Contraction rates and projection subspace estimation with Gaussian process priors in high dimension

Elie Odin, François Bachoc, Agnès Lagnoux

Contraction rates and projection subspace estimation with Gaussian process priors in high dimension

Elie Odin[1], **François Bachoc**[2], and **Agnès Lagnoux**[3]

[1]Institut de Mathématiques de Toulouse; UMR5219. Université de Toulouse; CNRS. UT3, F-31062 Toulouse, France.
elie.odin@math.univ-toulouse.fr

[2]Institut de Mathématiques de Toulouse; UMR5219. Université de Toulouse; CNRS. UT3, F-31062 Toulouse, France.
francois.bachoc@math.univ-toulouse.fr

[3]Institut de Mathématiques de Toulouse; UMR5219. Université de Toulouse; CNRS. UT2J, F-31058 Toulouse, France.
lagnoux@univ-tlse2.fr

02 2023

Abstract

This work explores the dimension reduction problem for Bayesian nonparametric regression and density estimation. More precisely, we are interested in estimating a functional parameter f over the unit ball in $\mathbb{R}^d$, which depends only on a d_0-dimensional subspace of $\mathbb{R}^d$, with $d_0 < d$. It is well-known that rescaled Gaussian process priors over the function space achieve smoothness adaptation and posterior contraction with near minimax-optimal rates. Moreover, hierarchical extensions of this approach, equipped with subspace projection, can also adapt to the intrinsic dimension d_0 ([Tok11]). When the ambient dimension d does not vary with n, the minimax rate remains of the order $n^{-\beta/(2\beta+d_0)}$. However, this is up to multiplicative constants that can become prohibitively large when d grows. The dependences between the contraction rate and the ambient dimension have not been fully explored yet and this work provides a first insight: we let the dimension d grow with n and, by combining the arguments of [Tok11] and [JT21], we derive a growth rate for d that still leads to posterior consistency with minimax rate. The optimality of this growth rate is then discussed. Additionally, we provide a set of assumptions under which consistent estimation of f leads to a correct estimation of the subspace projection, assuming that d_0 is known.

1 Introduction

With the ever-increasing availability of high-dimensional data in various fields of science and technology, dimension reduction methods have become more and more important, especially in non-parametric estimation, to counteract the curse of dimensionality. Suppose we want to estimate an unknown function $f : \mathbb{R}^d \to \mathbb{R}$ that depends only on a d_0-dimensional linear subspace $\mathcal{S} \subset \mathbb{R}^d$, with $d_0 \ll d$. For regression and density estimation problems, minimax rates without sparsity assumptions are both of the order $n^{-\beta/(2\beta+d)}$ where β is the smoothness of f and n is the sample size ([Bir86], [Sto82]). The aim of dimension reduction is to convert this d-dimensional problem into a d_0-dimensional one in order to obtain the way more attractive rate $n^{-\beta/(2\beta+d_0)}$.

As the above rates are given up to a multiplicative constant, which may itself depend on the *ambient dimension* d, another problem arises: determining if the number of available data is sufficient in regard to the problem's dimension. This is generally done by allowing the ambient dimension d to grow with n, letting $d = d_n$, and then observing which growth rate still permits minimax estimation at rate $n^{-\beta/(2\beta+d_0)}$. Note that the subspace $\mathcal{S}$ also depends on n, thus we write $\mathcal{S} = \mathcal{S}_n$.

For fixed *intrinsic dimension* d_0, we distinguish two cases, whether the subspace $\mathcal{S}_n$ is parallel to the axes or not. In the first case (when $\mathcal{S}_n$ is parallel to the axes), the dimension-reduction problem is referred to as *variable selection*. In this context, it is known that for non-parametric regression, the sparsity pattern can be consistently recovered when d_n grows exponentially with the sample size ([CD12], [YT15]). More precisely, [CD12] show that there exist two constants $c_* < c^*$ such that

- if $\frac{\log d_n}{n} < c_*$, there exists a consistent estimator of the sparsity pattern,
- if $\frac{\log d_n}{n} > c^*$, no such estimator exists.

This phase transition phenomenon seems to be similar in the linear regression framework (see [Ver12] and [Wai09]).

In the second case (when nothing is assumed on $\mathcal{S}_n$), the estimation of a minimal subspace which contains all the information on f is sometimes referred to as *sufficient dimension reduction* ([Coo98]). Among the various methods proposed for estimating $\mathcal{S}_n$, sliced inverse regression (SIR) ([Li91]) is one of the most studied. The first article including the framework of growing ambient dimension d_n shows the consistency of SIR only under $d_n = O(n^{1/2})$ ([ZMP06]). Later, [LZL18] show that the phase transition phenomenon occurs at a growth rate d_n in $o(n)$. In other words, SIR-based estimators are consistent only if $d_n/n \underset{n\to+\infty}{\longrightarrow} 0$ and this growth rate appears to be optimal ([Lin+21]).

The difference between growth rates encountered in variable selection and in sufficient dimension reduction has led recently to the emergence of methods combining both approaches. If f depends on a d_0-dimensional subspace $\mathcal{S}_n$ which can be described by linear combination of only a small number of variables, then we can perform both variable selection and sufficient dimension reduction over the selected variables. This method is studied for example in [Lin+21], [LZL19], [TSY20], and [ZMZ22] and allows a return to the exponential growth of the dimension d_n.

The aim of this article is to perform both function and subspace estimation in the case where no hypotheses are made on $\mathcal{S}_n$ and to derive the maximum dimension growth rate. Our analysis is done in the nonparametric Bayesian framework introduced by [GGV00]. Among the advantages of this approach, the use of very versatile priors, such as Gaussian processes [VV08a], allows to perform smoothness and dimension adaptability at near minimax rates ([VV09], [TZG10], [JT21]) with a single Bayesian procedure, and avoids the complications associated with kernel methods (see for example the introduction of [STG13]).

The work of Tokdar, Zhu, and Ghosh [TZG10] is one of the first to include a hierarchical prior with a parameter on the subspace.They use a uniform prior on the Grassmannian of dimension d_0 and a logistic Gaussian process prior for the conditional density function. The authors are able to derive posterior consistency for both the conditional density function and the subspace but they do not provide contraction rates. Near minimax contraction rates are then derived in [Tok11] by extending the framework introduced by [VV09]. Finally, [JT21] show that for variable selection, the estimation of the regression function and that of the sparsity pattern can be realized simultaneously at near minimax rates even with dimension d_n growing exponentially with the sample size. The growth rate is linked to the smoothness β of f via $\log(d_n) = O(n^{d_0/(2\beta+d_0)})$.

The paper is organized as follows. In Section 2, we introduce a hierarchical Gaussian process-based prior for both regression and density estimation models. This prior consists of a dimension parameter for d_0, an invariant prior over linear d_0-dimensional subspaces of $\mathbb{R}^{d_n}$, a d_0-dimensional Gaussian process, and a rescaling parameter to ensure smoothness adaptability. Our first result (Theorem 3.1 in Section 3) shows that, for the estimation problem of f, near minimax contraction rates can be achieved for dimensions d_n growing not faster than $n^{d_0/(2\beta+d_0)}$ which is interestingly the already mentioned growth rate where we drop out the exponential. We are not able to prove the optimality of this result but some clues are given below (see Remark 5.2); notably, this growth rate is equivalent to n when $\beta \to 0$, which is known to be the breakpoint of the consistency of the SIR estimator. In Section 4, we show that for fixed ambient dimension d, the hierarchical Bayes procedure contracts to a subspace that contains $\mathcal{S}$ and we conjecture that this subspace is exactly $\mathcal{S}$. Our estimation result of f combines the standard arguments used in [Tok11] and [JT21], which are based on [VV09]. To prove the contraction around the central subspace $\mathcal{S}$, we show that an error on the estimation of $\mathcal{S}$ leads to an error on the estimation of f from which we obtain a contradiction on the previously established minimax estimation of f. The proofs of the main results (Theorems 3.1 and 4.1) are postponed to Appendices 5.1 and 5.2 while Appendix 5.3 is dedicated to useful lemmas.

2 Problem formulation

2.1 Notation and definitions

The abundant technical notation used throughout this article make this section very useful. We begin with the definition of standard functional spaces.

Let K be a bounded convex subset of $\mathbb{R}^d$, with $d \in \mathbb{N}^*$. For $\alpha > 0$, write $\alpha = k + r$ with k a nonnegative integer and $r \in (0, 1]$. The *Hölder space* $\mathfrak{C}^\alpha(K)$ is the space of all functions $f : K \to \mathbb{R}$ that are k-times differentiable and whose partial derivatives of order $(k_1, \dots, k_d)$, with $k_1, \dots, k_d$ nonnegative integers such that $k_1 + \cdots + k_d = k$, are Lipshitz functions of order r, that is, there exists a constant D such that

$$\left| \frac{\partial^k}{\partial_1^{k_1} \cdots \partial_d^{k_d}} (f(x) - f(y)) \right| \leq D \left\| x - y \right\|^r ,$$

for all pairs $x, y \in K^2$ and where $\|\cdot\|$ is the Euclidean norm.

We use the following asymptotic notation: if f and g are two real functions over an arbitrary set S, then we write $f \lesssim g$ if there exists a constant c such that $|f(s)| \leq c \cdot |g(s)|$ for all $s \in S$. The notation $\gtrsim$ is defined in the same way and we write $f \asymp g$ when both $f \lesssim g$ and $f \gtrsim g$ hold.

To model the central subspace $\mathcal{S}$, we will use isometries instead of the Grassmannian. For $d \in \mathbb{N}^*$, we denote by $\mathcal{O}_d$ the space of linear isometries over $\mathbb{R}^d$. In addition, the introduction of canonical

subspaces and of "component filters" notation will be very convenient when dealing with the sparsity. For $x \in \mathbb{R}^d$ and $\mathbf{v} \in \{0,1\}^d$, we denote by $|\mathbf{v}|$ the number of ones in $\mathbf{v}$, by $x_{\mathbf{v}} := (x_j : v_j = 1,\ 1 \leq j \leq d) \in \mathbb{R}^{|\mathbf{v}|}$ the sub-vector with components selected according to $\mathbf{v}$, and for $y \in \mathbb{R}^{|\mathbf{v}|}$, by $y^{\mathbf{v}} := (\tilde{y}_j)_{1 \leq j \leq d}$ the vector in $\mathbb{R}^d$ with $\tilde{y}_j = 0$ if $v_j = 0$ and $\tilde{y}_j = y_i$ if v_j is the i-th one in $\mathbf{v}$.

Moreover, for any integer $b \in [\![1, d]\!]$, we denote by $\mathbf{b}$ the vector $\sum_{i=1}^{b} e_i$, where $\{e_i : 1 \leq i \leq d\}$ is the canonical basis on $\mathbb{R}^d$. The dimension d of the ambient space is implicit in this notation.

Finally, for $\mathbf{v} \in \{0,1\}^d$, we denote by $E_{\mathbf{v}}$ the linear span of $\{e_i : v_i = 1\}$ and by $E_{1-\mathbf{v}}$ the linear span of $\{e_i : v_i = 0\}$. Clearly, $E_{1-\mathbf{v}}$ is the orthogonal complement of $E_{\mathbf{v}}$.

The proof of Theorem 3.1 involves measuring the complexity of the space where the prior puts its mass. This measure is carried out via *metric entropy*. Given a subset B of a metric space (E, d) and a radius $\varepsilon > 0$, we can define the following numbers:

- the *ε-packing number* $D(\varepsilon, B, d)$ is the maximum number of points in B such that the distance between every pair is at least ε,
- the *ε-covering number* $N(\varepsilon, B, d)$ is the minimum number of balls of radius ε needed to cover B.

The logarithms of the packing and the covering number are called the *entropy* and the *metric entropy* respectively.

2.2 Bayesian framework for density estimation and regression

Our main result will be stated for two statistical settings: *density estimation* and *fixed* or *random design regression with Gaussian error*. As we will work with subspaces that are not orthogonal with the axes, the usual support $[0,1]^d$ for the density or the regression function will be replaced by the unit ball $\mathbb{U}_d := \{x \in \mathbb{R}^d, \|x\| \leq 1\}$. For a given number of observations n, the density or the regression function will be characterized by a functional parameter $f_n^* : \mathbb{U}_{d_n} \to \mathbb{R}$. The ambient dimension d_n is allowed to grow with n but f_n^* is supposed to depend only on a subspace $\mathcal{S}_n$ with fixed dimension d_0. A prior on d_0 and on the subspace itself will be later introduced to ensure the dimension adaptability. The prior on the true parameter f_n^* will consist of a projected Gaussian random variable W_n with values in the Banach space $(\mathcal{C}(\mathbb{U}_{d_n}), \|\cdot\|_\infty)$. Now let us describe the two previously introduced statistical settings.

Density estimation. Suppose we observe an i.i.d. sample $X_1, \ldots, X_n$ from a law P_n^* over $\mathbb{U}_{d_n}$, which admits a continuous density p_n^* relative to the Lebesgue measure on $\mathbb{R}^{d_n}$. The prior W_n puts its mass on a space that is far too large compared to the space of continuous densities. So to correctly retrieve p_n^*, we will work with the parametrized density p_{n,W_n} where, for $w \in \mathcal{C}(\mathbb{U}_{d_n})$,

$$p_{n,w}(x) := \frac{e^{w(x)}}{\int_{\mathbb{U}_{d_n}} e^{w(x)} dx}. \tag{2.1}$$

Here the exponential forces the prior to charge only nonnegative functions while the renormalization ensures that $p_{n,w}$ integrates to one. The true density p_n^* will then be encoded by the parameter $f_n^* \in \mathcal{C}(\mathbb{U}_{d_n})$ such that $p_n^* = p_{n,f_n^*}$. In this way, all the assumptions on the true parameter f_n^* can be transferred to the density p_n^*. That is, p_n^* is supposed to depend only on the d_0-dimensional subspace $\mathcal{S}_n$ of $\mathbb{R}^{d_n}$.

The natural metric between two densities p and p' is the Hellinger distance defined by $h(p,p') = \left\|\sqrt{p}-\sqrt{p'}\right\|_2$, where $\|\cdot\|_2$ is the L^2-norm with respect to the Lebesgue measure. Consequently, if the parameter space is embedded with a prior Π_n, we will say that the posterior *contracts* to p_n^* at rate $(\varepsilon_n)_{n\in\mathbb{N}}$ if, for any sufficiently large constant M,

$$\mathbb{P}_n^*\left[\Pi_n\left(f\in\mathcal{C}(\mathbb{U}_{d_n}) : h(p_{n,f},p_n^*) > M\varepsilon_n \mid X_1,\dots,X_n\right)\right] \underset{n\to+\infty}{\longrightarrow} 0, \tag{2.2}$$

where $\mathbb{P}_n^*$ is the joint law of $(X_1,\dots,X_n)$.

Regression with Gaussian error. In a regression problem, the covariates can be either predetermined for each observation, this is the *fixed design* case, or can be part of the observation themselves. In the later case, the covariates can be considered as random; this corresponds to the *random design* case. The notion of posterior contraction differs slightly between these two situations and some clarifications are in order.

Fixed design. In this setting, we consider a sample of n real observations $Y_1,\dots,Y_n$ satisfying the model $Y_i = f_n^*(x_i)+\epsilon_i$, with $\epsilon_i\sim\mathcal{N}(0,\sigma^2)$ where the $x_i\in\mathbb{U}_{d_n}$ for $i\in[\![1,n]\!]$ are n fixed covariates and where the ϵ_i are n i.i.d. univariate Gaussian random variables with zero mean and standard deviation σ. As previously, the regression function $f_n^*:\{x_i : i\in[\![1,n]\!]\}\to\mathbb{R}$ is supposed to depend only on a d_0-dimensional subspace of $\mathbb{R}^{d_n}$.

We will use W_n directly as a prior for the regression function because W_n can be viewed by restriction as a Gaussian process over the space $\mathcal{X}_n := \{x_i : i\in[\![1,n]\!]\}$ of design points. To quantify the posterior contraction, we introduce the design dependent semi-metric $\|\cdot\|_n$ defined as the $L^2(\mathbb{P}_n^x)$-norm for the empirical measure $\mathbb{P}_n^x = n^{-1}\sum_{i=1}^n\delta_{x_i}$ of the design points. If the space of regression functions over $\mathcal{X}_n$ is embedded with a prior Π_n, we will say that the posterior *contracts* to f_n^* at rate $(\varepsilon_n)_{n\in\mathbb{N}}$ if, for any sufficiently large constant M,

$$\mathbb{P}_n^*\left[\Pi_n\left(f\in\mathcal{C}(\mathcal{X}_n) : \|f-f_n^*\|_n > M\varepsilon_n \mid Y_1,\dots,Y_n\right)\right] \underset{n\to+\infty}{\longrightarrow} 0, \tag{2.3}$$

where $\mathbb{P}_n^*$ is the joint law of $(Y_1,\dots,Y_n)$.

Random design. Here, we observe n i.i.d. pairs $(X_1,Y_1),\dots,(X_n,Y_n)$ such that $Y_i = f_n^*(X_i)+\epsilon_i$, with i.i.d. $\epsilon_i\sim\mathcal{N}(0,\sigma^2)$, $\sigma\in[1,2]$, and where the X_i's are random covariates over $\mathbb{U}_{d_n}$ independent of the ϵ_i's and admitting a common density G_n that is bounded away from zero. For the sake of simplicity, the standard deviation σ is restricted to the interval $[1,2]$ but these bounds can be relaxed, see Remark 5.1.1 for details. Again, the regression function $f_n^*:\mathbb{U}_{d_n}\to\mathbb{R}$ is supposed to depend only on a d_0-dimensional subspace of $\mathbb{R}^{d_n}$. Moreover, we use W_n directly as a prior for the regression function. The natural metric for this problem is the $L^2(G_n)$-norm denoted by $\|\cdot\|_{2,G_n}$ where G_n is identified with the law of one covariate. This metric is not equivalent to the Hellinger metric, which is used in the proof of Theorem 3.1, unless all regression functions are uniformly bounded by a constant $Q>0$. This condition can be fulfilled by projecting the prior on the space of all functions uniformly bounded by Q, as proposed in [GN11], but this would force us to rewrite the proof of Theorem 3.1 only for this setting. Instead, we directly post-process the posterior to integrate this constraint as in [YD16]. Then, the formulation of posterior consistency becomes as follows. Considering a prior Π_n over the regression functions, we will say that the posterior *contracts* to f_n^* at rate $(\varepsilon_n)_{n\in\mathbb{N}}$ if, for

$Q > 0$ and any sufficiently large constants M,

$$\mathbb{P}_n^* \left[\Pi_n \left(f \in \mathcal{C}(\mathbb{U}_{d_n}) : \left\| f^Q - f_n^{*Q} \right\|_{2,G_n} > M\varepsilon_n \mid (X_1, Y_1), \ldots, (X_n, Y_n) \right) \right] \underset{n \to +\infty}{\longrightarrow} 0, \tag{2.4}$$

where $\mathbb{P}_n^*$ is the joint law of $(X_1, Y_1), \ldots, (X_n, Y_n)$ and where $f^Q := (f \vee -Q) \wedge Q$ is the truncated version of f.

3 Main result for the functional parameter

In order for the true parameter f_n^* to be recovered, we suppose that its restriction to the d_0-dimensional subspace $\mathcal{S}_n$ does not depend on the ambient dimension d_n.

Assumption 3.1 (Sparsity of the true parameter)**.** There exist $n_0, d_0 \in \mathbb{N}$, $f_0 \in \mathcal{C}(\mathbb{U}_{d_0})$, and a sequence of linear isometries $q_n^* \in \mathcal{O}_{d_n}$ such that for all $n \geq n_0$, we have $d_n \geq d_0$, and $f_n^*(x) = f_0\left((q_n^*(x))_{\mathbf{d_0}}\right)$, for all $x \in \mathbb{U}_{d_n}$.

In this way, each f_n^* can be viewed as a sparse continuation in dimension d_n of an underlying fixed function f_0 called the *core function*. The use of isometries instead of vector subspaces permits us to avoid the manipulation of the Grassmannian. We will use instead the more convenient orthogonal group $\mathcal{O}_{d_n}$. The next property is straightforward.

Property 3.1. For $n \geq n_0$, f_n^* is constant on the intersection between $\mathbb{U}_{d_n}$ and the affine subspaces $(q_n^*)^{-1}(E_{1-\mathbf{d_0}}) + x$, for $x \in \mathbb{R}^{d_n}$.

In parallel to the dimension adaptability, the present setting allows the core function f_0 to be arbitrarily smooth (in a Hölder sense) while maintaining near-minimax contraction rates.

Assumption 3.2 (Smoothness of f_0)**.** There exists $\beta > 0$ such that $f_0 \in \mathfrak{C}^\beta(\mathbb{U}_{d_0})$.

3.1 Prior specification

Here we specify the hierarchical prior on the parameter space. The true parameter f_n^* is characterized by a sparsity pattern (d_0, q_n^*), where the intrinsic dimension d_0 is the one of the relevant subspace and $q_n^* \in \mathcal{O}_{d_n}$ is an isometry for the orientation; its smoothness is modeled by a rescaling parameter, and the core function f_0 is modeled by a standard squared exponential Gaussian process which has infinitely smooth sample paths. Indeed, this process has proven to be fruitful in combination with a scale parameter and allows smoothness adaptation (see [VV09]).

For $n > 0$, let $W = (W(x) : x \in \mathbb{R}^{d_n})$ be a standard squared exponential Gaussian process on $\mathbb{R}^{d_n}$; that is, a centered Gaussian process with covariance kernel

$$\mathbb{E}[W(s)W(t)] \;=\; \exp\bigl(-\|s-t\|^2\bigr), \qquad \text{for all } s, t \in \mathbb{R}^{d_n},$$

where $\|\cdot\|$ is the Euclidean norm.

Let $a > 0$, $b \in [\![1, d_n]\!]$, and $q \in \mathcal{O}_{d_n}$. We define $W_x^{a,b,q} := W(a\,\mathrm{Diag}(\mathbf{b}) \cdot q(x))$ and $W^{a,b,q} := (W_x^{a,b,q} : x \in \mathbb{R}^{d_n})$ a rescaled Gaussian process with sparsity pattern (b, q), where $\mathrm{Diag}(\mathbf{b})$ is the diagonal matrix with diagonal vector $\mathbf{b}$. Then, the process $W^{a,b,q}$ is constant on affine subspaces $q^{-1}(E_{1-\mathbf{b}}) + x$, for

$x \in \mathbb{R}^{d_n}$ (as in Property 3.1) and if $R := q^{-1}\,\mathrm{Diag}(\mathbf{b})q$ is the orthogonal projection onto $q^{-1}(E_{\mathbf{b}})$, then $W_x^{a,b,q} = W_{Rx}^{a,b,q}$, for all $x \in \mathbb{R}^{d_n}$.

To work properly with $W^{a,b,q}$, we have to verify that its law identifies with the law of a b-dimensional standard squared exponential process. To do so, define

$$\begin{aligned} \phi : \mathbb{R}^b &\to q^{-1}(E_{\mathbf{b}}) \\ x &\mapsto \frac{1}{a} q^{-1}(x^{\mathbf{b}}), \end{aligned}$$

a bijection with inverse $\phi^{-1}(t) = a(qt)_{\mathbf{b}}$ for $t \in q^{-1}(E_{\mathbf{b}})$. Then, $W_{\phi(x)}^{a,b,q} = W(x^{\mathbf{b}})$ for all $x \in \mathbb{R}^b$.

Let us introduce $\widetilde{W} := (W_{\phi(x)}^{a,b,q}, x \in \mathbb{R}^b)$. Then, for all $x, y \in \mathbb{R}^b \times \mathbb{R}^b$, we have

$$\mathbb{E}[\widetilde{W}(x)\widetilde{W}(y)] \;=\; \mathbb{E}[W(x^{\mathbf{b}})W(y^{\mathbf{b}})] \;=\; e^{-\|x^{\mathbf{b}} - y^{\mathbf{b}}\|^2}.$$

So $\widetilde{W}$ is a standard squared exponential Gaussian process in dimension b that does not depend on a nor q. Moreover, we have $W_t^{a,b,q} = \widetilde{W}(\phi^{-1}(Rt))$.

From now on, $W^{a,b,q}$ will refer to the restriction on $\mathbb{U}_{d_n}$ of this process. Then, the hierarchical prior on the parameter $f \in \mathcal{C}(\mathbb{U}_{d_n})$ with stochastic subspace selection is defined as the law Π_n of $W^{A,\Gamma,\Theta}$, where A is the scaling parameter, $\Gamma \in [\![1, d_n]\!]$ is the prior on the subspace dimension, and Θ is the prior on the orientation.

Assumption 3.3. The intrinsic dimension d_0 of the subspace is assumed to be bounded by a known deterministic number $d_{\max}$.

Consequently, Γ is defined by a probability vector $(\pi_\Gamma(d) : 1 \le d \le d_{\max})$ with $\pi_\Gamma(d) > 0$ for all d. Moreover, we define the scaling parameter A such that there exists a collection of probability measures $\pi_{n,d}$ on $(0, \infty)$, $0 \le d \le d_{\max} \wedge d_n$, with $A \mid (\Gamma = d) \sim \pi_{n,d}$. We require the law of the stochastic isometry Θ to be translation invariant. That is, for all subset $\mathcal{Q}$ of $\mathcal{O}_{d_n}$ and for all $q \in \mathcal{O}_{d_n}$, we need $\mathbb{P}(\Theta \in q \cdot \mathcal{Q}) = \mathbb{P}(\Theta \in \mathcal{Q})$. Therefore, the law of Θ is taken as the unit Haar measure on $\mathcal{O}_{d_n}$, the only probability measure that is translation invariant on $\mathcal{O}_{d_n}$. In addition, all A, Γ, and Θ are supposed to be independent of W.

For convenience, the notation $\pi_{n,d}$ will refer to a probability measure as well as its density.

Assumption 3.4 (Rescaling measures)**.** There exist constants D_1, D_2, C_1, C_2, and $c > 1$ such that for all $n \in \mathbb{N}^*$ and $d < d_{\max} \wedge d_n$, the density $\pi_{n,d}$ satisfies

1. for all sufficiently large a, $\pi_{n,d}(a) \ge D_1 e^{-C_1 a^d (\log a)^{d+1}}$;
2. for all $a > c$, $\pi_{n,d}(a) \le D_2 e^{-C_2 a^d (\log a)^{d+1}}$;
3. $\pi_{n,d}([0, c]) = 0$.

Assumptions similar to Assumption 3.4 are standard, see for instance Equation (3.4) in [VV09] or Assumption 5 [JT21]. For example, this assumption is satisfied if, for all $n \in \mathbb{N}^*$ and $d < d_{\max} \wedge d_n$, $A^d(\log A)^{d+1} \mid (\Gamma = d)$ is the restriction to $(c, +\infty)$ of an exponential law with parameter independent of d and n (indeed, if $g(A)$ has density function f, with g differentiable and strictly increasing, then A has density function $(f \circ g) \cdot g'$).

The next section gives some precision about the reproducing kernel Hilbert space (RKHS) of $W^{a,b,q}$. The content is a bit technical and can be skipped at first reading.

3.2 Reproducing kernel Hilbert space of $W^{a,b,q}$

One of the advantages of choosing a Gaussian process prior is that the contraction rate depends explicitly on the small ball probability and on the relative position of the parameter with respect to the RKHS associated with the process. This section is dedicated to the basic properties of this space. For elementary definitions and for some precision about the link between the contraction rate and the RKHS, we refer the reader to [VV08a] and [VV08b].

Notation. We denote by $\mathcal{C}(\mathbb{U}_d \mid q^{-1}(E_{\mathbf{b}}))$ the space of continuous functions on $\mathbb{U}_d$ which are constant on affine subspaces $q^{-1}(E_{1-\mathbf{b}}) + x$, for $x \in \mathbb{U}_d$.

We introduce the operator

$$\Lambda : \begin{cases} \mathcal{C}(\mathbb{U}_b) \to \mathcal{C}(\mathbb{U}_d \mid q^{-1}(E_{\mathbf{b}})) \\ f \mapsto \Lambda f : \begin{cases} \mathbb{U}_d \to \mathbb{R} \\ x \mapsto f((qx)_{\mathbf{b}}), \end{cases} \end{cases}$$

so that $W^{a,b,q} = \Lambda(\widetilde{W}^a)$, where $\widetilde{W}^a = (\widetilde{W}_{at}, t \in \mathbb{U}_b)$ is the process $\widetilde{W}$ introduced above rescaled by a and restricted to $\mathbb{U}_b$. It is a bijective linear map and also an isometry if the domain and the codomain are endowed with the uniform norm. In particular, the map Λ is continuous. According to Lemma 7.1 in [VV08b], if $\widetilde{\mathbb{H}}_a$ is the RKHS of $\widetilde{W}^a$, then the RKHS $\mathbb{H}_{a,b,q}$ of $W^{a,b,q}$ is equal to $\Lambda(\widetilde{\mathbb{H}}_a)$. Let us detail its elements. The stochastic process RKHS of $\widetilde{W}^a$ (as defined in [VV08b]) is composed of functions $h : \mathbb{U}_b \to \mathbb{R}$ for which there exists $\psi \in L^2_{\mathbb{C}}(\mu^{se}_{a,b})$ such that

$$h(t) = \mathfrak{Re} \int_{\mathbb{R}^b} e^{-i\lambda \cdot t} \psi(\lambda) d\mu^{se}_{a,b}(\lambda), \quad t \in \mathbb{U}_b, \tag{3.1}$$

where $\mu^{se}_{a,b}$ is the spectral measure of the a-rescaled squared exponential process in dimension b with spectral density $f^{se}_{a,b} : t \mapsto (2a\sqrt{\pi})^{-b} \exp(-\frac{1}{4}\|t/a\|^2)$ (see Lemma 4.1 in [VV09], and the following discussion). We can view $\widetilde{W}^a$ as a random Gaussian element with values in the Banach space $(\mathcal{C}(\mathbb{U}_b), \|\cdot\|_\infty)$. Thus, according to Theorem 2.1 in [VV08b], the stochastic process RKHS and the Banach space RKHS coincide and we can apply Lemma 7.1 from the same reference. The space $\mathbb{H}_{a,b,q} = \Lambda(\widetilde{\mathbb{H}}_a)$ is then the set of functions

$$\overline{h} : x \in \mathbb{U}_d \mapsto \mathfrak{Re} \int_{\mathbb{R}^b} e^{-i\langle \lambda, (qx)_{\mathbf{b}} \rangle} \psi(\lambda) d\mu^{se}_{a,b}(\lambda), \tag{3.2}$$

where ψ runs through $L^2_{\mathbb{C}}(\mu^{se}_{a,b})$ and the RKHS norm is $\|\overline{h}\|_{\mathbb{H}_{a,b,q}} = \|\psi\|_{L^2(\mu^{se}_{a,b})}$.

We remark that functions of the RKHS of $W^{a,b,q}$ have the same sparsity-pattern as the trajectories of $W^{a,b,q}$.

Remark 3.1. Functions $\overline{h} \in \mathbb{H}_{a,b,q}$ are constant on affine subspaces $q^{-1}(E_{1-\mathbf{b}}) + x$ for $x \in \mathbb{U}_d$.

As mentioned at the beginning of this section, contraction rates under Gaussian process prior depend on two quantities: the small ball probability and the relative position of the parameter with respect to the RKHS. For a parameter $f \in \mathcal{C}(\mathbb{U}_d \mid q^{-1}(E_{\mathbf{b}}))$ and $\varepsilon > 0$, these two quantities define the *concentration function* $\phi^{a,b,q}_f$, with

$$\phi^{a,b,q}_f(\varepsilon) := \inf_{h \in \mathbb{H}_{a,b,q} : \|h - f\|_\infty < \varepsilon} \|h\|^2_{\mathbb{H}_{a,b,q}} - \log \mathbb{P}\left(\left\| W^{a,b,q} \right\|_\infty < \varepsilon \right). \tag{3.3}$$

3.3 Posterior consistency

Before we state the theorem, we need a last assumption, which determines how the ambient dimension d_n is allowed to grow with the sample size n.

Assumption 3.5 (Growth of d_n)**.** The ambient dimension d_n satisfies

$$d_n \;\leq\; C_D \cdot n^{\frac{d_0}{2\beta+d_0}} \cdot (\log n)^{2\kappa-1},$$

for some small constant $C_D > 0$ and where $\kappa = (d_0+1)\beta/(2\beta+d_0)$.

An examination of κ shows that $\kappa \geq 1/2$ if $\beta \geq 1/2$ and that $\kappa > \beta$ otherwise. Thereby, a standard rate of order $n^{1/2}$ for d_n is achieved with parameter $\beta = d_0/2$. The fastest rate tends to the order $n \cdot (\log n)^{-1}$ when β tends to zero. Although it is always possible to set β extremely close to zero in order to obtain the best rate for d_n, one should keep in mind that the contraction rate may then be suboptimal, as discussed at the end of this section.

Theorem 3.1. *Let $\varepsilon_n = C_\varepsilon \cdot \underline{\varepsilon}_n (\log n)^\kappa$ with $\underline{\varepsilon}_n = n^{-\beta/(2\beta+d_0)}$, C_ε a large constant that depends on f_0, and κ as in Assumption 3.5. Then, if the parameter space is embedded with the prior Π_n and under Assumptions 3.1-3.5, the posterior contracts at rate $(\varepsilon_n)_{n\in\mathbb{N}}$ for density estimation (as defined in (2.2)) as well as for regression with fixed or random design (as defined in (2.3) and (2.4)).*

An examination of ε_n shows that the contraction rate is improved as the smoothness β of f_0 grows, unlike d_n. This highlights a trade-off between the contraction rate and the growth of the design dimension: fast contraction rates imply slowly increasing dimension and conversely.

The proof of Theorem 3.1, postponed in the Appendix, in Section 5.1, combines the arguments of [Tok11] and [JT21].

4 Subspace recovery for the density estimation problem

In this section, we propose to recover the central subspace for the density estimation problem. To avoid identifiability issues caused by the spherical support, we suppose that the ambient dimension d_n does not depend on n. Hence, we denote the ambient dimension by d with $d \geq d_0$ and the central subspace by $\mathcal{S} := (q^*)^{-1}(E_{\mathbf{d_0}})$ where q^* corresponds to q_n^* in Assumption 3.1. This assumption is justified by the following considerations. If the ambient dimension grows with n, the Hellinger metric relative to the Lebesgue measure on $\mathbb{U}_{d_n}$ tends to give more importance to the center of the support, as n tends to infinity. For example, consider a parameter $f_0 : \mathbb{U}_2 \to \mathbb{R}$ in dimension two that is everywhere constant except in a small region on the border of $\mathbb{U}_2$, and such that the central subspace $\mathcal{S}_n$ is of dimension two. The importance of this small region in the support $\mathbb{U}_{d_n}$, in the Hellinger sense, decreases exponentially with n, way faster than the estimation of the true parameter f_n^* in Theorem 3.1. Consequently, for sufficiently large n, a constant function $f_0 : [0,1] \to \mathbb{R}$ together with some one-dimensional subspace $\mathcal{S}'$ characterize a density that is in the Hellinger ball of radius ε_n centered on f_n^*; so we have no hope of recovering the true subspace by simply using the posterior consistency.

As a consequence, the true density p^*, the parameter f^*, and the central subspace do not depend on n anymore. The true density $p^* = p_{f^*}$ is characterized by f^* via the transformation (2.1). Moreover, f^* is supposed to depend only on a d_0-dimensional subspace of $\mathbb{R}^d$ and can be viewed as the sparse continuation of an underlying function $f_0 \in \mathcal{C}(\mathbb{U}_{d_0})$. In the same way, p^* can be viewed as the sparse continuation of a function p_0 over $\mathbb{U}_{d_0}$, except that the renormalisation of p^* depends on d. Note that

p_0 is not necessarily a density on $\mathbb{U}_{d_0}$ so the notation $h(f,g)$ will designate from now on the L^2-distance between the square roots of f and g even if f and g are not densities.

Let us introduce a few more notation. Let $\mathcal{Q}^*$ be the set of all optimal isometries:

$$\mathcal{Q}^* \;:=\; \{q \in \mathcal{O}_d : q^{-1}(E_{\mathbf{d_0}}) = \mathcal{S}\},$$

and, for $d' > d_0$, let $\mathcal{Q}^*_{d'}$ be the set of isometries that send the subspace $E_{\mathbf{d}'}$ to a subspace containing $\mathcal{S}$:

$$\mathcal{Q}^*_{d'} \;:=\; \{q \in \mathcal{O}_d : q^{-1}(E_{\mathbf{d}'}) \supset \mathcal{S}\}.$$

Recovering $\mathcal{S}$ means the following: for some rate $\delta_n \to 0$,

$$\mathbb{P}^*_n \left[\Pi_n \left(\Gamma \neq d_0 \text{ or } \min_{q \in \mathcal{Q}^*} |||\Theta - q||| \geq \delta_n \mid X_1, \dots, X_n \right) \right] \underset{n \to +\infty}{\longrightarrow} 0,$$

where $|||\cdot|||$ is the operator norm with respect to the Euclidean distance in $\mathbb{R}^d$. However, under the assumptions of Theorem 3.1, the only information we have on the true subspace is posterior consistency to the density p^* with rate ε_n. This will only allow us to recover a subspace of $\mathbb{R}^d$ containing $\mathcal{S}$. A crucial assumption to eliminate the subspaces of dimension smaller than d_0 and the subspaces that do not contain $\mathcal{S}$ is to suppose that p_0 is non-constant in all directions. More precisely, the default of constancy for each direction has to be detectable in Hellinger distance, as formalized in the following assumption.

Assumption 4.1. There exist a constant D and a window size $L < 1$ such that for all vector line Δ in $\mathbb{R}^{d_0}$ (directed by a unit vector $\mathbf{\Delta}$), there exists $o \in \mathcal{B}_{d_0}(1-L)$ such that for all $0 < l \leq L$, for all $t \in \mathcal{B}_{d_0}(L/2) + o$, and for all constant $c > 0$,

$$h^2(p_{0|I}; c) \;\geq\; D \cdot l^2,$$

where $I :=]o + t - \frac{l}{2}\mathbf{\Delta}; o + t + \frac{l}{2}\mathbf{\Delta}[$.

Assumption 4.1 seems a bit technical at first glance but it can be shown that it is satisfied as soon as p_0 is differentiable over $\mathbb{U}_{d_0}$ with d_0 points such that the gradients at these points are linearly independent.

Theorem 4.1. *Under Assumption 4.1 and the assumptions of Theorem 3.1, we have, for some rate $(\delta_n)_n$ tending to zero,*

$$\Pi_n \left(\Gamma < d_0 \mid X_1, \dots, X_n \right) \underset{n \to +\infty}{\longrightarrow} 0, \qquad \textit{in } \mathbb{P}^*_n\textit{-probability}, \tag{4.1}$$

$$\Pi_n \left(\Gamma = d_0 \textit{ and } \min_{q \in \mathcal{Q}^*} |||\Theta - q||| \geq \delta_n \mid X_1, \dots, X_n \right) \underset{n \to +\infty}{\longrightarrow} 0, \qquad \textit{in } \mathbb{P}^*_n\textit{-probability}, \tag{4.2}$$

$$\Pi_n \left(\Gamma > d_0 \textit{ and } \min_{q \in \mathcal{Q}^*_\Gamma} |||\Theta - q||| \geq \delta_n \mid X_1, \dots, X_n \right) \underset{n \to +\infty}{\longrightarrow} 0, \qquad \textit{in } \mathbb{P}^*_n\textit{-probability}. \tag{4.3}$$

Theorem 4.1 ensures that the central subspace $\mathcal{S}$ can be recovered as soon as the intrinsic dimension d_0 is known. Subspaces of dimension smaller than d_0 are also eliminated but the theorem does not reject those of dimension greater than d_0. We conjecture that the prior mass on those spaces tends to vanish, for reasons similar to those exposed in [JT21]. Indeed, introducing a penalization on larger

dimensions if necessary, it should be possible to show that the posterior cannot contract as fast as the minimax rate for d_0 if a subspace of greater dimension is chosen. As discussed in the introduction of this section, the estimation of the central subspace is made under the assumption that d is fixed with n mainly because of the identifiability issue caused by the ellipsoid support. We believe that this restriction can be relaxed by extending the support $\mathbb{U}_d$ to the full ambient space $\mathbb{R}^d$, as in [JT21]. In this case, the square over which we integrate the Hellinger distance in the proof of Theorem 4.1 can be taken as the product space of a square of side L in directions $\mathbf{\Delta}$ and $\mathbf{\Lambda}$ times $\mathbb{R}^{d-2}$. Then, the integrated error should no longer depend on d and consistency to the true subspace should follow. Further investigations in this direction might be worthwhile.

The proof of Theorem 4.1 is postponed in Appendix 5.2.

Acknowledgments. We acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-21-CE40-0007 (GAP Project).

5 Appendix

5.1 Proof of Theorem 3.1

As a reminder, we first exhibit some facts about the convergence rate:

$$\varepsilon_n \;=\; C_\varepsilon \cdot n^{-\frac{\beta}{2\beta+d_0}} \cdot (\log n)^{\kappa}, \qquad n\varepsilon_n^2 \;=\; C_\varepsilon^2 \cdot n^{\frac{d_0}{2\beta+d_0}} \cdot (\log n)^{2\kappa}. \tag{5.1}$$

So ε_n is a large multiple of the minimax rate times a logarithm factor. The constant C_ε is chosen to be arbitrarily large in order to absorb undesired terms in the proof.

The proof of Theorem 3.1 is based on Theorem 2.1 in [GGV00]. The general outline is a combination of the arguments of [Tok11] (itself derived from [VV09]) and [JT21]. Concretely, it suffices to show that there exists a sequence of sets $\mathbb{B}_n \subset \mathcal{C}(\mathbb{U}_{d_n})$ (referred to as a *sieve*), such that the following three conditions hold for all sufficiently large n:

$$\Pi_n\left(\left\|W^{A,\Gamma,\Theta} - f_n^*\right\|_\infty \le 2\varepsilon_n\right) \;\ge\; \exp(-n\varepsilon_n^2), \tag{5.2}$$

$$\Pi_n\left(W^{A,\Gamma,\Theta} \notin \mathbb{B}_n\right) \;\le\; \exp(-5n\varepsilon_n^2), \tag{5.3}$$

$$\log N\left(3\varepsilon_n, \mathbb{B}_n, \|\cdot\|_\infty\right) \;\le\; n\varepsilon_n^2. \tag{5.4}$$

This is the purpose of the next sections. The first condition (5.2), referred to as *prior mass condition*, ensures that the prior puts a sufficient amount of mass around the true parameter. Condition (5.3), called *sieve condition*, forces the sieve $\mathbb{B}_n$ to capture most of the mass of the prior, while the *entropy condition* (5.4) constrains its size. These three conditions map one to one with the conditions of Theorem 2.1 in [GGV00], as showed in [VV08a] for density estimation and regression with fixed design. For regression with random design, we recall in the next section some arguments spread out in Bayesian literature.

5.1.1 Regression with random design

Here, we show that Theorem 2.1 in [GGV00] can be applied in the regression with random design setting, as soon as Conditions (5.2), (5.3), and (5.4) are satisfied. The procedure consists in showing

that the posterior contracts to the density of a pair (X_i, Y_i) and then to retrieve f_n^* from this density. For a function $f : \mathbb{U}_{d_n} \to \mathbb{R}$, we define $P_f : \mathbb{U}_{d_n} \times \mathbb{R} \to \mathbb{R}_+$, $(x, y) \mapsto G_n(x) \cdot \Phi_{f(x),\sigma}(y)$, where $\Phi_{\mu,\sigma}$ is the density of a univariate Gaussian variable with mean μ and standard deviation σ and G_n is the density of one covariate. Then, the density of one observation (X, Y) under regression with random design is $P_{f_n^*}$. We first prove that Condition (5.2) implies Condition (2.4) in [GGV00] with $C = 1$. Detailed calculations can be found in [FS23], Section A.2. We have to compare the uniform neighborhood of f_n^* with the Kullback-Leibler neighborhood

$$B_2(P_{f_n^*}; \varepsilon) := \{g : \mathrm{KL}(P_{f_n^*}, P_g) \leq \varepsilon^2,\ V_{2,0}(P_{f_n^*}, P_g) \leq \varepsilon^2\},$$

where $\mathrm{KL}(P_f, P_g) := P_f\left[\log(dP_f/dP_g)\right]$ is the Kullback-Leibler divergence between P_f and P_g and $V_{2,0}(P_f, P_g) := P_f\left[\log(dP_f/dP_g) - \mathrm{KL}(P_f, P_g)\right]^2$ is the Kullback-Leibler variation. Using the following identities from [FS23]

$$\begin{aligned}
\mathrm{KL}(P_f, P_g) &= \frac{1}{2\sigma^2}\|f - g\|_{2,G_n}^2, \\
V_2(P_f, P_g) &:= P_f\left[\log\left(\frac{dP_f}{dP_g}\right)^2\right] = \frac{1}{\sigma^2}\|f - g\|_{2,G_n}^2 + \left(\frac{1}{2\sigma^2}\left\|(f-g)^2\right\|_{2,G_n}\right)^2, \\
V_{2,0}(P_f, P_g) &= V_2(P_f, P_g) - \mathrm{KL}(P_f, P_g)^2,
\end{aligned}$$

we deduce that, if $\|f - g\|_\infty \leq 2\varepsilon$ with $2\varepsilon < 1$, then

$$\begin{aligned}
\mathrm{KL}(P_f, P_g) &\leq \frac{1}{2\sigma^2}\|f - g\|_\infty^2 \leq \frac{2\varepsilon^2}{\sigma^2}, \\
V_{2,0}(P_f, P_g) &\leq 4C_\sigma^2 \cdot \varepsilon^2,
\end{aligned}$$

where $C_\sigma := \sqrt{1/\sigma^2 + 1/(4\sigma^4)}$. Consequently, according to (5.2), and multiplying ε_n by $4C_\sigma^2$ if necessary, we have

$$\Pi_n\left(B_2(P_{f_n^*}; \varepsilon_n)\right) \geq \exp\left(-\frac{1}{4C_\sigma^2} n\varepsilon_n^2\right).$$

One can remark that for Condition (2.4) in [GGV00] to be satisfied, we must have $(4C_\sigma^2)^{-1} \leq 1$ which is the case as soon as $\sigma \leq 2$.

Condition (2.3) in [GGV00] is immediately deduced from (5.3). For Condition (2.4), we use the inequality

$$h(P_f, P_g) \leq \frac{1}{2\sigma}\|f - g\|_\infty, \tag{5.5}$$

see again [FS23] for details. Then, assuming that $\sigma \geq 1$, we have, according to (5.4) and multiplying ε_n by 3 if necessary,

$$D\left(\varepsilon_n, \mathbb{B}_n, h\right) \leq N\left(\frac{\varepsilon_n}{2}, \mathbb{B}_n, h\right) \leq N\left(\frac{\varepsilon_n}{2\sigma}, \mathbb{B}_n, h\right) \leq N\left(\varepsilon_n, \mathbb{B}_n, \|\cdot\|_\infty\right) \leq \exp(n\varepsilon_n^2),$$

where the first inequality comes from the definition of the packing number D and the covering number N and where the third inequality follows from (5.5). Theorem 2.1 in [GGV00] then ensures posterior consistency to $P_{f_n^*}$ at rate ε_n in Hellinger distance. Now, because we also have the converse inequality

$$h^2(P_f, P_g) \geq \frac{1}{4\sigma^2}\exp\left(-\frac{Q^2}{2\sigma^2}\right) \cdot \|f - g\|_{2,G_n}^2 \qquad \text{if } \|f\|_\infty \leq Q \text{ and } \|g\|_\infty \leq Q$$

and that $h(P_{f^Q}, P_{g^Q}) \leq h(P_f, P_g)$ when nothing is assumed on f and g with $f^Q = (f \vee -Q) \wedge Q$, we obtain posterior contraction to f_n^* at rate ε_n in the $L^2(G_n)$-distance:

$$\mathbb{P}_n^* \left[\Pi_n \left(g \in \mathcal{C}(\mathbb{U}_{d_n}) : \left\| f_n^{*Q} - g^Q \right\|_{2,G_n} > D_\sigma^Q \cdot \varepsilon_n \mid (X_1, Y_1), \ldots, (X_n, Y_n) \right) \right] \underset{n \to +\infty}{\longrightarrow} 0,$$

where $D_\sigma^Q := M \cdot 2\sigma \cdot \exp\left(Q^2/(4\sigma^2)\right)$.

Remark 5.1. The restriction to $[1, 2]$ for the standard deviation σ can be relaxed. In fact, if $\sigma > 2$, then it suffices to consider Theorem 2.1 in [GGV00] with $C = (4C_\sigma^2)^{-1}$. Condition (2.4) in [GGV00] is then immediately satisfied and, for Condition (2.3), the proof of (5.3) can be adapted to replace 5 by $4 + C$. On the contrary, if $0 < \sigma < 1$, Condition (2.2) in [GGV00] can be satisfied by multiplying ε_n by σ^{-1}.

5.1.2 Prior mass condition (5.2)

We verify here that $\Pi_n \left(\left\| W^{A,\Gamma,\Theta} - f_n^* \right\|_\infty \leq 2\varepsilon_n \right) \geq \exp(-n\varepsilon_n^2)$. Let us introduce the following notation.

Notation. For $q \in \mathcal{O}_{d_n}$, we denote by $f_{n,q} : \mathbb{U}_{d_n} \to \mathbb{R}$ the function such that $f_{n,q}(x) = f_0\left((qx)_{\mathbf{d_0}}\right)$, for all $x \in \mathbb{U}_{d_n}$. Hence, $f_n^* = f_{n,q_n^*}$.

We first reduce the problem to deterministic dimension and direction by conditioning with $\Gamma = d_0$ and integrating over $\mathcal{O}_{d_n}$:

$$\Pi_n \left(\left\| W^{A,\Gamma,\Theta} - f_n^* \right\|_\infty \leq 2\varepsilon_n \right) \geq \pi_\Gamma(d_0) \int_{\mathcal{O}_{d_n}} \Pi_n \left(\left\| W^{A,d_0,q} - f_n^* \right\|_\infty \leq 2\varepsilon_n \right) dq.$$

Now, we want to bound from below the integrand on a significant subset of $\mathcal{O}_{d_n}$. We remark that if $q \in \mathcal{O}_{d_n}$ is such that $\|f_n^* - f_{n,q}\|_\infty \leq \varepsilon_n$, then

$$\Pi_n \left(\left\| W^{A,d_0,q} - f_n^* \right\|_\infty \leq 2\varepsilon_n \right) \geq \Pi_n \left(\left\| W^{A,d_0,q} - f_{n,q} \right\|_\infty \leq \varepsilon_n \right).$$

We show that the right-hand side is bounded from below by $\exp(-\frac{1}{2}n\varepsilon_n^2)$ and then, we bound from below the measure of the set of $q \in \mathcal{O}_{d_n}$ satisfying $\|f_n^* - f_{n,q}\|_\infty \leq \varepsilon_n$.

From now on, we use without specification the constants of Lemmas 5.5, 5.6, and 5.7 and we fix $a_0 > 1$. Let $a \in [K_n, 2K_n]$ where $K_n = \left(\frac{2C_{f_0}}{\varepsilon_n}\right)^{1/\beta}$. We suppose n large enough so that $\varepsilon_n/2 < \min(\varepsilon_0^{a_0,d_0}; C_{f_0} \cdot a_0^{-\beta}; 1/2)$. Then,

$$K_n > \left(\frac{C_{f_0}}{C_{f_0} \cdot a_0^{-\beta}} \right)^{1/\beta} = a_0, \tag{5.6}$$

and, because $a \geq \left(\frac{2C_{f_0}}{\varepsilon_n}\right)^{1/\beta}$, we have

$$\frac{\varepsilon_n}{2} \geq C_{f_0} \cdot a^{-\beta}. \tag{5.7}$$

According to Lemma 5.3 in [VV08b], for $q \in \mathcal{O}_{d_n}$, we can write

$$\begin{aligned}\Pi_n\left(\left\|W^{A,d_0,q} - f_{n,q}\right\|_\infty \le \varepsilon_n\right) &\ge \int_{K_n}^{2K_n} \Pi_n\left(\left\|W^{a,d_0,q} - f_{n,q}\right\|_\infty \le \varepsilon_n\right) \pi_{n,d_0}(a)da \\ &\ge \int_{K_n}^{2K_n} \exp\left(-\phi^{a,d_0,q}_{f_{n,q}}(\varepsilon_n/2)\right) \pi_{n,d_0}(a)da,\end{aligned}$$

where $\phi^{a,d_0,q}_{f_{n,q}}$ is the concentration function in (3.3). Now we want to control the concentration function using Lemmas 5.5 and 5.7. The inequality (5.6) and the previous restriction on n ensure that the conditions of Lemma 5.7 are satisfied with $\varepsilon = \varepsilon_n/2$, while (5.7) and Lemma 5.5 give

$$\inf\left\{\left\|\overline{h}\right\|^2_{\mathbb{H}_{a,d_0,q}} : \overline{h} \in \mathbb{H}_{a,d_0,q},\ \left\|\overline{h} - f_{n,q}\right\|_\infty \le \varepsilon_n/2\right\} \ \le\ D_{f_0} \cdot a^{d_0}.$$

Using the expression (3.3) of the concentration function, a combination of the two lemmas gives

$$\begin{aligned}\phi^{a,d_0,q}_{f_{n,q}}(\varepsilon_n/2) &\le D_{f_0} \cdot a^{d_0} + C_{a_0,d_0} \cdot a^{d_0} \log(2a/\varepsilon_n)^{d_0+1} \\ &= \left(D_{f_0} \log(2a/\varepsilon_n)^{-d_0-1} + C_{a_0,d_0}\right) a^{d_0} \log(2a/\varepsilon_n)^{d_0+1} \\ &\le \left(D_{f_0} \log(a_0)^{-d_0-1} + C_{a_0,d_0}\right) a^{d_0} \log(2a/\varepsilon_n)^{d_0+1},\end{aligned}$$

where the last inequality holds because $a \ge a_0$ and $\varepsilon_n \le 1$ for n large enough. Let us define the constant $C_{f_0,a_0,d_0} := D_{f_0} \log(a_0)^{-d_0-1} + C_{a_0,d_0}$ and note that there exists a constant C'_{f_0} such that for sufficiently large n, $\log(4K_n/\varepsilon_n) = \log\left(4(2C_{f_0})^{1/\beta}\varepsilon_n^{-(1+1/\beta)}\right) \le C'_{f_0}\log(1/\varepsilon_n)$. Then, there exists a constant C'_{f_0,a_0,d_0} such that

$$\begin{aligned}\int_{K_n}^{2K_n} \exp\left(-\phi^{a,d_0,q}_{f_{n,q}}(\varepsilon_n/2)\right) \pi_{n,d_0}(a)da &\ge \int_{K_n}^{2K_n} \exp\left(-C_{f_0,a_0,d_0} \cdot a^{d_0} \log(2a/\varepsilon_n)^{d_0+1}\right) \pi_{n,d_0}(a)da \\ &\ge \exp\left(-C_{f_0,a_0,d_0}(2K_n)^{d_0} \log(4K_n/\varepsilon_n)^{d_0+1}\right) \pi_{n,d_0}(2K_n) \\ \text{(Assumption 3.4)} \qquad &\ge \exp\left(-C'_{f_0,a_0,d_0} \cdot \varepsilon_n^{-d_0/\beta} \log(1/\varepsilon_n)^{d_0+1}\right).\end{aligned}$$

With the help of the reminder (5.1), we see that

$$\varepsilon_n^{-\frac{d_0}{\beta}} = C_\varepsilon^{-\frac{d_0}{\beta}} \cdot n^{\frac{d_0}{2\beta+d_0}} \cdot (\log n)^{-\frac{(d_0+1)d_0}{2\beta+d_0}} \quad \text{and} \quad \left(\log\frac{1}{\varepsilon_n}\right)^{d_0+1} < (\log n)^{d_0+1},$$

hence $\varepsilon_n^{-d_0/\beta}\log(1/\varepsilon_n)^{d_0+1} < C_\varepsilon^{-d_0/\beta} n\varepsilon_n^2$. Then, by choosing C_ε such that $C_\varepsilon^{d_0/\beta} \ge 2C'_{f_0,a_0,d_0}$, we can achieve

$$\Pi_n\left(\left\|W^{A,d_0,q} - f_{n,q}\right\|_\infty \le \varepsilon_n\right) \ \ge\ \exp\left(-\frac{C'_{f_0,a_0,d_0} \cdot n\varepsilon_n^2}{C_\varepsilon^{d_0/\beta}}\right) \ \ge\ \exp\left(-\frac{1}{2}n\varepsilon_n^2\right). \tag{5.8}$$

At this point, the problem amount to bound from below the measure of the set of $q \in \mathcal{O}_{d_n}$ satisfying $\|f_n^* - f_{n,q}\|_\infty \le \varepsilon_n$. We denote by $\mathcal{A}_{\varepsilon_n}$ this set. The core function f_0 is continuous on the compact

subset $\mathbb{U}_{d_0}$, so there exists a constant $D_1 > 0$ such that f_0 is β-Hölder with Hölder constant D_1. Then, for all $q, q' \in \mathcal{O}_{d_n}$,

$$\|f_{n,q'} - f_{n,q}\|_\infty \;=\; \sup_{x \in \mathbb{U}_{d_n}} |f_0\left((q'x)_{\mathbf{d_0}}\right) - f_0\left((qx)_{\mathbf{d_0}}\right)| \;\leq\; D_1 \cdot |||q' - q|||^\beta.$$

From now on, it is apparently sufficient to compute the measure of a ball in $\mathcal{O}_{d_n}$ with radius $(\varepsilon_n/D_1)^{1/\beta}$. In fact, $B_{\mathcal{O}_{d_n}}\left(q_n^*,\ (\varepsilon_n/D_1)^{1/\beta}\right) \subset \mathcal{A}_{\varepsilon_n}$. However, this leads to a design dimension d_n not larger than $n^{d_0/(4\beta+2d_0)}$. To obtain d_n of order $n^{d_0/(2\beta+d_0)}$, we have to consider a larger subset.

Notation. Let $F \subset \mathbb{R}^{d_n}$ be a linear subspace of $\mathbb{R}^{d_n}$. We denote by $\mathcal{O}_{d_n}(F)$ the set of isometries that fix F:

$$\mathcal{O}_{d_n}(F) \;:=\; \{q' \in \mathcal{O}_{d_n} : q'_{|F} = \mathrm{Id}\}.$$

Then, for all $q' \in \mathcal{O}_{d_n}((q_n^*)^{-1}(E_{\mathbf{d_0}}))$, we have

$$f_{n,q_n^* q'} \;=\; f_{n,q_n^*} \circ q' \;=\; f_{n,q_n^*} \qquad \text{and} \qquad \|f_n^* - f_{n,q}\|_\infty \;=\; \left\|f_{n,q_n^* q'} - f_{n,q}\right\|_\infty \;\leq\; D_1 \cdot |||q_n^* q' - q|||^\beta.$$

For $\varepsilon > 0$, we define

$$\mathcal{Q}_{q_n^*,\varepsilon} \;:=\; \{q \in \mathcal{O}_{d_n} : \exists q' \in \mathcal{O}_{d_n}((q_n^*)^{-1}(E_{\mathbf{d_0}})),\ |||q_n^* q' - q||| \leq \varepsilon\}.$$

Then, $\mathcal{Q}_{q_n^*,(\varepsilon_n/D_1)^{1/\beta}} \subset \mathcal{A}_{\varepsilon_n}$. Since the Haar measure is translation invariant, it is sufficient to cover $\mathcal{O}_{d_n}$ with translations of $\mathcal{Q}_{q_n^*,\varepsilon}$ to obtain a lower bound on the measure of $\mathcal{Q}_{q_n^*,\varepsilon}$, that is, to cover $\mathcal{O}_{d_n}$ with sets $\overline{q}\mathcal{Q}_{q_n^*,\varepsilon}$ where $\overline{q}$ belongs to some net $\mathcal{R} \subset \mathcal{O}_{d_n}$ and then remark that $\mathbb{P}(\Theta \in \mathcal{Q}_{q_n^*,\varepsilon}) \geq 1/|\mathcal{R}|$.

Lemma 5.1. *We have,*

$$\mathbb{P}(\Theta \in \mathcal{Q}_{q_n^*,\varepsilon}) \;\geq\; \left(\frac{2}{\pi d_n}\right)^{\frac{d_0}{2}} \cdot \left(\frac{\varepsilon}{16\sqrt{d_0 d_n}}\right)^{d_0(d_n-1)}.$$

Proof of Lemma 5.1. Let $q'' \in \mathcal{O}_{d_n}$. The first step consists in constructing a net $\mathcal{R} \subset \mathcal{O}_{d_n}$ such that there exist $\overline{q} \in \mathcal{R}$ and $q \in \mathcal{Q}_{q_n^*,\varepsilon}$ with $q'' = \overline{q}q$. Let $(u_1, \ldots, u_{d_0}, u_{d_0+1}, \ldots, u_{d_n})$ be an orthonormal basis adapted to the direct sum $\mathbb{R}^{d_n} = (q_n^*)^{-1}(E_{\mathbf{d_0}}) \overset{\perp}{\bigoplus} (q_n^*)^{-1}(E_{1-\mathbf{d_0}})$.

For all d_0-tuple of orthonormal vectors $g = (g_1, \ldots, g_{d_0})$, we fix $r_g \in \mathcal{O}_{d_n}$ an isometry such that $r_g(q_n^* u_i) = g_i$ for all $i \in [\![1, d_0]\!]$. Moreover, we denote by $\mathcal{G}$ a set of d_0-tuples of orthonormal vectors in $\mathbb{R}^{d_n}$ such that, for all d_0-tuples $f = (f_1, \ldots, f_{d_0})$ of orthonormal vectors, there exists $g \in \mathcal{G}$ satisfying

$$\sup_{i \in [\![1,d_0]\!]} \|g_i - f_i\| \;\leq\; \frac{\varepsilon}{2\sqrt{d_0 d_n}}.$$

We claim that we can take $\mathcal{R} := \{r_g : g \in \mathcal{G}\}$. Indeed, there exists $g \in \mathcal{G}$ such that

$$\sup_{i \in [\![1,d_0]\!]} \|g_i - q''(u_i)\| \;\leq\; \frac{\varepsilon}{2\sqrt{d_0 d_n}}.$$

By Lemma 5.8, we can extend g in an orthonormal basis of $\mathbb{R}^{d_n}$ such that

$$\sup_{j \in [\![1,d_n]\!]} \|g_j - q''(u_j)\| \;\leq\; \frac{\varepsilon}{\sqrt{d_n}}. \tag{5.9}$$

Then, writing $\overline{q} = r_g$ and taking q such that $q(u_j) = r_g^{-1}(q'' u_j)$ for all $j \in [\![1, d_n]\!]$, we have $q'' = \overline{q} q$. Moreover, because $r_g^{-1}(g_j) \in E_{1-\mathbf{d_0}}$ and $(q_n^*)^{-1} r_g^{-1}(g_j) \in (q_n^*)^{-1}(E_{1-\mathbf{d_0}})$ for $j \in [\![d_0 + 1, d_n]\!]$, we can define q' such that

$$\begin{cases} q'(u_i) = u_i, & \text{if } i \in [\![1, d_0]\!], \\ q'(u_j) = (q_n^*)^{-1} r_g^{-1}(g_j), & \text{if } j \in [\![d_0 + 1, d_n]\!]. \end{cases}$$

Then, we have $q' \in \mathcal{O}_{d_n}((q_n^*)^{-1}(E_{\mathbf{d_0}}))$ and according to (5.9),

$$\|q_n^* q'(u_i) - q(u_i)\| \;=\; \left\|q_n^*(u_i) - r_g^{-1}(q'' u_i)\right\| \;=\; \|r_g q_n^*(u_i) - q''(u_i)\| \;\leq\; \frac{\varepsilon}{\sqrt{d_n}}, \qquad \text{for } i \in [\![1, d_0]\!],$$

and,

$$\|q_n^* q'(u_j) - q(u_j)\| \;=\; \left\|r_g^{-1}(g_j) - q(u_j)\right\| \;=\; \|g_j - r_g(q u_j)\| \;\leq\; \frac{\varepsilon}{\sqrt{d_n}}, \qquad \text{for } j \in [\![d_0 + 1, d_n]\!].$$

So $\|\!|q_n^* q' - q|\!\| \leq \varepsilon$ and the net $\mathcal{R} := \{r_g : g \in \mathcal{G}\}$ is appropriate. Finally, by taking $\mathcal{G}$ as in Lemma 5.10, we obtain

$$|\mathcal{R}| \;\leq\; \left(\frac{\pi d_n}{2}\right)^{\frac{d_0}{2}} \cdot \left(\frac{16\sqrt{d_0 d_n}}{\varepsilon}\right)^{d_0(d_n - 1)},$$

hence the result. □

Consequently, we have established that

$$\mathbb{P}(\Theta \in \mathcal{A}_{\varepsilon_n}) \;\geq\; \left(\frac{2}{\pi d_n}\right)^{\frac{d_0}{2}} \cdot \left(\left(\frac{\varepsilon_n}{D_1}\right)^{\frac{1}{\beta}} \frac{1}{16\sqrt{d_0 d_n}}\right)^{d_0(d_n - 1)}.$$

Recall that we have the following lower bound:

$$\Pi_n\left(\left\|W^{A,\Gamma,\Theta} - f_n^*\right\|_\infty \leq 2\varepsilon_n\right) \;\geq\; \pi_\Gamma(d_0) \cdot \mathbb{P}(\Theta \in \mathcal{A}_{\varepsilon_n}) \cdot \exp\left(-\frac{1}{2} n \varepsilon_n^2\right).$$

In order to establish the prior mass condition, it suffices to derive the greatest design dimension d_n for which we can reach

$$\mathbb{P}(\Theta \in \mathcal{A}_{\varepsilon_n}) \;\geq\; \pi_\Gamma(d_0)^{-1} \exp\left(-\frac{1}{2} n \varepsilon_n^2\right).$$

For n large enough, a design dimension d_n as specified in Assumption 3.5 is appropriate for sufficiently small constant C_D.

Remark 5.2. The exponent $d_0(d_n - 1)$ in Lemma 5.1 is probably not far to be optimal. In fact, ignoring the constants, changing this exponent to d_n^α with $\alpha < 1$ would lead to a growth rate of $n^{d_0/(\alpha(2\beta + d_0))}$ which, when β is close to zero, gives a growth rate with an order superior to n. The breakpoint of some popular subspace estimators, such as SIR, being the order n, it would be surprising to estimate a function faster than its central subspace.

5.1.3 Sieve condition (5.3)

The second condition can be verified similarly as in [JT21]. As in the previous section, we will first treat the case with deterministic rescaling parameter, dimension, and direction and then integrate according to A, Γ, and Θ.

We suppose that n is large enough so that $d_n > d_{\max}$. We introduce the quantities $M_n := C_M\sqrt{n\varepsilon_n^2}$ for some large constant C_M and, for $1 \le b \le d_{\max}$, the quantity $r_{n,b}$ such that $r_{n,b}^b(\log n)^{b+1} = C_r n\varepsilon_n^2$, for a large constant C_r. The sieve $\mathbb{B}_n$ is defined as follows:

$$\mathbb{B}_n \;:=\; \bigcup_{q\in\mathcal{O}_{d_n}} \mathcal{B}_{n,q},$$

with

$$\mathcal{B}_{n,q} \;:=\; \bigcup_{b=1}^{d_{\max}} \mathcal{B}_{n,b,q} \qquad \text{and} \qquad \mathcal{B}_{n,b,q} \;:=\; M_n\sqrt{r_{n,b}}\cdot \mathbb{H}_1^{r_{n,b},b,q} + \varepsilon_n B_1,$$

where B_1 is the unit ball in the Banach space $(\mathcal{C}^0(\mathbb{U}_{d_n}), \|\cdot\|_\infty)$.

The nesting property of Lemma 4.7 in [VV09] remains true in the present setting, that is, for $a \le a'$,

$$\sqrt{a}\cdot\mathbb{H}_1^{a,b,q} \;\subseteq\; \sqrt{a'}\cdot\mathbb{H}_1^{a',b,q}.$$

Consequently, if $1 \le a \le r_{n,b}$, then

$$M_n\mathbb{H}_1^{a,b,q} + \varepsilon_n B_1 \;\subseteq\; M_n\sqrt{\frac{r_{n,b}}{a}}\cdot\mathbb{H}_1^{r_{n,b},b,q} + \varepsilon_n B_1 \;\subseteq\; \mathcal{B}_{n,b,q}.$$

By Borell's inequality (see [VV08b], Theorem 5.1, or [Bor75]), for every $a \in [1, r_{n,b}]$,

$$\begin{aligned}
\Pi_n(W^{a,b,q} \notin \mathbb{B}_n) \;&\le\; \Pi_n(W^{a,b,q} \notin \mathcal{B}_{n,b,q}) \\
&\le\; \Pi_n(W^{a,b,q} \notin M_n\mathbb{H}_1^{a,b,q} + \varepsilon_n B_1) \\
&\le\; 1 - \Phi\left(\Phi^{-1}\left(\Pi_n\left(\left\|W^{a,b,q}\right\|_\infty \le \varepsilon_n\right)\right) + M_n\right),
\end{aligned}$$

where Φ is the cumulative distribution function of the standard normal distribution. Now, because

$$\Pi_n\left(\left\|W^{a,b,q}\right\|_\infty \le \varepsilon_n\right) \;\ge\; \Pi_n\left(\left\|W^{r_{n,b},b,q}\right\|_\infty \le \varepsilon_n\right) \;=\; \exp\left(-\phi_0^{r_{n,b},b,q}(\varepsilon_n)\right),$$

we have

$$\Pi_n(W^{a,b,q} \notin \mathbb{B}_n) \;\le\; 1 - \Phi\left(\Phi^{-1}\left(e^{-\phi_0^{r_{n,b},b,q}(\varepsilon_n)}\right) + M_n\right).$$

For n large enough, we have $\varepsilon_n \le \min\{\varepsilon_0^{a_0,b} : b \in [\![1, d_{\max}]\!]\}$ and $r_{n,b} \ge a_0$, so according to Lemma 5.7 and because $b \le d_{\max}$, we have

$$\phi_0^{r_{n,b},b,q}(\varepsilon_n) \;\lesssim\; r_{n,b}^b \log\left(\frac{r_{n,b}}{\varepsilon_n}\right)^{b+1} \;\lesssim\; r_{n,b}^b(\log n)^{b+1} \;\lesssim\; n\varepsilon_n^2,$$

for sufficiently large n. So by taking M_n^2 a very large multiple of $n\varepsilon_n^2$, we can reach $M_n \ge 4\sqrt{\phi_0^{r_{n,b},b,q}(\varepsilon_n)}$. The second assertion of Lemma 4.10 in [VV09] gives $M_n \ge -2\Phi^{-1}\left(\exp\left(-\phi_0^{r_{n,b},b,q}(\varepsilon_n)\right)\right)$ which leads to the upper bound

$$\Pi_n(W^{a,b,q} \notin \mathbb{B}_n) \;\le\; 1 - \Phi(M_n/2) \;\le\; \exp(-M_n^2/8).$$

Taking into account the random rescaling parameter A, we have, for sufficiently large n,

$$\begin{aligned}
\Pi_n(W^{A,b,q} \notin \mathbb{B}_n) &\leq \int_c^{r_{n,b}} \Pi_n(W^{a,b,q} \notin \mathbb{B}_n)\pi_{n,b}(a)da \;+\; \pi_{n,b}(A \geq r_{n,b}) \\
\text{(Assumption 3.4)} \quad &\leq \exp(-M_n^2/8) \;+\; D_2 \int_{r_{n,b}}^{\infty} \exp\left(-C_2 a^b (\log a)^{b+1}\right) da \\
&\leq \exp(-M_n^2/8) \;+\; D_2 \int_{r_{n,b}}^{\infty} C_2 a^{b-1}((b+1)\log^b a + b\log^{b+1} a)\exp\left(-C_2 a^b (\log a)^{b+1}\right) da \\
&\leq \exp(-M_n^2/8) \;+\; D_2 \exp\left(-C_2 r_{n,b}^b (\log r_{n,b})^{b+1}\right) \\
&\leq \frac{1}{2}\exp(-5n\varepsilon_n^2) \;+\; \frac{1}{2}\exp(-5n\varepsilon_n^2) \\
&= \exp(-5n\varepsilon_n^2),
\end{aligned}$$

where the last inequality holds because C_r and C_M are supposed to be large enough.

Now considering the prior on the sparsity pattern, we obtain

$$\Pi_n(W^{A,\Gamma,\Theta} \notin \mathbb{B}_n) \;\leq\; \sum_{b=1}^{d_{\max}} \Pi_n(\Gamma = b) \int_{\mathcal{O}_{d_n}} \Pi_n(W^{A,b,q} \notin \mathbb{B}_n)dq \;\leq\; \exp(-5n\varepsilon_n^2).$$

5.1.4 Entropy condition (5.4)

We use again the notation and quantities of the previous section. According to Lemma 5.6, for all $q \in \mathcal{O}_{d_n}$ and $b \in [\![1, d_{\max}]\!]$, the metric entropy of $\mathcal{B}_{n,b,q}$ is bounded as:

$$\begin{aligned}
\log N\left(2\varepsilon_n, M_n\sqrt{r_{n,b}}\mathbb{H}_1^{r_{n,b},b,q} + \varepsilon_n B_1, \|\cdot\|_\infty\right) &\leq \log N\left(\varepsilon_n, M_n\sqrt{r_{n,b}}\mathbb{H}_1^{r_{n,b},b,q}, \|\cdot\|_\infty\right), \\
&\lesssim r_{n,b}^b \log\left(M_n\sqrt{r_{n,b}}\varepsilon_n^{-1}\right)^{b+1}.
\end{aligned}$$

The simple estimation $\log\left(M_n\sqrt{r_{n,b}}\varepsilon_n^{-1}\right) \asymp \log n$ gives then

$$\log N\left(2\varepsilon_n, \mathcal{B}_{n,b,q}, \|\cdot\|_\infty\right) \;\lesssim\; n\varepsilon_n^2. \tag{5.10}$$

The metric entropy of $\mathcal{B}_{n,q}$ is derived as follows:

$$N\left(2\varepsilon_n, \mathcal{B}_{n,q}, \|\cdot\|_\infty\right) \;\leq\; \sum_{b=1}^{d_{\max}} N\left(2\varepsilon_n, \mathcal{B}_{n,b,q}, \|\cdot\|_\infty\right) \;\leq\; d_{\max} \max_{1\leq b\leq d_{\max}} N\left(2\varepsilon_n, \mathcal{B}_{n,b,q}, \|\cdot\|_\infty\right).$$

To extend these inequalities to the full sieve, we need the following lemma from [Tok11].

Lemma 5.2 (Tokdar 2011, Lemma 1). *Let $a > 0$, $b < d_n$ and $q, \tilde{q} \in \mathcal{O}_{d_n}$. Then*

$$\mathbb{H}_1^{a,b,q} \;\subseteq\; \mathbb{H}_1^{a,b,\tilde{q}} \;+\; a\sqrt{2b}\cdot|||q - \tilde{q}|||B_1,$$

where B_1 is the unit ball in $(\mathcal{C}^0(\mathbb{U}_{d_n}), \|\cdot\|_\infty)$.

By examining the representation result in (3.2) for $\mathbb{H}_{a,b,q}$, we see that, for all $q' \in \mathcal{O}_{d_n}(q^{-1}(E_{\mathbf{b}}))$, we have $\mathbb{H}_{a,b,q} = \mathbb{H}_{a,b,qq'}$. Hence, Lemma 5.2 gives

$$\mathbb{H}_1^{a,b,q} \subseteq \mathbb{H}_1^{a,b,\tilde{q}q'} + a\sqrt{2b} \cdot |||q - \tilde{q}|||B_1.$$

If $\mathcal{R}_n$ is a net over $\mathcal{O}_{d_n}$ such that for all $q \in \mathcal{O}_{d_n}$, there exist $q' \in \mathcal{O}_{d_n}(q^{-1}(E_{\mathbf{d_0}}))$ and $\overline{q} \in \mathcal{R}_n$ with $|||qq' - \overline{q}||| \leq \zeta_n$, where ζ_n is the minimum of $\varepsilon_n/(M_n r_{n,b}^{3/2}\sqrt{2d_n})$ when b runs through $[\![1, d_{\max}]\!]$, then

$$\begin{aligned} M_n\sqrt{r_{n,b}} \cdot \mathbb{H}_1^{r_{n,b},b,q} &\subseteq M_n\sqrt{r_{n,b}} \cdot \mathbb{H}_1^{r_{n,b},b,\overline{q}} + M_n r_{n,b}^{3/2}\sqrt{2b} \cdot |||qq' - \overline{q}|||B_1 \\ &\subseteq M_n\sqrt{r_{n,b}} \cdot \mathbb{H}_1^{r_{n,b},b,\overline{q}} + \varepsilon_n B_1 \\ &= \mathcal{B}_{n,b,\overline{q}}. \end{aligned}$$

This clearly implies

$$\mathcal{B}_{n,q} \subseteq \mathcal{B}_{n,\overline{q}} + \varepsilon_n B_1,$$

and hence

$$\mathbb{B}_n = \bigcup_{q \in \mathcal{O}_{d_n}} \mathcal{B}_{n,q} \subseteq \bigcup_{\overline{q} \in \mathcal{R}_n} (\mathcal{B}_{n,\overline{q}} + \varepsilon_n B_1).$$

Consequently, the $3\varepsilon_n$-entropy of $\mathbb{B}_n$ can be bounded by the cardinal of the net $\mathcal{R}_n$ times the maximal $2\varepsilon_n$-entropy of sets $\mathcal{B}_{n,b,q}$:

$$\begin{aligned} N(3\varepsilon_n, \mathbb{B}_n, \|\cdot\|_\infty) &\leq \sum_{\overline{q} \in \mathcal{R}_n} N(3\varepsilon_n, \mathcal{B}_{n,\overline{q}} + \varepsilon_n B_1, \|\cdot\|_\infty) \\ &\leq \sum_{\overline{q} \in \mathcal{R}_n} N(2\varepsilon_n, \mathcal{B}_{n,\overline{q}}, \|\cdot\|_\infty) \\ &\leq |\mathcal{R}_n| \cdot d_{\max} \max_{\substack{1 \leq b \leq d_{\max} \\ \overline{q} \in \mathcal{R}_n}} N(2\varepsilon_n, \mathcal{B}_{n,b,q}, \|\cdot\|_\infty). \end{aligned}$$

It only remains to bound the cardinal of $\mathcal{R}_n$.

Lemma 5.3. *For $\zeta > 0$, there exists a net $\mathcal{R}$ over $\mathcal{O}_{d_n}$ such that*

$$\bigcup_{\overline{q} \in \mathcal{R}} \mathcal{A}_{\overline{q}} = \mathcal{O}_{d_n},$$

where

$$\mathcal{A}_{\overline{q}} := \{q \in \mathcal{O}_{d_n} \mid \exists q' \in \mathcal{O}_{d_n}(q^{-1}(E_{\mathbf{d_0}})),\ |||qq' - \overline{q}||| \leq \zeta\},$$

and such that

$$|\mathcal{R}| \leq \left(\frac{\pi\sqrt{d_0 d_n}}{2}\right)^{d_0} \left(\frac{16\sqrt{d_0 d_n}}{\zeta}\right)^{d_0(d_n+d_0-2)}.$$

Proof. Firstly, we remark that

$$\mathcal{A}_{\overline{q}} = \left\{q \in \mathcal{O}_{d_n} \mid \exists q'' \in \mathcal{O}_{d_n},\ q''_{|q^{-1}(E_{\mathbf{d_0}})} = q_{|q^{-1}(E_{\mathbf{d_0}})} \text{ and } |||q'' - \overline{q}||| \leq \zeta\right\}.$$

Thus, for $q \in \mathcal{O}_{d_n}$, we search to construct $\overline{q}$ such that there exists $q'' \in \mathcal{O}_{d_n}$ satisfying $q''_{|q^{-1}(E_{\mathbf{d_0}})} = q_{|q^{-1}(E_{\mathbf{d_0}})}$ and $|||q'' - \overline{q}||| \leq \zeta$.

Let $(u_1, \dots, u_{d_0}, u_{d_0+1}, \dots, u_{d_n})$ be an orthonormal basis adapted to the direct sum $\mathbb{R}^{d_n} = (q)^{-1}(E_{\mathbf{d_0}}) \overset{\perp}{\bigoplus} (q)^{-1}(E_{1-\mathbf{d_0}})$. We introduce $\mathcal{F}$ a set of orthonormal basis of $E_{\mathbf{d_0}}$ such that, for all orthonormal basis f' of $E_{\mathbf{d_0}}$, there exists $f \in \mathcal{F}$ such that

$$\sup_{i \in [\![1, d_0]\!]} \|f_i - f'_i\| \ \leq \ \frac{\zeta}{2\sqrt{d_0 d_n}},$$

and we reuse the set $\mathcal{G}$ of Lemma 5.1, replacing ε by ζ. For all $g \in \mathcal{G}$ and $f \in \mathcal{F}$, we fix an isometry $r_{g,f} \in \mathcal{O}_{d_n}$ such that $r_{g,f}(g_i) = f_i$, for all $i \in [\![1, d_0]\!]$.

By construction, there exist $f \in \mathcal{F}$ and $g \in \mathcal{G}$ such that

$$\sup_{i \in [\![1, d_0]\!]} \|f_i - q(u_i)\| \ \leq \ \frac{\zeta}{2\sqrt{d_0 d_n}} \quad \text{and} \quad \sup_{i \in [\![1, d_0]\!]} \|g_i - u_i\| \ \leq \ \frac{\zeta}{2\sqrt{d_0 d_n}}.$$

Then we choose $\overline{q} = r_{g,f}$. Using Lemma 5.8, we extend g to an orthonormal basis over $\mathbb{R}^{d_n}$ such that $\sup_{j \in [\![1, d_n]\!]} \|g_j - u_j\| \ \leq \ \zeta/\sqrt{d_n}$ and we define $f_j := r_{g,f}(g_j) \in E^{\perp}_{\mathbf{d_0}}$, for $j \in [\![d_0 + 1, d_n]\!]$. Now we choose $q'' \in \mathcal{O}_{d_n}$ such that

$$\begin{cases} q''(u_i) = q(u_i) & \text{if } i \in [\![1, d_0]\!], \\ q''(u_j) = f_j & \text{if } j \in [\![d_0 + 1, d_n]\!]. \end{cases}$$

This leads to $\|q''(u_j) - \overline{q}(u_j)\| \leq \zeta/\sqrt{d_n}$, for all $j \in [\![1, d_n]\!]$, hence $|||q'' - \overline{q}||| \leq \zeta$. We can thus define the net $\mathcal{R}$ as the set of all isometries $r_{g,f}$ for $g \in \mathcal{G}$ and $f \in \mathcal{F}$. According to Lemma 5.10, this yields the upper bound

$$\begin{aligned} |\mathcal{R}| \ &= \ |\mathcal{G}| \cdot |\mathcal{F}| \\ &\leq \ \left(\frac{\pi d_n}{2}\right)^{\frac{d_0}{2}} \left(\frac{16\sqrt{d_0 d_n}}{\zeta}\right)^{d_0(d_n - 1)} \left(\frac{\pi d_0}{2}\right)^{\frac{d_0}{2}} \left(\frac{16\sqrt{d_0 d_n}}{\zeta}\right)^{d_0(d_0 - 1)}. \end{aligned}$$ □

Observing that the upper bound in (5.10) does not hide a constant depending on q, we can write

$$\max_{\substack{1 \leq b \leq d_{\max} \\ \overline{q} \in \mathcal{R}_n}} N\left(2\varepsilon_n, \mathcal{B}_{n,b,q}, \|\cdot\|_\infty\right) \ \lesssim \ n\varepsilon_n^2.$$

Then, the lemma yields the following inequality:

$$N\left(3\varepsilon_n, \mathbb{B}_n, \|\cdot\|_\infty\right) \ \lesssim \ \left(\frac{\pi\sqrt{d_0 d_n}}{2}\right)^{d_0} \left(\frac{16 M_n r_n^{3/2} d_n \sqrt{2 d_0}}{\varepsilon_n}\right)^{d_0(d_n + d_0 - 2)} d_{\max} \cdot n\varepsilon_n^2,$$

where $r_n := \max\{r_{n,b} : b \in [\![1, d_{\max}]\!]\}$, which, with the logarithm and for sufficiently large n, gives the desired result.

5.2 Proof of Theorem 4.1

5.2.1 Case $\Gamma < d_0$

The idea of the proof is to show that the non-constancy of p_0 in all directions results in a significant difference (in the Hellinger sense) between the true density p^* and any density that is more parcimonious

than p^*. If this difference can be bounded from below, then the set of over-parcimonious densities is expected to have an almost-null posterior mass as soon as the contraction rate falls below the lower bound.

Let $q \in \mathcal{O}_d$ and let $\tilde{p}$ be a density that satisfies the model with parameters Γ and q. Then, $\tilde{p}$ is constant on $q^{-1}(E_{1-\mathbf{\Gamma}}) + x$, for any $x \in \mathbb{U}_d$. Moreover, the intersection between $\mathcal{S}$ and $q^{-1}(E_{1-\mathbf{\Gamma}})$ is non-null so $\tilde{p}_{|\mathcal{S}}$ is constant in at least one direction, say $\mathbf{\Delta} \in \mathcal{S}$. We will use Assumption 4.1 and integrate the Hellinger distance over a small square inside the region where p^* is non-constant in $\mathbf{\Delta}$. As usual, we denote $\Delta := \mathrm{Span}(\mathbf{\Delta})$.

Let us introduce the operator

$$\begin{aligned} \Psi : \mathbb{R}^{d_0} &\to \mathcal{S} \\ x &\mapsto (q^*)^{-1}(x^{\mathbf{d_0}}). \end{aligned}$$

In particular, we have $p^* \circ \Psi = p_0$. We use the notation of Assumption 4.1 with $\Psi^{-1}(\mathbf{\Delta})$ instead of $\mathbf{\Delta}$.

Let $(\mathbf{\Delta}, u_1, \dots, u_{d_0-1}; v_1, \dots, v_{d-d_0})$ be an orthonormal basis adapted to the direct sum $\mathbb{R}^d = \Delta \oplus (\Delta^\perp \cap \mathcal{S}) \oplus \mathcal{S}^\perp$ and let R be a solid square with edges parallel to this basis, of size $L/\sqrt{d}$ and centered on $\Psi(o)$. Then, $R \subset \mathcal{B}_d(L/2) + \Psi(o)$ and the inequality of Assumption 4.1 is valid when $t \in R$. Considering the basis previously introduced, integrating over R amounts to integrate with respect to each variables. To simplify, we bundle these variables in three groups: a variable δ parallel to Δ, a variable u parallel to $\Delta^\perp \cap \mathcal{S}$ and a variable v parallel to $\mathcal{S}^\perp$. In this coordinate system, we can write $\Psi(o) = (\Psi(o)_1, \Psi(o)_2, 0)$ and we have $p^*(\delta, u, v) = p_0(\Psi^{-1}(\delta, u, 0))$.Then

$$\begin{aligned} h^2(p^*_{|R}\,;\,\tilde{p}_{|R}) &= \iiint_R \left|\sqrt{p^*(\delta,u,0)} - \sqrt{\tilde{p}(0,u,v)}\right|^2 d\delta\, du\, dv \\ &= \iint \left(\int \left|\sqrt{p_0(\Psi^{-1}(\delta,u,0))} - \sqrt{\tilde{p}(0,u,v)}\right|^2 d\delta \right) du\, dv \\ &= \iint h^2\left(p_{0|I_u}\,;\,\tilde{p}(0,u,v)\right) du\, dv, \end{aligned}$$

where I_u is the inverse image via Ψ of the range of the integral in δ. Hence

$$\Psi(I_u) = (\Psi(o)_1, u, 0) + \Big]-\frac{L}{2\sqrt{d}}\mathbf{\Delta}\,;\,\frac{L}{2\sqrt{d}}\mathbf{\Delta}\Big[\quad \text{with } u \in \Psi(o)_2 + \Big]-\frac{L}{2\sqrt{d}}\,;\,\frac{L}{2\sqrt{d}}\Big[^{d_0-1}.$$

Then because $\Psi^{-1}(\Psi(o)_1, u, 0) \in o + \mathcal{B}_{d_0}(L/2)$, there exists $t \in \mathcal{B}_{d_0}(L/2)$ such that

$$I_u = o + t + \Big]-\frac{L}{2\sqrt{d}}\Psi^{-1}(\mathbf{\Delta})\,;\,\frac{L}{2\sqrt{d}}\Psi^{-1}(\mathbf{\Delta})\Big[. \tag{5.11}$$

Now we can use Assumption 4.1 and bound from below the Hellinger distance in the last integral, which gives

$$h^2(p^*_{|R}\,;\,\tilde{p}_{|R}) \geq \iint D \cdot \frac{L^2}{d}\, du\, dv = D \cdot \left(\frac{L}{\sqrt{d}}\right)^{d+1}.$$

Finally, $\Pi_n(\Gamma < d_0 \mid X_1, \dots, X_n) = 0$ as soon as the contraction rate achieves $\varepsilon_n \leq \sqrt{D}\left(\frac{L}{\sqrt{d}}\right)^{\frac{d+1}{2}}$.

5.2.2 Case $\Gamma = d_0$

Case $\mathbf{\Gamma = d_0}$, with $\mathbf{d = 2}$ and $\mathbf{d_0 = 1}$. To simplify the presentation, we first restrict ourselves to the case $d = 2$ and $d_0 = 1$. Assumption 4.1 specializes as follows: for all $0 < l \leq L$, there exists $o \in [-1 + L, 1 - L]$ such that, for all $t \in [-l/2, l/2]$ and all constant $c > 0$,

$$h^2\left(p_{0|]o+t-\frac{l}{2};o+t+\frac{l}{2}[}\,;\, c\right) \;=\; \int_{o+t-\frac{l}{2}}^{o+t+\frac{l}{2}} \left|\sqrt{p_0(\lambda)} - \sqrt{c}\right|^2 d\lambda \;\geq\; D \cdot l^2.$$

We use the fact that the non-constancy of p^* over $\mathcal{S}$ induces a non-constancy over any one-dimensional space not parallel to $\mathcal{S}^\perp$. It is then possible to set a lower bound on the Hellinger distance between p^* and any density that is constant on a space not parallel to $\mathcal{S}^\perp$. For $q \in \mathcal{O}_2$, we denote $E := q^{-1}(E_{\mathbf{d_0}})$ and $F := E^\perp$. If q is not in $\mathcal{Q}^*$, then there exists $0 < \vartheta \leq \pi/2$ such that for all $\overline{q} \in \mathcal{Q}^*$, we have $|||\overline{q} - q||| > \vartheta$. Then, the intersections of F and $\mathcal{S}^\perp$ with the unit circle are separated by at least ϑ.

With this setting, any square of size $L/\sqrt{2}$ centered in $\Psi(o)$ is included in $\mathbb{U}_2$. Let R be a solid square of size $L/\sqrt{2}$, parallel to the line F and centered on $\Psi(o)$. The line $F + \Psi(o)$ intersects the border of R at two points (see Figure 1), and using arguments from geometry on the two-dimensional Euclidean space, we can show that the orthogonal projections of these points over $\mathcal{S}$ are at a distance $\zeta \geq \frac{L\vartheta}{4\sqrt{2}}\sqrt{4 - \vartheta^2}$ from $\Psi(o)$. Similarly, the line $E + \Psi(o)$ intersects the border of R at two points whose orthogonal projections on $\mathcal{S}$ are at a distance $\chi \leq \frac{L}{2\sqrt{2}}\sqrt{1 - \vartheta^2 + \vartheta^4/4}$ from $\Psi(o)$.

Let $(\mathbf{u}, \mathbf{v})$ be an orthogonal basis of $\mathbb{R}^2$ adapted to the decomposition $E \oplus F$ and such that $\mathrm{pr}_{\mathcal{S}}(\mathbf{u}) = \frac{2\sqrt{2}}{L}\chi \cdot \Psi(1)$ and $\mathrm{pr}_{\mathcal{S}}(\mathbf{v}) = \frac{2\sqrt{2}}{L}\zeta \cdot \Psi(1)$. In this system of coordinates, $\Psi(o)$ can be written (o_1, o_2) and for all $u, v \in \mathbb{R}^2$, we have

$$\Psi^{-1}\left(\mathrm{pr}_{\mathcal{S}}(u, v)\right) \;=\; \chi \cdot \frac{2\sqrt{2}}{L}u \;+\; \zeta \cdot \frac{2\sqrt{2}}{L}v.$$

We will also use the fact that $p^*(u, v) = p_0\left(\Psi^{-1}\left(\mathrm{pr}_{\mathcal{S}}(u, v)\right)\right)$. Then, for all density $\tilde{p}$ constant in the direction F, we have

$$\begin{aligned}
h^2(p^*_{|R}\,;\, \tilde{p}_{|R}) \;&=\; \iint_R |\sqrt{p^*(u,v)} - \sqrt{\tilde{p}(u,0)}|^2 du\, dv \\
&=\; \iint_R |\sqrt{p_0\left(\Psi^{-1}\left(\mathrm{pr}_{\mathcal{S}}(u,v)\right)\right)} - \sqrt{\tilde{p}(u,0)}|^2 du\, dv \\
&=\; \int_{o_1 - L/(2\sqrt{2})}^{o_1 + L/(2\sqrt{2})} \int_{o_2 - L/(2\sqrt{2})}^{o_2 + L/(2\sqrt{2})} \left| p_0\Big(\chi \cdot \frac{2\sqrt{2}}{L}u \;+\; \zeta \cdot \frac{2\sqrt{2}}{L}v\Big)^{1/2} - \sqrt{\tilde{p}(u,0)}\right|^2 dv\, du \\
&=\; \int_{-L/(2\sqrt{2})}^{L/(2\sqrt{2})} \int_{-L/(2\sqrt{2})}^{L/(2\sqrt{2})} \left| p_0\Big(o + \chi \cdot \frac{2\sqrt{2}}{L}u \;+\; \zeta \cdot \frac{2\sqrt{2}}{L}v\Big)^{1/2} - \sqrt{\tilde{p}(u,0)}\right|^2 dv\, du \\
&=\; \int_{-L/(2\sqrt{2})}^{L/(2\sqrt{2})} \frac{L}{2\zeta \cdot \sqrt{2}} \left(\int_{-\zeta}^{\zeta} \left| p_0\Big(o + \chi \cdot \frac{2\sqrt{2}}{L}u \;+\; w\Big)^{1/2} - \sqrt{\tilde{p}(u,0)}\right|^2 dw \right) du \\
\text{(Assumption 4.1)} \quad &\geq\; \int_{-L/(2\sqrt{2})}^{L/(2\sqrt{2})} \frac{L}{2\zeta \cdot \sqrt{2}} \cdot D \cdot 4\zeta^2\, du \;=\; DL^2 \cdot \zeta \;\geq\; D \cdot \frac{L^3}{4\sqrt{2}}\vartheta\sqrt{4 - \vartheta^2}.
\end{aligned}$$

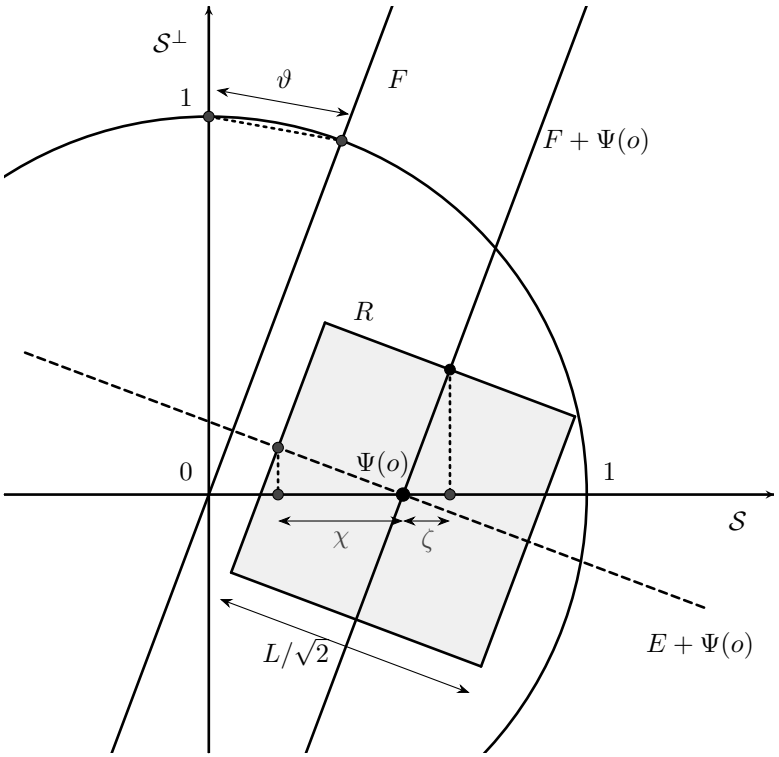

Figure 1: Illustration of the proof of Theorem 4.1 in the case $\Gamma = d_0$ with $d = 2$ and $d_0 = 1$.

Finally, $\Pi_n\left(\Gamma = d_0 \text{ and } \min_{q\in\mathcal{Q}^*} |||\Theta - q||| \geq \vartheta \mid X_1, \ldots, X_n\right) = 0$ as soon as $\varepsilon_n < \sqrt{DL^2 \cdot \zeta}$.

Case $\Gamma = d_0$, with arbitrary $\mathbf{d} > \mathbf{d_0}$. Given a non-optimal isometry q, we need to quantify how far from $\mathcal{S}^\perp$ the inverse image of the subspace $E_{1-\mathbf{d_0}}$ via q is. This result, elementary when $d = 2$, is stated for arbitrary $d > d_0$ in the following lemma. A proof is given in Appendix 5.3.

Lemma 5.4. *Let $q \in \mathcal{O}_d$. If for all $\overline{q} \in \mathcal{Q}^*$, we have $|||\overline{q} - q||| > \vartheta$, $0 < \vartheta \leq \pi/2$, then there exists $r \in E_{1-\mathbf{d_0}}$, $\|r\| = 1$, such that the distance between $q^{-1}(r)$ and $\mathcal{S}^\perp \cap \mathbb{S}_d$ is at least $\vartheta/2d =: \overline{\vartheta}$, where $\mathbb{S}_d := \{x \in \mathbb{R}^d : \|x\| = 1\}$.*

Now we work under the assumptions of Lemma 5.4. Let G be the linear span of $q^{-1}(r)$ and its orthogonal projection $\mathbf{\Lambda}$ on $\mathcal{S}^\perp$ (or any vector of $\mathcal{S}^\perp$ if the orthogonal projection is zero). Then G has a non-zero intersection with $\mathcal{S}$. Let Δ be this one-dimensional intersection.

Let R be a solid hypercube centered on $\Psi(o)$, with size $\overline{L} := L/\sqrt{d}$, and aligned with an orthogonal basis $(\mathbf{\Delta}, u_1, \ldots, u_{d_0-1}, \mathbf{\Lambda}, v_1, \ldots, v_{d-d_0-1})$ adapted to the direct sum $\mathbb{R}^d = \mathcal{S} \oplus \mathcal{S}^\perp$. With the restrictions on o, R is included in $\mathbb{U}_d$.

We will bound from below the quantity $h^2(p^*_{|R}; \tilde{p}_{|R})$ by using the preceding two-dimensional case on slices of R. For $t \in \{0\} \times \prod_{i=1}^{d_0-1}[o - \overline{L}/2 \cdot u_i; o + \overline{L}/2 \cdot u_i] \times \{0\} \times \prod_{j=1}^{d-d_0-1}[o - \overline{L}/2 \cdot v_j; o + \overline{L}/2 \cdot v_j]$,

the plane $G+t$ contains one element parallel to $\mathcal{S}$ and one element parallel to $\mathcal{S}^\perp$, so the situation is analogue to the previous case, replacing ζ by $\overline{\zeta} := \frac{\overline{L}}{2}\overline{\vartheta}\sqrt{4-\overline{\vartheta}^2}$ (Figure 2). With all this in mind, for all density $\tilde{p}$ constant in the direction $q^{-1}(r)$, one has

$$h^2(p^*_{|R}\,;\,\tilde{p}_{|R}) \;=\; \int_t h^2(p^*_{|R\cap(G+t)};\tilde{p}_{|R\cap(G+t)})dt \;\geq\; \int_t 2D\overline{L}^2\overline{\zeta}dt \;=\; 2D\overline{L}^d\overline{\zeta},$$

which is sufficient to conclude.

The case $\Gamma > d_0$ can be proven in a similar way.

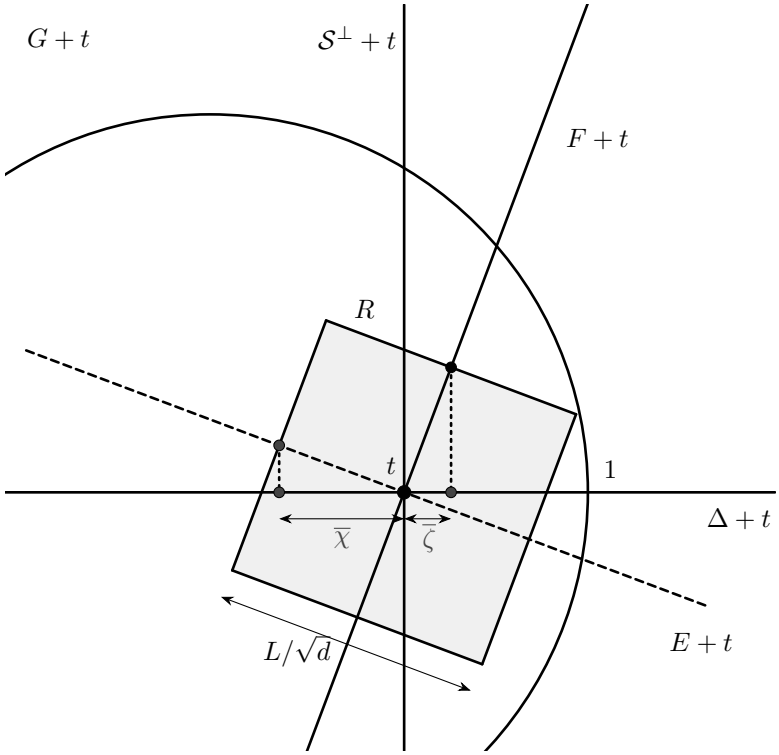

Figure 2: Illustration of the proof of Theorem 4.1 in the case $\Gamma = d_0$ for arbitrary $d > d_0$.

5.3 Lemmas

The next three lemmas are related to Lemmas 4.3, 4.5, and 4.6 in [VV09], hence their proofs can be omitted.

Lemma 5.5. *Let $n \in \mathbb{N}^*$ and $\beta > 0$. If $f_0 \in \mathfrak{C}^\beta(\mathbb{U}_{d_0})$, then, for all $a > 0$ and $q_n \in \mathcal{O}_{d_n}$, there exist constants C_{f_0} and D_{f_0} that depend only on f_0 such that*

$$\inf\left\{\left\|\overline{h}\right\|^2_{\mathbb{H}_{a,d_0,q_n}} \;:\; \overline{h}\in\mathbb{H}_{a,d_0,q_n},\; \left\|\overline{h}-f_{n,q_n}\right\|_\infty \leq C_{f_0}\cdot a^{-\beta}\right\} \;\leq\; D_{f_0}\cdot a^{d_0}.$$

Lemma 5.6. *Let $n \in \mathbb{N}^*$, $a > 0$, $b \leq d_{\max}$ and $q_n \in \mathcal{O}_{d_n}$. Then, there exists a constant L_b that depends only on b such that, for $\varepsilon < 1/2$,*

$$\log N(\varepsilon, \mathbb{H}_1^{a,b,q_n}, \|\cdot\|_\infty) \;\leq\; L_b \cdot a^b \left(\log \frac{1}{\varepsilon}\right)^{b+1}.$$

Lemma 5.7. *Let $n \in \mathbb{N}^*$, $b \leq d_{\max}$ and $q_n \in \mathcal{O}_{d_n}$. Then, for $a_0 > 0$, there exist constants $C_{a_0,b}$ and $\varepsilon_0^{a_0,b}$ that depends only on a_0 and b such that, for all $a \geq a_0$ and $\varepsilon < \varepsilon_0^{a_0,b}$,*

$$-\log \mathbb{P}\left(\left\|W^{a,b,q_n}\right\|_\infty \leq \varepsilon\right) \;\leq\; C_{a_0,b} \cdot a^b \left(\log \frac{a}{\varepsilon}\right)^{b+1}.$$

Lemma 5.8. *Let $n \in \mathbb{N}^*$ and let $(e_1, \dots, e_n)$ be an orthonormal basis of $\mathbb{R}^n$. For $d \leq n$, let $(g_1, \dots, g_d) \in \mathbb{R}^{n \times d}$ be a collection of orthonormal vectors in $\mathbb{R}^n$ such that*

$$\|e_i - g_i\| \leq \varepsilon, \quad \text{for all } i \in [\![1, d]\!].$$

Then we can complete this collection to obtain an orthonormal basis $(g_1, \dots, g_n)$ of $\mathbb{R}^n$ satisfying

$$\|e_j - g_j\| \leq 2\sqrt{d} \cdot \varepsilon, \quad \text{for all } j \in [\![1, n]\!].$$

Proof of Lemma 5.8. We denote by F the subspace $\mathrm{Span}(g_1, \dots, g_d)$. Let us determine the distance between a vector e_j and its orthogonal projection on $F^\perp$, for $j \in [\![d+1, n]\!]$. By Cauchy-Schwartz inequality, we have

$$|\langle e_j, g_i \rangle| \;\leq\; \|e_j\| \, \|g_i - e_i\| \;\leq\; \varepsilon,$$

for all $i \in [\![1, d]\!]$. Then

$$\|e_j - P_{F^\perp}(e_j)\| \;=\; \|P_F(e_j)\| = \left(\sum_{i=1}^{d} \langle e_j, g_i \rangle^2 \, \|g_i\|^2\right)^{1/2} \;\leq\; \sqrt{d} \cdot \varepsilon. \tag{5.12}$$

Thus the problem reduces to find a family of $n - d$ orthonormal vectors in $F^\perp$ with elements as close as possible to the vectors $P_{F^\perp}(e_j)$, for $j \in [\![d+1, n]\!]$. This is related to what is known as *procruste* problem. We denote by A the matrix $A := (P_{F^\perp}(e_{d+1}) | \cdots | P_{F^\perp}(e_n)) \in \mathbb{R}^{n \times n-d}$ and we use Theorem 4.1 stated in [Hig89]:

Theorem 5.9 ([Hig89])**.** *If A admits a polar decomposition $A = UH$, and if $Q \in \mathbb{R}^{n \times n-d}$ has orthonormal columns, then*

$$|||A - U|||_2 \;\leq\; |||A - Q|||_2.$$

Let us show that the columns of U can be chosen in $F^\perp$. A singular value decomposition of A can be written, $A = WD\,{}^tV$, where W has orthonormal columns, $V \in \mathcal{O}_{n-d}$, and $D \in \mathbb{R}^{n-d \times n-d}$ is diagonal. Therefore, $A = (W\,{}^tV)VD\,{}^tV$. Taking $U := W\,{}^tV$ and $H := VD\,{}^tV$, we have the polar decomposition $A = UH$ where U has orthonormal columns. Because $\mathrm{Im}(A) = \mathrm{Span}\,(P_{F^\perp}(e_j), j \in [\![d+1, n]\!]) \subset F^\perp$, it is possible to choose W with columns in $F^\perp$, whence the desired result.

Now, taking $Q = (e_{d+1} | \cdots | e_n)$, we have, for all unit vector $x \in \mathbb{R}^{n-d}$,

$$P_{F^\perp}(Qx) \;=\; Ax.$$

Moreover, using that $|\langle Qx, g_i\rangle| \leq \|Qx\| \, \|g_i - e_i\| \leq \varepsilon$ for all $i \in [\![1, d]\!]$, we finally have

$$\|Qx - Ax\|^2 \;=\; \|Qx - P_{F^\perp}(Qx)\|^2 \;\leq\; d\varepsilon^2,$$

thus $|||A - Q||| \leq \sqrt{d}\cdot\varepsilon$. According to Theorem 5.9, the last inequality is also true if we replace Q by U. Because the columns $u_{d+1}, \ldots, u_n$ of U are in $F^\perp$, the family $(g_1, \ldots, g_d, u_{d+1}, \ldots, u_n)$ is orthonormal and moreover satisfies (5.12) by the triangle inequality. □

Notation. Let $d, n \in \mathbb{N}^*$ with $d < n$ and let $\mathcal{B}^n_{\text{on}}(d)$ be the set of all d-tuples of orthonormal vectors in $\mathbb{R}^n$.

Lemma 5.10. *Let $d, n \in \mathbb{N}^*$ with $d \leq n$ and $0 < \varepsilon \leq 1$. Then there exists a set $\mathcal{G} \subset \mathcal{B}^n_{\text{on}}(d)$ such that for all $e \in \mathcal{B}^n_{\text{on}}(d)$, there exists $g \in \mathcal{G}$ such that*

$$\max_{i \in [\![1,d]\!]} \|e_i - g_i\|_2 \;\leq\; \varepsilon \qquad \text{and} \qquad |\mathcal{G}| \;\leq\; \left(\frac{\pi n}{2}\right)^{d/2} \left(\frac{8}{\varepsilon}\right)^{d(n-1)}.$$

Proof of Lemma 5.10. Let us construct $\mathcal{G}$. Let $\mathcal{T}$ be a set of balls in $\mathbb{R}^n$ with radius $\varepsilon/2$ which cover $\mathbb{S}^{n-1}$ and such that $|\mathcal{T}| = N(\mathbb{S}^{n-1}, \varepsilon/2, \|\cdot\|_2)$. We denote by $\overline{\mathcal{T}}^d$ the set of d-tuples of balls $(B_1, \ldots, B_d) \in \mathcal{T}^d$ such that $B_1 \times \cdots \times B_d$ contains at least one element of $\mathcal{B}^n_{\text{on}}(d)$. Then, for each $e \in \mathcal{B}^n_{\text{on}}(d)$, there exists $(B_1, \ldots, B_d) \in \overline{\mathcal{T}}^d$ such that $e \in B_1 \times \cdots \times B_d$. For each $B \in \overline{\mathcal{T}}^d$, choose one particular d-tuple $g \in \mathcal{B}^n_{\text{on}}(d)$ such that $g \in B$ and let $\mathcal{G}$ be the set of these d-tuples when B runs through $\overline{\mathcal{T}}^d$. It is clear that $\mathcal{G}$ satisfy the first condition of the lemma. Moreover,

$$|\mathcal{G}| \;=\; \left|\overline{\mathcal{T}}^d\right| \;\leq\; \left|\mathcal{T}^d\right| \;=\; N\left(\varepsilon/2, \mathbb{S}^{n-1}, \|\cdot\|_2\right)^d.$$

Let us estimate the last quantity. We use the inequality

$$N\left(\varepsilon, \mathbb{S}^{n-1}, \|\cdot\|_2\right) \;\leq\; D\left(\varepsilon, \mathbb{S}^{n-1}, \|\cdot\|_2\right),$$

where $D\left(\varepsilon, \mathbb{S}^{n-1}, \|\cdot\|_2\right)$ is the maximum number of disjoint balls with radius $\varepsilon/2$ and with center in $\mathbb{S}^{n-1}$. Recall that

$$\mathcal{A}\left(\mathbb{S}^{n-1}\right) \;=\; \frac{2\pi^{n/2}}{\Gamma(n/2)} \qquad \text{and} \qquad \mathcal{V}\left(B_{n-1}(\varepsilon)\right) \;=\; \frac{\pi^{\frac{n-1}{2}}\varepsilon^{n-1}}{\Gamma\left(\frac{n+1}{2}\right)}.$$

Consider the measure $\nu(\varepsilon/2)$ of the hyperspherical cap defined by the intersection of $\mathbb{S}^{n-1}$ and a ball with center in $\mathbb{S}^{n-1}$ and with radius $\varepsilon/2$. The colatitude angle of the cap is $\phi = 2\arcsin(\varepsilon/4)$ and, according to [Li11],

$$\nu(\varepsilon/2) \;=\; \frac{(n-1)\pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n+1}{2}\right)} \int_0^\phi \sin^{n-2}(\theta)d\theta.$$

Since $\phi \geq \varepsilon/2$,

$$\int_0^\phi \sin^{n-2}(\theta)d\theta \;\geq\; \int_0^\phi \left(\frac{\sin\phi}{\phi}\cdot\theta\right)^{n-2} d\theta \;=\; \left(\frac{\sin\phi}{\phi}\right)^{n-2} \frac{\phi^{n-1}}{n-1} \;\geq\; \left(\frac{\sin\phi}{\phi}\right)^{n-2} \frac{1}{n-1}\left(\frac{\varepsilon}{2}\right)^{n-1}$$

and, using the facts that $\varepsilon \leq 1$, $\phi \leq \varepsilon$, and $(\sin\phi)/\phi \geq 1/2$, we have

$$D\left(\varepsilon, \mathbb{S}^{n-1}, \|\cdot\|_2\right) \;\leq\; \frac{\mathcal{A}(\mathbb{S}^{n-1})}{\nu(\varepsilon/2)} \;<\; \frac{\mathcal{A}(\mathbb{S}^{n-1})}{\mathcal{V}\left(B_{n-1}(\varepsilon/2)\right)\cdot\left(\frac{1}{2}\right)^{n-2}} \;=\; \sqrt{\pi}\cdot\left(\frac{4}{\varepsilon}\right)^{n-1}\cdot\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma(n/2)}.$$

The ratio of two Gamma functions can be bounded as follows

$$\sqrt{x+1/4} \;<\; \frac{\Gamma(x+1)}{\Gamma(x+1/2)} \;<\; \sqrt{x+1/2},$$

for $x > -1/2$ (see [Wat59] and [LQ12], Section 2.3). Choosing $x = (n-1)/2$, we obtain

$$N\left(\varepsilon, \mathbb{S}^{n-1}, \|\cdot\|_2\right) \;<\; \sqrt{\frac{\pi n}{2}}\cdot\left(\frac{4}{\varepsilon}\right)^{n-1},$$

hence the result. □

Proof of Lemma 5.4. Suppose that, for all $r \in E_{\mathbf{d_0}}$, we have $d(q^{-1}(r), \mathcal{S}\cap\mathbb{S}_d) < \overline{\vartheta}$ and, for all $r' \in E_{1-\mathbf{d_0}}$, $d(q^{-1}(r'), \mathcal{S}^\perp\cap\mathbb{S}_d) < \overline{\vartheta}$. Let us show that for all vectors e_i of the canonical basis, $\left\|q^{-1}(e_i) - \overline{q}^{-1}(e_i)\right\| < 2\sqrt{d}\cdot\overline{\vartheta}$.

We begin with the first d_0 vectors $(e_1, \ldots, e_{d_0})$. Define $\mathrm{p}_\mathcal{S}$ an operator which maps $r \in E_{\mathbf{d_0}}$ to $\arg\min_{u\in\mathcal{S}\cap\mathbb{S}_d}\left\|q^{-1}(r) - u\right\|$. Then, for $i = 1, \ldots, d_0$, we have $\left\|q^{-1}(e_i) - \mathrm{p}_\mathcal{S}(q^{-1}(e_i))\right\| < \overline{\vartheta}$. Now, we reuse the arguments of the proof of Lemma 5.8, with $A := \left(\mathrm{p}_\mathcal{S}(q^{-1}(e_1))|\cdots|\,\mathrm{p}_\mathcal{S}(q^{-1}(e_{d_0}))\right)$. We can write $A = UH$ where $U = (u_1|\cdots|u_{d_0})$ is a rectangular matrix with orthonormal columns in $\mathcal{S}$ and where H is symmetric. Moreover, taking $Q := (q^{-1}(e_1)|\cdots|q^{-1}(e_{d_0}))$, and $x \in E_{\mathbf{d_0}}\cap\mathbb{S}_d$, $x = \sum_{i=1}^{d_0} a_i e_i$, we have

$$\|Qx - Ax\| \;=\; \left\|\sum_{i=1}^{d_0} a_i\cdot\left(q^{-1}(e_i) - \mathrm{p}_\mathcal{S}(q^{-1}(e_i))\right)\right\| \;<\; \sqrt{d}\cdot\overline{\vartheta}.$$

So, by Theorem 4.1 in [Hig89] (Theorem 5.9 in the present document), $|||A - U||| < \sqrt{d}\cdot\overline{\vartheta}$. Then, $(u_1, \ldots, u_{d_0})$ is an orthonormal basis of $\mathcal{S}$ such that $\left\|\mathrm{p}_\mathcal{S}(q^{-1}(e_i)) - u_i\right\| < \sqrt{d}\cdot\overline{\vartheta}$, for $i = 1, \ldots, d_0$. Let $\overline{q} \in Q^*$ be an isometry such that $\overline{q}(e_i) = u_i$, $i = 1, \ldots, d_0$. Then

$$\left\|q^{-1}(e_i) - \overline{q}^{-1}(e_i)\right\| \;\leq\; \left\|q^{-1}(e_i) - \mathrm{p}_\mathcal{S}(q^{-1}(e_i))\right\| + \left\|\mathrm{p}_\mathcal{S}(q^{-1}(e_i)) - \overline{q}(e_i)\right\| \;<\; 2\sqrt{d}\cdot\overline{\vartheta}, \qquad i = 1, \ldots, d_0.$$

The same reasoning occurs with the remaining vectors, $(e_{d_0+1}, \ldots, e_d)$, by replacing $\mathcal{S}$ by $\mathcal{S}^\perp$, and taking $A' = U'H'$, with $U' = (u_{d_0+1}|\cdots|u_d)$. The isometry $\overline{q} \in Q^*$ is now the one that maps e_i to u_i for $i = 1, \ldots, d$. As a result, for all $x \in \mathbb{S}_d$, $x = \sum_{i=1}^{d} a_i e_i$, we have

$$\left\|q^{-1}(x) - \overline{q}^{-1}(x)\right\| \;=\; \left\|\sum_{i=1}^{d} a_i\cdot\left(q^{-1}(e_i) - \overline{q}^{-1}(e_i)\right)\right\| \;<\; 2d\cdot\overline{\vartheta} \;=\; \vartheta,$$

which contradicts the fact that $|||\overline{q} - q||| > \vartheta$. Finally, $d(q, Q^*) < \vartheta$. □

Acknowledgement

We acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-21-CE40-0007 (GAP Project).

References

[Bir86] Lucien Birgé. "On estimating a density using Hellinger distance and some other strange facts". In: *Probability Theory and Related Fields* **71** (1986), pp. 271–291.

[Bor75] Christer Borell. "The Brunn–Minkowski inequality in Gauss space". In: *Inventiones mathematicae* **30** (1975), pp. 207–216.

[CD12] Laëtitia Comminges and Arnak S. Dalalyan. "Tight conditions for consistency of variable selection in the context of high dimensionality". In: *The Annals of Statistics* **40(5)** (2012), pp. 2667–2696.

[Coo98] R. Dennis Cook. *Regression graphics: Ideas for studying regressions through graphics*. John Wiley & Sons, 1998.

[FS23] Gianluca Finocchio and Johannes Schmidt-Hieber. "Posterior contraction for deep Gaussian process priors". In: *Journal of Machine Learning Research* **24(66)** (2023), pp. 1–49.

[GGV00] Subhashis Ghosal, Jayanta K. Ghosh, and Aad W. Van Der Vaart. "Convergence rates of posterior distributions". In: *The Annals of Statistics* **28(2)** (2000), pp. 500–531.

[GN11] Evarist Giné and Richard Nickl. "Rates of contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$". In: *The Annals of Statistics* **39(6)** (2011), pp. 2883–2911.

[Hig89] Nicholas J. Higham. "Matrix nearness problems and applications". In: *Applications of Matrix Theory*. Ed. by M. J. C. Gover and S. Barnett. Oxford University Press, 1989, pp. 1–27.

[JT21] Sheng Jiang and Surya T. Tokdar. "Variable selection consistency of Gaussian process regression". In: *The Annals of Statistics* **49(5)** (2021), pp. 2491–2505.

[Li11] Shengqiao Li. "Concise formulas for the area and volume of a hyperspherical cap". In: *Asian Journal of Mathematics and Statistics* **4(1)** (2011), pp. 66–70.

[Li91] Ker-Chau Li. "Sliced inverse regression for dimension reduction". In: *Journal of the American Statistical Association* **86(414)** (1991), pp. 316–327.

[Lin+21] Qian Lin et al. "On the optimality of sliced inverse regression in high dimensions". In: *The Annals of Statistics* **49(1)** (2021), pp. 1–20.

[LQ12] Qiu-Ming Luo and Feng Qi. "Bounds for the ratio of two gamma functions—From Wendel's and related inequalities to logarithmically completely monotonic functions". In: *Banach Journal of Mathematical Analysis* **6(2)** (2012), pp. 132–158.

[LZL18] Qian Lin, Zhigen Zhao, and Jun S. Liu. "On consistency and sparsity for sliced inverse regression in high dimensions". In: *The Annals of Statistics* **46(2)** (2018), pp. 580–610.

[LZL19] Qian Lin, Zhigen Zhao, and Jun S. Liu. "Sparse sliced inverse regression via lasso". In: *Journal of the American Statistical Association* **114(528)** (2019), pp. 1726–1739.

[STG13] Weining Shen, Surya T. Tokdar, and Subhashis Ghosal. "Adaptive Bayesian multivariate density estimation with Dirichlet mixtures". In: *Biometrika* **100(3)** (2013), pp. 623–640.

[Sto82] Charles J. Stone. "Optimal global rates of convergence for nonparametric regression". In: *The Annals of Statistics* **10(4)** (1982), pp. 1040–1053.

[Tok11] Surya T. Tokdar. "Dimension adaptability of Gaussian process models with variable selection and projection". Preprint . Available at arXiv:1112.0716. 2011.

[TSY20] Kai Tan, Lei Shi, and Zhou Yu. "Sparse SIR: Optimal rates and adaptive estimation". In: *The Annals of Statistics* **48(1)** (2020), pp. 64–85.

[TZG10] Surya T. Tokdar, Yu M. Zhu, and Jayanta K. Ghosh. "Bayesian density regression with logistic Gaussian process and subspace projection". In: *Bayesian Analysis* **5(2)** (2010), p. 319.

[Ver12] Nicolas Verzelen. "Minimax risks for sparse regressions: Ultra-high dimensional phenomenons". In: *Electronic Journal of Statistics* **6** (2012), pp. 38–90.

[VV08a] Aad W. Van Der Vaart and J. Harry Van Zanten. "Rates of contraction of posterior distributions based on Gaussian process priors". In: *The Annals of Statistics* **36(3)** (2008), pp. 1435–1463.

[VV08b] Aad W. Van Der Vaart and J. Harry Van Zanten. "Reproducing kernel Hilbert spaces of Gaussian priors". In: *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh. Inst. Math. Stat. (IMS) Collect.* **3** (2008), pp. 200–222.

[VV09] Aad W. Van Der Vaart and J. Harry Van Zanten. "Adaptative Bayesian estimation using a Gaussian random field with inverse gamma bandwidth". In: *The Annals of Statistics* **37(5B)** (2009), pp. 2655–2675.

[Wai09] Martin J. Wainwright. "Sharp thresholds for high-Dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)". In: *IEEE transactions on information theory* **55(5)** (2009), pp. 2183–2202.

[Wat59] G. N. Watson. "A note on gamma functions". In: *Edinburgh Mathematical Notes* **42** (1959), pp. 7–9.

[YD16] Yun Yang and David B. Dunson. "Bayesian manifold regression". In: *The Annals of Statistics* **44(2)** (2016), pp. 876–905.

[YT15] Yun Yang and Surya T. Tokdar. "Minimax-optimal nonparametric regression in high dimensions". In: *The Annals of Statistics* **43(2)** (2015), pp. 652–674.

[ZMP06] Lixing Zhu, Baiqi Miao, and Heng Peng. "On sliced inverse regression with high-dimensional covariates". In: *Journal of the American Statistical Association* **101(474)** (2006), pp. 630–643.

[ZMZ22] Jing Zeng, Qing Mai, and Xin Zhang. "Subspace estimation with automatic dimension and variable selection in sufficient dimension reduction". In: *Journal of the American Statistical Association* (2022), pp. 1–13.