. \tag{188}$$

Let $h_1, .., h_J \in B_1^{H^s([0,1]^d)}(0)$ be such that $B_1^{H^s([0,1]^d)}(0) \subset \bigcup_{j=1}^{J} B_{\varepsilon/L}^{L^\infty([0,1]^d)}(h_k)$. Then for any $f \in B_1^{H^s(\mathcal{M})}(0)$ there exists a sequence $\{j_l\}_{l=1}^{L} \subseteq \{1, .., J\}$ such that

$$\left\| f - \sum_{l=1}^{L} \chi_l (h_{j_l} \circ \phi_l) \right\|_{L^\infty(\mathcal{M})} < L \frac{\varepsilon}{L} = \varepsilon. \tag{189}$$

This shows that $N\left(\varepsilon, B_1^{H^s(\mathcal{M})}(0), \|\cdot\|_{L^\infty(\mathcal{M})}\right) \leq LJ$, where L is just the number of charts, thereby proving the claim. $\square$

For the related *diffusion spaces* [16], the RKHS corresponding to the heat (diffusion) kernels, Castillo et al. [12] uses the results of Coulhon et al. [14] to bound the entropy in terms of a wavelet frame instead of relying on charts. We believe this alternative proof scheme should work in our case as well. However, we could not, to the best of our effort, get a tight-enough bound for the Sobolev spaces by directly using the results of Coulhon et al. [14] and therefore we chose to rely on charts instead.

Having established regularity properties for our prior processes, we now turn to the *small ball problem*: we want to find sharp lower bounds on $\mathbb{P}(\|f\|_{L^\infty(\mathcal{M})} < \varepsilon)$ where $f \sim \Pi$ is our prior process. This will be crucial in order to control the *concentration functions* used in Appendix B. In fact, it is well-known that this problem is closely related to the estimation of the metric entropy of the unit ball of the RKHS of f with respect to the uniform norm: see Li and Linde [31] for details. Since we have already characterized the RKHS of our processes in Proposition 27 and Lemma 24, we are able to lower bound the small-ball probabilities. The technicality here involves getting a bound uniform in the truncation parameter for the truncated intrinsic Matérn process, as the truncated Matérn process is a sequence of priors rather than a fixed prior.

Lemma 33. *If $f \sim \Pi_n$ where Π_n is the prior in either Definition 4 and Theorem 6 or Definition 7 with smoothness parameter $\nu > d/2$, then there exist two constants $C, \varepsilon_0 > 0$ that do not depend on n such that for all $\varepsilon \leq \varepsilon_0$ we have $-\ln \mathbb{P}\left(\|f\|_{L^\infty(\mathcal{M})} < \varepsilon\right) \leq C\varepsilon^{-\frac{d}{\nu}}$.*

Proof. Because the processes are Gaussian random elements in $\mathcal{C}(\mathcal{M})$ by Lemma 18, their stochastic process RKHS given by Proposition 27 coincide with their *Gaussian random element RKHS* defined in van Zanten and van der Vaart [59]. Hence, for the non-truncated intrinsic and the extrinsic Matérn processes, the result follows by a direct application of Lemma 32 and Li and Linde [31], Theorem 1.2. For the intrinsic Matérn process truncated at J_n it is not immediately clear that the constants C, ε_0 can be taken independent of n, so we go through the proof of Li and Linde [31], Proposition 3.1 to see that this is, in fact, the case. We first need a crude upper bound of the form

$$-\ln \mathbb{P}\left(\|f\|_{L^\infty(\mathcal{M})} < \varepsilon\right) \leq c\varepsilon^{-c} \tag{190}$$

for some possibly large constant $c > 0$. To get such a bound, we use Castillo et al. [12], Proposition 3 which shows the existence of a universal constant $C > 0$ such that for all $\varepsilon \leq \min(1, 4\sigma(f))$

$$-\ln \mathbb{P}\Big(\|f\|_{L^\infty(\mathcal{M})} < \varepsilon\Big) \leq C n(\varepsilon) \ln\left(\frac{6n(\varepsilon)\max(1,\sigma(f))}{\varepsilon}\right) \tag{191}$$

where $\sigma(f) = \Big(\mathbb{E}_{f\sim\Pi_n}\|f\|^2_{L^\infty(\mathcal{M})}\Big)^{1/2}$ and $n(\varepsilon)$ is defined in the following way in Castillo et al. [12], page 684 using auxiliary quantities l_J that are defined in Li and Linde [31], page 1562, namely

$$n(\varepsilon) = \max\{J \geq 0 : 4l_J(f) \geq \varepsilon\}, \tag{192}$$

$$l_J(f) = \inf\Big\{\Big(\mathbb{E}_{f\sim\Pi_n}\Big\|\sum\nolimits_{j\geq J}\varepsilon_j h_j\Big\|^2_{L^\infty(\mathcal{M})}\Big)^{1/2} : f \overset{(d)}{=} \sum\nolimits_{j\geq 1}\varepsilon_j h_j\Big\} \tag{193}$$

with $\overset{(d)}{=}$ standing for the equality in distributions and the infimum being taken over all possible decompositions $\sum_{j\geq1}\varepsilon_j h_j$ with $h_j \in \mathcal{C}(\mathcal{M})$, ε_j being a sequence of IID $\mathrm{N}(0,1)$ random variables, and the series being required to converge uniformly almost surely.[9]

The function $f = \sum_{j=1}^{J_n+1}\big(\frac{2\nu}{\kappa^2}+\lambda_{j-1}\big)^{-\frac{\nu+d/2}{2}}\varepsilon_j f_{j-1}$ is a valid decomposition. Therefore

$$l_J(f) \leq \left(\mathbb{E}_{f\sim\Pi_n}\left\|\sum_{j=J}^{J_n+1}\left(\frac{2\nu}{\kappa^2}+\lambda_{j-1}\right)^{-\frac{\nu+d/2}{2}}\varepsilon_j f_{j-1}\right\|^2_{L^\infty(\mathcal{M})}\right)^{1/2}. \tag{194}$$

By the Sobolev Embedding Theorem and by Weyl's Law, given in Result 10, for every $d/2 < \gamma < \nu$ there exists a constant $C = C_{\gamma,\mathcal{M}}$ such that for all $J \in \mathbb{Z}_{>0}$, allowing C to change from line to line, we have

$$\begin{aligned}\mathbb{E}_{f\sim\Pi_n}\left\|\sum_{j=J}^{J_n+1}(1+\lambda_{j-1})^{-\frac{\nu+d/2}{2}}\varepsilon_j f_{j-1}\right\|^2_{L^\infty(\mathcal{M})} &\leq C^2\,\mathbb{E}_{f\sim\Pi_n}\left\|\sum_{j=J}^{J_n+1}(1+\lambda_{j-1})^{-\frac{\nu+d/2}{2}}\varepsilon_j f_{j-1}\right\|^2_{H^\gamma(\mathcal{M})} && (195)\\ &= C^2\sum_{j=J}^{J_n+1}(1+\lambda_{j-1})^{-(\nu+d/2-\gamma)} && (196)\\ &\leq C^2\sum_{j=J}^{J_n+1} j^{-(1+2(\nu-\gamma)/d)} && (197)\\ &\leq C^2\sum_{j\geq J} j^{-(1+2(\nu-\gamma)/d)} && (198)\\ &\leq C^2 J^{-2(\nu-\gamma)/d}. && (199)\end{aligned}$$

By choosing $J = 1$ this gives us $\sigma(f) \leq C$ independent of n. Moreover, by choosing $J \geq C\varepsilon^{-\frac{d}{2(\nu-\gamma)}}$, again for a comparison constant C independent of n, this gives us $n(\varepsilon) \leq C\varepsilon^{-\frac{d}{2(\nu-\gamma)}}$ for C independent of n. This implies using Castillo et al. [12], Proposition 3 that

$$-\ln \mathbb{P}\Big(\|f\|_{L^\infty(\mathcal{M})} < \varepsilon\Big) \leq c\varepsilon^{-c} \tag{200}$$

for $c > 0$ independent of n.

With this crude bound, we can now continue the proof of Li and Linde [31], Proposition 3.1. For this, we need a metric entropy estimate. For this notice that for all $J \in \mathbb{Z}_{>0} \cup \{\infty\}$ we have $B_1^{\mathbb{H}^J}(0) \subset B_1^{\mathbb{H}^\infty}(0) = B_1^{H^{\nu+d/2}(\mathcal{M})}(0)$, and therefore using Lemma 32, we have the metric entropy estimate

$$\ln N\Big(\varepsilon, B_1^{\mathbb{H}^J}(0), \|\cdot\|_{L^\infty(\mathcal{M})}\Big) \leq C\varepsilon^{-\frac{d}{\nu+d/2}} \tag{201}$$

for a constant $C > 0$ independent of J. Therefore following the proof of proposition 3.1 in Li and Linde [31] (with $J \equiv 1$) we find $-\ln\mathbb{P}\Big(\|f\|_{L^\infty(\mathcal{M})} < \varepsilon\Big) \leq C\varepsilon^{-\frac{d}{\nu}}$ for every $\varepsilon \leq \varepsilon_0$, where $C, \varepsilon_0 > 0$ are constants independent of n. □

[9] We consider $\sum_{j\geq1}$, unlike $\sum_{j\geq0}$ frequently used above, in order to follow the respective references.

F Expressions for Pointwise Worst-case Errors

Let k be a kernel on some abstract input domain $\mathcal{X}$, and let $\mathbb{H}_k$ be the respective RKHS. Consider n input values $\mathbf{X} \subseteq \mathcal{X}$ and let $\sigma_\varepsilon^2 > 0$ be the noise variance. Define

$$m_{k,\mathbf{X},f,\varepsilon}(t) = \mathbf{K}_{t\mathbf{X}}(\mathbf{K}_{\mathbf{X}\mathbf{X}} + \sigma_\varepsilon^2 \mathbf{I})^{-1}(f(\mathbf{X}) + \boldsymbol{\varepsilon}), \tag{202}$$

$$v^{(\mathrm{i})}(t) = v_{k,\mathbf{X}}(t) = k(t,t) - \mathbf{K}_{t\mathbf{X}}\left(\mathbf{K}_{\mathbf{X}\mathbf{X}} + \sigma_\varepsilon^2 \mathbf{I}\right)^{-1}\mathbf{K}_{\mathbf{X}t}. \tag{203}$$

Proposition 34. *With notation above*

$$v^{(\mathrm{i})}(t) = \sup_{f \in \mathcal{H}_k, \|f\|_{\mathcal{H}_k} \leq 1} \mathbb{E}_{\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I})} |f(t) - m_{k,\mathbf{X},f,\varepsilon}(t)|^2. \tag{204}$$

Proof. To simplify notation, we shorten $\mathbb{E}_{\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I})}$ to $\mathbb{E}$ and denote $\boldsymbol{\alpha} = \mathbf{K}_{t\mathbf{X}}(\mathbf{K}_{\mathbf{X}\mathbf{X}} + \sigma_\varepsilon^2 \mathbf{I})^{-1}$. First of all, by direct computation,

$$\mathbb{E}\, m_{k,\mathbf{X},f,\varepsilon}(t) = \boldsymbol{\alpha} f(\mathbf{X}), \tag{205}$$

$$\mathbb{E}\, m_{k,\mathbf{X},f,\varepsilon}(t)^2 = \boldsymbol{\alpha} f(\mathbf{X}) f(\mathbf{X})^\top \boldsymbol{\alpha}^\top + \sigma_\varepsilon^2 \boldsymbol{\alpha}\boldsymbol{\alpha}^\top. \tag{206}$$

Write

$$\begin{aligned}
\mathbb{E}|f(t) - m_{k,\mathbf{X},f,\varepsilon}(t)|^2 &= f(t)^2 - 2f(t)\,\mathbb{E}\, m_{k,\mathbf{X},f,\varepsilon}(t) + \mathbb{E}\, m_{k,\mathbf{X},f,\varepsilon}(t)^2 && (207)\\
&= f(t)^2 - 2f(t)\boldsymbol{\alpha} f(\mathbf{X}) + \boldsymbol{\alpha} f(\mathbf{X}) f(\mathbf{X})^\top \boldsymbol{\alpha}^\top + \sigma_\varepsilon^2 \boldsymbol{\alpha}\boldsymbol{\alpha}^\top && (208)\\
&= (f(t) - \boldsymbol{\alpha} f(\mathbf{X}))^2 + \sigma_\varepsilon^2 \boldsymbol{\alpha}\boldsymbol{\alpha}^\top && (209)\\
&= \left\langle k(t,\cdot) - \sum_{j=1}^n \alpha_j k(x_j,\cdot), f \right\rangle_{\mathbb{H}_k}^2 + \sigma_\varepsilon^2 \boldsymbol{\alpha}\boldsymbol{\alpha}^\top. && (210)
\end{aligned}$$

As $\|g\|_{\mathbb{H}_k} = \sup_{f \in \mathbb{H}_k, \|f\|_{\mathbb{H}_k} \leq 1} \langle g, f \rangle_{\mathbb{H}_k}$, implying $\sup_{f \in \mathbb{H}_k, \|f\|_{\mathbb{H}_k} \leq 1} \langle g, f \rangle_{\mathbb{H}_k}^2 = \|g\|_{\mathbb{H}_k}^2$, we have

$$\begin{aligned}
\sup_{\substack{f \in \mathcal{H}_k \\ \|f\|_{\mathcal{H}_k} \leq 1}} \mathbb{E}|f(t) - m_{k,\mathbf{X},f,\varepsilon}(t)|^2 &= \left\| k(t,\cdot) - \sum_{j=1}^n \alpha_j k(x_j,\cdot) \right\|_{\mathbb{H}_k}^2 + \sigma_\varepsilon^2 \boldsymbol{\alpha}\boldsymbol{\alpha}^\top && (211)\\
&= k(t,t) - 2\boldsymbol{\alpha}\mathbf{K}_{\mathbf{X}t} + \underbrace{\boldsymbol{\alpha}\mathbf{K}_{\mathbf{X}\mathbf{X}}\boldsymbol{\alpha}^\top + \sigma_\varepsilon^2 \boldsymbol{\alpha}\boldsymbol{\alpha}^\top}_{\boldsymbol{\alpha}\mathbf{K}_{\mathbf{X}t}} && (212)\\
&= k(t,t) - \boldsymbol{\alpha}\mathbf{K}_{\mathbf{X}t} = \underbrace{k(t,t) - \mathbf{K}_{t\mathbf{X}}(\mathbf{K}_{\mathbf{X}\mathbf{X}} + \sigma_\varepsilon^2 \mathbf{I})^{-1}\mathbf{K}_{\mathbf{X}t}}_{v_{k,\mathbf{X}}(t)}. && (213)
\end{aligned}$$

□

We now move to the misspecified case. Consider the RKHS $\mathcal{H}_c$ for some other kernel $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ instead of $\mathbb{H}_k$. Then, continuing from (210), write

$$\sup_{\substack{f \in \mathcal{H}_c \\ \|f\|_{\mathcal{H}_c} \leq 1}} \mathbb{E}|f(t) - m_{k,\mathbf{X},f,\varepsilon}(t)|^2 = \left\| c(t,\cdot) - \sum_{j=1}^n \alpha_j c(x_j,\cdot) \right\|_{\mathbb{H}_c}^2 + \sigma_\varepsilon^2 \boldsymbol{\alpha}\boldsymbol{\alpha}^\top. \tag{214}$$

The next question is how to compute the norm on the right-hand side. There is not much hope of doing this exactly in the misspecified case: thus, we consider approximations. To this end, we take some large set of locations $\mathbf{X}' \subseteq \mathcal{X}$. Then we use $\|g\|_{\mathbb{H}_c}^2 \approx g(\mathbf{X}')^\top \mathbf{C}_{\mathbf{X}'\mathbf{X}'}^{-1} g(\mathbf{X}')$ for $g(\cdot) = c(t,\cdot) - \sum_{j=1}^n \alpha_j c(x_j,\cdot)$. As a result, we obtain the approximation

$$\sup_{\substack{f \in \mathcal{H}_c \\ \|f\|_{\mathcal{H}_c} \leq 1}} \mathbb{E}|f(t) - m_{k,\mathbf{X},f,\varepsilon}(t)|^2 \approx g(\mathbf{X}')^\top \mathbf{C}_{\mathbf{X}'\mathbf{X}'}^{-1} g(\mathbf{X}') + \sigma_\varepsilon^2 \boldsymbol{\alpha}\boldsymbol{\alpha}^\top = \tilde{v}_{k,c,\mathbf{X}}(t) = v^{(\mathrm{e})}(t) \tag{215}$$

where $v^{(\mathrm{e})}(t)$ was first introduced in Section 4

To compute spatial averages of this quantity, let $g_t(\cdot) = c(t, \cdot) - \sum_{j=1}^{n} \alpha_j c(x_j, \cdot)$, the same as g before, but now with explicit dependence on t. Similarly, put $\boldsymbol{\alpha}_t = \mathbf{K}_{t\mathbf{X}}(\mathbf{K}_{\mathbf{X}\mathbf{X}} + \sigma_\varepsilon^2 \mathbf{I})^{-1}$. Then

$$g_t(\mathbf{X}') = \mathbf{C}_{\mathbf{X}'\,t} - \mathbf{C}_{\mathbf{X}'\,\mathbf{X}}\boldsymbol{\alpha}_t^\top = \mathbf{C}_{\mathbf{X}'\,t} - \mathbf{C}_{\mathbf{X}'\,\mathbf{X}}(\mathbf{K}_{\mathbf{X}\mathbf{X}} + \sigma_\varepsilon^2 \mathbf{I})^{-1}\mathbf{K}_{\mathbf{X}t} \tag{216}$$

$$g_t(\mathbf{X}')^\top \mathbf{C}_{\mathbf{X}'\mathbf{X}'}^{-1} g_t(\mathbf{X}') = (\mathbf{C}_{t\,\mathbf{X}'} - \boldsymbol{\alpha}_t \mathbf{C}_{\mathbf{X}\,\mathbf{X}'})\mathbf{C}_{\mathbf{X}'\mathbf{X}'}^{-1}\big(\mathbf{C}_{\mathbf{X}'\,t} - \mathbf{C}_{\mathbf{X}'\,\mathbf{X}}\boldsymbol{\alpha}_t^\top\big). \tag{217}$$

From here we can also deduce that

$$\frac{1}{|\mathbf{X}'|}\sum_{t\in\mathbf{X}'} \tilde{v}_{k,c,\mathbf{X}}(t) = \frac{1}{|\mathbf{X}'|}\sum_{t\in\mathbf{X}'} g_t(\mathbf{X}')^\top \mathbf{C}_{\mathbf{X}'\mathbf{X}'}^{-1} g_t(\mathbf{X}') \tag{218}$$

$$= \frac{1}{|\mathbf{X}'|}\operatorname{tr}\big(g_{\mathbf{X}'}(\mathbf{X}')^\top \mathbf{C}_{\mathbf{X}'\mathbf{X}'}^{-1} g_{\mathbf{X}'}(\mathbf{X}')\big) \tag{219}$$

where $g_{\mathbf{X}'}(\mathbf{X}') = \mathbf{C}_{\mathbf{X}'\,\mathbf{X}'} - \mathbf{C}_{\mathbf{X}'\,\mathbf{X}}(\mathbf{K}_{\mathbf{X}\mathbf{X}} + \sigma_\varepsilon^2 \mathbf{I})^{-1}\mathbf{K}_{\mathbf{X}\mathbf{X}'}$.

G Full Experimental Details

All of our kernels were computed using GPJAX [38] and the GEOMETRIC KERNELS library.[10] We use three manifolds, each represented by a mesh: (i) a dumbbell-shaped manifold represented as a mesh with 1556 nodes, (ii) a sphere represented by an icosahedral mesh with 2562 nodes, and (iii) the Stanford dragon mesh, preprocessed to keep only its largest connected component, which has 100179 nodes. For the sphere, we also considered a finer icosahedral mesh with 10242 nodes, but this was found to have virtually no effect on the computed pointwise expected errors.

We use extrinsic Matérn and Riemannian Matérn kernels with the following hyperparameters: $\sigma_f^2 = 1$ and $\sigma_\varepsilon^2 = 0.0005$. For the truncated Karhunen–Loève expansion, we used $J = 500$ eigenpairs obtained from the mesh. We selected smoothness values to ensure norm-equivalence of the intrinsic and extrinsic kernels' reproducing kernel Hilbert spaces, which was $\nu = 5/2$ for the intrinsic model, and $\nu = 5/2 + d/2$ for the extrinsic model, where d is the manifold's dimension. We used different length scales for each manifold: $\kappa = 200$ for the dumbbell, $\kappa = 0.25$ for the sphere, and $\kappa = 0.05$ for the dragon, selected to ensure that the Gaussian processes were neither approximately constant, nor white-noise-like. We considered data sizes of $N = 50$, $N = 500$, and $N = 1000$, respectively, for the dumbbell, sphere, and dragon, sampled uniformly from the mesh's nodes, which in each case resulted in a reasonably-uniform distribution of points across the manifold. Finally, for the extrinsic pointwise error approximation, we used a subset $\mathbf{X}'$ uniformly sampled from each mesh's nodes, of size equal to the data size. For each respective test set, we used the full mesh. Each experiment was repeated for 10 different seeds.

To set the length scales for the extrinsic process, we used maximum marginal likelihood optimization on the full data, except for the dumbbell whose full data size is small and for which we instead generated a larger set consisting of 500 points. We optimzied only the length scale, leaving all other hyperparameters fixed. We used ADAM with a learning rate of 0.005, and an initialization equal to the length scale κ of the intrinsic model, except for the dumbbell where this lead to divergence and we instead used an initial value of $\kappa/4$. We ran the optimizer for a total of 1000 steps. With these settings, we found empirically that maximum marginal likelihood optimization always converged.

[10] See HTTPS://GPJAX.READTHEDOCS.IO and HTTPS://GEOMETRIC-KERNELS.GITHUB.IO.

Statement of Authorship for joint/multi-authored papers for PGR thesis
To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Posterior Contraction Rates for Matérn Gaussian Processes on Riemannian Manifolds
Publication Status	Published
Publication Details	R., Borovitskiy, Terenin & Rousseau, *Posterior Contraction Rates for Matérn Gaussian Processes on Riemannian Manifolds* (2023), Proceedings of the Thirty-Seventh Annual Conference on Neural Information Processing Systems (NeurIPS 2023), New Orleans

Student Confirmation

Student Name:	Paul Rosa		
Contribution to the Paper	I am the first author of this paper. I did the theoretical asymptotic analysis and most of the writing (at the exception of the experimental Section), with Viacheslav Borovitskiy helping me for some technical points (in particular the Holder regularity regularity result for the intrinsic processes).		
Signature Paul Rosa		Date	06/01/2025

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Prof Judith Rousseau		
Supervisor comments	I confirm the contribution statement		
Signature		Date	06/01/2025

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 3

Nonparametric regression on random geometric graphs sampled from submanifolds

This is a joint work with Judith Rousseau and is currently accepted for publication at the *Journal of Machine learning Research* (JMLR) conditioned on minor revisions.

Journal of Machine Learning Research volume (year) 1-65 Submitted [DATE SUBMITTED]; Revised [DATE REVISED]; Published [DATE PUB

Nonparametric Regression on Random Geometric Graphs Sampled from Submanifolds

Paul Rosa PAUL.ROSA@STATS.OX.AC.UK
Department of Statistics
University of Oxford
24-29 St Giles', Oxford OX1 3LB, United Kingdom

Judith Rousseau ROUSSEAU@CEREMADE.DAUPHINE.FR
Department of Statistics
University of Oxford
24-29 St Giles', OX1 3LB Oxford, United Kingdom
and CEREMADE, CNRS
Université Paris-Dauphine, PSL University
Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France

Editor: TBD

Abstract

We consider the nonparametric regression problem when the covariates are located on an unknown compact submanifold of a Euclidean space. Under defining a random geometric graph structure over the covariates we analyse the asymptotic frequentist behaviour of the posterior distribution arising from Bayesian priors designed through random basis expansion in the graph Laplacian eigenbasis. Under Hölder smoothness assumption on the regression function and the density of the covariates over the submanifold, we prove that the posterior contraction rates of such methods are minimax optimal (up to logarithmic factors) for any positive smoothness index.

Keywords: graph Laplacian, Bayesian nonparametrics, manifold hypothesis, nonparametric regression, dimensionality reduction

1 Introduction

When the data is high dimensional, it is common practice to assume a lower dimensional structure. This idea dates back at least to Pearson's principal component analysis (Pearson, 1901) and the Johnson-Lindenstrauss lemma (Johnson, Lindenstrauss, and Schechtman, 1986; Ailon and Chazelle, 2009) where a linear projection onto a lower dimensional subspace of variables is performed while still retaining much of the statistical information. However, a linear constraint on the possible intrinsic shape of the data can be restrictive, and non-linear dimensionality reduction has received much attention during the past two decades. Examples include kernel PCA (Schölkopf, Smola, and Müller, 1998), Isomap (Tenenbaum, de Silva, and Langford, 2000), locally linear embeddings (Roweis and Saul, 2000), Laplacian eigenmaps and diffusion maps (Coifman and Lafon, 2006). These techniques typically take into account intrinsic geometric properties of the data.

Inference of intrinsic geometric properties of data is often believed to be a powerful way to gain insight in complex high dimensional statistical problems. Examples include its support (Genovese, Perone-Pacifico, Verdinelli, and Wasserman, 2012; Aamari and Levrard, 2019) under the Hausdorff metric, differential quantities (Aamari and Levrard, 2019; Aamari, Berenfeld, and Levrard, 2023), the geodesic distance (Aamari, Berenfeld, and Levrard, 2023; Bernstein, de Silva, Langford, and Tenenbaum), the intrinsic dimension (Kim, Rinaldo, and Wasserman, 2019; Denti, Doimo, Laio, and Mira, 2022) or even the Laplace-Beltrami operator eigenpairs (García Trillos, Gerlach, Hein, and Slepčev, 2020; Calder and Trillos, 2020; Dunson, Wu, and Wu, 2021; Wormell and Reich, 2021). In the same way, inferring the intrinsic topological structure in the data has also received a lot of interest since the modern development of topological data analysis as field within data science, with in particular the study of the persistence diagrams and modules (Chazal and Michel, 2021; Chazal, de Silva, Glisse, and Steve, 2016; Divol and Lacombe, 2021; Divol and Chazal, 2019; Loiseaux, Carrière, and Blumberg, 2023; Loiseaux, Scoccola, Carrière, Botnan, and Oudot, 2024).

The Laplacian eigenmaps, introduced initially by Belkin and Niyogi (2001, 2003) is a popular non linear method for dimensionality reduction, together with a tool for inference in high dimension setups. It has also been used successfully in the context of spectral clustering (Liu and Han, 2014; von Luxburg, 2007), supervised and semisupervised learning (Green, Balakrishnan, and Tibshirani, 2021; Shi, Balasubramanian, and Polonik, 2023; Sanz-Alonso and Yang, 2022) : in this paper we consider the use of Laplacian eigenmaps in the context of the reknown nonparametric regression problem where

$$y_i = f(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \quad i \leq n; \tag{1}$$

with covariates $x_i \in \mathbb{R}^D$ and D is possibly large. Then, a way to formalize a low dimensional structure to the problem is to assume that the covariates X_i's belong to some low dimensional submanifold $\mathcal{M}$ of $\mathbb{R}^D$, also known as the manifold hypothesis. This is the setup we are considering. We also consider both the supervised learning setup where the observations consist of $(y_i, x_i)_{i=1}^n$, under model 1; and the semi-supervised setup where in addition another sample $(x_i)_{i=n+1}^N$ of covariates are observed without their labels y_i.

Laplacian eigenmaps approaches are based on the construction of a graph $G = (V, E)$ whose vertices are the covariates $(x_i)_{i=1}^N$ and where an edge between two vertices x_i and x_j exists if and only if $\|x_i - x_j\| \leq h$ for some predefined threshold h. From that the Laplacian of the graph $L : L^2(V) \to L^2(V)$, is constructed together with its spectral decomposition $L = \sum_{j=1}^N \lambda_j \langle u_j | \cdot \rangle_{L^2(V)} u_j, \quad 0 \leq \lambda_1 \leq \cdots \leq \lambda_N$ for some appropriate Euclidian space structure on V to be defined in Section 2. Dimensionality reduction is then obtained by considering the projection operator $\sum_{j=1}^J \langle u_j | \cdot \rangle_{L^2(V)} u_j$ for some truncation level J. This dimension reduction can be used in different inferential contexts

In the context of the regression problem 1, the vector $f_N = (f(x_1), \cdots, f(x_N))^T$ is modelled as an element of $\mathrm{span}(u_1, \cdots, u_J)$. This construction has been studied with variants in the definition of the graph Laplacian L, for instance by Green, Balakrishnan, and Tibshirani (2021); Shi, Balasubramanian, and Polonik (2023), where the PCR-LE estimator $\hat{f}_N$, i.e. it is the minimizer of $\|Y - f_N\|^2$ over $\Sigma^J := \mathrm{span}(u_1, \cdots, u_J)$ (in the fully supervised case $N = n$) while Sanz-Alonso and Yang (2022); Fichera, Borovitskiy, Krause, and Billard

(2024); Dunson, Wu, and Wu (2022) are using a Bayesian estimation procedure based on Gaussian processes defined via L.

In this paper we are interested in understanding the capacity of graph Laplacian regression methods of capturing both the unknown low dimensional submanifold support $\mathcal{M}$ of the design, together with the unknown smoothness β of the regression functions.

1.1 Related works :

The asymptotic behaviour of least square estimators has been studied in Green, Balakrishnan, and Tibshirani (2021); Shi, Balasubramanian, and Polonik (2023) where they obtain minimax convergence rates when the support $\mathcal{M}$ of the covariates is an open set of $\mathbb{R}^D$ and under a Sobolev regularity condition. Extension to the case where $\mathcal{M}$ is a d dimensional submanifold of $\mathbb{R}^D$, with $d < D$, is also considered in Green, Balakrishnan, and Tibshirani (2021) however only for regression functions which have Sobolev smoothness restricted to the set $\{1, 2, 3\}$.

Bayesian estimators of f based on the graph Laplacian have been considered in the literature as well for instance by Sanz-Alonso and Yang (2022); Fichera, Borovitskiy, Krause, and Billard (2024); Dunson, Wu, and Wu (2022), however the asymptotic properties in terms of posterior contraction rates around the true regression function have only been derived in some restricted cases. Dunson, Wu, and Wu (2022) show that in the case $N = n$ and when the true regression vector $f_{0,N} = (f_0(x_i))_{i=1}^N$ belongs to Σ^J for $J \lesssim \left(\frac{n}{\ln n}\right)^{\frac{d}{2q+2d}}$ for some $q \geq \frac{d}{2}$, the posterior contraction rate (with respect to the empirical L^2 norm) is upper bounded by a multiple of $\left(\frac{\ln n}{n}\right)^{\frac{d}{2q+2d}}$, under some kind of separability assumption between the eigenvalues of the Laplace-Beltrami operator of $\mathcal{M}$. Assuming that $f_{0,N} \in \Sigma^J$ avoids studying the approximation of $f_{0,N}$ by an element in Σ^J, which is a nontrivial problem where one only has access to smoothness assumptions on the function $f_0 : \mathcal{M} \to \mathbb{R}$. A step in this direction has been made by Sanz-Alonso and Yang (2022), where they established that if the number of unlabeled covariates $\{x_{n+1}, \ldots, x_N\}$ is large enough, i.e at least $N \geq n^{2d}$, then posterior contraction at a minimax optimal rate (up to logarithmic factors) of the true regression vector $f_{0,N}$ is possible, provided that its smoothness β (more precisely : f_0 belongs to the Besov space $B^{\beta}_{\infty\infty}(\mathcal{M})$, see Coulhon, Kerkyacharian, and Petrushev (2012)) satisfies $\beta > d - \frac{1}{2}$ using a graph Matérn process. Notably, as is pointed out in the paper, their assumptions rule out the fully supervised setting and every other intermediate cases. Furthermore, adaptation to the smoothness β is not discussed. In both papers Dunson, Wu, and Wu (2022) and Sanz-Alonso and Yang (2022) control the contraction of the posterior distribution on f_N by controlling the convergence of the graph Laplacian (or the associated heat kernel) to the Laplacian-Beltrami operator of $\mathcal{M}$. As we will see, it is not fundamentally needed to study the nonparametric regression problem; this is also the point of view taken by Green, Balakrishnan, and Tibshirani (2021); Shi, Balasubramanian, and Polonik (2023).

Thus, the question the approximation of $f_{0,N}$ by elements of Σ^J for $J = o(N)$, when f_0 is a smooth function on an unknown manifold $\mathcal{M}$ remains unanswered. In this paper we therefore aim to bridge this gap in the literature.

1.2 Our contributions :

In this paper, our main aim is to study posterior contraction rates for priors based on the graph Laplacian for the recovery of the regression vector $f_{0,N}$. To do that we first show that any β Hölder function on $\mathcal{M}$, $\beta > 0$ can be well approximated (in some sense to be made precise later in the paper) by a function in Σ^J for $J \geq n^{d/(2\beta+d)}$ (up to $\log n$ terms) , where d is the intrinsic dimension of the submanifold $\mathcal{M}$. To do that we use a novel approximation argument by first constructing an approximation of $f_{0,N}$ by a vector of the form $e^{-tL/h^2} f_t$ for some well chosen $t > 0$ and $f_t \in \mathbb{R}^V$. Thanks to the exponential factor we can then control well the difference between $e^{-tL/h^2} f_t$ and its projection on Σ^J. This approximation argument is valid for any $\beta > 0$. This result has an interest in its own right; in particular we apply it to the frequentist PCR-LE estimator.

We also propose general classes of priors on f_N supported on Σ^J for which we characterize the posterior contraction rates, both for the empirical loss $\|f - f_0\|_n^2 = \sum_{i=1}^n (f(x_i) - f_0(x_i))^2/n$ and for the global loss $\|f - f_0\|_N^2 = \sum_{i=1}^N (f(x_i) - f_0(x_i))^2/N$, as functions of the truncation level J and the connectivity parameter h.

Finally by considering hyperpriors on h and J we propose a novel class of priors that achieve minimax adaptive posterior contraction rates (up to $\ln n$ terms), both under the $\|\cdot\|_n$ and the $\|\cdot\|_N$ (pseudo-)distances.

Hence, contrarywise to Sanz-Alonso and Yang (2022) we do not need a very large number of unlabelled covariates which were used to ensure an accurate discrete to continuum approximation of the graph Laplacian eigenpairs and compared to Green, Balakrishnan, and Tibshirani (2021); Shi, Balasubramanian, and Polonik (2023); Sanz-Alonso and Yang (2022) we significantly weaken the smoothness assumptions on f_0 (our assumptions actually match those considered in the continuous setting (see van der Vaart and van Zanten, 2011; Rosa et al., 2023). Also note that we do not require separability assumptions on the eigenvalues of the Laplace-Beltrami operator on $\mathcal{M}$, as opposed to Dunson, Wu, and Wu (2022).

1.3 Outline :

We start by describing our geometrical and statistical model as well as our notations in Section 2. We then describe our priors and the associated posterior contraction results in Section 3; with non adaptive results in Section 3.2.1 and adaptive results in Section 3.2.2. Convergence rates for the PCR-LE estimator are provided in Section 3.2.1. The approximation theory for $f_{0,N}$ by Σ^J is provided in Section 3.3. Main proofs are provided in Section 4. Finally, the appendix contains additional proofs.

2 Model and notations

In this paper we consider the semi-supervised nonlinear regression model where we observe labelled data $(y_i, x_i)_{i \leq n}$ together with unlabelled data $(x_i)_{i=n+1}^N$ (in particular $N = N_n \geq n$ depends on n). Note that $N = n$ corresponds to the supervised case. In this work we make the following growth assumption on N : there exists $b > 0$ such that $N_n \leq n^b$ for all n. Note that in particular we always have the inequality $\ln n \leq \ln N \leq b \ln n$. The nonparametric regression model is then

$$\begin{aligned} &y_i = f_0(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}\left(0, \sigma^2\right), \quad i = 1, \ldots, n, \\ &x_i \overset{i.i.d}{\sim} \mu_0, \quad i = 1, \ldots, N \end{aligned} \tag{2}$$

where $\sigma > 0$, $f_0 \in \mathcal{C}^{\beta}(\mathcal{M})$ with $\mathcal{M}$ an unknown compact connected submanifold of regularity $\alpha \geq (\beta + 3) \vee 6$ of $\mathbb{R}^D$ (see appendix A for a precise definition of α and the space $\mathcal{C}^{\beta}(\mathcal{M})$) and μ_0 is a probability distribution on $\mathcal{M}$ (the connectedness assumption could actually be dropped by working on each connected component separately). We assume that μ_0 is absolutely continuous with respect to the volume measure on $\mathcal{M}$, with density p_0 and $f_0, \mathcal{M}, p_0$ are unknown quantities. Without loss of generality we consider the case where σ is known, since the case of unknown σ can be easily derived from the results here, using the approach of Naulet and Barat (2018). Moreover, it would also be possible to handle non Gaussian errors as done by Ghosal and van der Vaart (2017, chap. 8.5.2).

We consider an unweighted random geometric graph $G = G^{(h)} = \left(V, E^{(h)}\right)$ based on $x_{1:N} = (x_i)_{i=1}^N$ generated as follows: the vertex set is $V = \{x_1, \ldots, x_N\}$ and the edges are defined by $E_{ij} = E_{ij}^{(h)} = 1$ if and only if $\|x_i - x_j\| < h$, for some $h > 0$, where $\|\cdot\|$ denotes the Euclidean distance in $\mathbb{R}^D$. Hereafter we denote $x_i \sim x_j \iff E_{ij} = 1$.

We define the *normalized graph Laplacian*

$$L = L^{(h)} = I - D^{-1} A$$

where $D = D^{(h)} = diag(\mu)$ is the *degrees matrix* $D_{xy} = \delta_x^y \mu_x$,

$$\mu_x = \mu_x^{(h)} = \# \{i : \|x - x_i\| < h\}$$

and $A = A^{(h)}$ is the *adjacency matrix* $A_{xy} = A_{xy}^{(h)} = \mu_{xy} = \mathbb{1}_{x \sim y}$. We also define the *normalized degree measure* $\nu = \nu^{(h)} = \frac{\mu}{\mu(V)}$. The graph Laplacian L is a nonnegative self-adjoint operator from $L^2(\nu)$ to itself, where the inner product is given by

$$\forall f, g \in L^2(\nu), \langle f | g \rangle_{L^2(\nu)} = \sum_{y \in V} f(y) g(y) \nu_y$$

and L is given explicitely by the formula

$$\forall f \in \mathbb{R}^V, (Lf)(x) = \frac{1}{\mu_x} \sum_{y \sim x} (f(x) - f(y))$$

We will often use the following identity

$$\forall f, g \in L^2(\nu), \langle f | Lg \rangle_{L^2(\nu)} = \frac{1}{2\mu(V)} \sum_{x \sim y} (f(x) - f(y)) (g(x) - g(y))$$

The negative graph Laplacian $-L$ satisfies

1. $\forall x \in V, -L_{xx} = \frac{\mu_{xy}}{\mu_x} - 1 \leq 0$

2. $\forall x \neq y \in V, -L_{xy} = \frac{\mu_{xy}}{\mu_x} \geq 0$

3. $\forall x \in V, \sum_{y \in V} -L_{xy} = \sum_{y \in V} \frac{\mu_{xy}}{\mu_x} - 1 = 0$

As a consequence e^{-tL} is a row stochastic matrix for any $t \geq 0$ (see Norris, 1997, chap. 2). It is the transition matrix in time t of the continuous time Markov chain $\left(\tilde{W}_t\right)_{t \geq 0}$ going from vertex x to vertex y with rate $\frac{\mu_{xy}}{\mu_x}$. This chain has a constant unit speed and a jump matrix given by $D^{-1}A$: equivalently $\tilde{W}_t = Y_{N_t}$ where N is a Poisson process with unit rate and $(Y_n)_{n \geq 0}$ is an independent Markov chain with transition matrix $D^{-1}A$. We will denote by $\mathbb{P}_x^{(h)}$ the probability distribution of the corresponding Markov chains (in either discrete or continuous times) starting from $x \in V$. For convenience in what follows we will define $\mathcal{L} = h^{-2}L$ and $W_t = \tilde{W}_{t/h^2}$.

It is shown for instance by García Trillos, Gerlach, Hein, and Slepčev (2020); Calder and Trillos (2020) that under appropriate conditions the graph Laplacian $\mathcal{L}$ converges in some sense to the true Laplace-Beltrami operator Δ of the limiting manifold $\mathcal{M}$ (up to a proportionality constant). Since W is a continuous time Markov chain with transition matrix $e^{-t\mathcal{L}}$, this shows that W can actually be seen in a way as a numerical approximation of a Brownian motion on $\mathcal{M}$, i.e an $\mathcal{M}-$valued continuous time Markov process with infinitesimal generator $-\Delta$.

For each $x \in V, \quad t \geq 0$ we define $p_t(x,\cdot) = p_t^{(h)}(x,\cdot)$ as the density of $e^{-t\mathcal{L}}(x,\cdot)$ with respect to ν

$$\forall y \in V, \quad p_t(x,y) = \frac{e^{-t\mathcal{L}}(x,y)}{\nu_y}$$

and we call the matrix p_t the *heat kernel of $\mathcal{L}$ with respect to ν*. If $(\lambda_j, u_j)_{j=1}^N = \left(\lambda_j^{(h)}, u_j^{(h)}\right)_{j=1}^N$ is an orthonormal eigendecomposition of $\mathcal{L}$ in $L^2(\nu)$, i.e.

$$\forall j, l \in \{1, \ldots, N\}, \quad \mathcal{L}u_j = \lambda_j u_j, \quad \langle u_j | u_l \rangle_\nu = \delta_j^l$$

(the existence of which is guaranteed by the finite dimensional spectral theorem) then we have

$$\forall t \geq 0, \quad e^{-t\mathcal{L}} = \sum_{j=1}^N e^{-t\lambda_j} \langle u_j | \cdot \rangle_\nu u_j = \frac{1}{\mu(V)} \sum_{j=1}^N e^{-t\lambda_j} u_j u_j^T$$

Hence

$$\forall t \geq 0, \quad x, y \in V, p_t(x,y) = \sum_{j=1}^N e^{-t\lambda_j} u_j(x) u_j(y)$$

In particular p_t is a symmetric matrix.

In addition to the $L^2(\nu)$ structure we will also use the uniform norm

$$\forall f \in \mathbb{R}^V, \quad \|f\|_{L^\infty(\nu)} = \max_{1 \leq i \leq N} |f(x_i)|$$

As well as the associated operator norm

$$\|A\|_{L^\infty(\nu)} = \sup_{\|f\|_{L^\infty(\nu)} \leq 1} \|Af\|_{L^\infty(\nu)}$$

defined for any linear endomorphism $A : L^\infty(\nu) \to L^\infty(\nu)$.

2.1 Notations

We denote by $\mathbb{P}_0$ the frequentist probability distribution of $\mathbb{X}^n := (x_{1:N}, y_{1:n})$, i.e

$$\mathbb{P}_0(dx_{1:N}, dy_{1:n}) = \prod_{i=1}^{N} p_0(x_i)\mu(dx_i) \prod_{i=1}^{n} \mathcal{N}\left(y_i | f_0(x_i), \sigma^2\right) dy_i$$

as well as $P_f = \bigotimes_{i=1}^n \mathcal{N}\left(f_i, \sigma^2\right)$ for any $f \in \mathbb{R}^n$, with the abuse of notation $P_{f_0} = P_{(f_0(x_i))_{i=1}^n}$. We will also denote the probability distribution of the infinite sequences $(x_i)_{i\geq 1}, \quad (y_i)_{i\geq 1}$ by P_0^∞. We denote the geodesic metric on $\mathcal{M}$ by $\rho : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_+$. For any $x \in \mathbb{R}^D, r > 0$ we define $B_{\mathbb{R}^D}(x, r) = \left\{y \in \mathbb{R}^D : \|x - y\| < r\right\}$ the Euclidean ball and $B_\rho(x, r) = \{y \in \mathcal{M} : \rho(x, y) < r\}$ the geodesic ball. For any measurable subset A of $\mathbb{R}^d, \quad d \geq 1$, $vol(A)$ will denote its $d-$dimensional Lebesgue measure.

For any open subset $\mathcal{U}$ of $\mathbb{R}^d, \quad d, k \geq 1$ and $f : \mathcal{U} \to \mathbb{R}$ we define its $\mathcal{C}^k(\mathcal{U})$ norm by $\|f\|_{\mathcal{C}^k(\mathcal{U})} = \max_{\alpha \in \mathbb{N}^d, |\alpha| \leq k} \sup_{x \in \mathcal{U}} \left|\frac{\partial^{|\alpha|} f}{\partial x^\alpha}(x)\right|$. We also define $\mathcal{C}^\infty(\mathcal{U}) = \cap_{k\geq 0} \mathcal{C}^k(\mathcal{U})$. For non integer $\beta \in (k, k+1), \quad k \in \mathbb{N}$ we define the Hölder class $\mathcal{C}^\beta(\mathcal{U})$ class as the space of all functions $f \in \mathcal{C}^k(\mathcal{U})$ satisfying

$$\|f\|_{\mathcal{C}^\beta(\mathcal{U})} := \|f\|_{\mathcal{C}^k(\mathcal{U})} + \max_{\alpha \in \mathbb{N}^d, |\alpha| = k} \sup_{x \neq y \in \mathcal{U}} \frac{\left|\frac{\partial^{|\alpha|} f}{\partial x^\alpha} - \frac{\partial^{|\alpha|} f}{\partial x^\alpha}\right|}{\|x - y\|^{\beta - k}} < +\infty$$

If V is a finite dimensional vector space and $\mathcal{U} \subset \mathbb{R}^d, \quad d \geq 1$, we say that a mapping $f : \mathcal{U} \to V$ is of class $\mathcal{C}^\beta(\mathcal{U})$ if $T \circ f : \mathcal{U} \to \mathbb{R} \in \mathcal{C}^\beta(\mathcal{U})$ for any $T \in V^*$ (linear form on V). Equivalently, each coordinate function of f (in any choice of basis) is of class $\mathcal{C}^\beta$.

It is immediate to check that any $f \in \mathcal{C}^\beta(\mathcal{U}), \quad k < \beta \leq k+1$ has a Taylor development of the form $f(y) = \sum_{l=0}^k \frac{d^l f(x).(y-x)^l}{l!} + R_k(x, y), \quad |R_k(x, y)| \leq C_f \|x - y\|^\beta$ with $C_f = \|f\|_{\mathcal{C}^\beta(\mathcal{U})}/k!$.

We write

$$\|f - f_0\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} \left(f(x_i) - f_0(x_i)\right)^2 \text{ and } \|f - f_0\|_N^2 = \frac{1}{N} \sum_{i=1}^{N} \left(f(x_i) - f_0(x_i)\right)^2.$$

In the following the symbols $\gtrsim, \quad \lesssim$ refer to inequalities up to constants that depend only on $\mathcal{M}, \quad p_0$ and f_0. We will also write $a \asymp b$ if both $a \lesssim b$ and $b \lesssim a$ hold. We will also often write $C((\lambda)_{\lambda\in\Lambda})$ or $C_{(\lambda)_{\lambda\in\Lambda}}$ generically to denote a positive constant depending only on the parameters $\lambda \in \Lambda$ whose value can change from line to line.

3 Main results

In the nonparametric regression model 2, the goal is to estimate the regression function f_0 evaluated on the covariates $x_i, \quad i = 1, \ldots, N$. Hence from a Bayesian point of view the goal is to design a prior distribution on an element $f : V \to \mathbb{R}$ leading to good frequentist guarantees a posteriori. More precisely, following Ghosal, Ghosh, and van der Vaart (2000); Ghosal and van der Vaart (2017) our goal is to derive *posterior contraction rates* for our suggested Bayesian methodologies : we wish to identify a sequence ε_n of positive real numbers satisfying

$$\Pi\left[\|f - f_0\|_n > \varepsilon_n | \mathbb{X}^n\right] \xrightarrow[n\to\infty]{P_0^\infty} 0$$

Deriving such rates typically requires to prove preliminary approximation results which is the object of section 3.3.

3.1 Priors

As done by Sanz-Alonso and Yang (2022); Fichera, Borovitskiy, Krause, and Billard (2024) we design our priors by random basis expansion in the graph Laplacian eigenbasis. Our construction depends on 2 degrees of freedom : the number of basis functions allowed in the sum, and the graph connectivity parameter h. This leads to a first family of priors :
Prior 1 Fixed J, h case:

We consider the prior $\Pi\left[\cdot|J,h\right]$ defined by the probability law of the random vector

$$f = \sum_{j=1}^{J} Z_j u_j, \quad Z_j \overset{i.i.d}{\sim} \Psi \tag{3}$$

for some $h > 0$, $J \in \{1, \dots, N\}$ and Ψ a probability distribution having a positive and continuous density ψ on $\mathbb{R}$ satisfying $\int_{|x|>z} \psi(x)dx \leq e^{-b_1 z^{b_2}}$, for all $z \geq z_0$ for some $b_1, b_2, z_0 > 0$. In particular, Ψ can be Gaussian, leading to the construction of a Gaussian vector f on V, as done by Dunson, Wu, and Wu (2022); Sanz-Alonso and Yang (2022).

As we shall see in section 3.2.1, this prior leads to explicit posterior contraction rates which happen to be minimax optimal (up to logarithmic factors) under appropriate choice of the hyperparameters $J = J_n, h = h_n$ that depends on the unknown regularity β of f_0 together with the dimension d of $\mathcal{M}$. In this sense the proposed prior is *non adaptive*.

It is then of interest to choose the hyperparameters J, h in a data driven way. From the Bayesian point of view this can be done by putting an additional prior layer over J and h. For practical reasons, since for each value of h a new Laplacian $\mathcal{L}$ needs to be computed, together with its eigenpairs (u_j, λ_j) we only consider discrete priors on h, i.e. we enforce $h \in \mathcal{H}$ where $\mathcal{H}$ is a discrete subset of $\mathbb{R}_+$, and $\pi_h(\cdot|J)$ denotes the prior probability mass function of h given J.
Prior 2 Adaptive prior:

We consider the prior Π defined by the probability law of the random vector

$$J \sim \pi_J, \quad h|J \sim \pi_h(\cdot|J), \quad f|J,h \sim \Pi\left[\cdot|J,h\right], \quad \text{denote } \mathcal{H}_J = \mathrm{supp}(\pi_h(\cdot|J)) \subset \mathcal{H}, \tag{4}$$

where π_h is a probability distribution supported on $\mathcal{H}_J$ (that may depends on J) and π_J is a probability distribution on $\mathbb{N}^*$ satisfying

$$\forall j \geq j_0, \quad e^{-a_1 j L_j} \leq \pi_J(j) \leq e^{a_2 j L_j} \text{ where } L_j = 1 \text{ or } L_j = \ln j, \text{ and } \#\mathcal{H}_J \leq K_1 e^{K_2 J \ln n}$$

for some constants $a_1, a_2, K_1, K_2 > 0$.

Note that the joint prior on (J, h) can be of very different forms. For instance if $J \sim \mathcal{P}(\lambda)$ then the condition on π_J holds with $L_j = \log j$ while if π_J is a Geometric distribution, then it holds with $L_j = 1$. Also we can choose for instance the support of $\pi_h(\cdot|J)$ to be $\mathcal{H}_J = \left\{\frac{J^{-1/d}}{\ln^\kappa N}\right\}$ for some $\kappa > 2$ and $\mathcal{H} = \cup_J \mathcal{H}_J$, where here the dimension d of $\mathcal{M}$ is assumed to be known. We will see that this leads to adaptive nearly minimax estimation rates. Alternatively and for more flexibility we can choose $\mathcal{H} = \left\{h_l := 2^l h_* : l = 0, \ldots, L\right\}$ with $h_* = \left(\frac{\ln n}{n}\right)^{1/d}, \quad \kappa > 2$ and $L \in \mathbb{N}$ such that $2^L h_* \leq 1 < 2^{L+1} h_*$, together with

$$\pi_h(h_l) \geq b_1 e^{-b_2 h_l^{-d}} \quad \forall h \in \mathcal{H}$$

for some constants $b_1, b_2 > 0$. We will detail the precise assumptions on $\Pi_J, \quad \pi_h(\cdot|J)$ in Theorems 6 & 7.

Remark 1. *Both **prior 1** and **prior 2**, depend on the covariates $x_{1:N}$ and possibly on the sample size n, but we keep the notation Π (and not $\Pi_{n,x}$) to avoid cumbersome notations.*

Having defined the two types of priors we now study their posterior contraction properties.

3.2 Posterior contraction rates

We present here results on posterior contraction for **prior 1** and **prior 2**. We first present the non adaptive posterior contraction rate as it highlights the role of J and h and then we present the posterior contraction rate derived from prior 2. The proof is based on the general prior mass and testing theorem (see Ghosal and van der Vaart, 2007, 2017; Rousseau, 2016). A key aspect of the proof is a new scheme to approximate f_0 by functions in the span of $u_1, \cdots, u_J$. This allows in particular to extend the results of Green, Balakrishnan, and Tibshirani (2021) to non linear manifolds when the regularity of f_0 not restricted to $\{1, 2, 3\}$. The approximation result is interesting in its own right and it allows also to derive minimax convergence rates for the frequentist estimator of Green, Balakrishnan, and Tibshirani (2021); hence it is presented in Section 3.3 and the convergence rate of the frequentist estimator is provided in Section 3.2.1. Moreover, the proof is very different from Green, Balakrishnan, and Tibshirani (2021) and is valid even when $N > n$.

3.2.1 Non adaptive rates: Prior 1 and PCR-LE estimator

Throughout this section we assume that the chosen connectivity parameter h_n is not too small :

Assumption 2. *The connectivity parameter h_n satisfies $h_n \to 0$ and $\frac{N h_n^d}{\ln N} \to \infty$.*

Our first main result identifies a posterior contraction rate associated with prior 3 if h satisfies assumption 2 and J is not too small. In particular, we recover minimax optimal rates up to logarithmic factors for appropriate choices of parameters :

Theorem 3. *Let (J_n, h_n) be a sequence of truncation and connectivity parameters satisfying $J_n \in \{1, \ldots, N\}, \quad J_n \geq \ln^\kappa N, \quad \kappa > d$ and h_n satisfying assumption 2. Consider a prior on f belonging to the **Prior 1** family. Then there exists $C > 0$ such that*

$$\Pi\left[\|f - f_0\|_n > C\varepsilon_n | \mathbb{X}^n, J_n, h_n\right] \xrightarrow[n\to\infty]{P_0^\infty} 0$$

where

$$\varepsilon_n = \varepsilon_n(J_n, h_n) = \sqrt{\frac{J_n \ln N}{n}} + \ln N^{\lceil \beta/2 \rceil} \max\left(1, \frac{\ln N}{J_n^{2/d} h_n^2}\right)^{\lceil \beta/2 \rceil} \left(h_n^\beta + \mathbb{1}_{\beta>1}\left(\frac{\ln N}{N h_n^d}\right)^{1/2} h_n\right) \tag{5}$$

In particular, for any $\tau > d/2$ *and*

$$h_n = n^{-\frac{1}{2\beta+d}} (\ln n)^{-\frac{1-\tau-2(1+2\tau/d)\lceil \beta/2 \rceil}{2\beta+d}}, \quad J_n = \frac{h_n^{-d}}{\ln^\tau N}$$

then

$$\varepsilon_n \asymp (\ln n)^{\frac{(2\tau+d)\lceil \beta/2 \rceil + (1-\tau)\beta}{2\beta+d}} n^{-\frac{\beta}{2\beta+d}}$$

The proof is given in section D.1. Theorem 3 highlights the trade - off between the complexity of the prior (term $\sqrt{J \ln N/n}$) and the approximation error bounded by

$$\ln N^{\lceil \beta/2 \rceil} \max\left(1, \frac{\ln N}{J_n^{2/d} h_n^2}\right)^{\lceil \beta/2 \rceil} \left[h_n^\beta + \mathbb{1}_{\beta>1} h_n \left(\frac{\ln N}{N h_n^d}\right)^{1/2}\right].$$

The first term of the approximation error has a bias flavour while the second, which appears only when $\beta > 1$ comes, roughly speaking, from stochastic deviations of $\mathcal{L} f_0 - \mathbb{E}(\mathcal{L} f_0)$, see Lemma 15. More discussion is given after Theorem 11.

Notice that theorem 3 only guarantees a posterior contraction with respect to the empirical L^2 norm $\|f - f_0\|_n$. Thus, it is not clear whether or not the posterior distribution correctly extrapolates outside of the design $\{x_1, \ldots, x_n\}$ in order to estimate the values $\{f_0(x_{n+1}), \ldots, f_0(x_N)\}$. This is the object of theorem 4 below, which requires stronger assumptions on J_n, h_n and β.

Theorem 4. *Assume that* $\beta > d/2$ *and let* h_n *satisfying* $h_n \geq (\ln N)^s n^{-\frac{1}{2\beta+d}}$ *for some* $s \in \mathbb{R}$ *as well as* $J = J_n \in \{1, \ldots, N\}$, $\frac{h_n^{-d}}{\ln^{\tau_1} N} \leq J_n \leq \frac{h_n^{-d}}{\ln^{\tau_2} N}$, $\tau_1 \geq \tau_2 > 2d$. *Then there exists* $C > 0$ *such that*

$$\Pi\left[\|f - f_0\|_N > C\varepsilon_n | \mathbb{X}^n, J_n, h_n\right] \xrightarrow[n\to\infty]{P_0^\infty} 0$$

where

$$\varepsilon_n = \sqrt{\frac{J_n \ln N}{n}} + \left(\frac{J_n^{-2/d} \ln N}{h_n^2}\right)^{\lceil \beta/2 \rceil} \left(h_n^\beta + \mathbb{1}_{\beta>1}\left(\frac{\ln N}{N h_n^d}\right)^{1/2} h_n\right)$$

In particular, choosing $J_n = \frac{h_n^{-d}}{\ln^\tau N}$, $\tau > 2d$ *and* $h_n = n^{-\frac{1}{2\beta+d}} (\ln N)^{-\frac{1-\tau-2\lceil \beta/2 \rceil(1+2\tau/d)}{2\beta+d}}$ *leads to a posterior contraction rate of order*

$$\varepsilon_n \asymp (\ln N)^{\frac{2\lceil \beta/2 \rceil (2\tau+d) + 2\beta(1-\tau)}{2\beta+d}} n^{-\frac{\beta}{2\beta+d}}$$

The proof is provided in Section 4.4 and relies on Theorem 3 together with a concentration inequality to upper bound $\|f - f_0\|_N$ by (a multiple of) $\|f - f_0\|_n$.

The main limitation of **prior 1**, with deterministic J, h is that the choice of J, h crucially impacts the asymptotic behaviour of the posterior distribution. In particular to obtain optimal behaviour (with respect to $\|\cdot\|_n$ or $\|\cdot\|_N$) one needs to choose them as a function of β and d. While there exist simple and consistent estimators of d (see Berenfeld and Hoffmann, 2021), it is very difficult to estimate β.

It is interesting to compare our results with those of Sanz-Alonso and Yang (2022). In this work a Matérn type Gaussian process of the form $f = \sum_{j=1}^{J_N} (1+\lambda_j)^{-\frac{\beta+d/2}{2}} Z_j u_j$, $Z_j \overset{i.i.d}{\sim} \mathcal{N}(0,1)$ (also for some deterministic values of J_n, h_n) is used in place of prior 3. In Sanz-Alonso and Yang (2022) the author obtain the minimax rate of convergence with respect to the empirical L^2 norm $\|\cdot\|_n$ by choosing J, N, h accordingly but under a more restrictive assumption on N, β, implying in particular $N >> n^{2d}$ and $\beta > d - 1/2$. Also, their proof technique is very different from ours as it is based on the convergence of the (discrete) Matérn Gaussian process prior to its continuous counterpart (associated to the Laplace-Beltrami operator on $\mathcal{M}$).

In the fully supervised $N = n$ case, our result can also be compared to the (frequentist) PCR-LE estimator of Green, Balakrishnan, and Tibshirani (2021). To be completely clear the estimator proposed in Green, Balakrishnan, and Tibshirani (2021) relies on the unnormalized graph Laplacian rather than the normalized one, but we believe our proof could be applied for the unnormalized graph Laplacian as well. The PCR-LE estimator is defined as

$$\hat{f} = \operatorname{argmin}_{f \in \operatorname{span}(u_1, \cdots, u_J)} \|Y - f\|_2, \quad Y = (y_1, \cdots, y_n), \quad f = (f(x_1), \cdots, f(x_n))$$

Green, Balakrishnan, and Tibshirani (2021) prove that when f_0 belongs to a $\beta-$ Sobolev class with $\beta \in \{1, 2, 3\}$ (where the Sobolev space is defined using the Laplacian operator weighted according to the covariates density), then $\hat{f}$ converges to f_0 at the rate $n^{-\beta/(2\beta+d)}$, if J and h are chosen accordingly. In comparison we obtain a near minimax rate for any $\beta-$ Hölder regularity assumption on $f_0, \quad \beta > 0$ (for the $\|\cdot\|_n$ loss) or $\beta > d/2$ (for the $\|\cdot\|_N$ loss). It should be noted that the restriction $\beta > d/2$ also appears in the continuous case when extending outside of the labelled points (see van der Vaart and van Zanten, 2011; Rosa et al., 2023).

As mentioned earlier, a key component of the proof of Theorems 3 & 4 is to show that f_0 can be well approximated by elements in the linear span of $(u_1, \cdots, u_J)$, see Theorem 11 . A consequence of Theorem 11 is thus that the PCR-LE estimator of Green, Balakrishnan, and Tibshirani (2021) achieves near minimax convergence rate over Hölder classes of any order β, thus extending the results of Green, Balakrishnan, and Tibshirani (2021).

Theorem 5. *Performance of PCR-LE for the random walk graph Laplacian Assume that $N = n$ and that the sequence (J_n, h_n) is such that $J_n \in \{1, \dots, n\}, \quad J_n \geq \ln^\kappa N, \quad \kappa > d$ and that $h_n > 0$ satisfies assumption 2. Define*

$$\hat{f} = \sum_{j=1}^{J_n} \langle u_j | Y \rangle_{L^2(\nu)} u_j$$

Then for any $q \geq 0$ there exists $C > 0$ such that

$$\mathbb{E}_0 \left[\left\| \hat{f} - f_0 \right\|_n^q \right] \leq C \varepsilon_n^q$$

where

$$\varepsilon_n = \sqrt{\frac{J_n}{n}} + (\ln n)^{\lceil \beta/2 \rceil} \max\left(1, \left(\frac{J_n^{-2/d} \ln n}{h_n^2}\right)\right)^{\lceil \beta/2 \rceil} \left(h_n^\beta + \mathbb{1}_{\beta>1} \left(\frac{\ln n}{n h_n^d}\right)^{1/2} h_n\right)$$

In particular for the choice $J_n = \frac{h_n^{-d}}{\ln^\tau n}, \quad \tau > d/2, \quad h_n = n^{-\frac{1}{2\beta+d}} (\ln n)^\rho, \rho \in \mathbb{R}$ *then for any* $q > 0$ *there exists* $C > 0$ *such that*

$$\mathbb{E}_0\left[\left\|\hat{f} - f_0\right\|_n^q\right] \leq C\varepsilon_n^q$$

where

$$\varepsilon_n = n^{-\frac{\beta}{2\beta+d}} (\ln n)^{(2\lceil \beta/2 \rceil (1+\tau/d)+\beta\rho) \vee (-\rho d/2 - \tau/2)}$$

The proof of Theorem 5 is given in Section C.3. Corollary 5 hence shows that in the case $N = n$ the PCR-LE estimator of Green, Balakrishnan, and Tibshirani (2021) can achieve optimal rates for particular choice of hyperparameters (up to logarithmic factors) for any smoothness index $\beta > 0$ if one is willing to work on the Hölder spaces scale, and not just for $\beta-$Sobolev functions, $\beta \in \{1, 2, 3\}$: the limitations of the manifold adaptivity was therefore only an artefact of the proof, as the non-negligible discrepancy between geodesic and Euclidean distances can be circumvented by considering a different approximation technique. Extending 5 to an adaptive estimator by Lepski's method as done by Shi, Balasubramanian, and Polonik (2023) in the Euclidean domain could also be of interest, which then would provide a frequentist alternative to our adaptive Bayesian prior 4.

In the case $N > n$, our theorems 11 & 12 can also potentially be used to design frequentist estimators in alternative to our priors 3 & 4. Indeed, as an example Theorem 11 can be reformulated as a high probability upper bound on $\left\|f_0 - Q_t^{(k)} f_0\right\|_{L^\infty(\nu)}$ where $Q_t^{(k)}$ is the kernel operator

$$\forall f : V \to \mathbb{R}, \left(Q_t^{(k)} f\right)(x) = \sum_{y \in V} \underbrace{\sum_{l=0}^{k} \frac{(t\mathcal{L})^l e^{-t\mathcal{L}}(x,y)}{l!}}_{=:Q_t^{(k)}(x,y)\nu_y} f(y)$$

This can be generalized by

$$\chi_t^{(k)}(x,y) = \sum_{l=0}^{k} \frac{(-t\mathcal{L})^l \chi^{(l)}(t\mathcal{L})(x,y)}{l!} = \sum_{l=0}^{k} \frac{1}{l!} \sum_{j=1}^{N} (-t\lambda_j)^l \chi^{(l)}(t\lambda_j) u_j(x) u_j(y)$$

for an arbitrary $\chi \in \mathcal{C}^k(\mathbb{R}_+, \mathbb{R})$ with $\chi(0) = 1$ which motivates the definition of kernel regression estimators of the form $\hat{f} = \sum_{i=1}^n \chi_t^{(k)}(\cdot, x_i) y_i \nu_{x_i}$: this is always tractable even when $N > n$ and analyzing the asymptotic properties of such estimators (even on other graph models) is therefore an interesting research direction.

3.2.2 Results for the adaptive priors

Having proved Theorems 3 & 4 it is straightforward to derive adaptive posterior contraction rates under the **prior 2**. This results in a Bayesian method able to achieve minimax optimal contraction rates (up to logarithmic factors) with a data driven choice for J, h.

Theorem 6. *Consider the prior* Π *on* f *belonging to the class* ***Prior 2*** *as defined by* (4) *and assume that for some* $K_1, K_2 > 0$ *and some* $j_0, a_1, a_2 > 0$

$$\forall j \geq j_0, \quad e^{-a_1 j L_j} \leq \pi_J(j) \leq e^{-a_2 j L_j}, \quad \textit{where} \quad L_j = 1 \quad \textit{or} \quad L_j = \ln j, \quad \textit{and } \#\mathcal{H}_J \leq K_1 e^{K_2 J \ln n} \tag{6}$$

Assume in addition that for some $\tau > d/2, \quad J_0, h_0 > 0$ *and*

$$J_n = J_0 \left(\ln N\right)^{\frac{2d(1+2\tau/d)\lceil \beta/2 \rceil}{2\beta+d}} n^{\frac{d}{2\beta+d}}$$

$h_n \in \mathcal{H}_{J_n} \cap \left[\frac{h_0 J_n^{-1/d}}{2\ln^{\tau/d} N}, \frac{h_0 J_n^{-1/d}}{\ln^{\tau/d} N}\right]$ *and* $c > 0$ *such that*

$$\pi_h(h_n|J_n) \geq e^{-ch_n^{-d}}, \tag{7}$$

Then there exists $C > 0$ *such that*

$$\Pi\left[\|f - f_0\|_n > C\varepsilon_n | \mathbb{X}^n\right] \xrightarrow[n\to\infty]{P_0^\infty} 0, \qquad \textit{with} \quad \varepsilon_n = n^{-\frac{\beta}{2\beta+d}} \left(\ln N\right)^{\frac{(2\tau+d)\lceil \beta/2 \rceil - \beta(\tau+2\beta/d)}{2\beta+d}}.$$

For clarity, we remind the reader that in Equation 7, $\pi_h\left(h_n|J_n\right)$ denotes the prior probability mass function of h given $J = J_n$ evaluated at h_n.

The proof of Theorem 6 is given in Section D.2. Notice that even if some particular values of J_n, h_n are part of a condition that **prior 2** (4), must satisfy theoretically, the precise values of J_n, h_n are not part of the definition of the prior itself.

As discussed above in section 4, there are many priors on J, h which satisfy (7). Consider two cases of particular interest

- If J follows a Poisson or a Geometric prior and, given J, $\quad h = \frac{J^{-1/d}}{\ln^{\tau/d} N}, \quad \tau > d/2$ then condition (7) is fulfilled (with $\mathcal{H}_J = \left\{\frac{J^{-1/d}}{\ln^{\tau/d} n}\right\}$).

- A more flexible alternative is to choose a discrete prior on $\mathcal{H} = \{2^l h_* : l = 0, \cdots, L\}$ for $h_* = \left(\frac{\ln n}{n}\right)^{1/D}$ and L such that $2^L h_* \leq 1 < 2^{L+1} h_*$, together with a probability mass function of the order of $e^{-\lambda h^{-1}} h^a$ for arbitrary $\lambda > 0$ and $a \geq 0$. For instance, such a prior can be constructed by first considering $\tilde{h} \sim IG(a, \lambda)$, an inverse Gamma random variable and then defining $h = \sum_{l=0}^L 2^l h_* \mathbb{1}_{I_l}(\tilde{h}), \quad I_0 = [0, h_*], \quad I_l =]2^l h_*, 2^{l+1} h_*]$ for $1 \leq l < L$ and $I_L =]2^L h_*, +\infty[$. In particular this prior is fully adaptive as it depends neither on β nor on d.

Similarly to the non adaptive prior, it is possible to obtain a posterior contraction rate in terms of the loss $\|\cdot\|_N$ from the result of Theorem 6, under additional assumptions on β and the prior on J, h.

Theorem 7. *Under the same conditions as in theorem 6 and if in addition we have* $\mathcal{H}_J \subset [h_*, \frac{J^{-1/d}}{\ln^{\tau/d} N}]$ $\Pi-$*almost surely for some* $\tau > 2d$, $h_* > 0$ *satisfying assumption 2 and if* $\beta > d/2$ *then there exists* $M > 0$ *such that*

$$\Pi\left[\|f - f_0\|_N > M\varepsilon_n | \mathbb{X}^n\right] \xrightarrow[n\to\infty]{P_0^\infty} 0$$

for the rate

$$\varepsilon_n = n^{-\frac{\beta}{2\beta+d}} (\ln N)^{\frac{(2\tau+d)\lceil \beta/2 \rceil - \beta(\tau + 2\beta/d)}{2\beta+d}}$$

Remark 8. *Theorem 7 thus shows that we can construct (near) minimax adaptive Bayesian procedures based on graph Laplacian decompositions, where the adaptivity is related to the (typically unknown) smoothness* β. *Similarly to theorem 6, many priors satisfy the assumptions of Theorem 7, for instance a Poisson or geometric prior on* J *can be chosen together with* $h = \frac{J^{-1/d}}{\ln^{\tau/d} n}$, $\quad \tau > d/2$. *Alternatively, as discussed below Theorem 6, we can choose a discrete prior on* $\mathcal{H}_J := \{2^l h_* : l = 0, \cdots, L\}$ *with* $h_* = m_n \left(\frac{\ln n}{n}\right)^{1/D}$ *and* L *such that* $2^L h_* \leq \frac{J^{-1/d}}{\ln^{\tau/d} N} < 2^{L+1} h_*$, *together with a probability mass function of the form of* $e^{-\lambda h^{-d}} h^a$ *for arbitrary* $m_n \to \infty, \lambda > 0$ *and* $a \geq 0$.

Compared to Theorem 6, we require $\mathcal{H}_J \subset \left[h_*, \frac{J^{-1/d}}{\ln^{\tau/d} N}\right]$ *for some* h_* *satisfying assumption 2 in order to apply Lemma 34 a posteriori (and importantly with constants that do not depend on* h, J*), see the detailed proof in Section 4.4. In fact the condition could be weakened to* $\mathcal{H}_J \subset \left[h_*, \frac{J^{-1/d}}{\ln^{\tau/d} N}\right]$ *on a set of posterior probability tending to* 1, *but this is not necessary for the examples of priors we provide.*

Remark 9. *The same condition* $\mathcal{H}_J \subset \left[h_*, \frac{J^{-1/d}}{\ln^{\tau/d} N}\right]$ *prevents us to design priors adaptive to* d, *contrariwise to Theorem 6. This condition is crucial in our setting in order to use Lemma 34. Whether or not such a restriction is necessary or is an artefact of our proof technique is therefore an interesting question. To design a procedure that is adaptive both to* β *and* d *one could plug in a consistent estimator of* d *(see Berenfeld and Hoffmann, 2021)—the procedure then becomes an empirical Bayes procedure)—and our theoretical results remain unchanged since the estimator only depends on* x *and not on* y.

Remark 10. *Theorems 6 & 7 hold for discrete priors on* h, *however we believe that this is not a restriction since a continuous prior would imply that for every value of* h, *one would need to compute the eigenpairs* (u_j, λ_j) *of the Laplacian corresponding to* h. *Obtaining a result which would hold for a continuous prior on* h, *or for empirical Bayes procedure, where optimization with respect to* h *is performed, is not straightforward since it would necessitate to prove the concentration lemma 15 uniformly over* $h \in (h_0, h_1)\varepsilon_n^{1/\beta}$.

Notice that our posterior contraction rates in Theorems 3 4 6 5 & 7 are all impacted by a power of $\ln N$: this can be large, for instance if N grows exponentially with n. However under the mild assumption $N \leq n^B, B > 0$ this factor become a power of $\ln n$ which is typically considered small from an asymptotic viewpoint, ignoring the multiplicative constant. Extending our results to the case $N >> n^B$ for all $B > 0$ would require more work, but we believe that in this case the proof strategy of Sanz-Alonso and Yang (2022) could be adapted.

3.3 Approximation result

All three theorems 3, 4 5, 7 & 6 are proved using approximation results of f_0 in the graph spectral domain, but since the smoothness assumption $f_0 \in \mathcal{C}^\beta(\mathcal{M})$ is only formulated at the continuous level this is a non trivial task. To prove that f_0 can be well approximated by an element of $\mathrm{span}(u_1, \cdots, u_J)$ we first find an approximation of f_0 (as a function on V) by an element of the form $e^{-t\mathcal{L}} f_t$ for some explicit f_t, with high probability with respect to the sampling distribution of the covariates $x_{1:N}$. Then we show that the projection of $e^{-t\mathcal{L}} f_t$ onto $\mathrm{span}(u_1, \cdots, u_J)$ is close to $e^{-t\mathcal{L}} f_t$ and thus to f_0.

Theorem 11. *Let $\beta > 0, \quad f_0 \in \mathcal{C}^\beta(\mathcal{M})$, assume that h_n satisfies assumption 2 and define for $t > 0$*

$$f_t = \sum_{l=0}^{k} \frac{1}{l!} (t\mathcal{L})^l f_0, \quad f_t : V \to \mathbb{R},$$

where $k = \lceil \beta/2 \rceil - 1$. Then

1. *If $0 < \beta \le 1$, there exists a constant $C(\beta, f_0)$ we have*

$$\left\| f_0 - e^{-t\mathcal{L}} f_t \right\|_{L^\infty(\nu)} \le C(\beta, f_0)\, t h_n^{\beta-2}$$

2. *if $\beta > 1, p_0 \in \mathcal{C}^{\beta-1}(\mathcal{M})$ then for any $H > 0$ there exists a constant $C(H, \beta, f_0, p_0, \mathcal{M})$, such that*

$$\mathbb{P}_0\left(\left\| f_0 - e^{-t\mathcal{L}} f_0 \right\|_{L^\infty(\nu)} > C(H, \beta, f_0, p_0, \mathcal{M}) \left(\frac{t}{h_n^2}\right)^{k+1} \left[h_n^\beta + \left(\frac{\ln N}{N h_n^d}\right)^{1/2} h_n \right] \right) \le N^{-H}$$

The proof of Theorem 11 is given in Section 4.1.

Interestingly, when $\beta > 1$,

$$h_n^\beta \ge h_n \sqrt{\frac{\ln N}{N h_n^d}}, \quad \Leftrightarrow h_n \ge \left(\frac{\ln N}{N}\right)^{\frac{1}{2(\beta-1)+d}}$$

which is—up to the $\ln N$ term—the condition required by Theorem 6 in Green, Balakrishnan, and Tibshirani (2021) in the case $N = n$. Although the proof techniques are very different, in both cases this condition is used to bound deviations of $\mathcal{L}f$ around its mean.

To prove that f_0 can be well approximated by a function of the form $f = \sum_{j=1}^{J_n} z_j u_j$ for some choice of truncation J_n, we now show that $e^{-t\mathcal{L}} f_t$ is close to $p_{J_n}(e^{-t\mathcal{L}} f_t)$, its $L^2(\nu)$ projection onto $\mathrm{span}(u_1, \cdots, u_{J_n})$. Since we are able to lower bound the eigenvalues of $\mathcal{L}$ with high probability, thanks to the exponential factor $e^{-t\mathcal{L}}$ this is actually straightforward.

Theorem 12. *Let $\beta > 0, \quad f_0 \in \mathcal{C}^\beta(\mathcal{M}), \quad h_n$ satisfying assumption 2 and $J_n \in \{1, \ldots, N\}$ satisfying $J_n \ge \ln^\kappa N, \quad \kappa > d$. Let $f_t : V \to \mathbb{R}$ be defined by*

$$f_t = \sum_{l=0}^{k} \frac{1}{l!} (t\mathcal{L})^l f_0,$$

where $k = \lceil \beta/2 \rceil - 1$. *Then there exists* $c > 0$ *such that with* $t = c\lambda_{J_n}^{-1} \ln N$ *and* $\tilde{\varepsilon}_n(J_n, h_n)$ *the rate*

$$\tilde{\varepsilon}_n(J_n, h_n) = (\ln N)^{\lceil \beta/2 \rceil} \max\left(1, \frac{J_n^{-2/d} \ln N}{h_n^2}\right)^{k+1} \left(h_n^\beta + \mathbb{1}_{\beta > 1} \left(\frac{\ln N}{N h_n^d}\right)^{1/2} h_n\right)$$

we have

1. *If* $0 < \beta \leq 1$ *: there exists a constant* $C(\beta, f_0, \rho)$ *such that*

$$\left\| f_0 - p_{J_n}\left(e^{-t\mathcal{L}} f_t\right) \right\|_{L^\infty(\nu)} \leq C(\beta, f_0) \tilde{\varepsilon}_n(J_n, h_n)$$

2. *If* $\beta > 1, p_0 \in \mathcal{C}^{\beta-1}(\mathcal{M}), H > 0$ *: for some constant* $C(H, \beta, p_0, f_0, \mathcal{M}, \rho)$

$$\mathbb{P}_0\left(\left\| f_0 - p_{J_n}\left(e^{-t\mathcal{L}} f_t\right) \right\|_{L^\infty(\nu)} \geq C(H, \beta, p_0, f_0, \mathcal{M}) \tilde{\varepsilon}_n(J_n, h_n)\right) \leq N^{-H}$$

The proof of Theorem 12 is given in Section 4.2.

Both theorem 11 & 12 are useful results in their own rights, and can be used to design and study the convergence properties of frequentist estimators, as we have done in Theorem 5.

4 Main proofs

In this Section we present the proofs of Theorems 11 and 12. To control the approximation of f_0 by $e^{-t\mathcal{L}} f_t$ we introduce a deterministic approximation of the operator $\mathcal{L}f$. More precisely, recall that

$$\mathcal{L}f = \frac{1}{\mu_x h^2} \sum_{y \sim x} f(x) - f(y).$$

By analogy this leads us to the following definition :

Definition 13. *For every* $f \in \mathcal{C}(\mathcal{M})$ *and* $h > 0$ *we define*

$$T_h f(x) = \frac{1}{h^2 P_0(B_{\mathbb{R}^D}(x, h))} \int_{B_{\mathbb{R}^D}(x,h) \cap \mathcal{M}} (f(x) - f(y))\, p_0(y) \mu(dy)$$

where $P_0\left(B_{\mathbb{R}^D}(x, h)\right) = \int_{B_{\mathbb{R}^D}(x,h) \cap \mathcal{M}} p_0(y) \mu(dy)$.

T_h acts as a second order nonlocal differential operator and is a deterministic approximation of the operator $\mathcal{L}$ as shown in Lemma 15 below.

4.1 Proof of theorem 11

By a Taylor expansion of $t \mapsto e^{-\lambda t}$ (with $\lambda \geq 0$) we have

$$1 = e^{-\lambda t} \sum_{l=0}^{k} \frac{(\lambda t)^l}{l!} + R_k(t)$$

where the remainder is given by

$$R_k(t) = \frac{1}{k!}\int_0^1 (\lambda x t)^{k+1} e^{-\lambda x t} \frac{dx}{x}$$

Hence by (finite dimensional) functional calculus we find

$$f_0 = e^{-t\mathcal{L}} f_t + \frac{1}{k!}\int_0^1 (xt\mathcal{L})^{k+1} e^{-xt\mathcal{L}} f_0 \frac{dx}{x}$$

Therefore, for all $x \in V$, using $\left\|e^{-t\mathcal{L}}\right\|_{L^\infty(\nu)} = 1$, we bound

$$\begin{aligned}
\left\|f_0 - e^{-t\mathcal{L}} f_t\right\|_{L^\infty(\nu)} \leq & \frac{1}{k!}\int_0^1 \left\|(xt\mathcal{L})^{k+1} e^{-xt\mathcal{L}} f_0\right\|_{L^\infty(\nu)} \frac{dx}{x} = \frac{1}{k!}\int_0^t \left\|(s\mathcal{L})^{k+1} e^{-s\mathcal{L}} f_0\right\|_{L^\infty(\nu)} \frac{ds}{s} \\
\leq & \frac{1}{k!}\int_0^t \left\|(s\mathcal{L})^{k+1} f_0\right\|_{L^\infty(\nu)} \frac{ds}{s} = \frac{\left\|(t\mathcal{L})^{k+1} f_0\right\|_{L^\infty(\nu)}}{(k+1)!}.
\end{aligned}$$

We now bound $\left\|(t\mathcal{L})^{k+1} f_0\right\|_{L^\infty(\nu)}$, considering the three cases: $\beta \leq 1$, $\quad 1 < \beta \leq 2$ and $\beta > 2$.

- If $\beta \leq 1$ then the result is trivial as

$$|(t\mathcal{L})\, f_0(x)| = t\left|\frac{1}{h_n^2 \mu_x}\sum_{y\sim x} f_0(x) - f_0(y)\right| \lesssim \|f_0\|_{\mathcal{C}^\beta(\mathcal{M})}\, t h_n^{\beta-2}.$$

- If $1 < \beta \leq 2$, we have

$$\|t\mathcal{L}f_0\|_{L^\infty(\nu)} \leq t\,\|\mathcal{L}f_0 - T_{h_n} f_0\|_{L^\infty(\nu)} + t\,\|T_{h_n} f_0\|_{L^\infty(\nu)}\,.$$

  Then lemma 14 implies

$$\|T_{h_n} f_0\|_{L^\infty(\mathcal{M})} \leq C(\beta, p_0)\, \|f_0\|_{\mathcal{C}^\beta(\mathcal{M})}\, h_n^{\beta-2}$$

  and Lemma 15 implies that for any $H > 0$ there exists $M_0 > 0$ such that

$$\mathbb{P}_0\left(\|\mathcal{L}f_0 - T_{h_n} f_0\|_{L^\infty(\nu)} > M_0 \left(\frac{\ln N}{N}\right)^{1/2} h_n^{-(1+d/2)}\right) \leq N^{-H}.$$

  Hence, for some $C(H, \beta, f_0, p_0) > 0$

$$\mathbb{P}_0\left(\left\|h_n^2 \mathcal{L} f_0\right\|_{L^\infty(\nu)} > C(H,\beta,p_0,f_0)\left(\left(\frac{\ln N}{N}\right)^{1/2} h_n^{1-d/2} + h_n^\beta\right)\right) \leq N^{-H}.$$

- When $\beta > 2$ we use an inductive argument. We have shown that for $k = \lceil \beta/2 \rceil - 1 = 0$, if $f \in \mathcal{C}^\beta$,

$$\left\|h_n^2 \mathcal{L} f\right\|_{L^\infty(\nu)} \leq C_0(H, \beta, f_0, p_0, \mathcal{M})[h_n^\beta + \Delta_n(f)]$$

where $\mathbb{P}_0\left(|\Delta_n(f)| > M_0\left(\frac{\ln N}{Nh_n^d}\right)^{1/2} h_n\right) \leq N^{-H}$. Assume that for all $k' < k$, if $f \in \mathcal{C}^{\beta'}$ with $2k' < \beta' \leq 2(k'+1)$, then

$$\left\|(h_n^2 \mathcal{L} f)^{k'+1}\right\|_{L^\infty(\nu)} \leq C_{k'}(H, \beta, p_0, f_0, \mathcal{M})[h_n^{\beta'} + \Delta_{n,k'}(f)]$$

where $\mathbb{P}_0\left(|\Delta_{n,k'}(f)| > M_0\left(\frac{\ln N}{Nh_n^d}\right)^{1/2} h_n\right) \leq N^{-H}$. We prove that the same holds for k. Let $f \in \mathcal{C}^\beta$ with $2k < \beta \leq 2(k+1)$. Since $\beta > 2$ by lemma 14 we find $f_1, \ldots, f_k$ (that depend on h_n) such that for all $l \in \{1, \ldots, k\}$

$$\|f_l\|_{\mathcal{C}^{\beta-2l}(\mathcal{M})} \lesssim \|f\|_{\mathcal{C}^\beta(\mathcal{M})}, \quad \left\|T_{h_n} f - \sum_{l=1}^k h_n^{2(l-1)} f_1^{(l)}\right\|_{L^\infty(\mathcal{M})} \leq C_l(p_0, \mathcal{M}) \|f\|_{\mathcal{C}^\beta(\mathcal{M})} h_n^{\beta-2}$$

Using $\left\|h_n^2 \mathcal{L}\right\|_{L^\infty(\nu)} \leq 2$ which holds since for any $f \in \mathbb{R}^V$ and $x \in V$ we have

$$\left|\left(h_n^2 \mathcal{L} f\right)(x)\right| = \left|\frac{1}{\mu_x} \sum_{y \sim x} f(x) - f(y)\right| \leq \frac{2\|f\|_{L^\infty(\nu)}}{\mu_x} \#\{y : y \sim x\} = 2\|f\|_{L^\infty(\nu)},$$

we get

$$\begin{aligned}
&\left\|\left(h_n^2 \mathcal{L}\right)^{k+1} f_0\right\|_{L^\infty(\nu)} \\
&= h_n^2 \left\|\left(h_n^2 \mathcal{L}\right)^k (T_{h_n} f_0 + \mathcal{L} f_0 - T_{h_n} f_0)\right\|_{L^\infty(\nu)} \\
&\leq h_n^2 \left\|\left(h_n^2 \mathcal{L}\right)^k T_{h_n} f_0\right\|_{L^\infty(\nu)} + h_n^2 \left\|\left(h_n^2 \mathcal{L}\right)^k (\mathcal{L} f_0 - T_{h_n} f_0)\right\|_{L^\infty(\nu)} \\
&\lesssim h_n^2 \left\|\left(h_n^2 \mathcal{L}\right)^k T_{h_n} f_0\right\|_{L^\infty(\nu)} + h_n^2 \|\mathcal{L} f_0 - T_{h_n} f_0\|_{L^\infty(\nu)} \\
&= h_n^2 \left\|\left(h_n^2 \mathcal{L}\right)^k \left(\sum_{l=1}^k h_n^{2(l-1)} f_l + T_{h_n} f_0 - \sum_{l=1}^k h_n^{2(l-1)} f_l\right)\right\|_{L^\infty(\nu)} + h_n^2 \|\mathcal{L} f_0 - T_{h_n} f_0\|_{L^\infty(\nu)} \\
&\leq h_n^2 \sum_{l=1}^k h_n^{2(l-1)} \left\|\left(h_n^2 \mathcal{L}\right)^k f_l\right\|_{L^\infty(\nu)} + h_n^2 2^k C_l(p_0, \mathcal{M}) \|f\|_{\mathcal{C}^\beta(\mathcal{M})} h_n^{\beta-2} + \Delta_n(f_0).
\end{aligned}$$

Now,

$$\begin{aligned}
\left\|\left(h_n^2 \mathcal{L}\right)^k f_l\right\|_{L^\infty(\nu)} &\leq 2^{l-1} \left\|\left(h_n^2 \mathcal{L}\right)^{k-l+1} f_l\right\|_{L^\infty(\nu)} \\
&\leq C_{k-l}(H, \beta, p_0, f_0, \mathcal{M})[h_n^{\beta-2l} + \Delta_{n,k-l}(f)]
\end{aligned}$$

so that for all $f \in \mathcal{C}^\beta$

$$\begin{aligned}
\left\|\left(h_n^2 \mathcal{L}\right)^{k+1} f\right\|_{L^\infty(\nu)} &\leq h_n^2 \sum_{l=1}^k C_{k-l}(\|f\|_{\mathcal{C}^\beta}, p_0, \mathcal{M}) h_n^{2(l-1)} h_n^{\beta-2l} \\
&\quad + \sum_{l=0}^k C_{k-l}(\|f\|_{\mathcal{C}^\beta}, p_0, \mathcal{M}) \Delta_{n,k-l}(f_0) \\
&\leq C'(\beta, f_0, p_0, \mathcal{M})[h_n^\beta + \Delta_{n,k}(f)]
\end{aligned}$$

where $\mathbb{P}_0\left(|\Delta_{n,k}(f)| > M_0 \left(\frac{\ln N}{N h_n^d}\right)^{1/2} h_n\right) \leq N^{-H}$ for M_0 large enough. This concludes the inductive argument. Therefore

$$\left\|f_0 - e^{-t\mathcal{L}} f_t\right\|_{L^\infty(\nu)} \leq C'(H, \beta, f_0, p_0, \mathcal{M}) \left(\frac{t}{h_n^2}\right)^{k+1} [h_n^\beta + \Delta_{n,k}(f_0)],$$

where $\mathbb{P}_0\left(|\Delta_{n,k}(f_0)| > M_0 \left(\frac{\ln N}{N h_n^d}\right)^{1/2} h_n\right) \leq N^{-H}$.

4.2 Proof of Theorem 12

We bound $\left\|f_0 - p_{J_n}\left(e^{-t\mathcal{L}} f_t\right)\right\|_{L^\infty(\nu)}$ by showing that $\left\|e^{-t\mathcal{L}} f_t - p_{J_n}\left(e^{-t\mathcal{L}} f_t\right)\right\|_{L^\infty(\nu)}$ is small, using a lower bound on the eigenvalues $\lambda_j' s$ given by lemma 32 : since h_n satisfies assumption 2, choosing $h_- = h_n$ in the definition of A_N, together with Lemma 32, we have that $\lambda_j^{(h_n)} \geq b_3 j^{2/d}$ for all $b_1 \ln^d N \leq j \leq t_0(h_n) = b \frac{h_n^{-d}}{\ln^{d/2} N}$, so that since $\ln^\kappa N \leq J$ with $\kappa > d$, we have $\lambda_J = \lambda_J^{(h_n)} \geq b_3 J^{2/d}$ when $J \leq b \frac{h_n^{-d}}{\ln^{d/2} N} := \bar{J}_n$.

Moreover, because the $\lambda_j' s$ are ordered we have

$$\forall J \geq \bar{J}_n, \quad \lambda_J \geq \lambda_{\lfloor \bar{J}_n \rfloor} \geq b_3 b^{2/d} \frac{h_n^{-2}}{\ln N}$$

Therefore

$$\begin{aligned}
\left\|f_0 - p_J\left(e^{-t\mathcal{L}} f_t\right)\right\|_{L^\infty(\nu)} &\leq \left\|f_0 - e^{-t\mathcal{L}} f_t\right\|_{L^\infty(\nu)} + \left\|\sum_{j>J} \left\langle u_j | e^{-t\mathcal{L}} f_t\right\rangle_{L^2(\nu)} u_j\right\|_{L^\infty(\nu)} \\
&= \left\|f_0 - e^{-t\mathcal{L}} f_t\right\|_{L^\infty(\nu)} + \left\|\sum_{j>J} e^{-t\lambda_j} \left\langle u_j | f_t\right\rangle_{L^2(\nu)} u_j\right\|_{L^\infty(\nu)} \\
&\leq \left\|f_0 - e^{-t\mathcal{L}} f_t\right\|_{L^\infty(\nu)} + \sum_{j>J} e^{-t\lambda_j} \left|\left\langle u_j | f_t\right\rangle_{L^2(\nu)}\right| \|u_j\|_{L^\infty(\nu)} \\
&\leq \left\|f_0 - e^{-t\mathcal{L}} f_t\right\|_{L^\infty(\nu)} + \|f_t\|_{L^2(\nu)} e^{-t\lambda_J} \sum_{j>J} \|u_j\|_{L^\infty(\nu)}
\end{aligned}$$

By lemma 18 we have $\max_{1\leq j\leq N} \|u_j\|_{L^\infty(\nu)} \leq N$ and therefore

$$\left\|f_0 - p_J\left(e^{-t\mathcal{L}} f_t\right)\right\|_{L^\infty(\nu)} \lesssim \left\|f_0 - e^{-t\mathcal{L}} f_t\right\|_{L^\infty(\nu)} + N^2 \|f_t\|_{L^2(\nu)} e^{-t\lambda_J}$$

Moreover, since $f_t = \sum_{l=0}^k \frac{1}{l!} (t\mathcal{L})^l f_0$ and $\left\|h_n^2 \mathcal{L}\right\|_{\mathcal{L}(L^\infty(\nu))} \leq 2$ we get

$$\|f_t\|_{L^\infty(\nu)} \leq e^2 \left(\frac{t}{h_n^2}\right)^k \|f_0\|_{L^\infty(\mathcal{M})},$$

which implies

$$\left\|f_0 - p_J\left(e^{-t\mathcal{L}} f_t\right)\right\|_{L^\infty(\nu)} \lesssim \left\|f_0 - e^{-t\mathcal{L}} f_t\right\|_{L^\infty(\nu)} + N^2 e^2 \left(\frac{t}{h_n^2}\right)^k e^{-t\lambda_J}.$$

Thus for any $K > 0$, choosing $t = c\lambda_J^{-1} \ln N$ for $c > 0$ large enough yields

$$\left\| f_0 - p_J \left(e^{-t\mathcal{L}} f_t\right)\right\|_{L^\infty(\nu)} \lesssim \left\| f_0 - e^{-t\mathcal{L}} f_t\right\|_{L^\infty(\nu)} + N^{-K}.$$

Finally we bound $\left\| f_0 - e^{-t\mathcal{L}} f_t\right\|_{L^\infty(\nu)}$ using Theorem 11 which gives us for $K > 0$ large enough and arbitrarily large $H > 0$

$$\mathbb{P}_0 \left(\left\| f_0 - p_J \left(e^{-t\mathcal{L}} f_t\right)\right\| > C(H, \beta, f_0, p_0, \mathcal{M})\tilde{\varepsilon}_n\right) \le N^{-H}$$

where

$$\tilde{\varepsilon}_n = \left(\frac{t}{h_n^2}\right)^{\lceil \beta/2 \rceil} \left(h_n^\beta + \mathbb{1}_{\beta>1} \left(\frac{\ln N}{N h_n^d}\right)^{1/2} h_n\right) \lesssim \left(\frac{\lambda_J^{-1} \ln N}{h_n^2}\right)^{\lceil \beta/2 \rceil} \left(h_n^\beta + \mathbb{1}_{\beta>1} \left(\frac{\ln N}{N h_n^d}\right)^{1/2} h_n\right)$$

$$\lesssim \begin{cases} \left(\frac{J^{-2/d} \ln N}{h_n^2}\right)^{\lceil \beta/2 \rceil} \left(h_n^\beta + \mathbb{1}_{\beta>1} \left(\frac{\ln N}{N h_n^d}\right)^{1/2} h_n\right) & \text{if } J \le \bar{J}_n \\ (\ln N)^{2\lceil \beta/2 \rceil} \left(h_n^\beta + \mathbb{1}_{\beta>1} \left(\frac{\ln N}{N h_n^d}\right)^{1/2} h_n\right) & \text{if } J > \bar{J}_n \end{cases}$$

which concludes the proof.

4.3 Lemmas 14 and 15

Below we present two Lemmas which control the terms $\Delta_n(f)$ together with the behaviour of $T_h f$.

Lemma 14. *Let $a_{\mathcal{M}}, C_0, h_+$ the constants defined in Appendix A and B. Then*

1. *If $0 < \beta \le 2$, $h \le \frac{\pi a_{\mathcal{M}}}{C_0}$, $f \in \mathcal{C}^\beta(\mathcal{M})$ (with the additional assumption $p_0 \in \mathcal{C}^{\beta-1}(\mathcal{M})$ if $1 < \beta \le 2$), then $\|T_h f\|_{L^\infty(\mathcal{M})} \lesssim \|f\|_{\mathcal{C}^\beta(\mathcal{M})} h^{\beta-2}$*

2. *If $\beta > 2$, $f \in \mathcal{C}^\beta(\mathcal{M})$, $p_0 \in \mathcal{C}^{\beta-1}(\mathcal{M})$, $h \le h_+$, then there exists $\left(g_h^{(l)}\right)_{l=1}^k$, $g_h^{(l)} \in \mathcal{C}^{\beta-2l}(\mathcal{M})$ such that*

$$\left\| g_h^{(l)} \right\|_{\mathcal{C}^{\beta-2l}(\mathcal{M})} \lesssim \|f\|_{\mathcal{C}^\beta(\mathcal{M})}, \quad \left\| T_h f - \sum_{l=1}^k h^{2(l-1)} g_h^{(l)} \right\|_{L^\infty(\mathcal{M})} \lesssim \|f\|_{\mathcal{C}^\beta(\mathcal{M})} h^{\beta-2}$$

Lemma 15. *Let $h_n > 0$ satisfying assumption 2. Then*

- *If $f \in \mathcal{C}^\beta(\mathcal{M}), 0 < \beta \le 1$ we have*

$$\|\mathcal{L}f\|_{L^\infty(\nu)} \lesssim h_n^{\beta-2}, \|T_{h_n} f\|_{L^\infty(\mathcal{M})} \lesssim h_n^{\beta-2}$$

- *If $f \in \mathcal{C}^1(\mathcal{M})$ then any $H > 0$ there exists $M_0 > 0$ such that*

$$\mathbb{P}_0 \left(\|\mathcal{L}f - T_{h_n} f\|_{L^\infty(\nu)} > M_0 \left(\frac{\ln N}{N}\right)^{\frac{1}{2}} h_n^{-(1+d/2)} \right) \le N^{-H}$$

The proofs of both lemmas are provided in Sections C.1 and C.2 respectively.

4.4 Proof of Theorems 4 and 7

We now show how to extend the posterior contraction rates in terms of $\|f - f_0\|_n$ obtained in Theorems 3 and 6 to rates in terms of $\|f - f_0\|_N$.

As in van der Vaart and van Zanten (2011); Rosa, Borovitskiy, Terenin, and Rousseau (2023) where a posterior contraction rate with respect to the empirical $\|f - f_0\|_n$ norm is used to show a contraction rate with respect to the continuous L^2 norm, we will use a concentration inequality : indeed, by exchangeability the variables $(x_i)_{i=1}^n$ can be considered as sampled uniformly without replacement from the full sample $(x_i)_{i=1}^N$, and therefore the quantity $\|f - f_0\|_n^2 = \frac{1}{n}\sum_{i=1}^n (f(x_i) - f_0(x_i))$ is close to its mean $\frac{1}{N}\sum_{i=1}^N (f(x_i) - f_0(x_i)) = \|f - f_0\|_N^2$. We make this informal argument rigorous below.

We start by the non adaptive case, i.e the proof of theorem 4. Since $h_n \geq (\ln N)^s\, n^{-\frac{1}{2\beta+d}}$, it satisfies assumption 2, and since $\frac{h_n^{-d}}{\ln^{\tau_1} N} \leq J_n \leq \frac{h_n^{-d}}{\ln^{\tau_2} N}, \quad \tau_1 \geq \tau_2 > 2d$, Theorem 3 implies that, for some $C > 0$

$$\Pi\left[\|f - f_0\|_n > C\varepsilon_n(J_n, h_n)|\mathbb{X}^n, J_n, h_n\right] \xrightarrow[n\to\infty]{P_0^\infty} 0$$

where $\epsilon_n(J_n; h_n)$ is given by (5). Note that the choice of J_n implies that

$$\varepsilon_n(J_n, h_n) = \sqrt{\frac{J_n \ln N}{n}} + \left(\frac{J_n^{-2/d}\ln N}{h_n^2}\right)^{\lceil \beta/2\rceil}\left(h_n^\beta + \mathbb{1}_{\beta>1}\left(\frac{\ln N}{N h_n^d}\right)^{1/2} h_n\right)$$

Let

$$f_n = p_{J_n}\left(e^{-t_n\mathcal{L}} f_{t_n}\right), \quad f_{t_n} = \sum_{l=0}^{\lceil \beta/2\rceil - 1}\frac{(t_n\mathcal{L})^l f_0}{l!}, \quad t_n = c\lambda_{J_n}^{-1}\ln N, c > 0.$$

Then Theorem 12 implies that for some $c > 0$ and any $H > 0$, choosing C large enough,

$$\mathbb{P}_0\left(\|f_n - f_0\|_{L^\infty(\nu)} \leq C\varepsilon_n(J_n, h_n)\right) = 1 + O(N^{-H}).$$

Now if $M > C$, then on $\{\|f_n - f_0\|_{L^\infty(\nu)} \leq C\varepsilon_n\} = V_n$,

$$\begin{aligned}
&\Pi\left[\|f - f_0\|_N > M\varepsilon_n(J_n, h_n)|\mathbb{X}^n, J_n, h_n\right] && (8)\\
\leq &\Pi\left[\|f - f_n\|_N > (M - C)\,\varepsilon_n(J_n, h_n)|\mathbb{X}^n, J_n, h_n\right] &&\\
&\leq \Pi\left[\|f - f_n\|_n > (C+1)\,\varepsilon_n(J_n, h_n)|\mathbb{X}^n, J_n, h_n\right] && (9)\\
&+ \Pi\left[\|f - f_n\|_N > (M - C)\,\varepsilon_n(J_n, h_n) \geq \frac{M - C}{C+1}\|f - f_n\|_n\,|\mathbb{X}^n, J_n, h_n\right]. && (10)
\end{aligned}$$

Since

$$\|f - f_n\|_n \leq \|f - f_0\|_n + \|f_0 - f_n\|_{L^\infty(\nu)} \leq \|f - f_0\|_n + C\varepsilon_n(J_n, h_n),$$

we have on V_n

$$\Pi\left[\|f - f_n\|_n > (C+1)\,\varepsilon_n(J_n, h_n)|\mathbb{X}^n, J_n, h_n\right] \leq \Pi\left[\|f - f_0\|_n > \varepsilon_n(J_n, h_n)|\mathbb{X}^n, J_n, h_n\right] = o_{\mathbb{P}_0}(1).$$

Hence it remains to bound the second term of the right hand side of (8). For this notice that a consequence of the proof of Theorem 35 (see Ghosal and van der Vaart, 2007) is that there exists $c' > 0$ such that

$$\mathbb{P}_0\left[D_n \leq e^{-c'n\varepsilon_n(J_n,h_n)^2}\right] = o(1), \quad D_n = \int_{\Sigma^{J_n}} e^{\ell_n(f)-\ell_n(f_0)} d\Pi(f|J_n,h_n)$$

where $\ell_n(f)$ is the log-likelihood at f. As a consequence,

$$\begin{aligned}
&\mathbb{E}_0\left[\Pi\left[\|f-f_n\|_N \geq \frac{M-C}{C+1}\|f-f_n\|_n |\mathbb{X}^n, J_n, h_n\right]\right] \\
\leq& o(1) + e^{c'n\varepsilon_n^2(J_n,h_n)}\mathbb{E}_0\left[\int_{\Sigma^{J_n}} \mathbb{1}_{\|f-f_n\|_N > \frac{M-C}{C+1}\|f-f_n\|_n} \mathbb{1}_{\|f-f_n\|_n<(C+1)\epsilon_n} e^{\ell_n(f)-\ell_n(f_0)} d\Pi(f|J_n,h_n)\right] \\
\leq& o(1) + e^{c'n\varepsilon_n^2(J_n,h_n)}\mathbb{E}_0 \int_{f\in\Sigma^{J_n}} \mathbb{1}_{\|f-f_n\|_N > \frac{M-C}{C+1}\|f-f_n\|_n} \Pi(df)
\end{aligned}$$

where the last expectation is now only with respect to the distribution of $x_{1:N}$. Notice that by exchangeability of the x_i's, for any permutation $\tau \in \mathcal{S}_N$ the set of permutations of $\{1,\ldots,N\}$ we have $(x_i)_{i=1}^N \overset{(d)}{\equiv} \left(x_{\tau(i)}\right)_{i=1}^N$ and therefore, writing $\|f-f_n\|_{\tau,n}^2 = \frac{1}{n}\sum_{i=1}^n \left(f(x_{\tau(i)}) - f_n(x_{\tau(i)})\right)^2$,

$$\mathbb{E}_0 \int_{f\in\Sigma^{J_n}} \mathbb{1}_{\|f-f_n\|_N > \frac{M-C}{C+1}\|f-f_n\|_n} \Pi(df) = \mathbb{E}_0 \int_{f\in\Sigma^{J_n}} \mathbb{1}_{\|f-f_n\|_N > \frac{M-C}{C+1}\|f-f_n\|_{\tau,n}} \Pi(df).$$

Hence

$$\begin{aligned}
&\mathbb{E}_0 \int_{f\in\Sigma^{J_n}} \mathbb{1}_{\|f-f_n\|_N > \frac{M-C}{C+1}\|f-f_n\|_n} \Pi(df) \\
=&\frac{1}{N!}\sum_{\tau\in\mathcal{S}_N} \mathbb{E}_0 \int_{f\in\Sigma^{J_n}} \mathbb{1}_{\|f-f_n\|_N > \frac{M-C}{C+1}\|f-f_n\|_{\tau,n}} \Pi(df) \\
=&\mathbb{E}_{\tau\sim\mathcal{U}(\mathcal{S}_N)}\mathbb{E}_0 \int_{f\in\Sigma^{J_n}} \mathbb{1}_{\|f-f_n\|_N > \frac{M-C}{C+1}\|f-f_n\|_{\tau,n}} \Pi(df) \\
=&\mathbb{E}_0 \int_{f\in\Sigma^{J_n}} \mathbb{P}_{\tau\sim\mathcal{U}(\mathcal{S}_N)}\left(\|f-f_n\|_N > \frac{M-C}{C+1}\|f-f_n\|_{\tau,n}\right) \Pi(df).
\end{aligned}$$

Define for $i = 1,\ldots,N$, $W_i = (f(x_i) - f_n(x_i))^2$ and $W_i' = \left(f(x_{\tau(i)}) - f_n(x_{\tau(i)})\right)^2$. Then given $(x_i)_{i=1}^N$, $(W_i')_{i=1}^n$ is a uniform sampling without replacement from $(W_i)_{i=1}^N$. Hoeffding's lemma for random variables sampled without replacement from a finite population (see Hoeffding, 1963) then yields, for $M = 3C + 2 > C$

$$
\begin{aligned}
&\mathbb{P}_{\tau \sim \mathcal{U}(\mathcal{S}_N)}\left(\|f - f_n\|_N > \frac{M-C}{C+1} \|f - f_n\|_{\tau,n} \right) \\
=&\mathbb{P}_{\tau \sim \mathcal{U}(\mathcal{S}_N)}\left(\frac{1}{N}\sum_{i=1}^{N} W_i^2 > \frac{4}{n}\sum_{i=1}^{n} W_i'^2 \right) \\
=&\mathbb{P}_{\tau \sim \mathcal{U}(\mathcal{S}_N)}\left(\frac{-3}{4}\frac{1}{N}\sum_{i=1}^{N} W_i^2 > \frac{1}{n}\sum_{i=1}^{n} W_i'^2 - \frac{1}{N}\sum_{i=1}^{N} W_i^2 \right) \\
=&\mathbb{P}_{\tau \sim \mathcal{U}(\mathcal{S}_N)}\left(\frac{-1}{n}\sum_{i=1}^{n} W_i'^2 + \frac{1}{N}\sum_{i=1}^{N} W_i^2 > \frac{3}{4}\frac{1}{N}\sum_{i=1}^{N} W_i^2 \right) \\
\leq& \exp\left(-\frac{3n \|f - f_n\|_N^2}{2 \|f - f_n\|_{L^\infty(\nu)}^2} \right).
\end{aligned}
$$

Using theorem 34 yields, on $V_n \cap A_N$ (which is an event of probability tending to 1), with $h_- = h_n$ and if $b_4 \leq J_n \leq b_5 \frac{t_0^{-d/2}}{\ln^{3d/2} N}$ (where $t_0 = 64 h_n^2 \ln\left(2c_- N h_n^d\right)$)

$$
\forall f \in \Sigma^{J_n}, \quad \frac{\|f - f_n\|_N^2}{\|f - f_n\|_{L^\infty(nu)}^2} \geq e^{-1} a_3^{-1} b_6^{-d/2} J_n^{-1} \ln^{-3d/2} N.
$$

Note that since by assumption $\frac{h_n^{-d}}{\ln^{\tau_1} N} \leq J_n \leq \frac{h_n^{-d}}{\ln^{\tau_1} N}, \quad \tau_1 \geq \tau_2 > 2d$, then $b_4 \leq J_n \leq b_5 \frac{t_0^{-d/2}}{\ln^{3d/2} N}$. All in all we get

$$
\begin{aligned}
&\mathbb{E}_0\left[\Pi\left[\|f - f_n\|_N > (M - C)\,\varepsilon_n \geq \frac{M-C}{C+1} \|f - f_n\|_n \,|\mathbb{X}^n, J_n, h_n \right]\right] \\
&\leq o(1) + e^{c' n \varepsilon_n^2(J_n,h_n)} \mathbb{E}_0\left[\mathbb{1}_{A_N \cap V_n} \int_{f \in \Sigma^{J_n}} \exp\left(-\frac{3n \|f - f_n\|_N^2}{2\|f - f_n\|_{L^\infty(\nu)}^2} \right) \Pi(df) \right] \\
&\leq o(1) + e^{c' n \varepsilon_n^2(J_n,h_n)} \mathbb{E}_0 \int_{f \in \Sigma^{J_n}} \exp\left(-\frac{3n J_n^{-1}}{2 e a_3 b_6^{d/2} \ln^{3d/2} N} \right) \Pi(df) \\
&= o(1) + \exp\left(c' n \varepsilon_n^2(J_n, h_n) - \frac{3n J_n^{-1}}{2 e a_3 b_6^{d/2} \ln^{3d/2} N} \right).
\end{aligned}
$$

Moreover

$$
J_n^{-1} \ln^{-3d/2} N \geq h_n^d \ln^{\tau_2 - 3d/2} N
$$

and by definition of $\varepsilon_n(J_n, h_n)$ we have

$$
\varepsilon_n^2(J_n, h_n) \lesssim \frac{J_n \ln N}{n} + \left(\frac{J_n^{-2/d} \ln N}{h_n^2} \right)^{2\lceil \beta/2 \rceil} h_n^{2\beta} \leq \frac{h_n^{-d} \ln^{1-\tau_2} N}{n} + h_n^{2\beta} \ln^{2\lceil \beta/2\rceil (1 + 2\tau_1/d)} N.
$$

But since by assumption $h_n \geq (\ln N)^s n^{-\frac{1}{2\beta+d}}$ for some $s \in \mathbb{R}$ and $\beta > d/2$, this actually implies $n J_n^{-1} \ln^{-3d/2} N >> n \varepsilon_n^2(J_n, h_n)$, which concludes the proof.

The proof of the adaptive case, i.e of theorem 7, is fairly similar. Again, we use Theorem 6, so that there exists $C > 0$

$$\mathbb{E}_0 \left(\Pi \left[\|f - f_0\|_n > C\varepsilon_n | \mathbb{X}^n\right]\right) = o(1).$$

where $\varepsilon_n = n^{-\beta/(2\beta+d)} (\ln N)^{\frac{(2\tau+d)\lfloor\beta/2\rfloor - \beta(\tau+2\beta/d)}{2\beta+d}}$.

Moreover, since by assumption on Π_J for any J_n, k we have $\Pi\left[J > kJ_n\right] \lesssim e^{-a_2 k J_n}$, the remaining mass Theorem (see Ghosal and van der Vaart, 2017, chap. 8) together with our prior thickness result 36 yields, for some $k > 0$

$$\Pi\left[J > kJ_n | X^n\right] \xrightarrow[n\to\infty]{P_0^\infty} 0$$

for $J_n = n\varepsilon_n^2$.

Therefore, a similar proof as in the non adaptive case above leads to the upper bound: on $V_n \cap A_N$,

$$\begin{aligned}
&\Pi\left[\|f - f_0\|_N > M\varepsilon_n | \mathbb{X}^n\right] \\
\leq & o_{\mathbb{P}_0}(1) + e^{c' n \varepsilon_n^2 (J_n, h_n)} \sum_{J \leq kJ_n} \mathbb{E}_0 \int_{f \in \Sigma^J} \exp\left(-\frac{3n\,\|f - f_n\|_N^2}{2\,\|f - f_n\|_{L^\infty(\nu)}^2}\right) d\Pi(f|F)\Pi_J(J)
\end{aligned}$$

for some $c' > 0$. Moreover, since by construction $\Pi-$almost surely we have $J \leq \frac{h^{-d}}{\ln^{\tau/d} N}$, $\tau > 2d$, we can still apply theorem 34 to get on $V_n \cap A_N$,

$$\begin{aligned}
&\Pi\left[\|f - f_0\|_N > M\varepsilon_n | \mathbb{X}^n\right] \\
&\leq o_{\mathbb{P}_0}(1) + e^{c' n \varepsilon_n^2} \sum_{J \leq kJ_n} \mathbb{E}_0 \int_{f \in \Sigma^J} \exp\left(-\frac{3nJ^{-1}}{2ea_3 b_6^{d/2} \ln^{3d/2} N}\right) d\Pi(f|J)\Pi_J(J) \\
&\leq o_{\mathbb{P}_0}(1) + e^{c' n \varepsilon_n^2} \sum_{J \leq kJ_n} \exp\left(-\frac{3nJ^{-1}}{2ea_3 b_6^{d/2} \ln^{3d/2} N}\right) \\
&\leq o_{\mathbb{P}_0}(1) + J_n \exp\left(c' n \varepsilon_n^2 - \frac{3nJ_n^{-1}}{2ea_3 b_6^{d/2} \ln^{3d/2} N}\right) \\
&= o_{\mathbb{P}_0}(1) + \exp\left(c' n \varepsilon_n^2 - \frac{3}{2eka_3 b_6^{d/2}} \varepsilon_n^{-2} \ln^{-3d/2} N\right)
\end{aligned}$$

Given the expression of ε_n and since $\beta > d/2$, the right hand side is $o_{\mathbb{P}_0}(1)$, which concludes the proof of the adaptive case.

Acknowledgments and Disclosure of Funding

The project leading to this work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 834175).

Appendix A. Regularity assumption on the manifold and definition of the Hölder spaces

In this section we give details on the regularity of the submanifold $\mathcal{M}$ and the definition of the Hölder spaces. It is usually assumed in Riemannian geometry that the manifolds are $\mathcal{C}^\infty-$smooth but here our results also apply for manifolds with finite Hölder regularity if the latter is large enough, hence we precise our notations and recall some general facts. We assume that $\mathcal{M}$ to be a $d-$dimensional $\mathcal{C}^\alpha$ compact and connected submanifold of $\mathbb{R}^D$ in the following sense : we can find a family $(\varphi_i,\mathcal{U}_i)_{i\in I}$ such that $(\mathcal{U}_i)_{i\in I}$ is an open cover of $\mathcal{M}$ (that is, each $\mathcal{U}_i$ is an open set of $\mathcal{M}$ and $\mathcal{M}=\cup_{i\in I}\mathcal{U}_i$), $\varphi_i:\varphi_i^{-1}(\mathcal{U}_i)\subset\mathbb{R}^d\to\mathcal{U}_i\subset\mathcal{M}\subset\mathbb{R}^D$ is injective, of class $\mathcal{C}^\alpha$, $\quad \alpha\geq 1$ and its differential $d\varphi_i:\mathbb{R}^d\to\mathbb{R}^D$ is everywhere injective (hence φ_i is a $\mathcal{C}^\alpha$ immersion). Since $\mathcal{M}$ is compact we can assume without loss of generality that I is finite, but the regularity results below stay true in the general case, at least locally.

The geodesics on the submanifold are defined as the curves $\gamma(t)$ on $\mathcal{M}$ satisfying the geodesics equation : in local coordinates $y=\varphi^{-1}(\gamma)$ (we drop the subscript i for convenience) it takes the form

$$\forall k=1,\dots,d,\quad \ddot{y}^k+\Gamma^k_{ij}\dot{y}^i\dot{y}^j=0$$

where the Christoffel symbols are defined by

$$\Gamma^k_{ij}=\frac{1}{2}g^{kl}\left(\partial_i g_{jl}+\partial_j g_{il}-\partial_l g_{ij}\right)$$

Here both the metric tensor g, its inverse $\left(g^{kl}\right)$ and the Christoffel symbols depend on the point $\gamma(t)=\varphi(y(t))$. Moreover, with $\left(\partial_i:=\frac{\partial\varphi}{\partial y_i}\right)_{i=1}^d$ the coordinates vector fields and $\varphi=\left(\varphi^l\right)_{l=1}^D$, the metric tensor is given by the Gram matrix

$$\forall p\in\mathcal{M},\quad g_{ij}(p)=\langle\partial_i|\partial_j\rangle_{T_p\mathcal{M}}=\sum_{l=1}^D\frac{\partial\varphi^l}{\partial y^i}(\varphi^{-1}(p))\frac{\partial\varphi^l}{\partial y^j}(\varphi^{-1}(p))=\left(\frac{\partial\varphi}{\partial y_i}\right)^T\left(\frac{\partial\varphi}{\partial y_j}\right)(\varphi^{-1}(p))$$

and at $p=\gamma(t)$

$$\partial_l g_{ij}=\frac{\partial\left(g_{ij}\circ\varphi\right)}{\partial y^l}(y(t))=\left[\frac{\partial}{\partial y^l}\sum_{k=1}^D\frac{\partial\varphi^k}{\partial y^i}\frac{\partial\varphi^k}{\partial y^j}\right](y(t))$$

Hence for each i,j,k, Γ^k_{ij} is a $\mathcal{C}^{\alpha-2}$ Hölder function of $y(t)$. Therefore by standard ODE theory, if $\alpha\geq 3$ existence and uniqueness of geodesics $t\mapsto\gamma_{x,v}(t)$ defined on maximal intervals and such that $\varphi(y(0))=\gamma_{x,v}(0)=x$, $\frac{\partial\varphi}{\partial y}(\varphi^{-1}(x))\dot{y}(0)=\gamma'_{x,v}(0)=v$ for arbitrary $x\in\mathcal{M},\quad v\in T_x\mathcal{M}$ is guaranteed. Moreover the flow $(x,v,t)\mapsto\left(\gamma_{x,v}(t),\gamma'_{x,v}(t)\right)$ is of class $\mathcal{C}^{\alpha-2}$, in the sense that

$$(y_0,\dot{y}_0,t)\mapsto\left(\gamma_{\varphi(y_0),\frac{\partial\varphi}{\partial y}(y_0)\dot{y}_0}(t),\gamma'_{\varphi(y_0),\frac{\partial\varphi}{\partial y}(y_0)\dot{y}_0}(t)\right)$$

is of class $\mathcal{C}^{\alpha-2}$ on its domain. In particular the exponential map defined as $\exp_x(v)=\gamma_{x,v}(1)$ for all $\|v\|_{T_x\mathcal{M}}<r_x$ (the injectivity radius of $\mathcal{M}$ at x) is in this sense of class $\mathcal{C}^{\alpha-2}$ on its

domain, as seen as a function with values in $\mathbb{R}^D$. Since $\mathcal{M}$ is compact it has a positive global injectivity radius $r_{\mathcal{M}} := \inf_{x\in\mathcal{M}} r_x$ and therefore for any $x \in \mathcal{M}$ the exponential map $\exp_x$ is well defined from $B_{T_x\mathcal{M}}(0, r_{\mathcal{M}})$ to $\mathcal{M}$ and is a diffeomorphism onto its image. Moreover, by compactness of $\mathcal{M}$ the Hopf-Rinow theorem implies that for each $x \in \mathcal{M}$ the exponential map $\exp_x$ is actually defined on the whole of $T_x\mathcal{M}$ (while obviously no longer being guaranteed to be injective). Hence by the above, for any $i \in I$ the following map

$$\left(\begin{array}{ccc} \varphi_i^{-1}(\mathcal{U}_i) \times \mathbb{R}^d & \to & \mathbb{R}^D \\ (y_0, \dot{y}_0) & \mapsto & \exp_{\varphi_i(y_0)}\left(\frac{\partial \varphi_i}{\partial y}(y_0)\dot{y}_0\right) \end{array}\right)$$

is of class $\mathcal{C}^{\alpha-2}$.

Finally let us define the different Hölder spaces over $\mathcal{M}$ used in the paper. Let $(\chi_i)_{i\in I}$ be a partition of the unity subordinated to the open cover $(\mathcal{U}_i)_{i\in I}$. We say that $f : \mathcal{M} \to \mathbb{R}$ is of class $\mathcal{C}^\beta$ (β-Hölder) on $\mathcal{M}$ if $f \circ \varphi_i : \varphi_i^{-1}(\mathcal{U}_i) \to \mathbb{R}$ is of class $\mathcal{C}^\beta$ for any $i \in I$. For a finite dimensional vector space V we say that $f : \mathcal{M} \to V$ is of class $\mathcal{C}^\beta$ (β-Hölder) on $\mathcal{M}$ if $T \circ f \circ \varphi_i : \varphi_i^{-1}(\mathcal{U}_i) \to \mathbb{R}$ is of class $\mathcal{C}^\beta$ for any $i \in I$ and linear form $T \in V^*$. Equivalently the coordinates of f in any choice of basis of V are all of class $\mathcal{C}^\beta$. We also define $\mathcal{C}^\infty(\mathcal{M}) = \cap_{\beta\geq 0}\mathcal{C}^\beta(\mathcal{M})$. Furthermore we define the $\mathcal{C}^\beta$ norm of f by $\|f\|_{\mathcal{C}^\beta(\mathcal{M})} = \max_{i\in I}\left\|(\chi_i f) \circ \varphi_i^{-1}\right\|_{\mathcal{C}^\beta(\mathcal{U}_i)}$. The definition of the Hölder space $\mathcal{C}^\beta(\mathcal{M}, V)$ can be seen to be independent of the chosen family $(\varphi_i, \mathcal{U}_i, \chi_i)_{i\in I}$ as long as $\alpha \geq \beta \vee 1$, with equivalence of the resulting Hölder norms.

Recall that we can always define regular local orthonormal frames on $T\mathcal{M}$: on $\mathcal{U}_i$ simply apply the Gram-Schmidt orthonormalization algorithm to the coordinates vector fields $x \mapsto \frac{\partial \varphi}{\partial y_i}$, which gives a family $\left(e_i^j\right)_{i\in I, j=1,\ldots,d}$ of orthonormal vector fields whose components in local coordinates are rational functions of the components of the coordinates vector fields $\frac{\partial \varphi_i}{\partial y_i}$. As $\varphi_i : \varphi_i^{-1}(\mathcal{U}_i) \subset \mathbb{R}^d \mapsto \mathcal{U}_i \subset \mathbb{R}^D$ is $\mathcal{C}^{\alpha-1}$, the maps: $x \in \varphi_i^{-1}(\mathcal{U}_i) \subset \mathbb{R}^d \mapsto e_i^j(x) \in \mathcal{U}_i \subset \mathbb{R}^D$ are all $\mathcal{C}^{\alpha-1}$. Moreover by construction the function $\omega \mapsto \psi_i(x, \omega) := \sum_j \omega_j e_i^j(x)$ is a linear isometry from $\in \mathbb{R}^d$ to $T_x\mathcal{M}$ for each $x \in \mathcal{U}_i$ where $\mathbb{R}^d$ is equipped with its standard Euclidean structure and $T_x\mathcal{M}$ its inner product induced by the metric.

In particular if $\alpha - 2 \geq \beta \vee 1 \iff \alpha \geq (\beta + 2) \vee 3$ then any $f \in \mathcal{C}^\beta(\mathcal{M})$ satisfies

$$\begin{aligned}
&\sup_{x\in\mathcal{U}_i} \|f \circ \exp_x(\psi_i(x,\cdot))\|_{\mathcal{C}^\beta\left(B_{\mathbb{R}^d}(0,r_{\mathcal{M}})\right)} = \sup_{x\in\mathcal{U}_i}\left\|\sum_j (\chi_j f) \circ \exp_x(\psi_i(x,\cdot))\right\|_{\mathcal{C}^\beta\left(B_{\mathbb{R}^d}(0,r_{\mathcal{M}})\right)} \\
&\leq \sum_j \sup_{x\in\mathcal{U}_i} \|(\chi_j f) \circ \exp_x(\psi_i(x,\cdot))\|_{\mathcal{C}^\beta\left(B_{\mathbb{R}^d}(0,r_{\mathcal{M}})\right)} \\
&= \sum_j \sup_{x\in\mathcal{U}_i}\left\|((\chi_j f) \circ \varphi_j) \circ \left(\varphi_j^{-1} \circ \exp_x(\psi_i(x,\cdot))\right)\right\|_{\mathcal{C}^\beta\left(B_{\mathbb{R}^d}(0,r_{\mathcal{M}})\right)} \\
&\lesssim \max_j \|(\chi_j f) \circ \varphi_j\|_{\mathcal{C}^\beta(\mathcal{U}_j)} \sum_j \sup_{x\in\mathcal{U}_i}\left\|\varphi_j^{-1} \circ \exp_x(\psi_i(x,\cdot))\right\|_{\mathcal{C}^\beta\left(B_{\mathbb{R}^d}(0,r_{\mathcal{M}}),\varphi_j^{-1}(\mathcal{U}_i)\right)} \lesssim \|f\|_{\mathcal{C}^\beta(\mathcal{M})},
\end{aligned}$$

the last step using resulting from the compactness of $\mathcal{M}$, finiteness of I and the fact that $\alpha - 2 \geq \beta$. We will use this property in our approximation results.

We recall that we assume $\alpha \geq (\beta+3) \vee 6$. This is for technical reasons that will appear later in the proofs.

We finish this section by showing a differential geometry lemma that will be useful for the proofs of the results in section 3 : very roughly speaking, this lemma states that for small h we can find a function $t = t_h(x,v)$ of $x \in \mathcal{M}, v \in T_x\mathcal{M}$ such that $\|\exp_x(thv) - x\|_{\mathbb{R}^D} = h$, and that moreover this function can be written as $t_h = 1 + h^2 s_h$ where s_h is Hölder regular and has bounded derivatives of order less than $\alpha - 5$, even when h is close to 0. First, recall that the radius of curvature (see Bernstein et al.) $a_{\mathcal{M}}$ of $\mathcal{M}$ is positive by compactness of $\mathcal{M}$.

Lemma 16. *There exists $r'_{\mathcal{M}} > 0$ such that for each $x \in \mathcal{M}, \quad v \in T_x\mathcal{M}, \quad \|v\| = 1$ the function $r \mapsto \|\exp_x(rv) - x\|$ is increasing on $[0, r'_{\mathcal{M}}]$ and satisfies $\|\exp_x(rv) - x\|^2 \geq \frac{r^2}{4}, \quad r \in [0, r'_{\mathcal{M}}]$.*

As a consequence for any $h \in [0, \frac{r'_{\mathcal{M}}}{2}]$ there is a unique $t = t_h(x,v) \in [0, h^{-1}r'_{\mathcal{M}}]$ satisfying $\|\exp_x(thv) - x\| = h$. Moreover $t_h(x,v) \in [1, 1 + \frac{\pi^3 h^2}{192 a_{\mathcal{M}}^2}]$ and there exists $h_+ > 0$ such that

$$\max_{\mathbf{k} \in \mathbb{N}^d, |\mathbf{k}| \leq \alpha - 5} \sup_{\substack{i \in I \\ y \in \varphi_i^{-1}(\mathcal{U}_i) \\ 0 < h < h'_+ \\ \|v\|_{T_{\varphi_i(y)}\mathcal{M}} = 1}} \left| \frac{\partial^{|\mathbf{k}|}}{\partial y^{\mathbf{k}}} \left(\frac{t_h(\varphi_i(y), v) - 1}{h^2} \right) \right| < +\infty$$

Proof We have for $r < r_{\mathcal{M}}$,

$$\begin{aligned}
&\frac{\partial}{\partial r} \|\exp_x(rv) - x\|_{\mathbb{R}^D}^2 \\
=&2 \langle d\exp_x(rv)v | \exp_x(rv) - x \rangle_{\mathbb{R}^D} \\
=&2 \langle v + d\exp_x(rv)v - v | rv + \exp_x(rv) - x - rv \rangle_{\mathbb{R}^D} \\
=&2 \langle v | rv \rangle_{\mathbb{R}^D} + 2 \langle d\exp_x(rv)v - v | rv \rangle_{\mathbb{R}^D} + 2 \langle \exp_x(rv) - x - rv | v \rangle_{\mathbb{R}^D} \\
&\quad + 2 \langle d\exp_x(rv)v - v | \exp_x(rv) - x - rv \rangle_{\mathbb{R}^D} \\
=&2r + \mathcal{O}\left(r^2\right)
\end{aligned}$$

where the last equality uses

$$\exp_x(rv) - x - rv = r d\exp_x(0)v - rv + O(r^2) = O(r^2),$$

$$d\exp_x(rv)v - v = d\exp_x(0)v - v + O(r^2) = O(r^2).$$

By compactness of $\mathcal{M}$ and since $(x,v) \mapsto \exp_x(v) \in \mathcal{C}^{\alpha-3}, \alpha \geq 3$ this implies that

$$\sup_{\substack{0 < r < r_{\mathcal{M}} \\ x \in \mathcal{M} \\ \|v\|_{T_x\mathcal{M}} = 1}} \frac{\left| \frac{\partial}{\partial r} \|\exp_x(rv) - x\|_{\mathbb{R}^D}^2 - 2r \right|}{r^2} < +\infty,$$

In particular there exists $r = r'_{\mathcal{M}} > 0$ such that for any $r \in [0, r'_{\mathcal{M}}]$ we have

$$\forall x \in \mathcal{M}, \quad \forall v \in T_x\mathcal{M}, \quad \|v\|_{T_x\mathcal{M}} = 1, \quad \frac{\partial}{\partial r} \|\exp_x(rv) - x\|_{\mathbb{R}^D}^2 \geq \frac{r}{2}.$$

As a consequence, the function $r \mapsto \|\exp_x(rv) - x\|^2_{\mathbb{R}^D}$ is increasing on $[0, r'_{\mathcal{M}}]$ and so is $r \mapsto \|\exp_x(rv) - x\|_{\mathbb{R}^D}$. Moreover by integrating we get

$$\forall r \in [0, r'_{\mathcal{M}}], \quad \|\exp_x(rv) - x\|^2_{\mathbb{R}^D} \geq \frac{r^2}{4},$$

so that, if $h \in [0, \frac{r'_{\mathcal{M}}}{2}]$, then $r = 2h \in [0, r'_{\mathcal{M}}]$ and $\|\exp_x(rv) - x\|^2_{\mathbb{R}^D} \geq \frac{(2h)^2}{4} = h^2$. By continuity and strict monotonicity there exists a unique $r = r_h(x, v) \in [0, 2h]$ solution of $\|\exp_x(rv) - x\| = h$ on $[0, r'_{\mathcal{M}}]$. Defining $t_h(x, v) = h^{-1} r_h(x, v)$ gives the first part of the statement.

Let $i \in I, \quad x \in \mathcal{U}_i, \quad \omega \in \mathbb{R}^d, \quad \|\omega\| = 1$ and consider the Taylor expansion (justified since $\alpha \geq 5$ and $\exp \in \mathcal{C}^{\alpha-2}$)

$$\begin{aligned}&\exp_x(th\psi_i(x, \omega))\\ =&x + th\psi_i(x, \omega) + \frac{t^2 h^2}{2} d^2 \exp_x(0).\psi_i(x, \omega) + h^3 \int_0^t \frac{(t-s)^2}{2} d^3 \exp_x(sh\psi_i(x, \omega))\psi_i(x, \omega)^3 ds.\end{aligned}$$

Changing of variables with $y \in \varphi_i^{-1}(\mathcal{U}_i), \quad x = \varphi_i(y)$, we then have

$$\begin{aligned}&F(y, t)\\ :=&h^{-2} \|\exp_x(th\psi_i(x, \omega)) - x\|^2 - 1\\ =&h^{-2} \left\| th\psi_i(x, \omega) + \frac{t^2 h^2}{2} d^2 \exp_x(0).\psi_i(x, \omega) + h^3 \int_0^t \frac{(t-s)^2}{2} d^3 \exp_x(sh\psi_i(x, \omega))\psi_i(x, \omega)^3 ds \right\|^2 - 1\\ =&\left\| t\psi_i(x, \omega) + \frac{t^2 h}{2} d^2 \exp_x(0).\psi_i(x, \omega) + h^2 \int_0^t \frac{(t-s)^2}{2} d^3 \exp_x(sh\psi_i(x, \omega))\psi_i(x, \omega)^3 ds \right\|^2 - 1\end{aligned}$$

Therefore we have

$$F(y, t) = t^2 - 1 + h^2 G(y, t)$$

where, with

$$A = t\psi_i(x, \omega), \quad B = \frac{t^2}{2} d^2 \exp_x(0).\psi_i(x, \omega), \quad C = \int_0^t \frac{(t-s)^2}{2} d^3 \exp_x(sh\psi_i(x, \omega))\psi_i(x, \omega)^3 ds$$

we have defined

$$G(y, t) = G_{i,h,\omega}(y, t) = 2h^{-1} \langle A|B \rangle + 2 \langle A|C \rangle + \|B\|^2 + 2h \langle B|C \rangle + h^2 \|C\|^2$$

Moreover, since $\|t\psi_i(x, \omega)\|^2 = t^2$ and $d^2 \exp_x(0).\psi_i(x, \omega) \in N_x\mathcal{M}$ (as the second derivative at time 0 of the geodesic $t \mapsto \exp_x(t\psi_i(x, \omega)))$ this implies $\langle A|B \rangle = 0$ and therefore

$$G(y, t) = 2 \langle A|C \rangle + \|B\|^2 + 2h \langle B|C \rangle + h^2 \|C\|^2$$

Furthermore since the exponential map is of class $\mathcal{C}^{\alpha-2}$ we see that G is of class $\mathcal{C}^{\alpha-5}$ and that

$$M := \max_{\substack{i \in I \\ (\mathbf{k}, l) \in \mathbb{N}^{d+1} \\ |\mathbf{k}| + l \leq \alpha - 5}} \sup_{\substack{y \in \varphi_i^{-1}(\mathcal{U}_i) \\ 1 \leq t \leq 1 + \frac{\pi}{2} \\ 0 < h < r'_{\mathcal{M}}/2 \\ \|\omega\| = 1}} \left| \frac{\partial^{|\mathbf{k}| + l} G}{\partial y^{|\mathbf{k}|} \partial t^l} \right| < +\infty.$$

Differentiating yields

$$\frac{\partial F}{\partial y} = h^2 \frac{\partial G}{\partial y}, \frac{\partial F}{\partial t} = 2t + h^2 \frac{\partial G}{\partial t}.$$

Also, notice that $1 \leq t_h(x,v) \leq 1 + \frac{\pi^3 h^2}{192 a_{\mathcal{M}}^2} \wedge \frac{\pi}{2}$: indeed, without loss of generality $r'_{\mathcal{M}} \leq \pi a_{\mathcal{M}}$ and therefore if $h \leq r'_{\mathcal{M}}/2$ then for $t = t_h(x,v) \in [0,2]$ we have $th = \rho\left(x, \exp_x(th\psi_i(v))\right) \leq r'_{\mathcal{M}} \leq \pi a_{\mathcal{M}}$ and

$$\frac{2}{\pi} th = \frac{2}{\pi}\rho(x, \exp_x(th\psi_i(x,\omega))) \leq \|\exp_x(th\psi_i(x,\omega)) - x\| = h \leq \rho(x, \exp_x(th\psi_i(x,\omega))) = th$$

$$\implies 1 \leq t_h(x,v) \leq \frac{\pi}{2}.$$

Reusing this gives

$$\begin{aligned}
& th - \frac{(th)^3}{24 a_{\mathcal{M}}^2} \\
=& \rho\left(x, \exp_x(th\psi_i(x,\omega))\right) - \frac{\rho\left(x, \exp_x(th\psi_i(x,\omega))\right)^3}{24 a_{\mathcal{M}}^2} \\
\leq& \|\exp_x(th\psi_i(x,\omega)) - x\| \\
=& h \leq \rho\left(x, \exp_x(th\psi_i(x,\omega))\right) \\
=& th,
\end{aligned}$$

which implies

$$1 \leq t_h(x,v) \leq 1 + \frac{t^3 h^2}{24 a_{\mathcal{M}}^2} \leq 1 + \frac{\pi^3 h^2}{192 a_{\mathcal{M}}^2}.$$

In particular, using $\alpha \geq 6$

$$\begin{aligned}
& \sup_{\substack{i \in I \\ y \in \varphi_i^{-1}(\mathcal{U}_i) \\ 0 < h < r'_{\mathcal{M}}/2 \\ \|v\|_{T_{\varphi_i^{-1}(y)}\mathcal{M}} = 1}} \frac{\left|\frac{\partial F}{\partial t}(\varphi_i^{-1}(y), t_h(\varphi_i^{-1}(y), v)) - 2t_h(\varphi_i^{-1}(y), v)\right|}{h^2} \\
=& \sup_{\substack{i \in I \\ y \in \varphi_i^{-1}(\mathcal{U}_i) \\ 0 < h < r'_{\mathcal{M}}/2 \\ \|v\|_{T_{\varphi_i^{-1}(y)}\mathcal{M}} = 1}} \left|\frac{\partial G}{\partial t}\left(y, t_h(\varphi_i^{-1}(y), v)\right)\right| \\
\leq& M < +\infty
\end{aligned}$$

Since $2t_h(x,v) \geq 2$, there exists $0 < h_+ \leq r'_{\mathcal{M}}/2$ such that for any $i \in I, \quad h \leq h_+, \quad x \in \mathcal{U}_i, \quad \|v\|_{T_x\mathcal{M}} = 1$ we have $\frac{\partial F}{\partial t} \neq 0$. Hence the implicit function theorem implies that $y \mapsto t_h(\varphi_i(y), v)$ is smooth on $\varphi_i^{-1}(\mathcal{U}_i)$ with differential given by

$$\frac{\partial}{\partial y} t_h(\varphi_i(y), v) = -\frac{\frac{\partial F}{\partial y}}{\frac{\partial F}{\partial t}} = -\frac{h^2 \frac{\partial G}{\partial y}}{2t_h(\varphi_i(y), v) + h^2 \frac{\partial G}{\partial t}} \in \mathcal{L}\left(\mathbb{R}^d, \mathbb{R}\right)$$

To conclude notice first that since $t_h(x,v) \in [1, 1 + \frac{\pi^3 h^2}{192 a_{\mathcal{M}}^2}]$ we have, setting $s_h(x,v) = \frac{t_h(x,v)-1}{h^2}$

$$\sup_{\substack{x \in \mathcal{M} \\ 0<h<r'_{\mathcal{M}}/2 \\ \|v\|_{T_x\mathcal{M}}=1}} |s_h(x,v)| \leq \frac{\pi^3}{192 a_{\mathcal{M}}^2}$$

Moreover for the derivatives we get, for any $y \in \varphi_i^{-1}(\mathcal{U}_i)$

$$\begin{aligned}\frac{\partial}{\partial y} s_h(\varphi_i(y), v) = h^{-2} \frac{\partial}{\partial y} t_h(\varphi_i(y), v) &= -h^{-2} \frac{h^2 \frac{\partial G}{\partial y}}{2 t_h(\varphi_i(y), v) + h^2 \frac{\partial G}{\partial t}} \\ &= -\frac{\frac{\partial G}{\partial y}}{2 t_h(\varphi_i(y), v) + h^2 \frac{\partial G}{\partial t}} = -\frac{\frac{\partial G}{\partial y}}{2 + 2h^2 s_h(\varphi_i(y), v) + h^2 \frac{\partial G}{\partial t}}.\end{aligned}$$

Hence for any $y \in \varphi_i^{-1}(\mathcal{U}_i)$, $\quad \|v\|_{T_{\varphi_i(y)}\mathcal{M}} = 1, \quad 0 < h < h'_+ = h_+ \wedge M^{-1/2}$

$$\forall 1 \leq j \leq d, \quad \left| \frac{\partial}{\partial y_j} s_h(\varphi_i(y), v) \right| = \frac{\left| \frac{\partial G}{\partial y_j} \right|}{\left| 2 t_h(\varphi_i(y), v) + h^2 \frac{\partial G}{\partial t} \right|} \leq \frac{M}{2 - h^2 M} \leq M$$

In the same way by induction we can prove that

$$\max_{\alpha \in \mathbb{N}^d, |\alpha| \leq \alpha - 5} \sup_{\substack{i \in I \\ y \in \varphi_i^{-1}(\mathcal{U}_i) \\ 0<h<h'_+ \\ \|v\|_{T_{\varphi_i(y)}\mathcal{M}}=1}} \left| \frac{\partial^{|\alpha|}}{\partial y^\alpha} s_h(\varphi_i(y), v) \right| < +\infty$$

■

Remark 17. *We have stated our lemma with $r'_{\mathcal{M}}$ such that $\|\exp_x(rv) - x\|^2 \geq \frac{r^2}{4}$ for any $r \in [0, r'_{\mathcal{M}}]$, but more generally for any $c \in (0,1)$ we could have found $r'_{\mathcal{M},c} > 0$ such that $r \in [0, r'_{\mathcal{M},c}]$ implies $\|\exp_x(rv) - x\|^2 \geq c^2 r^2$.*

Appendix B. Geometrical and analytical properties of the random graph

In this section we prove/recall useful facts on the graph coming from geometrical properties of the manifold.

We start by a simple but useful property satisfied by the graph Laplacian eigenvectors :

Proposition 18. *For any $h > 0$, $\quad j \in \{1, \dots, N\}$ we have $\|u_j\|_{L^\infty(\nu)} \leq N$.*

Proof For all $x \in V$ we have $\nu_x = \frac{\mu_x}{\sum_{y \in V} \mu_y} \geq \frac{1}{N^2}$, which implies

$$\forall x \in V, u_j(x)^2 \leq \frac{1}{\nu_x} \sum_{y \in V} u_j(y)^2 \nu_y = \frac{1}{\nu_x} \|u_j\|^2_{L^2(\nu)} = \frac{1}{\nu_x} \leq N^2$$

■

As we will see below, we actually have $\nu_x \asymp N^{-1}$ on a high probability event, so that the result of proposition 18 could actually be improved to $\max_{1\leq j\leq N} \|u_j\|_{L^\infty(\nu)} \lesssim \sqrt{N}$, but this would not improve the final rates anyway.

In the next subsections we will be discussing various properties of the random geometric graph, namely volume regularity, on diagonal heat kernel bounds, Weyl type upper and lower bounds on the graph Laplacian eigenvalues and finally a norm comparison theorem. These properties will be used in the proof of Theorems 3,4,6 & 7. More precisely, the proof of the Kullback-Leibler prior mass condition (Lemma 36, which is then subsequently used in the proof of Theorems 3 & 6) uses a lower bound on the eigenvalues of the graph Laplacian $\mathcal{L}^{(h_n)}$ for a determined h_n (the one appearing in the statements of Theorems 3 & 6). To prove contraction rates with respect to $\|\cdot\|_N$ (Theorems 4 & 7), we combine the results for $\|\cdot\|_n$ together with Lemma 34, which states an inequality of the form : with $\mathbb{P}_0-$probability going to 1

$$\forall h_- \leq h \leq h_0, \quad J_0 \leq J \leq J_1 h^{-d}, \quad f \in \operatorname{span}\{u_1,\dots,u_J\}, \quad \|f\|_{L^\infty(\nu)} \lesssim \sqrt{J}\|f\|_{L^2(\nu)} \tag{11}$$

for some constants $h_0, J_0, J_1 > 0$, up to logarithmic factors, as long as h_- satisfies assumption 2. To prove theorem 7 (adaptive posterior contraction rates with respect to $\|\cdot\|_N$) we use inequality 11 for any h in the support of the posterior distribution. Thus in this section we analyse the properties of the random graph associated to the connectivity parameter h for h belonging to $[h_-, h_0]$. Note that these properties will also be used in the proof of the other results but for a specific value h_n in $[h_-, h_0]$.

B.1 Volume regularity

In this subsection we establish a volume regularity property for the random geometric graph : roughly speaking, with high probability, for every suitable r, h and geodesic ball of radius r we have $\nu^{(h)}(B) \asymp r^d$. To start with notice that since $\mathcal{M}$ is compact it has a positive radius of curvature $a_{\mathcal{M}}$ and in particular by lemma 3 in Bernstein, de Silva, Langford, and Tenenbaum for any $x, y \in \mathcal{M}$ with $\rho(x,y) \leq \pi a_{\mathcal{M}}$ we find

$$\frac{2}{\pi}\rho(x,y) \leq 2a_{\mathcal{M}} \sin\left(\rho(x,y)/2ra_{\mathcal{M}}\right) \leq \|x-y\| \leq \rho(x,y) \tag{12}$$

and, using $\sin(\rho/2a_{\mathcal{M}}) \geq \frac{\rho}{2ra_{\mathcal{M}}} - \frac{1}{6}\left(\frac{\rho}{2a_{\mathcal{M}}}\right)^3$,

$$\rho(x,y) - \frac{\rho(x,y)^3}{24a_{\mathcal{M}}^2} \leq \|x-y\| \leq \rho(x,y).$$

Moreover the set $\mathcal{M}^2 \backslash \left\{(x,y) \in \mathcal{M}^2 : \rho(x,y) < \pi a_{\mathcal{M}}\right\}$ is compact with respect to the topology inherited from the Euclidean distance or equivalently the geodesic distance. Hence the function $(x,y) \mapsto \frac{\rho(x,y)}{\|x-y\|}$ is bounded on $\mathcal{M}^2 \backslash \left\{(x,y) \in \mathcal{M}^2 : \rho(x,y) < \pi a_{\mathcal{M}}\right\}$ by some constant C_0, while it satisfies $\frac{\rho(x,y)}{\|x-y\|} \leq \frac{\pi}{2}$ when $\rho(x,y) \leq \pi a_{\mathcal{M}}$. Combining the two cases and assuming without loss of generality $C_0 \geq \frac{\pi}{2}$ we find the existence of $C_0 > 0$ such that $\|x-y\| \leq \rho(x,y) \leq C_0 \|x-y\|$ on $\mathcal{M}$.

Theorem 19. *Let h_- satisfy Assumption 2, i.e. $Nh_-^d >> \ln N$, and A_N be the event*

$$A_N = \left\{ \forall i = 1, \ldots, N, \forall h_- \leq r \leq diam_\rho(\mathcal{M})/C_0, c_- r^d \leq \frac{\mu_{X_i}^{(r)}}{N} \leq c_+ r^d \right\},$$

where $\mu_{X_i}^{(r)} = \mu(B(X_i, r))$. Then, for some $c_-, c_+, c > 0$ we have $\mathbb{P}(A_N) \geq 1 - e^{-cNh_-^d}$. Moreover, on A_N we have $\frac{c_-/c_+}{N} \leq \nu_y^{(h)} \leq \frac{c_+/c_-}{N}$ for any $h_- \leq h \leq diam_\rho(\mathcal{M})$ and $y \in V$.

A consequence of Theorem 19 is that on A_N for all $r \geq h_-$, $\quad \nu^{(h)}(B(X_i, r)) \geq r^d c_-^2 / c_+$ for all $i \leq N$.

Proof Using

$$\forall x \in \mathcal{M}, r > 0, B_\rho(x, r) \subset B_{\mathbb{R}^D}(x, r) \subset B_\rho(x, C_0 r),$$

Theorem 3.8 in Göbel and Blanchard (2020) implies that for some $b_-, b_+, c > 0$, for all $h_- \leq r \leq diam_\rho(\mathcal{M})/C_0$, $\quad diam_\rho(\mathcal{M}) := \max_{x,y \in \mathcal{M}} \rho(x, y)$, $\quad i = 1, \ldots, N$, with probability at least $1 - e^{-cNh_-^d}$

$$b_- r^d \leq \frac{\mu_{X_i}^{(r)} - 1}{N - 1} \leq b_+ r^d.$$

Using $\frac{Nh_-^d}{\ln N} \to \infty$ and a union bound we find that the last inequality holds simultaneously for all $i \in \{1, \ldots, N\}$ with probability at least $1 - e^{-cNh_-^d}$ for some $c > 0$. To make the result uniform in r, define $r_k = 2^k h_-$, $\quad k_+ = \lfloor \frac{\ln \frac{diam(\mathcal{M})}{h_-}}{\ln 2} \rfloor \leq c_{\mathcal{M}}^{-1} \ln N$ (without loss of generality for small $c_{\mathcal{M}}$, because $h_- >> \left(\frac{\ln N}{N}\right)^{1/d}$). Using a union bound we find that with probability $1 - (k_+ + 1) e^{-c_{\mathcal{M}} Nh_-^d} \geq 1 - c_{\mathcal{M}}^{-1} \ln N e^{-cNh_-^d}$, for every $i = 1, \ldots, N$ and $k = 0, \ldots, k_+ + 1$,

$$b_- r_k^d \leq \frac{\mu_{X_i}^{(r_k)} - 1}{N - 1} \leq b_+ r_k^d.$$

But then, for every $k = 0, \ldots, k_+ + 1$, $\quad r_k \leq r \leq r_{k+1}$ and $i \in \{1, \ldots, N\}$ we have (assuming $N \geq 2$)

$$\begin{aligned}
&\frac{N-1}{N} \frac{\mu_{X_i}^{(r)} - 1}{N-1} \leq \frac{\mu_{X_i}^{(r)}}{N} = \frac{1}{N} + \frac{\mu_{X_i}^{(r)} - 1}{N-1} \frac{N-1}{N} \\
\implies &\frac{N-1}{N} \frac{\mu_{X_i}^{(r_k)} - 1}{N-1} \leq \frac{\mu_{X_i}^{(r)}}{N} \leq \frac{1}{N} + \frac{\mu_{X_i}^{(r_{k+1})} - 1}{N-1} \frac{N-1}{N} \\
\implies &\frac{1}{2} \frac{\mu_{X_i}^{(r_k)} - 1}{N-1} \leq \frac{\mu_{X_i}^{(r)}}{N} \leq \frac{1}{N} + \frac{\mu_{X_i}^{(r_{k+1})} - 1}{N-1} \\
\implies &\frac{1}{4} b_- r_{k+1}^d = \frac{1}{2} b_- r_k^d \leq \frac{\mu_{X_i}^{(r)}}{N} \leq \frac{1}{N} + b_+ r_{k+1}^d = \frac{1}{N} + 2 b_+ r_k^d \\
\implies &\frac{1}{4} b_- r^d \leq \frac{\mu_{X_i}^{(r)}}{N} \leq h^d + 2 b_+ r^d \quad \implies \frac{1}{4} b_- r^d \leq \frac{\mu_{X_i}^{(r)}}{N} \leq (2b_+ + 1) r^d.
\end{aligned}$$

Therefore, using again $\frac{Nh_-^d}{\ln N} \to \infty$ with probability $1 - e^{-cNh_-^d}$, for every $i = 1, \dots, N$ and $h_- \le r \le diam_\rho(\mathcal{M})$,

$$\frac{1}{4} b_- r^d \le \frac{\mu_{X_i}^{(r)}}{N} \le (2b_+ + 3)\, r^d,$$

which proves the first statement with $c_- = \frac{1}{4} b_-, c_+ = 2b_+ + 3$. The second statement is an immediate consequence of the first one. Indeed, using again

$$\forall x \in \mathcal{M}, \quad r > 0, \quad B_\rho(x, r) \subset B_{\mathbb{R}^D}(x, r) \subset B_\rho(x, C_0 r)$$

we get for any $y \in V$ and $h_- \le h \le diam_\rho(\mathcal{M})/C_0$,,

$$\frac{c_-/c_+}{N} = \frac{c_- N r^d}{c_+ N^2 r^d} \le \nu_y^{(h)} = \frac{\mu_y^{(h)}}{\sum_{x \in V} \mu_x^{(h)}} \le \frac{c_+ N r^d}{c_- N^2 r^d} = \frac{c_+/c_-}{N}.$$

■

In what follows we take $h_0 \le diam_\rho(\mathcal{M})\,, \quad r_0 \le diam_\rho(\mathcal{M})/C_0$ (that we will actually reduce along the proof) and we fix h_- satisfying assumption 2 in order to work on the corresponding A_N.

B.2 Heat kernel bounds

The goal of this section is to show the following result : we consider A_N and h_- as defined in Theorem 19.

Theorem 20. *There exist constants $a_0, a_1, a_2, a_3, h_0 > 0$ (that also depend on $\mathcal{M}, p_0$) such that, on A_N, for all $h_- \le h \le h_0$ and $t_0(h) := a_0 h^2 \ln\left(Nh^d\right) \le t \le \frac{a_1}{\ln^2 N}$*

$$\forall x \in V, \quad a_2 \frac{t^{-d/2}}{\ln^d N} \le p_t^{(h)}(x, x) \le a_3 t^{-d/2}.$$

This result is a very weak form (it is often called an "on diagonal" upper bound) of the heat kernel bound used in Castillo, Kerkyacharian, and Picard (2014); Coulhon, Kerkyacharian, and Petrushev (2012) over continuous spaces, but fortunately will be enough for our purposes. It is potentially possible to apply further techniques in Barlow (2017) and get full off diagonal bounds (i.e, bounds on $p_t^{(h)}(x, y), x \ne y$) yielding a Gaussian type behaviour of the heat kernel, but we won't need this. It should be noted that the restriction $t \gtrsim h^2$ (up to a logarithmic factor) for the heat kernel bound to be valid substantially complicates the analysis of the situation. Indeed, if instead we could prove the bound for every $0 < t \le \frac{a_1}{\ln^2 N}$ then we could conduct our analysis by combining the techniques from Coulhon and Sikora (2008); Coulhon, Kerkyacharian, and Petrushev (2012), our approximation results and the proof strategy of Castillo, Kerkyacharian, and Picard (2014). Since we require $t \gtrsim h^2 \ln\left(Nh^d\right)$ with h not arbitrarily small (otherwise our approximation results 1112 become vacuous) this is not possible and we have to prove things differently.

As in section 2, we emphasize the dependence of the graph (and in particular the graph Laplacian, the heat kernel and its eigendecomposition) with respect to h with a (h) exponent. While the notation become heavier, this is important to keep in mind as theorem 20 deals with a high probability control of the heat kernels associated with different values of *h simultaneously.*

B.2.1 On diagonal upper bound up to t_1

Lemma 21. *With A_N, c_-, c_+ the event and constant defined in Theorem 19 we have, on A_N*

$$\forall h_- \leq h \leq diam_\rho(\mathcal{M}), \quad t \geq 16h^2 \ln\left(2c_- N h^d\right), \quad x, y \in V, \quad p_t^{(h)}(x,y) \leq 2\frac{c_+}{c_-} h^{-d}$$

Proof Recall that for all $x \in V, \quad t > 0, \quad \nu_y p_t^{(h)}(x,y) = \mathbb{P}_x^{(h)}(W_t = y)$. On the event A_N we have $\nu_y^{(h)} \geq \frac{c_-}{c_+ N}$ for any $h_- \leq h \leq h_0$, which implies

$$\forall x, y \in V, \quad p_t^{(h)}(x,y) \leq \frac{c_+}{c_-} N \mathbb{P}_x^{(h)}(W_t = y) = \frac{c_+}{c_-} N \sum_{l \geq 0} \mathbb{P}\left(N_{t/h^2} = l\right) \mathbb{P}_x^{(h)}(Y_l = y)$$

where the last equality uses the independence of N and Y. Moreover, using Lemma 5.13 (e) from Barlow (2017) we find

$$\forall \rho \geq 2, \quad \mathbb{P}\left(N_\rho \leq \frac{\rho}{2}\right) \leq \mathbb{P}\left(|N_\rho - \rho| \geq \frac{\rho}{2}\right) \leq 2e^{-\rho/16}$$

which implies for any $x, y \in V$,

$$\begin{aligned} p_t^{(h)}(x,y) \leq& \frac{c_+}{c_-} N \mathbb{P}_x^{(h)}(W_t = y) \\ \leq& \frac{c_+}{c_-} N \left\{ 2e^{-t/16h^2} + \sum_{l \geq t/2h^2} \mathbb{P}\left(N_{t/h^2} = l\right) \mathbb{P}_x^{(h)}(Y_l = y) \right\} \\ \leq& \frac{c_+}{c_-} N \left\{ 2e^{-t/16h^2} + \sup_{l \geq 1} \mathbb{P}_x^{(h)}(Y_l = y) \right\} \end{aligned}$$

For each $l \geq 1$ we have $\mathbb{P}_x^{(h)}(Y_l = y) = \mathbb{E}_x^{(h)}\left[\mathbb{P}_x^{(h)}(Y_l = y | Y_{l-1})\right]$, and given Y_{l-1} the probability of going from Y_{l-1} to $Y_l = y$ is by definition 0 (if Y_{l-1} and y are not neighbours) or $\frac{1}{\mu_x^{(h)}}$ (if Y_{l-1} and y are neighbours). In either case on A_N we have $\mu_x^{(h)} \geq c_- N h^d$ which implies

$$\forall x, y \in V, \quad p_t^{(h)}(x,y) \leq \frac{c_+}{c_-} N \left\{ 2e^{-t/16h^2} + \frac{1}{c_- N h^d} \right\}.$$

All in all on A_N we get

$$\forall h_- \leq h \leq h_0, \quad t \geq 16h^2 \ln\left(2c_- N h^d\right), \quad x, y \in V, \quad p_t^{(h)}(x,y) \leq 2\frac{c_+}{c_-} h^{-d}.$$

■

The goal is now, roughly speaking, to refine the inequality $p_t^{(h)}(x,y) \lesssim h^{-d}$ to $p_t^{(h)}(x,y) \lesssim t^{-d/2}$ (up to a logarithmic factor) by using the approach of Barlow (2017). This requires a local Poincare inequality, i.e an inequality of the form:

$$\forall x \in V, \quad r \in [r_-, r_+], \quad B = B_\rho(x,r), \quad f : B \to \mathbb{R},$$

$$\sum_{y \in B} (f(y) - f_B)\, \nu_y^{(h)} \le Cr^2 \sum_{x,y \in B, y \sim x} (f(x) - f(y))^2, \quad \text{where } f_B = \frac{\sum_{y \in B} f(y) \nu_y^{(h)}}{\nu^{(h)}(B)}.$$

where the constants r_-, r_+, C need to be controlled. Göbel and Blanchard (2020) provide a way to prove such an inequality using the construction of random Hamming paths, however we give below a detailed application of their results to our setting in order to control the dependence of the constant C of the inequality in the parameters N, h.

Theorem 22. *Poincare inequality*
Let $A'_N = \left\{G^{(h_-)} \text{ is connected}\right\}$. Then there exist constants

$$r_0 = r_0(\mathcal{M}), \quad h_0 = h_0(\mathcal{M}), \quad C = C(\mathcal{M}) > 0$$

such that on A_N, for any $x \in V, \quad 0 < r < r_0, \quad h_- \le h < h_0$ and $f : B = B_\rho(x,r) \to \mathbb{R}$, on $A_N \cap A'_N$ we have

$$\sum_{y \in B} (f(y) - f_B)\, \nu_y^{(h)} \le C \frac{(r/h)^2}{N^2 h^d} \sum_{z \in B, y \in B, y \sim x} (f(z) - f(y))^2$$

where $f_B = \sum_{y \in B} f(y) \nu_y^{(h)} / \nu^{(h)}(B)$. Moreover, $\mathbb{P}_0(A_N \cap A'_N) \ge 1 - e^{-cNh_-^d}$ for some $c > 0$.

Proof Let $h_- > 0$ and A_N, A'_N defined accordingly. The event $A_N \cap A'_N$ satisfies $\mathbb{P}_0(A_N \cap A'_N) \ge 1 - e^{-cNh_-^d}$ for some $c > 0$, see remark 23 below. For any $h \ge h_-, \quad x \in V, \quad B = B_\rho(x,r)$, Corollary 5.6 in Göbel and Blanchard (2020) shows that

$$\forall f : B \to \mathbb{R}, \quad \sum_{y \in B} (f(y) - f_B)\, \nu_y^{(h)} \le \tilde{\kappa}_B r^2 \sum_{z \in B, y \in B, y \sim x} (f(z) - f(y))^2$$

where

$$\tilde{\kappa}_B = \frac{\max_{z \in B} \nu^{(h)}(y)^2}{2\nu^{(h)}(B)} l_{\max}\left(\tilde{\Gamma}_B\right) b_{\max}\left(\tilde{\Gamma}_B\right)$$

and $l_{\max}\left(\tilde{\Gamma}_B\right), b_{\max}\left(\tilde{\Gamma}_B\right)$ are quantities defined in Göbel and Blanchard (2020). Notice that Göbel and Blanchard (2020) requires connectedness of the subgraph restricted to B. This is shown in remark 23 below.

By following the proof of Corollary 5.13 in Göbel and Blanchard (2020) we find the existence of $C = C(\mathcal{M}, p_0), \quad r_0 = r_0(\mathcal{M}), \quad h_0 = h_0(\mathcal{M}) > 0$ such that on A_N, for any $0 < r < r_0, \quad x \in V, \quad B = B_\rho(x,r)$

$$l_{\max}\left(\tilde{\Gamma}_B\right) \le Ch^{-1}, b_{\max}\left(\tilde{\Gamma}_B\right) \le Ch^{-(d+1)}$$

(in Göbel and Blanchard (2020) the result is proved with high probability over a single ball, but an inspection of the proof shows that the only thing needed is a control of the number of points in the balls on the graph for different values of r and h, which is precisely what A_N is for as shown in lemma 19, hence the uniformity). This shows the desired inequality

$$\sum_{y\in B}\left(f(y)-f_B\right)\nu_y^{(h)} \lesssim \frac{(r/h)^2}{N^2h^d}\sum_{z\in B, y\in B, x\sim y}\left(f(z)-f(y)\right)^2$$

■

Remark 23. *We show the following simple result : if h_- satisfies assumption 2 then the probability that the resulting random geometric graph $G^{(h_-)}$ is connected (i.e the probability of the event A'_N) converges to* 1. *In particular, by monotonicity the probability that all random geometric graphs with connectivity parameter $h \geq h_-$ are connected converges to* 1 *as well.*

Indeed, since $\mathcal{M}$ is connected, consider $\{y_1,\dots,y_p\}$ an $h/8-$net of $\mathcal{M}$ with respect to the Euclidean distance $\|\cdot\|$. By standard arguments and compactness of $\mathcal{M}$ we can assume that $p=\mathcal{O}(h^{-d})$. Then the event

$$A'_n = \{\forall l,\ \exists i_l,\ \|y_l - x_{i_l}\| < h/8\}$$

satisfies

$$\mathbb{P}\left((A'_n)^c\right) \lesssim p\left(1-ch^d\right)^N = \mathcal{O}\left(\frac{1}{h^d}e^{-c8^{-d}Nh^d}\right) \lesssim e^{-c'Nh^d}$$

for some $c'>0$ by assumption 2. Here $c>0$ is small enough such that $\int_{B_{\mathbb{R}^D}(x,h)} p_0(x)\mu(dx) \geq ch^d$ for all $x\in\mathcal{M}$ (c exists since $p_0 \geq p_{min} > 0$).

Moreover by connectedness of $\mathcal{M}$ (which implies path connectedness since $\mathcal{M}$ is a submanifold of $\mathbb{R}^D$) for each $i,j\in\{1,\dots,N\}$ there exists a continuous path $c:[0,1]\to\mathcal{M}$ with $c(0)=x_i, c(1)=x_j$. Hence for each $t\in[0,1]$ there exists $l(t)\in\{1,\dots,p\}$ such that $\left\|c(t)-y_{l(t)}\right\| < h/8$, and therefore on A'_n for each $t\in[0,1]$ there exists $i_t = i_{l(t)}$ such that

$$\|c(t)-x_{i_t}\| \leq \left\|c(t)-y_{l(t)}\right\| + \left\|y_{l(t)} - x_{i_{l(t)}}\right\| < h/4$$

By continuity of c, this implies the existence of $K\geq 1$ and $1\leq i_k \leq N$ for $k=1,\dots,K$ such that $\left\|x_{i_k}-x_{i_{k+1}}\right\| \leq h/2$ for any $1\leq k<K$ and such that $c(0)=x_i\sim x_{i_1},\quad c(1)=x_j=x_{i_K}$. Therefore the path $x_i, x_{i_1},\dots,x_{i_{K-1}},x_j$ connects x_i and x_j in the graph which implies connectedness on the event A'_n.

Notice that this is a very basic and coarse result about connectivity and more generally connected components of random geometric graphs. More details and results can be found in Penrose (2003).

With the Poincare inequality 22 we can now apply the proof strategy of Barlow (2017). We first prove an upper bound of the form

$$\forall x,y\in V,\quad p_t^{(h)}(x,y)\lesssim t^{-d/2}$$

for $t \gtrsim h^2\ln\left(Nh^d\right)$ (up to a logarithmic factor) but less than a random time $t_1=t_1(x,h)$ that we shall lower bound adequately later.

Lemma 24. *There exists $h_0 = h_{0|\mathcal{M},p_0}, \quad K = K_{\mathcal{M},p_0}, \quad K' = K'_{\mathcal{M},p_0} > 0$ such that, on A_N, for every $h_- \leq h \leq h_0, \quad x_0 \in V$, with the random times*

$$t_1(x,h) = \inf\left\{t > 0 : p_{2t}^{(h)}(x,x) \leq K'\right\}$$

and

$$t_1(h) = \min_{x \in V} t_1(x,h),$$

we have, with $t_0(h) = 64h^2 \ln\left(2c_- N h^d\right)$,

$$\forall h_- \leq h < h_0, \quad t \in [t_0(h), 2t_1(h)], \quad x,y \in V, \quad p_t^{(h)}(x,y) \leq K t^{-d/2}.$$

Proof First consider the case $x = y$. Take $x_0 \in V$ and let $\phi(t) = p_{2t}^{(h)}(x_0,x_0) = \sum_{x\in V} f_t(x)^2 \nu_x^{(h)}, \quad f_t(x) = p_t^{(h)}(x_0,x)$. Then since $\nu_y^{(h)}\mathcal{L}_{yx} = \nu_x^{(h)}\mathcal{L}_{xy}$, $\frac{d}{dt}f_t = -\mathcal{L}f_t$ we have

$$-\phi'(t) = 2\left\langle f_t | \mathcal{L} f_t\right\rangle_{L^2(\nu)} = \frac{1}{\mu(V)\,h^2}\sum_{x\sim y}\left(f_t(x) - f_t(y)\right)^2 \geq \frac{1}{c_+ N^2 h^{d+2}}\sum_{x\sim y}\left(f_t(x)-f_t(y)\right)^2,$$

where we have used theorem 19 to bound $\mu(V) \leq c_+ N^2 h^d$. This proves that ϕ is non-increasing. Taking a covering of V by balls $B_i = B_\rho(x_i, r), \quad h_- \leq r \leq diam_\rho(\mathcal{M})/C_0$ given by lemma 25 below, since each x_i belongs to at most $M = 3^d (c_+/c_-)^3$ of the balls we have

$$-\phi'(t) \geq \frac{1}{c_+ N^2 h^{d+2}}\sum_{x\sim y}\left(f_t(x) - f_t(y)\right)^2 \geq \frac{1}{c_+ M N^2 h^{d+2}}\sum_i \sum_{x\sim y \in B_i}\left(f_t(x) - f_t(y)\right)^2.$$

On A_N, if $h_- < r < r_0$, $h_- \leq h < h_0$ and using Theorem 22 in each of the balls we find

$$-\phi'(t) \geq \frac{r^{-2}}{MC_{c_+}}\sum_i \sum_{x\in B_i}\left(f_t(x) - f_{t|B_i}\right)^2 \nu_x^{(h)}.$$

Defining $\nu_{B_i}^{(h)} = \frac{\nu_{|B_i}^{(h)}}{\nu^{(h)}(B_i)}$ the normalised restriction of $\nu^{(h)}$ to B_i and $Var(X) = EX^2 - (EX)^2$ we get

$$\begin{aligned}
-\phi'(t) \geq& \frac{r^{-2}}{MCc_+}\sum_i\sum_{x\in B_i}\left(f_t(x) - f_{t|B_i}\right)^2\nu_x^{(h)} \\
=& \frac{r^{-2}}{MCc_+}\sum_i \nu^{(h)}(B_i)\sum_{x\in B_i}\left(f_t(x) - f_{t|B_i}\right)^2 \nu_{B_i}^{(h)}(x) \\
=& \frac{r^{-2}}{MCc_+}\sum_i \nu^{(h)}(B_i)\left\{\sum_{x\in B_i} f_t(x)^2 \nu_{B_i}^{(h)}(x) - \left(\sum_{x\in B_i} f_t(x)\nu_{B_i}^{(h)}(x)\right)^2\right\} \\
=& \frac{r^{-2}}{MCc_+}\sum_i\left\{\sum_{x\in B_i} f_t(x)^2\nu_x - \nu^{(h)}(B_i)\left(\sum_{x\in B_i} f_t(x)\nu_{B_i}^{(h)}(x)\right)^2\right\}.
\end{aligned}$$

But the balls form a covering of V, therefore

$$-\phi'(t) \geq \frac{r^{-2}}{MCc_+}\left\{\sum_{x\in V} f_t(x)^2\nu_x - \sum_i \nu^{(h)}(B_i)\left(\sum_{x\in B_i} f_t(x)\nu^{(h)}_{B_i}(x)\right)^2\right\}.$$

We have

$$\sum_{x\in V} f_t(x)^2\nu^{(h)}_x = \phi(t)$$

and, still on A_N

$$\begin{aligned}
\sum_i \nu^{(h)}(B_i)\left(\sum_{x_j\in B_i} f_t(x_j)\nu^{(h)}_{B_i}(x_j)\right)^2 &= \sum_i \nu^{(h)}(B_i)^{-1}\left(\sum_{x_j\in B_i} f_t(x_j)\nu^{(h)}_{x_j}\right)^2\\
&\leq \sum_i \frac{c_+}{c_-^2} r^{-d}\left(\sum_{x_j\in B_i} f_t(x_j)\nu^{(h)}_{x_j}\right)^2 \leq \frac{c_+}{c_-^2} r^{-d}\left(\sum_i\sum_{x_j\in B_i} f_t(x_j)\nu^{(h)}_{x_j}\right)^2\\
&=\frac{c_+}{c_-^2} r^{-d}\left(\sum_{x_j\in V} f_t(x_j)\nu^{(h)}_{x_j}\sum_{i:x_j\in B_i} 1\right)^2\\
&\leq\frac{c_+}{c_-^2} r^{-d}\left(\sum_{x_j\in V} f_t(x_j)\nu^{(h)}_{x_j} M\right)^2 = \frac{M^2c_+}{c_-^2} r^{-d}
\end{aligned}$$

where $h_- \leq r \leq r_0, \quad h_- \leq h \leq h_0$ and we have used Theorem 19 to bound $\nu^{(h)}(B_i) \geq r^d c_-^2/c_+$. Hence

$$-\phi'(t) \geq \frac{r^{-2}}{MCc_+}\left\{\phi(t) - \frac{c_+M^2}{c_-^2} r^{-d}\right\},$$

as long as $h_- \leq r \leq r_0, \quad h_- \leq h \leq h_0$. Let $r = r(t) = \left(\frac{\phi(t)}{K}\right)^{-1/d}$, $K = 2c_+M^2/c_-^2 \vee 2c_+/c_-$. Define $t_1(x,h)$ the random time

$$t_1(x,h) = \inf\{t \geq 0 : r(t) \geq r_0\} = \inf\left\{t \geq 0 : \phi(t) \leq Kr_0^{-d}\right\}, \quad t_1(h) = \inf_x t_1(x,h);$$

now, using Lemma 21 and since $K \geq 2c_+/c_-$

$$\forall t \geq t_0(h)/2, \quad \phi(t) \leq \frac{2c_+}{c_-} h^{-d} \leq Kh^{-d} \leq Kh_-^{-d},$$

which in turns implies that $r(t) \geq h_-$. Moreover by definition of $t_1(h)$ we have

$$\forall t \leq t_1, \quad \phi(t) \geq Kr_0^{-d} \quad \text{so that} \quad r(t) \leq r_0.$$

Therefore $h_- \leq r(t) \leq r_0$ whenever $t_0(h)/2 \leq t \leq t_1(h)$.

Then on A_N we obtain

$$\forall t \in [t_0(h)/2, t_1(h)], \quad -\phi'(t) \geq \frac{(\phi(t)/K)^{2/d}}{MCc_+}\frac{1}{2}\phi(t) = \frac{1}{2MCK^{2/d}c_+}\phi(t)^{1+2/d}.$$

Integrating yields, for any $t_1(h) \geq t \geq t_0(h)/2$

$$\left[\frac{d}{2}\phi^{-2/d}\right]_{t_0(h)/2}^{t} \geq \frac{1}{2MCK^{2/d}c_+^2}\left(t - t_0(h)/2\right) \geq \frac{t}{4MCK^{2/d}c_+^2}$$

which implies in particular

$$\frac{d}{2}\phi(t)^{-2/d} \geq \frac{d}{2}\phi\left(t_0(h)/2\right)^{-2/d} + \frac{t}{4MCK^{2/d}c_+^2} \geq \frac{t}{4MCK^{2/d}c_+^2}$$

i.e

$$\forall t \in [t_0(h)/2, t_1(h)], \quad \phi(t) \leq \left(\frac{MCc_+^2K^{2/d}}{dt}\right)^{d/2}.$$

To conclude we apply Cauchy-Schwarz inequality to find

$$\begin{aligned}
\forall x, y \in V, \quad p_t(x,y) &= \sum_{j=1}^{N} e^{-t\lambda_j}u_j(x)u_j(x) \\
&\leq \left(\sum_{j=1}^{N} e^{-t\lambda_j}u_j(x)^2\right)^{1/2}\left(\sum_{j=1}^{N} e^{-t\lambda_j}u_j(y)^2\right)^{1/2} \\
&= \sqrt{p_t(x,x)p_t(y,y)} \\
&\leq \max_{x\in V} p_t(x,x)
\end{aligned}$$

■

Lemma 25. *For all $h_- \leq h \leq h_0, \quad h \leq r \leq r_0$, on A_N there exists a covering of V by balls $B_\rho(x_i, r)$ such that the balls $B_\rho(x_i, r/2)$ are disjoints and each x_i belongs to at most $3^d\left(c_+/c_-\right)^3$ balls $B_\rho(x_j, r)$.*

Proof Take a maximal packing $\left(B_\rho(x_i, r/2)\right)$. Then the balls $B_\rho(x_i, r/2)$ cover V : indeed, if there exists $x_j \in V$ such that $\rho(x_i, x_j) \geq r$ for all i, then this contradicts the maximal packing property by considering the ball $B_\rho(x_j, r/2)$. Moreover if $y \in V$ belong to M of the balls $B_\rho(x_i, r)$ (and we may assume without loss of generality that these balls are those corresponding to $i = 1, \dots, M$), using theorem 19 we have

$$\begin{aligned}
\frac{Mc_-^2}{c_+N}(r/2)^d \leq & M\min_i \nu\left(B_\rho(x_i, r/2)\right) \leq \sum_{i=1}^{M}\nu\left(B_\rho(x_i, r/2)\right) = \nu\left(\bigsqcup_{i=1}^{M} B_\rho(x_i, r/2)\right) \\
\leq & \nu\left(\bigsqcup_{i=1}^{M} B_\rho(y, 3/2r)\right) = \nu\left(B_\rho(y, 3/2r)\right) \leq \frac{c_+^2}{c_-N}(3/2r)^d,
\end{aligned}$$

which implies the result. ■

B.2.2 Escape from the origin for the Markov process W_t

In this section we find an upper bound on the quantities $m(t,x) = \mathbb{E}_x^{(h)}\left[\rho\left(x, W_t\right)\right]$. Let $q(t,x) = \mathbb{E}_x^{(h)}\left[\ln p_t^{(h)}(x, W_t)\right] = -\sum_{y\in V} p_t^{(h)}(x,y)\ln p_t^{(h)}(x,y)\nu_y^{(h)}$ as well as $M(s,x) = m(st_0,x)/\sqrt{t_0}$, $\quad Q(s,x) = q(st_0,x)$, $\quad t_0(h) = 64h^2\ln\left(2c_- N h^d\right)$.

Lemma 26. *For all $x \in V$,*

1.
$$\forall t>0, \quad q'(t,x) \geq \frac{1}{4C_0^2} m'(t,x)^2, \qquad \textit{and} \qquad \forall s \geq 1,\ Q'(s,x) \geq \frac{1}{4C_0^2} M'(s)^2$$

2. *On A_N we have*
$$\forall 1 \leq s \leq t_1(h)/t_0(h), \quad Q(s,x) \geq -\left(\ln K - \frac{d}{2}\ln\left(st_0(h)\right)\right) = \frac{d}{2}\ln s + \frac{d}{2}\ln t_0(h) - \ln K$$

3.
$$1 + M(s,x) \geq e^{-\frac{e+1}{ed}} e^{Q(s)/d}$$

Proof

1. Let $f_t(y) = p_t^{(h)}(x,y)$, $\quad b_t(y,z) = f_t(y) + f_t(z)$. Throughout the proof we write $m(t), M(t), q(t), Q(t)$ in place of $m(t,x), M(t,x), q(t,x), Q(t,x)$ and $t_0 = t_0(h)$ for shortness sake. We have

$$\begin{aligned}
&|m'(t)| \\
=&\left|\frac{\partial}{\partial t}\sum_{y\in V}\rho\left(x,y\right)f_t(y)\nu_y^{(h)}\right| = \left|\sum_{y\in V}\rho\left(x,y\right)\left(\mathcal{L}f_t\right)(y)\nu_y^{(h)}\right| \\
=&\left|\left\langle \rho\left(x,\cdot\right)|\mathcal{L}f_t\right\rangle_{L^2(\nu)}\right| = \left|\frac{1}{2h^2\mu\left(V\right)}\sum_{y\sim z}\left(\rho(x,y)-\rho(x,z)\right)\left(f_t(y)-f_t(z)\right)\right| \\
\leq&\frac{C_0}{h\mu\left(V\right)}\sum_{y\sim z}|f_t(y)-f_t(z)| = \frac{C_0}{h\mu\left(V\right)}\sum_{y\sim z}\frac{|f_t(y)-f_t(z)|}{|f_t(y)+f_t(z)|^{1/2}}\left(f_t(y)+f_t(z)\right)^{1/2} \\
\leq&\frac{C_0}{h\mu\left(V\right)}\left\{\sum_{y\sim z}\frac{\left(f_t(y)-f_t(z)\right)^2}{f_t(y)+f_t(z)}\right\}^{1/2}\left\{\sum_{y\sim z}\left(f_t(y)+f_t(z)\right)\right\}^{1/2} \\
=&\frac{C_0}{h\mu\left(V\right)}\left\{\sum_{y\sim z}\frac{\left(f_t(y)-f_t(z)\right)^2}{f_t(y)+f_t(z)}\right\}^{1/2}\left\{2\mu\left(V\right)\right\}^{1/2} = \frac{C_0\sqrt{2}}{h\mu\left(V\right)^{1/2}}\left\{\sum_{y\sim z}\frac{\left(f_t(y)-f_t(z)\right)^2}{f_t(y)+f_t(z)}\right\}^{1/2}.
\end{aligned}$$

Now using the inequality (page 157 in Barlow (2017))

$$\forall u,v>0, \quad \frac{(u-v)^2}{u+v} \leq (u-v)\left(\ln u - \ln v\right)$$

We find

$$m'(t)^2 \leq \frac{2C_0^2}{h^2\mu\left(V\right)}\sum_{y\sim z}\left(f_t(y)-f_t(z)\right)\left(\ln f_t(y) - \ln f_t(z)\right)$$

On the other hand

$$\begin{aligned}
q'(t) =& \frac{\partial}{\partial t} \sum_{y \in V} -f_t(y) \ln f_t(y) \nu_y^{(h)} \\
=& -\sum_{y \in V} \left\{ \left(-\mathcal{L} f_t\right)(y) \ln f_t(y) + f_t(y) \frac{\left(-\mathcal{L} f_t\right)(y)}{f_t(y)} \right\} \nu_y^{(h)} \\
=& \sum_{y \in V} \left(\mathcal{L} f_t\right)(y) \left\{1 + \ln f_t(y)\right\} \nu_y^{(h)} \\
=& \frac{1}{2h^2 \mu(V)} \sum_{y \sim z} \left(f_t(y) - f_t(z)\right) \left(\ln f_t(y) - \ln f_t(z)\right)
\end{aligned}$$

Hence

$$q'(t) \geq \frac{1}{4C_0^2} m'(t)^2$$

To conclude we use the fact that $Q'(s) = t_0 q'(st_0), \quad M'(s) = m'(st_0)\sqrt{t_0}$ which implies via $t = st_0$

$$q'(t) \geq \frac{1}{4C_0^2} m'(t)^2 \quad \text{and} \quad Q'(s) \geq \frac{t_0}{4C_0^2} \left[M'(s)/\sqrt{t_0}\right]^2 = \frac{1}{4C_0^2} M'(s)^2$$

2. Using lemma 24, for any $h_- \leq h \leq h_0, \quad x, y \in V$

$$\forall 1 \leq s \leq t_1(h)/t_0, \quad p_{st_0}(x, y) \leq K \left(st_0\right)^{-d/2}.$$

Therefore

$$\forall 1 \leq s \leq t_1(h)/t_0, \quad Q(s) \geq -\left(\ln K - \frac{d}{2} \ln\left(st_0\right)\right) = \frac{d}{2} \ln s + \frac{d}{2} \ln t_0 - \ln K$$

3. Let $a \leq 1, \quad b \geq 0$. Then

$$\begin{aligned}
-Q(s) + aM(s) + b =& \mathbb{E}_x \left[\ln p_{st_0}^{(h)}(x, W_{st_0}) + a\rho\left(x, W_{st_0}\right)/\sqrt{t_0} + b\right] \\
=& \sum_{y \in V} \nu_y^{(h)} p_{st_0}^{(h)}(x, y) \left\{\ln p_{st_0}^{(h)}(x, y) + a\rho\left(x, y\right)/\sqrt{t_0} + b\right\}
\end{aligned}$$

Using the inequality $u\left(\ln u + \lambda\right) \geq -e^{-1-\lambda}$ and setting $a = \left(1 + M(s)\right)^{-1}, \quad b = d \ln\left(1 + M(s)\right)$ we find

$$\begin{aligned}
-Q(s) + aM(s) + b =& \sum_{y \in V} \nu_y^{(h)} p_{st_0}^{(h)}(x, y) \left\{\ln p_{st_0}^{(h)}(x, y) + a\rho\left(x, y\right)/\sqrt{t_0} + b\right\} \\
\geq& -\sum_{y \in V} \exp\left\{-1 - a\rho\left(x, y\right)/\sqrt{t_0} - b\right\} \nu_y^{(h)} \\
=& -e^{-1-b} \sum_{y \in V} \exp\left\{-a\rho\left(x, y\right)/\sqrt{t_0}\right\} \nu_y^{(h)} \\
\geq& -e^{-1-b} \sum_{y \in V} \nu_t^{(h)} = -e^{-1-b} \geq -e^{-1}.
\end{aligned}$$

Hence

$$-Q(s) + 1 + d\ln(1 + M(s)) \geq -Q(s) + \frac{M(s)}{1 + M(s)} + d\ln(1 + M(s)) \geq -e^{-1}$$

and rearranging yields

$$1 + M(s) \geq e^{-\frac{e+1}{ed}} e^{Q(s)/d}$$

$\blacksquare$

Lemma 27. *There exists $c = c_{\mathcal{M},p_0,d}$ such that, on A_N we have, for every $h_- \leq h \leq h_0$ and $t_0(h) \leq t \leq t_1(h)$*

$$m(t,x) \leq c \ln N\sqrt{t}, \quad \forall x \in V$$

Proof We follow the proof of Lemma 6.13 in Barlow (2017) (it is stated for $s \geq 1$, but the same proof remains valid for $1 \leq s \leq t_1(h)/t_0$). As done in the previous computations we drop x in the notations $m(t,x), M(t,x), q(t,x), Q(t,x)$. From Lemma 26 we bound $M'(s) \leq 2C_0\sqrt{Q'(s)}$ and defining $r(s) = \frac{1}{d}\left(Q(s) + \ln K - \frac{d}{2}\ln(st_0(h))\right) \geq 0$, we bound $\forall 1 \leq s \leq t_1(h)/t_0(h)$,

$$\begin{aligned}
e^{-\frac{e+1}{ed}} K^{-1/d}\sqrt{st_0(h)}e^{r(s)} \leq& 1 + M(s) \\
\leq& 1 + M(1) + 2C_0\sqrt{d}\int_1^s \left(r'(\sigma) + \frac{1}{2\sigma}\right)^{1/2} d\sigma \\
\leq& 1 + C_0\sqrt{t_0(h)}/h + 2C_0\sqrt{2ds}\,(1 + r(s)) \\
=& 1 + 8C_0\sqrt{\ln(2c_- Nh^d)} + 2C_0\sqrt{2ds}\,(1 + r(s))
\end{aligned}$$

where we have used the bound

$$m(t_0) = \mathbb{E}_x^{(h)}[\rho(x, W_{t_0(h)})] \leq C_0 h\mathbb{E}(N_{t_0(h)/h^2}) = \frac{C_0 t_0(h)}{h}.$$

Dividing by $\sqrt{s} \geq 1$ we get

$$\left\{e^{-\frac{e+1}{ed}} K^{-1/d}\sqrt{t_0}\right\} e^{r(s)} \leq \frac{1 + M(s)}{\sqrt{s}} \leq \left(1 + C_0\sqrt{64\ln(2c_- Nh^d)}\right) + 2C_0\sqrt{2d}\,(1 + r(s))$$

In particular, weakening the constants we find c that depend on d, K, C_0, c_- such that

$$\begin{aligned}
e^{r(s)} \leq ct_0^{-1/2}\left(1 + \sqrt{\ln(Nh^d)} + r(s)\right) \iff& \frac{e^{1+\sqrt{\ln\left(Nh^d\right)}+r(s)}}{1 + \sqrt{\ln(Nh^d)} + r(s)} \leq cet_0^{-1/2}\exp\left(\sqrt{\ln(Nh^d)}\right) \\
\iff& \underbrace{\frac{e^{\frac{1}{2}\left(1+\sqrt{\ln\left(Nh^d\right)}+r(s)\right)}}{\frac{1}{2}\left(1 + \sqrt{\ln(Nh^d)} + r(s)\right)}}_{\geq 2e^{1/2}} \frac{1}{2} e^{\frac{1}{2}\left(1+\sqrt{\ln\left(Nh^d\right)}+r(s)\right)} \\
&\leq cet_0^{-1/2}\exp\left(\sqrt{\ln(Nh^d)}\right),
\end{aligned}$$

which implies

$$e^{\frac{1}{2}\left(1+\sqrt{\ln\left(Nh^d\right)}+r(s)\right)} \le ce^{1/2}t_0(h)^{-1/2}\exp\left(\sqrt{\ln\left(Nh^d\right)}\right).$$

In particular

$$r(s) \le 2\left\{\ln\left(ce^{1/2}\right)+\ln\left(1/\sqrt{t_0(h)}\right)+\sqrt{\ln\left(Nh^d\right)}\right\}$$

and using $t_0(h) = 64h^2\ln\left(2c_-Nh^d\right)$ and $h_- \le h \le h_0$ we find, for some $c = c_{d,c_-,K,C_0}$

$$r(s) \le c\left(\ln\frac{1}{h_-} \vee \sqrt{\ln N}\right)$$

Since $h_- >> \left(\frac{\ln N}{N}\right)^{1/d}$ this implies $\ln\frac{1}{h_-} \le \sqrt{\ln N}$, and therefore

$$M(s) \le c\ln N\sqrt{s}$$

■

B.2.3 On diagonal lower bound

In this section we set again $t_0 = t_0(h) = 64h^2\ln\left(2c_-Nh^d\right)$.

Definition 28. *For every $x \in V$ and $r > 0$ we define the exit time $\tau(x,r)$ of the ball $B_\rho(x,r)$ by*

$$\tau(x,r) = \inf\{t \ge 0 : \rho(x,W_t) > r\} = \inf\{t \ge 0 : W_t \notin B_\rho(x,r)\}$$

Lemma 29. *There exists $c = c_{\mathcal{M},p_0,d}$ such that on A_N and with $2t_0 \le t \le t_1(h)$, $\quad r = c\ln N\sqrt{t}$ we have*

$$\mathbb{P}_x^{(h)}\left(\tau(x,r) \le t\right) \le \frac{1}{2}$$

Proof We have, with $\tau = \tau(x,r)\,, \quad t_0 \le t \le t_1(h), \quad r > 0$

$$\begin{aligned}
r\mathbb{P}_x^{(h)}(\tau \le t) =& r\mathbb{E}_x^{(h)}\left[\mathbb{1}_{\tau\le t}\right]\\
\le& \mathbb{E}_x^{(h)}\left[r\mathbb{1}_{\rho(x,W_{t\wedge\tau})>r}\right]\\
\le& \mathbb{E}_x^{(h)}\left[\rho\left(x,W_{t\wedge\tau}\right)\right]\\
\le& \mathbb{E}_x^{(h)}\left[\rho\left(x,W_{2t}\right)\right]+\mathbb{E}_x^{(h)}\left[\rho\left(W_{2t},W_{t\wedge\tau}\right)\right]\\
=& \mathbb{E}_x^{(h)}\left[\rho\left(x,W_{2t}\right)\right]+\mathbb{E}_x^{(h)}\left[\mathbb{E}_x^{(h)}\left[\rho\left(W_{2t},W_{t\wedge\tau}\right)|W_{t\wedge\tau}\right]\right]\\
\le& \mathbb{E}_x^{(h)}\left[\rho\left(x,W_{2t}\right)\right]+\sup_{z\in V,s\le t}\mathbb{E}_z^{(h)}\left[\rho\left(z,W_{2t-s}\right)\right]
\end{aligned}$$

Now using lemma 27 we find, as long as $t_0(h) \le t\frac{1}{2}t_1(h)$

$$\begin{aligned}
r\mathbb{P}_x^{(h)}(\tau \le t) \le& m(2t,x)+\sup_{z\in V,s\le t} m(2t-s,z)\\
\le& c\ln N\left(\sqrt{t}+\sqrt{2t-s}\right) \le c\ln N\sqrt{t}
\end{aligned}$$

the value of c changing from line to line. Hence by setting $r=r(t)=2c\ln N\sqrt{t}$ we have

$$\mathbb{P}_x^{(h)}\left(\tau(x,r)\leq t\right)\leq\frac{1}{2}.$$

■

Lemma 30. *There exists $c=c_{\mathcal{M},p_0,d}$ such that on A_N, for any $t_0(h)\leq t\leq t_1(h)\wedge\left(\frac{r_0}{c\ln N}\right)^2$ we have*

$$\forall h_-\leq h\leq h_0,\forall x\in V,p_t^{(h)}(x,x)\geq c\frac{t^{-d/2}}{\ln^d N}$$

Proof We have, with $\tau=\tau\left(x,r\right),\quad t_0/2\leq t\leq t_1(h),\quad r=r(t)$ from Lemma 29 and $B=B_\rho\left(x,r\right)$

$$\mathbb{P}_x^{(h)}\left(W_t\notin B\right)\leq\mathbb{P}_x^{(h)}\left(\tau(x,r)\leq t\right)\leq\frac{1}{2}$$

Hence

$$\begin{aligned}\frac{1}{2}\leq&\mathbb{P}_x^{(h)}\left(W_t\in B\right)=\sum_{y\in B}p_t^{(h)}\left(x,y\right)\nu_y^{(h)}\\&\leq\left(\nu^{(h)}\left(B\right)\right)^{1/2}\left(\sum_{y\in B}p_t^{(h)}(x,y)^2\nu_y^{(h)}\right)^{1/2}\leq\left(\nu^{(h)}\left(B\right)\right)^{1/2}\left(\sum_{y\in V}p_t^{(h)}(x,y)^2\nu_y^{(h)}\right)^{1/2}\\&=\left(\nu^{(h)}\left(B\right)\right)^{1/2}p_{2t}^{(h)}(x,x)^{1/2},\end{aligned}$$

and therefore we have the inequality

$$p_{2t}^{(h)}(x,x)\geq\frac{1}{4\nu^{(h)}\left(B\right)}.$$

We now want to conclude using Theorem 19. For this we need to guarantee that

$$h_-\leq r(t)\leq r_0\iff\left(\frac{h_-}{c\ln N}\right)^2\leq t\leq\left(\frac{r_0}{c\ln N}\right)^2.$$

Since $2t\geq t_0=64h^2\ln\left(2c_-Nh^d\right)\geq 64h_-^2\ln\left(2c_-Nh_-^d\right)$ and h_- satisfies 2, we have $t\geq\left(\frac{h_-}{c\ln N}\right)^2$. Since by assumption $t\leq\left(\frac{r_0}{c\ln N}\right)^2$, we conclude that $h_-\leq r(t)\leq r_0$. Hence theorem 19 implies

$$\nu^{(h)}\left(B\right)\leq\frac{c_+^2}{c_-}r(t)^d$$

and therefore

$$p_{2t}^{(h)}(x,x)\geq\frac{c_-}{4c_+^2}r(t)^{-d}=\frac{c_-c^{-d}}{4c_+^2}\frac{t^{-d/2}}{\ln^d N}.$$

■

Lemma 31. *There exists $c = c_{\mathcal{M},p_0,d}$ such that on A_N we have, for any $h_- \leq h \leq h_0$*

$$t_1(h) \geq \frac{c}{\ln^2 N}$$

Proof If $t_1(h) < \left(\frac{r_0}{c \ln N}\right)^2$ then for any $t_0/2 \leq t \leq t_1(h)$ using lemma 30 we have

$$p_{2t}^{(h)}(x,x) \geq \frac{c_- c^{-d}}{4c_+^2} \frac{t^{-d/2}}{\ln^d N}.$$

In particular

$$K' = p_{2t_1(h)}^{(h)}(x,x) \geq \frac{c_- c^{-d}}{4c_+^2} \frac{t_1(h)^{-d/2}}{\ln^d N},$$

i.e.

$$t_1(h) \geq \left(\frac{c_- c^{-d}}{4c_+^2} \frac{K'^{-1}}{\ln^d N}\right)^{2/d}.$$

Hence

$$t_1(h) \geq \left(\frac{r_0}{c \ln N}\right)^2 \wedge \left(\frac{c_- c^{-d}}{4c_+^2} \frac{K'^{-1}}{\ln^d N}\right)^{2/d},$$

which concludes the proof. ■

B.3 Control of the Laplacian spectrum

Recall that here $t_0 = t_0(h) = 64h^2 \ln\left(2c_- N h^d\right)$ and $h_0 = h_{0|\mathcal{M},p_0,d}$ is a theoretical constant.

Lemma 32. *There exists constants $b_1, b_2, b_3 > 0$ such that, on A_N, for any $h_- \leq h \leq h_0$*

$$\forall b_1 \ln^d N \leq j \leq b_2 t_0^{-d/2}, \quad \lambda_j^{(h)} \geq b_3 j^{2/d}$$

Proof Using the inequality $\mathbb{1}_{\lambda \leq \Lambda} \leq e^{1-\lambda/\Lambda}$ we find

$$\begin{aligned}
\#\left\{1 \leq j \leq N : \lambda_j^{(h)} \leq \Lambda\right\} &= \sum_{j=1}^N \mathbb{1}_{\lambda_j^{(h)} \leq \Lambda} \\
&= \sum_{j=1}^N \mathbb{1}_{\lambda_j^{(h)} \leq \Lambda} \|u_j\|_N^2 \\
&= \sum_{j=1}^N \mathbb{1}_{\lambda_j^{(h)} \leq \Lambda} \sum_{x \in V} u_j(x)^2 \nu_x \\
&= \sum_{x \in V} \left\{\sum_{j=1}^N \mathbb{1}_{\lambda_j^{(h)} \leq \Lambda} u_j(x)^2\right\} \nu_x \leq \sum_{x \in V} \left\{\sum_{j=1}^N e^{1-\lambda_j^{(h)}/\Lambda} u_j(x)^2\right\} \nu_x \\
&= \sum_{x \in V} e p_{\Lambda^{-1}}^{(h)}(x,x) \nu_x \\
&\leq e \max_{x \in V} p_{\Lambda^{-1}}^{(h)}(x,x)
\end{aligned}$$

We now use theorem 20 with $t = \Lambda^{-1}$: there exist constants $a_2, a_3, h_0 > 0$ (that also depend on $\mathcal{M}, p_0$) such that, on A_N, for all $h_- \leq h \leq h_0$ and $a_1^{-1} \ln^2 N \leq \Lambda \leq t_0^{-1}$

$$\forall x \in V, \quad p_t^{(h)}(x,x) \leq a_3 \Lambda^{d/2}$$

which implies

$$\#\left\{1 \leq j \leq N : \lambda_j \leq \Lambda\right\} \leq e a_3 \Lambda^{d/2}$$

In particular $j = 2ea_3\Lambda^{d/2} \iff \Lambda = (j/2ea_3)^{2/d}$ yields

$$\lambda_j^{(h)} > \Lambda = (j/2ea_3)^{2/d}$$

This is valid if

$$a_1^{-1} \ln^2 N \leq (j/2ea_3)^{2/d} \leq t_0^{-1}$$

which gives the result by inverting the inequality. ■

Lemma 33. *There exist constants $b_4, b_5, b_6 > 0$ such that on A_N, for any $h_- \leq h \leq h_0$*

$$\forall b_4 \leq j \leq b_5 \frac{t_0^{-d/2}}{\ln^{3d/2} N}, \quad \lambda_j^{(h)} \leq b_6 j^{2/d} \ln^3 N$$

Proof As in in the proof of Lemma 3.19 in Coulhon, Kerkyacharian, and Petrushev (2012) we start by the following inequality

$$e^{-t\lambda} = \mathbb{1}_{[0,\Lambda]}(\lambda) e^{-t\lambda} + \sum_{l \geq 0} \mathbb{1}_{[2^l\Lambda, 2^{l+1}\Lambda]}(\lambda) e^{-t\lambda} \leq \mathbb{1}_{[0,\Lambda]}(\lambda) + \sum_{l \geq 0} e^{-t2^l\Lambda} \mathbb{1}_{[2^l\Lambda, 2^{l+1}\Lambda]}(\lambda)$$

which holds for any $t, \Lambda > 0$. Hence

$$\begin{aligned} p_t^{(h)}(x,x) &= \sum_{j=1}^{N} e^{-t\lambda_j} u_j(x)^2 \\ &\leq \sum_{j=1}^{N} \mathbb{1}_{[0,\Lambda]}\left(\lambda_j^{(h)}\right) u_j(x)^2 + \sum_{l \geq 0} e^{-t2^l\Lambda} \sum_{j=1}^{N} \mathbb{1}_{[2^l\Lambda, 2^{l+1}\Lambda]}\left(\lambda_j^{(h)}\right) u_j(x)^2 \end{aligned}$$

Since for any $h > 0, \quad j = 1, \ldots, N$ we have $\lambda_j^{(h)} \leq 2h^{-2}$ (because $\|\mathcal{L}\|_{L^\infty(\nu)} \leq 2h^{-2}$), we can stop the sum at $L \in \mathbb{N}$ such that $2^L\Lambda \leq 2h^{-2} \leq 2^{L+1}\Lambda$:

$$p_t^{(h)}(x,x) \leq \sum_{j=1}^{N} \mathbb{1}_{[0,\Lambda]}\left(\lambda_j^{(h)}\right) u_j(x)^2 + \sum_{l=0}^{L} e^{-t2^l\Lambda} \sum_{j=1}^{N} \mathbb{1}_{[2^l\Lambda, 2^{l+1}\Lambda]}\left(\lambda_j^{(h)}\right) u_j(x)^2.$$

Using theorem 20 we find, on A_N and for $t_0 \leq t \leq \frac{a_1}{\ln^2 N}$

$$p_t^{(h)}(x,x) \geq a_2 \frac{t^{-d/2}}{\ln^d N} \implies a_2 \frac{t^{-d/2}}{\ln^d N} \leq \sum_{j=1}^{N} \mathbb{1}_{[0,\Lambda]}\left(\lambda_j^{(h)}\right) u_j(x)^2 + \sum_{l=0}^{L} e^{-t2^l\Lambda} \sum_{j=1}^{N} \mathbb{1}_{[2^l\Lambda, 2^{l+1}\Lambda]}\left(\lambda_j^{(h)}\right) u_j(x)^2$$

We now define $K \in \mathbb{N}$ such that $2^{K+1}\Lambda \leq t_0^{-1} \leq 2^{K+2}\Lambda$. If $a_1^{-1}\ln^2 N \leq \Lambda \leq t_0^{-1}$ then for $1 \leq l \leq K$ we have $a_1^{-1}\ln^2 N \leq \Lambda \leq 2^{l+1}\Lambda \leq t_0^{-1}$, and therefore

$$\begin{aligned}\sum_{l=0}^{K} e^{-t2^l\Lambda}\sum_{j=1}^{N}\mathbb{1}_{[2^l\Lambda,2^{l+1}\Lambda]}\left(\lambda_j^{(h)}\right)u_j(x)^2 \leq& \sum_{l=0}^{K} e^{-t2^l\Lambda}\sum_{j=1}^{N}\mathbb{1}_{[0\Lambda,2^{l+1}\Lambda]}\left(\lambda_j^{(h)}\right)u_j(x)^2\\ \leq& \sum_{l=0}^{K} e^{-t2^l\Lambda}\sum_{j=1}^{N} e^{1-2^{-(l+1)}\Lambda^{-1}\lambda_j^{(h)}}u_j(x)^2\\ =& ep^{(h)}_{2^{-(l+1)}\Lambda^{-1}}(x,x) \leq ea_3 2^{\frac{d(l+1)}{2}}\Lambda^{d/2},\end{aligned}$$

which implies

$$\begin{aligned}a_2\frac{t^{-d/2}}{\ln^d N} \leq& \sum_{j=1}^{N}\mathbb{1}_{[0,\Lambda]}\left(\lambda_j^{(h)}\right)u_j(x)^2 + ea_3 2^{d/2}\sum_{l\geq 0} e^{-t2^l\Lambda}\left(2^l\Lambda\right)^{d/2} + \sum_{j=1}^{N}\sum_{l=K+1}^{L} e^{-t2^l\Lambda}u_j(x)^2\\ \leq& \sum_{j=1}^{N}\mathbb{1}_{[0,\Lambda]}\left(\lambda_j^{(h)}\right)u_j(x)^2 + ea_3 2^{d/2}\sum_{j=1}^{N}\left(2^l\Lambda\right)^{d/2} + Le^{-t2^{K+1}\Lambda}\sum_{j=1}^{N}u_j(x)^2.\end{aligned}$$

Taking the expectation of this quantity with respect to $x \sim \nu$ and using $\sum_{x\in V} u_j(x)^2\nu_x = 1$ then yields

$$a_2\frac{t^{-d/2}}{\ln^d N} \leq \#\left\{j : \lambda_j^{(h)} \leq \Lambda\right\} + ea_3 2^{d/2}t^{-d/2}\sum_{l\geq 0} e^{-t2^l\Lambda}\left(2^l t\Lambda\right)^{d/2} + LNe^{-t2^{K+1}\Lambda}$$

Using an integral comparison (see Barlow, 2017, Appendix. A) we find, for any $x \geq 1$

$$\sum_{l\geq 0} e^{-2^l x}\left(2^l x\right)^{d/2} \leq c_d x^{d/2-1}e^{-x/2} \leq c_d c'_d e^{-x/4}, \quad c'_d = \sup_{x\geq 1} x^{d/2-1}e^{-x/4} < +\infty.$$

In particular, if $t \geq \Lambda^{-1}$, with $c = c_d c'_d$

$$\sum_{l\geq 0} e^{-t2^l\Lambda}\left(2^l t\Lambda\right)^{d/2} \leq ce^{-t\Lambda/4}.$$

Therefore we obtain : for any $a_1^{-1}\ln^2 N \leq \Lambda^{-1} \leq t \leq t_0^{-1}$

$$\begin{aligned}\#\left\{j : \lambda_j^{(h)} \leq \Lambda\right\} \geq& t^{-d/2}\left(a_2\frac{1}{\ln^d N} - cea_3 2^{d/2}e^{-t\Lambda/4} - LNt^{d/2}e^{-t2^{K+1}\Lambda}\right)\\ \geq& t^{-d/2}\left(a_2\frac{1}{\ln^d N} - cea_3 2^{d/2}e^{-t\Lambda/4} - LNt^{d/2}e^{-t\Lambda}\right)\end{aligned}$$

Let $t = \kappa\Lambda^{-1}\ln N, \kappa > 0$. Then

$$\#\left\{j : \lambda_j^{(h)} \leq \Lambda\right\} \geq \frac{\Lambda^{d/2}}{\ln^{d/2} N}\left(a_2\frac{1}{\ln^d N} - cea_3 2^{d/2}N^{-\kappa/4} - LN\Lambda^{d/2}\kappa^{-d/2}\left(\ln N\right)^{-d/2}N^{-\kappa}\right)$$

But notice that $L \leq \frac{\ln\left(2h^{-1}\Lambda^{-1}\right)}{\ln 2} \leq \frac{\ln\left(2h_{-}^{-1}a_1(\ln N)^{-2}\right)}{\ln 2} \lesssim \ln N$, which implies that for $\kappa > 0$ sufficiently large we find

$$\#\left\{j : \lambda_j^{(h)} \leq \Lambda\right\} \geq \frac{a_2\Lambda^{d/2}}{2\ln^{3d/2} N}$$

To conclude we only need to guarantee the conditions on t, Λ, κ :

$$t_0 \leq \Lambda^{-1} \leq t = \kappa\Lambda^{-1}\ln N \leq \frac{a_1}{\ln^2 N} \iff \kappa a_1^{-1}\ln^3 N \leq \Lambda \leq t_0^{-1}$$

i.e we have obtained

$$\forall \kappa a_1^{-1}\ln^3 N \leq \Lambda \leq t_0^{-1}, \quad \#\left\{j : \lambda_j^{(h)} \leq \Lambda\right\} \geq \frac{a_2\Lambda^{d/2}}{2\ln^{3d/2} N}$$

In particular with $j = \frac{a_2\Lambda^{d/2}}{2\ln^{3d/2} N}$ so that $\Lambda = (2j/a_2)^{2/d}\ln^3 N$ we get

$$\forall \frac{a_2}{2}\left(\kappa a_1^{-1}\right)^{d/2} \leq j \leq \frac{a_2}{2}\frac{t_0^{-d/2}}{\ln^{3d/2} N}, \quad \lambda_j^{(h)} \leq (2j/a_2)^{2/d}\ln^3 N$$

This concludes the proof. ■

B.4 $L^\infty(\nu) - L^2(\nu)$ comparison theorem

Again recall that here $t_0 = t_0(h) = 32h^2\ln\left(2c_- N h^d\right)$.

Lemma 34. *On A_N, for any $h_- \leq h \leq h_0$ with*

$$\Sigma_\Lambda = span\left\{u_j^{(h)} : \lambda_j^{(h)} \leq \Lambda\right\}$$

we have

$$\forall a_1^{-1}\ln^2 N \leq \Lambda \leq t_0^{-1}, \quad f \in \Sigma_\Lambda, \quad \|f\|_{L^\infty(\nu)}^2 \leq e a_3\Lambda^{d/2}\|f\|_{L^2(\nu)}^2$$

Moreover, with

$$\Sigma^J = span\left\{u_j^{(h)} : j \leq J\right\}$$

we have

$$\forall b_4 \leq J \leq b_5\frac{t_0^{-d/2}}{\ln^{3d/2} N}, \quad f \in \Sigma^J, \quad \|f\|_{L^\infty(\nu)}^2 \leq e a_3 b_6^{d/2} J \ln^{3d/2} N\|f\|_{L^2(\nu)}^2$$

Proof Since $f \in \Sigma_\Lambda$ we have $\mathbb{1}_{[0,\Lambda]}(\mathcal{L})(f) = f$, therefore, with $\mathbb{1}_{[0,\Lambda]}(\mathcal{L})(x,y)$ the kernel of $\mathbb{1}_{[0,\Lambda]}(\mathcal{L})$ with respect to $\nu^{(h)}$ is given by

$$\mathbb{1}_{[0,\Lambda]}(\mathcal{L})(x,y) = \sum_{j=1}^{N}\mathbb{1}_{[0,\Lambda]}\left(\lambda_j^{(h)}\right)u_j^{(h)}(x)u_j^{(h)}(y).$$

Then for any $x \in V$,

$$
\begin{aligned}
f(x)^2 &= \left[\mathbb{1}_{[0,\Lambda]}\left(\mathcal{L}\right)(f)\right](x)^2 = \left[\sum_y \mathbb{1}_{[0,\Lambda]}\left(\mathcal{L}\right)(x,y) f(y)\nu_y^{(h)}\right]^2 \\
&\leq \left[\sum_y \mathbb{1}_{[0,\Lambda]}\left(\mathcal{L}\right)(x,y)^2 \nu_y\right]\left[\sum_y f(y)^2 \nu_y^{(h)}\right] \leq \|f\|_N^2 \max_y \mathbb{1}_{[0,\Lambda]}\left(\mathcal{L}\right)(x,y)^2 \\
&= \|f\|_N^2 \max_y \left[\sum_{j=1}^N \mathbb{1}_{[0,\Lambda]}(\lambda_j) u_j^{(h)}(x) u_j^{(h)}(y)\right]^2 \\
&\leq \|f\|_N^2 \max_y \left[\sum_{j=1}^N \mathbb{1}_{[0,\Lambda]}(\lambda_j) u_j^{(h)}(x)^2\right]\left[\sum_{j=1}^N \mathbb{1}_{[0,\Lambda]}(\lambda_j) u_j^{(h)}(y)^2\right] \\
&= \|f\|_N^2 \max_y \mathbb{1}_{[0,\Lambda]}\left(\mathcal{L}\right)(x,x) \mathbb{1}_{[0,\Lambda]}\left(\mathcal{L}\right)(y,y) \leq \|f\|_N^2 \max_x \mathbb{1}_{[0,\Lambda]}\left(\mathcal{L}\right)(x,x)^2 \\
&= \|f\|_N^2 \max_x \sum_{j=1}^N \mathbb{1}_{[0,\Lambda]}\left(\lambda_j\right) u_j^{(h)}(x)^2 \\
&\leq \|f\|_N^2 \max_x \sum_{j=1}^N e^{1-\lambda_j/\Lambda} u_j^{(h)}(x)^2 = \|f\|_N^2 \max_x e p_{\Lambda^{-1}}^{(h)}(x,x)
\end{aligned}
$$

Using theorem 20 we thus find

$$
\forall a_1^{-1} \ln^2 N \leq \Lambda \leq t_0^{-1}, \quad \forall f \in \Sigma_\Lambda, \quad \|f\|_{L^\infty(\nu)}^2 \leq e a_3 \Lambda^{d/2} \|f\|_{L^2(\nu)}^2 .
$$

Now if $b_4 \leq J \leq b_5 \frac{t_0^{-d/2}}{\ln^{3d/2} N}$, lemma 33 implies $f \in \Sigma_\Lambda, \quad \Lambda = b_6 J^{2/d} \ln^3 N$, which concludes the proof. ■

Appendix C. Proofs of Lemmas 14 and 15

C.1 Proof of Lemma 14

We first bound $T_h f$ for $\beta \leq 1$, then we treat the case $1 < \beta \leq 2$ followed by the case $\beta > 2$. Recall that $h \leq h_0 \wedge \pi r_0 / C_0$.

Case $\beta \leq 1$**.** In this case the result is trivial since $\forall x \in \mathcal{M}$,

$$
\begin{aligned}
|T_h f(x)| = &\left|\frac{1}{h^2 P_0(B_{\mathbb{R}^D}(x,h))} \int_{B_{\mathbb{R}^D}(x,h)\cap\mathcal{M}} \left(f(x) - f(y)\right) p_0(y)\mu(dy)\right| \\
\leq & h^{-2} \sup_{y \in B_{\mathbb{R}^D}(x,h)\cap\mathcal{M}} |f(x) - f(y)| \leq h^{-2} \sup_{y \in B_{\mathbb{R}^D}(x,h)} \sum_{i\in I} |\chi_i f(x) - \chi_i f(y)| \\
\leq & h^{-2} \max_{i \in I} \sup_{y \in B_{\mathbb{R}^D}(x,h)} \left|(\chi_i f \circ \phi_i^{-1})(\phi_i(y)) - (\chi_i f \circ \phi_i^{-1})(\phi_i(y))\right| \\
\leq & h^{-2} \|f\|_{\mathcal{C}^\beta(\mathcal{M})} \max_{i\in I} \sup_{y \in B_{\mathbb{R}^D}(x,h)} \|\phi_i(y) - \phi_i(x)\|^\beta
\end{aligned}
$$

But since $h \leq h_0$, for any $y \in B_{\mathbb{R}^D}(x,h) \cap \mathcal{M}$ we have $\rho(x,y) \leq C_0 \|x-y\| \leq C_0 h$. Hence since the functions ϕ_i are Lipschitz

$$|T_h f(x)| \leq h^{-2} \|f\|_{\mathcal{C}^\beta(\mathcal{M})} \max_{i\in I} \sup_{y \in B_\rho(x,h)} \|\phi_i(y) - \phi_i(x)\|^\beta \lesssim \|f\|_{\mathcal{C}^\beta(\mathcal{M})} h^{\beta-2}$$

Case $1 < \beta \leq 2$**.** Recall that C_0 is such that $\rho(x,y) \leq C_0 \|x-y\|$ for any $x, y \in \mathcal{M}$. Recall that $f_0 \in \mathcal{C}^\beta(\mathcal{M})$, so that using a Taylor expansion of $f_x = f \circ \phi_x$, with $\phi_x = \exp_x(\psi(x,\cdot))$ at some $x \in V$ and for any $\|x-y\| < h$ we have $\rho(x,y) \leq C_0\|x-y\| \leq C_0 h$ and therefore

$$f_0(y) = f(x) + df_x(0).\omega + R(x,y)$$

where $\exp_x(\psi(x,\omega)) = y$ and

$$\max_{x \neq y \in \mathcal{M}} \frac{|R(x,y)|}{\rho(x,y)^\beta} \lesssim \|f\|_{\mathcal{C}^\beta(\mathcal{M})}.$$

This, together with the fact that when $x \sim y$ then $\|x-y\| < h$, implies that

$$\left| T_h f(x) - \frac{1}{h^2 P_0(B_{\mathbb{R}^D}(x,h))} \int_{\exp_x \circ \phi_x^{-1} B_{\mathbb{R}^D}(x,h)} df_x(0).\omega p_x(\omega) J_x(\omega) d\omega \right| \lesssim \frac{\|f\|_{\mathcal{C}^\beta} h^\beta}{h^2}.$$

where $p_x = p_0 \circ \phi_x, \quad J_x = |g(\exp_x(\psi(x,\omega)))|^{1/2}$. It remains to bound

$$\frac{1}{h^2 P_0(B_{\mathbb{R}^D}(x,h))} \int_{\exp_x \circ \phi_x^{-1} B_{\mathbb{R}^D}(x,h) \cap \mathcal{M}} df_x(0).\omega p_x(\omega) J_x(\omega) d\omega$$

For this notice that since $p_0 \in \mathcal{C}^{\beta-1}(\mathcal{M})$ and $J_x(\omega) = 1 + \mathcal{O}(\|\omega\|^2)$ we have

$$p_x(\omega) J(\omega) = \left(p_0(x) + \mathcal{O}\left(h^{\beta-1}\right)\right)\left(1 + \mathcal{O}\left(h^2\right)\right) = p_0(x) + \mathcal{O}\left(h^{\beta-1}\right)$$

Therefore

$$\begin{aligned}\frac{\int_{\exp_x \circ \phi_x^{-1} B_{\mathbb{R}^D}(x,h)} df_x(0).\omega p_x(\omega) J_x(\omega) d\omega}{P_0(B_{\mathbb{R}^D}(x,h))} &= \frac{\int_{\exp_x \circ \phi_x^{-1}\left(B_{\mathbb{R}^D}(x,h)\right)} df_x(0).\omega \left(p_0(x) + \mathcal{O}\left(h^{\beta-1}\right)\right) d\omega}{P_0(B_{\mathbb{R}^D}(x,h))} \\ &= p_0(x) \frac{\int_{\exp_x \circ \phi_x^{-1}\left(B_{\mathbb{R}^D}(x,h)\right)} df_x(0).\omega d\omega}{P_0(B_{\mathbb{R}^D}(x,h))} + \mathcal{O}\left(\|f\|_{\mathcal{C}^\beta(\mathcal{M})} h^\beta\right)\end{aligned}$$

Note that since $\exp_x \circ \phi_x^{-1}(B_\rho(x,h)) = \{\omega \in T_x\mathcal{M} : \|\omega\| < h\}$ is a symmetric set, we get

$$\int_{\exp_x \circ \phi_x^{-1}(B_\rho(x,h))} df_x(0).\omega d\omega = 0,$$

therefore to control the first term of the right hand side of the above term, it is enough to show that $B_{\mathbb{R}^D}(x,h)$ is close to $B_\rho(x,h)$.

Since $h \leq \frac{\pi a_{\mathcal{M}}}{C_0}$, for all $x, y \in \mathcal{M}, \|x-y\| < h$ we have $\rho(x,y) \leq C_0\|x-y\| \leq C_0 h \leq \pi a_{\mathcal{M}}$. Hence, using (12), $\rho(x,y) - \frac{\rho(x,y)^3}{24 a_{\mathcal{M}}^2} \leq \|x-y\| \leq \rho(x,y)$. Therefore if

$y \in \mathcal{M}, \|x-y\| < h$ but $\rho(x,y) \geq h$ then $\rho(x,y) \leq \|x-y\| + \frac{\rho(x,y)^3}{24a_{\mathcal{M}}^2} \leq h + \frac{C_0^3}{24a_{\mathcal{M}}^2}h^3$. This implies

$$\begin{aligned}
vol\left(\{\omega : \exp_x \circ \phi_x(\omega) \in B_\rho(x,h)^c \cap B_{\mathbb{R}^D}(x,h)\}\right) &\leq vol\left(\left\{\omega : h \leq \|\omega\| < h + \frac{C_0^3}{24a_{\mathcal{M}}^2}h^3\right\}\right) \\
&= vol\left(\{\omega : \|\omega\| \leq 1\}\right)\left(\left(h + \frac{C_0^3}{24a_{\mathcal{M}}^2}h^3\right)^d - h^d\right) \lesssim h^d\left(\left(1 + \frac{C_0^3}{24a_{\mathcal{M}}^2}h^2\right)^d - 1\right) \\
&\leq C_{\mathcal{M}} h^{d+2}
\end{aligned}$$

where the constant $C_{\mathcal{M}}$ depends on $\mathcal{M}$ through $a_{\mathcal{M}}, h_0, C_0$ only. Therefore,

$$\begin{aligned}
&\int_{\exp_x \circ \phi_x^{-1}\left(B_{\mathbb{R}^D}(x,h)\right)} df_x(0).\omega d\omega \\
&= \int_{\exp_x \circ \phi_x^{-1}(B_\rho(x,h))} df_x(0).\omega d\omega + \int_{\exp_x^{-1}\left(B_{\mathbb{R}^D}(x,h) \cap B_\rho(x,h)^c\right)} df_x(0).\omega d\omega \\
&= \mathcal{O}\left(\|f\|_{\mathcal{C}^\beta(\mathcal{M})} h^{d+3}\right) \leq \mathcal{O}\left(\|f\|_{\mathcal{C}^\beta(\mathcal{M})} h^{d+\beta}\right)
\end{aligned}$$

where we have used that $\beta \leq 2 \leq 3$. Finally since $P_0(B(x,h)) \gtrsim h^d$ we obtain

$$\frac{1}{h^2 P_0(B(x,h))} \int_{\exp_x \circ \phi_x^{-1} B_{\mathbb{R}^D}(x,h)} df_x(0).\omega p_x(\omega) J_x(\omega) d\omega = \mathcal{O}\left(\|f\|_{\mathcal{C}^\beta(\mathcal{M})} h^{\beta-2}\right),$$

which in turns implies that

$$\|T_h f_0\|_{L^\infty(\mathcal{M})} = \mathcal{O}\left(\|f\|_{\mathcal{C}^\beta(\mathcal{M})} h^{\beta-2}\right).$$

Case $\beta > 2$**.** Let $k = \lceil \beta/2 \rceil - 1 \geq 1$. Since $f \in \mathcal{C}^\beta(\mathcal{M})$ and $p_0 \in \mathcal{C}^{\beta-1}(\mathcal{M})$, then for any $i \in I, \quad x \in \mathcal{U}_i, \quad y \in B_\rho(x, r_0)$ (with $I, (\mathcal{U}_i)_{i\in I}, \quad (\psi_i)_{i\in I}, \quad (\varphi_i)_{i \in I}$ the objects defined in section A)

$$f_i(y) = f_i(x) + \sum_{l=1}^{k}\left\{\frac{d^{2l-1}f_x(0).\exp_x^{-1}(y)^{2l-1}}{(2l-1)!} + \frac{d^{2l}f_x(0).\exp_x^{-1}(y)^{2l}}{(2l)!}\right\} + R_{k,1}(x,y)$$

$$\begin{aligned}
p_{0,i}(y) =& p_{0,i}(x) + \sum_{l=1}^{k-1}\left\{\frac{d^{2l-1}p_x(0).\exp_x^{-1}(y)^{2l-1}}{(2l-1)!} + \frac{d^{2l}p_x(0).\exp_x^{-1}(y)^{2l}}{(2l)!}\right\} \\
&+ \frac{d^{2k-1}p_x(0).\exp_x^{-1}(y)^{2k-1}}{(2k-1)!} + R_{k,2}(x,y)
\end{aligned}$$

$$\begin{aligned}
\sqrt{|g|}_i(y) =& \sqrt{|g|}_i(x) + \sum_{l=1}^{k-1}\left\{\frac{d^{2l-1}J_x(0).\exp_x^{-1}(y)^{2l-1}}{(2l-1)!} + \frac{d^{2l}J_x(0).\exp_x^{-1}(y)^{2l}}{(2l)!}\right\} \\
&+ \frac{d^{2k-1}p_x(0).\exp_x^{-1}(y)^{2k-1}}{(2k-1)!} + R_{k,3}(x,y)
\end{aligned}$$

where

$$f_i = \chi_i f, p_{0,i} = \chi_i p_0, \quad f_x = (\chi_i f) \circ \exp_x \circ \psi_i(x, \cdot)$$
$$p_x = (\chi_i p_0) \circ \exp_x \circ \psi_i(x, \cdot), \quad J_x = \left(\chi_i \sqrt{|g|}\right) \circ \exp_x \circ \psi_i(x, \cdot)$$

(we drop the index i for convenience), and the remainders satisfy

$$|R_{k,1}(x, y)| \leq \|f\|_{\mathcal{C}^\beta(\mathcal{M})} \rho(x, y)^\beta,$$

$$|R_{k,2}(x, y)| \leq \|p_0\|_{\mathcal{C}^{\beta-1}(\mathcal{M})} \rho(x, y)^{\beta-1},$$

$$|R_{k,3}(x, y)| \leq C_{\mathcal{M}} \rho(x, y)^{\beta-1},$$

where $C_{\mathcal{M}}$ depends on the Hölder constant of the exponential map. Thus, by Riemannian change of variables, there exist coefficients

$$A_{i,m} \in \mathcal{C}^{\beta-m}\left(\mathcal{U}_i, \mathcal{L}\left(\left(\mathbb{R}^d\right)^{\otimes m}, \mathbb{R}\right)\right)$$

and remainders

$$|B_{i,k}(x, \omega)| \leq C(\beta, \mathcal{M}, p_0) \|f\|_{\mathcal{C}^\beta(\mathcal{M})} \|\omega\|^\beta$$

such that

$$\begin{aligned}
&\int_{B_{\mathbb{R}^D}(x,h)} (f_0(x) - f_0(y)) p_0(y) \mu(dy) \\
&= \int_{\psi_i(x,\cdot)^{-1} \circ \exp_x^{-1}\left(B_{\mathbb{R}^D}(x,h)\right)} (f_x(0) - f_x(\omega)) p_x(\omega) J_x(\omega) \, d\omega \\
&= \sum_{i \in I} \int_{\psi_i(x,\cdot)^{-1} \circ \exp_x^{-1}\left(B_{\mathbb{R}^D}(x,h)\right)} \left\{ \sum_{m=1}^{2k} A_{i,m}(x) \omega^m + B_{i,k}(x, \omega) \right\} d\omega \\
&= \sum_{i \in I} \left\{ \sum_{m=1}^{2k} A_{i,m}(x) \int_{\psi_i(x,\cdot)^{-1} \circ \exp_x^{-1}\left(B_{\mathbb{R}^D}(x,h)\right)} \omega^m d\omega + \int_{\psi_i(x,\cdot)^{-1} \circ \exp_x^{-1}\left(B_{\mathbb{R}^D}(x,h)\right)} B_{i,k}(x, \omega) d\omega \right\} \\
&= \sum_{i \in I} \sum_{m=1}^{2k} A_{i,m}(x) \int_{\psi_i(x,\cdot)^{-1} \circ \exp_x^{-1}\left(B_{\mathbb{R}^D}(x,h)\right)} \omega^m d\omega + \mathcal{O}\left(\|f\|_{\mathcal{C}^\beta(\mathcal{M})} h^{d+\beta}\right).
\end{aligned}$$

We thus get

$$T_h f(x) = \sum_{i \in I} \sum_{m=1}^{2k} A_{i,m}(x) \frac{h^{-(d+2)} \int_{\psi_i(x,\cdot)^{-1} \circ \exp_x^{-1}\left(B_{\mathbb{R}^D}(x,h)\right)} \omega^m d\omega}{h^{-d} P_0\left(B_{\mathbb{R}^D}(x, h)\right)} + \mathcal{O}\left(\|f\|_{\mathcal{C}^\beta(\mathcal{M})} h^{\beta-2}\right)$$

By lemma 16 we have, with $s_h(x, v) := \frac{t_h(x,v)-1}{h^2}$

$$\psi_i(x, \cdot)^{-1} \circ \exp_x^{-1}\left(B_{\mathbb{R}^D}(x, h)\right) = \left\{thv : v \in \mathbb{R}^d, \|v\| = 1, 0 \leq t \leq 1 + h^2 s_h(x, v)\right\}$$

And using the spherical coordinates,

$$
\begin{aligned}
&h^{-(d+2)} \int_{\psi_i(x,\cdot)^{-1} \circ \exp_x^{-1}\left(B_{\mathbb{R}^D}(x,h)\right)} \omega^m d\omega \\
=&h^{-(d+2)} \int_0^{\pi} \dots \int_0^{\pi} \int_0^{2\pi} \int_0^{h+h^3 s_h(x,v)} \left(re_\theta\right)^m r^{d-1} \sin\left(\theta_1\right)^{d-2} \sin\left(\theta_2\right)^{d-3} \dots \sin\left(\theta_{d-1}\right) dr d\theta_1 \dots d\theta_{d-1} \\
=&h^{-(d+2)} \int_0^{\pi} \dots \int_0^{\pi} \int_0^{2\pi} \int_0^{1+h^2 s_h(x,v)} \left(the_\theta\right)^m (th)^{d-1} \sin\left(\theta_1\right)^{d-2} \sin\left(\theta_2\right)^{d-3} \dots \sin\left(\theta_{d-1}\right) h dt d\theta_1 \dots d\theta_{d-1} \\
=&h^{m-2} \int_0^{\pi} \dots \int_0^{\pi} \int_0^{2\pi} \int_0^{1+h^2 s_h(x,v)} \left(te_\theta\right)^m t^{d-1} \sin\left(\theta_1\right)^{d-2} \sin\left(\theta_2\right)^{d-3} \dots \sin\left(\theta_{d-1}\right) dt d\theta_1 \dots d\theta_{d-1} \\
=&h^{m-2} \int_{\mathbb{A}^d} \int_0^{1+h^2 s_h(x,v)} \left(te_\theta\right)^m t^{d-1} K(\theta) dt d\theta \\
=&\frac{h^{m-2}}{m+d} \int_{\mathbb{A}^d} e_\theta^m \left(1+h^2 s_h(x,e_\theta)\right)^{m+d} K(\theta) d\theta
\end{aligned}
$$

where $\mathbb{A}^d = (0,\pi)^{d-2} \times (0,2\pi)$ and

$$
e_\theta = \begin{bmatrix} \cos(\theta_1) \\ \sin(\theta_1)\cos(\theta_2) \\ \sin(\theta_1)\sin(\theta_2)\cos(\theta_3) \\ \vdots \\ \sin(\theta_1)\sin(\theta_2)\dots\sin(\theta_{d-2})\cos(\theta_{d-1}) \\ \sin(\theta_1)\sin(\theta_2)\dots\sin(\theta_{d-2})\sin(\theta_{d-1}) \end{bmatrix}, K(\theta) = \sin\left(\theta_1\right)^{d-2} \sin\left(\theta_2\right)^{d-3} \dots \sin\left(\theta_{d-1}\right)
$$

Moreover if m is odd, by symmetry

$$
h^{-(d+2)} \int_{\psi_i(x,\cdot)^{-1} \circ \exp_x^{-1}\left(B_{\mathbb{R}^D}(x,h)\right)} \omega^m d\omega = h^{-(d+2)} \int_{\psi_i(x,\cdot)^{-1} \circ \exp_x^{-1}\left(B_{\mathbb{R}^D}(x,h) \setminus B_\rho(0,h)\right)} \omega^m d\omega
$$

and therefore the same computations show that

$$
\begin{aligned}
h^{-(d+2)} \int_{\psi_i(x,\cdot)^{-1} \circ \exp_x^{-1}\left(B_{\mathbb{R}^D}(x,h)\right)} \omega^m d\omega =&\frac{h^{m-2}}{m+d} \int_{\mathbb{A}^d} e_\theta^m \left(\left(1+h^2 s_h(x,e_\theta)\right)^{m+d} - 1\right) K(\theta) d\theta \\
=&\frac{h^{m-2}}{m+d} \int_{\mathbb{A}^d} e_\theta^m \left(\sum_{r=1}^{m+d} \binom{m+d}{r} h^{2r} s_h(x,e_\theta)^r\right) K(\theta) d\theta \\
=&\frac{h^m}{m+d} \sum_{r=1}^{m+d} \binom{m+d}{r} h^{2r-2} \int_{\mathbb{A}^d} e_\theta^m s_h(x,e_\theta)^r K(\theta) d\theta
\end{aligned}
$$

All in all for each $l \in \{1, \dots, k\}, \quad m = 2l-1$ or $2l$ we have

$$
\begin{aligned}
h^{-(d+2)} \int_{\psi_i(x,\cdot)^{-1} \circ \exp_x^{-1}\left(B_{\mathbb{R}^D}(x,h)\right)} \omega^m d\omega =&\frac{h^{2(l-1)}}{m+d} \sum_{r=1}^{m+d} \binom{m+d}{r} h^{2r-1} \int_{\mathbb{A}^d} e_\theta^m s_h(x,e_\theta)^r K(\theta) d\theta \text{ if } m = 2l-1 \\
h^{-(d+2)} \int_{\psi_i(x,\cdot)^{-1} \circ \exp_x^{-1}\left(B_{\mathbb{R}^D}(x,h)\right)} \omega^m d\omega =&h^{2(l-1)} \int_{\mathbb{A}^d} e_\theta^m \left(1+h^2 s_h(x,e_\theta)\right)^{m+d} K(\theta) d\theta \text{ if } m = 2l
\end{aligned}
$$

This gives

$$T_h f(x) = \frac{1}{h^{-d} P_0\left(B\left(x,h\right)\right)} \sum_{i\in I} \sum_{l=1}^{k} h^{2(l-1)} \left\{A_{i,2l-1}(x) V_{i,h,2l-1}(x) + A_{i,2l}(x) V_{i,h,2l}(x)\right\} + \mathcal{O}\left(\|f\|_{\mathcal{C}^{\beta}(\mathcal{M})} h^{\beta-2}\right)$$

where by lemma 16, composition and integration we have

$$\max_{\alpha\in\mathbb{N}^d, |\alpha|\le \alpha-5} \sup_{\substack{i\in I \\ y\in\varphi_i^{-1}(\mathcal{U}_i) \\ 0<h<h_+ \\ \|v\|_{T_{\varphi_i(y)}\mathcal{M}}=1}} \left\| \frac{\partial^{|\alpha|}}{\partial x^{\alpha}} V_{i,h,m}(\varphi_i(y)) \right\|_{\left(\mathbb{R}^d\right)^{\otimes m}} < +\infty$$

To conclude it suffices to show that $x \mapsto h^{-d} P_0(B_{\mathbb{R}^D}(x,h))$ is lower bounded and satisfies

$$\sup_{0<h<h_+} \left\| h^{-d} P_0(B_{\mathbb{R}^D}(\cdot,h)) \right\|_{\mathcal{C}^{\beta-1}(\mathcal{M})} < +\infty$$

Since p_0 is lower bounded and $\mu\left(B_{\mathbb{R}^D}(x,h)\right) \asymp h^d$ (as can be seen by the Riemannian change of variables formula), we get

$$h^{-d} \int_{B_{\mathbb{R}^D}(x,h)} p_0(y)\mu(dy) \ge \inf_{\mathcal{M}} p_0 \times h^{-d} \mu\left(B_{\mathbb{R}^D}(x,h)\right) \gtrsim \inf_{\mathcal{M}} p_0$$

hence $x \mapsto h^{-d} P_0\left(B_{\mathbb{R}^D}(x,h)\right)$ is indeed lower bounded. For the derivatives, doing the same computations as above

$$\begin{aligned}
h^{-d} P_0\left(B_{\mathbb{R}^D}(x,h)\right) =& \sum_{i\in I} h^{-d} \int_{B_{\mathbb{R}^D}(x,h)} p_0(y)\mu(dy) \\
=& h^{-d} \int_{\varphi_i^{-1}\circ \exp_x^{-1}\left(B_{\mathbb{R}^D}(x,h)\right)} (\chi_i p_0) \circ \exp_x\left(\psi_i(x,\omega)\right) \sqrt{|g(\exp_x(\psi_i(x,\omega)))|} d\omega \\
=& h^{-d} \int_{\varphi_i^{-1}\circ \exp_x^{-1}\left(B_{\mathbb{R}^D}(x,h)\right)} p_x(\omega) J_x(\omega) d\omega \\
=& h^{-d} \int_{\mathbb{A}^d} \int_0^{1+h^2 s_h(x,v)} p_x(the_\theta) J_x(the_\theta) \left(th\right)^{d-1} K(\theta) h dt d\theta \\
=& \int_{\mathbb{A}^d} \int_0^{1+h^2 s_h(x,v)} p_x(the_\theta) J_x(the_\theta) t^{d-1} K(\theta) dt d\theta
\end{aligned}$$

And again, under this form, by composition and integration, $x \mapsto h^{-d} P_0(B_{\mathbb{R}^D}(x,h))$ has indeed the regularity of p_0, i.e is $\mathcal{C}^{\beta-1}\left(\mathcal{M}\right)$ with uniform bound on the derivatives as $h \to 0$.

This concludes the proof by setting

$$g_h^{(l)} = \frac{1}{h^{-d} P_0\left(B_{\mathbb{R}^D}(x,h)\right)} \sum_{i\in I} h A_{i,2l-1}(x) V_{i,h,2l-1}(x) + A_{i,2l}(x) V_{i,h,2l}(x)$$

which is $\mathcal{C}^{\beta-2l}$ since by assumption $\alpha \ge \beta+3$ which implies $\alpha-5 \ge \beta-2 \ge \beta-2l$ for any $l \in \{1,\ldots,k\}$. ■

C.2 Proof of Lemma 15

First if $\beta \leq 1$, $\|T_{h_n} f\|_{L^\infty(\mathcal{M})} \lesssim h_n^{\beta-2}$ follows from lemma 14. For $\|\mathcal{L}f\|_{L^\infty(\nu)}$ notice that for each $x \in V$ we have

$$|(\mathcal{L}f)(x)| = \frac{1}{h_n^2 \mu_x} \left| \sum_{i=1}^{N} (f(x) - f(x_i)) \mathbb{1}_{x_i \sim x} \right| \lesssim \|f\|_{\mathcal{C}^\beta(\mathcal{M})} h_n^{\beta-2}.$$

Now, if $\beta > 1$,

$$\forall x \in V, h_n^2 \mu_x \mathcal{L}f(x) = \sum_{i=1}^{N} (f(x) - f(x_i)) \mathbb{1}_{x_i \sim x},$$

and therefore by Bernstein's inequality (see Van der Vaart and Wellner, 1996, Part. 2), using $|f(x) - f(x_i)| \mathbb{1}_{x \sim x_i} \lesssim \|f\|_{\mathcal{C}^\beta(\mathcal{M})} h_n$,

$$\mathbb{P}_0 \left(\left| h_n^2 \mu_x \mathcal{L}f(x) - N h_n^2 P_0(B_{\mathbb{R}^D}(x, h_n)) T_{h_n} f(x) \right| > u \right) \leq \exp\left(-c \frac{u^2}{N h_n^{d+2} + h_n u} \right)$$

and

$$\begin{aligned}
\mathbb{P}_0 \left(|\mu_x - N P_0(B_{\mathbb{R}^D}(x, h_n))| > u/h_n^2 \right) &\leq \exp\left(-c \frac{u^2/h_n^4}{N h_n^d + u/h_n^2} \right) \\
&\leq \exp\left(-c \frac{u^2}{N h_n^{d+4} + u h_n^2} \right).
\end{aligned}$$

Hence

$$\begin{aligned}
\mathbb{P}_0 \left(\left| h_n^2 \mu_x \mathcal{L}f(x) - h_n^2 \mu_x T_{h_n} f(x) \right| > 2u \right) &\leq \mathbb{P}_0 \left(\left| h_n^2 \mu_x \mathcal{L}f(x) - N h_n^2 P_0(B_{\mathbb{R}^D}(x, h_n)) T_{h_n} f(x) \right| > u \right) \\
+ \quad & \mathbb{P}_0 \left(\left| N h_n^2 P_0(B_{\mathbb{R}^D}(x, h_n)) T_{h_n} f(x) - h_n^2 \mu_x T_{h_n} f(x) \right| > u \right) \\
\leq \exp\left(-c \frac{u^2}{N h_n^{d+2} + h_n u} \right) &+ \mathbb{P}_0 \left(|N P_0(B_{\mathbb{R}^D}(x, h_n)) - \mu_x| > u/h_n^2 \|f\|_{\mathcal{C}^\beta(\mathcal{M})} \right) \\
\leq \exp\left(-c \frac{u^2}{N h_n^{d+2} + h_n u} \right) &+ \exp\left(-c \frac{u^2}{N h_n^{d+4} + u h_n^2} \right)
\end{aligned}$$

Therefore, using the fact that $\mu_x \lesssim N h_n^d$ on $V_N(c)$

$$\begin{aligned}
\mathbb{P}_0 \left(|\mathcal{L}f(x) - T_{h_n} f(x)| > u \right) \leq & \mathbb{P}_0 \left(V_N(c)^c \right) + \mathbb{P}_0 \left(\left| h_n^2 \mu_x \mathcal{L}f(x) - h_n^2 \mu_x T_{h_n} f(x) \right| > c N h_n^{d+2} u \right) \\
\leq & \mathbb{P}_0 \left(V_N(c)^c \right) + \exp\left(-c \frac{\left(N h_n^{d+2} u \right)^2}{N h_n^{d+2} + N h_n^{d+3} u} \right) + \exp\left(-c \frac{\left(N h_n^{d+2} u \right)^2}{N h_n^{d+4} + N h_n^{d+4} u} \right) \\
= & \mathbb{P}_0 \left(V_N(c)^c \right) + \exp\left(-c \frac{N^2 h_n^{2d+4} u^2}{N h_n^{d+2} + N h_n^{d+3} u} \right) + \exp\left(-c \frac{N^2 h_n^{2d+4} u^2}{N h_n^{d+4} + N h_n^{d+4} u} \right) \\
= & \mathbb{P}_0 \left(V_N(c)^c \right) + \exp\left(-c \frac{N h_n^{d+2} u^2}{1 + h_n u} \right) + \exp\left(-c \frac{N h_n^d u^2}{1 + u} \right)
\end{aligned}$$

For any $0 < h_n \le 1, u > 0$ we have $\frac{h_n^2}{1+h_n u} \le \frac{1}{1+u}$, therefore

$$\mathbb{P}_0\left(|\mathcal{L}f(x) - T_{h_n}f(x)| > u\right) \le \mathbb{P}_0\left(V_N(c)^c\right) + 2\exp\left(-c\frac{Nh_n^{d+2}u^2}{1+h_n u}\right)$$

By a union bound this gives

$$\mathbb{P}_0\left(\|\mathcal{L}f - T_{h_n}f\|_{L^\infty(\nu)} > u\right) \le N\mathbb{P}_0\left(V_N(c)^c\right) + 2N\exp\left(-c\frac{Nh_n^{d+2}u^2}{1+h_n u}\right)$$

We have also $\mathbb{P}_0\left(V_N(c)^c\right) \le e^{-cNh_n^d}$, so that for any $H > 0$, taking $u = M_0\left(\frac{\ln N}{N}\right)^{\frac{1}{2}} h_n^{-(1+d/2)}$ with M_0 large enough (that depends on H) we obtain $h_n u = o(1)$ by assumption 2 and

$$\mathbb{P}_0\left(\|\mathcal{L}f - T_{h_n}f\|_{L^\infty(\nu)} > M_0\left(\frac{\ln N}{N}\right)^{\frac{1}{2}} h_n^{-(1+d/2)}\right) \le N^{-H}$$

■

C.3 Proof of Theorem 5

Recall that $\hat{f} = \sum_{j=1}^J \langle u_j|Y\rangle_{L^2(\nu)} u_j$ so that

$$\hat{f} - f_0 = \sum_{j=1}^J \langle u_j|f_0\rangle_{L^2(\nu)} u_j - f_0 + \sum_{j=1}^J \langle u_j|\underline{\varepsilon}\rangle_{L^2(\nu)} u_j, \quad \underline{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$$

where I_n is the identity matrix in $\mathbb{R}^n$. Therefore

$$\|\hat{f} - f_0\|_{L^2(\nu)} \le \|\sum_{j=1}^J \langle u_j|f_0\rangle_{L^2(\nu)} u_j - f_0\|_{L^2(\nu)} + \|\sum_{j=1}^J \langle u_j|\underline{\varepsilon}\rangle_{L^2(\nu)} u_j\|_{L^2(\nu)}.$$

We have by orthonormality of $(u_1, \cdots, u_{J_n})$

$$\left\|\sum_{j=1}^J \langle u_j|\underline{\varepsilon}\rangle_{L^2(\nu)} u_j\right\|_{L^2(\nu)}^2 = \sum_{j=1}^{J_n} \langle u_j|\underline{\varepsilon}\rangle_{L^2(\nu)}^2$$

and also that

$$\mathbb{E}_0\left[\langle u_j|\varepsilon\rangle_{L^2(\nu)}^2 \middle| x_{1:n}\right] = \mathbb{E}_0\left[\left(u_j^T Diag\left(\nu\right)\underline{\varepsilon}\right)^2 \middle| x_{1:n}\right] = \sigma^2 u_j^T Diag\left(\nu^2\right) u_j = \sigma^2 \sum_{i=1}^n u_j(x_i)^2 \nu_{x_i}^2$$

On A_N defined in lemma 19, with $N = n$ and $h_- = h_n$ which satisfies the assumption 2 we have $\nu_{x_i} \lesssim \frac{1}{n}$ so that

$$\mathbb{E}_0\left[\langle u_j|\underline{\varepsilon}\rangle_{L^2(\nu)}^2 |x_{1:n}\right] \lesssim \frac{1}{n}\|u_j\|_{L^2(\nu)}^2 = \frac{1}{n}.$$

This implies that

$$\sum_{i=1}^{J_n} \langle u_j | \varepsilon \rangle_{L^2(\nu)}^2 \lesssim \frac{1}{n} \sum_{j=1}^{J_n} Z_j(x_{1:n})^2, \quad \text{where} \quad Z_j(x_{1:n}) = \frac{\langle u_j | \varepsilon \rangle_{L^2(\nu)}}{\mathbb{E}_0 \left[\langle u_j | \varepsilon \rangle_{L^2(\nu)}^2 \Big| x_{1:n} \right]^{1/2}}$$

In particular conditionally on $x_{1:n}$, $Z_j(x_{1:n}) \overset{i.i.d}{\sim} \mathcal{N}(0,1)$. Hence as in Green et al. (2021, Appendix. C1), by a result of Laurent and Massart (2000) we get for some $c > 0$

$$\begin{aligned}\mathbb{P}_0 \left(\left\| \sum_{j=1}^{J_n} \langle u_j | \varepsilon \rangle_{L^2(\nu)} u_j \right\|_{L^2(\nu)}^2 > c\frac{J_n}{n} \right) &\leq \mathbb{P}_0 \left(A_N^c \right) + \mathbb{E}_0 \left[\mathbb{1}_{A_N} \mathbb{P}_0 \left(\left\| \sum_{j=1}^{J_n} \langle u_j | \varepsilon \rangle_{L^2(\nu)} u_j \right\|_{L^2(\nu)}^2 > c\frac{J_n}{n} \middle| x_{1:n} \right) \right] \\ &\leq \mathbb{P}_0 \left(A_N^c \right) + \exp\left(-J_n \right)\end{aligned} \tag{13}$$

Moreover

$$\|f_{0J_n} - f_0\|_{L^2(\nu)}^2 \leq \left\| p_{J_n} \left(e^{-t\mathcal{L}} f_t \right) - f_0 \right\|_{L^2(\nu)}^2$$

where $p_{J_n}\left(e^{-t\mathcal{L}} f_t\right)$ is defined in Theorem 12. Therefore, using Theorem 12, we deduce that choosing t accordingly,

$$\|f_{0J_n} - f_0\|_{L^2(\nu)}^2 \leq \left\| p_{J_n} \left(e^{-t\mathcal{L}} f_t \right) - f_0 \right\|_{L^2(\nu)}^2 \leq \tilde{\varepsilon}_n(J_n)$$

Finally combining this with (13), we obtain that some $C > 0$

$$\mathbb{P}_0 \left(\left\| \hat{f} - f_0 \right\|_{L^2(\nu)} > C\varepsilon_n \right) \to 0$$

where

$$\varepsilon_n = \sqrt{\frac{J_n}{n}} + (\ln N)^{\lceil \beta/2 \rceil} \max\left(1, \left(\frac{J_n^{-2/d} \ln N}{h_n^2}\right)\right)^{\lceil \beta/2 \rceil} \left(h_n^\beta + \mathbb{1}_{\beta > 1} \left(\frac{\ln N}{N h_n^d} \right)^{1/2} h_n \right)$$

In order to get a result in expectation, simply notice that

$$\mathbb{E}_0 \left[\left\| \hat{f} - f_0 \right\|_n^q \right] \leq c\varepsilon_n^q + \mathbb{E}_0 \left[\left\| \hat{f} - f_0 \right\|_n^q \mathbb{1}_{\left\| \hat{f} - f_0 \right\|_n > c\varepsilon_n} \right]$$

Since by definition

$$\left\| \hat{f} - f_0 \right\|_n = \min_{f \in \operatorname{span}(u_1, \ldots, u_J)} \|f - f_0\|_n \underbrace{\leq}_{f=0} \|f_0\|_n \leq \|f_0\|_{L^\infty(\mathcal{M})}$$

This implies

$$\mathbb{E}_0 \left[\left\| \hat{f} - f_0 \right\|_n^q \right] \leq c\varepsilon_n^q + \|f_0\|_{L^\infty(\mathcal{M})}^q \mathbb{P}_0 \left[\left\| \hat{f} - f_0 \right\|_n > c\varepsilon_n \right]$$

By the first part of the proof, for any $H > 0$ we have

$$\mathbb{P}_0 \left[\left\| \hat{f} - f_0 \right\|_n > c\varepsilon_n \right] \lesssim n^{-H} + e^{-J_n} + \mathbb{P}_0 \left(A_N^c \right)$$

Since h_n satisfies assumption 2 and $J_n \geq \ln^\kappa n, \quad \kappa > d \geq 1$, for $H > 0$ large enough we get

$$\mathbb{P}_0 \left[\left\| \hat{f} - f_0 \right\|_n > c\varepsilon_n \right] \lesssim \varepsilon_n^q$$

which concludes the proof. ■

Appendix D. Proofs of section 3.2

We first recall an adaptation to our setting of the general prior mass and testing approach of Ghosal, Ghosh, and van der Vaart (2000); Ghosal and van der Vaart (2007) that we use to prove posterior contraction rates with respect to the empirical L^2 distance $\|\cdot\|_n$. Note that we apply the approach on a high probability event A_N (which does not change the result), and since in the Gaussian regression setting the Kullback-Leibler divergence and variation are both proportional to the (squared) empirical L^2 norm $\|\cdot\|_n^2$, the prior mass condition on the Kullback-Leibler neighbourhood is merely a condition on the prior mass of the $\|\cdot\|_n-$balls centered at f_0 (see Lemma 2.7 in Ghosal and van der Vaart (2017) as well as Chapter 8 for more details on the prior mass and testing approach).

Theorem 35. *Contraction rate theorem*
Assume that $\varepsilon_n \to 0$, $\quad n\varepsilon_n^2 \to \infty$ *and* A_N *be an event of the form* $A_N = \{x_{1:N} \in \Omega_N\}$ *for some* $\Omega \subset \left(\mathbb{R}^D\right)^N$ *satisfying* $\mathbb{P}_0\left(A_N\right) \to 1$. *If there exists fixed constants* $C_1 > 1 + \frac{1}{2\sigma^2}$, $\quad C_2 > 4\sqrt{1+2\sigma^2}$, $\quad C_3 > 0$ *and* $\mathcal{F}_n \subset \mathbb{R}^V$ *such that, on* A_N

- $\Pi\left[\|f - f_0\|_n \leq \varepsilon_n\right] \geq \exp\left(-n\varepsilon_n^2\right)$
- $\Pi\left[\mathcal{F}_n^c\right] \leq e^{-C_1 n\varepsilon_n^2}$
- $\ln N\left(\frac{C_2\varepsilon_n}{2}, \mathcal{F}_n, \|\cdot\|_n\right) \leq C_3 n\varepsilon_n^2$

where $N\left(\varepsilon, \mathcal{F}, \|\cdot\|_n\right)$ *denotes the covering number of the set* $\mathcal{F}$ *at scale* ε *with respect to the* $\|\cdot\|_n-$*metric, then for* $M > 0$ *large enough*

$$\mathbb{E}_0\left[\Pi\left[\|f - f_0\|_n > M\varepsilon_n | \mathbb{X}^n\right]\right] \to 0$$

D.1 Proof of Theorem 3

As shown by Theorem 35, the proof is based on first bounding from below the prior mass of neighbourhoods of f_0, which is done in Lemma 36 below and then to control the entropy by a sieve sequence, which is done in Lemma 37. Recall that

$$\varepsilon_n(J_n, h_n) = \sqrt{\frac{J_n \ln N}{n}} + (\ln N)^{\lceil \beta/2 \rceil} \max\left(1, \left(\frac{J_n^{-2/d} \ln N}{h_n^2}\right)\right)^{\lceil \beta/2 \rceil} \left(h_n^\beta + \mathbb{1}_{\beta > 1}\left(\frac{\ln N}{N h_n^d}\right)^{1/2} h_n\right)$$

with $J_n \geq \ln^\kappa N$, $\quad \kappa > d$ and h_n satisfying assumption 2.

With $z = z_n = z_1\left(n\varepsilon_n^2\right)^{1/b_2}$ with $z_1 > 0$ and $\mathcal{F}_n = \mathcal{F}_{z_n, J_n}$. Its $L^\infty(\nu) -$ metric entropy is then, according to lemma 37, bounded by

$$\forall \zeta > 0, \quad \ln N\left(\zeta\varepsilon_n, \mathcal{F}_n, \|\cdot\|_{L^\infty(\nu)}\right) \lesssim J_n \ln N \lesssim n\varepsilon_n^2.$$

Moreover by the choice of z_n we have, for some $c > 0$

$$\Pi\left[\mathcal{F}_n^c | J_n, h_n\right] \leq J_n e^{-b_1 z_n^{b_2}} \leq e^{-c z_1^{b_2} n\varepsilon_n^2}$$

By taking $z_1 > 0$ large enough we conclude the proof of Theorem 3 by applying Lemma 36 together with Theorem 35. ∎

D.2 Proof of Theorem 6

The proof is based on that of Theorem 3, but showing that J and h can be chosen in a data dependent way. To do that notice that by the non adaptive case, for any $H > 0$ there exists a constant $C > 0$ such that

$$\mathbb{P}_0\left(\Pi\left[\|f - f_0\|_{L^\infty(\nu)} > C\varepsilon_n(J_n, h_n)|J_n, h_n\right] < \exp\left(-nC^2\varepsilon_n^2\right)\right) \leq N^{-H}$$

where $\varepsilon_n(J_n, h_n)$ is given by (5). If for some $J_0, h_0, h_1 > 0$

$$J_n = J_0\left(\ln N\right)^{\frac{2d(1+2\tau/d)\lceil\beta/2\rceil}{2\beta+d}} n^{\frac{d}{2\beta+d}}, \quad h_n \in \left[\frac{h_0 J_n^{-1/d}}{2\ln^{\tau/d} N}, \frac{h_1 J_n^{-1/d}}{\ln^{\tau/d} N}\right]$$

then

$$\varepsilon_n(J_n, h_n) \lesssim n^{-\frac{\beta}{2\beta+d}}\left(\ln N\right)^{\frac{(2\tau+d)\lceil\beta/2\rceil - \beta(\tau+2\beta/d)}{2\beta+d}}$$

Since

$$\Pi\left[\|f - f_0\|_{L^\infty(\nu)} \leq C\varepsilon_n\right] \geq \pi_J(J_n)\pi_h\left(h_n|J_n\right)\Pi\left[\|f - f_0\|_{L^\infty(\nu)} > C\varepsilon_n|J_n, h_n\right]$$

By assumption we have for some a_1, b_1, b_2

$$\pi_J(J_n) \geq e^{-a_1 J_n L_{J_n}} \geq e^{-a_1 J_n \ln N} \text{ and } \pi_h\left(h_n|J_n\right) \geq b_1 e^{-b_2 h_n^{-d}}$$

Therefore, since $\varepsilon_n \geq \left(nh_n^d\right)^{-1/2} + \sqrt{\frac{J_n \ln N}{n}}$, for some $C' > 0$ we have

$$\mathbb{P}_0\left(\Pi\left[\|f - f_0\|_{L^\infty(\nu)} \leq C'\varepsilon_n\right] < e^{-nC'^2\varepsilon_n^2}\right) \leq N^{-H}$$

To verify the entropy condition, consider for some $k > 0$ the set $\mathcal{F}_n = \cup_{\substack{J\leq kJ_n\\ h\in\mathcal{H}_J}}\mathcal{F}_{z_n,J,h}$, where $\mathcal{F}_{z,J,h}$ is defined in lemma 37 and $z_n = u(n\varepsilon_n^2)^{1/b_2}$ with u, k large enough. Then

$$\Pi\left[\mathcal{F}_n^c\right] \leq \Pi_J(J > kJ_n) + \sum_{\substack{J\leq kJ_n\\ h\in\mathcal{H}_J}} Je^{-b_1 z_n^{b_2}} \leq e^{-CJ_n} + (kJ_n)^2 e^{-b_1 u^{b_2} n\varepsilon_n^2} \max_J \#\mathcal{H}_J \leq e^{-Cn\varepsilon_n^2}$$

for any $C > 0$ by choosing k, u large enough, since $\#\mathcal{H}_J \leq K_1 \exp\left(K_2 J \ln N\right) \leq K_1 \exp\left(K_2' n\varepsilon_n^2\right)$ for some $K_1, K_2, K_2' > 0$. Finally for all $\zeta > 0$

$$N(\zeta\varepsilon_n, \mathcal{F}_n, \|\cdot\|_{L^\infty(\nu)}) \lesssim n\varepsilon_n^2$$

which concludes the proof. ■

D.3 Lemmas 36 and 37

Lemma 36. *Prior thickness*
Let $J_n \in \{1, \ldots, N\}$, $\quad J_n \geq \ln^\kappa N$, $\quad \kappa > d$ *and* h_n *satisfying assumption 2. Consider*

the prior defined by (3)*. Then for any* $H > 0$ *there exists* $c > 0$ *such that with the rate* $\epsilon_n(J_n, h_n)$ *as defined in* (5)*, We have*

$$\mathbb{P}_0\left(\Pi\left[\|f - f_0\|_{L^\infty(\nu)} \le c\varepsilon_n(J_n, h_n)|J_n, h_n\right] < \exp\left(-n\varepsilon_n(J_n, h_n)^2\right)\right) \le N^{-H}.$$

In particular, for any $\tau > d/2$ *and*

$$h_n = n^{-\frac{1}{2\beta+d}} (\ln n)^{-\frac{1-\tau-2(1+2\tau/d)\lceil\beta/2\rceil}{2\beta+d}}, J_n = \frac{h_n^{-d}}{\ln^\tau N}$$

then

$$\varepsilon_n(J_n, h_n) \asymp (\ln n)^{\frac{(2\tau+d)\lceil\beta/2\rceil+(1-\tau)\beta}{2\beta+d}} n^{-\frac{\beta}{2\beta+d}}.$$

Proof Throughout the proof we write $\varepsilon_n = \varepsilon_n(J_n, h_n)$. We use theorem 12 and distinguish the cases : for some $k > 0$ large enough, with $t = t_n = k\lambda_{J_n}^{-1} \ln N$

1. If $\beta \le 2$ we have

$$\begin{aligned}\forall f : V \to \mathbb{R}, \|f - f_0\|_{L^\infty(\nu)} &\le \left\|f - p_J\left(e^{-t\mathcal{L}} f_0\right)\right\|_{L^\infty(\nu)} + \left\|f_0 - p_J\left(e^{-t\mathcal{L}} f_0\right)\right\|_{L^\infty(\nu)} \\ &\lesssim \left\|f - p_J\left(e^{-t\mathcal{L}} f_0\right)\right\|_{L^\infty(\nu)} + O_{P_0}\left(h_n^\beta\left(\lambda_{J_n}^{-1} h_n^{-2} \ln N\right)\right).\end{aligned}$$

Hence for all $c > 0$, there exists $c' > 0$, such that

$$\left\|f - p_J\left(e^{-t\mathcal{L}} f_0\right)\right\|_{L^\infty(\nu)} \le c' h_n^\beta \lambda_{J_n}^{-1} h_n^{-2} \ln N \implies \|f - f_0\|_{L^\infty(\nu)} \le \varepsilon_n.$$

Moreover if $f = \sum_{j=1}^J z_j u_j$, using proposition 18 we get

$$\left\|f - p_J\left(e^{-t\mathcal{L}} f_0\right)\right\|_{L^\infty(\nu)} = \left\|\sum_{j=1}^J \left(z_j - \left\langle u_j | e^{-t\mathcal{L}} f_0\right\rangle_{L^2(\nu)}\right) u_j\right\|_{L^\infty(\nu)} \le N \max_{1\le j\le N} \left|z_j - \left\langle u_j | e^{-t\mathcal{L}} f_0\right\rangle_{L^2(\nu)}\right|.$$

Hence,

$$\begin{aligned}&\Pi\left[\left\|f - p_{J_n}\left(e^{-t\mathcal{L}} f_0\right)\right\|_{L^\infty(\nu)} \le c' h_n^\beta \lambda_{J_n}^{-1} h_n^{-2} \ln N \middle| J_n, h_n\right] \\ \ge &\Pi\left[\max_{1\le j\le N} \left|z_j - \left\langle u_j | e^{-t\mathcal{L}} f_0\right\rangle_{L^2(\nu)}\right| \le \frac{c' h_n^\beta \lambda_{J_n}^{-1} h_n^{-2} \ln N}{N}\right] \\ \ge &\prod_{j=1}^{J_n} \Pi\left[\left|z_j - \left\langle u_j | e^{-t\mathcal{L}} f_0\right\rangle_{L^2(\nu)}\right| \le \frac{c' h_n^\beta \lambda_{J_n}^{-1} h_n^{-2} \ln N}{N}\right] \\ \ge &\prod_{j=1}^{J_n} \Pi\left[\left|z_j - \left\langle u_j | e^{-t\mathcal{L}} f_0\right\rangle_{L^2(\nu)}\right| \le \frac{\frac{c'}{2} h_n^\beta \ln N}{N}\right] \\ \ge &\left(\frac{c' h_n^\beta \ln N}{N} \inf_{[-K,K]} \Psi\right)^{J_n} \mathrm{vol}(B_{\mathbb{R}^{J_n}}(0, 1)),\end{aligned}$$

where we have used $\lambda_{J_n} \leq 2h_n^{-2}$, which holds because $\|\mathcal{L}\|_{\mathcal{L}(L^\infty(\nu))} \leq 2$, and where $K > 0$ is chosen to be larger than

$$\max_{1\leq j\leq J_n} \left|\langle u_j | e^{-t\mathcal{L}} f_0 \rangle_{L^2(\nu)}\right| + \frac{c' h_n^\beta \ln N}{N} \lesssim 1 + \underbrace{\left\|e^{-t\mathcal{L}}\right\|_{\mathcal{L}(L^\infty(\nu))}}_{=1} \|f_0\|_{L^\infty(\mathcal{M})} < +\infty.$$

Hence

$$-\ln \Pi \left[\left\| f - p_J\left(e^{-t\mathcal{L}} f_0\right)\right\|_{L^\infty(\nu)} \leq c' h_n^\beta \lambda_{J_n}^{-1} h_n^{-2} \ln N \,\middle|\, J_n, h_n \right] \lesssim J_n \ln N.$$

Now $\varepsilon_n \geq \sqrt{\frac{J_n \ln N}{n}} \iff J_n \ln N \leq n\varepsilon_n^2$, therefore

$$\begin{aligned}
&\mathbb{P}_0\left(\Pi\left[\left\| f - p_J\left(e^{-t\mathcal{L}} f_0\right)\right\|_{L^\infty(\nu)} \leq \varepsilon_n \,\middle|\, J_n, h_n\right] < \exp\left(-n\varepsilon_n^2\right)\right) \\
\leq &\mathbb{P}_0\left(\left\| f_0 - p_{J_n}\left(e^{-t\mathcal{L}} f_0\right)\right\|_{L^\infty(\nu)} > \varepsilon_n/2\right) \\
\leq &n^{-H}
\end{aligned}$$

for any $H > 0$.

2. The case $\beta > 2$ is similar, the only difference being that we use the bound

$$\left|\langle u_j | e^{-t\mathcal{L}} f_t\rangle_{L^2(\nu)}\right| \leq \left\| \sum_{l=0}^k \frac{(t\mathcal{L})^l e^{-t\mathcal{L}}}{l!} \right\|_{\mathcal{L}(L^2(\nu))} \|f_0\|_{L^2(\nu)} \leq \|f_0\|_{L^\infty(\mathcal{M})} \sup_{s>0} \sum_{l=0}^k \frac{s^l e^{-s}}{l!} < +\infty$$

■

Lemma 37. *For some $z \geq 1$ and $J \in \{1, \ldots, N\}$ let*

$$\mathcal{F}_{z,J,h} = \left\{ \sum_{j=1}^J a_j u_j(h) : |a_j| \leq z \right\} \subset \mathbb{R}^V, \quad z > 0$$

Then

$$\Pi\left[\mathcal{F}_{z,J,h}^c | J, h\right] \leq J e^{-b_1 z^{b_2}}$$

and, there exists $\varepsilon_1, C_1 > 0$ such that for any $\varepsilon \leq \varepsilon_1$

$$N\left(\varepsilon, \mathcal{F}_{z,J,h}, \|\cdot\|_{L^\infty(\nu)}\right) \leq \left(\frac{C_1 z N}{\varepsilon}\right)^{2J}$$

Proof Proof of lemma 37
We drope the subscript h for ease of notation. Using the representation $f = \sum_{j=1}^J Z_j u_j$, the upper bound

$$\Pi\left[\mathcal{F}_{z,J}^c | J, h\right] \leq \sum_{j=1}^J \Pi\left[|Z_j| > z | J, h\right] \leq J\Psi\left([-z, z]^c\right) \leq J e^{-b_1 z^{b_2}}$$

follows by the assumption on the tails of Ψ. Moreover by proposition 18, if $\mathcal{E}$ is an $\frac{\varepsilon}{JN}$-net of $[-z,z]^J$ then $\left\{\sum_{j=1}^J f_j u_j : f \in \mathcal{E}\right\}$ is an ε-net of $\mathcal{F}_{z,J}$ (in $L^\infty(\nu)$). Since we can always take

$$\#\mathcal{E} \le \left(\frac{C_1 zJN}{\varepsilon}\right)^J \le \left(\frac{C_1 zN^2}{\varepsilon}\right)^J \le \left(\frac{C_1^{1/2} z^{1/2} N}{\varepsilon^{1/2}}\right)^{2J} \le \left(\frac{C_1^{1/2} cN}{\varepsilon}\right)^{2J}$$

for $\varepsilon \le \varepsilon_1$ and some $C_1, \varepsilon_1 > 0$ this gives the result by changing $C_1^{1/2}$ into a new C_1. ∎

References

Eddie Aamari and Clément Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. *The Annals of Statistics*, 47(1):177–204, February 2019. ISSN 0090-5364, 2168-8966. doi: 10.1214/18-AOS1685.

Eddie Aamari, Clément Berenfeld, and Clément Levrard. Optimal reach estimation and metric learning. *The Annals of Statistics*, 51(3):1086–1108, June 2023. ISSN 0090-5364, 2168-8966. doi: 10.1214/23-AOS2281.

Nir Ailon and Bernard Chazelle. The Fast Johnson–Lindenstrauss Transform and Approximate Nearest Neighbors. *SIAM Journal on Computing*, 39(1):302–322, January 2009. ISSN 0097-5397. doi: 10.1137/060673096.

Martin T. Barlow. *Random Walks and Heat Kernels on Graphs*. London Mathematical Society Lecture Note Series. Cambridge University Press, Cambridge, 2017.

Mikhail Belkin and Partha Niyogi. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

Mikhail Belkin and Partha Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, 15(6):1373–1396, June 2003. ISSN 0899-7667. doi: 10.1162/089976603321780317.

Clément Berenfeld and Marc Hoffmann. Density estimation on an unknown submanifold. *Electronic Journal of Statistics*, 15(1):2179–2223, January 2021. ISSN 1935-7524, 1935-7524. doi: 10.1214/21-EJS1826.

Mira Bernstein, Vin de Silva, John C Langford, and Joshua B Tenenbaum. Graph approximations to geodesics on embedded manifolds.

Jeff Calder and Nicolas Garcia Trillos. Improved spectral convergence rates for graph Laplacians on epsilon-graphs and k-NN graphs, June 2020.

Ismaël Castillo, Gérard Kerkyacharian, and Dominique Picard. Thomas Bayes' walk on manifolds. *Probability Theory and Related Fields*, 158(3):665–710, April 2014. ISSN 1432-2064. doi: 10.1007/s00440-013-0493-0.

Frédéric Chazal and Bertrand Michel. An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. *Frontiers in Artificial Intelligence*, 4, September 2021. ISSN 2624-8212. doi: 10.3389/frai.2021.667963.

Frédéric Chazal, Vin de Silva, Marc Glisse, and Oudot Steve. *The Structure and Stability of Persistence Modules*. Springer, 2016.

Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, July 2006. ISSN 1063-5203. doi: 10.1016/j.acha.2006.04.006.

Thierry Coulhon and Adam Sikora. Gaussian heat kernel upper bounds via the Phragmén-Lindelöf theorem. *Proceedings of the London Mathematical Society. Third Series*, 96(2):507–544, 2008. ISSN 0024-6115. doi: 10.1112/plms/pdm050.

Thierry Coulhon, Gerard Kerkyacharian, and Pencho Petrushev. Heat kernel generated frames in the setting of Dirichlet spaces, June 2012.

Francesco Denti, Diego Doimo, Alessandro Laio, and Antonietta Mira. The generalized ratios intrinsic dimension estimator. *Scientific Reports*, 12(1):20005, November 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-20991-1.

Vincent Divol and Frédéric Chazal. The density of expected persistence diagrams and its kernel based estimation. *Journal of Computational Geometry*, 10(2):127–153, 2019. ISSN 1920-180X. doi: 10.20382/jocg.v10i2a7.

Vincent Divol and Théo Lacombe. Understanding the topology and the geometry of the space of persistence diagrams via optimal partial transport. *Journal of Applied and Computational Topology*, 5(1):1–53, March 2021. ISSN 2367-1734. doi: 10.1007/s41468-020-00061-z.

David B. Dunson, Hau-Tieng Wu, and Nan Wu. Spectral Convergence of Graph Laplacian and Heat Kernel Reconstruction in L^∞ from Random Samples, June 2021.

David B. Dunson, Hau-Tieng Wu, and Nan Wu. Graph based Gaussian processes on restricted domains. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(2):414–439, 2022. ISSN 1467-9868. doi: 10.1111/rssb.12486.

Bernardo Fichera, Viacheslav Borovitskiy, Andreas Krause, and Aude Billard. Implicit Manifold Gaussian Process Regression, February 2024.

Nicolás García Trillos, Moritz Gerlach, Matthias Hein, and Dejan Slepčev. Error Estimates for Spectral Convergence of the Graph Laplacian on Random Geometric Graphs Toward the Laplace–Beltrami Operator. *Foundations of Computational Mathematics*, 20(4):827–887, August 2020. ISSN 1615-3383. doi: 10.1007/s10208-019-09436-w.

Christopher Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Minimax Manifold Estimation. *Journal of Machine Learning Research*, 13(43):1263–1291, 2012. ISSN 1533-7928.

S. Ghosal and A. van der Vaart. Convergence rates of posterior distributions for non iid observations. *aos*, 35(1):192–223, 2007.

Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2017. ISBN 978-0-521-87826-5. doi: 10.1017/9781139029834.

Subhashis Ghosal, Jayanta K. Ghosh, and Aad W. van der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531, April 2000. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1016218228.

Franziska Göbel and Gilles Blanchard. Volume Doubling Condition and a Local Poincar\'e Inequality on Unweighted Random Geometric Graphs, March 2020.

Alden Green, Sivaraman Balakrishnan, and Ryan Tibshirani. Minimax Optimal Regression over Sobolev Spaces via Laplacian Regularization on Neighborhood Graphs. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 2602–2610. PMLR, March 2021.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 01621459, 1537274X.

William B. Johnson, Joram Lindenstrauss, and Gideon Schechtman. Extensions of lipschitz maps into Banach spaces. *Israel Journal of Mathematics*, 54(2):129–138, June 1986. ISSN 1565-8511. doi: 10.1007/BF02764938.

Jisu Kim, Alessandro Rinaldo, and Larry Wasserman. Minimax Rates for Estimating the Dimension of a Manifold. *Journal of Computational Geometry*, pages 42–95 Pages, February 2019. doi: 10.20382/jocg.v10i1a3.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, October 2000. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1015957395.

Jialu Liu and Jiawei Han. Spectral Clustering. In *Data Clustering*. Chapman and Hall/CRC, 2014. ISBN 978-1-315-37351-5.

David Loiseaux, Mathieu Carrière, and Andrew J. Blumberg. A Framework for Fast and Stable Representations of Multiparameter Persistent Homology Decompositions, June 2023.

David Loiseaux, Luis Scoccola, Mathieu Carrière, Magnus Bakke Botnan, and Steve Oudot. Stable Vectorization of Multiparameter Persistent Homology using Signed Barcodes as Measures, February 2024.

Zacharie Naulet and Éric Barat. Some Aspects of Symmetric Gamma Process Mixtures. *Bayesian Analysis*, 13(3):703–720, September 2018. ISSN 1936-0975, 1931-6690. doi: 10.1214/17-BA1058.

J. R. Norris. *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1997. ISBN 978-0-521-63396-3. doi: 10.1017/CBO9780511810633.

Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space. November 1901. doi: 10.1080/14786440109462720.

Mathew Penrose. *Random Geometric Graphs*. Oxford University Press, May 2003. ISBN 978-0-19-850626-3. doi: 10.1093/acprof:oso/9780198506263.001.0001.

Paul Rosa, Slava Borovitskiy, Alexander Terenin, and Judith Rousseau. Posterior Contraction Rates for Matérn Gaussian Processes on Riemannian Manifolds. *Advances in Neural Information Processing Systems*, 36:34087–34121, December 2023.

Judith Rousseau. On the frequentist properties of bayesian nonparametric methods. *Annual Review of Statistics and Its Application*, 3(Volume 3, 2016):211–231, 2016. ISSN 2326-831X. doi: 10.1146/annurev-statistics-041715-033523.

Sam T. Roweis and Lawrence K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, December 2000. doi: 10.1126/science.290.5500.2323.

Daniel Sanz-Alonso and Ruiyi Yang. Unlabeled Data Help in Graph-Based Semi-Supervised Learning: A Bayesian Nonparametrics Perspective. *Journal of Machine Learning Research*, 23(97):1–28, 2022. ISSN 1533-7928.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10(5):1299–1319, July 1998. ISSN 0899-7667. doi: 10.1162/089976698300017467.

Zhaoyang Shi, Krishnakumar Balasubramanian, and Wolfgang Polonik. Adaptive and non-adaptive minimax rates for weighted Laplacian-eigenmap based nonparametric regression, October 2023.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, December 2000. doi: 10.1126/science.290.5500.2319.

Aad van der Vaart and Harry van Zanten. Information Rates of Nonparametric Gaussian Process Methods. *Journal of Machine Learning Research*, 12(60):2095–2119, 2011. ISSN 1533-7928.

Aad Van der Vaart and Jon Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, December 2007. ISSN 1573-1375. doi: 10.1007/s11222-007-9033-z.

Caroline L. Wormell and Sebastian Reich. Spectral Convergence of Diffusion Maps: Improved Error Bounds and an Alternative Normalization. *SIAM Journal on Numerical Analysis*, 59(3):1687–1734, January 2021. ISSN 0036-1429. doi: 10.1137/20M1344093.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Nonparametric regression on random geometric graphs sampled from submanifolds
Publication Status	Submitted for Publication
Publication Details	**R. & Rousseau,** *Nonparametric regression on random geometric graphs sampled from submanifolds* (2024)

Student Confirmation

Student Name:	Paul Rosa		
Contribution to the Paper	I am the first author of this paper. I did most of the theoretical analysis (approximation results, entropy bounds, heat kernel bounds) and most of the writing, in concertation with and under the supervision of my supervisor Judith Rousseau.		
Signature Paul Rosa		Date	06/01/2025

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Prof Judith Rousseau			
Supervisor comments I confirm the contribution statement			
Signature		Date	06/01/2025

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 4

Estimating a density near an unknown manifold: a Bayesian nonparametric approach

Berenfeld, R. & Rousseau, *Estimating a density near an unknown manifold: a Bayesian nonparametric approach* (2024), The Annals of Statistics

The Annals of Statistics
2024, Vol. 52, No. 5, 2081–2111
https://doi.org/10.1214/24-AOS2423

ESTIMATING A DENSITY NEAR AN UNKNOWN MANIFOLD: A BAYESIAN NONPARAMETRIC APPROACH

BY CLÉMENT BERENFELD[1,a], PAUL ROSA[2,b] AND JUDITH ROUSSEAU[2,c]

[1]*Institut für Mathematik, Universität Potsdam,* [a]*berenfeld@uni-potsdam.de*
[2]*Department of Statistics, University of Oxford,* [b]*paul.rosa@jesus.ox.ac.uk,* [c]*judith.rousseau@stats.ox.ac.uk*

We study the Bayesian density estimation of data living in the offset of an unknown submanifold of the Euclidean space. In this perspective, we introduce a new notion of anisotropic Hölder for the underlying density and obtain posterior rates that are minimax optimal and adaptive to the regularity of the density, to the intrinsic dimension of the manifold, and to the size of the offset, provided that the latter is not too small—while still allowed to go to zero. Our Bayesian procedure, based on location-scale mixtures of Gaussians, appears to be convenient to implement and yields good practical results, even for quite singular data.

1. Introduction.

1.1. *Manifold density estimation.* In many high dimensional statistical problems it is common to consider that the data has an intrinsic low dimensional structure. More precisely, statistics and computer sciences have seen a growing interest in the so-called *manifold hypothesis* where the data is believed to be supported (or near supported) on a low dimensional submanifold M of an ambient space (see [44] for an introduction).

There are good intuitive reasons to believe that real world data (such as natural images, sounds, texts, etc) belong to the vicinity of a low dimensional submanifold, often due to physical constraints; see, for instance, [42] or [23]. Empirical evidence has also been shown in a number of important cases such as texts data sets [5], sounds [5, 40], images and videos [39, 64] or more recently in Covid data [48]. Analysing such data sets is often called manifold learning (see [44] for an introduction). Manifold learning deal with either nonlinear dimension reduction techniques, manifold estimation or the construction of generative models and the estimation of the distribution on or near an unknown manifold. These problems are strongly connected. Dimension reduction consists in finding low dimensional representations of the data. This is typically done by constructing mappings as in Kernel PCA [59] or graph based methods such as Isomap [62], Locally Linear Embeddings [58] or Laplacian Eigenmaps [6]. Instead of estimating an embedding, the problem of reconstructing the manifold is another popular aspect of manifold learning; see [1, 18, 24, 55] or [21] among others. Finally, the estimation of the distributions on or near manifolds and the construction of generative models have received recent wide interests in the statistics and machine learning communities, specially with the developments of deep learning algorithms. There is a growing literature on generative models under the manifold hypothesis with many methodological developments around variational autoencoders (see [33], Section 14.6, or [39]), Generative adversarial networks (see [3, 4, 34] among others) or recent versions of normalizing flows (see [36]). The theoretical results associated to these approaches control the error between the true generative process and the estimated generative models typically under adversarial

Received September 2022; revised June 2024.
MSC2020 subject classifications. Primary 62G07, 62G20; secondary 53A07.
Key words and phrases. Density estimation, Bayesian nonparametrics, minimax adaptive estimation, posterior concentration rates, manifold learning.

losses such as the Wasserstein distance, as in [61], since the focus is more on generating interesting samples than on estimating the distribution per se.

In this paper, we study the estimation of the density in the vicinity of an unknown submanifold M of the ambient space $\mathbb{R}^D$, of dimension $1 \le d \le D-1$. Density estimation is a canonical problem in statistics and machine learning and in addition to being of interest in itself can be used as an intermediate steps in many tasks of unsupervised or supervised learning such as clustering, prediction and classification, dimensionality reduction or in ridge estimation; see, for instance, [15, 25, 49] among many others. Density or distribution estimation under the exact manifold hypothesis (assuming that the data belong to a submanifold) has been studied theoretically for instance, in [54] or [19] under Wasserstein losses and in [7] under the pointwise loss. Assuming that the data belong exactly to a smooth manifold may be too restrictive since signals are often corrupted by noise. Hence in this paper, we assume that the data belong to a neighbourhood $M^\delta = \{x \in \mathbb{R}^D; d(x, M) \le \delta\}$, where neither M nor δ or the density f are known. This problem is studied in [14] in the special case of data corrupted with Gaussian noise and [49] proposes a Bayesian nonparametric method to estimate a density on M^δ based on mixtures of Fisher—Gaussian distributions for which they prove consistency under the assumption that the width δ of the tube M^δ is fixed.

1.2. *Our approach and contributions*. As far as we are aware, there is no theoretical results on convergence rates—either from a frequentist or a Bayesian approach—for estimating a density on tubes M^δ when M and δ are unknown and δ is possibly small. In this paper, we bridge this gap and we propose a Bayesian nonparametric method based on specific families of location-scale mixtures of Gaussian distributions. We study the posterior concentration rates associated to these priors, that is, the smallest possible ε_n such that

$$\Pi\big(d(f_0, f) \le \varepsilon_n | X_1, \dots, X_n\big) \to 1,$$

in probability when the data $X_1, \dots, X_n$ are a n sample from f_0 and where $\Pi(\cdot | X_1, \dots, X_n)$ denotes the posterior distribution; see [27]. As is well known, when the distance $d(.,.)$ is the Hellinger or the L_1 metric, this posterior concentration rates induces also a convergence rate ε_n for the posterior mean $\hat{f}$; see, for instance, [27]. Typically, the rate ε_n depends on regularity properties of the density f_0 and on the prior.

To do so, we first define a general mathematical framework describing regularity properties of densities defined on possibly small neighbourhoods M^δ of submanifolds M, with the idea that the density has a given smoothness β_0 along the manifold M and another smoothness $\beta_\perp$ along the normal to the manifold. This manifold driven anisotropic smoothness is defined in Section 2.2 and is an extension to anisotropic Hölder functions along coordinate axes. Because the corruption of a signal by some small noise can be modelled in many different ways, with additive homoscedastic noise being the simplest and most restrictive one, our framework only assumes that the density lives on M^δ and does not model the noise. As shown in Section 2, our framework encompasses common models of noisy data.

Building on that we show that the posterior concentration rate depends on β_0, $\beta_\perp$ together with the dimension d of M and the width δ of the tube. Interestingly the prior Π does not need to depend on β_0, $\beta_\perp$, d, δ or M which makes the approach fully adaptive and the rate we obtain, at least when δ is not too small is of order (up to a polylog term)

$$n^{-\gamma} \quad \text{with } \gamma = \frac{\beta_0}{2\beta_0 + d + (D-d)\beta_0/\beta_\perp},$$

where we recall that D is the dimension of the ambient space, and is minimax.

Nonparametric location mixtures of Gaussians are known to be flexible models for densities, and adaptive minimax rates of convergence on Hölder types spaces have been obtained using Bayesian or frequentist estimation procedures based on location mixtures of normals; see [41, 60], and [28] for Bayesian methods and [46] for a penalized likelihood approach. However, location mixtures are not versatile enough since the covariance matrix remains fixed across the components, so we instead take advantage of the flexibility of location-scale mixtures of Gaussians. In [12], the authors derive a suboptimal posterior concentration rate for isotropic positive Hölder densities on $\mathbb{R}^D$, while [46] obtained minimax convergence rates for penalized maximum likelihood methods based on the same type of location-scale mixtures and [50] obtained also minimax posterior concentration rates using a hybrid location-scale mixture prior in the regression model. These results thus indicate that one has to be careful in designing the prior in nonparametric location-scale mixtures of Gaussians. The priors we consider in this paper are variants of location-scale mixture priors, see Section 2.3, which are flexible enough to adapt to the nonlinear or manifold driven smoothness of the class of densities studied here. This prior construction can also be seen as tiling the manifold by low-rank *Gaussian pancakes*, a method that is similar to mixtures of factors analyzers [20, 26] or manifold Parzen windows [63] where, however, no theoretical guarantees on the estimation of the density were proven.

Hence our contributions are both methodological and theoretical. From a methodological point of view, we provide with a family of versatile priors (see Section 2) that are shown empirically and theoretically to perform very well in modelling data that are singularly supported near submanifolds. In particular, we show empirically that these variants of location-scale mixtures of Gaussian priors behave much better than the standard *conjuguate* location-scale mixture of Gaussian prior; see Section 4. From a theoretical point of view, we introduce a new notion of Hölder smoothness along a submanifold (see Section 2.2) which is proving to be adequate for the study of such almost-degenerate densities, and we derive posterior concentration rates for this new model (Section 3.1). The rates are optimal if the data do not collapse too quickly towards the manifold. These results rely on an intermediate result in approximation theory which has interests in its own right and is provided in Section 3.2.

1.3. *Organisation of the paper.* In Section 2, we define manifold-anisotropic Hölder function, together with the families of priors we consider in the paper. Section 3 contains the main theoretical results and Section 4 the empirical, numerical results. We provide in Section 5 proofs of the main results, namely the contraction rate Theorem 3.1 and the approximation Theorem 3.4. Some useful facts on manifolds are presented in the Appendix. The Supplementary Material [8] contains additional proofs and lemmata, as well as details on the numerical setting of Section 4. References to equations, theorems, propositions, lemmata and sections in [8] are preceded with the letter S, to avoid repeated references with this manuscript.

1.4. *Notations.* For a multi-index $k = (k_1, \dots, k_D) \in \mathbb{N}^D$, we set $|k| = k_1 + \cdots + k_D$ and $k! = k_1! \dots k_D!$. For $x \in \mathbb{R}^D$, we write $x^k = x_1^{k_1} \dots x_D^{k_D} \in \mathbb{R}$ and $x_{\max}$ (resp. $x_{\min}$) to be the maximal (resp. minimal) value of its entries. For any two indices $i, j \in \{1, \dots, D\}$ with $i \le j$, we set $x_{i:j} = (x_i, \dots, x_j) \in \mathbb{R}^{j-i+1}$. Finally, for a sufficiently regular function $f : \mathbb{R}^D \to \mathbb{R}$, we define its kth partial derivative as

$$\mathrm{D}^k f(x) = \frac{\partial^{|k|} f}{\partial x_1^{k_1} \dots \partial x_D^{k_D}}(x).$$

When $M \subset \mathbb{R}^D$ is a measurable subset with Hausdorff dimension d, μ_M denotes the Borel measure $\mu_M = \mathcal{H}^d(\cdot \cap M)$ where $\mathcal{H}^d$ is the d-dimensional Hausdorff measure on $\mathbb{R}^D$. For $r > 0$ and $x \in \mathbb{R}^D$, let $\mathsf{B}_M(x, r) = \mathsf{B}(x, r) \cap M$ where $\mathsf{B}(x, r)$ is the usual Euclidean ball of

$\mathbb{R}^D$. If M is closed, then pr_M defines the (possible multi-valued) orthonormal projection from $\mathbb{R}^D$ to M.

We will denote by $\|\cdot\|$ the usual Euclidean norm of $\mathbb{R}^k$ for any $k \in \mathbb{N}^*$. When $\mathcal{L}$ is a linear map between such spaces, we write $\|\mathcal{L}\|_{\mathrm{op}}$ for the operator norm associated with the Euclidean norms. The notation $\|\cdot\|_1$ (resp. $\|\cdot\|_\infty$) will refer to both the L^1-norm (resp sup-norm) for vectors of $\mathbb{R}^k$ for any $k \in \mathbb{N}^*$, and to the L^1-norm (resp sup-norm) for measurable functions from $\mathbb{R}^k$ to $\mathbb{R}$ for any $k \in \mathbb{N}^*$. The brackets $\langle \cdot, \cdot \rangle$ will be used to denote the usual Euclidean product in $\mathbb{R}^k$ for any $k \in \mathbb{N}^*$. For any matrix $A \in \mathbb{R}^{k\times k}$, the notation $\|\cdot\|_A^2$ will refer to the quadratic form over $\mathbb{R}^k$ defined by $x \mapsto \langle Ax, x\rangle$, which is a squared norm if A is positive definite. The set of orthogonal transform of $\mathbb{R}^D$ will be denoted by $\mathcal{O}(D, \mathbb{R})$, or sometimes simply $\mathcal{O}(D)$.

For two positive functions $f, g : \mathbb{R}^D \to \mathbb{R}$, we write the Hellinger distance as

$$\mathrm{d}_{\mathrm{H}}(f, g) = \left\{ \int_{\mathbb{R}^D} \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx \right\}^{1/2}.$$

In this paper, M will designate a closed submanifold of $\mathbb{R}^D$ of dimension $1 \leq d \leq D - 1$. For any point $x \in M$, the tangent and normal spaces of M at x will be denoted $T_x M$ and $N_x M$, and the corresponding bundles TM and NM. We write $\exp_x : (T_x M, 0) \to (M, x)$ for the exponential map of M at point x. We let $d_M(x, y)$ denote the intrinsic distance between x and y in M.

Finally, we will use throughout the symbols $\simeq$, $\lesssim$ and $\gtrsim$ to denote equalities or inequalities up to a constant, when the constant is not important.

2. Model: Distributions concentrated near manifolds. We assume that we observe $X_1, \ldots, X_n$ independent and identically distributed from P_0 on $\mathbb{R}^D$ with density f_0 with respect to Lebesgue measure. We assume that there is a low dimensional structure underlying our observations, that is, that f_0 has support concentrated near a low dimensional manifold M which is unknown. More precisely there exists $\delta > 0$ unknown and typically small such that $P_0(M^\delta) = 1$, where M^δ is the δ-offset of M: it is the set of points that are at distance less than δ from M,

$$M^\delta := \bigcup_{x \in M} \mathsf{B}(x, \delta) = \left\{ z \in \mathbb{R}^D \mid d(z, M) \leq \delta \right\}.$$

A typical example is when the observations are noisy versions of data whose support is M: $X = Y + Z$ with $Y \in M$ and $|Z| \leq \delta$ almost surely. We would like to underline here that δ is to be understood as $\delta = \delta_n$ which can either stay constant (and smaller than the reach of M, defined in the Appendix) or goes to 0 as n goes to infinity, in which case $f_0 = f_{0,n}$ and $X_j = X_{j,n}$ is a triangular array of data. This dependence on n for $f_{0,n}$ does not impact our results and in order to make the notation lighter, we intentionally drop the extra n from the indices, and will recall this dependency at any time when it might create confusion. When δ is fixed, the density is not degenerate but its support has a nontrivial geometry. Allowing $\delta = \delta_n = o(1)$ is a way to model very concentrated densities (for a given n) and is similar in spirit to high dimensional regression where the number of covariates p is modelled as a function of n. In other words, it is a way to take into account the impact of δ in the concentration properties of the posterior distribution. When the noise Z has a density smoother than the density of Y on M (with respect to the Hausdorff measure), the density of X is anisotropic with a smoothness along the manifold M smaller than that along the normal directions. In this paper, we thus aim at constructing priors which are flexible enough to lead to *good* estimation of f_0 in situations where the density has a complex anisotropic structure in that it has an unknown smoothness β_0 along an unknown manifold M and a different (larger) smoothness $\beta_\perp$, also unknown, along the normal spaces of the manifold. In this context, since the anisotropy varies spatially, it is therefore important to consider priors which adapt spatially to such *non linear smoothness*. In Section 2.3 below, we consider two families of location-scale mixtures with a new and careful modelling of the prior on the variance of the components and we show in Section 3 that these priors are well behaved for densities with manifold driven smoothness.

To begin with, we define what we think is a new notion of anisotropic Hölder spaces on the Euclidean space $\mathbb{R}^D$, and which happens to be a natural extension of the usual notion of (isotropic) Hölder smoothness. We are aware that there exist various notions of anisotropic smoothness (see, for instance, [17, 31, 32, 35, 37]) with most of them stemming from the anisotropic smoothness as defined in [52]. In all the aforementionned references, the anisotropy was consistently defined as a control of the variations of the partial derivatives along each axis separately, with no control of the cross-derivatives (and no guarantee that they, in fact, exist). While this is enough in a Euclidean framework, we argue that, to the best of our effort, we could not make such assumptions sufficient in our nonlinear setting, as the proofs presented in Section 5 or in the Supplementary Material might highlight. Instead, we come out with a new notion of Hölder anisotropy, in the footsteps of what [60] already sketched in their paper, that handles cross-derivatives in the same way that the usual notion of (isotropic) Hölder smoothness does, and which in fact coincides with the latter when the anisotropy vector is isotropic. This new class is defined in the following subsection and its main properties reviewed in [8], Section S2.1. We would also like to mention the notion of mixed smoothness as introduced in [16], which is a stronger notion of regularity than ours: the classes defined below can be seen as nested between the ones of [52] and [16].

2.1. *General anisotropic Hölder functions*. An anisotropic Hölder functions $f:\mathbb{R}^D\to\mathbb{R}$ is, informally, a function whose smoothness is different along each axis of $\mathbb{R}^D$. Letting $\boldsymbol{\beta}=(\beta_1,\dots,\beta_D)\in(\mathbb{R}_+^*)^D$, which will represent the regularity indices along each axis, we define

$$\boldsymbol{\alpha}=(\alpha_1,\dots,\alpha_D)\quad\text{where } \alpha_i=\beta/\beta_i\in[0,D]\quad\text{and}\quad \beta^{-1}=\frac{1}{D}\sum_i\beta_i^{-1}. \tag{1}$$

The coefficient β acts as the effective smoothness of the function f. Notice that $\alpha_1+\dots+\alpha_D=D$. In this section, we define the spaces of anisotropic functions over bounded open subset of $\mathbb{R}^D$. We defer to [8], Section S2.1, the introduction of the same class over general open subsets. We let $\mathcal{U}\subset\mathbb{R}^D$ be a bounded open subset and $L:\mathcal{U}\to\mathbb{R}_+$ be any nonnegative function.

DEFINITION 2.1. The anisotropic Hölder spaces $\mathcal{H}_{\mathrm{an}}^{\boldsymbol{\beta}}(\mathcal{U},L)$ is the set of all functions $f:\mathcal{U}\to\mathbb{R}$ satisfying:

(i) For any multi-index $k\in\mathbb{N}^D$ such that $\langle k,\boldsymbol{\alpha}\rangle<\beta$, the partial derivative $\mathrm{D}^k f$ is well defined on $\mathcal{U}$ and $|\mathrm{D}^k f(x)|\le L(x)$ for all $x\in\mathcal{U}$;

(ii) For any multi-index $k\in\mathbb{N}^D$ such that $\beta-\alpha_{\max}\le\langle k,\boldsymbol{\alpha}\rangle<\beta$, there holds

$$\left|\mathrm{D}^k f(y)-\mathrm{D}^k f(x)\right|\le L(x)\sum_{i=1}^D|y_i-x_i|^{\frac{\beta-\langle k,\boldsymbol{\alpha}\rangle}{\alpha_i}\wedge 1}\quad\forall x,y\in\mathcal{U}. \tag{2}$$

See Figure 1 for a graphical representation of the quantities at stake. The function L acts as an upper-bound for the localized and anisotropic version of the usual Hölder-norm:

$$\Big\{\max_{\langle k,\boldsymbol{\alpha}\rangle<\beta}|\mathrm{D}^k f(x)|\Big\}\vee\max_{\beta-\alpha_{\max}\le\langle k,\boldsymbol{\alpha}\rangle<\beta}\sup_{y\in\mathcal{U}}\frac{|\mathrm{D}^k f(y)-\mathrm{D}^k f(x)|}{\sum_{i=1}^D|y_i-x_i|^{\frac{\beta-\langle k,\boldsymbol{\alpha}\rangle}{\alpha_i}\wedge 1}}.$$

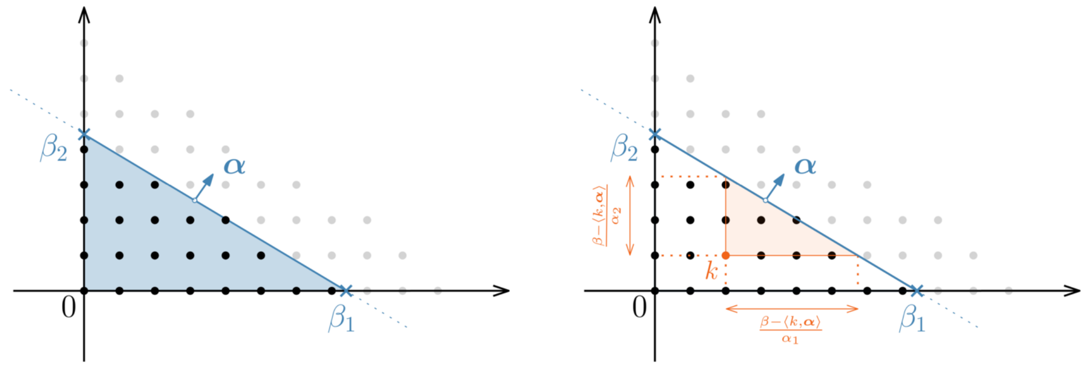


FIG. 1. *An example in dimension $D = 2$. The vector $\boldsymbol{\alpha}$ is the only vector of 1-norm D which has positive coordinates and which is orthogonal to the simplex of vertices $\{\beta_i e_i\}_{1\le i\le D}$. In black are the points k of $\mathbb{N}^2$ such that $\langle k, \boldsymbol{\alpha}\rangle < \beta$.*

Note that constraint (i) on the intermediate derivatives $\mathrm{D}^k f(x)$ for $0 < \langle k, \boldsymbol{\alpha}\rangle < \beta$ may seem superfluous since some Kolmogorov-Landau type inequalities would yield some bounds on these derivatives, but we add them nonetheless to our functional class to simplify some notations. We list and prove in [8], Section S2.1, various useful properties of functions in the anisotropic Hölder class. Also the function L in the definition of $\mathcal{H}^{\boldsymbol{\beta}}_{\mathrm{an}}(\mathcal{U}, L)$ can be constant, in which case we will typically denote it C, to make it more explicit (leading to $\mathcal{H}^{\boldsymbol{\beta}}_{\mathrm{an}}(\mathcal{U}, C)$).

REMARK 1. The usual isotropic Hölder spaces are special cases of our definition of $\mathcal{H}^{\boldsymbol{\beta}}_{\mathrm{an}}(\mathcal{U}, L)$ corresponding to $\boldsymbol{\beta} = (\beta, \dots, \beta)$ with $\beta > 0$. In this case, we write

$$\mathcal{H}^{\beta}_{\mathrm{iso}}(\mathcal{U}, L) := \mathcal{H}^{\boldsymbol{\beta}}_{\mathrm{an}}(\mathcal{U}, L) \quad \text{for } \boldsymbol{\beta} = (\beta, \dots, \beta).$$

As a final remark, we will use the same notations for the spaces of multivalued functions when their coordinate functions are all in the corresponding space. For instance, if $\Psi : \mathcal{U} \to \mathbb{R}^{D_1}$, for $D_1 > 1$, then

$$\Psi = (\Psi_1, \dots, \Psi_{D_1}) \in \mathcal{H}^{\boldsymbol{\beta}}_{\mathrm{an}}(\mathcal{U}, L) \quad \underset{\text{def}}{\Leftrightarrow} \quad \Psi_i \in \mathcal{H}^{\boldsymbol{\beta}}_{\mathrm{an}}(\mathcal{U}, L) \text{ for all } i \in \{1, \dots, D_1\},$$

and the same holds for the other spaces defined in this subsection.

2.2. *Manifold anisotropic Hölder functions.* We now consider functions whose smoothness directions at point $x \in \mathbb{R}^D$ are dependent on the position of x with respect to a given submanifold $M \subset \mathbb{R}^D$ of dimension $1 \le d \le D - 1$. More specifically, we extend the above notions of anisotropy to functions with a given regularity in the tangential directions of M, and of another regularity in the normal directions of M. We call such functions *manifold-anisotropic Hölder*, or sometimes simply *M-anisotropic*. To define such a class of functions, we assume that M is a closed submanifold with reach bounded from below by $\tau > 0$ (see the Appendix for definition and properties of the reach) and we consider local parametrizations at any $x_0 \in M$

$$\Psi_{x_0} : \mathcal{V}_{x_0} \to M,$$

where $\mathcal{V}_{x_0}$ is a neighborhood of 0 in $T_{x_0}M$. The maps Ψ_{x_0} can be taken in a wide class of parametrizations of M. For instance, one could consider taking Ψ_{x_0} to be (close to) the inverse projection over $M \to T_{x_0}M$ where $T_{x_0}M$ is seen as an affine subspace of $\mathbb{R}^D$ going through x_0; see, for instance, [1] or [19]. For purely practical matter, we choose Ψ_{x_0} to be the exponential map $\exp_{x_0}$, although the results in this paper could be carried out with other well-behaved parametrizations, such as the one mentioned above. In particular, in the case of the exponential maps, we can define the domain of Ψ_{x_0} to be $\mathsf{B}_{T_{x_0}M}(0,\pi\tau)$; see the Appendix. In the rest of this paper, we set

$$\mathcal{V}_{x_0} := \mathsf{B}_{T_{x_0}M}(0,\tau/8),$$

with factor $1/8$ being there for technical reasons. If all the maps Ψ_{x_0} are of regularity $\beta_M > 1$, meaning that there exists a constant $C_M > 0$ such that

$$\Psi_{x_0} \in \mathcal{H}^{\beta_M}_{\text{iso}}(\mathcal{V}_{x_0}, C_M), \quad \forall x_0 \in M \tag{3}$$

(in particular M is at least $\mathcal{C}^k$ with $k = \lceil \beta_M - 1 \rceil$), then one can construct a map

$$\bar{\Psi}_{x_0} : \begin{cases} \mathcal{V}_{x_0} \times N_{x_0}M \to \mathbb{R}^D, \\ (v,\eta) \mapsto \Psi_{x_0}(v) + N_{x_0}(v,\eta), \end{cases}$$

where $N_{x_0}(v,\cdot)$ is an isometry from $N_{x_0}M$ to $N_{\Psi_{x_0}(v)}M$ and where $v \mapsto N_{x_0}(v,\cdot) \in \mathcal{H}^{\beta_M-1}_{\text{iso}}(\mathcal{V}_{x_0}, C^{\perp}_M)$ for some other constant $C^{\perp}_M$ depending on C_M, τ and β_M. We refer to the Appendix for further details concerning the construction of $\bar{\Psi}_{x_0}$ and the proof of its regularity. When restricting the latter map, one gets a local parametrization of the offset $M^{\tau/2}$ around x_0

$$\bar{\Psi}_{x_0} : \mathcal{V}_{x_0} \times \mathsf{B}_{N_{x_0}M}(0,\tau/2) \to M^{\tau/2}$$

as shown in Lemma A.2. This parametrization is such that $\mathrm{pr}_M(\bar{\Psi}_{x_0}(v,\eta)) = \Psi_{x_0}(v)$ for any $(v,\eta) \in \mathcal{V}_{x_0} \times \mathsf{B}_{N_{x_0}M}(0,\tau/2)$ and $\bar{\Psi}_{x_0}$ is a diffeomorphism from $\mathcal{V}_{x_0} \times \mathsf{B}_{N_{x_0}M}(0,\tau/2)$ to its image which satisfies

$$\bar{\Psi}_{x_0} \in \mathcal{H}^{\beta_M-1}_{\text{iso}}(\mathcal{V}_{x_0} \times \mathsf{B}_{N_{x_0}M}(0,\tau/2), C^*_M) \tag{4}$$

for some $C^*_M > 0$ depending on C_M, τ and β_M. See Figure 2 for a visual interpretation of this parametrizations.

For any $\delta > 0$, we define $\bar{\Psi}_{x_0,\delta}(v,\eta) := \bar{\Psi}_{x_0}(v,\delta\eta)$ to be the rescaled version of $\bar{\Psi}_{x_0}$ in the normal directions. It is a well defined parametrization of $M^{\tau/2}$ on the set $\mathcal{W}_{x_0,\delta} := \mathcal{V}_{x_0} \times \mathsf{B}_{N_{x_0}M}(0,\tau/2\delta)$. We let β_0, $\beta_\perp$ be two positive real numbers, and define the vector

$$\boldsymbol{\beta}_{0,\perp} = (\underbrace{\beta_0,\dots,\beta_0}_{d}, \underbrace{\beta_\perp,\dots,\beta_\perp}_{D-d}) \in \mathbb{R}^D$$

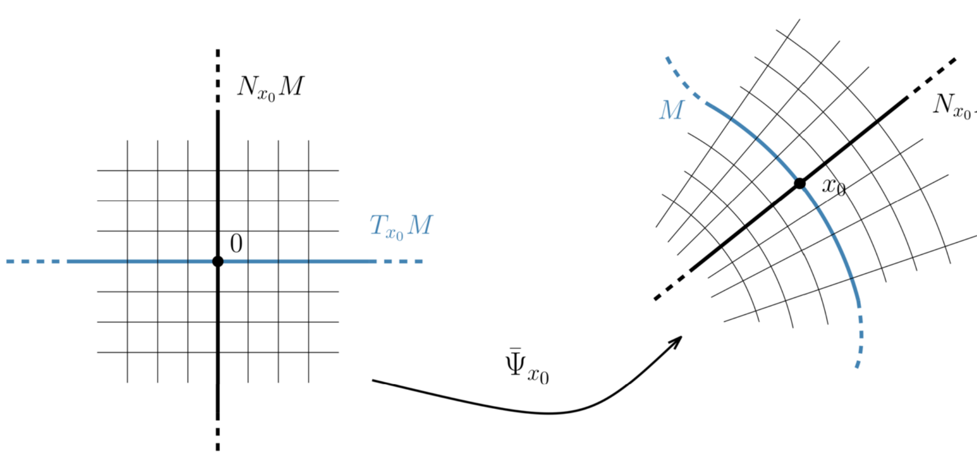


FIG. 2. *A visual interpretation of the parametrization $\bar{\Psi}_{x_0}$.*

and define $\boldsymbol{\alpha}$ to be the counterpart of $\boldsymbol{\beta}_{0,\perp}$ as in (1):

$$\boldsymbol{\alpha} = (\alpha_0, \dots, \alpha_0, \alpha_\perp, \dots, \alpha_\perp) \quad \text{with } \alpha_0 = \beta/\beta_0, \quad \alpha_\perp = \beta/\beta_\perp, \tag{5}$$
$$\text{and } \beta^{-1} = \frac{d}{D}\beta_0^{-1} + \frac{D-d}{D}\beta_\perp^{-1}.$$

Now for any function $L : \mathbb{R}^D \to \mathbb{R}_+$, we define the following.

DEFINITION 2.2. Let $L : \mathbb{R}^D \to \mathbb{R}_+$ be a function; the class $\mathcal{H}_\delta^{\beta_0,\beta_\perp}(M, L)$ is the set of all functions $f : \mathbb{R}^D \to \mathbb{R}$ which satisfy:

(i) f is supported on M^δ;
(ii) For any $x_0 \in M$, set $\bar{f}_{x_0,\delta} := \delta^{D-d} f \circ \bar{\Psi}_{x_0,\delta}$ and $L_{x_0,\delta} := \delta^{D-d} L \circ \bar{\Psi}_{x_0,\delta}$, then

$$\bar{f}_{x_0,\delta} \in \mathcal{H}_{\mathrm{an}}^{\boldsymbol{\beta}_{0,\perp}}(\mathcal{W}_{x_0,\delta}, L_{x_0,\delta}). \tag{6}$$

Informally, such a function is β_0-Hölder along the manifold M, and $\beta_\perp$-Hölder normal to the manifold M. The normalization δ^{D-d} accounts for the scaling $\eta \mapsto \delta\eta$ along the normal spaces (which are of dimension $D-d$) in the definition of $\bar{\Psi}_{x_0,\delta}$. Its presence is natural and can be understood as follows: when f is a density supported on M^δ, the typical magnitude of its values is of order $1/\delta^{D-d}$, and the absence of normalization would whence make the above functional class irrelevant to describe the regularity of such densities.

REMARK 2. Strictly speaking $\mathcal{H}_\delta^{\beta_0,\beta_\perp}(M, L)$ depends on the choice of the parametrizations $\bar{\Psi}$ and should be written as $\mathcal{H}_\delta^{\beta_0,\beta_\perp}(M, L, \bar{\Psi})$. We show however in [8], Proposition S1.3, that

$$\mathcal{H}_\delta^{\beta_0,\beta_\perp}(M, L, \bar{\Psi}) \subset \mathcal{H}_\delta^{\beta_0,\beta_\perp}(M, C_1 L, \bar{\Psi}') \subset \mathcal{H}_\delta^{\beta_0,\beta_\perp}(M, C_2 L, \bar{\Psi})$$

for some constant $C_1, C_2 > 0$ independent of Ψ and Ψ', as soon as $\bar{\Psi}'$ satisfies (4) also. Hence throughout the paper, we denote this Hölder space $\mathcal{H}_\delta^{\beta_0,\beta_\perp}(M, L)$.

M-anisotropic functions happen to be a convenient way to describe the regularity of a number of densities that are naturally supported around M. To illustrate this, take $f_* : M \to \mathbb{R}$ to be a β_0-Hölder density, meaning that there exists $L_0 : M \to \mathbb{R}$ such that for any $x_0 \in M$,

$$f_* \circ \Psi_{x_0} \in \mathcal{H}_{\mathrm{iso}}^{\beta_0}(\mathcal{V}_{x_0}, L_0 \circ \Psi_{x_0}).$$

Now take $K : \mathbb{R}^D \to \mathbb{R}$ to be a normalized positive smooth isotropic kernel supported on $\mathsf{B}(0,1)$. We introduce $c_\perp^{-1} = \int K(\varepsilon)\,\mathrm{d}\mu_E(\varepsilon)$ where E is any (through isometry) $(D-d)$-dimensional subspaces of $\mathbb{R}^D$. We also assume that $K \in \mathcal{H}_{\mathrm{iso}}^{\beta_\perp}(\mathbb{R}^D, L_\perp)$ for some function $L_\perp$ which is also rotationally invariant.

PROPOSITION 2.3. *Let f be the density of a random variable $Z = X + \delta\mathcal{E}$ where $X \sim f_*(x)\mu_M(\mathrm{d}x)$ and $0 < \delta < \tau$. Then*:

1. (Orthonormal noise) *If $\beta_\perp \le \beta_M - 1$, and if $\mathcal{E}|X \sim c_\perp K(\varepsilon)\mu_{N_X M}(\mathrm{d}\varepsilon)$, then*

$$f \in \mathcal{H}_\delta^{\beta_0,\beta_\perp}(M, L), \quad \textit{with } L(x) := C\delta^{-(D-d)} L_0(\mathrm{pr}_M x) \times L_\perp(x - \mathrm{pr}_M x), \quad C > 0.$$

2. (Isotropic noise) *If $\delta < \tau/32$ and $\beta_0 \le \beta_\perp \le \beta_M - 1$, and if $\mathcal{E} \sim K(\varepsilon)\,\mathrm{d}\varepsilon$, independently of X, then*

$$f \in \mathcal{H}_\delta^{\beta_0,\beta_\perp}(M, L), \quad \textit{with } L(x) := C\delta^{-D} \int_M L_\perp\left(\frac{x-y}{\delta}\right) L_0(y)\mu_M(\mathrm{d}y), \quad C > 0.$$

In both cases C depends on C_M, τ, β_M, β_0, $\beta_\perp$.

See Section A.2 for a proof of this result.

REMARK 3. Throughout the paper, we assume that the true density $f_0 \in \mathcal{H}_\delta^{\beta_0,\beta_\perp}(M, L)$, which implies that $P_0(M^\delta) = 1$. However it is enough to assume that $P_0(M^\delta) \geq 1 - o(1/n)$ where n is the number of observations. This makes no difference in terms of the results presented in Section 3. The weaker assumption $P_0(M^\delta) \geq 1 - o(1/n)$ is for instance fulfilled in the additive noise model (see Proposition 2.3) with $Z = X + \frac{\delta}{\sqrt{C \log n}}\mathcal{E}$ with the Gaussian kernel $K(x) := (2\pi)^{D/2} \exp(-\|x\|^2/2)$.

REMARK 4. Note that under the model $Z = X + \delta\mathcal{E}$ with isotropic noise and K the distribution of the noise belonging to $C^{\beta_\perp}$ with $\|K\|_{C^{\beta_\perp}} \lesssim 1$, then $f \in \mathcal{H}_\delta^{\beta_0,\beta_\perp}(M, L) \cap C^\beta$, with $\|f\|_{C^{\beta_\perp}} \lesssim \delta^{-\beta-D+d}$. Hence for Hölder constant is strongly impacted by δ and D. The geometric framework of M anisotropy, allows to handle this dependence on δ and is therefore useful for such models when δ is small.

In the following section, we describe the family of priors which we use to estimate the above family of densities.

2.3. *Location-scale mixtures of normal priors.* We model the manifold-anisotropic Hölder densities using location-scale mixtures of normals. We parametrize the covariances of the components by $\Sigma = O^T \Lambda O$ where O is a unitary matrix and $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_D)$ is diagonal. Location-scale mixtures can then be written as

$$
\begin{aligned}
f_P(x) &= \int_{\mathbb{R}^D} \varphi_{O^T \Lambda O}(x - \mu)\, \mathrm{d}P(\mu, O, \Lambda), \\
P &= \sum_{k=1}^{K} p_k \delta_{(\mu_k, O_k, \Lambda_k)}, \quad K \in \mathbb{N} \cup \{+\infty\},
\end{aligned}
\tag{7}
$$

where, for any positive definite matrix Σ,

$$
\varphi_\Sigma(z) := \frac{1}{\det^{1/2}(2\pi\Sigma)} \exp\left\{-\frac{1}{2}\|z\|_{\Sigma^{-1}}^2\right\},
$$

is the density of a centered Gaussian with covariance matrix Σ. The two most well-known families of priors on P are Dirichlet process priors and mixtures with random number of components, also known as mixtures of finite mixtures. Recall that if P follows a Dirichlet process priors with parameters A and H where $A > 0$ and H is a probability measure on some measurable space Θ, then

$$
P = \sum_{k=1}^{\infty} p_k \delta_{\theta_k} \quad \text{with } p_k = V_k \prod_{i<k} (1 - V_i), \quad V_i \overset{iid}{\sim} \operatorname{Beta}(1, A) \quad \text{and} \quad \theta_k \overset{iid}{\sim} H.
$$

If P follows a mixture of finite mixtures prior of parameters α_K and π_K where $\alpha_K > 0$ and π_K is a probability measure on $\mathbb{N}$, then

$$
P = \sum_{k=1}^{K} p_k \delta_{\theta_k} \quad \text{with } K \sim \pi_K, \quad (p_1, \dots, p_K) \,|K \sim \mathcal{D}(\alpha_K, \dots, \alpha_K) \quad \text{and} \quad \theta_k \overset{iid}{\sim} H.
$$

In both cases (Dirichlet process and mixture of finite mixtures) we call H the base probability measure. Obviously in the case of mixtures of finite mixtures the conditional prior on $(p_1, \dots, p_K)$ and $\theta_1, \dots, \theta_K$ could be different but we consider this setup for the sake of simplicity.

As shown empirically by [49] location-scale Dirichlet process mixtures with base measure constructed from the conjuguate prior of the Gaussian model are not well adapted to the problem at hand. We show however that if particular care is given to the choice of H, the posterior on manifold-anisotropic density is well behaved. In particular, we consider the two following types of location-scale mixtures:

- *Partial location-scale mixtures*: The eigenvalues Λ of the covariance of the Gaussians are common accross components,

$$(8)\quad f_{\Lambda,P}(x) = \int_{\mathbb{R}^D} \varphi_{O^T\Lambda O}(x-\mu)\,\mathrm{d}P(\mu, O), \qquad P = \sum_{k=1}^{K} p_k \delta_{(\mu_k, O_k)}, \qquad K \in \mathbb{N} \cup \{+\infty\},$$

  where P is a probability distribution on $\mathbb{R}^D \times \mathcal{O}(D)$ (where $\mathcal{O}(D)$ is the set of unitary matrices in $\mathbb{R}^D$) and is either a Dirichlet process prior or a mixture of finite mixtures.
- *Hybrid location-scale mixtures*: The density f_P is written as (7) where P conditionally on a probability Q_2 on $\mathbb{R}_+^D$ follows a Dirichlet process mixture or a mixture of finite mixtures with base measure $H_0(\mathrm{d}\mu, \mathrm{d}O, \mathrm{d}\lambda) = H_1(\mathrm{d}\mu, \mathrm{d}O) \otimes Q_2(\mathrm{d}\lambda)$, and Q_2 follows a distribution $\widetilde{\Pi}_\Lambda$.

We denote by Π the prior on the parameter and we consider the following assumptions on Π. These conditions differs wether Π is assumed to come from a partial location-scale mixture prior or a hybrid location-scale mixture prior.

Conditions on the partial location-scale mixtures. f is modelled as in (8) and P follows either a Dirichlet process with base measure H or a mixture of finite mixtures with base measure H and prior on K satisfying

$$(9)\qquad -\log \Pi_K(K = x) \simeq x(\log x)^r, \qquad r = 0, 1.$$

Here $r = 0$ corresponds for instance to the geometric and $r = 1$ to the Poisson priors on K. The base measure $H(\mathrm{d}\mu, \mathrm{d}O) = h(\mu, O)\,\mathrm{d}\mu\,\mathrm{d}O$ where $\mathrm{d}\mu$ designates the Lebesgue measure on $\mathbb{R}^D$ and $\mathrm{d}O$ the Haar measure on $\mathcal{O}(D)$, and we further assume that there exist $c_1, b_1 > 0$ and $b_2 > 2D - 1$ such that

$$(10)\qquad e^{-c_1\|\mu\|^{b_1}} \lesssim h(\mu, O) \lesssim \big(1 + \|\mu\|\big)^{-b_2} \quad \text{with } \forall \mu, O.$$

We also assume that Λ is drawn from a probability measure Π_Λ that has a density π_Λ with respect to Lebesgue measure on $\mathbb{R}^D$, and that this density satisfies: there exist $c_2, c_3, b_3 > 0$ and $b_4 > D(D-1)/2$ such that

$$(11)\qquad \begin{aligned} & e^{-c_2\sum_{i=1}^{D}\lambda_i^{-d/2}} \lesssim \pi_\Lambda(\lambda_1, \dots, \lambda_D) \quad \text{for small } \lambda_1, \dots, \lambda_D \in \big(\mathbb{R}_+^*\big)^D, \\ & \Pi_\Lambda\Big(\min_{1\le i\le D} \lambda_i < x\Big) \lesssim e^{-c_3 x^{-b_3}} \quad \text{for small } x > 0, \\ \text{and} \quad & \Pi_\Lambda\Big(\max_{1\le i\le D} \lambda_i > x\Big) \lesssim x^{-b_4} \quad \text{for large } x > 0. \end{aligned}$$

Condition (10) is weak and is for instance, satisfied as soon as μ and O are independent under H with positive and continuous density for O and positive density for μ with weak tail assumptions. Condition (11) is also weak and common in the case of location Gaussian mixtures and is verified in particular if the $\sqrt{\lambda_i}$'s are independent inverse Gammas under Π_Λ, or if the λ_i's are independent inverse Gammas and $d \geq 2$.

Conditions on the hybrid location-scale mixtures. H_1 satisfies (10) and Q_2 is random with distribution $\tilde{\Pi}_\lambda$ which satisfies: for all $b > 0$ there exists $B_0, c_2 > 0$ such that for $x_1^2 \leq x_2$

TABLE 1
Summary table of the required conditions depending on the type of mixture and the type of scale sampling

	MFM	DPM	Partial	Hybrid
Conditions	(9) + (10)	(10)	(11)	(12) + (13)

both small,

$$\tilde{\Pi}_\Lambda\big[Q_2([x_1, x_1(1+x_1^b)]^d \times [x_2, x_2(1+x_1^b)]^{D-d}) \geq x_1^{B_0}\big] \gtrsim e^{-c_2 x_1^{-d/2}}. \tag{12}$$

Moreover, we assume that for some positive constant $c_3, b_3, c_4, b_4 > 0$ such that

$$\begin{aligned} &\mathbb{E}_{\tilde{\Pi}_\Lambda}\Big[Q_2\Big(\min_{1\leq i\leq D} \lambda_i \leq x\Big)\Big] \lesssim e^{-c_3 x^{-b_3}} \quad \text{for small } x, \\ &\mathbb{E}_{\tilde{\Pi}_\Lambda}\Big[Q_2\Big(\max_{1\leq i\leq D} \lambda_i > x\Big)\Big] \lesssim e^{-c_4 x^{b_4}} \quad \text{for large } x. \end{aligned} \tag{13}$$

REMARK 5. One can view the partial location-scale mixture as a special instance of the hybrid location-scale mixture defined above: take Q_2 to be the Dirac mass at a value Λ where $\Lambda \sim \Pi_\Lambda$.

Conditions (12) and (13) are in particular satisfied if Q_2 comes from a Dirichlet process. More precisely, if Q_2 is of the form

$$Q_2(\mathrm{d}\lambda) = \prod_{i=1}^{D} Q_0(\mathrm{d}\lambda_i) \quad \text{with } Q_0 \sim \mathrm{DP}(BH_\lambda),$$

with $B > 0$, then (12) and (13) are satisfied for reasonable choices of probability distributions H_λ on $\mathbb{R}_+$. We show in the next proposition that this is in particular true when H_λ is a square-root- inverse Gamma.

PROPOSITION 2.4. *Assume that $Q_2 = Q_0^{\otimes D}$ where $Q_0 \sim \mathrm{DP}(BH_\lambda)$, where $B > 0$ and H_λ has density h_λ verifying*

$$e^{-c_2\lambda^{-1/4}} \mathbb{1}_{\lambda\leq 1} \lesssim h_\lambda(\lambda) \lesssim e^{-c_3\lambda^{-b_3}} e^{-c_4\lambda^{b_4}},$$

then conditions (12) and (13) are satisfied.

A proof of Proposition 2.4 can be found in [8], Section S2.3. For instance, if $\lambda^{1/4}$ follows an inverse Gamma truncated on [0, R] for arbitrarily large R under H_λ, then Proposition 2.4 holds.

REMARK 6. Although the conditions on the prior, (11) and (12), depend on d, they are satisfies for all d by setting $d = 1$ and in particular they are verified for all $d \geq 1$ if $\lambda_i^{1/4}$ follow an inverse Gamma under the base measure, which is agnostic to d.

3. Main results.

3.1. *Posterior contraction rates*. Recall that $X_1, \dots, X_n$ is an n sample drawn from a distribution P_0 with density f_0. Recall that $f_0 = f_{0,n}$ might depend on n, but that all the constants appearing in the subsequent conditions do not. This density is concentrated around a submanifold M, with a a given smoothness β_0 along the manifold and a typically much larger smoothness $\beta_\perp$ along the normal spaces. More precisely, we will assume:

- *Conditions on* M: the submanifold M is of dimension d and has a reach greater than $\tau > 0$. Furthermore, there exists $\beta_M > 4$ and $C_M > 0$ such that $\Psi_{x_0} \in \mathcal{H}^{\beta_M}_{\mathrm{iso}}(\mathcal{V}_{x_0}, C_M)$, $\forall x_0 \in M$. In particular, M also satisfies (4).
- *Conditions on* f_0: the density f_0 is in $\mathcal{H}^{\beta_0,\beta_\perp}_\delta(M, L)$ with $\delta = \delta_n \le \tau/2$. Furthermore, there exists $c_5, \kappa > 0$,

$$f_0(x) \lesssim e^{-c_5\|x\|^\kappa} \quad \forall x \in \mathbb{R}^D, \tag{14}$$

and for some $\omega > 6\beta$ and $C_0 < \infty$,

$$\int_{\mathcal{W}_{x_0,\delta}} \left|\frac{\mathrm{D}^k \bar{f}_{x_0,\delta}}{\bar{f}_{x_0,\delta}}\right|^{\omega/\langle k,\boldsymbol{\alpha}\rangle} \bar{f}_{\delta,x_0} \le C_0 \quad \text{and} \quad \int_{\mathcal{W}_{x_0,\delta}} \left|\frac{L_{x_0,\delta}}{\bar{f}_{x_0,\delta}}\right|^{\omega/\beta} \bar{f}_{\delta,x_0} \le C_0 \tag{15}$$

for all δ small, $x_0 \in M$ and all $0 \le \langle k, \boldsymbol{\alpha}\rangle < \beta$. Recall that the quantities $\boldsymbol{\alpha}$ and β are defined through β_0 and $\beta_\perp$ through (5).

In the rest of this paper the symbols $\simeq$, $\lesssim$ and $\gtrsim$ denote equalities or inequalities up to a constant depending on D, d, τ, β_M, C_M, β_0, $\beta_\perp$ and all the other constants appearing in conditions (9) to (15).

THEOREM 3.1. *Let $X_1, \dots, X_n$ be a n-sample from f_0 satisfying* (14) *and* (15) *with $\omega > 6\beta + (2\beta + D)(D - d)\log(1/\delta)/\log n$ and that $\beta_0 \le \beta_\perp \le \beta_M - 4$. Consider the partial location scale mixture or the hybrid location scale mixture prior* [*either the mixture of finite mixtures or the Dirichlet process mixture*] *satisfying the conditions displayed in Table* 1.

Then, under the conditions stated above,

$$\mathbb{E}_0^n\big[\Pi\big(\mathrm{d_H}(f_P, f_0) \ge \varepsilon_n \mid \mathcal{X}_n\big)\big] = o(1),$$

where

$$\varepsilon_n \simeq \log^p n \times \left\{\frac{1}{\sqrt{n\delta^{\frac{D}{2\alpha_0-\alpha_\perp}}}} \vee n^{-\frac{\beta}{2\beta+D}}\right\}, \tag{16}$$

with $p > 0$ depending on D, κ and β, and where we recall that β, α_0 and $\alpha_\perp$ are defined through (5).

A few remarks are in order.

REMARK 7. Note that the conditions regarding the prior Π do not involve M, δ, L, β, or τ: they are regarded as unknown in this framework. Also, while the conditions (11) or (12) involve the intrinsic dimension d, one can always choose Π independent of d and satisfying these two conditions for all $d \ge 1$, as noted in Remark 6. Hence, the Bayesian procedure is fully adaptive.

REMARK 8. The rates obtained in Theorem 3.1 are minimax (up to a $\log n$ term) as soon as $\delta \gtrsim n^{-(2\alpha_0-\alpha_\perp)/(2\beta+D)} := \delta_n(\beta_0, \beta_\perp)$. To see this, note that the case of densities concentrating near the linear subspace spanned by the d first vectors of the canonical basis of $\mathbb{R}^D$, described as

$$\mathcal{F}_{\mathrm{lin}}(\delta) := \big\{f_\delta(x) = f(x_1, x_2/\delta)/\delta^{D-d},\ f \in \mathcal{H}^{\boldsymbol{\beta}}_{\mathrm{an}}(\mathcal{U}, \mathbb{R})\big\},$$

where $x_1 \in \mathbb{R}^d$, $x_2 \in \mathbb{R}^{D-d}$ are restricted to a compact subset $\mathcal{U}$ of $\mathbb{R}^D$, is a special case of the assumption (14). Even if $\delta = \delta_n = o(1)$, the minimax rate for estimating, in L_1 norm, densities belonging to $\mathcal{F}_{\text{lin}}(\delta)$ is bounded from below by $n^{-\beta/(2\beta+D)}$. This is a consequence of fact that for two densities $f_{\delta,1}$ and $f_{\delta,2}$, $\|f_{\delta,1} - f_{\delta,2}\|_1 = \|f_{1,1} - f_{1,2}\|_1$ is independent of δ and a look at the construction of the lower bound in [32] — proof of Theorem 4 (ii) for $p = 1$, $r = (\infty, \dots, \infty)$, $s = \infty$ and $\theta = 1$ (tail dominance from (14)) — shows that the lower bound is not impacted by δ. Obviously the result would be different if other metrics, such as the L_2 norm were considered. However these distances, being non transformation invariant, are less natural in the context of density estimation. Hence, at least when $\delta \gtrsim \delta_n(\beta_0, \beta_\perp)$ the rate ε_n is the minimax estimation rate (up to a $\log n$ term).

REMARK 9 (Relations to existing results). The result of Theorem 3.1 is in line with the many existing results regarding Bayesian multivariate density estimation, as in [12, 28, 29, 41, 60], and extends them in this new *near manifold* framework. We illustrate these extensions with simple cases and examples below.

- *Linear case*: As explained above, the case where M is a linear subspace spanned by vectors of the canonical basis of $\mathbb{R}^D$ and where δ is constant (not going to zero) recovers the classical Euclidean anisotropic setting and our result thus encompasses in particular those of [60] regarding the Bayesian estimation of multivariate, anisotropic Hölder densities. The rates we obtain are the same and are minimax up to logarithmic factors. The technical assumptions we do are roughly identical to theirs — although in our context, the regularity vectors can only take two values β_0 and $\beta_\perp$ because of the geometry of the problem. The linear anisotropic case with regularity taking more than 2 values can be treated similarly to what we have done in Theorem 3.4, thus extending the results of [60] to linear anisotropy along different axes than the canonical basis.
- *Isotropic case* When $\beta_0 = \beta_\perp = \beta$, the density f_0 is globally β-Hölder on $\mathbb{R}^D$, and the general results of [60] would also apply in that case. However, because f_0 has magnitude $\delta^{-(D-d)}$ in a tubular neighborhood of size δ, its β-Hölder norm is of order $\delta^{-(D-d+\beta)}$. This yields a rate of order
$$\frac{1}{\delta^{D-d+\beta}} n^{-\frac{\beta}{2\beta+D}}.$$
The rate we find in Theorem 3.1 is:
$$\frac{1}{\sqrt{n\delta^D}} \vee n^{-\frac{\beta}{2\beta+D}}$$
which is always better if the density is sufficiently smooth, that is, $\beta \geq (d - D/2)_+$, or for δ not too small otherwise, that is, $\delta \gg n^{-D/((2\beta+D)(2d-D-2\beta))}$. That is because our procedure takes advantage of the geometry of the problem in this context.
- *Fixed width case*: The case where $\delta > 0$ is fixed and do not go to zero has been investigated by [49] using Fisher–Gaussian mixtures. However, little theoretical results or rates are derived as they were not interested in providing a general model for such densities. This case encompasses many natural situations such as for instance, the ones considered in [13] where the data can be naturally described using spherical coordinates with an unknown center point. *Diverging width case*: The case where $\delta \to \infty$ implies in particular that the reach of M becomes infinite, which can only be the case if M is an affine subspace of $\mathbb{R}^D$, as the reach lower-bounds the minimal radius of curvature of M. We thus fall back in the Linear case described at the beginning of this remark. In this case, there is no need to rescale with δ to define the smoothness class and we can consider the alternative smoothness class $\mathcal{H}^{\beta_0,\beta_\perp}(M, L)$ of function $f : \mathbb{R}^D \to \mathbb{R}$ verifying $f \circ \bar{\Psi} \in \mathcal{H}_{\text{an}}^{\boldsymbol{\beta}_{0,\perp}}(M, L)$ where $\bar{\Psi}$ is the rotation aligning with $M, M^\perp$. Notice however that our functional classes, as defined in Definition 2.2, becomes trivial for $\delta = \infty$, and one would need to threshold

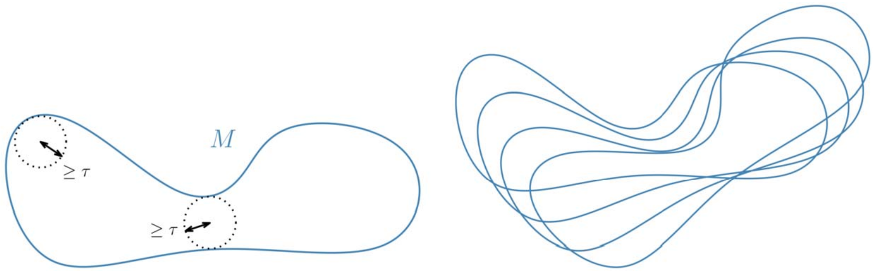


FIG. 3. *(Left) An example of smooth submanifolds with a reach constrained to be greater than τ and (Right) a finite union of such manifolds. Both subsets are admissible as (near) supports for a density f_0 satisfying the contraction rates displayed Proposition 3.2.*

at $\delta = \log^t n$ to use this definition, where t is driven by the exponential tail of the density, thus losing logarithmic terms in the final contraction rate. It is therefore more interesting in this case to use the alternative class $\mathcal{H}^{\beta_0,\beta_\perp}(M, L)$.

- *Decaying width case*: Finally, the case where $\delta = \delta_n \to 0$ is, to our knowledge, new and the rate we obtain, in its dependence to δ_n, does not relate to any rate that we know of.

REMARK 10. Because the approximation results of Section 3.2 are stable under finite mixtures, so do the results of Theorem 3.1. In particular, the support of f_0 can be a finite union of vicinity of submanifolds M_i with arbitrary intersections and with each M_i fulfilling (3); see Figure 3 for a diagram of such a situation. Hence, the assumption on the lower bound on the reach can be significantly weakened, as shown in the following proposition.

PROPOSITION 3.2. *Assume that there exist M_j, $j \leq J$ smooth manifolds with dimensions d_j, $j \leq J$ and verifying the conditions on M; assume that there exist probability densities $f_{0,j}$ $j \leq J$ verifying assumptions (14) and (15) with respect to $(\beta_{0j}, \beta_{\perp j})$ and $p_{0j} > 0$, $j \leq J$ such that $\sum_j p_{0j} = 1$. Then under the hybrid location scale mixture prior satisfying the conditions displayed in Table 1*,

$$\mathbb{E}_0^n\big[\Pi\big(\mathrm{d_H}(f_P, f_0) \geq \varepsilon_n \mid \mathcal{X}_n\big)\big] = o(1), \quad \textit{where}$$

$$\varepsilon_n \simeq \log^p n \times \max_j \left\{ \frac{1}{\sqrt{n\delta_j^{\frac{D}{2\alpha_{0j}-\alpha_{\perp j}}}}} \vee n^{-\frac{\beta_j}{2\beta_j+D}} \right\},$$

If in addition for all j, $(d_j, \beta_{0j}, \beta_{\perp j}, \delta_j) = (d, \beta_0, \beta_\perp, \delta)$ then the above result holds also true under the partial location scale mixture prior.

Proposition 3.2 is a direct consequence of the proof of Theorem 5.1 and of the fact that with $\mathrm{KL}(f_0, f) = P_0(\log(f_0/f))$ and $f = \sum_{j=1}^J p_{0j} f_j$,

$$KL(f_0, f) \leq \sum_{j=1}^J p_{0j} KL(f_{0j}, f_j);$$

see Section 5 for more details.

The case of collections of manifolds has also been considered in [38], who bound the variance term of a Kernel estimator, in sup norm, under a condition on the support which include unions of manifolds.

We consider such an example in the simulations of Section 4. Interestingly, to allow for different β_{0i} or different d_i across $i = 1, \dots, N$, we need to consider the hybrid location—scale prior, since we need to allow for different $\Delta_{\sigma_i,\delta_i}$, for $i = 1, \dots, N$ in the approximation theory presented in Theorem 3.4. We stress here that the accuracy of the density estimation does not decrease near the intersections. It is possible that locally better rates can be achieved (depending on β_j, d_j, δ_j in particular, however, deriving local rates is much harder than global Hellinger or L_1 rates (see [56]) and is beyond the scope of this paper.

REMARK 11. The case where $\beta_\perp$ is infinite is particularly of interest. Then, $\beta \to D\beta_0/d$, $\alpha_0 \to D/d$, $\alpha_\perp \to 0$ and the rate ε_n becomes, up to $\log n$ terms

$$\frac{1}{\sqrt{n\delta^{d/2}}} \vee n^{-\frac{\beta_0}{2\beta_0+d}},$$

which is, when δ is not too small (i.e., $\delta \gtrsim n^{-2/(2\beta_0+d)}$), the minimax rate for estimating a β_0 Hölder density in $\mathbb{R}^d$. Here, again, the strength of our result lies in that the manifold (and thus the support of f_0) is unknown and the prior depends neither on $\boldsymbol{\beta}$ nor on δ or d (or M).

Assumptions (14) and (15) are common assumptions in density estimation based on mixtures of Gaussians; see, for instance, [41] or [60] for the Bayesian approaches and [46] for the frequentist approaches. They are rather weak asssumptions. The difficulty with (15) is that it is expressed on $\bar{f}_{x_0,\delta}$, which is natural in our context since the smoothness assumption on f_0 is expressed in terms of $\bar{f}_{x_0,\delta}$, but is not so intuitive. However, a careful examination of (15) shows that this assumption is for instance implied by the stronger, but chart-independent assumption:

$$\int_{\mathbb{R}^D} \left\{ \frac{L(x)}{f_0(x)} \right\}^{\omega_*} f_0(x)\,\mathrm{d}x \le C_0,$$

for some large ω^*. Moreover, to understand better what (15) means in terms of f_0, we illustrate it in our archetypal model where f_0 is the density of $X = Y + \delta\mathcal{E}$ as in Proposition 2.3. We then have the following result.

LEMMA 3.3. *Under the conditions of Proposition* 2.3 *and if*

$$\int_M \left\{ \frac{L_0(x)}{f_*(x)} \right\}^{\omega_*} f_*(x)\,\mathrm{d}\mu_M(x) < \infty \quad \textit{and} \quad \int_{\mathsf{B}(0,1)} \left\{ \frac{L_\perp(\eta)}{K(\eta)} \right\}^{\omega_*} K(\eta)\,\mathrm{d}\eta < \infty,$$

for some ω_* large. Then (15) *is satisfied in both the orthonormal noise model and in the isotropic noise model.*

The proof of Lemma 3.3 can be found in [8], Section S2.2. Note that the conditions in Lemma 3.3 are fulfilled by a number of natural kernels or densities such as $K(\eta) \approx (1 - \|\eta\|^2)_+^p$ and $L_\perp \approx (1 - \|\eta\|^2)_+^{p-\beta_\perp}$ with large p or $K(\eta) \approx \exp(-(1 - \|x\|^2)_+^{-1})$ and $L_\perp(\eta) \approx (1 - \|\eta\|^2)^{-\beta_M} K(\eta)$; see [8], Lemma S1.4, for further details.

3.2. *Approximating M-anisotropic densities*. Theorem 3.1 is proved using the approach of [28], with a control on the prior mass of Kullback-Leibler neighbourhoods of f_0 and on the entropy of the support of the prior. The main difficulty in our setup is in proving the Kullback-Leibler prior mass condition. To do so, we need to construct an *efficient* approximation of f_0 by mixtures of Gaussian. This construction is of interest in its own as it sheds light on the behaviour of such mixtures and on the geometry of M-anisotropic densities.

To explain the construction, we denote, for any $x \in M^\tau$, $T_x = T_{\mathrm{pr}_M(x)}M$ and $N_x = N_{\mathrm{pr}_M(x)}M$. We set $\mathcal{B}_x^0$ an orthonormal basis of T_x, $\mathcal{B}_x^\perp$ an orthonormal basis of N_x, and let O_x

be the matrix of the linear map $z \mapsto (\mathrm{pr}_{T_x} z, \mathrm{pr}_{N_x} z)$ from the canonical basis to the concatenated basis $(\mathcal{B}_x^0, \mathcal{B}_x^\perp)$. We finally write $\Sigma(x) = O_x^\top \Delta_{\sigma,\delta}^2 O_x$ where

$$\Delta_{\sigma,\delta} = \begin{pmatrix} \sigma^{\alpha_0} \mathrm{Id}_d & 0 \\ 0 & \delta\sigma^{\alpha_\perp} \mathrm{Id}_{D-d} \end{pmatrix}.$$

Note that $\Sigma(x)$ does not depends on the choice of the bases $\mathcal{B}_x^0$ and $\mathcal{B}_x^\perp$ since for any orthonormal basis change P that preserves T_x (whence N_x), one has $P^\top \Delta_{\sigma,\delta} P = \Delta_{\sigma,\delta}$. For any function $f : \mathbb{R}^D \to \mathbb{R}$, we define

$$K_\Sigma f(x) := \int_{M^\tau} \varphi_{\Sigma(y)}(x-y) f(y)\,\mathrm{d}y,$$

where we recall that $\varphi_{\Sigma(y)}$ is the density of a centered Gaussian with variance $\Sigma(y)$. The idea behind the construction of the approximation of f_0 by a mixture of Gaussians is to show that $K_\Sigma f_0(x)$ is close to f_0 and then to define a perturbation f_1 of f_0 such that $K_\Sigma f_1(x) - f_0(x) = O(L(x)\sigma^\beta)$. Compared to the construction proposed by [41] in the univariate case or [60] in the multivariate case where $K_\Sigma f = \varphi_\Sigma * f$, in our construction Σ varies with the location y. Note that in particular $\int \varphi_{\Sigma(y)}(x-y)\,\mathrm{d}y$ may be different from 1. This dependence in y is crucial to adapt to the geometry of the manifold but considerably complicates the proof as the underlying kernel integral operator can no longer be written as a convolution. We first show that f_0 can be efficiently approximated pointwise.

THEOREM 3.4. *Assume that $\sigma, \delta \le 1$, that $f_0 \in \mathcal{H}_\delta^{\beta_0,\beta_\perp}(M, L)$ satisfies (14) and (15), that $\sigma^{2\alpha_0 - \alpha_\perp} = o(\delta)$ and that the manifold satisfies (3) with $\beta_0 \le \beta_\perp \le \beta_M - 4$. Then there exists a function $g : \mathbb{R}^D \to \mathbb{R}$ such that, for any $H > 0$,*

$$\bigl|K_\Sigma g(x) - f_0(x)\bigr| \lesssim \sigma^\beta L(x) \mathbb{1}_{M^\tau} + \bigl(H \log(1/\sigma)\bigr)^{D/\kappa} \sigma^H \|L\|_\infty \quad \forall x \in \mathbb{R}^D.$$

The function g has the form

$$g(x) = d_0(x, \sigma, \delta) f_0(x) + \frac{1}{\delta^{D-d}} \sum_{0 < \langle k, \boldsymbol{\alpha}\rangle < \beta} \sigma^{\langle k, \boldsymbol{\alpha}\rangle} \sum_{j=1}^J d_{j,k}(x, \sigma, \delta)\, \mathrm{D}_z^k \overline{(\chi_j f_0)}_{x_j,\delta}(z_{j,x}),$$

where $z_{j,x} := \Delta_{1,\delta}^{-1} \bar{\Psi}_{x_j}^{-1}(x)$, where $(\chi_j)_{j \le J}$ is a partition of unity, defined in [8], Section S3.1, of the set $M^\tau \cap \mathsf{B}(0, R_0(\log(1/\sigma))^{1/\kappa})$ associated with a $\tau/64$-packing $(x_j)_{j \le J}$ of $M^\tau \cap \mathsf{B}(0, R_0(\log(1/\sigma))^{1/\kappa})$ and where $d_{j,k}(x, \sigma, \delta)$ are smooth and bounded functions depending on χ_j and M.

We then establish that the previous bound translates to a control in terms of Hellinger distance.

COROLLARY 3.5. *In the context of Theorem 3.4, if f_0 also satisfies (15) and if $\sigma^{\omega - 6\beta} = o(\delta^{D-d})$ and $\delta\sigma^{\alpha_\perp} = o(|\log \sigma|^{-1/2})$, the probability density $\tilde{h} \propto g \mathbb{1}_{g \ge f_0/2} + f_0/2 \mathbb{1}_{g < f_0/2}$ verifies*

$$\mathrm{d_H}(K_\Sigma \tilde{h}, g)^2 \lesssim \sigma^{2\beta} |\log \sigma|^{16\beta + D/\kappa}. \tag{17}$$

Theorem 3.4 is proven in Section 5.2 while the proof of Corollary 3.5, which appears to be a nontrivial consequence of Theorem 3.4, is delayed to [8], Section S3.1.

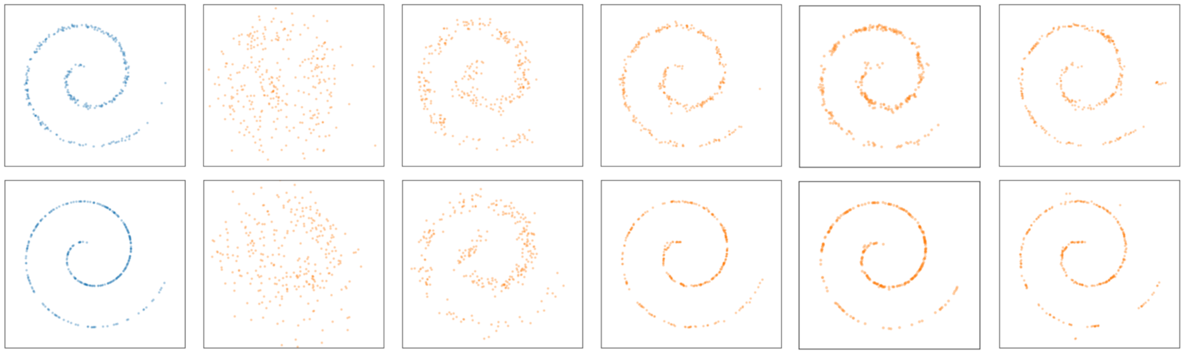


FIG. 4. (*First column*) *We observe* $n = 300$ *points drawn under* P_0 *for the* 2*D-spiral with* $\delta = 0.1$ (*Top row*) *and* $\delta = 0.01$ (*Bottom row*). *We then sample the same amount of points using the Gibbs sampler with* (*Second column*) *Inverse-Wishart prior*, (*Third column*) *model* (18) *with* $b_j = 1$, (*Forth column*) *model* (18) *with* $b_j = 0.001$ *and* (*Fifth column*) *model* (18) *with the exponential hyperprior on* b_j. *Finally*, *we used in* (*Sixth column*) *the approximate MAP distribution*.

4. Numerical experiments. In this section we present some implementations and visual results of the location-scale mixtures of Section 2.3. The aim of this section is to illustrate the ability of the aforementioned priors to describe in a relevant way data that can be very singular. More precisely, we consider the hierarchical model:

$$(y_i)_{i=1}^n \mid (\mu_i, O_i)_{i=1}^n, \Lambda \sim \bigotimes_{i=1}^n \mathcal{N}(\cdot|\mu_i, O_i \Lambda O_i^T) \quad \text{with } \Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_D),$$

$$(\mu_i, O_i)_{i=1}^n \mid P \sim P^{\otimes n}, \quad P \sim \mathrm{DP}(\alpha P_0),\ P_0 = \mathcal{N}(\cdot|\mu_0, \Sigma_0) \otimes \mathrm{Unif}(\mathcal{O}(D)) \tag{18}$$

$$(\lambda_j)_{j=1}^D | (b_j)_{j=1}^D \sim \bigotimes_{j=1}^D \mathrm{Inv}\Gamma(a_j, b_j), \quad (b_j)_{j=1}^D \sim \bigotimes_{j=1}^D \mathrm{Exp}(\kappa_j).$$

We set in our experiments the value of the hyperparameters as $\alpha = 1$, $a_j = \kappa_j = 1$ for all $1 \leq j \leq D$ and $\mu_0 = 0$, $\Sigma_0 = I_D$. Alternatively, we also consider the previous model with the b_j's being fixed ahead.

We present two different inference approaches: Gibbs sampling (using [51], Algo 8, and example 5.1 in [30]) and stochastic approximation of the Maximum A Posteriori estimator (MAP) using the `python` package *pyro* [9]. The former is an asymptotically exact MCMC sampler while the latter is an approximate (but way easier to compute) point estimator. We refer the reader to the [8], Section S5, for further details of the implementation.

The manifolds we use for the experiments are a 2D spiral, two intersecting circles, a 3D spiral and a torus; see Figures 4–7. Precise parametrizations are given in [8], Section S5.1. As for the data generating distributions, we use the additive isotropic noise model (as described in Proposition 2.3), with noise kernel $K_{\beta_\perp}(x) \propto (1 - \|x\|^2)_+^{\beta_\perp}$, and base densities chosen as

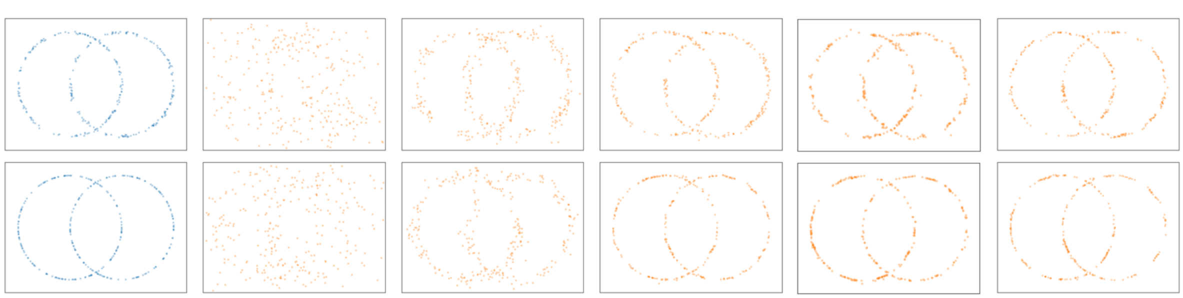


FIG. 5. *Same experiment as in Figure* 4 *but for the two overlapping circles*.

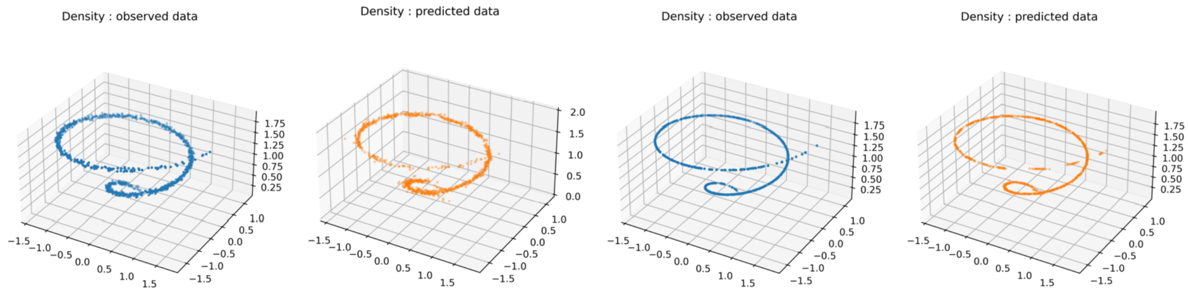


FIG. 6. *We observe $n = 1000$ points* (*blue plots*) *and sample the same amount through the approximate MAP distribution* (*orange plots*). *The width δ is set to* (*Left two plots*) 0.1 *and* (*Right two plots*) *to* 0.01.

follows: we take the uniform distribution for the two-circles and for the torus, whereas for the 2D and 3D spiral, we push forward through the embeddings (described in [8], Section S5.1) the probability distribution g_{β_0} on [0, 1] given by $g_{\beta_0}(t) \propto (1 - (1 - 2t)^{\beta_0})\mathbb{1}_{t\in[0,1/2]} + (1 - (2t - 1)^{\beta_0+1})\mathbb{1}_{t\in[1/2,1]}$. In the experiments, the regularities were set to $\beta_0 = 2$ and $\beta_\perp = 6$.

We test different values for δ and challenge the model (18) for various choices of b_j's against the conjugate Normal Inverse-Wishart prior on these datasets (for which we choose to use the `R` package *dirichletprocess* [45]). Since the Gibbs sampling algorithm is computationally very intensive, we compare this algorithm with the much lighter approximate MAP on the 2-dimensional datasets, with a sample of size $n = 300$; see Figures 4 and 5.

As seen in these figures, the hierarchical prior (18) seem to perform much better than the prior with fixed b_j (unless these are chosen very carefully) and much better than the Normal Inverse-Wishart, which is also consistent with the empirical results of [49]. Our theoretical results together with our empirical results strongly suggest that this conjuguate DP mixture is suboptimal is this case.

Thanks to the low computational cost of approximating the MAP, we were able to conduct experiments on 3D datasets, the 3D-spiral and the torus. See Figures 6 and 7.

5. Proofs of the main results.

5.1. *Proof of Theorem* 3.1. Theorem 3.1 is proved using [28], Theorem 5, which relies on two things: making sure that the prior probability distribution puts enough mass around the true density f_0 (Kullback-Leibler condition), and ensuring that most of its probability mass is concentrated on a subset of manageable entropy (entropy condition). Notice that [28], Theorem 5, is expressed for a fixed true distribution P_0, but a careful look at the statement and at the proof of the theorem shows that this results extends straightforwardly to a true distribution $P_{0,n}$ depending on n, as their result is nonasymptotic in nature.

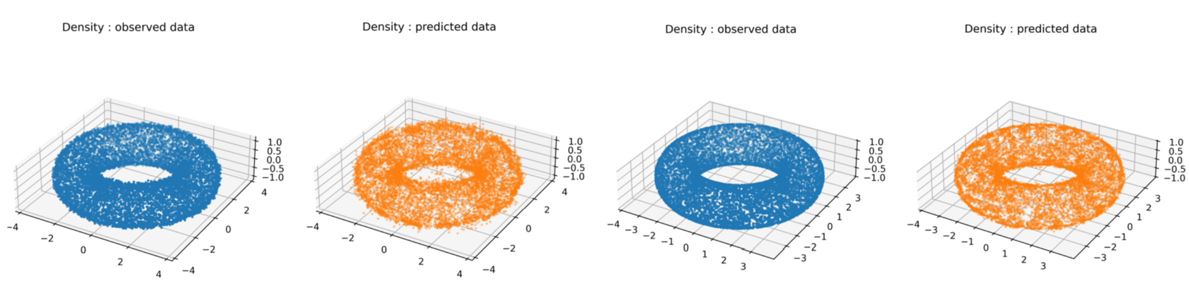


FIG. 7. *Same experiment as in Figure 6 for the torus with $n = 10000$ points observed and sampled and with respectively*, (*Left two plots*) $\delta = 0.5$ *and* (*Right two plots*) $\delta = 0.05$.

PROOF OF THEOREM 3.1. The most challenging part of the proof is the Kullback-Leibler condition. To verify it we define, for any $\varepsilon > 0$,

$$\mathsf{B}(f_0, \varepsilon) = \left\{ f : \mathbb{R}^D \to \mathbb{R} \mid \mathbb{P}_0\left(\log \frac{f_0}{f}\right) \le \varepsilon^2 \text{ and } \mathbb{P}_0\left(\log^2 \frac{f_0}{f}\right) \le \varepsilon^2 \right\}, \tag{19}$$

and introduce

$$\tilde{\varepsilon}_n := \delta^{\frac{\beta}{2\alpha_0 - \alpha_\perp}} \wedge n^{-\frac{\beta}{2\beta + D}} \quad \text{and} \quad \varepsilon_n := \left\{ \frac{C^{1/2}}{\sqrt{n}} \tilde{\varepsilon}_n^{-D/2\beta} \log^{t/2}(1/\tilde{\varepsilon}_n) \right\} \vee \{ \tilde{\varepsilon}_n \log^s(1/\tilde{\varepsilon}_n) \},$$

where s, t and C are introduced in Lemma 5.2. The sequence $\tilde{\varepsilon}_n$ goes to 0 and is such that $\tilde{\varepsilon}_n^{2\alpha_0} \le \delta^\beta \tilde{\varepsilon}_n^{\alpha_\perp}$ so that $\mathsf{B}(f_0, \tilde{\varepsilon}_n \log^s(1/\tilde{\varepsilon}_n)) \subset \mathsf{B}(f_0, \varepsilon_n)$. The control of $\Pi(f_P \in \mathsf{B}(f_0, \varepsilon_n))$ relies first on Theorem 3.4 which constructs explicitly an approximation of f_0 by a continuous location scale mixture of Gaussians $K_\Sigma f_1$ with f_1 close to f_0 and then on a discrete approximation of $K_\Sigma f_1$; it is formally stated in Lemma 5.2 below and implies that $\Pi(f_P \in \mathsf{B}(f_0, \tilde{\varepsilon}_n \log^s(1/\tilde{\varepsilon}_n)) \gtrsim \exp(-C\tilde{\varepsilon}_n^{-D/\beta} \log^t(1/\tilde{\varepsilon}_n)) \gtrsim \exp(-n\varepsilon_n^2)$..

To verify the entropy condition of [28], Theorem 5, we define, for any sequence ε_n going to 0, $\mathcal{F}_n = \mathcal{F}_n(\varepsilon_n, R_0, H_0, \sigma_0, \sigma_1)$ to be the set of all probability density function f_P with $P = \sum_{h \ge 1} \pi_h \delta_{\mu_h, U_h, \Lambda_h}$ such that

$$\sum_{h > H_n} \pi_h \le \varepsilon_n,$$

$$\forall h \le H_n, \mu_h \in B(0, R_n) \quad \text{and} \quad \forall h \le H_n, \Lambda_h \in \mathcal{Q}_n = [\underline{\sigma}_n^2, \bar{\sigma}_n^2]^D, U_h \in \mathcal{O}(D),$$

where $R_n = \exp(R_0 n\varepsilon_n^2)$, $H_n = \lfloor H_0(n\varepsilon_n^2)/\log n \rfloor$, $\underline{\sigma}_n^2 = \sigma_0^2(n\varepsilon_n^2)^{-1/b_3}$ for some positive constant R_0, H_0 and σ_0 and where, for some $\sigma_1 > 0$:

- $\bar{\sigma}_n^2 = \exp(\sigma_1^2 n\varepsilon_n^2)$ in the case of the Partial location-scale mixture
- $\bar{\sigma}_n^2 = \sigma_1^2(n\varepsilon_n^2)^{1/b_4}$ in the case of the Hybrid location-scale mixture.

We then show in Lemma 5.1 below that $\Pi(\mathcal{F}_n^c) \lesssim \exp(-c_1 n\varepsilon_n^2)$ as soon as $n\varepsilon_n^2 \ge n^\omega$ for some $\omega > 0$.

We now define, for $\mathbf{j} = (j_h, h \le H_n) \in \mathbb{N}^{H_n}$,

$$\mathcal{F}_{n,\mathbf{j}} := \{ f_P \in \mathcal{F}_n \mid \forall h \le h_n,\ j_h\sqrt{n} < \|\mu_h\| \le (j_h + 1)\sqrt{n} \}$$

along with the following refinement:

$$\mathcal{F}_{n,\mathbf{j},0} := \left\{ f_P \in \mathcal{F}_{n,\mathbf{j}} \mid \frac{\max_i \lambda_i}{\min_i \lambda_i} \le n \right\}$$

$$\text{and} \quad \forall \ell \ge 1,\ \mathcal{F}_{n,\mathbf{j},\ell} := \left\{ f_P \in \mathcal{F}_{n,\mathbf{j}} \mid n^{2^{\ell-1}} < \frac{\max_i \lambda_i}{\min_i \lambda_i} \le n^{2^\ell} \right\}.$$

In Lemma 5.1 below, we show that for the partial location scale prior

$$\sum_{\mathbf{j},\ell} \sqrt{\Pi(\mathcal{F}_{n,\mathbf{j},\ell}) N(\varepsilon_n, \mathcal{F}_{n,\mathbf{j},\ell}, \|\cdot\|_1)} e^{-M_0 n\varepsilon_n^2} = o(1);$$

while for the hybrid one the same holds true with $\mathcal{F}_{n,\mathbf{j}}$.

Hence [28], Theorem 5, implies that the posterior contracts at rate ε_n, so long as $n\varepsilon_n^2 \ge n^{t_0}$ for some $t_0 > 0$. We now distinguish three cases. Denoting $p = s \vee t/2$, there holds:

1. If first $\delta^{\frac{D}{2\alpha_0 - \alpha_\perp}} \le n^{-1}$, then the results of Theorem 3.1 is trivial and there is nothing to show (because the contraction rate goes to ∞ instead of 0);

2. If $\delta^{\frac{D}{2\alpha_0-\alpha_\perp}} \geq n^{-\frac{D}{2\beta+D}}$, then $\tilde{\varepsilon}_n = n^{-\frac{\beta}{2\beta+D}}$ and one easily gets that $\varepsilon_n \simeq n^{-\frac{\beta}{2\beta+D}} \log^p n$. In particular, $n\varepsilon_n^2 \gg n^{t_0}$ for $t_0 = D/(2\beta + D)$.

3. If finally $n^{-1} < \delta^{\frac{D}{2\alpha_0-\alpha_\perp}} < n^{-\frac{D}{2\beta+D}}$, then $\tilde{\varepsilon}_n = \delta^{\frac{\beta}{2\alpha_0-\alpha_\perp}}$ and $\log(1/\delta) \gtrsim \log n$ so that

$$n\varepsilon_n^2 \geq C\tilde{\varepsilon}_n^{-D/\beta} \log^t(1/\tilde{\varepsilon}_n) \gtrsim \delta^{-\frac{D}{2\alpha_0-\alpha_\perp}} \log^t(n) \gg n^{\frac{D}{2\beta+D}}.$$

Finally to understand how ε_n depends on δ in this last case, note that by assumption, $\log(1/\tilde{\varepsilon}_n) \lesssim \log(1/\delta) \lesssim \log(n)$ and $\tilde{\varepsilon}_n < n^{-\beta/(2\beta+D)}$, so that

$$\varepsilon_n \lesssim \log^p n \times \left\{\frac{\tilde{\varepsilon}_n^{-D/2\beta}}{\sqrt{n}} \vee \tilde{\varepsilon}_n\right\} = \log^p n \times \frac{1}{\sqrt{n\delta^{D/(2\alpha_0-\alpha_\perp)}}}. \qquad \square$$

We end this section with the presentations of the technical lemmata that are keys to the proof of Theorem 3.1.

LEMMA 5.1. *Under assumptions of Table 1, for any sequence $\varepsilon_n \to 0$ such that $n\varepsilon_n^2 \geq n^{t_0}$, for some $\omega > 0$, and for all $c_1 > 0$, if R_0, H_0, σ_1^2 are large enough and σ_0^2 is small enough, there exists $M_0 > 0$ such that*:

(i) $\Pi(\mathcal{F}_n^c) \lesssim \exp(-c_1 n\varepsilon_n^2)$;

(ii) *In the case of the partial location-scale prior*

$$\sum_{\mathbf{j},\ell} \sqrt{\Pi(\mathcal{F}_{n,\mathbf{j},\ell}) N(\varepsilon_n, \mathcal{F}_{n,\mathbf{j},\ell}, \|\cdot\|_1)} e^{-M_0 n\varepsilon_n^2} = o(1);$$

(iii) *In the case of the hybrid location-scale prior*

$$\sum_{\mathbf{j}} \sqrt{\Pi(\mathcal{F}_{n,\mathbf{j}}) N(\varepsilon_n, \mathcal{F}_{n,\mathbf{j}}, \|\cdot\|_1)} e^{-M_0 n\varepsilon_n^2} = o(1).$$

The proof of Lemma 5.1 can be found in [8], Section S4.1. The last elementary brick in the proof of Theorem 3.1 is the control of the probability of small balls around $\mathbb{P}_0$, which is stated below.

LEMMA 5.2. *Let $\varepsilon > 0$ and assume that it is small enough so that $\varepsilon^{2\alpha_0} \leq \delta^\beta \varepsilon^{\alpha_\perp}$ and that $\varepsilon^{\alpha_\perp} \ll \log^{-1}(1/\varepsilon)$. Then, in the context of Theorem 3.1, there holds*

$$\Pi\big(f_P \in \mathsf{B}(f_0, \tilde{\varepsilon})\big) \gtrsim \exp\{-C\varepsilon^{-D/\beta} \log^t(1/\varepsilon)\},$$

where $\mathsf{B}(f_0, \cdot)$ is defined in (19), where the constant C depends on the parameters, where $\tilde{\varepsilon} \simeq \varepsilon \log^s(1/\varepsilon)$ and with $s, t > 0$ depending on D, β and κ.

The complete proof is given in [8], Section S4.2, and is sketched as follow: the first step is the Hellinger approximation of f_0 by $K_\Sigma \tilde{h}$ as expressed in Corollary 3.5. Then we exhibit an ε-approximation of $K_\Sigma \tilde{h}$ by a discrete location scale mixture with a controlled number of atoms through the use of Lemma 5.3 below. The result then follows from similar arguments as in [41, 60] or [50].

LEMMA 5.3. *Let $\varepsilon > 0$ such that $\delta\sigma^{\alpha_\perp} \log(1/\varepsilon) \ll 1$. For any density g on M^δ satisfying (14), there exists a discrete probability measure G on $\mathbb{R}^D$ with at most $N \simeq \sigma^{-D} \log^D(1/\varepsilon)$ atoms such that*

$$\|K_\Sigma G - K_\Sigma g\|_\infty \lesssim \frac{\varepsilon}{\sigma^D \delta^{D-d}} \quad \textit{and} \quad \|K_\Sigma G - K_\Sigma g\|_1 \lesssim \varepsilon \log^{D/2}(1/\varepsilon).$$

The atoms of G are in M^δ and are $\sigma^{2\alpha_0}\varepsilon$-apart.

The proof of Lemma 5.3 can be found in [8], Section S4.3. We underline that although it uses similar ideas to [29], Lem 3.1, it is not a straightforward adaptation of it, since in K_Σ the covariances depend on the locations of the mixture in a complicated way.

5.2. *Proof of Theorem* 3.4. As explained in Section 3.2 a key ingredient of the proof of Theorem 3.1 is the pointwise approximation of f_0 by $K_\Sigma \tilde{g}$ where g is close to f_0 and is explicited in the proof of Theorem 3.4 below.

PROOF OF THEOREM 3.4. Let $x_0 \in M$ and define

$$\mathcal{W}_{x_0}^j := \mathsf{B}_{T_{x_0}}\left(0, \frac{2+j}{16}\tau\right) \times \mathsf{B}_{N_{x_0}}\left(0, \frac{6+j}{8}\tau\right) \quad \text{for } j \in \{0,1,2\},$$

and $\mathcal{O}_{x_0}^j = \bar{\Psi}_{x_0}(\mathcal{O}_{x_0}^j)$ for $j \in \{0,1,2\}$. We have $\mathcal{W}_{x_0}^0 \subset \mathcal{W}_{x_0}^1 \subset \mathcal{W}_{x_0}^2$ and $\mathcal{O}_{x_0}^0 \subset \mathcal{O}_{x_0}^1 \subset \mathcal{O}_{x_0}^2$. Furthermore, the sets $\mathcal{O}_{x_0}^0$ for $x_0 \in M$ forms a covering of $M^{3\tau/4}$; see [8], Section S1.3, for more details.

We now drop the x_0 from the notation. Let $f : \mathbb{R}^D \to \mathbb{R}$ be in $\mathcal{H}_\delta^{\beta_0,\beta_\perp}(M, L)$ and supported on $\mathcal{O}^0$ — it is to be thought of as f_0 multiplied by a smooth function supported on $\mathcal{O}^0$. Take $x \in \mathcal{O}^1$ and compute

$$K_\Sigma f(x) := \int_{\mathbb{R}^D} \varphi_{\Sigma(u)}(x-u) f(u)\,\mathrm{d}u = \int_{M^\delta \cap \mathcal{O}^0} \varphi_{\Sigma(u)}(x-u) f(u)\,\mathrm{d}u.$$

We first prove that we can construct a function g such that $K_\Sigma g$ is close to f and we then apply this result to $f = f_0 \chi_j$ with χ_j the partition of unity defined in [8], Lemma S1.4.

The idea is to use the fact that $\Sigma(u) = o(1)$ and the smoothness of $u \mapsto \Sigma(u)$ so that

$$K_\Sigma f(x) \approx \int_{M^\delta \cap \mathcal{O}^0} \varphi_{\Sigma(x)}(x-u) f(u)\,\mathrm{d}u \approx f(x).$$

We now write down the approximation rigorously and quantify the error, taking into accound the geometry of the manifold M. In all that follows, we use the notation $z = (v, \eta)$ for points belonging to $\mathsf{B}_{T_{x_0}M}(0, \tau/16) \times \mathsf{B}_{N_{x_0}M}(0, \tau/2)$, while throughout $w = (v, \delta\eta) \in \mathsf{B}_{T_{x_0}M}(0, \tau/16) \times \mathsf{B}_{N_{x_0}M}(0, \delta\tau/2)$. We first make the change of variable $w = \bar{\Psi}^{-1}(u)$, yielding

$$K_\Sigma f(x) = \int_{\bar{\Psi}^{-1}(M^\delta \cap \mathcal{O}^0)} \varphi_{\Sigma(\bar{\Psi}(w))}\big(x - \bar{\Psi}(w)\big) f\big(\bar{\Psi}(w)\big) \big|\det \mathrm{d}\bar{\Psi}(w)\big|\,\mathrm{d}w.$$

Then, denoting by $w_x = \bar{\Psi}^{-1}(x)$, we write

$$w = \Delta_{\sigma,\delta} z + w_x = \Delta_{\sigma,\delta} z + \Delta_{1,\delta} z_x, \quad w_x = (v_x, \delta\eta_x), z_x = (v_x, \eta_x),$$

in the integral above, giving

$$K_\Sigma f(x) = \frac{1}{(2\pi)^{D/2}\delta^{D-d}} \int_{\Delta_{\sigma,1}^{-1}(\mathcal{W}^0 - z_x)} e^{-B_\sigma(z_x,z)} \bar{f}_\delta(\Delta_{\sigma,1} z + z_x) \zeta(\Delta_{\sigma,\delta} z + w_x)\,\mathrm{d}z, \tag{20}$$

with

$$B_\sigma(z_x, z) := \frac{1}{2}\big\|x - \bar{\Psi}(\Delta_{\sigma,\delta} z + w_x)\big\|^2_{\Sigma^{-1}(\bar{\Psi}(\Delta_{\sigma,\delta} z + w_x))}$$

$$\text{and} \quad \zeta(\Delta_{\sigma,\delta} z + w_x) := \big|\det \mathrm{d}\bar{\Psi}(\Delta_{\sigma,\delta} z + w_x)\big|,$$

and where we used the fact that $|\det \Sigma(u)|$ is constantly $\sigma^D \delta^{D-d}$ for $u \in M^\tau$. Notice that $\bar{f}_\delta$ is zero outside of $\Delta_{\sigma,1}^{-1}(\mathcal{W}^0 - z_x)$ so we can replace the latter by $T_{x_0}M \times N_{x_0}M \simeq \mathbb{R}^D$. We denote

$$\bar{K}_\Sigma \bar{f}_\delta(z_x) = \frac{1}{(2\pi)^{D/2}} \int e^{-B_\sigma(z_x,z)} \bar{f}_\delta(\Delta_{\sigma,1} z + z_x) \zeta(\Delta_{\sigma,\delta} z + w_x)\,\mathrm{d}z.$$

We now develop each term separately. First

$$\bar{f}_\delta(\Delta_{\sigma,1}z + z_x) = \underbrace{\bar{f}_\delta(z_x)}_{=\delta^{D-d}f(x)} + \sum_{0<\langle k,\alpha\rangle<\beta} \frac{z^k}{k!}\sigma^{\langle k,\alpha\rangle}\,\mathrm{D}^k\bar{f}_\delta(z_x) + R_\sigma^f(x,z) \tag{21}$$

with $R_\sigma^f(z_x,z) \le DL_\delta(z_x)\sigma^\beta\|1\vee z\|_1^{\beta_{\max}} = D\delta^{D-d}L(x)\sigma^\beta\|1\vee z\|_1^{\beta_{\max}}$ in application of Corollary A.3. We can thus write

$$\bar{K}_\Sigma\bar{f}_\delta(z_x) = \bar{f}_\delta(z_x)I_0(z_x) + \sum_{0<\langle k,\alpha\rangle<\beta}\frac{\mathrm{D}^k\bar{f}_\delta(z_x)}{k!}\sigma^{\langle k,\alpha\rangle}I_k(z_x) + R_\sigma^K(z_x), \tag{22}$$

with

$$I_k(z_x) = \frac{1}{(2\pi)^{D/2}}\int e^{-B_\sigma(z_x,z)}\zeta(\Delta_{\sigma,\delta}z + w_x)z^k\,\mathrm{d}z$$

$$\text{and}\quad R_\sigma^K(z_x) = \frac{1}{(2\pi)^{D/2}}\int e^{-B_\sigma(z_x,z)}\zeta(\Delta_{\sigma,\delta}z + w_x)R_\sigma^f(z_x,z)\,\mathrm{d}z.$$

We will show that:

(i) $R_\sigma^K(z_x) = O(\sigma^\beta\delta^{D-d}L(x))$ uniformly;
(ii) For all k, $I_k(z_x) \in \mathcal{H}_{\mathrm{iso}}^{\beta_M-3}(\mathcal{W}_{x_0}^0, C)$ for some constant $C>0$;
(iii) $I_0(z_x) = 1 + o(1)$ uniformly.

If these three assertions holds, then we can recursively define a function g such that

$$\bar{K}_\Sigma\bar{g}(x) = \bar{f}_\delta(x) + O\big(\sigma^\beta L_\delta(x)\big)$$

uniformly on x. The recursion goes as follows: we let $\mathcal{Q} = \{0 < \langle k,\alpha\rangle \mid \langle k,\alpha\rangle < \beta\} = \{q_1 < \cdots < q_N\}$ and let $q_0 = 0$ and define

$$\bar{g}_0(z_x) = \bar{f}_\delta(z_x)/I_0(z_x),$$

$$\bar{g}_1(z_x) = \bar{g}_0(z_x) - \frac{\sigma^{q_1}}{I_0(z_x)}\sum_{\langle k,\alpha\rangle=q_1}\frac{1}{k!}\mathrm{D}^k(\bar{g}_0)(z_x)I_k(z_x),$$

then using [8], Proposition S2.1, for all $\langle k,\alpha\rangle = q_1$, $\mathrm{D}^k(\bar{g}_0) \in \mathcal{H}_{\mathrm{an}}^{\boldsymbol{\beta}(1-q_1/\beta)}(\tilde{C}_kL)$ for some $\tilde{C}_k > 0$ and we apply (22) to $\bar{g}_1$ which leads to

$$\begin{aligned}
\bar{K}_\Sigma\bar{g}_1(z_x) &= \bar{f}_\delta(z_x) - \sigma^{q_1}\sum_{\langle k,\alpha\rangle=q_1}\frac{1}{k!}\mathrm{D}^k(\bar{g}_0)(z_x)I_k(z_x) + \sum_{q_1\le\langle k,\alpha\rangle<\beta}\sigma^{\langle k,\alpha\rangle}\frac{\mathrm{D}^k\bar{g}_0(z_x)}{k!}I_k(z_x)\\
&\quad - \sigma^{q_1}\sum_{\langle k,\alpha\rangle=q_1}\frac{1}{k!}\sum_{\langle k',\alpha\rangle<\beta-q_1}\frac{\mathrm{D}^{k'}(\mathrm{D}^k(\bar{g}_0)I_k)(z_x))}{k'!}I_{k'}(z_x) + \tilde{R}_\sigma^K(z_x),\\
&= \bar{f}_\delta(z_x) + \sum_{q_2\le\langle k,\alpha\rangle<\beta}\sigma^{\langle k,\alpha\rangle}\frac{\mathrm{D}^k\bar{g}_0(z_x)}{k!}I_k(z_x)\\
&\quad - \sigma^{q_1}\sum_{\langle k,\alpha\rangle=q_1}\frac{1}{k!}\sum_{\langle k',\alpha\rangle<\beta-q_1}\sigma^{\langle k',\alpha\rangle}\frac{\mathrm{D}^{k'}(\mathrm{D}^k(\bar{g}_0)I_k)(z_x))}{k'!}I_{k'}(z_x) + \tilde{R}_\sigma^K(z_x)\\
&= \bar{f}_\delta(z_x) + \sum_{q_2\le\langle k,\alpha\rangle<\beta}\sigma^{\langle k,\alpha\rangle}h_k(z_x)I_k(z_x) + \tilde{R}_\sigma^K(z_x),
\end{aligned}$$

where $\tilde{R}^K_\sigma(z_x)$ also satisfies (i) above and $h_k \in \mathcal{H}_{\mathrm{an}}^{\boldsymbol{\beta}(1-\beta_k/\beta)}(CL)$, $\beta_k = \langle k, \alpha\rangle$ and $C > 0$. All in all, $\bar{g}$ will be of the form

$$\bar{g}(z_x) = d_0(z_x, \sigma, \delta)\bar{f}_\delta(z_x) + \frac{1}{\delta^{D-d}} \sum_{0<\langle k,\boldsymbol{\alpha}\rangle<\beta} \sigma^{\langle k,\boldsymbol{\alpha}\rangle} d_k(z_x, \sigma, \delta)\,\mathrm{D}^k \bar{f}_\delta(z_x)$$

with all d_k uniformly bounded and $d_0(z_x, \sigma, \delta) = 1 + O(\sigma^{2\alpha_0-\alpha_\perp}/\delta)$. Let us now prove (i), (ii) and (iii). There holds

$$\begin{aligned}\bar{\Psi}(\Delta_{\sigma,\delta} z + w_x) &= \Psi\big(v_x + \sigma^{\alpha_0} v\big) + \delta N\big(v_x + \sigma^{\alpha_0} v\big)\big[\eta_x + \sigma^{\alpha_\perp}\eta\big] \\ &= x + \sigma^{\alpha_0}\,\mathrm{d}\Psi(v_x)[v] + \delta\sigma^{\alpha_\perp} N(v_x)\eta + \delta\sigma^{\alpha_0}\,\mathrm{d}N(v_x)[v]\eta_x + R^{\bar{\Psi}}_\sigma(z_x, z),\end{aligned}$$

where, with for $s \in (0, 1)$ $v_s = v_x + s\sigma^{\alpha_0} v$,

$$\begin{aligned}R^{\bar{\Psi}}_\sigma(z_x, z) &= \sigma^{2\alpha_0}\int_0^1 (1-s)\big[\mathrm{d}^2\Psi(v_s)\big[v^{\otimes 2}\big] + \delta\,\mathrm{d}_v^2 N(v_s)\big[v^{\otimes 2}\big]\big(\eta_x + \sigma^{\alpha_\perp}\eta\big)\big]\,\mathrm{d}s \\ &\quad + \delta\sigma^{\alpha_0+\alpha_\perp}\,\mathrm{d}N(v_x)[v]\eta \\ &= O(r_\sigma) \mathrm{with} r_\sigma := \sigma^{2\alpha_0} + \delta\sigma^{\alpha_0+\alpha_\perp}.\end{aligned}$$

The above equation shows that $r_\sigma^{-1} R^{\bar{\Psi}}_\sigma(z_x, z) \in \mathcal{H}_{\mathrm{iso}}^{\beta_M-3}(\mathcal{W}^0, C)$ as a function of z_x, where C depends on $\|\bar{\Psi}\|_{C^{\beta_M-1}}$, uniformly on z. Now notice that in appropriate orthogonal basis, the differential of $\bar{\Psi}$ in w_x writes

$$\begin{pmatrix} \mathrm{d}\Psi(v_x) + \mathrm{pr}_{T_x}\,\mathrm{d}N(v_x)[\cdot](\delta\eta_x) & 0 \\ \mathrm{pr}_{N_x}\,\mathrm{d}N(v_x)[\cdot](\delta\eta_x) & N(v_x)\end{pmatrix}$$

so that in particular, $|\det \mathrm{d}\Psi(w_x)| = |\det\{\mathrm{d}\Psi(v_x) + \mathrm{pr}_{T_x}\,\mathrm{d}N(v_x)[\cdot](\delta\eta_x)\}|$. Let now $u = \bar{\Psi}(\Delta_{\sigma,\delta} z + w_x)$. The projection pr_{T_u} only depends smoothly on $\mathrm{d}\Psi(v_x + \sigma^{\alpha_0} v)$ so that there $R^T_\sigma(z_x, v)$ such that

$$\mathrm{pr}_{T_u}[y] = \mathrm{pr}_{T_x}[y] + R^T_\sigma(z_x, v)(y)$$

with

$$R^T_\sigma(z_x, v)(y) = \sigma^{\alpha_0}\int_0^1 (1-s)\,\mathrm{d}_v\,\mathrm{pr}_{T_{v_s}}[v](y)\,\mathrm{d}s \lesssim \sigma^{\alpha_0}\|y\|,$$

and that for all $y \neq 0$, $\|y\|^{-1}\sigma^{-\alpha_0} R^T_\sigma(\cdot, v)(y) \in \mathcal{H}_{\mathrm{iso}}^{\beta_M-2}(\mathcal{W}^0, C)$. This also implies that $\mathrm{pr}_{N_u}[y] = \mathrm{pr}_{N_x}[y] - R^T_\sigma(x, v)[y]$. As a results, we find that

$$\mathrm{pr}_{T_u}(u - x) = \sigma^{\alpha_0}\,\mathrm{d}\Psi(v_x)v + \delta\sigma^{\alpha_0}\,\mathrm{pr}_{T_x}\,\mathrm{d}N(v_x)[v]\eta_x + \widetilde{R}^T_\sigma(z_x, z)$$

with $r_\sigma^{-1}\widetilde{R}^T_\sigma(x, z) \in \mathcal{H}_{\mathrm{iso}}^{\beta_M-3}(\mathcal{W}^0, C)$ as a function of z_x, uniformly in z. Likewise,

$$\mathrm{pr}_{N_u}(u - x) = \delta\sigma^{\alpha_0}\,\mathrm{pr}_{N_x}\,\mathrm{d}N(v_x)[v]\eta_x + \delta\sigma^{\alpha_\perp} N(v_x)\eta + \widetilde{R}^N_\sigma(z_x, z)$$

with $r_\sigma\widetilde{R}^N_\sigma(x, z) \in \mathcal{H}_{\mathrm{iso}}^{\beta_M-3}(\mathcal{W}^0, C)$ as a function of z_x, uniformly in z. All in all, we get that

$$\begin{aligned}B_\sigma(z_x, z) &= \frac{1}{2\sigma^{2\alpha_0}}\|\mathrm{pr}_{T_u}(u - x)\|^2 + \frac{1}{2\delta^2\sigma^{2\alpha_\perp}}\|\mathrm{pr}_{N_u}(u - x)\|^2 \\ &= \frac{1}{2}\|A_\sigma(z_x, z)\|^2 + R^B_\sigma(z_x, z),\end{aligned}$$

where

$$A_\sigma(z_x, z) = \mathrm{d}\Psi(v_x)v + \delta\,\mathrm{pr}_{T_x}\,\mathrm{d}N(v_x)[v]\eta_x + \sigma^{\alpha_0-\alpha_\perp}\,\mathrm{pr}_{N_x}\,\mathrm{d}N(v_x)[v]\eta_x + N(v_x)\eta, \tag{23}$$

and where $R^B_\sigma(z_x, z) = O(\tilde{r}_\sigma)$ with $\tilde{r}_\sigma = \sigma^{\alpha_0} + \delta\sigma^{\alpha_\perp} + \sigma^{2\alpha_0}/(\delta\sigma^{\alpha_\perp}) = o(1)$, together with $\tilde{r}_\sigma^{-1} R^B_\sigma(z_x, z) \in \mathcal{H}^{\beta_M - 3}_{\mathrm{iso}}(\mathcal{W}^0, C)$ as a function of z_x, uniformly in z. Finally, let us write that

$$\zeta(w_x + \Delta_{\sigma,\delta} z) = \left|\det \mathrm{d}\bar{\Psi}(w_x)\right| + R^\zeta_\sigma(z_x, z)$$

and where $R^\zeta_\sigma(z_x, z)$ is $(\beta_M - 2)$-Hölder as a function of z_x, uniformly in z. Let us now show that (i) and (ii) holds. We show that if $|F(z_x, z)| \le A\|z\|^r$, then

$$\int e^{-\frac{1}{2}\|A_\sigma(z_x,z)\|^2} \left|F(z_x, z)\right| \mathrm{d}z \le AC_r \tag{24}$$

for some constant $C_r > 0$ depending on r and C_M. To do so, let us notice that $A_\sigma(x, \cdot)$ is an invertible linear map such that $\mathrm{spec}(A_\sigma(z_x, \cdot)) = \mathrm{spec}(\mathrm{d}\bar{\Psi}(w_x))$ and thus

$$\begin{aligned}\int e^{-\frac{1}{2}\|A_\sigma(z_x,z)\|^2} \left|F(x, z)\right| \mathrm{d}z &= A\left|\det A_\sigma(z_x, \cdot)\right|^{-1} \int_{\mathbb{R}^D} e^{-\frac{1}{2}\|z\|^2} \left\|A_\sigma(z_x, \cdot)^{-1} z\right\|^r \mathrm{d}z \\ &= A|\det \mathrm{d}\bar{\Psi}(w_x)|^{-1} \|\mathrm{d}\bar{\Psi}(w_x)^{-1}\|^r_{\mathrm{op}} \int_{\mathbb{R}^D} e^{-\frac{1}{2}\|z\|^2} \|z\|^r \mathrm{d}z \\ &\le AC_r,\end{aligned}$$

where C_r depends only on r and C_M. Using this newly proven (24), we have immediately that (i) and (ii) holds. For (iii), notice that

$$\begin{aligned}I_0(x) &= \frac{1}{(2\pi)^{D/2}} \int e^{-\frac{1}{2}\|A_\sigma(z_x,z)\|^2} e^{-\frac{1}{2}R^B_\sigma(z_x,z)} \left(\left|\det \mathrm{d}\bar{\Psi}(w_x)\right| + R^\zeta_\sigma(z_x, z)\right) \mathrm{d}z \\ &= \frac{|\det \mathrm{d}\bar{\Psi}(w_x)|}{(2\pi)^{D/2}} \int e^{-\frac{1}{2}\|A_\sigma(z_x,z)\|^2} \mathrm{d}z + \frac{1}{(2\pi)^{D/2}} \int e^{-\frac{1}{2}\|A_\sigma(z_x,z)\|^2} R^I_\sigma(z_x, z) \,\mathrm{d}z,\end{aligned}$$

where $R^I_\sigma(z_x, z) = (e^{-\frac{1}{2}R^B_\sigma(z_x,z)} - 1)|\det \mathrm{d}\bar{\Psi}(w_x)| + e^{-\frac{1}{2}R^B_\sigma(z_x,z)} R^\zeta_\sigma(z_x, z) = o(1)$ uniformly. Now there holds

$$\int e^{-\frac{1}{2}\|A_\sigma(z_x,z)\|^2} \mathrm{d}z = (2\pi)^{D/2}\left|\det\left(\mathrm{d}\bar{\Psi}(w_x)\right)\right|^{-1} + o(1),$$

and using again (24)

$$\frac{1}{(2\pi)^{D/2}} \int e^{-\frac{1}{2}\|A_\sigma(z_x,z)\|^2} R^I_\sigma(z_x, z) \,\mathrm{d}z = o(1)$$

so that in total $I_0(x) = 1 + o(1)$ as expected.

We are now ready to prove Theorem 3.4. We let

$$R = \left\{H \log(1/\sigma)\right\}^{1/\kappa}$$

and $\chi_1, \dots, \chi_J$ be the functions defined at [8], Lemma S1.4, from a $\tau/64$-packing of $M \cap \mathsf{B}(0, R)$. Recall that we can always chose J of order less than R^D. In light of the point (iv) of [8], Lemma S1.4, the function $f_j := \chi_j f_0$ is still in $\mathcal{H}^{\beta_0,\beta_\perp}_\delta(M, CL)$ for some constant C depending on τ and $\beta_\perp$. Since $\mathrm{supp}\, f_j \subset \mathcal{O}^0_{x_j}$ (point (i) of [8], Lemma S1.4), the first part of this proof yields that there exists some functions g_j supported on $\mathcal{O}^0_{x_j}$, such that

$$\left|K_\Sigma g_j(x) - f_j(x)\right| \lesssim L(x)\sigma^\beta \tag{25}$$

uniformly on $\mathcal{O}^1_{x_j}$. Now notice that for x outside of $\mathcal{O}^1_{x_j}$, we have $d(x, \mathcal{O}^0_{x_j}) > (\sigma^{\alpha_0} \vee \delta\sigma^{\alpha_\perp})\sqrt{(H + D)\log(\sigma)}$ so that

$$\left|K_\Sigma g_j(x)\right| \le \int_{\mathcal{O}^0_{x_j}} \left|\varphi_{\Sigma(u)}(x - u) g_j(u)\right| \mathrm{d}u \le \frac{\sigma^H}{(2\pi)^{D/2}\delta^{D-d}} \int_{\mathcal{O}^0_{x_j}} \left|g_j(u)\right| \mathrm{d}u \lesssim \sigma^H \sup_{\mathcal{O}^0_{x_j}} L$$

and the equality (25) extends to the whole set $\mathbb{R}^D$ with the bound $\sigma^H\|L\|_\infty$ on $\mathbb{R}^D \setminus \mathcal{O}^1_{x_j}$. Using the linearity of K_Σ, we thus find that for $g = \sum_{j=1}^J g_j$, and for any $x \in \mathbb{R}^D$, there holds,

$$\begin{aligned}
\left|K_\Sigma g(x) - f_0(x)\right| &\le \sum_{j=1}^J \left|K_\Sigma g_j(x) - f_j(x)\right| \\
&\le \sum_{j\in J(x)} \left|K_\Sigma g_j(x) - f_j(x)\right| + \sum_{j\notin J(x)} \left|K_\Sigma g_j(x)\right| \\
&\lesssim \left|J(x)\right| \times L(x)\sigma^\beta + \left(J - \left|J(x)\right|\right) \times \sigma^H \|L\|_\infty \\
&\lesssim \sigma^\beta L(x) + \left\{H \log(1/\sigma)\right\}^{D/\kappa} \sigma^H \|L\|_\infty,
\end{aligned}$$

where we denoted $J(x) = \{1 \le j \le J \mid x \in \mathcal{O}^1_{x_j}\}$, and used the fact that $J(x)$ is bounded from above by something depending on D and τ only, ending the proof. $\square$

6. Discussion. With the aim of developing Bayesian procedures in the framework of manifold learning, we exhibited a new family of priors based on location-scales of Dirichlet mixture of Gaussians, and described a general setting for studying density supported near a submanifold. The latter relies on two things: first, a parametrization of the offset of the manifold and second, an anisotropic class of Hölder functions. In this model, we obtained concentration rates in Theorem 3.1 for the associated posterior distribution that are adaptive to the regularity of the underlying density while being totally agnostic of the underlying submanifolds and their main features. Our procedure is therefore fully adaptive.

An interesting feature of our theoretical framework is that it allows to express the rate in terms of the smoothnesses of the density and the manifold together with the thickness δ of the support around the manifold. When δ is fixed, our results can be viewed as an extension of minimax rates for regular anisotropic densities to manifold driven anisotropic densities. But we are also considering the regime where $\delta = o(1)$ which corresponds to the manifold learning problem. The rates obtained in Theorem 3.1 have two regimes: one when δ is not too small with rate $n^{-\beta/(2\beta+D)}$ (up to $\log n$ terms) and the concentrated regime where δ is very small ($\delta = o(n^{-2/(2\beta+D)})$) where we obtain the rate $(n\delta^{D/(2\alpha_0-\alpha_\perp)})^{-1/2}$. It is not clear if this latter rate is optimal or not. The case $\delta = 0$ would correspond to the observations belonging to the manifold M (and for which we would expect the rate $n^{-\beta_0/(2\beta_0+d)}$) but cannot be thought as a limiting case of our problem since then the distribution has density with respect to the Hausdorff measure on M and not with respect to the Lebesgue measure on $\mathbb{R}^D$. When M is unknown the model is not dominated and our approach is not applicable. The problem of posterior contraction rates when the distribution lives on an unknown manifold remains open, although some interesting ideas in [61] or [11] could be used to address it.

An important special case of our framework is the deconvolution model as defined in (ii) of Proposition 2.3. For such a model, our results address the problem of the predictive density of the observations, which is the direct problem. Other quantities are also of interest in deconvolution models, such as the distribution of the unobserved signal or the underlying manifold M. These are important question and we believe that our framework can be used as a first step to answer these questions, possibly using inversion inequalities as in [57]

Another interesting output of our results is that if nonparametric mixtures of normal densities define a versatile and flexible model for smooth densities, the structure of the prior on the mixing distribution is crucial. In this paper we propose two classes of priors which we believe enjoy many strong theoretical properties while remaining reasonably simple to implement. Moreover variational Bayes algorithms can be easily implemented and for which the same theoretical guarantees hold. Moreover, in order to make the MCMC algorithm more scalable to the dimension D and to the number of observations n, we could use a variant based on parallel computing for Dirichlet process mixtures (see, for instance, [47]). It is quite possible that other nonparametric mixture models such as the Fisher–Gaussian kernels of [49] would enjoy the same theoretical guarantees and we believe that our approximation result can be useful to study the theoretical properties of mixtures of Fisher–Gaussian kernels which are strongly related to Gaussian kernels.

APPENDIX: SOME FACTS ON SUBMANIFOLDS WITH BOUNDED REACH

A.1. The geometry of submanifolds with bounded reach. The reach τ_K of a closed subset $K \subset \mathbb{R}^D$, initially introduced by [22], Def 4.1 p. 432, is the supremum of all the $r \geq 0$ such that the orthogonal projection from the r-offset $K^r = \{x \in \mathbb{R}^D \mid d(x, K) \leq r\}$ to K is well-defined, namely

$$\tau_K := \sup\{r \geq 0 \mid \forall x \in K^r, \forall y, z \in K, d(x, K) = \|x - y\| = \|x - z\| \Rightarrow y = z\}.$$

When the reach of a closed submanifold $M \subset \mathbb{R}^D$ is bounded away from zero, M enjoys a number useful properties, that we list and prove below. In all the results stated hereafter, the reach of M is bounded from below by some $\tau > 0$. Lemma A.1 provide already existing results from the literature, sometimes slightly rephrased to better suit our needs. We start off with a control of the exponential map over submanifold with bounded curvature together with a comparison between the intrinsic distance on M, denoted by $d_M(\cdot, \cdot)$, and the ambient Euclidean distance. Recall that for any $x, y \in M$, $d_M(x, y)$ is the infimum of the length of all continuous path between x and y in M and if x and y are in two separate path-connected components, then $d_M(x, y) = \infty$.

LEMMA A.1. *The following facts hold true*:

(i) *For any $x \in M$, the exponential map* $\exp_x$ *is a diffeomorphism from $B_{T_xM}(0, \pi\tau)$ to* $\exp_x\{B_{T_{x_0}M}(0, \pi\tau)\}$.

(ii) *It is double Lipschitz from $\mathsf{B}_{T_xM}(0, \tau/4)$ to its image with*

$$\forall v, w \in \mathsf{B}_{T_xM}(0, \tau/4), \quad \frac{11}{16}\|v - w\| \leq \|\exp_x(v) - \exp_x(w)\| \leq \frac{21}{16}\|v - w\|. \tag{26}$$

(iii) *Letting $\kappa : \gamma \mapsto 2(1 - \sqrt{1 - \gamma})/\gamma$. If $\|x - y\| \leq \gamma\tau/2$ with $\gamma \leq 1$, then*

$$\|x - y\| \leq d_M(x, y) \leq \kappa(\gamma)\|x - y\|. \tag{27}$$

(iv) *Finally, if $\|x - y\| \leq \tau/2$, there holds*

$$\| \mathrm{pr}_{T_xM} - \mathrm{pr}_{T_yM} \|_{\mathrm{op}} \leq \frac{d_M(x, y)}{\tau} \leq \frac{2}{\tau}\|x - y\|. \tag{28}$$

PROOF. The first result on $\exp_x$ is an application of [2], Theorem 1.3. For (ii), denoting $R_x(v) = \exp_x(v) - x - v$, there holds that

$$\big|\|\exp_x(v) - \exp_x(w)\| - \|v - w\|\big| \leq \|R_x(v) - R_x(w)\| \leq \frac{5}{16}\|v - w\|$$

where we used [1], Lem 1. Finally (iii) comes from the monotonicity of κ and [53], Prp 6.3, and using [10], Lem 6, there holds $\| \mathrm{pr}_{T_xM} - \mathrm{pr}_{T_yM} \|_{\mathrm{op}} \leq d_M(x, y)/\tau$, which, together with (iii) for $\gamma = 1$, leads to (iv). □

We now wish to define the parametrization of the $\tau/2$-offset of M that we introduced in Section 2.2. This requires to identify in a nonambiguous way every normal fiber N_xM to the base fiber $N_{x_0}M$ for x in the vicinity of x_0. A natural way to do that would be to use parallel transport, and define

$$N_{x_0}(v,\eta) := t_\gamma(\eta),$$

where $t_\gamma : N_{x_0}M \to N_{\exp_{x_0}(v)}M$ is the parallel transport along the path $\gamma(s) := \exp_{x_0}(sv)$. We refer to [43], Section 4, for a formal introduction to parallel transport.

In order to make things more comprehensible for the reader who is unfamiliar with parallel transports, and in order to have clear and quantitative controls and the quantity at stake, we suggest another, more elementary approach. We assert that the two approaches yields similar regularity classes as introduced in Section 2.2. We start off with a few notations. For a matrix $A \in \mathbb{R}^{D\times D}$ and $1 \le k \le D$, we let $V_k(A)$ be the vector space spanned by the first k columns of A. We denote by $\mathrm{Norm} : x \mapsto x/\|x\|$. We let

$$\mathcal{G} : GL(D,\mathbb{R}) \to O(D,\mathbb{R})$$

be the Gram-Schmidt process, defined recursively on the columns of any invertible matrix $A = (A_1,\dots,A_D)$ as $\mathcal{G}(A) = (\mathcal{G}_1(A),\dots,\mathcal{G}_D(A))$ with

$$\mathcal{G}_1(A) := \mathrm{Norm}(A_1) \quad \text{and} \quad \forall 1 \le j \le D-1,$$
$$\mathcal{G}_{j+1}(A) := \mathrm{Norm}\big(\bar{\mathcal{G}}_{j+1}(A)\big) \quad \text{where } \bar{\mathcal{G}}_{j+1}(A) := A_{j+1} - \sum_{1\le i\le j} \langle A_{j+1}, \mathcal{G}_i(A)\rangle \mathcal{G}_i(A).$$

Because $\mathcal{G}$ is such that $V_k(A) = V_k(\mathcal{G}(A))$ for every $1 \le k \le D$, there holds that $\bar{\mathcal{G}}_{k+1}(A) = \mathrm{pr}_{V_k(A)^\perp}(A_{k+1})$, and that $\bar{\mathcal{G}}_j(A)$ is thus non zero everywhere, so that $\mathcal{G}$ is a well-defined, smooth application. In order to bound its derivatives, we need to control how $\bar{\mathcal{G}}_k$ is far away from zero. We introduce

$$GL_\varepsilon(D,\mathbb{R}) := \big\{A \in GL(D,\mathbb{R}) \mid d\big(A_{k+1}, V_k(A)\big) \ge \varepsilon \ \forall 1 \le k \le D\big\},$$

so that $\|\bar{\mathcal{G}}_k(A)\| \ge \varepsilon$ for every k and any $A \in GL_\varepsilon(D,\mathbb{R})$, and thus straightforwardly all the derivatives, up to any order, of $\mathcal{G}$ are bounded on $GL_\varepsilon(D,\mathbb{R})$. We let B_0 be an arbitrary basis of $T_{x_0}M$, $B_\perp$ be an arbitrary basis of $N_{x_0}M$ and let $B = (B_0, B_\perp)$. We define, for $v \in \mathsf{B}_{T_{x_0}M}(0,\tau/4)$, $A_{x_0}(B,v) := (\mathrm{d}\Psi_{x_0}(v)[B_0], B_\perp)$. Note that since Ψ_{x_0} is a diffeomorphism, there holds

$$V_d\big(A_{x_0}(B,v)\big) = \mathrm{Vect}\big(\mathrm{d}\Psi_{x_0}(v)[B_0]\big) = T_{\Psi_{x_0}(v)}M. \tag{29}$$

Set

$$N^B_{x_0}(v,\cdot) := \mathcal{G}\big(A_{x_0}(B,v)\big)[\cdot], \qquad \bar{\Psi}^B_{x_0}(v,\eta) := \Psi_{x_0}(v) + N^B_{x_0}(v,\eta).$$

We show in Lemma A.2 that $N^B_{x_0}$ and $\bar{\Psi}^B_{x_0}$ are well defined and smooth, which combined with (29) yields in particular that $N^B_{x_0}(v,\cdot)$ is an isometry between $N_{x_0}M$ and $N_{\Psi_{x_0}(v)}M$.

LEMMA A.2. *For any $v \in \mathsf{B}_{T_{x_0}M}(0,\tau/4)$, there holds that $A_{x_0}(B,v) \in GL_\varepsilon(D,\mathbb{R})$ with $\varepsilon = 1/2^d$. Consequently, the map $v \mapsto N^B_{x_0}(v,\cdot)$ is in $\mathcal{H}^{\beta_M-1}_{\mathrm{iso}}(\mathsf{B}_{T_{x_0}M}(0,\tau), C)$ for some constant C_M, τ, β_M and D (and not on B).*

Moreover for any basis B, it holds that $|\det \mathrm{d}\bar{\Psi}^B_{x_0}(v,\eta)| \ge (3/16)^d$ for any $v \in \mathsf{B}_{T_{x_0}M}(0, \tau/4)$ and $\eta \in \mathsf{B}_{N_{x_0}M}(0,\tau/2)$.

The proof of Lemma A.2 can be found in [8], Section S1.1. An important feature of the parametrizations $\bar{\Psi}^B_{x_0}$ is that the subsequent Hölder classes as defined in Definition 2.2 do not depend, up to a universal constant, on the choice of the collection of basis $(B_{x_0})_{x_0\in M}$, nor on the choice of the parametrisation. This is shown in [8], Section S1.2.

A.2. Taylor expansion of M-anisotropic Hölder functions. In this section, we derive a Taylor expansion for manifold-anisotropic Hölder functions. Recall from Section 3 that for any $\sigma, \delta > 0$,

$$\Delta_{\sigma,\delta} = \begin{pmatrix} \sigma^{\alpha_0}\,\mathrm{Id}_d & 0 \\ 0 & \delta\sigma^{\alpha_\perp}\,\mathrm{Id}_{D-d} \end{pmatrix}.$$

COROLLARY A.3. *Let $f \in \mathcal{H}_\delta^{\beta_0,\beta_\perp}(M, L)$. Then, for any $x_0 \in M$, any $w \in \mathcal{W}_{x_0,\delta}$, and any $z \in T_{x_0}M \times N_{x_0}M$ such that $w + \Delta_{\sigma,1}z \in \mathcal{W}_{x_0,\delta}$, there holds*

$$\bar{f}_{x_0,\delta}(w + \Delta_{\sigma,1}z) = \bar{f}_{x_0,\delta}(w) + \sum_{0<\langle k,\boldsymbol{\alpha}\rangle<\beta} \sigma^{\langle k,\alpha\rangle} \frac{z^k}{k!} \mathrm{D}^k \bar{f}_{\delta,x_0}(w) + R(w,z),$$

where the remainder R satisfies the following bound:

$$\left|R(w,z)\right| \le D\sigma^\beta \|1 \vee z\|_1^{\beta_{\max}} L_{x_0,\delta}(w).$$

PROOF OF COROLLARY A.3. Simply applying [8], Proposition S2.2, to $\bar{f}_{x_0,\delta}$ yields

$$\bar{f}_{x_0,\delta}(w + \Delta_{\sigma,1}z) = \bar{f}_{x_0,\delta}(w) + \sum_{0<\langle k,\boldsymbol{\alpha}\rangle<\beta} \frac{(\Delta_{\sigma,1}z)^k}{k!} \mathrm{D}^k \bar{f}_{\delta,x_0}(w) + R_{x_0,\delta}(w, w + \Delta_{\sigma,1}z),$$

where $R_{x_0,\delta}(w, \Delta_{\sigma,1}z)$ satisfies

$$\begin{aligned} \left|R_{x_0,\delta}(w, w + \Delta_{\sigma,1}z)\right| &\le L_{x_0,\delta}(w) \sum_{\beta-\alpha_{\max}\le\langle k,\alpha\rangle<\beta} \frac{|\Delta_{\sigma,1}z|^k}{k!} \sum_{j=1}^D \left|(\Delta_{\sigma,1}z)_j\right|^{\frac{\beta-\langle k,\boldsymbol{\alpha}\rangle}{\alpha_j}} \\ &= L_{x_0,\delta}(w) \sum_{\beta-\alpha_{\max}\le\langle k,\boldsymbol{\alpha}\rangle<\beta} \sigma^{\langle k,\boldsymbol{\alpha}\rangle} \frac{|z|^k}{k!} \sum_{j=1}^D \sigma^{\beta-\langle k,\boldsymbol{\alpha}\rangle} |z_j|^{\frac{\beta-\langle k,\boldsymbol{\alpha}\rangle}{\alpha_j}} \\ &\le D\sigma^\beta L_{x_0,\delta}(w)\|1 \vee z\|_1^{\beta_{\max}}, \end{aligned}$$

where we used the fact that $k_j + \frac{\beta-\langle k,\boldsymbol{\alpha}\rangle}{\alpha_j} \le \beta_j \le \beta_{\max}$ for any $1 \le j \le D$. $\square$

Acknowledgments. We are very thankful to Marc Hoffmann and to Yann Chaubet for helpful discussions, as well as to Leo Zhang for spotting an initial mistake in the Gibbs sampling algorithm. We would also like to express our gratitude to the referees and the associate editor for the great care and attention shown to the manuscript, which helped us improve the results and their exposition.

Funding. This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 834175).

SUPPLEMENTARY MATERIAL

Supplementary Material to "Estimating a density near an unknown manifold: A Bayesian nonparametric approach" (DOI: 10.1214/24-AOS2423SUPP; .pdf). In the supplementary material [8], we present further technical results regarding the geometric framework and general anisotropic Hölder functions. We also provide intermediary results and proofs leading to the proofs of Theorems 3.1 and 3.4. The supplement ends with details on the numerical experiments carried in this paper.

REFERENCES

[1] AAMARI, E. and LEVRARD, C. (2019). Nonasymptotic rates for manifold, tangent space and curvature estimation. *Ann. Statist.* **47** 177–204.

[2] ALEXANDER, S. B. and BISHOP, R. L. (2006). Gauss equation and injectivity radii for subspaces in spaces of curvature bounded above. *Geom. Dedicata* **117** 65–84.

[3] ARJOVSKY, M. and BOTTOU, L. (2017). Towards principled methods for training generative adversarial networks. https://doi.org/10.48550/ARXIV.1701.04862

[4] ARJOVSKY, M., CHINTALA, S. and BOTTOU, L. (2017). Wasserstein GAN. https://doi.org/10.48550/ARXIV.1701.07875

[5] BELKIN, M. and NIYOGI, P. (2001). Laplacian eigenmaps and spectral techniques for embedding and clustering. *Adv. Neural Inf. Process. Syst.* **14**.

[6] BELKIN, M. and NIYOGI, P. (2001). Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems* (T. Dietterich, S. Becker and Z. Ghahramani, eds.) **14**. MIT Press.

[7] BERENFELD, C. and HOFFMANN, M. (2021). Density estimation on an unknown submanifold. *Electron. J. Stat.* **15** 2179–2223.

[8] BERENFELD, C., ROSA, P. and ROUSSEAU, J. (2024). Supplement to "Estimating a density near an unknown manifold: a Bayesian nonparametric approach." https://doi.org/10.1214/24-AOS2423SUPP

[9] BINGHAM, E., CHEN, J. P., JANKOWIAK, M., OBERMEYER, F., PRADHAN, N., KARALETSOS, T., SINGH, R., SZERLIP, P., HORSFALL, P. and GOODMAN, N. D. (2019). Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.* **20** 973–978.

[10] BOISSONNAT, J.-D., LIEUTIER, A. and WINTRAECKEN, M. (2019). The reach, metric distortion, geodesic convexity and the variation of tangent spaces. *J. Appl. Comput. Topol.* **3** 29–58. https://doi.org/10.1007/s41468-019-00029-8

[11] CAMERLENGHI, F., DOLERA, E., FAVARO, S. and MAININI, E. (2022). Wasserstein posterior contraction rates in non-dominated Bayesian nonparametric models. Preprint. Available at arXiv:2201.12225.

[12] CANALE, A. and DE BLASI, P. (2017). Posterior asymptotics of nonparametric location-scale mixtures for multivariate density estimation. *Bernoulli* **23** 379–404.

[13] CAPITAO-MINICONI, J. and GASSIAT, É. (2023). Deconvolution of spherical data corrupted with unknown noise. *Electron. J. Stat.* **17** 607–649.

[14] CHAE, M., KIM, D., KIM, Y. and LIN, L. (2021). A likelihood approach to nonparametric estimation of a singular distribution using deep generative models. Preprint. Available at arXiv:2105.04046.

[15] CHEN, Y.-C., GENOVESE, C. R. and WASSERMAN, L. (2015). Asymptotic theory for density ridges. *Ann. Statist.* **43** 1896–1928.

[16] CLEANTHOUS, G., GEORGIADIS, A. G. and PORCU, E. (2019). Minimax density estimation on Sobolev spaces with dominating mixed smoothness. Preprint. Available at arXiv:1906.06835.

[17] COMTE, F. and LACOUR, C. (2013). Anisotropic adaptive kernel deconvolution. In *Annales de l'IHP Probabilités et statistiques* **49** 569–609.

[18] DIVOL, V. (2020). Minimax adaptive estimation in manifold inference. https://doi.org/10.48550/ARXIV.2001.04896

[19] DIVOL, V. (2021). Reconstructing measures on manifolds: An optimal transport approach. Preprint. Available at arXiv:2102.07595.

[20] DUNSON, D., CHEN, M., SILVA, J., PAISLEY, J., WANG, C. and CARIN, L. (2010). Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds. *IEEE Trans. Signal Process.* **58** 6140–6155.

[21] DUNSON, D. B. and WU, N. (2021). Inferring manifolds from noisy data using Gaussian processes. Preprint. Available at arXiv:2110.07478.

[22] FEDERER, H. (1959). Curvature measures. *Trans. Amer. Math. Soc.* **93** 418–491.

[23] FEFFERMAN, C., MITTER, S. and NARAYANAN, H. (2016). Testing the manifold hypothesis. *J. Amer. Math. Soc.* **29** 983–1049.

[24] GENOVESE, C. R., PERONE-PACIFICO, M., VERDINELLI, I. and WASSERMAN, L. (2012). Manifold estimation and singular deconvolution under Hausdorff loss. *Ann. Statist.* **40**. https://doi.org/10.1214/12-aos994

[25] GENOVESE, C. R., PERONE-PACIFICO, M., VERDINELLI, I. and WASSERMAN, L. (2014). Nonparametric ridge estimation. *Ann. Statist.* **42** 1511–1545.

[26] GHAHRAMANI, Z., HINTON, G. E. et al. (1996). The EM algorithm for mixtures of factor analyzers Technical Report, Technical Report CRG-TR-96-1, University of Toronto.

[27] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531.

[28] GHOSAL, S. and VAN DER VAART, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann*. *Statist*. 697–723.

[29] GHOSAL, S. and VAN DER VAART, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann*. *Statist*. **29** 1233–1263.

[30] GHOSAL, S. and VAN DER VAART, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

[31] GOLDENSHLUGER, A. and LEPSKI, O. (2011). Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality. *Ann*. *Statist*. **39** 1608–1632.

[32] GOLDENSHLUGER, A. and LEPSKI, O. (2014). On adaptive minimax density estimation on Rd. *Probab. Theory Related Fields* **159** 479–543.

[33] GOODFELLOW, I., BENGIO, Y. and COURVILLE, A. (2016). *Deep Learning*. *Adapt*. *Comput*. *Mach*. *Learn*. MIT Press, Cambridge, MA.

[34] GOODFELLOW, I. J., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A. and BENGIO, Y. (2014). Generative adversarial networks. https://doi.org/0.48550/ARXIV.1406.2661

[35] HOFFMAN, M. and LEPSKI, O. (2002). Random rates in anisotropic regression (with a discussion and a rejoinder by the authors). *Ann*. *Statist*. **30** 325–396.

[36] HORVAT, C. and PFISTER, J.-P. (2021). Density estimation on low-dimensional manifolds: An inflation-deflation approach. Preprint. Available at arXiv:2105.12152.

[37] KERKYACHARIAN, G., LEPSKI, O. and PICARD, D. (2001). Nonlinear estimation in anisotropic multi-index denoising. *Probab*. *Theory Related Fields* **121** 137–170.

[38] KIM, J., SHIN, J., RINALDO, A. and WASSERMAN, L. (2019). Uniform convergence rate of the kernel density estimator adaptive to intrinsic volume dimension. 3398–3407.

[39] KINGMA, D. P. and WELLING, M. (2013). Auto-encoding variational Bayes. https://doi.org/10.48550/ARXIV.1312.6114

[40] KLEIN, W., PLOMP, R. and POLS, L. C. (1970). Vowel spectra, vowel spaces, and vowel identification. *J*. *Acoust*. *Soc*. *Am*. **48** 999–1009.

[41] KRUIJER, W., ROUSSEAU, J. and VAN DER VAART, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electron*. *J*. *Stat*. **4** 1225–1257.

[42] LEE, J. A. and VERLEYSEN, M. (2007). *Nonlinear Dimensionality Reduction* **1**. Springer, Berlin.

[43] LEE, J. M. (2006). *Riemannian Manifolds*: *An Introduction to Curvature* **176**. Springer, Berlin.

[44] MA, Y. and FU, Y. (2012). *Manifold Learning Theory and Applications* **434**. CRC Press, Boca Raton.

[45] MARKWICK, D. and ROSS, G. J. (2020). dirichletprocess: Build Dirichlet process objects for Bayesian modelling. https://cran.r-project.org/package=dirichletprocess.

[46] MAUGIS-RABUSSEAU, C. and MICHEL, B. (2013). Adaptive density estimation for clustering with Gaussian mixtures. *ESAIM Probab. Stat*. **17** 698–724.

[47] MEGUELATI, K., FONTEZ, B., HILGERT, N. and MASSEGLIA, F. (2019). Dirichlet process mixture models made scalable and effective by means of massive distribution. In *Proceedings of the* 34*th ACM/SIGAPP Symposium on Applied Computing* 502–509.

[48] MENGERSEN, K., SANTOS-FERNANDEZ, E., DENTI, F., MIRA, A. and VARGHESE, A. (2022). On the intrinsic dimensionality of Covid-19 data: A global perspective. https://doi.org/10.48550/ARXIV.2203.04165

[49] MUKHOPADHYAY, M., LI, D., DUNSON, D. B. et al. (2020). Estimating densities with non-linear support by using Fisher–Gaussian kernels. *J*. *Roy*. *Statist*. *Soc*. *Ser*. *B* **82** 1249–1271.

[50] NAULET, Z. and ROUSSEAU, J. (2017). Posterior concentration rates for mixtures of normals in random design regression. *Electron*. *J*. *Stat*. **11** 4065–4102.

[51] NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J*. *Comput*. *Graph*. *Statist*. **9** 249–265.

[52] NIKOL'SKII, S. M. (2012). *Approximation of Functions of Several Variables and Imbedding Theorems* **205**. Springer, Berlin.

[53] NIYOGI, P., SMALE, S. and WEINBERGER, S. (2008). Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput*. *Geom*. **39** 419–441.

[54] OZAKIN, A. and GRAY, A. (2009). Submanifold density estimation. *Adv*. *Neural Inf*. *Process*. *Syst*. **22**.

[55] PUCHKIN, N. and SPOKOINY, V. G. (2022). Structure-adaptive manifold estimation. *J*. *Mach*. *Learn*. *Res*. **23** 1–62.

[56] ROČKOVÁ, V. and ROUSSEAU, J. (2023). Ideal Bayesian spatial adaptation. *J*. *Amer*. *Statist*. *Assoc*. **0** 1–14. https://doi.org/10.1080/01621459.2023.2241705

[57] ROUSSEAU, J. and SCRICCIOLO, C. (2023). Wasserstein convergence in Bayesian and frequentist deconvolution models. Preprint. Available at arXiv:2309.15300.

[58] Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290** 2323–2326. https://doi.org/10.1126/science.290.5500.2323

[59] Schölkopf, B., Smola, A. and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput*. **10** 1299–1319.

[60] Shen, W., Tokdar, S. T. and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika* **100** 623–640.

[61] Tang, R. and Yang, Y. (2022). Minimax rate of distribution estimation on unknown submanifold under adversarial losses. Preprint. Available at arXiv:2202.09030.

[62] Tenenbaum, J. B., de Silva, V. and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290** 2319–2323.

[63] Vincent, P. and Bengio, Y. (2002). Manifold parzen windows. *Adv. Neural Inf. Process. Syst*. **15**.

[64] Weinberger, K. Q. and Saul, L. K. (2006). Unsupervised learning of image manifolds by semidefinite programming. *Int. J. Comput. Vis*. **70** 77–90.

SUPPLEMENTARY MATERIAL TO "ESTIMATING A DENSITY NEAR AN UNKNOWN MANIFOLD: A BAYESIAN NONPARAMETRIC APPROACH"

BY CLÉMENT BERENFELD[1,a], PAUL ROSA[2,b] AND JUDITH ROUSSEAU[2,c]

[1]*Institut für Mathematik, Universität Potsdam,* [a]*berenfeld@uni-potsdam.de*

[2]*Department of Statistics, University of Oxford,* [b]*paul.rosa@jesus.ox.ac.uk;* [c]*judith.rousseau@stats.ox.ac.uk*

In this supplement, we present further technical results regarding the geometric framework and general anisotropic Hölder functions. We also provide intermediary results and proofs leading to the proofs of Theorems 3.1 and 3.4. The supplementary ends with details on the numerical experiments carried in the main manuscript.

S1. Additional results and proofs on the geometry of submanifolds.

S1.1. *Proof of Lemma A.2.*

PROOF OF LEMMA A.2. First notice that for $d \le k \le D-1$ and $v \in \mathsf{B}_{T_{x_0}M}(0,\tau/4)$, there holds, letting $A = A_{x_0}(B,v)$, and because of (29) and the fact that $B_\perp$ is an orthonormal frame of $N_{x_0}M$,

$$d(A_{k+1}, V_k(A)) = d(A_{k+1}, T_{\Psi_{x_0}(v)}M) \ge 1 - \| \mathrm{pr}_{T_{x_0}M} - \mathrm{pr}_{T_{\Psi_{x_0}(v)}M} \|_{\mathrm{op}} \ge 3/4$$

where we used (28). Now we let $1 \le k \le d-1$ and let $V_0(A) = \{0\}$. Letting $Q = (A_1, \dots, A_d)$, here holds that

$$\prod_{k=0}^{d-1} d(A_{k+1}, V_k(A))^2 = \det Q^\top Q = \det \mathrm{d}\Psi_{x_0}(v)^\top \mathrm{d}\Psi_{x_0}(v).$$

Using [1, Lem 1], it holds $\| \mathrm{d}\Psi_{x_0}(v) - \iota\|_{\mathrm{op}} \le 5/16$ for $v \in \mathsf{B}_{T_{x_0}M}(0,\tau/4)$ where $\iota : T_{x_0}M \to \mathbb{R}^D$ is the inclusion so that for any $h \in T_{x_0}M$, $\| \mathrm{d}\Psi_{x_0}(v)[h]\| \ge (1-5/16)\|h\| = 11/16\|h\|$. In particular $\det \Psi_{x_0}(v)^\top \mathrm{d}\Psi_{x_0}(v) \ge (11/16)^{2d}$. Since now

$$d(A_{k+1}, V_k(A))^2 \le \|A_{k+1}\|^2 \le \| \mathrm{d}\Psi_{x_0}(v)\|_{\mathrm{op}}^2 \le (21/16)^2$$

there holds that for any $1 \le k \le d$,

$$d(A_{k+1}, V_k(A))^2 \ge (16/21)^{2d}(11/16)^{2d} \ge (11/21)^{2d} \ge 1/2^{2d}.$$

Using that the Gram-Schmidt transform is smooth on $GL_\varepsilon(D,\mathbb{R})$ with $\varepsilon = 1/2^d$ and Proposition S2.7, we obtain that the map $v \mapsto N_{x_0}^B(v,\cdot)$ is in $\mathcal{H}_{\mathrm{iso}}^{\beta_M-1}(\mathsf{B}_{T_{x_0}M}(0,\tau),C)$.

Also, in B and in orthonormal bases of $T_{\Psi_{x_0}(v)}M$ and $N_{\Psi_{x_0}(v)}$, the Jacobian of $\bar\Psi_{x_0}^B(v,\eta)$ write

$$\begin{pmatrix} \mathrm{d}\Psi_{x_0}(v) + \mathrm{pr}_{T_{\Psi_{x_0}(v)}M} \circ \mathrm{d}_v N_{x_0}^B(v,\eta) & 0 \\ \mathrm{pr}_{N_{\Psi_{x_0}(v)}M} \circ \mathrm{d}_v N_{x_0}^B(v,\eta) & N_{x_0}^B(v,\cdot) \end{pmatrix}$$

so that $|\det \mathrm{d}\bar\Psi_{x_0}^B(v,\eta)| = \left|\det\left\{\mathrm{d}\Psi_{x_0}(v) + \mathrm{pr}_{T_{\Psi_{x_0}(v)}M} \circ \mathrm{d}_v N(v,\eta)\right\}\right|$. We saw earlier in the proof that $\| \mathrm{d}\Psi_{x_0}(v)\|_{\mathrm{op}} \ge 11/16$. Furthermore, using (28), we find that for any small $w \in$

$T_{x_0}M$,

$$\| \operatorname{pr}_{T_{\Psi_{x_0}(v)}} \left\{ N^B_{x_0}(v+w,\eta) - N^B_{x_0}(v,\eta) \right\} \| = \| (\operatorname{pr}_{T_{\Psi_{x_0}(v+w)}} - \operatorname{pr}_{T_{\Psi_{x_0}(v)}}) N^B_{x_0}(v+w,\eta) \| \\ \leq \frac{\|w\|}{\tau} \|\eta\|,$$

and consequently $\| \operatorname{pr}_{T_{\Psi_{x_0}(v)}} \circ \mathrm{d}_v N^B_{x_0}(v,\eta) \|_{\mathrm{op}} \leq \|\eta\|/\tau \leq 1/2$. Thus, for any $h \in T_{x_0}M$, $\| \mathrm{d}\Psi_{x_0}(v)[h] + \operatorname{pr}_{T_{\Psi_{x_0}(v)}M} \circ \mathrm{d}_v N(v,\eta)[h] \|_{\mathrm{op}} \geq (11/16 - 1/2)\|h\| \geq 3/16\|h\|$ and thus

$$|\det \mathrm{d}\bar{\Psi}^B_{x_0}(v,\eta)| \geq (3/16)^d.$$

□

S1.2. *Stability by a change of basis.* We now show that a change of basis does not interfere with the anisotropic regularity of a map seen through $\bar{\Psi}^B_{x_0}$.

LEMMA S1.1. *For any orthonormal basis $B' = (B'_0, B'_\perp)$ subordinated to $T_{x_0}M$ and $N_{x_0}M$, and for any $\delta > 0$, it holds that*

$$\left(\bar{\Psi}^B_{x_0,\delta}\right)^{-1} \circ \bar{\Psi}^{B'}_{x_0,\delta}(v,\eta) = (v, C_{B,B'}(v)\eta) \tag{S1}$$

where $C_{B,B'}$ is independent of δ and is in $\mathcal{H}^{\beta_M-1}_{\mathrm{iso}}(\mathsf{B}_{T_{x_0}M}(0,\tau), C)$ for some constant C depending on C_M, τ, β_M and D (and not on B and B').

PROOF. Short and simple computations shows that $C_{B,B'}(v) \coloneqq N^B_{x_0}(v,\cdot)^\top N^{B'}_{x_0}(v,\cdot)$ so that an application of Lemma A.2 with Proposition S2.6 immediately yields the result. □

COROLLARY S1.2. *In the context of Subsection 2.2, assume that $\beta_0 \leq \beta_\perp \leq \beta_M - 1$. Then, there exists a constant C depending on C_M, τ, β_M and D such that, if there exists a basis B such that $f \circ \bar{\Psi}^B_{x_0,\delta} \in \mathcal{H}^{\boldsymbol{\beta}}_{\mathrm{an}}(\mathcal{W}_{x_0,\delta}, L^B_{x_0,\delta})$, then, for any other orthonormal basis B', $f \circ \bar{\Psi}^{B'}_{x_0,\delta} \in \mathcal{H}^{\boldsymbol{\beta}}_{\mathrm{an}}(\mathcal{W}_{x_0,\delta}, CL^{B'}_{x_0,\delta})$.*

PROOF. Using the lemma above, there holds

$$f \circ \bar{\Psi}^{B'}_{x_0,\delta} = f \circ \bar{\Psi}^B_{x_0,\delta}(v, C_{B,B'}(v)\eta) = f \circ \bar{\Psi}^B_{x_0,\delta} \circ J(v,\eta)$$

where J is defined through (S1). We denote for short $f^{B'} = f \circ \bar{\Psi}^{B'}_{x_0,\delta}$, $f^B = f \circ \bar{\Psi}^{B'}_{x_0,\delta}$ so that $f^{B'} = f^B \circ J$. Taking $\langle k, \boldsymbol{\alpha} \rangle < \beta$ and using the multivariate Faa di Bruno formula [3], we find that $\mathrm{D}^k f^{B'}$ is a sum of products of the form

$$\mathrm{D}^\ell (f^B) \circ J \times \prod (\mathrm{D}^{k^{(j)}} J)^{\ell^{(j)}}$$

subject to $|\ell| \leq |k|$, $\sum_j \ell^{(j)} = \ell$ and $\sum_j |\ell^{(j)}| k^{(j)} = k$ with $k^{(j)} \neq 0$. Now notice that $(\mathrm{D}^{k^{(j)}} J)_i = 0$ for $1 \leq i \leq d$ as soon as $k^{(j)}_\perp \neq 0$. For a configuration of $\ell, \ell^{(j)}$ and $k^{(j)}$ such that the above product is not zero, there thus holds

$$|\ell_0| = \sum_j |\ell^{(j)}_0| = \sum_{k^{(j)}_\perp = 0} |\ell^{(j)}_0| \leq \sum_{k^{(j)}_\perp = 0} |\ell^{(j)}_0| |k^{(j)}_0| \leq |k_0|$$

which, together with $|\ell| \leq |k$ and $\alpha_0 \geq \alpha_\perp$, yields that $\langle \ell, \boldsymbol{\alpha} \rangle \leq \langle k, \boldsymbol{\alpha} \rangle$ whenever the above product is non zero. We conclude with a telescopic argument with Lemma S1.1. □

More generally, as noted in Remark 2 the definition of $\mathcal{H}_\delta^{\beta_0,\beta_\perp}(M,L)$ through the construction given in Appendix A does not depend on the choice of Ψ_{x_0} as long as the parametrization Ψ_{x_0} is β_M regular.

PROPOSITION S1.3. *Let $(\Psi'_{x_0})_{x_0\in M}$ be a set of parametrizations of M satisfying $\Psi'_{x_0}\in\mathcal{H}_{\text{iso}}^{\beta_M}(\mathcal{V}_{x_0},C_M)$ for all $x_0\in M$. Let $\beta_0\le\beta_\perp\le\beta_M-3$. We let $\bar\Psi$ and $\bar\Psi'$ be the extensions of Ψ and Ψ' constructed in Appendix A. Then $\mathcal{H}_\delta^{\beta_0,\beta_\perp}(M,L,\bar\Psi)\subset\mathcal{H}_\delta^{\beta_0,\beta_\perp}(M,CL,\bar\Psi')$ for some $C>0$ depending on $\beta_0,\beta_\perp,D$ and C_M.*

PROOF. Let $f\in\mathcal{H}_\delta^{\beta_0,\beta_\perp}(M,L,\bar\Psi)$ and $x_0\in M$. We need to prove that $\delta^{D-d}f\circ\bar\Psi'_{x_0}\in\mathcal{H}_{\text{an}}(\mathcal{W}_{x_0,\delta},\delta^{D-d}L\circ\bar\Psi'_{x_0,\delta})$. We have $\delta^{D-d}f\circ\bar\Psi'_{x_0,\delta}=\bar f_{x_0,\delta}\circ\left[\bar\Psi_{x_0,\delta}^{-1}\circ\bar\Psi'_{x_0,\delta}\right]=\bar f_{x_0,\delta}\circ\tau_{x_0,\delta}$ where we have defined $\tau_{x_0,\delta}=\bar\Psi_{x_0,\delta}^{-1}\circ\bar\Psi'_{x_0,\delta}$. By proposition S2.7 we find that $\tau_{x_0,\delta}\in\mathcal{H}_{\text{iso}}^{\beta_M-1}(\mathcal{W}_{x_0,\delta},C_M^*)$ for some $C_M^*>0$. Moreover by assumption we have $\bar f_{x_0,\delta}\in\mathcal{H}_{\text{an}}^{\boldsymbol{\beta}_{0,\perp}}(W_{x_0,\delta},L_{x_0,\delta})$, and $\tau_{x_0,\delta}$ can be expressed as

$$\begin{aligned}\tau_{x_0,\delta}(v,\eta)&=\left(\Psi_{x_0}^{-1}\circ\Psi'_{x_0}(v),\delta N_{x_0}(\Psi_{x_0}^{-1}\circ\Psi'_{x_0}(v),\cdot)^{-1}(\delta N'_{x_0}(\Psi_{x_0}^{-1}\circ\Psi_{x'_0}(v),\eta))\right)\\&=\left(\tau^{(1)}(v),\tau^{(2)}(v,\eta)\right)\end{aligned}$$

With $\tau^{(1)}\in\mathcal{H}_{\text{iso}}^{\beta_M}(\mathcal{V}_{x_0},C)$ and $\tau^{(2)}\in\mathcal{H}_{\text{iso}}^{\beta_M-1}\left(\mathsf{B}_{N_{x_0}M}(0,\tau/2\delta),C\right)$ for some $C>0$. What is crucial here is that $\tau^{(1)}$ takes only v as argument, which implies that for every multi-index k such that $\langle k,\boldsymbol{\alpha}\rangle<\beta$, the partial derivative $\mathrm{D}^k\left(\bar f_{x_0,\delta}\circ\tau_{x_0,\delta}\right)$ exists and is a linear combination of terms of the form

$$\mathrm{D}^\ell\bar f_{x_0,\delta}(\tau_{x_0,\delta})\prod_{i=1}^{|\ell|}\left(\mathrm{D}^{r_i}\tau_{x_0,\delta}\right)^{q_i}$$

where $|\ell|\le|k|$, $\langle\ell,\boldsymbol{\alpha}\rangle<\beta$, $r_1,\dots,r_{|\ell|}\in\mathbb{N}^D$, $q_1,\dots,q_{|\ell|}\in\mathbb{N}$, and $\sum_{i=1}^D|r_i|q_i=|\ell|$. Since the derivatives of $\tau_{x_0,\delta}$ are bounded this implies that for some $C>0$ we have

$$\left|\mathrm{D}^k\left(\bar f_{x_0,\delta}\circ\tau_{x_0,\delta}\right)\right|\le CL_{x_0,\delta}\circ\tau_{x_0,\delta}=C\delta^{D-d}L\circ\bar\Psi'_{x_0,\delta}$$

Moreover, for every $\beta-\alpha_{\max}\le\langle k,\boldsymbol{\alpha}\rangle<\beta$ and every $x,y\in\mathcal{W}_{x_0,\delta}$, as above we can upper bound $\left|\mathrm{D}^k\left(\bar f_{x_0,\delta}\circ\tau_{x_0,\delta}\right)(x)-\mathrm{D}^k\left(\bar f_{x_0,\delta}\circ\tau_{x_0,\delta}\right)(y)\right|$ by a linear combination of terms of the forms

$$\left|\mathrm{D}^\ell\bar f_{x_0,\delta}(\tau_{x_0,\delta}(x))-\mathrm{D}^\ell\bar f_{x_0,\delta}(\tau_{x_0,\delta}(y))\right|\times\prod_{i=1}^{|\ell|}\left|(\mathrm{D}^{r_i}\tau_{x_0,\delta})^{q_i}(y)\right| \tag{S2}$$

and

$$\left|\prod_{i=1}^{|\ell|}(\mathrm{D}^{r_i}\tau_{x_0,\delta})^{q_i}(x)-\prod_{i=1}^{|\ell|}(\mathrm{D}^{r_i}\tau_{x_0,\delta})^{q_i}(y)\right|\times\left|\mathrm{D}^\ell\bar f_{x_0,\delta}(\tau_{x_0,\delta})(x)\right| \tag{S3}$$

The functions $(\mathrm{D}^{r_i}\tau_{x_0,\delta})^{q_i}$ are in $\mathcal{H}_{\text{iso}}^{\beta_M-1}\left(\mathsf{B}_{N_{x_0}M}(0,\tau/2\delta),C\right)$ by Proposition S2.6 for some $C>0$. Hence, as $\beta_M-1\ge\beta_\perp+2\ge1$ and $|\mathrm{D}^\ell\bar f_{x_0,\delta}(\tau_{x_0,\delta}(x))|\le L_{x_0,\delta}(\tau_{x_0,\delta}(x))$ the term in (S3) is bounded by a multiple of

$$\delta^{D-d}L\circ\bar\Psi'_{x_0,\delta}(x)\times\sum_{i=1}^D|x_i-y_i|\le\delta^{D-d}L\circ\bar\Psi'_{x_0,\delta}(x)\times\sum_{i=1}^D|x_i-y_i|^{1\wedge\frac{\beta-\langle k,\boldsymbol{\alpha}\rangle}{\alpha_i}}$$

Whereas since $(\mathrm{D}^{r_i}\tau_{x_0,\delta})^{q_i}$ are bounded and $\tau_{x_0,\delta}$ is Lipschitz, the term in (S2) is upper bounded by a multiple of

$$\left|\mathrm{D}^{\ell}\bar{f}_{x_0,\delta}(\tau_{x_0,\delta}(x)) - \mathrm{D}^{\ell}\bar{f}_{x_0,\delta}(\tau_{x_0,\delta}(y))\right| \le L_{x_0,\delta}(\tau_{x_0,\delta}(x))\sum_{i=1}^{D}|\tau_i(x)-\tau_i(y)|^{1\wedge\frac{\beta-\langle k,\boldsymbol{\alpha}\rangle}{\alpha_i}}$$

As noticed above, only $(x_i)_{i\le d}$ are involved in the computation of $(\tau_i(x))_{i\le d}$. Hence, using in addition that $\tau_{x_0,\delta}$ is Lipschitz and that $\alpha_i = \alpha_0$ for every $i \le d$, we have

$$\sum_{i=1}^{d}|\tau_i(x)-\tau_i(y)|^{1\wedge\frac{\beta-\langle k,\boldsymbol{\alpha}\rangle}{\alpha_i}} \lesssim \sum_{i=1}^{d}|x_i-y_i|^{1\wedge\frac{\beta-\langle k,\boldsymbol{\alpha}\rangle}{\alpha_i}}$$

Moreover since for every $i > d$ we have $\alpha_i = \alpha_\perp \le \alpha_0$, there holds

$$\sum_{i>d}|g_i(x)-g_i(y)|^{1\wedge\frac{\beta-\langle k,\boldsymbol{\alpha}\rangle}{\alpha_i}} \lesssim \sum_{i=1}^{D}|x_i-y_i|^{1\wedge\frac{\beta-\langle k,\boldsymbol{\alpha}\rangle}{\alpha_i}}$$

which ends the proof. □

S1.3. *Partitions of unity and packings.* The approximation result uses a particular covering of an offset of the manifold M, which we describe here. Take $x_0 \in M$ and define

$$\mathcal{W}^j_{x_0} \coloneqq \mathsf{B}_{T_{x_0}}\left(0, \frac{2+j}{16}\tau\right) \times \mathsf{B}_{N_{x_0}}\left(0, \frac{6+j}{8}\tau\right) \quad \text{for } j \in \{0,1,2\},$$

and $\mathcal{O}^j_{x_0} = \bar{\Psi}_{x_0}(\mathcal{O}^j_{x_0})$ for $j \in \{0,1,2\}$. We have

$$\mathcal{W}^0_{x_0} \subset \mathcal{W}^1_{x_0} \subset \mathcal{W}^2_{x_0} \quad \text{and} \quad \mathcal{O}^0_{x_0} \subset \mathcal{O}^1_{x_0} \subset \mathcal{O}^2_{x_0}.$$

Furthermore, the sets $\mathcal{O}^0_{x_0}$ for $x_0 \in M$ forms a covering of $M^{3\tau/4}$. See Figure S1 for an illustration of these open sets.

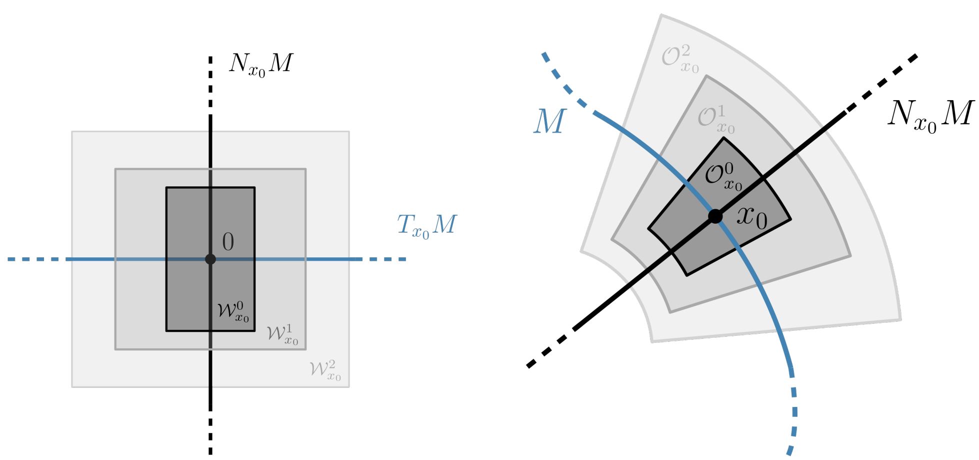


Fig S1: A visual representation of the sets used in the proof of Section 3.

In what follows, we will need the notion of packing. An ε-packing of a subset $\mathcal{A} \subset \mathbb{R}^D$ is a set $\{y_1,\dots,y_J\}$ of points of $\mathcal{A}$ such that $\|x_j - x_k\| > \varepsilon$ for any $1 \le i \ne j \le J$ and such that no set of $J+1$ points has this property. We denote by $\mathrm{pk}(\mathcal{A},\varepsilon) \coloneqq J$ the ε-packing number of $\mathcal{A}$. By maximality of J, it is straightforward to see that a $\mathcal{A}$ is covered by the union of the balls $\mathsf{B}(x_j,\varepsilon)$. Furthermore, the balls $\mathsf{B}(x_j,\varepsilon/2)$ must be disjoint by definition of a packing so that

$$\mathrm{pk}(\mathcal{A},\varepsilon)\times\min_{x\in\mathcal{A}}\mathrm{vol}(\mathcal{A}\cap\mathsf{B}(x,\varepsilon/2))\le\mathrm{vol}\left\{\mathcal{A}\cap\bigcup_{1\le j\le J}\mathsf{B}(x_j,\varepsilon/2)\right\}\le\mathrm{vol}\,\mathcal{A}.$$

In the case where $\mathcal{A}$ is a ball of radius R with R large before ε, it is straightforward to see that $\mathrm{vol}(\mathcal{A}\cap\mathsf{B}(x,\varepsilon/2))\gtrsim\varepsilon^D$ for any $x\in\mathcal{A}$ so that $\mathrm{pk}(\mathcal{A},\varepsilon)\lesssim(R/\varepsilon)^D$. We now let

$$\rho(x):=\exp\left\{-\frac{1}{(1-\|x\|^2)_+}\right\}$$

which is an infinitely differentiable, radially symmetric function from $\mathbb{R}^D$ to $[0,1]$ supported on $\mathsf{B}(0,1)$. For any $x_0\in M$, we define

$$\rho_{x_0}(x):=\rho\left(32\frac{x-x_0}{\tau}\right)$$

For any large $R>0$, one can take a $\tau/64$-packing of $M\cap\mathsf{B}(0,R)$, say $\{x_1,\dots,x_J\}$ with J of order less than $\mathrm{pk}(\mathsf{B}(0,R),\tau/64)\lesssim R^D$. We can define

$$\chi_j(x):=\frac{\rho_{x_j}(x)}{\sum_{i=1}^J\rho_{x_i}(x)}.$$

In a similar fashion as what is done in [5], we first review a few properties satisfied by the maps $\{\chi_j\}_{1\le j\le J}$, which forms a partition of unity associated with the covering $\{\mathsf{B}(x_j,\tau/32)\}_{1\le j\le J}$ of $M\cap\mathsf{B}(0,R)$. See Figure S2 for a geometric interpretation of the situation.

LEMMA S1.4. *The following assertions hold true:*

i) *For any $1\le j\le J$, $\mathrm{supp}\,\chi_j\subset\mathcal{O}^0_{x_j}$;*

ii) *There exists a numeric constant $\gamma>0$ such that $M^{\gamma\tau}\cap\mathsf{B}(0,R)\subset\mathrm{supp}\sum_j\chi_j$;*

iii) *There exists a numeric constant $\nu>0$, such that for any $x\in M^{\gamma\tau}\cap\mathsf{B}(0,R)$, there holds $\sum_{j=1}^J\rho_{x_j}(x)\ge\nu$;*

iv) *For any $|k|\le K$, there holds that $\|\mathrm{D}^k\chi_j\|_\infty\le C<\infty$ with C depending on K and τ;*

v) *For any $|k|\le K$, there exists a non-negative function I_K such that, for any $1\le j\le J$,*

$$|\mathrm{D}^k\chi_j(x)|\le I_K(x-x_j)\chi_j(x),\quad\forall x\in M^{\gamma\tau}\cap\mathsf{B}(0,R)$$

with I_K being such that for any $\omega>0$

$$\sup_{x\in M^{\gamma\tau}}(I_K(x-x_j))^\omega\chi_j(x)\le C<\infty$$

where C depends on K, τ and ω.

We stress out that the constants appearing in Lemma S1.4 do not depend on R or J at all, so that it can be applied for a family of covering $\{x_1,\dots,x_J\}$ indexed by $R\to\infty$, as it will be the case in the proofs below.

PROOF OF LEMMA S1.4. We start with proving i). Let $x\in\mathsf{B}(x_j,\tau/32)$ and let $z=\mathrm{pr}_M\,x$. There holds that

$$\|z-x_j\|^2=\|z-x\|^2+\|x-x_j\|^2+2\langle z-x,x-x_j\rangle=\|x-x_j\|^2-\|z-x\|^2+2\langle z-x,z-x_j\rangle.$$

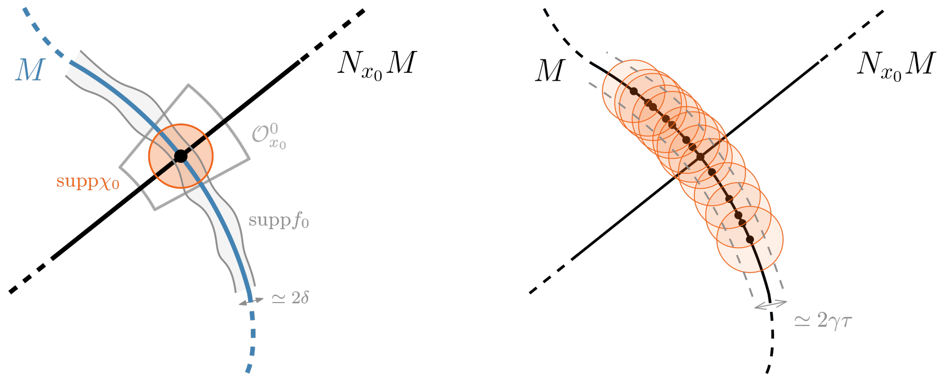


Fig S2: A visual interpretation of the framework and results of Lemma S1.4.

But now using [6, Thm 4.7 (8)], there holds that $2\langle z-x, z-x_j\rangle \le \|z-x_j\|^2\|z-x\|/\tau$ so that in the end

$$\|z-x_j\|^2 \le \frac{\|z-x\|^2}{1-\|z-x\|/\tau} \le \left\{\sqrt{\frac{32}{31}}\frac{\tau}{32}\right\}^2,$$

where we used the fact that $\|z-x\| \le \|x_j - x\| \le \tau/32$. In particular, using (27) in Lemma A.1, we obtain

$$d_M(z,x_j) \le \kappa\left(\sqrt{\frac{32}{31}}\frac{1}{16}\right)\|z-x_j\| \le \kappa\left(\sqrt{\frac{32}{31}}\frac{1}{16}\right)\sqrt{\frac{32}{31}}\frac{\tau}{32} < \tau/16$$

where the last inequality was checked with a calculator, so that $z = \Psi_{x_j}(v)$ with $v \in \mathsf{B}_{T_{x_j}M}(0,\tau/16)$. Finally $\|x-z\| \le \tau/32$ so that $x - z = N_{x_j}(v,\eta)$ for some $\eta \in \mathsf{B}_{N_{x_j}M}(0,\tau/32)$. In the, end $x = \bar{\Psi}_{x_j}(v,\eta)$ with $(v,\eta) \in \mathcal{W}^0_{x_j}$ so $x \in \mathcal{O}^0_{x_j}$, ending the proof of the assertion i).

We continue with proving ii) and iii) with $\gamma = 1/128$ and $\nu = \exp\{-16/7\}$. Take $x \in M^{\gamma\tau} \cap \mathsf{B}(0,R)$ and $z = \mathrm{pr}_M(x)$. There exists x_j such that $\|z - x_j\| \le \tau/64$, leading to $\|x-x_j\| \le 3\tau/128$ and thus $x \in \operatorname{supp}\chi_{x_j}$. Furthermore,

$$\rho_{x_j}(x) = \exp\left\{-\frac{1}{(1-(32\|x-x_j\|/\tau)^2)_+}\right\} \ge \exp\left\{-\frac{1}{(1-(3/4)^2)_+}\right\} = \nu,$$

so that points ii) and iii) are proven.

We finish with the proof of iv) and v). We let $R(x) = (\sum_{i=1}^J \rho_{x_i}(x))^{-1}$, which is well defined on $M^{\gamma\tau} \cap \mathsf{B}(0,R)$ and bounded from above by ν^{-1} so that $\chi_j(x) = R(x)\rho_{x_j}(x)$ and

$$\mathrm{D}^k\chi_j(x) = \sum_{|\ell|\le k}\binom{k}{\ell}\mathrm{D}^{k-\ell}R(x)\,\mathrm{D}^\ell\rho_{x_j}(x).$$

Outside of $\mathsf{B}(x_j,\tau/32)$ one can take $I_K = 0$. Now if $x \in \mathsf{B}(x_j,\tau/32)$, one get that for any $|\ell| < k$, using the Faa Di Bruno formula yields that $\mathrm{D}^\ell\rho_{x_j}(x)$ is a sum of term of the form

$$(-1)^s\prod_{i=1}^s \mathrm{D}^{r_i}\Upsilon(x)\times\rho_{x_j}(x) \quad \text{with} \quad \Upsilon(x) := \frac{1}{1-\frac{32^2}{\tau^2}\|x-x_j\|^2} \quad \text{and} \quad \sum_{i=1}^s r_i = \ell.$$

In particular, we get that $\mathrm{D}^\ell\rho_{x_j}(x)$ is bounded from above by some constant depending on K and τ on $\mathsf{B}(x_j,\tau/32)$. Furthermore, for any $|\ell| < k$, the derivative $\mathrm{D}^\ell R(x)$ is a sum of terms of the form

$$R(x)^r \prod_{i=1}^{s} \mathrm{D}^{r_i} R^{-1}(x) \quad \text{with} \quad \sum_{i=1}^{s} r_i \le \ell \quad \text{and} \quad r + \sum_{i=1}^{s} |r_i| = |\ell| + 2,$$

and $\mathrm{D}^{r_i} R^{-1}(x) = \sum_q \mathrm{D}^{r_i} \rho_{x_q}(x)$ which is bounded from above by something depending on τ and K. All in all, we get the uniform boundedness of $\mathrm{D}^k \chi_j$ and that

$$|\mathrm{D}^k \chi_j(x)| \le I_K(x - x_j)\chi_j(x)$$

$$\text{with} \quad I_K(x) := \begin{cases} C_{K,\tau} \left\{1 - \dfrac{32^2}{\tau^2}\|x\|^2\right\}^{-(K+1)} & \text{if } x \in \mathsf{B}(0, \tau/32), \\ 0 & \text{otherwise,} \end{cases}$$

where $C_{K,\tau}$ depends on K and τ, ending the proof. □

S2. Supplement to Section 2.

S2.1. *Auxiliary results on general anisotropic Hölder functions.* We first extend Definition 2.1 to functions defined on general open sets of $\mathbb{R}^D$. For any open set $\mathcal{U} \subset \mathbb{R}^D$, any function $L : \mathcal{U} \to \mathbb{R}_+$ and any positive real number $\zeta > 0$, we define the anisotropic Hölder spaces $\mathcal{H}^{\boldsymbol{\beta}}_{\mathrm{an}}(\mathcal{U}, L, \zeta)$ as the set of functions $f : \mathcal{U} \to \mathbb{R}^D$ satisfying that

i) For any multi-index $k \in \mathbb{N}^D$ such that $\langle k, \boldsymbol{\alpha}\rangle < \beta$ the partial derivative $\mathrm{D}^k f$ is well defined on $\mathcal{U}$ and $|\mathrm{D}^k f(x)| \le L(x)$ for all $x \in \mathcal{U}$;

ii) For any multi-index $k \in \mathbb{N}^D$ such that $\beta - \alpha_{\max} \le \langle k, \boldsymbol{\alpha}\rangle < \beta$, there holds

$$|\mathrm{D}^k f(y) - \mathrm{D}^k f(x)| \le L(x) \sum_{i=1}^{D} |y_i - x_i|^{\frac{\beta - \langle k, \boldsymbol{\alpha}\rangle}{\alpha_i} \wedge 1} \quad \forall x, y \in \mathcal{U},\ \|x - y\| \le \zeta. \tag{S4}$$

Like in Subsection 2.1, we define in a similar fashion $\mathcal{H}^{\beta}_{\mathrm{iso}}(\mathcal{U}, L, \zeta)$, $\mathcal{H}^{\boldsymbol{\beta}}_{\mathrm{an}}(\mathcal{U}, C, \zeta)$ and $\mathcal{H}^{\beta}_{\mathrm{iso}}(\mathcal{U}, C, \zeta)$ for some constant $C > 0$. We now list a number of useful results which hold for our definition of anisotropy. The proofs of these results are provided below.

PROPOSITION S2.1. *Let $f \in \mathcal{H}^{\boldsymbol{\beta}}_{\mathrm{an}}(\mathcal{U}, L, \zeta)$. Then for any $k \in \mathbb{N}^D$ such that $\langle k, \boldsymbol{\alpha}\rangle < \beta$, the partially differentiated function $\mathrm{D}^k f$ is in $\mathcal{H}^{\boldsymbol{\beta}^{(k)}}_{\mathrm{an}}(\mathcal{U}, L, \zeta)$ with*

$$\boldsymbol{\beta}^{(k)} = \left\{1 - \frac{\langle k, \boldsymbol{\alpha}\rangle}{\beta}\right\} \boldsymbol{\beta}.$$

Anisotropic Hölder functions enjoy the same convenient Taylor expansion as usual Hölder function.

PROPOSITION S2.2. *Let $f \in \mathcal{H}^{\boldsymbol{\beta}}_{\mathrm{an}}(\mathcal{U}, L, \zeta)$. Then, for any $x, y \in \mathcal{U}$ with $\|x - y\| \le \zeta$,*

$$f(y) = f(x) + \sum_{0 < \langle k, \boldsymbol{\alpha}\rangle < \beta} \frac{(y - x)^k}{k!} \mathrm{D}^k f(x) + R(x, y),$$

where the remainder R satisfies the following bound

$$|R(x, y)| \le L(x) \sum_{\beta - \alpha_{\max} \le \langle k, \alpha\rangle < \beta} \frac{|y - x|^k}{k!} \sum_{i=1}^{D} |y_i - x_i|^{\frac{\beta - \langle k, \boldsymbol{\alpha}\rangle}{\alpha_i}}.$$

PROPOSITION S2.3. *Let $f \in \mathcal{H}_{\mathrm{an}}^{\boldsymbol{\beta}}(\mathcal{U}, L, \zeta)$. Then, for any $k \in \mathbb{N}^D$ such that $\langle k, \boldsymbol{\alpha}\rangle < \beta$, and any $x, y \in \mathcal{U}$ with $\|x-y\| \le \zeta$,*

$$|\mathrm{D}^k f(y) - \mathrm{D}^k f(x)| \le C L(x) \left\{ \sum_{i=1}^{D} |y_i - x_i|^{\frac{\beta - \langle k, \boldsymbol{\alpha}\rangle}{\alpha_i} \wedge 1} \right\}.$$

with C depending on k,ζ and D.

PROPOSITION S2.4. *Let $f \in \mathcal{H}_{\mathrm{an}}^{\boldsymbol{\beta}}(\mathcal{U}, L, \zeta)$. Then, for any $\boldsymbol{\beta}' \le \boldsymbol{\beta}$, $f \in \mathcal{H}_{\mathrm{an}}^{\boldsymbol{\beta}'}(\mathcal{U}, CL, \zeta)$ with C depending on $\boldsymbol{\beta}$, ζ and D.*

PROPOSITION S2.5. *Let $f \in \mathcal{H}_{\mathrm{an}}^{\boldsymbol{\beta}_1}(\mathcal{U}_1, L_1, \zeta_1)$ and $g \in \mathcal{H}_{\mathrm{an}}^{\boldsymbol{\beta}_2}(\mathcal{U}_2, L_2, \zeta_2)$ with $\mathcal{U}_1 \subset \mathbb{R}^{D_1}$ and $\mathcal{U}_2 \subset \mathbb{R}^{D_2}$. Then*

$$f \otimes g \in \mathcal{H}_{\mathrm{an}}^{(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)}(\mathcal{U}_1 \times \mathcal{U}_2, C L_1 \otimes L_2, \zeta)$$

where $f \otimes g(x,y) = f(x)g(y)$ and $\zeta = \zeta_1 \wedge \zeta_2$, and where C depends on $D_1, D_2, \zeta, \boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$.

PROPOSITION S2.6. *Let $f \in \mathcal{H}_{\mathrm{an}}^{\boldsymbol{\beta}_1}(\mathcal{U}, L_1, \zeta)$ and $g \in \mathcal{H}_{\mathrm{an}}^{\boldsymbol{\beta}_2}(\mathcal{U}, L_2, \zeta)$. Then $f \times g \in \mathcal{H}_{\mathrm{an}}^{\boldsymbol{\beta}}(\mathcal{U}, L, \zeta)$ with $\boldsymbol{\beta} = \boldsymbol{\beta}_1 \wedge \boldsymbol{\beta}_2$, and where, for some constant C depending on ζ, $\boldsymbol{\beta}$ and D,*

$$L(x) = C L_1(x) L_2(x).$$

PROPOSITION S2.7. *Let $f \in \mathcal{H}_{\mathrm{iso}}^{\beta_1}(\mathcal{U}_1, L_1)$ and $g \in \mathcal{H}_{\mathrm{iso}}^{\beta_0}(\mathcal{U}_0, C_0)$ where g takes its value in $\mathcal{U}_1$, and where $\mathcal{U}_0$ is bounded. Assume furthermore that $\beta_0 \ge 1$. Then $f \circ g \in \mathcal{H}_{\mathrm{iso}}^{\beta}(\mathcal{U}_0, L_1 \circ g)$ where $\beta = \beta_0 \wedge \beta_1$ and where C depends on C_0, C_1 and $\mathrm{diam}\,\mathcal{U}_0$.*

PROOF OF PROPOSITION S2.1. Let $k \in \mathbb{N}^D$ such that $\langle k, \alpha\rangle < \beta$. Note that

$$\boldsymbol{\alpha} = \beta \frac{1}{\boldsymbol{\beta}} = \beta^{(k)} \frac{1}{\boldsymbol{\beta}^{(k)}}.$$

Let $\ell \in \mathbb{N}^D$ such that $\langle \ell, \boldsymbol{\alpha}\rangle < \beta^{(k)}$. Then $\langle k+\ell, \alpha\rangle < \beta$ so that by definition, $\mathrm{D}^\ell \mathrm{D}^k f = \mathrm{D}^{k+\ell} f$ exists and is bounded from above by L. If now ℓ is such that $\beta^{(k)} - \alpha_{\max} \le \langle \ell, \boldsymbol{\alpha}\rangle < \beta^{(k)}$, then $\beta - \alpha_{\max} \le \langle k+\ell, \boldsymbol{\alpha}\rangle < \beta$ and for any $x, y \in \mathcal{U}$ that are at most ζ-apart,

$$|\mathrm{D}^\ell \mathrm{D}^k f(y) - \mathrm{D}^\ell \mathrm{D}^k f(x)| \le L(x) \sum_{i=1}^{D} |y_i - x_i|^{\frac{\beta - \langle k+\ell, \boldsymbol{\alpha}\rangle}{\alpha_i} \wedge 1} = L(x) \sum_{i=1}^{D} |y_i - x_i|^{\frac{\beta^{(k)} - \langle \ell, \boldsymbol{\alpha}\rangle}{\alpha_i} \wedge 1},$$

ending the proof. □

PROOF OF PROPOSITION S2.2. We let for any $0 \le i \le D$,

$$z^{(i)} = (y_1, \ldots, y_i, x_{i+1}, \ldots, x_D) \in \mathbb{R}^D$$

so that $z^{(0)} = x$ and $z^{(D)} = y$. We prove the results recursively on the integer $N = \|\lceil \boldsymbol{\beta} - 1 \rceil\|$. If $N = 0$, then every coefficient β_i are strictly less than 1, and the results follow immediately from the definition of being $\boldsymbol{\beta}$-Hölder (there are no k that satisfies $\langle k, \boldsymbol{\alpha}\rangle < \beta$ except $k = 0$). If $N \ge 1$, we can order without loss of generality and for ease of notations

$$\beta_1 \le \cdots \le \beta_k < 1 \le \beta_{k+1} \le .. \le \beta_D$$

with $k \le D-1$ because $N \ge 1$. We write

$$f(y) - f(x) = f(z^{(k)}) - f(x) + \sum_{i=k}^{D-1} f(z^{(i+1)}) - f(z^{(i)}).$$

The first term is simply bounded from above by

$$|f(z^{(k)}) - f(x)| \le L(x) \sum_{i=1}^{k} |y_i - x_i|^{\beta/\alpha_i}.$$

Recall we write $e_i = (0, \dots, 0, 1, 0, \dots, 0)$ with 1 at the i-th position. For the other terms, we do the Taylor expansion with integral remaining term:

$$f(z^{(i)}) - f(z^{(i-1)}) = \sum_{\ell=1}^{L_i-1} \frac{(y_i - x_i)^\ell}{\ell!} \mathrm{D}^{\ell e_i} f(z^{(i-1)}) + R_i(y,x), \quad L_i = \lceil \beta_i - 1 \rceil \ge 1$$

(S5) $$R_i(y,x) = \frac{(y_i - x_i)^{L_i}}{(L_i-1)!} \int_0^1 (1-t)^{L_i-1} \mathrm{D}^{L_i e_i} f(z^{(i)}(t))\, \mathrm{d}t,$$
$$\text{with} \quad z^{(i)}(t) = t z^{(i)} + (1-t) z^{(i-1)}.$$

Note that $L_i \alpha_i = \frac{\lceil \beta_i - 1 \rceil}{\beta_i} \beta$ and for all $k = (k_{1:i}, 0) \ne 0$, $L_i\alpha_i + \langle k, \boldsymbol{\alpha} \rangle \ge \beta$. Indeed, because the coefficient of $\boldsymbol{\beta}$ are ordered, there holds

$$L_i \alpha_i + \langle k, \boldsymbol{\alpha} \rangle = \frac{\beta}{\beta_i} \left[\lceil \beta_i - 1 \rceil + \sum_{j \le i} k_j \beta_i / \beta_j \right] \ge \frac{\beta}{\beta_i} (\lceil \beta_i - 1 \rceil + 1) \ge \beta.$$

Therefore

(S6) $$\left| \mathrm{D}^{L_i e_i} f(z^{(i)}(t)) - \mathrm{D}^{L_i e_i} f(x) \right| \le L(x) \sum_{j=1}^{i} |y_j - x_j|^{\frac{\beta - L_i \alpha_i}{\alpha_j}}$$

and in particular

$$R_i(x,y) = \frac{(y_i - x_i)^{L_i}}{L_i!} \mathrm{D}^{L_i e_i} f(x) + \widetilde{R}_i(x,y)$$
$$\text{where} \quad |\widetilde{R}_i(x,y)| \le L(x) \frac{|y_i - x_i|^{L_i}}{L_i!} \sum_{j=1}^{i} |y_j - x_j|^{\frac{\beta - L_i \alpha_i}{\alpha_j}}.$$

Now we use the induction hypothesis on $f_{i,\ell} = \mathrm{D}^{\ell e_i} f$ which belongs to $\mathcal{H}^{\boldsymbol{\beta}^{(\ell e_i)}}(\mathcal{U}, L, \zeta)$ (according to Proposition S2.1):

$$f_{i,\ell}(z^{(i-1)}) - f_{i,\ell}(x) = \sum_{\substack{0 < \langle k, \boldsymbol{\alpha} \rangle < \beta - \ell \alpha_i \\ k_{i:D} = 0}} \frac{(y-x)^k}{k!} \mathrm{D}^{\ell e_i + k} f(x) + R_{i,\ell}(x,y),$$

where the remainder R satisfies the following bound

$$|R_{i,\ell}(x,y)| \le L(x) \sum_{\substack{\beta - \alpha_{\max} \le \langle k + \ell e_i, \alpha \rangle < \beta \\ k_{i:D} = 0}} \frac{|y-x|^k}{k!} \sum_{i=1}^{D} |y_j - x_j|^{\frac{\beta - \langle k + \ell e_i, \boldsymbol{\alpha} \rangle}{\alpha_j}}.$$

All in all, gathering all the developments yields

$$
\begin{aligned}
& f(z^{(i)}) - f(z^{(i-1)}) \\
&= \sum_{\ell=1}^{L_i-1} \frac{(y_i - x_i)^\ell}{\ell!} \left\{ \mathrm{D}^{\ell e_i} f(x) + \sum_{\substack{0<\langle k,\boldsymbol{\alpha}\rangle<\beta-\ell\alpha_i \\ k_{i:D}=0}} \frac{(y-x)^k}{k!} \mathrm{D}^{\ell e_i + k} f(x) + R_{i,\ell}(x,y) \right\} + R_i(y,x) \\
&= \sum_{\ell=1}^{L_i-1} \left\{ \sum_{\substack{0<\langle k+\ell e_i,\boldsymbol{\alpha}\rangle<\beta \\ k_{i:D}=0}} \frac{(y-x)^{k+\ell e_i}}{(k+\ell e_i)!} \mathrm{D}^{\ell e_i + k} f(x) + \frac{(y_i - x_i)^\ell}{\ell!} R_{i,\ell}(y,x) \right\} + R_i(y,x) \\
&= \sum_{\substack{0<\langle k,\boldsymbol{\alpha}\rangle<\beta \\ k_i \neq 0,\ k_{(i+1):D}=0}} \frac{(y-x)^k}{k!} \mathrm{D}^k f(x) + \left\{ \widetilde{R}_i(y,x) + \sum_{\ell=1}^{L_i-1} \frac{(y_i - x_i)^\ell}{\ell!} R_{i,\ell}(y,x) \right\}
\end{aligned}
$$

so that

$$
\begin{aligned}
f(y) - f(x) = & \sum_{0<\langle k,\boldsymbol{\alpha}\rangle<\beta} \frac{(y-x)^k}{k!} \mathrm{D}^k f(x) \\
& + \underbrace{\sum_{i \geq k+1} \left\{ \widetilde{R}_i(y,x) + \sum_{\ell=1}^{L_i-1} \frac{(y_i - x_i)^\ell}{\ell!} R_{i,\ell}(y,x) \right\} + R_0(y,x)}_{:=R(y,x)}
\end{aligned}
$$

with $R_0(y,x) = f(z^{(k)}) - f(x)$, and with $R(y,x)$ being exactly bounded from above by

$$
|R(y,x)| \leq L(x) \sum_{\beta-\alpha_{\max}\leq\langle k,\alpha\rangle<\beta} \frac{|y-x|^k}{k!} \sum_{j=1}^{D} |y_j - x_j|^{\frac{\beta-\langle k,\boldsymbol{\alpha}\rangle}{\alpha_j}}
$$

and Proposition S2.2 is proved. □

PROOF OF PROPOSITION S2.3. Either $\langle k,\boldsymbol{\alpha}\rangle \geq \beta - \alpha_{\max}$ and then there is nothing to show, or $\langle k,\boldsymbol{\alpha}\rangle < \beta - \alpha_{\max}$, in which case $\langle k+e_i,\boldsymbol{\alpha}\rangle < \beta$ for any i and thus $\mathrm{d}\,\mathrm{D}^k f$ is well defined. There thus exist $z \in [x,y]$ such that $\mathrm{D}^k f(y) - \mathrm{D}^k f(x) = \mathrm{d}\,\mathrm{D}^k f(z)[y-x]$ leading to

$$
|\mathrm{D}^k f(y) - \mathrm{D}^k f(x)| \leq \sum_{i=1}^{D} |\mathrm{D}^{k+e_i} f(z)| \times |y_i - x_i|.
$$

Using an induction argument, there exists $C_i > 0$ such that

$$
|\mathrm{D}^{k+e_i} f(z) - \mathrm{D}^{k+e_i} f(x)| \leq C_i L(x) \left\{ \sum_{i=1}^{D} |y_i - x_i|^{\frac{\beta-\langle k+e_i,\boldsymbol{\alpha}\rangle}{\alpha_i}\wedge 1} \right\} \leq C_i D(1\vee\zeta) L(x)
$$

leading the the right results with $C = 1 + D(1\vee\zeta)\max_{1\leq i\leq D} C_i$. □

PROOF OF PROPOSITION S2.4. Let β' be the harmonic mean of $\boldsymbol{\beta}'$ and $\boldsymbol{\alpha}' = \beta'/\boldsymbol{\beta}'$. For any $k \in \mathbb{N}^D$ such that $\langle k,\boldsymbol{\alpha}'\rangle < \beta'$, we have $\langle k,\boldsymbol{\alpha}\rangle < \beta$ so that $\mathrm{D}^k f$ is well defined and bounded from above by L. What's more, using Proposition S2.3, we have

$$
|\mathrm{D}^k f(y) - \mathrm{D}^k f(x)| \leq C L(x) \left\{ \sum_{i=1}^{D} |y_i - x_i|^{\frac{\beta-\langle k,\boldsymbol{\alpha}\rangle}{\alpha_i}\wedge 1} \right\}.
$$

Now notice that

$$\frac{\beta - \langle k, \boldsymbol{\alpha}\rangle}{\alpha_i} = \beta_i \{1 - \langle k, 1/\boldsymbol{\beta}\rangle\} \geq \beta_i' \{1 - \langle k, 1/\boldsymbol{\beta}'\rangle\} = \frac{\beta' - \langle k, \boldsymbol{\alpha}'\rangle}{\alpha_i'}$$

so that

$$|y_i - x_i|^{\frac{\beta - \langle k, \boldsymbol{\alpha}\rangle}{\alpha_i} \wedge 1} \leq (1 \vee \zeta) \times |y_i - x_i|^{\frac{\beta' - \langle k, \boldsymbol{\alpha}'\rangle}{\alpha_i'} \wedge 1}$$

yielding the result. □

PROOF OF PROPOSITION S2.5. Let $k = (k_1, k_2) \in \mathbb{N}^{D_1 + D_2}$. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and β, β_1, β_2 be the harmonic means of $\boldsymbol{\beta}, \boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Let also $\boldsymbol{\alpha} = \beta / \boldsymbol{\beta}$ and $\boldsymbol{\alpha}_i = \beta_i / \boldsymbol{\beta}_i$ for $i \in \{1, 2\}$. If $\langle k, \boldsymbol{\alpha}\rangle < \beta$, then

$$\langle k_1, 1/\boldsymbol{\beta}_1\rangle + \langle k_2, 1/\boldsymbol{\beta}_2\rangle < 1$$

so that both term is stricly less that one $\mathrm{D}^{k_1} f$ and $\mathrm{D}^{k_2} g$ are well defined and $\mathrm{D}^k (f \otimes g) = \mathrm{D}^{k_1} f \otimes \mathrm{D}^{k_2} g$ is well defined as well. Furthermore, if $x = (x_1, x_2)$ and $y = (y_1, y_2)$, is such that $\|x - y\| \leq \zeta$, then $\|x_1 - y_1\| \leq \zeta \leq \zeta_1$ and $\|x_2 - y_2\| \leq \zeta \leq \zeta_2$. Using Proposition S2.3, one find that

$$\begin{aligned}
&|\mathrm{D}^k (f \otimes g)(x) - \mathrm{D}^k (f \otimes g)(y)| \\
&\leq |\mathrm{D}^{k_2} g(x_2)| \times |\mathrm{D}^{k_1} f(x_1) - \mathrm{D}^{k_1} f(y_1)| + |\mathrm{D}^{k_1} f(y_1)| \times |\mathrm{D}^{k_2} g(x_2) - \mathrm{D}^{k_2} f(y_2)| \\
&\leq C L_2(x_2) L_1(x_1) \left\{ \sum_{i=1}^{D_1} |y_{1,i} - x_{1,i}|^{\frac{\beta_1 - \langle k_1, \boldsymbol{\alpha}_1\rangle}{\alpha_{1,i}} \wedge 1} \right\} \\
&+ C |\mathrm{D}^{k_1} f(y_1)| L_2(x_2) \left\{ \sum_{i=1}^{D_2} |y_{2,i} - x_{2,i}|^{\frac{\beta_2 - \langle k_2, \boldsymbol{\alpha}_2\rangle}{\alpha_{2,i}} \wedge 1} \right\}.
\end{aligned}$$

for some constant $C > 0$ depending on D_1, D_2, ζ and $\boldsymbol{\beta}$. Now notice first that

$$|\mathrm{D}^{k_1} f(y_1)| \leq |\mathrm{D}^{k_1} f(y_1) - \mathrm{D}^{k_1} f(x_1)| + |\mathrm{D}^{k_1} f(x_1)| \leq (CD(1 \vee \zeta) + 1) L_1(x_1),$$

and notice furthermore that of either $j \in \{1, 2\}$,

$$\frac{\beta_j - \langle k_j, \boldsymbol{\alpha}_j\rangle}{\alpha_{j,i}} = \boldsymbol{\beta}_{j,i} \{1 - \langle k_j, 1/\boldsymbol{\beta}_j\rangle\} \geq \boldsymbol{\beta}_{j,i} \{1 - \langle k, 1/\boldsymbol{\beta}\rangle\} = \frac{\beta - \langle k, \boldsymbol{\alpha}\rangle}{\alpha_i} \quad \left(= \frac{\beta - \langle k, \boldsymbol{\alpha}\rangle}{\alpha_{i+D_1}} \text{ if } j = 2 \right)$$

so that

$$\begin{aligned}
&|\mathrm{D}^k (f \otimes g)(x) - \mathrm{D}^k (f \otimes g)(y)| \\
&\leq (1 \vee \zeta) C L_2(x_2) L_1(x_1) \left\{ \sum_{i=1}^{D_1} |y_{1,i} - x_{1,i}|^{\frac{\beta - \langle k, \boldsymbol{\alpha}\rangle}{\alpha_i} \wedge 1} \right\} \\
&+ (1 \vee \zeta) C (CD(1 \vee \zeta) + 1) L_1(x_1) L_2(x_2) \left\{ \sum_{i=1}^{D_2} |y_{2,i} - x_{2,i}|^{\frac{\beta - \langle k, \boldsymbol{\alpha}\rangle}{\alpha_{i+D_1}} \wedge 1} \right\} \\
&\leq (1 \vee \zeta) C (CD(1 \vee \zeta) + 1)(L_1 \otimes L_2)(x) \left\{ \sum_{i=1}^{D_1 + D_2} |y_i - x_i|^{\frac{\beta - \langle k, \boldsymbol{\alpha}\rangle}{\alpha_i} \wedge 1} \right\}
\end{aligned}$$

ending the proof. □

PROOF OF PROPOSITION S2.6. We let $\boldsymbol{\alpha} = \beta/\boldsymbol{\beta}$ with β the harmonic mean of $\boldsymbol{\beta}$. Now for any multi-index k such that $\langle k, \boldsymbol{\alpha}\rangle < \beta$, we have $\langle \ell, \boldsymbol{\alpha}\rangle < \beta$ for any $\ell \le k$ so that $\mathrm{D}^k fg$ is indeed well-defined and

$$\mathrm{D}^k fg(x) = \sum_{\ell \le k} \binom{k}{\ell} \mathrm{D}^\ell f(x)\, \mathrm{D}^{k-\ell} g(x)$$

so that $|\mathrm{D}^k fg(x)| \le 2^{|k|} L_1(x) L_2(x)$ and

$$|\mathrm{D}^k fg(y) - \mathrm{D}^k fg(x)| \le L_2(x) \sum_{\ell \le k} \binom{k}{\ell} |\mathrm{D}^\ell f(y) - \mathrm{D}^\ell f(x)| + L_1(x) \sum_{\ell \le k} \binom{k}{\ell} |\mathrm{D}^\ell g(y) - \mathrm{D}^\ell g(x)|.$$

Using now Proposition S2.3, we find that there exists $C > 0$ depending on $\boldsymbol{\beta}$, D and ζ such that

$$\begin{aligned}|\mathrm{D}^k fg(y) - \mathrm{D}^k fg(x)| &\le 2C L_1(x) L_2(x) \sum_{\ell \le k} \binom{k}{\ell} \sum_{i=1}^{D} |y_i - x_i|^{\frac{\beta - \langle \ell, \boldsymbol{\alpha}\rangle}{\alpha_i} \wedge 1} \\ &\le 2^{|k|+1} (1 \vee \zeta) C L_1(x) L_2(x) \sum_{i=1}^{D} |y_i - x_i|^{\frac{\beta - \langle k, \boldsymbol{\alpha}\rangle}{\alpha_i} \wedge 1}\end{aligned}$$

where we used again that

$$|y_i - x_i|^{\frac{\beta - \langle \ell, \boldsymbol{\alpha}\rangle}{\alpha_i} \wedge 1} \le (1 \vee \zeta) \times |y_i - x_i|^{\frac{\beta - \langle k, \boldsymbol{\alpha}\rangle}{\alpha_i} \wedge 1},$$

for $\ell \le k$, ending the proof. □

PROOF OF PROPOSITION S2.7. A simple use of the multivariate Faa Di Bruno formula [3] yields that, for any $k \in \mathbb{N}^D$ with $|k| \le \beta$, $\mathrm{D}^k (f \circ g)(x)$ for $x \in \mathcal{U}_0$ is a sum a term of the form

$$\mathrm{D}^\ell f(g(x)) \prod_{j=1}^{s} (\mathrm{D}^{k^{(j)}} g(x))^{\ell^{(j)}}$$

with $|\ell| \le |k|$, $s \le |k|$, $\sum_j \ell^{(j)} = \ell$, and $\sum_j |\ell^{(j)}| k^{(j)} = k$. In particular, it is bounded by $C L_1(g(x))$ where C depends on C_0 and β. Likewise, a telescopic argument would yield that

$$|\mathrm{D}^k (f \circ g)(y) - \mathrm{D}^k (f \circ g)(x)| \le C L_1(g(x)) \prod_{i=1}^{D} |x - y|^{(\beta - |k|) \wedge 1}$$

where C depends on C_0, β and $\mathrm{diam}\,\mathcal{U}_0$, ending the proof. □

S2.2. *Proofs associated to the examples of Proposition 2.3.*

PROOF OF PROPOSITION 2.3. Take $x_0 \in M$. In the orthonormal noise model, the density f takes the very simple form

$$\begin{aligned}\bar{f}_{x_0,\delta}(v, \eta) &= \delta^{D-d} f \circ \bar{\Psi}_{x_0,\delta}(v, \eta) = \delta^{D-d} \times f_0(\Psi_{x_0}(v)) \times \delta^{-(D-d)} c_\perp K\left(\frac{1}{\delta} N_{x_0}(v, \delta\eta)\right) \\ &= f_{x_0}(v) \times c_\perp K(\eta)\end{aligned}$$

where we used the isotropy of K. Thus, $\bar{f}_{x_0,\delta}$ lies in $\mathcal{H}^{\boldsymbol{\beta}}_{\mathrm{an}}(\mathcal{W}_{x_0,\delta}, L_{x_0,\delta})$ where L was defined in the statement of Proposition 2.3, through Proposition S2.5 and the isotropy of $L_\perp$.

In the isotropic noise model, one can write

$$\bar{f}_{x_0,\delta}(v,\eta) = \delta^{D-d}\int_M \delta^{-D} K\left(\frac{\bar{\Psi}_{x_0,\delta}(v,\eta)-x}{\delta}\right) f_0(x)\,\mathrm{d}\mu(x).$$

If $\|\eta\| \geq 1$, the integrand is trivially 0, so one may focus on $\|\eta\| \leq 1$. Now if $x \in M$ is at least 2δ apart from $\Psi_{x_0}(v)$, there holds

$$\|x - \bar{\Psi}_{x_0,\delta}(v,\eta)\| \geq \|x - \Psi_{x_0}(v)\| - \delta \geq \delta$$

so that the integrand in the integral above is zero for x outside of $B(\Psi_{x_0}(v), 2\delta)$. Doing the variable change $x = \Psi_{x_0}(w)$, we can write

$$\begin{aligned}\bar{f}_{x_0,\delta}(v,\eta) &= \frac{1}{\delta^d}\int_{\exp_{x_0}^{-1}\mathsf{B}(\Psi_{x_0}(v),2\delta)} K\left(\frac{\bar{\Psi}_{x_0,\delta}(v,\eta)-\Psi_{x_0}(w)}{\delta}\right) f_{x_0}(w)|\det \mathrm{d}\Psi_{x_0}(w)|\,\mathrm{d}w \\ &= \int_{\mathcal{Z}_\delta} \underbrace{K\left(\frac{\Psi_{x_0}(v)-\Psi_{x_0}(v+\delta s)}{\delta} + N_{x_0}(v,\eta)\right)}_{=:K\circ g_{\delta,s}(v,\eta)} \underbrace{f_{x_0}(v+\delta s)|\det \mathrm{d}\Psi_{x_0}(v+\delta s)|}_{=:h_{\delta,s}(v)}\,\mathrm{d}s\end{aligned}$$

where

$$\mathcal{Z}_\delta = \frac{1}{\delta}\left\{\exp_{x_0}^{-1}\mathsf{B}(\Psi_{x_0}(v),2\delta) - v\right\}.$$

Now notice that, using the first inequality of (27), we have $\mathsf{B}(\Psi_{x_0}(v),2\delta) \subset \mathsf{B}(x_0, \tau/8+2\delta) \subset \mathsf{B}(x_0, 3\tau/16)$, so that, using the second inequality of (27) with $\gamma = 3/8$, we find

$$\exp_{x_0}^{-1}\mathsf{B}(\Psi_{x_0}(v),2\delta) \subset \mathsf{B}_{T_{x_0}M}(0,\kappa(3/8)3\tau/16) \subset \mathsf{B}_{T_{x_0}M}(0,\tau/4)$$

so that in particular, according to (26), $\exp_{x_0}^{-1}$ is $16/11$-Lipschitz on $\mathsf{B}(\Psi_{x_0}(v),2\delta)$ and consequently, $\mathcal{Z}_\delta \subset \mathsf{B}_{T_{x_0}M}(0,32/11)$. Furthermore, notice that

$$v + \delta\mathsf{B}_{T_{x_0}M}(0,32/11) \subset \mathsf{B}_{T_{x_0}M}(0,19\tau/88)$$

with $19\tau/88$ being smaller than the injectivity radius, so that finally

$$\bar{f}_{x_0,\delta}(v,\eta) = \int_{\mathsf{B}_{T_{x_0}M}(0,32/11)} K\circ g_{\delta,s}(v,\eta)\times h_{\delta,s}(v)\,\mathrm{d}s \tag{S7}$$

It is straighforward to see that

$$v \mapsto \frac{\Psi_{x_0}(v)-\Psi_{x_0}(v+\delta s)}{\delta}$$

has derivatives bounded from above, in virtue of Proposition S2.3, by something depending on C_M and β_M only (and not on δ) and up to order $\lceil \beta_M - 1\rceil \geq \beta_\perp$. Furthermore, $(v,\eta) \mapsto N_{x_0}(v,\eta)$ is (β_M-1)-Hölder by construction. It so happens that $g_{\delta,s} \in \mathcal{H}^{\beta_\perp}_{\mathrm{iso}}(\mathcal{W}_{x_0,\delta}, C_g)$ for some g depending on C_M and β_M. Using Proposition S2.7, we get that $K\circ g_{\delta,s} \in \mathcal{H}^{\beta_\perp}_{\mathrm{iso}}(\mathcal{W}_{x_0,\delta}, C_1 L_\perp \circ g_{\delta,s})$ for some constant C_1 depending on C_g, C_K and $\beta_\perp$. Similar reasoning and the use and Proposition S2.6 would lead to $h_{\delta,s} \in \mathcal{H}^{\beta_0}_{\mathrm{iso}}(\mathcal{V}_{x_0}, C_2 L_0 \circ \Psi_{x_0} \circ \tau_{\delta s})$ where $\tau_{\delta s}(v) = v + \delta s$ and for some constant C_2 depending on C_M, τ and β_0. Now seeing $h_{\delta,s}$ as a function of (v,η) would trivially lead to $h_{\delta,s} \in \mathcal{H}^{\boldsymbol{\beta}}_{\mathrm{an}}(\mathcal{W}_{x_0,\delta}, C_2 L_0 \circ \Psi_{x_0} \circ \tau_{\delta s} \circ \pi_1)$ with $\boldsymbol{\beta} = (\beta_0,\dots,\beta_0,\beta_\perp,\dots,\beta_\perp)$ and $\pi_1(v,\eta) = v$, and an application of Proposition S2.6 together with the observation that $\beta_0 \leq \beta_\perp$ yields

$$K\circ g_{\delta,s}\times h_{\delta,s} \in \mathcal{H}^{\boldsymbol{\beta}}_{\mathrm{an}}(\mathcal{W}_{x_0,\delta}, CL_\perp \circ g_{\delta,s} \times L_0 \circ \Psi_{x_0} \circ \tau_{\delta s} \circ \pi_1)$$

for some constant C depending on C_2, C_1 and $\beta_0, \beta_\perp$. We conclude by using the linearity of the integral, so that integrating (S7) gives $\bar{f}_{x_0,\delta} \in \mathcal{H}_{\mathrm{an}}^{\boldsymbol{\beta}}(\mathcal{W}_{x_0,\delta}, \widetilde{L})$, where

$$\begin{aligned}
\widetilde{L}(v,\eta) &= C \int_{\mathsf{B}_{T_{x_0}M}(0,32/11)} L_\perp \circ g_{\delta,s}(v,s) \times L_0 \circ \Psi_{x_0}(v+\delta s)\,\mathrm{d}s \\
&\lesssim \int_{\mathsf{B}_{T_{x_0}M}(0,32/11)} L_\perp \circ g_{\delta,s}(v,s) \times L_0 \circ \Psi_{x_0}(v+\delta s) \times |\det \mathrm{d}\Psi_{x_0}(v+\delta s)|\,\mathrm{d}s \\
&= \delta^{-d} \int_M L_\perp\left(\frac{\bar{\Psi}_{x_0,\delta}(v,\eta) - x}{\delta}\right) \times L_0(x)\,\mathrm{d}\mu_M(x)
\end{aligned}$$

so that

$$L(y) = C\delta^{-D} \int_M L_\perp\left(\frac{y-x}{\delta}\right) \times L_0(x)\,\mathrm{d}\mu_M(x)$$

where L was defined in the statement of Proposition 2.3, and where we used Lemma A.2 to justify the introduction of $|\det \mathrm{d}\Psi_{x_0}(v+\delta s)|$. This ends the proof. □

PROOF OF LEMMA 3.3. In the orthogonal noise model, recall (see proof of Proposition 2.3) that

$$\begin{cases} f_0(x) = f_*(\mathrm{pr}_M(x)) \times c_\perp \delta^{-(D-d)} K((x - \mathrm{pr}_M x)/\delta) \\ L(x) = C\delta^{-(D-d)} L_0(\mathrm{pr}_M(x)) \times L_\perp((x-\mathrm{pr}_M x)/\delta), \end{cases}$$

for some $C > 0$. There thus holds, letting $\mathcal{O}_{x_0} = \bar{\Psi}_{x_0,1}(\mathcal{W}_{x_0,1})$, and using Cauchy-Schwartz inequality

$$\begin{aligned}
&\left\{\int_{\mathcal{O}_{x_0}} \left\{\frac{L(x)}{f_0(x)}\right\}^{\omega^*/2} f_0(x)\,\mathrm{d}x\right\}^2 \\
&\lesssim \int_{\mathcal{O}_{x_0}} \left\{\frac{L_0(\mathrm{pr}_M x)}{f_*(\mathrm{pr}_M x)}\right\}^{\omega^*} f_0(x)\,\mathrm{d}x \int_{\mathcal{O}_{x_0}} \left\{\frac{L_\perp((x-\mathrm{pr}_M x)/\delta)}{K((x-\mathrm{pr}_M x)/\delta)}\right\}^{\omega^*} f_0(x)\,\mathrm{d}x.
\end{aligned}$$

The first integral in the RHS is simply, up to the constant $c_\perp$,

$$\int_{\Psi_{x_0}(\mathcal{V}_{x_0})} \left\{\frac{L_0(z)}{f_*(z)}\right\}^{\omega^*} f_*(z)\,\mathrm{d}\mu_M(z),$$

which is bounded by assumption. The second integral is exactly

$$\int_{\Psi_{x_0}(\mathcal{V}_{x_0})} f_*(z)\,\mathrm{d}\mu_M(z) \times \int_{\mathsf{B}_{\mathbb{R}^{D-d}}(0,1)} \left\{\frac{L_\perp(\eta)}{K(\eta)}\right\}^{\omega^*} K(\eta)\,\mathrm{d}x.$$

which is also bounded by assumption. Hence (15) holds true in this case.

In the isotropic noise model, there holds this time

$$\begin{cases} f_0(x) = \delta^{-D} \int_{M\cap\mathsf{B}(x,\delta)} K((y-x)/\delta) f_*(y)\,\mathrm{d}\mu_M(y) \\ L(x) = C\delta^{-D} \int_{M\cap\mathsf{B}(x,\delta)} L_\perp((y-x)/\delta) L_0(y)\,\mathrm{d}\mu_M(y), \end{cases}$$

for some $C > 0$. Using the fact that for any integrable function $\phi, \psi \geq 0$ and any $\omega \geq 1$, there holds

$$\left\{\int \phi\right\}^\omega \leq \left\{\int \frac{\phi^\omega}{\psi^{\omega-1}}\right\} \times \left\{\int \psi\right\}^{\omega-1},$$

(which is simple consequence of the Hölder inequality), one find that

$$\begin{aligned}
&\int_{\mathcal{O}_{x_0}} \left\{\frac{L(x)}{f_0(x)}\right\}^{\omega^*/2} f_0(x)\,\mathrm{d}x \\
&\lesssim \int_{\mathcal{O}_{x_0}} \int_M \left\{\frac{L_\perp((y-x)/\delta)L_0(y)}{K((y-x)/\delta)f_*(y)}\right\}^{\omega^*/2} \delta^{-D} K((y-x)/\delta) f_*(y)\,\mathrm{d}\mu_M(y)\,\mathrm{d}x \\
&\leq \delta^{-D} I_1^{1/2} I_2^{1/2},
\end{aligned}$$

where

$$\begin{aligned}
I_1 &:= \int_{\mathcal{O}_{x_0}} \int_M \left\{\frac{L_0(y)}{f_*(y)}\right\}^{\omega^*} K((y-x)/\delta) f_*(y)\,\mathrm{d}\mu_M(y)\,\mathrm{d}x \\
&\lesssim \delta^D \int_M \left\{\frac{L_0(y)}{f_*(y)}\right\}^{\omega^*} f_*(y)\,\mathrm{d}\mu_M(y) \lesssim \delta^D,
\end{aligned}$$

by assumption, and where

$$\begin{aligned}
I_2 &:= \int_{\mathcal{O}_{x_0}} \int_M \left\{\frac{L_\perp((y-x)/\delta)}{K((y-x)/\delta)}\right\}^{\omega^*} K((y-x)/\delta) f_*(y)\,\mathrm{d}\mu_M(y)\,\mathrm{d}x \\
&\lesssim \delta^D \int_M f\,\mathrm{d}\mu_M \int_{\mathsf{B}(0,1)} \left\{\frac{L_\perp(\eta)}{K(\eta)}\right\}^{\omega^*} K(\eta)\,\mathrm{d}\eta,
\end{aligned}$$

so that indeed $\delta^{-D} I_1^{1/2} I_2^{1/2} \lesssim 1$, which ends the proof. □

S2.3. *Proof of Proposition 2.4*. If $Q_2 = Q_0^{\otimes D}$ with $Q_0 \sim \mathrm{DP}(BH_\lambda) := \Pi_{\mathrm{DP}}$ then, for $0 < x_1 \leq x_2/2$ and $x_1 < 1$, letting $A_1 = [x_1, x_1(1+x_1^b)]$, $A_2 = [x_2, x_2(1+x_1^b)]$ and $A_3 = (A_1 \cup A_2)^c$, we have

$$(Q_0(A_1), Q_0(A_2), Q_0(A_3)) \sim \mathcal{D}(BH_\lambda(A_1), BH_\lambda(A_2), BH_\lambda(A_3)),$$

and

$$Q_2(A_1^d \times A_2^{D-d}) = Q_0(A_1)^d Q_0(A_2)^{D-d}.$$

We let $\gamma = \Gamma(B)/(\Gamma(BH_\lambda(A_1))\Gamma(BH_\lambda(A_2))\Gamma(BH_\lambda(A_3))$. Noticing that when x_1, x_2 are small, $H_\lambda(A_3) = 1 - o(1)$, there holds, letting $i = 1$ if $x_1 \leq x_2$ and $i = 2$ otherwise,

$$\begin{aligned}
&\tilde{\Pi}_\Lambda\left[Q_2\left(A_1^d A_2^{D-d}\right) \geq x_1^{B_0}\right] \geq \Pi_{DP}\left(Q_0(A_1) > x_1^{B_0/D}, Q_0(A_2) > x_1^{B_0/D}\right) \\
&\geq \frac{\gamma}{2^{(B-1)_+}} \int_{x_1^{B_0/D}}^{1/4} x^{BH_\lambda(A_1)-1} dx \int_{x_1^{B_0/D}}^{1/4} x^{BH_\lambda(A_2)-1} dx \\
&\geq \frac{\Gamma(B)}{(1+o(1))2^{(B-1)_+}} \left[4^{-BH_\lambda(A_1)} - x_1^{\frac{BH_\lambda(A_1)B_0}{D}}\right]\left[4^{-BH_\lambda(A_2)} - x_1^{\frac{BH_\lambda(A_2)B_0}{D}}\right] \\
&\geq C_1\left[1 - C_2 e^{-\frac{BH_\lambda(A_i)B_0}{D}}\right]^2,
\end{aligned}$$

for some constant $C_1, C_2 > 0$, where we used the fact that for x_1 and x_2 small, $H_\lambda(A_i) = \min\{H_\lambda(A_1), H_\lambda(A_2)\}$ and $H_\lambda(A_1)$ and $H_\lambda(A_2)$ are both small. Under the assumption on H_λ near 0,

$$H_\lambda(A_i) \gtrsim e^{-c_2 x_i^{-1/4}} x_i^{1+b} \gtrsim e^{-c_2' x_i^{-1/4}}$$

for some $c_2' > 0$, which in turn implies

$$\tilde{\Pi}_\Lambda\big[Q_2\left(A_1^d A_2^{D-d}\right)\big] \gtrsim e^{-c_2' x_i^{-1/4}} \geq e^{-c_2' x_1^{-d/2}}$$

where we used that $x_i^{-1/4} \leq x_1^{-1/2} \leq x_i^{-d/2}$, so that (12) holds. Condition (13) is verified similarly: first note that $Q_2(\min_{i\leq D}\lambda_i \leq x) \leq DQ_0((0,x])$ and using the fact that $Q_0((0,x])$ follows a Beta variable with parameters $(BH_\lambda((0,x]), B(1+o(1)))$ we obtain that

$$\mathbb{E}_{\tilde{\Pi}_\Lambda}\left[Q_2\left(\min_{i\leq D}\lambda_i \leq x\right)\right] \lesssim H_\lambda((0,x]) \lesssim e^{-c_3 x^{-b_3}},$$

for small x. Similar computation terminates the proof of (13).

S3. Supplement to Section 5.2: proof and framework of Theorem 3.4.

S3.1. *Technical Lemmata.* We let $\chi_1,\dots,\chi_J$ be the partition of unity defined in Section S1.3 associated with a $\tau/64$-packing of $\{x_1,\dots,x_J\}$ of $M \cap \mathsf{B}(0,R)$. We recall that $J \lesssim R^D$ and we take $R = (H\log(1/\sigma))^{1/\kappa}$. We define

$$f_j := c_j^{-1}\chi_j \times f_0 \quad \text{where} \quad c_j = \int_{\mathbb{R}^D}\chi_j(x) f_0(x)\,\mathrm{d}x.$$

Before we give the proof of Corollary 3.5 we need a few technical Lemmata.

LEMMA S3.1. *For all $1 \leq j \leq J$, $f_j \in \mathcal{H}_\delta^{\beta_0,\beta_\perp}(M, L_j)$ with $L_j(x) := Cc_j^{-1} I_K(x - x_j)\chi_j(x)L(x)$ where $C > 0$ is a constant depending on τ and where we took $K = \lceil \beta_\perp \rceil$.*

PROOF. The proof is a direct consequence of Proposition S2.6 and Proposition S2.7 applied to $\chi_j(x)$, along with the assertion iv) from Lemma S1.4. □

We let $L_{j,\delta} := \delta^{D-d} L_j \circ \bar{\Psi}_{x_j,\delta}$ and

$$\bar{f}_{j,\delta} := f_j \circ \bar{\Psi}_{x_j,\delta} : \mathcal{W}_{x_j,\delta} \to \mathbb{R}^D.$$

and also introduce the functions $g_j : \mathbb{R}^D \to \mathbb{R}$ from the proof of Theorem 3.4. Recall that these functions are of the form

$$g_j(x) := f_j(x) + \frac{1}{\delta^{D-d}} \sum_{0<\langle k,\boldsymbol{\alpha}\rangle<\beta} \sigma^{\langle k,\boldsymbol{\alpha}\rangle} d_{j,k}(x,\sigma,\delta)\,\mathrm{D}^k \bar{f}_{j,\delta}(z_x) \quad \text{where} \ \ z_x := \Delta_{1,\delta}^{-1}\bar{\Psi}_{x_j}^{-1}(x)$$

and satisfy

1. They are supported on $\mathcal{O}_{x_j}^0$;
2. The functions $d_{j,k}$ are uniformly bounded by a constant C depending on C_M;
3. $|K_\Sigma g_j(x) - f_j(x)| \lesssim \sigma^\beta L_j(x)$ on $\mathcal{O}_{x_j}^1$;
4. For all $H > 0$, $|K_\Sigma g_j(x) - f_j(x)| \lesssim \sigma^H \|L_j\|_\infty$ outside of $\mathcal{O}_{x_j}^1$.

LEMMA S3.2. *Under (15), there holds, for any $\varepsilon \in (0, \omega - 2\beta)$, any $\langle k,\boldsymbol{\alpha}\rangle < \beta$ and any $1 \leq j \leq J$,*

$$\int \left(\frac{|\mathrm{D}^k \bar{f}_{j,\delta}(z)|}{\bar{f}_{j,\delta}(z)}\right)^{\frac{2\beta+\varepsilon}{\langle k,\boldsymbol{\alpha}\rangle}} \bar{f}_{j,\delta}(z)\,\mathrm{d}z \lesssim c_j^{-1} \quad \textit{and} \quad \int \left(\frac{L_{j,\delta}(z)}{\bar{f}_{j,\delta}(z)}\right)^{\frac{2\beta+\varepsilon}{\beta}} \bar{f}_{j,\delta}(z)\,\mathrm{d}z \lesssim c_j^{-1},$$

up to a constant depending on the parameters.

PROOF. We denote by $\bar{\chi}_{j,\delta} = \chi_j \circ \bar{\Psi}_{x_j,\delta}$ and $\bar{I}_{j,\delta} := I_K(\bar{\Psi}_{x_j,\delta}(\cdot) - x_j)$. Writing that

$$\left|\frac{\mathrm{D}^k \bar{f}_{j,\delta}}{\bar{f}_{j,\delta}}\right| \le \sum_{\ell \le k} \binom{k}{\ell} \left|\frac{\mathrm{D}^{k-\ell} \bar{\chi}_{j,\delta}}{\bar{\chi}_{j,\delta}}\right| \times \left|\frac{\mathrm{D}^\ell \bar{f}_{x_j,\delta}}{\bar{f}_{x_j,\delta}}\right| \lesssim \bar{I}_{j,\delta} \sum_{\ell \le k} \left|\frac{\mathrm{D}^\ell \bar{f}_{x_j,\delta}}{\bar{f}_{x_j,\delta}}\right|$$

we easily see that

$$\begin{aligned}\int \left|\frac{\mathrm{D}^k \bar{f}_{j,\delta}}{\bar{f}_{j,\delta}}\right|^{\frac{2\beta+\varepsilon}{\langle k,\boldsymbol{\alpha}\rangle}} \bar{f}_{j,\delta} &\lesssim c_j^{-1} \sum_{\ell \le k} \int \left\{ \bar{I}_{j,\delta} \left|\frac{\mathrm{D}^\ell \bar{f}_{x_j,\delta}}{\bar{f}_{x_j,\delta}}\right| \right\}^{\frac{2\beta+\varepsilon}{\langle k,\boldsymbol{\alpha}\rangle}} \bar{\chi}_{j,\delta} \bar{f}_{x_j,\delta} \\ &\lesssim c_j^{-1} \sum_{\ell \le k} \int \left|\frac{\mathrm{D}^\ell \bar{f}_{x_j,\delta}}{\bar{f}_{x_j,\delta}}\right|^{\frac{2\beta+\varepsilon}{\langle k,\boldsymbol{\alpha}\rangle}} \bar{f}_{x_j,\delta} \lesssim c_j^{-1}\end{aligned}$$

where we used both Lemma S1.4 iv) and (15) with the fact that $\langle \ell, \boldsymbol{\alpha}\rangle \le \langle k, \boldsymbol{\alpha}\rangle$ for $\ell \le k$. The bound on the second integral follows from the same line of reasoning. □

LEMMA S3.3. *We let $\mathcal{A}_{j,\sigma}$ be the event of all $x \in \mathcal{O}^2_{x_j}$ such that*

$$\forall 0 < \langle k, \boldsymbol{\alpha}\rangle < \beta, \frac{|\mathrm{D}^k \bar{f}_{j,\delta}(z_x)|}{\bar{f}_{j,\delta}(z_x)} \le \sigma^{-\langle k,\alpha\rangle} \log(1/\sigma)^{-\langle k,\alpha\rangle/2} \ \textit{and}\ \frac{L_{j,\delta}(z_x)}{\bar{f}_{j,\delta}(z_x)} \le \sigma^{-\beta} \log(1/\sigma)^{-\beta/2}.$$

Then, the following assertions hold true for σ small enough

i) $g_j(x) \ge f_j(x)/2$ *for all* $x \in \mathcal{A}_{j,\sigma}$;

ii) $P_{f_j}(\mathcal{A}^c_{j,\sigma}) \lesssim c_j^{-1} \{\sigma \log(1/\sigma)\}^{2\beta+\varepsilon}$;

iii) *For any* $\langle k, \boldsymbol{\alpha}\rangle < \beta$, $\int_{\mathcal{A}^c_{j,\sigma}} |\mathrm{D}^k \bar{f}_{j,\delta}(z_x)| \,\mathrm{d}x \lesssim c_j^{-1} \delta^{D-d} \{\sigma \log(1/\sigma)\}^{2\beta+\varepsilon-\langle k,\boldsymbol{\alpha}\rangle}$.

PROOF. We start by proving i). For any $x \in \mathcal{A}_{j,\sigma}$, there holds, using the fact that the functions $d_{j,k}$ are uniformly bounded,

$$\begin{aligned}|g_j(x) - f_j(x)| &= \left|\frac{1}{\delta^{D-d}} \sum_{0<\langle k,\alpha\rangle<\beta} \sigma^{\langle k,\alpha\rangle} d_k(y,\sigma,\delta)\, \mathrm{D}^k \bar{f}_{j,\delta}(z_y)\right| \\ &\lesssim \frac{\bar{f}_{j,\delta}(z_x)}{\delta^{D-d}} \sum_{0<\langle k,\alpha\rangle<\beta} \sigma^{\langle k,\alpha\rangle} \left|\frac{\mathrm{D}^k \bar{f}_{j,\delta}(z_x)}{\bar{f}_{j,\delta}(z_x)}\right| \\ &\lesssim f_j(x) \sum_{0<\langle k,\alpha\rangle<\beta} \log(1/\sigma)^{-\langle k,\alpha\rangle/2} \lesssim \log^{-\alpha_\perp/2}(1/\sigma) f_j(x) < \frac{1}{2} f_j(x)\end{aligned}$$

provided that σ is chosen small enough. For ii), notice that

$$\begin{aligned}P_{f_j}(\mathcal{A}^c_{j,\sigma}) &\le \sum_{0<\langle k,\boldsymbol{\alpha}\rangle<\beta} P_{f_j}\left(\frac{|\mathrm{D}^k \bar{f}_{j,\delta}(z_x)|}{\bar{f}_{j,\delta}(z_x)} > \sigma^{-\langle k,\alpha\rangle} \log(1/\sigma)^{-\langle k,\alpha\rangle/2}\right) \\ &\le \sum_{0<\langle k,\boldsymbol{\alpha}\rangle<\beta} \sigma^{2\beta+\varepsilon} \log(1/\sigma)^{2\beta+\varepsilon} \int_{\mathcal{O}^0_{x_j}} \left(\frac{|\mathrm{D}^k \bar{f}_{j,\delta}(z_x)|}{\bar{f}_{j,\delta}(z_x)}\right)^{\frac{2\beta+\varepsilon}{\langle k,\boldsymbol{\alpha}\rangle}} f_j(x)\,\mathrm{d}x \\ &\lesssim c_j^{-1} \sigma^{2\beta+\varepsilon} \log(1/\sigma)^{2\beta+\varepsilon}\end{aligned}$$

where we used Lemma S3.2 after a variable change $z = z_x$. Finally, for iii), there holds

$$\int_{\mathcal{A}_{j,\sigma}^c} |\mathrm{D}^k \bar{f}_{j,\delta}(z_x)| \,\mathrm{d}x = \int \frac{|\mathrm{D}^k \bar{f}_{j,\delta}(z_x)|}{\bar{f}_{j,\delta}(z_x)} \bar{f}_{j,\delta}(z_x) \mathbb{1}_{\mathcal{A}_{j,\sigma}^c}(x)\,\mathrm{d}x$$

$$\leq \left\{ \int \left\{ \frac{|\mathrm{D}^k \bar{f}_{j,\delta}(z_x)|}{\bar{f}_{j,\delta}(z_x)} \right\}^{\frac{2\beta+\varepsilon}{\langle k,\boldsymbol{\alpha}\rangle}} \bar{f}_{j,\delta}(z_x) \mathbb{1}_{\mathcal{A}_{j,\sigma}^c}(x)\,\mathrm{d}x \right\}^{\frac{\langle k,\boldsymbol{\alpha}\rangle}{2\beta+\varepsilon}} \times \left\{ \delta^{D-d} P_{f_j}(\mathcal{A}_{j,\sigma}^c) \right\}^{1-\frac{\langle k,\boldsymbol{\alpha}\rangle}{2\beta+\varepsilon}}$$

$$\lesssim \left\{ c_j^{-1} \delta^{D-d} \right\}^{\frac{\langle k,\boldsymbol{\alpha}\rangle}{2\beta+\varepsilon}} \times \left\{ c_j^{-1} \delta^{D-d} \left\{ \sigma \log(1/\sigma) \right\}^{2\beta+\varepsilon} \right\}^{1-\frac{\langle k,\boldsymbol{\alpha}\rangle}{2\beta+\varepsilon}}$$

$$\lesssim c_j^{-1} \delta^{D-d} \left\{ \sigma \log(1/\sigma) \right\}^{2\beta+\varepsilon-\langle k,\boldsymbol{\alpha}\rangle}.$$

ending the proof. □

S3.2. *Proof of Corollary 3.5.* Let define

$$\widetilde{h}_j := g_j \mathbb{1}_{g_j > f_j/2} + \frac{1}{2} f_j \mathbb{1}_{g_j \leq f_j/2}$$

and $h_j = \widetilde{h}_j / I_j$ where $I_j = \int \widetilde{h}_j$. The function h_j is a probability measure on $\mathcal{O}_{x_j}^0$ and furthermore, using the convexity of $f \to \mathrm{d}_{\mathrm{H}}^2(f, f_0)$, there holds

$$\mathrm{d}_{\mathrm{H}}^2(K_\Sigma(h), f_0) \leq \sum_{j=1}^J c_j \,\mathrm{d}_{\mathrm{H}}^2(K_\Sigma(h_j), f_j) \quad \text{where } h = \sum_{j=1}^J c_j h_j.$$

We will control each term separately. First notice that

$$\sum_{j=1}^J c_j \,\mathrm{d}_{\mathrm{H}}^2(K_\Sigma(h_j), f_j) \leq J\sigma^{2\beta} + \sum_{c_j \geq \sigma^{2\beta}} c_j \,\mathrm{d}_{\mathrm{H}}^2(K_\Sigma(h_j), f_j).$$

Now take $1 \leq j \leq J$ such that $c_j \geq \sigma^{2\beta}$ and define $\mathcal{U}_{j,\sigma} = \left\{ x \in \mathbb{R}^D \mid f_j(x) \geq \sigma^{H_1} \right\}$ for some $H_1 > 0$ to be specified later. Then there holds

$$\mathrm{d}_{\mathrm{H}}^2(K_\Sigma h_j, f_j) \leq \int_{\mathcal{U}_{j,\sigma} \cap \mathcal{O}_{x_j}^1} \left(\sqrt{K_\Sigma(h_j)} - \sqrt{f_j} \right)^2 + \int_{(\mathcal{O}_{x_j}^1)^c} K_\Sigma(h_j) + \int_{\mathcal{U}_{j,\sigma}^c} [K_\Sigma(h_j) + f_j]$$

and

$$\int_{\mathcal{U}_{j,\sigma} \cap \mathcal{O}_{x_j}^1} \left(\sqrt{K_\Sigma(h_j)} - \sqrt{f_j} \right)^2 \leq \int_{\mathcal{U}_{j,\sigma} \cap \mathcal{O}_{x_j}^1} \frac{(K_\Sigma(h_j) - f_j)^2}{K_\Sigma(h_j) + f_j}$$

$$\leq \int_{\mathcal{U}_{j,\sigma} \cap \mathcal{O}_{x_j}^1} \frac{K_\Sigma(h_j - \tilde{h}_j)^2}{K_\Sigma(h_j) + f_j} + \int_{\mathcal{U}_{j,\sigma} \cap \mathcal{O}_{x_j}^1} \frac{K_\Sigma(\tilde{h}_j - g_j)^2}{K_\Sigma(h_j) + f_j} + \int_{\mathcal{U}_{j,\sigma} \cap \mathcal{O}_{x_j}^1} \frac{(K_\Sigma(g_j) - f_j)^2}{K_\Sigma(h_j) + f_j}$$

and therefore

$$\mathrm{d}_{\mathrm{H}}^2(f_j, K_\Sigma(h_j)) \leq \int_{(\mathcal{O}_{x_j}^1)^c} K_\Sigma(h_j) \tag{S8}$$

$$+ \int_{\mathcal{U}_{j,\sigma}^c} [K_\Sigma(h_j) + f_j] \tag{S9}$$

$$+ (1 - I_j)^2 \tag{S10}$$

$$+ \int_{\mathcal{U}_{j,\sigma} \cap \mathcal{O}_{x_j}^1} \frac{K_\Sigma\left((f_j/2 - g_j) \mathbb{1}_{g_j < f_j/2} \right)^2}{K_\Sigma(h_j) + f_j} \tag{S11}$$

$$+\int_{\mathcal{U}_{j,\sigma}\cap\mathcal{O}^1_{x_j}}\left(\frac{K_\Sigma(g_j)-f_j}{f_j}\right)^2 f_j \tag{S12}$$

where each term will be bounded independently.

1. For (S8), notice that for any $y\in\mathcal{O}^0_{x_j}$ and any $x\notin\mathcal{O}^1_{x_j}$, there holds that $\|x-y\|^2_{\Sigma_y^{-1}}\geq C/\rho^2(\delta,\sigma($ where $\rho(\delta,\sigma)=\max(\delta\sigma^{\alpha_\perp},\sigma^{\alpha_0})$ for some C depending on τ. We can thus write

$$\int_{(\mathcal{O}^1_{x_j})^c}K_\Sigma(h_j)=\int_{(O^1_{x_j})^c}\int h_j(y)\varphi_{\Sigma(y)}(x-y)\,\mathrm{d}y\,\mathrm{d}x$$
$$=\int_{(O^1_{x_j})^c}\int_{\mathcal{O}^0_{x_j}}h_j(y)\frac{\exp\left(-\frac{1}{2}\|x-y\|^2_{\Sigma^{-1}(y)}\right)}{(2\pi)^{D/2}\sigma^D\delta^{D-d}}\,\mathrm{d}y\,\mathrm{d}x$$
$$\leq\exp\left\{-\frac{C^2}{4\delta^2\sigma^{2\alpha_\perp}}\right\}\int_{(\mathcal{O}^1_{x_j})^c}\int_{\mathcal{O}^0_{x_j}}h_j(y)\frac{\exp\left(-\frac{1}{4}\|x-y\|^2_{\Sigma^{-1}(y)}\right)}{(2\pi)^{D/2}\sigma^D\delta^{D-d}}\,\mathrm{d}y\,\mathrm{d}x$$
$$\lesssim\exp\left\{-\frac{C^2}{4\rho(\delta,\sigma)^2}\right\}\int_{(O^1_{x_j})^c}K_{2\Sigma}(h_j)\lesssim\exp\left\{-\frac{C^2}{4\rho(\delta,\sigma)^2}\right\}\lesssim\sigma^{2\beta},$$

because $\rho(\delta,\sigma)=o(|\log(\sigma)|^{1/2})$ by assumption.

2. For (S12), there holds

$$\int_{\mathcal{U}_{j,\sigma}\cap\mathcal{O}^1_{x_j}}\left(\frac{K_\Sigma(g_j)-f_j}{f_j}\right)^2 f_j\leq\int_{\mathcal{O}^1_{x_j}}\left(\frac{L_j\sigma^\beta}{f_j}\right)^2 f_j\leq\sigma^{2\beta}\left\{\int_{\mathcal{O}^1_{x_j}}\left(\frac{L_j}{f_j}\right)^{\frac{2\beta+\varepsilon}{\beta}}f_j\right\}^{\frac{2\beta}{2\beta+\varepsilon}}$$
$$\leq\sigma^{2\beta}\left\{\int_{\mathcal{W}^1_{x_j}}\left(\frac{L_{j,\delta}}{\bar{f}_{j,\delta}}\right)^{\frac{2\beta+\varepsilon}{\beta}}\bar{f}_{j,\delta}\frac{|\det\mathrm{d}\bar{\Psi}_{x_j,\delta}|}{\delta^{D-d}}\right\}^{\frac{2\beta}{2\beta+\varepsilon}}$$
$$\lesssim\sigma^{2\beta}\left\{\int_{\mathcal{W}^1_{x_j}}\left(\frac{L_{j,\delta}}{\bar{f}_{j,\delta}}\right)^{\frac{2\beta+\varepsilon}{\beta}}\bar{f}_{j,\delta}\right\}^{\frac{2\beta}{2\beta+\varepsilon}}\lesssim c_j^{-\frac{2\beta}{2\beta+\varepsilon}}\sigma^{2\beta}\lesssim c_j^{-1}\sigma^{2\beta},$$

where we used that $|\det\mathrm{d}\bar{\Psi}_{x_j,\delta}|\lesssim\delta^{D-d}$ and Lemma S3.2.

3. For (S10), notice that

$$I_j=\int g_j\mathbb{1}_{\{f_j<2g_j\}}+\int\frac{f_j}{2}\mathbb{1}_{\{f_j\geq 2g_j\}}=\int g_j+\int\frac{(f_j-2g_j)}{2}\mathbb{1}_{\{f_j\geq 2g_j\}}$$

and that $\int g_j=\int K_\Sigma(g_j)$. Moreover

$$\int_{\mathcal{O}^1_{x_j}}K_\Sigma(g_j)=\int_{\mathcal{O}^1_{x_j}}f_j+\int_{\mathcal{O}^1_{x_j}}(K_\Sigma(g_j)-f_j)=1+\int_{\mathcal{O}^1_{x_j}}(K_\Sigma(g_j)-f_j)$$

and since

$$\left|\int_{\mathcal{O}^1_{x_j}}K_\Sigma(g_j)-f_j\right|\leq\int_{\mathcal{O}^1_{x_j}}L_j\sigma^\beta=\sigma^\beta\int_{\mathcal{W}^1_{x_j}}L_{j,\delta}\times\frac{|\det\mathrm{d}\bar{\Psi}_{x_j,\delta}|}{\delta^{D-d}}$$
$$\lesssim\sigma^\beta\left\{\int_{\mathcal{W}^1_{x_j}}\left(\frac{L_{j,\delta}}{\bar{f}_{j,\delta}}\right)^{\frac{2\beta+\varepsilon}{\beta}}\bar{f}_{j,\delta}\right\}^{\frac{\beta}{2\beta+\varepsilon}}\lesssim c_j^{-\frac{\beta}{2\beta+\varepsilon}}\sigma^\beta\lesssim c_j^{-1/2}\sigma^\beta.$$

where we again used that $|\det \mathrm{d}\bar{\Psi}_{x_j,\delta}| \lesssim \delta^{D-d}$ and Lemma S3.2, we have that

$$\int_{\mathcal{O}^1_{x_j}} K_\Sigma(g_j) = 1 + O(c_j^{-1/2}\sigma^\beta).$$

We also have, as in (S8) that

$$\begin{aligned}\int_{(\mathcal{O}^1_{x_j})^c} K_\Sigma(|g_j|) &\lesssim e^{-\frac{C^2}{4\delta^2\sigma^{2\alpha_\perp}}} \int_{\mathcal{O}^1_{x_j}} |g_j|(y)dy \\ &\lesssim e^{-\frac{C^2}{4\delta^2\sigma^{2\alpha_\perp}}} \|L_j\|_\infty = o(c_j^{-1}\sigma^{2\beta}) = o(c_j^{-1/2}\sigma^\beta),\end{aligned}$$

where the two last inequality comes from $\rho^2(\delta,\sigma) = o(1/\log(1/\sigma))$ and $c_j \geq \sigma^{2\beta}$. Therefore $|1 - \int g_j| \lesssim c_j^{-1/2}\sigma^\beta$. Moreover, using this time Lemma S3.3

$$\begin{aligned}\int_{f_j > 2g_j} (f_j - 2g_j) &\lesssim P_{f_j}(f_j > 2g_j) + \sum_{0<\langle k,\boldsymbol{\alpha}\rangle<\beta} \frac{\sigma^{\langle k,\boldsymbol{\alpha}\rangle}}{\delta^{D-d}} \int_{f_j>2g_j} |\mathrm{D}^k \bar{f}_{j,\delta}(z_x)|dx \\ &\lesssim c_j^{-1}\sigma^{2\beta+\varepsilon}.\end{aligned}$$

All in all, we find that

$$\begin{aligned}(1 - I_j)^2 &\lesssim \exp\left\{-C^2/(4\delta^2\sigma^{2\alpha_\perp})\right\} + \left\{c_j^{-1}\sigma^{2\beta}\right\} \vee \left\{c_j^{-2}\left\{\sigma\log(1/\sigma)\right\}^{4\beta+2\varepsilon}\right\} \\ &\lesssim \exp\left\{-C^2/(4\delta^2\sigma^{2\alpha_\perp})\right\} + c_j^{-1}\sigma^{2\beta}\end{aligned}$$

where the last inequality again holds because $c_j \geq \sigma^{2\beta}$.

4. For (S9), we start with taking $\langle k,\boldsymbol{\alpha}\rangle < \beta$. Notice that, thanks to Proposition S2.3, there exists a constant $C > 0$ such that for any $x, y \in \mathcal{O}^0_{x_j}$, $|\mathrm{D}^k \bar{f}_{j,\delta}(z_x) - \mathrm{D}^k \bar{f}_{j,\delta}(z_y)| \leq C\delta^{D-d}L_j(x)$. Then, using (20), together with $|\det \mathrm{d}\bar{\Psi}(w)| \leq C$,

$$\begin{aligned}|K_\Sigma(\mathrm{D}^k \bar{f}_{j,\delta}(z_\square))(x)| &\lesssim \int_{\Delta^{-1}_{\sigma,1}(\mathcal{W}^0 - z_x)} e^{-B_\sigma(x,z)} |\mathrm{D}^k \bar{f}_{j,\delta}(\Delta_{\sigma,1}z + z_x)|dz \\ &\lesssim \delta^{D-d}L_j(x) \int_{\Delta^{-1}_{\sigma,1}(\mathcal{W}^0 - z_x)} e^{-B_\sigma(x,z)}dz \lesssim \delta^{D-d}L_j(x).\end{aligned}$$

Now notice that

$$\begin{aligned}&\int_{\mathcal{U}^c_{j,\sigma}} |\mathrm{D}^k \bar{f}_{j,\delta}(z_x)|\,\mathrm{d}x \\ &\leq \left\{\int_{\mathcal{U}^c_{j,\sigma}} \left\{\frac{|\mathrm{D}^k \bar{f}_{j,\delta}(z_x)|}{\bar{f}_{j,\delta}(z_x)}\right\}^{\frac{2\beta+\varepsilon}{\langle k,\boldsymbol{\alpha}\rangle}} \bar{f}_{j,\delta}(z_x)\,\mathrm{d}x\right\}^{\frac{\langle k,\boldsymbol{\alpha}\rangle}{2\beta+\varepsilon}} \times \left\{\int_{\mathcal{U}^c_{j,\sigma}} \bar{f}_{j,\delta}(z_x)\,\mathrm{d}x\right\}^{1-\frac{\langle k,\boldsymbol{\alpha}\rangle}{2\beta+\varepsilon}} \\ &\lesssim c_j^{-1}\delta^{D-d}\sigma^{H_1\times\{1-\langle k,\boldsymbol{\alpha}\rangle/(2\beta+\varepsilon)\}} \lesssim c_j^{-1}\delta^{D-d}\sigma^{H_1/2}\end{aligned}$$

and likewise, $\int_{\mathcal{U}^c_{j,\sigma}} L_j(x)\,\mathrm{d}x \lesssim c_j^{-1}\delta^{D-d}\sigma^{H_1\times\{1-\beta/(2\beta+\varepsilon)\}} \lesssim \delta^{D-d}\sigma^{H_1/2}$. We thus have shown that

$$\int_{\mathcal{U}^c_{j,\sigma}} |K_\Sigma(\mathrm{D}^k \bar{f}_{j,\delta}(z_\square))(x)| \lesssim c_j^{-1}\delta^{D-d}\sigma^{H_1/2}. \tag{S13}$$

Coming back to (S9), there immediately holds that $\int_{\mathcal{U}^c_{j,\sigma}} f_j \lesssim \sigma^{H_1}$, and furthermore, noticing that

$$h_j \lesssim \widetilde{h}_j \lesssim f_j + \frac{1}{\delta^{D-d}} \sum_{0\leq\langle k,\boldsymbol{\alpha}\rangle<\beta} \sigma^{\langle k,\boldsymbol{\alpha}\rangle} |\mathrm{D}^k \bar{f}_{j,\delta}(z_\square)|$$

and using (S13) and the monotonicity of K_Σ, we find that $\int_{\mathcal{U}_{j,\sigma}^c} K_\Sigma(h_j) \lesssim c_j^{-1}\sigma^{H_1/2}$.

5. Finally, for (S11), notice that $f_j/2 - g_j$ is a sum of terms of the form $\sigma^{\langle k,\alpha\rangle}\delta^{-(D-d)}\,\mathrm{D}^k \bar f_{j,\delta}(z_\square)$ for $0 \le \langle k,\alpha\rangle < \beta$. The term $K_\Sigma((f_j/2-g_j)\mathbb{1}_{g_j\le f_j/2})^2$ is thus upper bounded by a sum of terms of the form $\sigma^{2\langle k,\boldsymbol{\alpha}\rangle}\delta^{-2(D-d)}K_\Sigma(\mathrm{D}^k\bar f_{j,\delta}(z_\square)\mathbb{1}_{g_j\le f_j/2})^2$. Now there holds
$$\int_{\mathcal{U}_{j,\sigma}\cap\mathcal{O}^1_{x_j}} \frac{K_\Sigma(\mathrm{D}^k\bar f_{j,\delta}(z_\square)\mathbb{1}_{g_j\le f_j/2})^2}{K_\Sigma(h_j)+f_j}$$
$$\lesssim \sigma^{-H_1}\|K_\Sigma(\mathrm{D}^k\bar f_{j,\delta}(z_\square)\mathbb{1}_{g_j\le f_j/2})\|_\infty \times \int K_\Sigma(|\mathrm{D}^k\bar f_{j,\delta}(z_\square)|\mathbb{1}_{g_j\le f_j/2})$$
$$\lesssim \frac{\sigma^{-H_1}}{c_j}\int_{\mathcal{O}^1_{x_j}} |\mathrm{D}^k\bar f_{j,\delta}(z_y)|\mathbb{1}_{g_j\le f_j/2})dy \lesssim \frac{\sigma^{-H_1}}{c_j^2}\delta^{D-d}\{\sigma\log(1/\sigma)\}^{2\beta+\varepsilon-\langle k,\boldsymbol{\alpha}\rangle}$$
where we used Lemma S3.3. Summing all the bounds in k, we obtain that (S11) is bounded from above, up to constant, by
$$\max_k c_j^{-2}\sigma^{2\beta+\varepsilon-H_1+2\langle k,\boldsymbol{\alpha}\rangle}\delta^{-(D-d)}\log(1/\sigma)^{2\beta+\varepsilon} \lesssim c_j^{-1}\delta^{-(D-d)}\sigma^{\varepsilon-H_1}\log(1/\sigma)^{2\beta+\varepsilon}$$
where we used that $c_j \ge \sigma^{2\beta}$.

Collecting all the bounds on (S8-S12) together with $\delta^2\sigma^{2\alpha_\perp} = o(1/\log(1/\sigma))$, we get that
$$c_j\,\mathrm{d}_{\mathrm{H}}^2(K_\Sigma(h_j),f_j) \lesssim c_j\left[\sigma^{2\beta}\log(1/\sigma)^{4\beta+2\varepsilon} + \sigma^{H_1/2} + \sigma^{\varepsilon-H_1}\delta^{-(D-d)}\log(1/\sigma)^{2\beta+\varepsilon}\right].$$
Choosing $H_1 = 4\beta$ and $\varepsilon \ge 6\beta+\varepsilon_1$ where $\sigma^{\varepsilon_1}\le\delta^{D-d}$, we obtain the result.

S4. Supplement to Section 5.1: proof of Theorem 3.1. For any probability distribution P on $\mathbb{R}^D\times\mathcal{S}^{++}(D,\mathbb{R})$, one defines the probability density function on $\mathbb{R}^D$
$$f_P(x) := \int \varphi_\Sigma(x-y)\,\mathrm{d}P(y,\Sigma).$$
Note that when g is a probability distribution on $\mathbb{R}^D$, then
$$K_\Sigma g(x) = f_P(x) \quad\text{with}\quad \mathrm{d}P(y,\Sigma) = \delta_{\Sigma(y)}(\Sigma)g(y)\,\mathrm{d}y.$$

LEMMA S4.1. *Let $V_0,\dots,V_N$ be a partition of $\mathbb{R}^D$ and let $P=\sum_{j=1}^N \pi_j\delta_{z_j,\Sigma_j}$ with $z_j\in V_j$. Then, for any probability measure Q on $\mathbb{R}^D\times\mathcal{S}^{++}(D,\mathbb{R})$,*
$$\|f_Q-f_P\|_1 \le 2\sum_{j=1}^N|Q_1(V_j)-\pi_j| + \frac12\sup_{1\le j\le N}\|\Sigma_j^{-1/2}\|_{\mathrm{op}}\operatorname{diam}V_j + \frac32\sup_{1\le j\le N}\sup_{\Sigma\in\mathcal{S}_j}\sqrt{\operatorname{tr}(\Sigma_j^{-1}\Sigma-\mathrm{Id})^2},$$
where Q_1 is the first marginal of Q and $\mathcal{S}_j$ is the support of the second marginal of $\mathbb{1}_{V_j}(y)\,\mathrm{d}Q(y,\Sigma)$.

PROOF. We write that $f_Q(x)-f_P(x)$ is
$$\int_{V_0}\varphi_\Sigma(x-y)\,\mathrm{d}Q(y,\Sigma) + \sum_{j=1}^N\int_{V_j}\left\{\varphi_\Sigma(x-y)-\varphi_{\Sigma_j}(x-z_j)\right\}\mathrm{d}Q(y,\Sigma)$$
$$+\sum_{j=1}^N(Q_1(V_j)-\pi_j)\varphi_{\Sigma_j}(x-z_j),$$

The last term can be readily bounded in L^1-norm by

$$\sum_{j=1}^{N} |Q_1(V_j) - \pi_j| \int \varphi_{\Sigma_j}(x - z_j)\,\mathrm{d}x = \sum_{j=1}^{N} |Q_1(V_j) - \pi_j|,$$

and the first one by $Q_1(V_0) = 1 - \sum_{j=1}^N Q_1(V_j) \leq \sum_{j=1}^N |Q_1(V_j) - \pi_j|$. For the second term, notice that each term of the sum is upper-bounded in L^1-norm by

$$\begin{aligned}
&\int_{V_j} \|\varphi_\Sigma(\cdot - y) - \varphi_{\Sigma_j}(\cdot - z_j)\|_1 \,\mathrm{d}Q(y,\Sigma) \\
&\leq \int_{V_j} \|\varphi_{\Sigma_j}(\cdot - y) - \varphi_{\Sigma_j}(\cdot - z_j)\|_1 \,\mathrm{d}Q(y,\Sigma) + \int_{V_j} \|\varphi_\Sigma(\cdot - y) - \varphi_{\Sigma_j}(\cdot - y)\|_1 \,\mathrm{d}Q(y,\Sigma) \\
&\leq \frac{1}{2}\int_{V_j} \|\Sigma_j^{-1/2}(y - z_j)\| \,\mathrm{d}Q_1(y) + \frac{3}{2}\int_{V_j} \sqrt{\mathrm{tr}(\Sigma_j^{-1}\Sigma - \mathrm{Id})^2}\,\mathrm{d}Q(y,\Sigma) \\
&\leq \frac{1}{2} Q_1(V_j)\|\Sigma_j^{-1/2}\|_{\mathrm{op}}\,\mathrm{diam}\,V_j + \frac{3}{2}Q_1(V_j)\sup_{\Sigma\in\mathcal{S}_j}\sqrt{\mathrm{tr}(\Sigma_j^{-1}\Sigma - \mathrm{Id})^2},
\end{aligned}$$

where we used Prp 2.1 and Thm 1.1 of [4]. □

S4.1. *Proof of Lemma 5.1*. Throughout the proof, C denotes a generic constant whose value depends only on D. We start with bounding the probability measure of $\mathcal{F}_n^c$. There holds

$$\Pi(\mathcal{F}_n^c) \leq \Pi\{\exists h \leq H_n, \mu_h \notin B(0,R_n)\} + \Pi\left\{\sum_{h>H_n} \pi_h \geq \varepsilon_n\right\} + \Pi(\exists h \leq H_n, \Lambda_h \notin [\underline{\sigma}_n^2, \bar{\sigma}_n^2]^D).$$

The first mass is bounded using (10):

$$\Pi\{\exists h \leq H, \mu_h \notin B(0,R_n)\} \lesssim H_n \int \|\mu\|^{-b_2}\mathbb{1}_{\|\mu\|\geq R_n}\,\mathrm{d}\mu \lesssim H_n R_n^{-b_2} \lesssim e^{-c_1 n\varepsilon_n^2},$$

as soon as $b_2 R_0 \geq 2c_1$. The second term is bounded in [12, p. 15] by

$$\Pi\left\{\sum_{h>H_n} \pi_h \geq \varepsilon_n\right\} \leq \left\{\frac{eB}{H_n}\log(1/\varepsilon)\right\}^{H_n} \lesssim e^{-c_1 n\varepsilon_n^2},$$

as soon as H_0 is large enough. Finally, to bound the last probability, we consider separately the partial and hybrid location-scale priors.

- Partial location-scale: $\Lambda_h = \Lambda_1$ for all h and using (11),

  $$\Pi(\Lambda_1 \notin \mathcal{Q}_n) \leq \Pi_\Lambda\left\{\min_i \Lambda_i \leq \underline{\sigma}_n^2\right\} + \Pi_\Lambda\left\{\max_i \Lambda_i \geq \bar{\sigma}_n^2\right\} \lesssim e^{-c_3\sigma_0^{-2b_3}n\varepsilon_n^2} + \bar{\sigma}_n^{-2b_4} \leq e^{-c_1 n\varepsilon_n^2},$$

  as soon as σ_1^2 is large enough and σ_0^2 is small enough.
- Hybrid location-scale prior

$$\begin{aligned}
\Pi(\exists h \leq H_n,\, \Lambda_h \notin \mathcal{Q}_n) &\leq H_n \mathbb{E}_{\tilde{\Pi}_\Lambda}\left[Q_2(\min_i \Lambda_i \leq \underline{\sigma}_n^2)\right] + H_n \mathbb{E}_{\tilde{\Pi}_\Lambda}\left[Q_2(\max_i \Lambda_i > \bar{\sigma}_n^2)\right] \\
&\leq H_n e^{-c_3\sigma_0^{-2b_3}n\varepsilon_n^2} + H_n e^{-c_4\sigma_1^{2b_4}n\varepsilon_n^2} \lesssim e^{-c_1 n\varepsilon_n^2}.
\end{aligned}$$

We then turn on bounding the entropy of the partions of $\mathcal{F}_n$. Note that on $\mathcal{F}_n$ $\min_i \lambda_i \geq \underline{\sigma}_n^2$ so that

$$\frac{\max_i \lambda_i}{\min_i \lambda_i} \leq \max_i \lambda_i \underline{\sigma}_n^{-2} \leq \sigma_0^{-2} n^{2t_0/b_3} \times \max_i \lambda_i.$$

From [2], the covering number of $\mathcal{F}_{n,\mathbf{j},\ell}$ is bounded by

$$N(\varepsilon_n, \mathcal{F}_{n,\mathbf{j},\ell}, \|\cdot\|_1) \le \exp\left(CH_n[\log n + \log(1/\varepsilon_n)] + (D-1)\sum_{h\le H_n} \log(j_h+1) + D(D-1)2^{\ell-2} H_n \log n \right),$$

and the covering number of $\mathcal{F}_{n,\mathbf{j}}$ is bounded by

$$N(\varepsilon_n, \mathcal{F}_{n,\mathbf{j}}, \|\cdot\|_1) \le \exp\left(CH_n[\log n + \log(1/\varepsilon_n)] + \sum_{h\le H_n} \log(j_h+1) + DH_n \log \bar\sigma_n^2 \right)$$
$$\lesssim \exp\left(CH_n[\log n + \log(1/\varepsilon_n)] + \sum_{h\le H_n} \log(j_h+1) \right).$$

We bound $\Pi(\mathcal{F}_{n,\mathbf{j},\ell})$ in the case of the partial location-scale prior, for $\ell \ge 1$:

$$\Pi(\mathcal{F}_{n,\mathbf{j},\ell}) \lesssim \prod_{h\le H_n} (j_h\sqrt{n})^{-(b_2-D)\mathbb{1}_{j_h\ge 1}} \Pi_\Lambda\left(\max_i \lambda_i > \sigma_0^2 n^{2^{\ell-1}-2t_0/b_3} \right)$$
$$\lesssim \prod_{h\le H_n} (j_h\sqrt{n})^{-(b_2-D)\mathbb{1}_{j_h\ge 1}} n^{-b_4(2^{\ell-1}-2t_0/b_3)}$$
$$= \exp\left\{ -(b_2-D)\sum_{h\le H_n} \mathbb{1}_{j_h\ge 1}\log(j_h) - \frac{1}{2}(b_2-D)nH_n - b_4(2^{\ell-1}-2t_0/b_3)H_n \log n \right\}.$$

We bound $\Pi(\mathcal{F}_{n,\mathbf{j}})$ in the case of the hybrid location-scale prior:

$$\Pi(\mathcal{F}_{n,\mathbf{j}}) \lesssim \prod_{h\le H_n} (j_h\sqrt{n})^{-(b_2-D)\mathbb{1}_{j_h\ge 1}}$$
$$= \exp\left\{ -(b_2-D)\sum_{h\le H_n} \mathbb{1}_{j_h\ge 1}\log(j_h) - \frac{1}{2}(b_2-D)H_n \log n \right\}.$$

This implies in particular that for the partial location-scale prior

$$\sum_{\mathbf{j},\ell} \sqrt{\Pi(\mathcal{F}_{n,\mathbf{j},\ell}) N(\varepsilon_n, \mathcal{F}_{n,\mathbf{j},\ell}, \|\cdot\|_1)} \lesssim \exp\left(\frac{1}{2} CH_n[\log n + \log(1/\varepsilon_n)] \right)$$
$$\times \sum_{\mathbf{j},\ell} \exp\Big(\frac{1}{2}\sum_h [(D-1)\log(j_h+1) - (b_2-D)(\log(j_h)\mathbb{1}_{j_h\ge 1} - 1)]$$
$$+ \mathbb{1}_{\ell\ge 2} 2^{\ell-1}\log n[D(D-1)/2 - b_4] \Big)$$
$$\lesssim \exp\left(CH_n[\log n + \log(1/\varepsilon_n)]/2 \right),$$

since $b_4 > D(D-1)/2$ and $b_2 > 2D-1$. Therefore, by choosing $M_0 > 0$ large enough

$$\sum_{\mathbf{j},\ell} \sqrt{\Pi(\mathcal{F}_{n,\mathbf{j},\ell}) N(\varepsilon_n, \mathcal{F}_{n,\mathbf{j},\ell}, \|\cdot\|_1)} e^{-M_0 n\varepsilon_n^2} = o(1).$$

We obtain a similar result for the hybrid location-scale prior.

S4.2. *Proof of Lemma 5.2.* We let again $R = (H\log(1/\varepsilon)/C_2)^{1/\kappa}$ and define $\sigma := \varepsilon^{1/\beta}$. Thanks to Corollary 3.5, we know there exists a density function g supported on M^δ such that $\mathrm{d}_{\mathrm{H}}^2(K_\Sigma g, f_0) \lesssim \sigma^{2\beta}\log^q(1/\sigma)$ for some $a > 0$. We can in turn, thanks to Lemma 5.3, find a discrete probability measure G on $M^\delta \cap \mathsf{B}(0,R)$ with N atoms at least $\sigma^{\alpha_1}\varepsilon^2$-apart such that

$$\|K_\Sigma g - K_\Sigma G\|_1 \lesssim \varepsilon^2 \log^{D/2}(1/\varepsilon) \quad \text{and} \quad N \lesssim \sigma^{-D}\log^D(1/\varepsilon).$$

We thus have $\mathrm{d}_{\mathrm{H}}^2(K_\Sigma G, f_0) \lesssim \sigma^{2\beta}\log^q(1/\sigma) + \varepsilon^2\log^{D/2}(1/\varepsilon) \lesssim \varepsilon^2\log^{2r}(1/\varepsilon)$ with $r := q/2 \vee D/4$. We let $z_1,\dots,z_N$ be the atoms of G and denote by $p_j = G(z_j)$. We let V_j be the ball centered around z_j with radius $\sigma^{2\alpha_0}\varepsilon^2/2$. We complete $V_1,\dots,V_N$ with sets $V_{N+1},\dots,V_J$ that forms a partition of $M^\tau \cap \mathsf{B}(0,R)$ with V_j included in balls of the form $\left\{x \in \mathbb{R}^D \mid \|x - z_j\|_{\Sigma^{-1}(z_j)} \le 1\right\}$ for some $z_j \in M^\tau \cap \mathsf{B}(0,R)$, so that we can take $J \lesssim N + (R/\sigma)^D \lesssim \sigma^{-D}\log^{D/\kappa}(1/\varepsilon)$. We then set V_0 to be the complementary set of the reunion of the V_j and set further $p_j = 0$ for j greater than $N+1$.

We write under the prior Π, $P = \sum_{h=1}^\infty \pi_h \delta_{\mu_h,U_h,\Lambda_h}$ and $\Sigma_h = U_h^\top \Lambda_h U_h$. We use the convention that $\Lambda_h = \Lambda$ for all h in the case of the Partial location scale prior and $\pi_h = 0$ for $h \ge K$ for the mixture of finite mixtures prior. Set $\widetilde{N} = \log(1/\varepsilon) \times N$, we consider the following events

$$\begin{aligned}
\mathcal{P}_J &= \left\{ \sum_{j=1}^J |p_j - P_\mu(V_j)| \le \varepsilon^2 \ \text{ and } \ \min_{1\le j\le J} P_\mu(V_j) \ge \varepsilon^4 \right\},\\
\mathcal{F}_{\widetilde{N}} &= \left\{ \sum_{h\le \widetilde{N}} \pi_h \ge 1 - \varepsilon^8 \right\},\\
\mathcal{O}_{\widetilde{N}} &= \left\{ \forall 1\le h\le \widetilde{N},\ \|U_h O_{\mu_h}^\top - \mathrm{Id}\,\|_{\mathrm{op}} \le \sigma^{2\alpha_0}\varepsilon^2 \right\},\\
\text{and } \ \mathcal{L}_{\widetilde{N}} &= \left\{ \forall 1\le h\le \widetilde{N},\ \Lambda_h \in \mathcal{S}(\sigma^{\alpha_0})^d \times \mathcal{S}(\delta\sigma^{\alpha_1})^{D-d} \right\}\\
\text{where } \ \mathcal{S}(t) &= \left\{ s \mid t^2 \le s \le t^2(1+\sigma^{2\beta}) \right\}.
\end{aligned}$$

We first show that if $P \in \mathcal{P}_J \cap \mathcal{F}_{\widetilde{N}} \cap \mathcal{O}_{\widetilde{N}} \cap \mathcal{L}_{\widetilde{N}}$, then

$$\mathrm{d}_{\mathrm{H}}(f_P, f_0) \lesssim \varepsilon \log^r(1/\varepsilon).$$

Indeed, we have

$$\mathrm{d}_{\mathrm{H}}(f_P,f_0) \le \mathrm{d}_{\mathrm{H}}(f_P, K_\Sigma G) + \mathrm{d}_{\mathrm{H}}(K_\Sigma G, f_0) \le \mathrm{d}_{\mathrm{H}}(f_{\widehat{P}}, f_{\widehat{G}}) + \mathrm{d}_{\mathrm{H}}(f_{\widehat{P}}, f_P) + \varepsilon\log^r(1/\varepsilon),$$

where $\mathrm{d}\widehat{G}(y,\Sigma) = \delta_{\Sigma(y)}\,\mathrm{d}G(y)$ and $\widehat{P} = \sum_{h\ge1}\pi_h\delta_{\mu_h,\widehat{\Sigma}_h}$ with

$$\widehat{\Sigma}_h = \begin{cases} \Sigma_h & \text{if } h \le \widetilde{N} \\ \Sigma(z_h) & \text{if } h > \widetilde{N} \text{ and } \mu_h \in V_j \text{ with } j \le N; \\ 0 & \text{otherwise,} \end{cases}$$

where, conventionally, $\varphi_\Sigma(\cdot - z)\,\mathrm{d}z = \delta_z$ when $\Sigma = 0$. Since $P \in \mathcal{F}_{\widetilde{N}}$, there easily holds $\|f_P - f_{\widehat{P}}\|_1 \le 2\varepsilon^2$ and, using Lemma S4.1, we find that

$$\begin{aligned}
&\|f_{\widehat{P}} - f_{\widehat{G}}\|_1\\
&\le 2\sum_{j=1}^N |P_\mu(V_j) - p_j| + \frac{1}{2}\sigma^{-\alpha_0}\sup_{1\le j\le N}\operatorname{diam} V_j + \frac{3}{2}\sup_{1\le j\le N}\sup_{\mu_h\in V_j}\sqrt{\operatorname{tr}(\Sigma(z_j)^{-1}\widehat{\Sigma}_h - \mathrm{Id})^2},\\
&\lesssim \varepsilon^2 + \sup_{1\le j\le N}\sup_{\mu_h\in V_j}\|\Sigma(z_j)^{-1}\widehat{\Sigma}_h - \mathrm{Id}\,\|_{\mathrm{op}}.
\end{aligned}$$

where we used that $P \in \mathcal{P}_J$ and that $\operatorname{diam} V_j \lesssim \sigma^{\alpha_0}\varepsilon^2$ for $j \le N$. In the last supremum, if $h > \widetilde{N}$ and $\mu_h \in V_j$, then $\Sigma_h = \Sigma(z_j)$ so we only need to handle the case when $h \le \widetilde{N}$. Moreover,

$$\begin{aligned}\Sigma(z_j)^{-1}\widehat{\Sigma}_h - \mathrm{Id} &= O_{z_j}^\top \Delta_{\sigma,\delta}^{-2} O_{z_j} U_h^\top \Lambda_h U_h - \mathrm{Id}\\ &= O_{z_j}^\top \Delta_{\sigma,\delta}^{-2}(O_{z_j}U_h^\top - \mathrm{Id})\Lambda_h U_h + O_{z_j}^\top \Delta_{\sigma,\delta}^{-2}\Lambda_h U_h - \mathrm{Id}\\ &= O_{z_j}^\top \Delta_{\sigma,\delta}^{-2}(O_{z_j}U_h^\top - \mathrm{Id})\Lambda_h U_h + O_{z_j}^\top (\Delta_{\sigma,\delta}^{-2}\Lambda_h - \mathrm{Id})U_h + O_{z_j}^\top U_h - \mathrm{Id},\end{aligned}$$

so that

$$\begin{aligned}\|\Sigma(z_j)^{-1}\widehat{\Sigma}_h - \mathrm{Id}\|_{\mathrm{op}} &\lesssim \|\Delta_{\sigma,\delta}^{-2}\|_{\mathrm{op}}\|\Lambda_h\|_{\mathrm{op}}\|O_{z_j}U_h^\top - \mathrm{Id}\|_{\mathrm{op}} + \|\Delta_{\sigma,\delta}^{-2}\Lambda_h - \mathrm{Id}\|_{\mathrm{op}} + \|O_{z_j}^\top U_h - \mathrm{Id}\|_{\mathrm{op}}\\ &\lesssim \varepsilon^2 + \sigma^{2\beta} \simeq \varepsilon^2,\end{aligned}$$

where we used both that $h \le \widetilde{N}$ and $P \in \mathcal{O}_{\widetilde{N}} \cap \mathcal{L}_{\widetilde{N}}$.

Using [12, Lem B2], we find that for $\lambda > 0$ small enough

$$\begin{aligned}P_0 \log\frac{f_0}{f_P} &\lesssim \mathrm{d}_{\mathrm{H}}^2(f_0, f_P)(1 + \log(1/\lambda)) + P_0\left\{\log\frac{f_0}{f_P}\mathbb{1}_{f_P/f_0<\lambda}\right\}\\ P_0\left(\log\frac{f_0}{f_P}\right)^2 &\lesssim \mathrm{d}_{\mathrm{H}}^2(f_0, f_P)(1 + \log^2(1/\lambda)) + P_0\left\{\left(\log\frac{f_0}{f_P}\right)^2\mathbb{1}_{f_P/f_0<\lambda}\right\},\end{aligned}$$

up to numeric constant. Notice that by assumption, f_0 is bounded from above by $\|L\|_\infty$. Let $x \in M^\tau \cap \mathsf{B}(0,R)$, and let $1 \le j \le J$ be such that $x \in V_j$. Notice that, since $P \in \mathcal{F}_{\widetilde{N}} \cap \mathcal{P}_J$, there holds

$$\varepsilon^4 \le P_\mu(V_j) = \sum_{\mu_h \in V_j}\pi_h \le \sum_{\substack{\mu_h \in V_j\\ h \le \widetilde{N}}}\pi_h + \varepsilon^8 \quad \text{so that} \quad \sum_{\substack{\mu_h \in V_j\\ h \le \widetilde{N}}}\pi_h \ge \varepsilon^4/2.$$

Now we can write

$$f_P(x) \ge \sum_{\substack{\mu_h \in V_j\\ h \le \widetilde{N}}}\pi_h \varphi_{\Sigma_h}(x - \mu_h) \gtrsim \frac{\varepsilon^4}{\delta^{D-d}\sigma^D},$$

where we used that $|\det \Sigma_h|^{1/2} \le (1 + \sigma^{2\beta})^{D/2}\delta^{D-d}\sigma^D \lesssim \delta^{D-d}\sigma^D$ and that $\|x - y\|_{\Sigma^{-1}(y)} \lesssim 1$ for any $x, y \in V_j$ with a similar line of reasoning as in the proof of Lemma 5.3, relying on $P \in \mathcal{O}_{\widetilde{N}} \cap \mathcal{L}_{\widetilde{N}}$. Furthermore, since $f_P(x) \le \varepsilon^8 + \sum_{h \le \widetilde{N}}\pi_h\varphi_{\Sigma_h}(x - \mu_h)$, there must be some $h \le \widetilde{N}$ such that both $\pi_h \ge \varepsilon^2$ and $\mu_h \in \mathsf{B}(0,R)$ holds, otherwise one would get a contradiction looking at the mass of f_0 since one would get

$$\begin{aligned}\|f_0\|_1 &= \|f_0\mathbb{1}_{\mathsf{B}(0,R/2)}\|_1 + \|f_0\mathbb{1}_{\mathsf{B}(0,R/2)^c}\|_1\\ &\le \underbrace{\|f_P\mathbb{1}_{\mathsf{B}(0,R/2)}\|_1}_{\lesssim \varepsilon^2 + \varepsilon^H} + \underbrace{\|(f_P - f_0)\mathbb{1}_{\mathsf{B}(0,R/2)}\|_1}_{\lesssim \varepsilon + \sigma^\beta} + \underbrace{\|f_0\mathbb{1}_{\mathsf{B}(0,R/2)^c}\|_1}_{\lesssim \varepsilon^H},\end{aligned}$$

where the inequalities occurs up to log-term. For $x \notin \mathsf{B}(0,R)$, notice that, for this particular $h \le \widetilde{N}$, $f_P(x) \gtrsim \varepsilon^2\varphi_{\Sigma_h}(x - \mu_h)$. Taking $\lambda = C\varepsilon^4/(\delta^{D-d}\sigma^D)$ for some small constant C, one get for any $\ell \ge 1$,

$$\begin{aligned}&P_0\left\{\left(\log\frac{f_0}{f_P}\right)^\ell\mathbb{1}_{f_P/f_0<\lambda}\right\} \le P_0\left\{\left(\log\frac{f_0}{f_P}\right)^\ell\mathbb{1}_{\mathsf{B}(0,R)^c}\right\}\\ &\lesssim \log^\ell\frac{\varepsilon^2}{\sigma^D\delta^{D-d}} \times P_0(\mathsf{B}(0,R)^c) + \int_{\mathsf{B}(0,R)^c}\|x - \mu_h\|_{\Sigma_h^{-1}}^{2\ell} f_0(x)\,\mathrm{d}x\end{aligned}$$

$$\lesssim \varepsilon^H \log^\ell \frac{\varepsilon^2}{\sigma^D \delta^{D-d}} + \sigma^{-2\ell\alpha_0} \left\{ \int \|x\|^{4\ell} f_0(x)\,\mathrm{d}x \right\}^{1/2} P_0(\mathsf{B}(0,R)^c)^{1/2}$$
$$\lesssim \varepsilon^2 \log^2(1/\varepsilon),$$

where the last inequality holds for $\ell \in \{1,2\}$, provided that we chose $H \geq 8\alpha_0 + 4\beta$. This shows that $f_P \in \mathsf{B}(f_0, \tilde{\varepsilon})$ for $\tilde{\varepsilon} \approx \varepsilon \log^s(1/\varepsilon)$ with $s = r \vee 1$. It only remains to lower bound the prior mass of the event $\mathcal{P}_J \cap \mathcal{F}_{\widetilde{N}} \cap \mathcal{O}_{\widetilde{N}} \cap \mathcal{L}_{\widetilde{N}}$.

For this, one can use (10) and the fact that the scales are drawn independently to rest to find that

$$\Pi(\mathcal{P}_J \cap \mathcal{F}_{\widetilde{N}} \cap \mathcal{O}_{\widetilde{N}} \cap \mathcal{L}_{\widetilde{N}}) \geq \Pi(\mathcal{P}_J \cap \mathcal{F}_{\widetilde{N}}) \times c_o^{\widetilde{N}} \prod_{h \leq \widetilde{N}} O\left\{\|UO_{\mu_h} - \mathrm{Id}\,\|_{\mathrm{op}} \leq \sigma^{2\alpha_0}\varepsilon^2\right\} \times \Pi(\mathcal{L}_{\widetilde{N}}).$$

We easily get that $O\left\{\|UO_{\mu_h} - \mathrm{Id}\,\|_{\mathrm{op}} \leq \sigma^{2\alpha_0}\varepsilon^2\right\} \gtrsim (\sigma^{2\alpha_0}\varepsilon^2)^{D(D-1)/2}$. For $\Pi(\mathcal{L}_{\widetilde{N}})$, we use (11) or (12) along with a simple Markov inequality to find that

$$\Pi(\mathcal{L}_{\widetilde{N}}) = \mathbb{E}_\Pi \left\{ Q_2\left([\sigma^{2\alpha_0}, (1+\sigma^{2\beta})\sigma^{2\alpha_0}]^d \times [\delta^2\sigma^{2\alpha_\perp}, (1+\sigma^{2\beta})\delta^2\sigma^{2\alpha_\perp}]^{D-d} \right)^{\widetilde{N}} \right\}$$
$$\gtrsim \exp(-2\alpha_0 B_0 \log(1/\sigma)\widetilde{N}) \exp\left\{-c_2 D \sigma^{-D}\right\}.$$

For $\Pi(\mathcal{P}_J \cap \mathcal{F}_{\widetilde{N}})$, we write

$$\Pi(\mathcal{P}_J \cap \mathcal{F}_{\widetilde{N}}) \geq \Pi(\mathcal{P}_J) - \Pi(\mathcal{F}^c_{\widetilde{N}}) \gtrsim e^{-CJ\log(1/\varepsilon)} - e^{-\widetilde{N}\log(\widetilde{N})},$$

where we used [7, Lem 10] and the bound [12, p. 15]. We conclude by noticing that $e^{-\widetilde{N}\log(\widetilde{N})} \ll e^{-N\log(1/\varepsilon)}$ so that in the end

$$\Pi(\mathcal{P}_J \cap \mathcal{F}_{\widetilde{N}} \cap \mathcal{O}_{\widetilde{N}} \cap \mathcal{L}_{\widetilde{N}}) \gtrsim \exp(-CN \log^t(1/\varepsilon))),$$

for some C depending on the parameters and $t = D/\kappa + 2$, ending the proof.

S4.3. *Proof of Lemma 5.3.*

PROOF OF LEMMA 5.3. Let $g : \mathbb{R}^D \to \mathbb{R}$ be a density supported on M^δ and satisfying (14) and (15). Let $R = (H\log(1/\varepsilon)/C_2)^{1/\kappa}$, so that g is less than $C_1\varepsilon^H$ outside of $\mathsf{B}(0,R)$. We consider the functions $\{\chi_i\}_{i\in\mathcal{I}}$ introduced in Lemma S1.4 associated with a $\tau/64$-packing $\{x_i\}_{i\in\mathcal{I}}$ of $M \cap \mathsf{B}(0,R)$. Recall that the number $|\mathcal{I}|$ of functions is less than of order R^D. We can write

$$K_\Sigma g(x) = K_\Sigma(g\mathbb{1}_{\mathsf{B}^c(0,R)})(x) + \sum_{i\in\mathcal{I}} K_\Sigma(\chi_i g \mathbb{1}_{\mathsf{B}(0,R)})(x) = K_\Sigma(g\mathbb{1}_{\mathsf{B}^c(0,R)})(x) + \sum_{i\in\mathcal{I}} c_i K_\Sigma g_i(x)$$

where $c_i = \|\chi_i g \mathbb{1}_{\mathsf{B}(0,R)}\|_1$ and $g_i = \chi_i g \mathbb{1}_{\mathsf{B}(0,R)}/c_i$ is a density supported on $\mathcal{O}^0_{x_i}$. Notice that for any $x \in \mathbb{R}^D$

$$K_\Sigma(g\mathbb{1}_{\mathsf{B}^c(0,R)})(x) \leq C_1\varepsilon^H K_\Sigma(\mathbb{1}_{\mathsf{B}^c(0,R)})(x) \leq C_1\varepsilon^H$$

and that

$$\|K_\Sigma(g\mathbb{1}_{\mathsf{B}^c(0,R)})\|_1 = \|g\mathbb{1}_{\mathsf{B}^c(0,R)}\|_1 \leq \int_{\|x\|\geq R} C_1 e^{-C_2\|x\|^\kappa} \lesssim \int_{r\geq R} e^{-C_2 r^\kappa} r^{D-1}\,\mathrm{d}r$$
$$\lesssim \log^{(D-2)/\kappa+1}(1/\varepsilon) \times \varepsilon^H,$$

where we used a NP-bound on the incomplete Gamma function in the last inequality. Consequently, the term $K_\Sigma(g\mathbb{1}_{\mathsf{B}^c(0,R)})$ need not be discretized. Take now $i \in \mathcal{I}$. As in the proof of Theorem 3.4, there holds that

$$\varphi_{\Sigma(y)}(x-y) \lesssim \frac{\varepsilon^H}{\sigma^D\delta^{D-d}} \quad \forall x \notin \mathcal{O}^1_{x_i}, y \in \mathcal{O}^0_{x_i},$$

provided that $\tau \gtrsim R\delta\sigma^{\alpha_\perp}$. This means that $|K_\Sigma g_i(x)| \lesssim \frac{\varepsilon^H}{\sigma^D\delta^{D-d}}$ for all $x \notin \mathcal{O}^1_{x_i}$. If now $x \in \mathcal{O}^1_{x_i}$, there holds

$$\begin{aligned}K_\Sigma g_i(x) &= \int_{\mathcal{W}^0_{x_i}} \varphi_{\Sigma(\bar\Psi_{x_i}(w))}(x - \bar\Psi_{x_i}(w))\bar g_i(w)\,\mathrm{d}w\\ &= \sum_{j\in\mathcal{J}_i} \bar c_{i,j}\int_{\mathcal{W}_{i,j}} \varphi_{\Sigma(\bar\Psi_{x_i}(w))}(x - \bar\Psi_{x_i}(w))\bar g_{i,j}(w)\,\mathrm{d}w\end{aligned}$$

where $\bar g_i(w) = g_i(\bar\Psi_{x_i}(w)) \times |\det \mathrm{d}\bar\Psi_{x_i}(w)|$, $\bar c_{i,j} = \int_{\mathcal{W}_{i,j}} \bar g_i$ and $\bar g_{i,j} = \bar g_i/\bar c_{i,j}$. The sets $\mathcal{W}_{i,j}$ form a partition of $\mathcal{W}^0_{x_i}$ that are included in

$$\prod_{\ell=1}^d [w^0_{t_\ell}, w^0_{t_\ell+1}) \times \prod_{\ell=d+1}^D [w^\perp_{t_\ell}, w^\perp_{t_\ell+1}) \cap \mathcal{W}^0_{x_i},$$

for $(w^0_{t_1},\dots,w^0_{t_d},w^\perp_{t_{d+1}},\dots,w^\perp_{t_D})$ vary on a grid of size $(\sigma^{\alpha_0},\dots,\sigma^{\alpha_0},\delta\sigma^{\alpha_\perp},\dots,\delta\sigma^{\alpha_\perp})$. Notice that

$$\operatorname{Card}\mathcal{J}_i \simeq \frac{1}{\sigma^{\alpha_0 d}} \times \left(\frac{\delta}{\delta\sigma^{\alpha_\perp}}\right)^{D-d} = \sigma^{-D}.$$

Let $i \in \mathcal{I}$ and $j \in \mathcal{J}_i$ be fixed and denote for short $\bar\Psi = \bar\Psi_{x_i}$ and $\bar\Sigma = \Sigma \circ \bar\Psi$. We let $\Gamma > 0$ and we distinguish two cases: $\inf_{w\in\mathcal{W}_{i,j}} \|x - \bar\Psi(w)\|^2_{\bar\Sigma^{-1}(w)} \geq \Gamma\log(1/\varepsilon)$ and $\inf_{w\in\mathcal{W}_{i,j}} \|x - \bar\Psi(w)\|^2_{\bar\Sigma^{-1}(w)} < \Gamma\log(1/\varepsilon)$. In the former,

$$\int_{\mathcal{W}_{i,j}} \varphi_{\bar\Sigma(w)}(x - \bar\Psi(w))\bar g_{i,j}(w)\,\mathrm{d}w \lesssim \frac{\varepsilon^{\Gamma/2}}{\sigma^D\delta^{D-d}}.$$

While if $\inf_{w\in\mathcal{W}_{i,j}} \|x - \bar\Psi(w)\|^2_{\bar\Sigma^{-1}(w)} \leq \Gamma\log(1/\varepsilon)$, we first show that $\sup_{w\in\mathcal{W}_{i,j}} \|x - \bar\Psi(w)\|^2_{\bar\Sigma^{-1}(w)} \leq \Gamma'\log(1/\varepsilon)$ for some $\Gamma' > \Gamma$. We let $w_0 \in \mathcal{W}_{i,j}$ such that $\|x - \bar\Psi(w_0)\|^2_{\bar\Sigma^{-1}(w_0)} \leq 2\Gamma\log(1/\varepsilon)$ and take $w \in \mathcal{W}_{i,j}$. Using (28), we get that there $\|\operatorname{pr}_{T_{\bar\Psi(w)}} - \operatorname{pr}_{T_{\bar\Psi(w_0)}}\|_{\mathrm{op}} \lesssim \sigma^{\alpha_0}$ and the same holds for $\operatorname{pr}_{N_{\bar\Psi(w)}} - \operatorname{pr}_{N_{\bar\Psi(w_0)}}$. Let us denote $y = \bar\Psi(w_0)$ and $z = \bar\Psi(w)$. First notice that if $x \notin \mathcal{W}^1_{x_i}$, then $\|x - y\|_{\Sigma^{-1}(y)} \gtrsim \sigma^{-\alpha_0} \wedge \delta^{-1}\sigma^{-\alpha_\perp} > \Gamma\log(1/\varepsilon)$ so that necessarily $x \in \mathcal{W}^1_{x_i}$. In particular, $\|x - y\| \leq B$ for some constant $B > 0$ not depending on Γ. Now recall that

$$\|x - z\|_{\Sigma^{-1}(z)} = \left\|\frac{1}{\sigma^{\alpha_0}}\operatorname{pr}_{T_z}(x - z) + \frac{1}{\sigma^{\alpha_\perp}\delta}\operatorname{pr}_{N_z}(x - z)\right\|.$$

Furthermore, we can write

$$x - y = \Psi(v_x) - \Psi(v_y) + \delta\left(N(v_x)\eta_x - N(v_y)\eta_x\right) \tag{S14}$$

$$= \mathrm{d}\Psi(v_y)[v_x - v_y] + \delta\,\mathrm{d}N(v_y)[v_x - v_y]\eta_x + \delta N(v_y)[\eta_x - \eta_y] + O(\sigma^{2\alpha_0}). \tag{S15}$$

Using (S14), there holds

$$\begin{aligned}\operatorname{pr}_{T_y}(x - y) &= \operatorname{pr}_{T_y}(\Psi(v_x) - \Psi(v_y) + \delta N(v_x)\eta_x)\\ &= \operatorname{pr}_{T_y}(\Psi(v_x) - \Psi(v_z) + \delta N(v_x)\eta_x) + O(\sigma^{\alpha_0})\\ &= \operatorname{pr}_{T_z}(\Psi(v_x) - \Psi(v_{x_{i,j}}) + \delta N(v_x)\eta_x) + O(\sigma^{\alpha_0})\\ &= \operatorname{pr}_{T_z}(x - z) + O(\sigma^{\alpha_0})\end{aligned}$$

where $O(\sigma^{\alpha_0})$ is independant of Γ. Now using (S15), we find that

$$\begin{aligned}\operatorname{pr}_{N_y}(x - y) &= \delta N(v_y)[\eta_x - \eta_y] + \delta\operatorname{pr}_{N_y}(\mathrm{d}N(v_y)[v_x - v_y]\eta_x) + O(\sigma^{2\alpha_0})\\ &= \delta N(v_z)[\eta_x - \eta_z] + \delta\operatorname{pr}_{N_z}(\mathrm{d}N(v_y)[v_x - v_{x_{i,j}}]\eta_z) + O(\delta\sigma^{\alpha_\perp}) + O(\sigma^{2\alpha_0})\\ &= \operatorname{pr}_{N_z}(x - z) + O(\delta\sigma^{\alpha_\perp} + \sigma^{2\alpha_0})\end{aligned}$$

where we used that $\alpha_0 \geq \alpha_\perp$, and where again $O(\delta\sigma^{\alpha_\perp} + \sigma^{2\alpha_0})$ does not depend on A. All in all, we find that

$$\begin{aligned}\|x-y\|_{\Sigma^{-1}(y)} &= \left\| \frac{1}{\sigma^{\alpha_0}} \operatorname{pr}_{T_x}(x-y) + \frac{1}{\sigma^{\alpha_\perp}\delta} \operatorname{pr}_{N_y}(x-y) \right\| \\ &= \left\| \frac{1}{\sigma^{\alpha_0}} \operatorname{pr}_{T_z}(x-z) + \frac{1}{\sigma^{\alpha_\perp}\delta} \operatorname{pr}_{N_z}(x-z) \right\| + O\left(1+\sigma^{2\alpha_0}/(\delta\sigma^{\alpha_\perp})\right) \\ &= \|x-z\|_{\Sigma^{-1}(z)} + O\left(1+\sigma^{2\alpha_0}/(\delta\sigma^{\alpha_\perp})\right). \end{aligned} \tag{S16}$$

Now using that $\sigma^{2\alpha_0} = o(\delta\sigma^{\alpha_\perp})$, there thus holds that $\|x-\bar\Psi(w)\|_{\bar\Sigma^{-1}(w)} \leq 2\Gamma\log(1/\varepsilon)$ for Γ large enough.

Denote $R_T(u) = \exp(u) - \sum_{t=0}^{T-1} u^t/t!$, then $|R_T(u)| \leq e^u|u|^T/T!$ and

$$\exp\left\{-\frac{1}{2}\|x-\bar\Psi(w)\|^2_{\bar\Sigma^{-1}(w)}\right\} = \sum_{t=0}^{T-1} \frac{(-1)^t}{2^t t!}\|x-\bar\Psi(w)\|^{2t}_{\bar\Sigma^{-1}(w)} + \underbrace{R_T\left(-\|x-\bar\Psi(w)\|^2_{\bar\Sigma^{-1}(w)}/2\right)}_{:=\,R_T(x,w)}.$$

Note that $R_T(x,w)$ is uniformly bounded by

$$|R_T(x,w)| \leq e^{5\Gamma\log(1/\varepsilon)}\frac{(5\Gamma\log(1/\varepsilon))^T}{T!} \approx \frac{\varepsilon^{-5\Gamma}\log^T(1/\varepsilon)}{T!}.$$

Set,

$$A_{i,j}(w) := \langle e_i, \bar\Sigma^{-1}(w)e_j\rangle, \quad B_i(w) := \langle e_i, \bar\Sigma^{-1}(w)\bar\Psi(w)\rangle, \quad \text{and} \quad C(w) := \|\bar\Psi(w)\|^2_{\bar\Sigma^{-1}(w)},$$

so that all functions $A_{i,j}$, B_i and C are continuous functions of w and if $x = (x_1,\dots,x_D)$,

$$\begin{aligned}\|x-\bar\Psi(w)\|^{2t}_{\bar\Sigma^{-1}(w)} &= \left\{\|x\|^2_{\bar\Sigma^{-1}(w)} - 2\langle x, \bar\Sigma^{-1}(w)\bar\Psi(w)\rangle + \|\bar\Psi(w)\|^2_{\bar\Sigma^{-1}(w)}\right\}^t \\ &= \sum_{|k|=t}\binom{t}{k}(-2)^{k_2}\left\{\sum_{1\leq i,j\leq D} x_i x_j A_{i,j}(w)\right\}^{k_1}\left\{\sum_{i=1}^D x_i B_i(w)\right\}^{k_2} C(w)^{k_3} \\ &= \sum_{|k|=t}\binom{t}{k}(-2)^{k_2}C(w)^{k_3}\sum_{|\ell|=k_1}\binom{k_1}{\ell}\prod_{1\leq i,j\leq D}(x_i x_j)^{\ell_{i,j}}A_{i,j}^{\ell_{i,j}}(w)\sum_{|m|=k_2}\binom{k_2}{m}\prod_{i=1}^D x_i^{m_i}B_i^{m_i}(w) \\ &= \sum_{(p,\ell,m)\in\mathcal{G}_t} P_{p,\ell,m}(x) \times \underbrace{C(w)^p\prod_{1\leq i,j\leq D}A_{i,j}^{\ell_{i,j}}(w)\prod_{1\leq i\leq D}B_i^{m_i}(w)}_{:=\,Q_{p,\ell,m}}\end{aligned}$$

where $P_{p,\ell,m}(x)$ are polynomial functions of x, $Q_{p,\ell,m}(w)$ are continuous functions of w, and where $\mathcal{G}_t$ is the set $\{(p,\ell,m) \mid p+|\ell|+|m| = t\} \subset \mathbb{N}^{D^2+D+1}$. According to [8, Lem 3.1], one can always find an atomic probability measure $G_{i,j}$ such that

$$\int Q_{p,\ell,m}(w)\bar g_{i,j}(w)\,\mathrm{d}w = \int Q_{p,\ell,m}(w)G_{i,j}(\mathrm{d}w)$$

for all p,ℓ,m such that $p+|\ell|+|m| \leq T-1$. Since there are less than T^{D^2+D+1} such triplets, the probability measure $G_{i,j}$ can be taken to have less than T^{D^2+D+1} atoms. Note that then, this measure satisfies that

$$\begin{aligned}&\left|\int_{\mathcal{W}_{i,j}}\varphi_{\bar\Sigma(w)}(x-\bar\Psi(w))(\bar g_{i,j}(w)\,\mathrm{d}w - G_{i,j}(\mathrm{d}w))\right| \\ &\qquad = \frac{1}{(2\pi)^{D/2}\delta^{D-d}\sigma^D}\left|\int_{\mathcal{W}_{i,j}} R_T(x,w)(\bar g_{i,j}(w)\,\mathrm{d}w - G_{i,j}(\mathrm{d}w))\right| \lesssim \frac{\varepsilon^{-5\Gamma}\log^T(1/\varepsilon)}{\delta^{D-d}\sigma^D T!}.\end{aligned}$$

All in all, $G_{i,j}$ is such that

$$\Big| \int_{\mathcal{W}_{i,j}} \varphi_{\bar{\Sigma}(w)}(x - \bar{\Psi}(w))\bar{g}_{i,j}(w)\,\mathrm{d}w - \int_{\mathcal{W}_{i,j}} \varphi_{\bar{\Sigma}(w))}(x - \bar{\Psi}(w))G_{i,j}(\mathrm{d}w)\Big|$$
$$\lesssim \begin{cases} \dfrac{1}{\sigma^D \delta^{D-d}}\varepsilon^H & \text{if } x \notin \mathcal{O}^1_{x_i}; \\ \dfrac{1}{\sigma^D \delta^{D-d}}\varepsilon^{\Gamma/2} & \text{if } x \in \mathcal{O}^1_{x_i} \text{ and } \inf_{w\in J}\|x - \bar{\Psi}(w)\|^2_{\bar{\Sigma}^{-1}(w)} \ge \Gamma\log(1/\varepsilon); \\ \dfrac{1}{\sigma^D \delta^{D-d}}\dfrac{1}{T!}\varepsilon^{-5\Gamma}\log^T(1/\varepsilon) & \text{if } x \in \mathcal{O}^1_{x_i} \text{ and } \inf_{w\in J}\|x - \bar{\Psi}(w)\|^2_{\bar{\Sigma}^{-1}(w)} \le \Gamma\log(1/\varepsilon). \end{cases}$$

Taking $H \ge 1$, $\Gamma \ge 2$ and $T \ge 5\Gamma\log(1/\varepsilon)$ yields a bound of order $\varepsilon/\sigma^D\delta^{D-d}$ in every case. Then the probability measure

$$G = \sum_{i\in\mathcal{I}} \alpha_i \sum_{j\in\mathcal{J}_i} \bar{c}_{i,j}(\bar{\Psi}_{x_i})_{\#}G_{i,j},$$

is a discrete measure on M^δ with at most

$$N_\varepsilon = T^{D^2+D+1}\sum_{i\in\mathcal{I}} \operatorname{Card}\mathcal{J}_i \approx \log^{D^2+D+1}(1/\varepsilon)R^D\sigma^{-D} \le \sigma^{-D}\log^{D^2+D/(\kappa\wedge 2)}(1/\varepsilon)$$

atoms and such that $\|K_\Sigma G - K_\Sigma g\|_\infty \lesssim \varepsilon/(\sigma^D\delta^{D-d})$.

We now turn to bounding the L^1 norm. Let again pick $i \in \mathcal{I}$ and $j \in \mathcal{J}_i$. We let $x_{i,j} = \bar{\Psi}_{x_i}(w_{i,j})$ for some $w_{i,j} \in \mathcal{W}_{i,j}$ and for $T > 0$ we define

$$\mathcal{H}_A = \left\{x \in \mathbb{R}^D \mid \|x - x_{i,j}\|_{\Sigma^{-1}(x_{i,j})} \le A\right\}.$$

There holds that $\operatorname{Leb}\mathcal{H}_A \simeq \delta^{D-d}\sigma^D T^D$. Furthermore, we have, denoting respectively $P_{i,j}$ and $Q_{i,j}$ the push-forwards of $\bar{g}_{i,j}$ and $G_{i,j}$ through $\bar{\Psi}_{x_i}$

$$\begin{aligned} \|K_\Sigma P_{i,j} - K_\Sigma Q_{i,j}\|_1 &= \int_{\mathcal{H}_A} |K_\Sigma P_{i,j} - K_\Sigma Q_{i,j}| + \int_{\mathcal{H}_A^c} |K_\Sigma P_{i,j} - K_\Sigma Q_{i,j}| \\ &\lesssim \delta^{D-d}\sigma^D A^D \|K_\Sigma P_{i,j} - K_\Sigma Q_{i,j}\|_\infty + \int_{\mathcal{H}_A^c} [K_\Sigma Q_{i,j} + K_\Sigma P_{i,j}] \\ &\lesssim A^D\varepsilon + \int_{\mathcal{H}_A^c} [K_\Sigma P_{i,j} + K_\Sigma Q_{i,j}]. \end{aligned} \tag{S17}$$

Recall that the support of $P_{i,j}$ and $Q_{i,j}$ are in $\bar{\Psi}_{x_i}(\mathcal{W}_{i,j})$. Furthermore, since

$$\operatorname{diam}\operatorname{pr}_M \bar{\Psi}_{x_i}(\mathcal{W}_{i,j}) \lesssim \sigma^{\alpha_0},$$

there holds, using (28), that $\|\operatorname{pr}_{T_y} - \operatorname{pr}_{T_{y'}}\|_{\mathrm{op}} \lesssim \sigma^{\alpha_0}$ for any two $y, y' \in \bar{\Psi}_{x_i}(\mathcal{W}_{i,j})$. Let $y \in \bar{\Psi}_{x_i}(\mathcal{W}_{i,j})$ and $x \in \mathcal{H}_A^c$, we now show that $\|x - y\|_{\Sigma^{-1}(y)} \ge A/2$ if A is chosen large enough. First notice again that if $x \notin \mathcal{W}^1_{x_i}$, then $\|x - y\|_{\Sigma^{-1}(y)} \gtrsim \sigma^{-\alpha_0} \wedge \delta^{-1}\sigma^{-\alpha_\perp} \ge A$ for all A so that we can take $x \in \mathcal{W}^1_{x_i}$. Now using (S16) applied to $z = x_{i,j}$, we find that $\|x - y\|_{\Sigma^{-1}(y)} = \|x - x_{i,j}\|_{\Sigma^{-1}(x_{i,j})} + O(1 + \sigma^{2\alpha_0}/(\delta\sigma^{\alpha_\perp})) \ge A/2$, where we used that $\sigma^{2\alpha_0} = o(\delta\sigma^{\alpha_\perp})$, and provided that A is chosen big enough.

This yields that

$$\begin{aligned} \int_{\mathcal{H}_A^c} K_\Sigma P_{i,j} &= \int_{\bar{\Psi}_{x_i}(\mathcal{W}_{i,j}} \int_{\mathcal{H}_A^c} \varphi_{\Sigma(y)}(x - y)\,\mathrm{d}x P_{i,j}(\mathrm{d}y) \\ &= \frac{1}{(2\pi)^{D/2}\sigma^D\delta^{D-d}} \int_{\bar{\Psi}_{x_i}(\mathcal{W}_{i,j})} \int_{\mathcal{H}_A^c} \exp\left\{-\frac{1}{2}\|x - y\|^2_{\Sigma^{-1}(y)}\right\} \mathrm{d}x P(\mathrm{d}y) \\ &\le \frac{1}{(2\pi)^{D/2}} \int_{\|z\|\ge A/2} \exp\left\{-\frac{1}{2}\|z\|^2\right\} \mathrm{d}z \lesssim \exp(-A^2/8), \end{aligned}$$

where we made the variable change $z = \Sigma^{-1/2}(y)(x-y)$ in the second to last inequality. The same holds for $K_\Sigma Q_{i,j}$. Setting $A = (8\log(1/\varepsilon))^{1/2}$ and combining with (S17), $\|K_\Sigma P_{i,j} - K_\Sigma Q_{i,j}\|_1 \lesssim \log^{D/2}(1/\varepsilon)\varepsilon$. We finally obtain that

$$\|K_\Sigma g - K_\Sigma G\|_1 \lesssim \sum_i \alpha_i \sum_{j\in\mathcal{J}_i} \bar{c}_{i,j} \log^{D/2}(1/\varepsilon)\varepsilon \lesssim \log^{D/2}(1/\varepsilon)\varepsilon.$$

Also we can choose the atoms of G to be $\sigma^{\alpha_0} \wedge \delta\sigma^{\alpha_\perp} \times \varepsilon$ apart, thanks to Lemma S4.1 together with the following bound on $\Sigma(z)^{-1}\Sigma(y) - \mathrm{Id}$ when $\|z-y\| \le \sigma^{\alpha_0} \wedge \delta\sigma^{\alpha_\perp}\varepsilon$:

$$\begin{aligned}\Sigma(z)^{-1}\Sigma(y) - \mathrm{Id} &= O_z^\top \Delta_{\sigma,\delta}^{-2} O_z O_y^\top \Delta_{\sigma,\delta}^2 O_y - \mathrm{Id}\\ &= O_z^\top \Delta_{\sigma,\delta}^{-2}(O_z O_y^\top - \mathrm{Id})\Delta_{\sigma,\delta}^2 O_y + O_z O_y^\top - \mathrm{Id},\end{aligned}$$

so that

$$\begin{aligned}&\sqrt{\mathrm{tr}(\Sigma(z)^{-1}\Sigma(y) - \mathrm{Id})^2}\\ &\le \sqrt{D}\|\Sigma(z)^{-1}\Sigma(y) - \mathrm{Id}\,\|_{\mathrm{op}} \lesssim \|\Delta_{\sigma,\delta}^{-2}\|_{\mathrm{op}}\|O_z O_y^\top - \mathrm{Id}\,\|_{\mathrm{op}}\|\|\Delta_{\sigma,\delta}^2\|_{\mathrm{op}} + \|O_z O_y^\top - \mathrm{Id}\,\|_{\mathrm{op}}\\ &\lesssim \sigma^{-2\alpha_0} \vee \delta^{-2}\sigma^{-2\alpha_\perp}\|O_z - O_y\|_{\mathrm{op}} \lesssim [\sigma^{-2\alpha_0} \vee \delta^{-2}\sigma^{-2\alpha_\perp}]\|\,\mathrm{pr}_{T_z} - \mathrm{pr}_{T_y}\,\|_{\mathrm{op}}\\ &\quad\lesssim [\sigma^{-2\alpha_0} \vee \delta^{-2}\sigma^{-2\alpha_\perp}]\|z-y\| \lesssim \varepsilon,\end{aligned}$$

where we used (28) in the second to last inequality, ending the proof.

□

S5. Supplement to Section 4.

S5.1. *Additional details on the numerical setting.* The manifolds used in the numerical experiments are given through the following equations:

- *The two circles*: it is the union of $\mathcal{C}_1$ and $\mathcal{C}_2$ with equation $(x-x_i)^2 + (y-y_i)^2 = r_i^2$ for $i \in \{1,2\}$. In the experiements, we chose $(x_1,y_1) = (0,0)$, $(x_2,y_2) = (2,0)$ and $r_1 = r_2 = 2$.
- *The 2D-spiral*: it is given by the parametric embedding

$$\varphi_2 : t \in [0,1] \mapsto \begin{cases} R(\omega t + \theta_0)\cos(\omega t + \theta_0)\\ R(\omega t + \theta_0)\sin(\omega t + \theta_0).\end{cases}$$

  In the numerical studies, we chose $R = 1/2\pi$, $\omega = 7\pi/2$ and $\theta_0 = \pi/2$.
- *The 3D-spiral*: it is given by the parametric embedding

$$\varphi_3 : t \in [0,1] \mapsto \begin{cases} R(\omega t + \theta_0)\cos(\alpha t + \theta_0)\\ R(\omega t + \theta_0)\sin(\alpha t + \theta_0)\\ \nu t.\end{cases}$$

  In the numerical studies, we chose $R = 1/2\pi$, $\omega = 7\pi/2$, $\theta_0 = \pi/2$ and $\nu = 2$.
- *The torus*: it is the set $\mathcal{T}$ given by the equation $(\sqrt{(x-x_0)^2 + (y-y_0)^2} - R)^2 + (z-z_0)^2 = r^2$. In the experiements, we chose $(x_0,y_0,z_0) = (0,0,0)$, $R = 3$ and $r = 1$.

S5.2. *Details on the Gibbs sampling algorithm.* This first section presents the implementation of the partial location-scale prior (8) with Gibbs sampling for a particular choice of measure H.

In order to simplify the notations we will write indifferently Λ as a D-dimensional diagonal matrix or a D-dimensional vector.

We use the latent cluster representation of the $(\mu_i O_i)$ and denote by c_i the cluster allocation of observation i:

$$c_i = c_j \Leftrightarrow y_i \text{ and } y_j \text{ are in the same cluster}$$

and let N be the number of clusters, so that $1 \le c_i \le N$. We write $\phi_c = (\mu_c^*, O_c^*)_{1\le c\le N}$ for the unique values of the cluster parameters.

PROPOSITION S5.1. *In the model (18), we have :*

$$p(\Lambda|Y,(\mu_i,O_i),b) = \bigotimes_{j=1}^{D} \mathrm{Inv}\Gamma\left(a_j + n/2, b_j + \frac{1}{2}\sum_{i=1}^{n}\langle O_i^j|y_i-\mu_i\rangle^2\right)$$

$$p(b|Y,(\mu_i,O_i),\Lambda) = \bigotimes_{j=1}^{D} \Gamma(a_j+1, \kappa_j + \lambda_j^{-1})$$

and :

$$p(\mu_c^*|O_c^*,Y,(c_i),\Lambda) = \mathcal{N}(\cdot|\hat{\mu}_c, A^{-1}),\ c\le N$$

$$p(O_c^*|\mu_c^*,Y,(c_i),\Lambda) = p_{\mathrm{BMF}}(\cdot|S, -\frac{1}{2}\Lambda^{-1}, M_0)$$

where

$$A = \Sigma_0^{-1} + n_c O_c^* \Lambda^{-1}(O_c^*)^T, \qquad n_c = \mathrm{Card}\{i : c_i = c\},$$

$$\hat{\mu}_c = A^{-1}[\Sigma_0^{-1}\mu_0 + O_c^*\Lambda^{-1}(O_c^*)^T \sum_{c_i=c} y_i] \text{ and } S = \sum_{c_i=c}(y_i-\mu_c^*)(y_i-\mu_c^*)^T.$$

PROOF OF PROPOSITION S5.1. • For the variances Λ we have

$$p(\Lambda|Y,(\mu_i,O_i),b) \propto p(\Lambda|b)\prod_i \mathcal{N}(y_i|\mu_i, O_i\Lambda O_i^T) \propto \left[\prod_{j=1}^{D}\lambda_j^{-(a+1)}e^{-b_j/\lambda_j}\right]\prod_{i=1}^{n}\mathcal{N}(y_i|\mu_i,O_i\Lambda O_i^T).$$

But by orthogonality :

$$\begin{aligned}
\mathcal{N}(y_i|\mu_i,O_i\Lambda O_i^T) &\propto \left[\prod_{j=1}^{D}\lambda_j^{-1/2}\right]\exp\left(-\frac{1}{2}\langle y_i-\mu_i|O_i\Lambda^{-1}O_i^T(y_i-\mu_i)\rangle\right)\\
&\propto \left[\prod_{j=1}^{D}\lambda_j^{-1/2}\right]\exp\left(-\frac{1}{2}\langle O_i^T(y_i-\mu_i)|\Lambda^{-1}O_i^T(y_i-\mu_i)\rangle\right)\\
&\propto \left[\prod_{j=1}^{D}\lambda_j^{-1/2}\right]\exp\left(-\frac{1}{2}\sum_{j=1}^{D}\lambda_j^{-1}\langle O_i^j|y_i-\mu_i\rangle^2\right)\\
&\propto \prod_{j=1}^{D}\left[\lambda_j^{-1/2}\exp\left(-\frac{1}{2}\lambda_j^{-1}\langle O_i^j|y_i-\mu_i\rangle^2\right)\right],
\end{aligned}$$

where O_i^j is the j-th column of O_i. Therefore,

$$\begin{aligned}
p(\Lambda|Y,(\mu_i,O_i),b) &\propto \prod_{j=1}^{D}\left[\lambda_j^{-(a+1)}e^{-b_j/\lambda_j}\lambda_j^{-n/2}\exp\left(-\frac{1}{2}\lambda_j^{-1}\sum_{i=1}^{n}\langle O_i^j|y_i-\mu_i\rangle^2\right)\right]\\
&\propto \prod_{j=1}^{D}\left[\lambda_j^{-(a+1+n/2)}\exp\left(-\lambda_j^{-1}\left[b_j+\frac{1}{2}\sum_{i=1}^{n}\langle O_i^j|y_i-\mu_i\rangle^2\right]\right)\right],
\end{aligned}$$

i.e conditionally on the rest of the variables, the $(\lambda_j)_{j=1}^D$ are independent with distribution $\mathrm{Inv}\Gamma(a+n/2, b_j + \frac{1}{2}\sum_{i=1}^n \langle O_i^j | y_i - \mu_i\rangle^2)$.

- For the hyperparameters $(b_j)_{j=1}^D$ we have

$$p(b|Y, (\mu_i, O_i), \Lambda) \propto p(b)p(\Lambda|b) \propto \bigotimes_{j=1}^{D} \Gamma(a_j + 1, \kappa_j + \lambda_j^{-1})$$

i.e conditionally on the rest of the variables, the $b_j's$ are independent with distribution $\Gamma(a_j+1, \kappa_j + \lambda_j^{-1})$.

- For the cluster locations $(\mu_c^*)_{1\le c\le K}$ we have

$$p(\mu_c^*|O_c^*, Y, (c_i), \Lambda) \propto p(\mu_c^*) \prod_{c_i=c} \mathcal{N}(y_i|\mu_c^*, O_c^*\Lambda(O_c^*)^T)$$
$$\propto \exp\left(-\frac{1}{2}\left\{\langle \mu_c^* - \mu_0|\Sigma_0^{-1}(\mu_c^* - \mu_0)\rangle + \sum_{c_i=c}\langle y_i - \mu_c^*|O_c^*\Lambda^{-1}(O_c^*)^T(y_i - \mu_c^*)\rangle\right\}\right),$$

and since the prior $p(\mu_c^*)$ is Gaussian with mean μ_0 and variance Σ_0 the above conditional posterior distribution is a Gaussian distribution with mean $\hat{\mu}_c = A^{-1}[\Sigma_0^{-1}\mu_0 + (O_c^*)\Lambda^{-1}(O_c^*)^T\sum_{c_i=c} y_i]$ and variance $A^{-1} = [\Sigma_0^{-1} + n_c(O_c^*)\Lambda^{-1}(O_c^*)^T]^{-1}$.

- For the cluster orientations $(O_c^*)_{1\le c\le K}$ we have

$$p(O_c^*|\mu_c^*, Y, (c_i), \Lambda) \propto \exp\left(\mathrm{tr}\left\{M_0^T O_c^*\right\}\right) \prod_{c_i=c} \mathcal{N}(y_i|\mu_c^*, O_c^*\Lambda(O_c^*)^T)$$
$$\propto \exp\left(\mathrm{tr}\left\{M_0^T O_c^*\right\}\right) \exp\left(-\frac{1}{2}\sum_{c_i=c}\langle y_i - \mu_c^*|O_c^*\Lambda^{-1}(O_c^*)^T(y_i - \mu_c^*)\rangle\right)$$
$$\propto \exp\left(\mathrm{tr}\left\{M_0^T O_c^* - \frac{1}{2}\sum_{c_i=c}(y_i - \mu_c^*)(y_i - \mu_c^*)^T O_c^*\Lambda^{-1}(O_c^*)^T\right\}\right)$$
$$\propto \exp\left(\mathrm{tr}\left\{M_0^T O_c^* - \frac{1}{2}\Lambda^{-1}(O_c^*)^T S O_c^*\right\},\right).$$

We thus obtain

$$O_c^* \sim p_{\mathrm{BMF}}(O_c^*|S, -\frac{1}{2}\Lambda^{-1}, M_0) \mid \mu_c^*, Y, (c_i), \Lambda.$$

□

Proposition S5.1 is the key to implement algorithm 8 in [11].

REMARK 1. Here we can see that when $\sum_{i=1}^n \langle O_i^j|y_i - \mu_i\rangle^2 = o(n^{3/2})$, $\mathrm{Inv}\Gamma(a+n/2, b_j + \frac{1}{2}\sum_{i=1}^n\langle O_i^j|y_i - \mu_i\rangle^2)$ is tightly concentrated around its mean

$$\frac{b_j + \frac{1}{2}\sum_{i=1}^n\langle O_i^j|y_i - \mu_i\rangle^2}{a + n/2 - 1} \ge \frac{b_j}{a+n/2-1};$$

and if we fix b_j this may induce a rather strong penalization on small values of λ_j for finite n.

Even though to the best of our knownledge no direct samplers are available for p_{BMF}, it is still possible to perform a Gibbs sampling update over the columns of O_c^* (cf [9] and the associated package `rstiefel`). Another more involved (but more efficient) option would be to perform Hamiltonian Monte Carlo via polar expansion as suggested in [10]. With a slight abuse of notation we will write

$$O_c^* \sim p_{\mathrm{BMF,Gibbs}}(\cdot|A,B,C,O_c^*)$$

for a Gibbs sampling scan on the column of O_c^* starting from O_c^*. All in all, this leads us to Algorithm 1 below. It uses the conditional allocation distributions :

$$P(c_i=l|c_{-i},y_i,(\phi_l)_{1\le l\le h}) \propto \begin{bmatrix} n_{-i,c}\mathcal{N}(y_i|\mu_l,O_l\Lambda O_l^T) & \forall 1\le l\le k \\ \frac{\alpha}{m}\mathcal{N}(y_i|\mu_l,O_l\Lambda O_l^T) & \forall k<l\le h \end{bmatrix} \tag{S18}$$

Algorithm 1: Gibbs sampling algorithm for Partial location scale DP mixture

Input : a dataset $Y=(y_i)_{i=1}^n$, a current partition $c=(c_i)_{i=1}^n$, a current state $\phi=(\mu_l^*,O_l^*)_{l=1}^K$ (with K the current number of clusters), current variances $\Lambda=(\lambda_j)_{j=1}^D$, current hyperparameters $(b_j)_{j=1}^D$, a number of auxilliary parameters $m\in\mathbb{N}^*$,
$\alpha>0,\mu_0\in\mathbb{R}^D,\Sigma_0\in\mathcal{S}_D^{++}(\mathbb{R}),M_0\in\mathcal{M}_D(\mathbb{R})$ the parameters for the base distribution of the Dirichlet process, $(a_j)_{j=1}^D$ the prior parameters for the variances

Output: an updated state (c,ϕ,Λ,b)

for $i=1,\dots,n$ **do**
- $k\leftarrow \#\{j\neq i: c_j=c_i\}$; $h\leftarrow k+m$; and label $(c_j)_{j\neq i}$ with $\{1,\dots,k\}$;
- **if** $\exists j\neq i: c_i=c_j$ **then**
  - $(\phi_l)_{k<l\le h} \overset{i.i.d}{\sim} G_0$;
- **else**
  - **if** $\forall j\neq i: c_i\neq c_j$ **then**
    - $c_i\leftarrow k+1$; $(\phi_l)_{k+1<l\le h} \overset{i.i.d}{\sim} G_0$;
  - **end**
- **end**
- $n_{-i,c}\leftarrow \#\{c_j: j\neq i\}$; draw c_i from $\{1,\dots,h\}$ with probability (S18);
  discard the (ϕ_l) not associated with any observations;
- **for** $l\in\{c_1,\dots,c_n\}$ **do**
  - $n_l\leftarrow \#\{i: c_i=l\}$; $A=\Sigma_0^{-1}+n_lO_l^*\Lambda^{-1}(O_l^*)^T$;
    $\hat\mu_l=A^{-1}[\Sigma_0^{-1}\mu_0+O_l^*\Lambda^{-1}(O_l^*)^T\sum_{c_i=l}y_i]$;
  - $\mu_l\leftarrow\mu_l\sim\mathcal{N}(\hat\mu_l,A^{-1})$; $S\leftarrow\sum_{c_i=l}(y_i-\mu_l^*)(y_i-\mu_l^*)^T$;
  - $O_l^*\leftarrow O_l^*\sim p_{BMF,Gibbs}(O_l^*|S,-\frac{1}{2}\Lambda^{-1},M_0,O_l^*)$;
- **end**
- ;
- $(\lambda_j)_{j=1}^D\leftarrow(\lambda_j)_{j=1}^D \overset{ind}{\sim} p(\lambda_j|(y_i),c,(\phi_l))\propto Inv\Gamma(a+n/2,b+\frac{1}{2}\sum_{i=1}^n\langle(O_{c_i}^*)^j|y_i-\mu_{c_i}^*\rangle^2)$;
- $b\leftarrow(b_j)_{j=1}^D \overset{ind}{\sim}\Gamma(a_j+1,\kappa_j+\lambda_j^{-1})$

end

REFERENCES

[1] AAMARI, E. and LEVRARD, C. (2019). Nonasymptotic rates for manifold, tangent space and curvature estimation. *The Annals of Statistics* **47** 177–204.

[2] CANALE, A. and DE BLASI, P. (2017). Posterior asymptotics of nonparametric location-scale mixtures for multivariate density estimation. *Bernoulli* **23** 379–404.

[3] CONSTANTINE, G. and SAVITS, T. (1996). A multivariate Faa di Bruno formula with applications. *Transactions of the American Mathematical Society* **348** 503–520.

[4] DEVROYE, L., MEHRABIAN, A. and REDDAD, T. (2018). The total variation distance between high-dimensional Gaussians. *arXiv preprint arXiv:1810.08693*.

[5] DIVOL, V. (2021). Reconstructing measures on manifolds: an optimal transport approach. *arXiv preprint arXiv:2102.07595*.

[6] FEDERER, H. (1959). Curvature measures. *Transactions of the American Mathematical Society* **93** 418–491.

[7] GHOSAL, S. and VAN DER VAART, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics* 697–723.

[8] GHOSAL, S. and VAN DER VAART, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics* **29** 1233–1263.

[9] HOFF, P. D. (2009). Simulation of the matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics* **18** 438–456.

[10] JAUCH, M., HOFF, P. D. and DUNSON, D. B. (2021). Monte Carlo simulation on the Stiefel manifold via polar expansion. *Journal of Computational and Graphical Statistics* **30** 622–631.

[11] NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics* **9** 249–265.

[12] SHEN, W., TOKDAR, S. T. and GHOSAL, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika* **100** 623–640.

Statement of Authorship for joint/multi-authored papers for PGR thesis
To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Estimating a density near an unknown manifold: a Bayesian nonparametric approach
Publication Status	Published
Publication Details	Berenfeld, R. & Rousseau, *Estimating a density near an unknown manifold: a Bayesian nonparametric approach* (2024)

Student Confirmation

Student Name:	Paul Rosa		
Contribution to the Paper	The order of the authors is alphabetical. I had the idea of the manifold anisotropic model, the theoretical results have been obtained as a collaboration between C. Berenfeld, J. Rousseau and myself, together with the writing of the paper. I also wrote and implemented the Gibbs sampling algorithm.		
Signature Paul Rosa		Date	06/01/2025

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Prof Judith Rousseau		
Supervisor comments	I confirm the contribution statement		
Signature		Date	06/01/2025

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 5

Discussion and future directions

We conclude this thesis by summarising in Section 5.1 the contributions of in Chapters 2 3 & 4, and ending in Section 5.2 by a discussion of possible extensions.

5.1 Summary

In Chapters 2 & 3, we have investigated the asymptotic behaviour of some classes of Bayesian nonparametric procedures in the nonparametric regression model with covariates located on an unknown submanifold of a Euclidean space. We have first focused in Chapter 2 on Gaussian process priors of Matérn type for the unknown regression function f_0, by studying the intrinsic Matérn processes, which are defined by random basis expansions in the eigenbasis of the Laplace-Beltrami operator of the submanifold. Although such a prior is not always tractable since it requires to work directly on the (possibly unknown) submanifold in question—in particular to compute the eigenfunctions of the Laplace-Beltrami operator—it serves as a benchmark that we compare with the so-called extrinsic Matérn processes, defined by the restriction of an ambient, Euclidean, standard Mátern process to the submanifold in question. The extrinsic Matérn processes are precisely the priors used implicitely if one uses a standard Euclidean Matérn process prior without knowing that the covariates are actually supported on a submanifold. Surprisingly, we have shown that both types of priors can lead to optimal posterior contraction rates under smoothness assumption on f_0, even though empirical arguments tend to indicate that intrinsic methods are preferable in practice. Indeed, as opposed to the extrinsic ones, the intrinsic processes do not introduce spurious correlations for the value of the regression function at point being close with respect to the ambient distance but far away with respect to the Euclidean distance.

In Chapter 3 we have considered an alternative solution to the problem considered in Chapter 2. In a possibly semi-supervised setting, we have constructed an approximation to the intrinsic processes over the submanifold by a random basis expansion the eigenbasis of the graph Laplacian of a random geometric graph constructed over the covariates. Since the construction of the random geometric graph is always tractable—as it only necessitates the knowledge of the pairwise Euclidean distance between the covariates—this provides a graph based prior that can be interpreted as a numerical approximation to the original intrinsic processes. As opposed to Chapter 2 we have not limited ourselves to Gaussian processes but actually consider general priors on the basis coefficients. Under Hölder smoothness assumption on f_0 and for appropriate choice of hyperparameters we have shown near minimax optimal posterior contraction rates for such priors. Moreover, we have explained how to construct Bayesian procedures adaptive to the smoothness of f_0 by considering hierarchical priors, and where the hyperparameters are data-driven. The proof of our posterior contraction rates crucially relies on a novel approximation argument for the restriction of smooth functions to random geometric graphs in the graph Laplacian eigenbasis, which can be used in other contexts (for instance we deduce a similar rate of convergence for a least squares estimator from our approximation result).

In Chapter 4 we have considered the problem of estimating a probability density f_0 supported on a small tubular neighbourhood of a submanifold of $\mathbb{R}^D$ based on an *i.i.d.* $n-$sample. By constructing a new class nonparametric Gaussian mixture models (where the prior on the covariance matrices explicitely separates orientation and eigenvalues) and a new class of manifold anisotropic Hölder functions (where the smoothness anisotropy is measured along a coordinate frame moving along the submanifold), we have described how such priors can lead to minimax optimal posterior contraction rates for the estimation of f_0 under the Hellinger distance, provided that the width δ_n of the tubular neighbourhood is not too small. In particular, a width δ_n satisfying $\delta_n \to 0$ is compatible with our assumptions, provided that this decay is not too fast. We have also shown that the density f_0 belongs to our new anisotropic smoothness class in natural additive noise models. Finally we have complemented our theoretical analysis by an MCMC and MAP (maximum a posteriori) implementation of our methods, demonstrating a clear advantage over standard priors when the probability distribution is nearly singular.

5.2 Limitations and future directions

There are several ways in which our contributions can be extended. First of all, in Chapter 3 we have provided a solution to the problem of adaptive estimation by considering a hierarchical prior, which adapts to the unknown smoothness of the regression function but (potentially) not to the intrinsic dimension (in the case of the global loss $\|\cdot\|_N$). Since we do not present lower bounds it is not clear whether or not this is an artefact of the proof and if these hierarchical priors can actually adapt both to the smoothness and the intrinsic dimension when considering the $\|\cdot\|_N$ loss. The issue stems from the fact that, in order for our proof to go through, the truncation level J in the random basis expansion and the connectivity parameter h of the random geometric graph must satisfy an inequality of the type $h \lesssim J^{-1/d}$ a posteriori, ignoring log factors. As pointed out in the paper, a simple solution could be to plug in a consistent intrinsic dimension estimator as in [17], but even if this solution is satisfactory theoretically this may result in suboptimal behaviour in the unfortunate case where the dimension estimator fails. Thus refining our proof strategy in order to design priors adaptive both to the smoothness and the intrinsic dimension could be of interest.

The adaptation problem was not considered in Chapter 2. When using the intrinsic processes, it is assumed that the submanifold is known (since we need to access the Laplace-Beltrami eigenpairs), and therefore the intrinsic dimension is considered known too. The adaptation to the smoothness can then be achieved using standard approaches, for instance by considering a hierarchical random truncation as in Chapter 3. For the extrinsic processes, the results contained in [117] & [106] indicate that it may be possible to adapt simultaneously to the smoothness of the regression function and the intrinsic dimension, by putting an appropriate empirical Bayes prior on the ambient Matérn process length scale.

Moreover, as we pointed out in Chapter 2, looking at the posterior contraction rates alone (i.e. the power of n appearing in the rate) was not able to discriminate between intrinsic and extrinsic processes, even though they demonstrate different behaviours in practice. It may be interesting to investigate in which way the posterior contraction rate depends of the submanifold itself, in particular making the dependence of the $\mathcal{O}$ constant in the extrinsic curvature explicit.

Our result regarding the extrinsic processes in Chapter 2 is a direct consequence of the trace and extension theorem between submanifold and ambient Sobolev spaces ([97] Section 10.2). It turns out that a similar theorem exists for Besov spaces and an analog result for a class of rescaled extrinsic Besov priors defined by random basis expansion in a wavelet basis [3, 57] can then be proved using the spatial localisation of the basis under Besov smoothness assumption on the true regression function.

For the density estimation problem studied in Chapter 4, a lot of questions remain unanswered. For instance, our posterior contraction rates are minimax optimal provided that the width of the tubular neighbourhood does not decay too quickly, but it is not clear if the corresponding threshold is optimal or not. In particular, it may be that this lower bound condition on the width is a consequence of our proof technique (and therefore potentially an artefact of the proof itself) and/or can be lowered when considering other kernels for the mixture models. A natural candidate is the Fisher-Gaussian kernel, as it can potentially adapt better to non linear submanifolds thanks to the curved shape of the kernel itself. Using this kernel might therefore help to weaken the condition on the width, but no formal results are known yet.

In Chapter 4 we focused on posterior contraction rates with respect to the Hellinger distance, which is compatible with our model of probability distribution supported in a tubular neighbourhood around a submanifold (since in this case the model is dominated by the Lebesgue measure). Still in the *i.i.d.* sampling model, if the true probability distribution of the observations is supported on an unknown (possibly non smooth) subset of $[0,1]^D$ of Minkowski dimension $d \leq D$, using a Dirichlet process prior (which is conjugate for the likelihood), it is possible to prove a result similar to the one obtained in [115] for the empirical distribution and deduce near minimax optimal posterior contraction rates under the Wasserstein$-p$ metrics with automatic adaptation to the intrinsic dimension d by taking advantage of the explicit form of the posterior distribution. For more involved models and priors however, the prior mass and testing approach presented in Section 1.3 of Chapter 1 might not lead to (near) minimax rates in general, and no alternative proofs giving near optimal rates are known so far. Therefore the question of optimal Wasserstein posterior contraction rates is still an open problem.