Journal of Machine Learning Research 13 (2012) 2735-2772 Submitted 8/11; Revised 1/12; Published 9/12

Linear Regression With Random Projections

Odalric-Ambrym Maillard ODALRIC.MAILLARD@INRIA.FR
Rémi Munos REMI.MUNOS@INRIA.FR
INRIA Lille Nord Europe
SequeL Project
40 avenue Halley
59650 Villeneuve d'Ascq, France

Editor: Sanjoy Dasgupta

Abstract

We investigate a method for regression that makes use of a randomly generated subspace $\mathcal{G}_P \subset \mathcal{F}$ (of finite dimension P) of a given large (possibly infinite) dimensional function space $\mathcal{F}$, for example, $L_2([0,1]^d;\mathbb{R})$. $\mathcal{G}_P$ is defined as the span of P random features that are linear combinations of a basis functions of $\mathcal{F}$ weighted by random Gaussian i.i.d. coefficients. We show practical motivation for the use of this approach, detail the link that this random projections method share with RKHS and Gaussian objects theory and prove, both in deterministic and random design, approximation error bounds when searching for the best regression function in $\mathcal{G}_P$ rather than in $\mathcal{F}$, and derive excess risk bounds for a specific regression algorithm (least squares regression in $\mathcal{G}_P$). This paper stresses the motivation to study such methods, thus the analysis developed is kept simple for explanations purpose and leaves room for future developments.

Keywords: regression, random matrices, dimension reduction

1. Introduction

We consider a standard regression problem. Thus let us introduce $\mathcal{X}$ an input space, and $\mathcal{Y} = \mathbb{R}$ the real line. We denote by $\mathcal{P}$ an unknown probability distribution over the product space $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$ and by $\mathcal{P}_\mathcal{X}$ its first marginal, that is, $d\mathcal{P}_\mathcal{X}(x) = \int_\mathbb{R} \mathcal{P}(x,dy)$. In order for this quantity to be well defined we assume that $\mathcal{X}$ is a Polish space (i.e., metric, complete, separable), see Dudley (1989, Th. 10.2.2). Finally, let $L_{2,\mathcal{P}_\mathcal{X}}(\mathcal{X};\mathbb{R})$ be the space of real-valued functions on $\mathcal{X}$ that are squared integrable with respect to (w.r.t.) $\mathcal{P}_\mathcal{X}$, equipped with the quadratic norm

$$\|f\|_{\mathcal{P}_\mathcal{X}} \stackrel{\text{def}}{=} \sqrt{\int_\mathcal{X} f^2(x) dP_\mathcal{X}(x)}\,.$$

In this paper, we consider that $\mathcal{P}$ has some structure corresponding to a model of regression with random design; there exists a (unknown) function $f^\star : \mathcal{X} \to \mathbb{R}$ such that if $(x_n, y_n)_{n \leqslant N} \in \mathcal{X} \times \mathbb{R}$ are independently and identically distributed (i.i.d.) according to $\mathcal{P}$, then one can write

$$y_n = f^\star(x_n) + \eta_n\,,$$

where η_n is a centered noise, independent from $\mathcal{P}_\mathcal{X}$, introduced for notational convenience. In terms of random variables, we will often simply write $Y = f^\star(X) + \eta$ where $(X,Y) \sim \mathcal{P}$.

Let $\mathcal{F} \subset L_{2,\mathcal{P}_X}(\mathcal{X};\mathbb{R})$ be some given class of functions. The goal of the statistician is to build, from the observations only, a regression function $\widehat{f} \in \mathcal{F}$ that is closed to the so-called target function $f^\star$, in the sense that it has a low excess risk $R(f) - R(f^\star)$, where the risk of any $f \in L_{2,\mathcal{P}_X}(\mathcal{X};\mathbb{R})$ is defined as

$$R(f) \stackrel{\text{def}}{=} \int_{\mathcal{X}\times\mathbb{R}} (y - f(x))^2 d\mathcal{P}(x,y)\,.$$

Similarly, we introduce the empirical risk of a function f to be

$$R_N(f) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N} [y_n - f(x_n)]^2\,,$$

and we define the empirical norm of f as $\|f\|_N \stackrel{\text{def}}{=} \sqrt{\frac{1}{N}\sum_{n=1}^{N} f(x_n)^2}$.

Function spaces and penalization. In this paper, we consider that $\mathcal{F}$ is an infinite dimensional space that is generated by the span over a denumerable family of functions $\{\varphi_i\}_{i\geqslant 1}$ of $L_{2,\mathcal{P}_X}(\mathcal{X};\mathbb{R})$: We call the $\{\varphi_i\}_{i\geqslant 1}$ the *initial features* and thus refer to $\mathcal{F}$ as to the initial feature space:

$$\mathcal{F} \stackrel{\text{def}}{=} \Big\{ f_\alpha(x) \stackrel{\text{def}}{=} \sum_{i\geqslant 1} \alpha_i \varphi_i(x), \|\alpha\| < \infty \Big\}.$$

Examples of initial features include Fourier basis, multi-resolution basis such as wavelets, and also less explicit features coming from a preliminary dictionary learning process.

In the sequel, for the sake of simplicity we focus our attention to the case when the target function $f^\star = f_{\alpha^\star}$ belongs to the space $\mathcal{F}$, in which case the excess risk of a function f can be written as $R(f) - R(f^\star) = \|f - f^\star\|_{\mathcal{P}_X}$. Since $\mathcal{F}$ is an infinite dimensional space, empirical risk minimization in $\mathcal{F}$ defined by $\operatorname{argmin}_{f\in\mathcal{F}} R_N(f)$ is certainly subject to overfitting. Traditional methods to circumvent this problem consider penalization techniques, that is, one searches for a function that satisfies

$$\widehat{f} = \arg\min_{f\in\mathcal{F}} R_N(f) + \text{pen}(f),$$

where typical examples of penalization include $\text{pen}(f) = \lambda \|f\|_p^p$ for $p = 1$ or 2, where λ is a parameter and usual choices for the norm are ℓ_2 (ridge-regression: Tikhonov 1963) and ℓ_1 (LASSO: Tibshirani 1994).

Motivation. In this paper we follow a complementary approach introduced in Maillard and Munos (2009) for finite dimensional space, called Compressed Least Squares Regression, and extended in Maillard and Munos (2010), which considers generating *randomly* a space $\mathcal{G}_P \in \mathcal{F}$ of finite dimension P and then returning an empirical estimate in $\mathcal{G}_P$. The empirical risk minimizer in $\mathcal{G}_P$, that is, $\arg\min_{g\in\mathcal{G}_P} R_N(g)$ is a natural candidate, but other choices of estimates are possible, based on traditional literature on regression when $P < N$ (penalization, projection, PAC-Bayesian estimates...). The generation of the space $\mathcal{G}_P$ makes use of random matrices, that have already demonstrated their benefit in different settings (see for instance Zhao and Zhang 2009 about spectral clustering or Dasgupta and Freund 2008 about manifold learning).

Our goal is first to give some intuition about this method by providing approximation error and simple excess risk bounds (which may not be the tightest possible ones as explained in Section 4.2)

for the proposed method, and also by providing links to other standards approaches, in order to encourage research in that direction, which, as showed in the next section, has already been used in several applications.

Outline of the paper. In Section 2, we quickly present the method and give practical motivation for investigating this approach. In Section 3, we give a short overview of Gaussian objects theory (Section 3.1), which enables us to show how to relate the choice of the initial features $\{\varphi_i\}_{i\geqslant 1}$ to the construction of standard function spaces via Gaussian objects (Section 3.2), and we finally state a useful version of the Johnson-Lindenstrauss Lemma for our setting (Section 3.3).

In Section 4, we describe a typical algorithm (Section 4.1), and then provide some quick survey of classical results in regression while discussing the validity of their assumptions in our setting (Section 4.2). Then our main results are stated in Section 4.3, where we provide bounds on approximation error of the random space $\mathcal{G}_P$ in the framework of regression with deterministic and random designs, and in Section 4.4, where we derive excess risk bounds for a specific estimate.

Section 5 provides some discussion about existing results and finally appendix A contains the proofs of our results.

2. Summary Of The Random Projection Method

From now on, we assume that the set of features $\{\varphi_i\}_{i\geqslant 1}$ are continuous and satisfy the assumption that,

$$\sup_{x\in\mathcal{X}} \|\varphi(x)\|^2 < \infty, \text{ where } \varphi(x) \stackrel{\text{def}}{=} (\varphi_i(x))_{i\geqslant 1} \in \ell_2 \text{ and } \|\varphi(x)\|^2 \stackrel{\text{def}}{=} \sum_{i\geqslant 1} \varphi_i(x)^2.$$

Let us introduce a set of P *random features* $(\psi_p)_{1\leqslant p\leqslant P}$ defined as linear combinations of the initial features $\{\varphi_i\}_{1\geqslant 1}$ weighted by random coefficients:

$$\psi_p(x) \stackrel{\text{def}}{=} \sum_{i\geqslant 1} A_{p,i}\varphi_i(x), \text{ for } 1 \leqslant p \leqslant P, \tag{1}$$

where the (infinitely many) coefficients $A_{p,i}$ are drawn i.i.d. from a centered distribution (e.g., Gaussian) with variance $1/P$. Then let us define $\mathcal{G}_P$ to be the (random) vector space spanned by those features, that is,

$$\mathcal{G}_P \stackrel{\text{def}}{=} \Big\{ g_\beta(x) \stackrel{\text{def}}{=} \sum_{p=1}^{P} \beta_p \psi_p(x), \beta \in \mathbb{R}^P \Big\}.$$

In the sequel, $\mathcal{P}_\mathcal{G}$ will refer to the law of the Gaussian variables, $\mathcal{P}_\eta$ to the law of the observation noise and $\mathcal{P}_\mathcal{Y}$ to the law of the observations. Remember also that $\mathcal{P}_\mathcal{X}$ refers to the law of the inputs.

One may naturally wish to build an estimate $g_{\widehat{\beta}}$ in the linear space $\mathcal{G}_P$. For instance in the case of deterministic design, if we consider the ordinary least squares estimate, that is, $\widehat{\beta} = \arg\min_{\beta\in\mathbb{R}^P} R_N(g_\beta)$, then we can derive the following result (see Section 4.4 for a similar result with random design):

Theorem 1 (Deterministic design) *Assuming that the random variable Y is such that $|Y| \leqslant B$, then for all $P \geqslant 1$, for all $\delta \in (0,1)$ there exists an event of $\mathcal{P}_\mathcal{Y} \times \mathcal{P}_\mathcal{G}$-probability larger than $1-\delta$ such*

that on this event, the excess risk of the least squares estimate $g_{\widehat{\beta}}$ is bounded as

$$\|g_{\widehat{\beta}} - f^\star\|_N^2 \leqslant \frac{12\log(8N/\delta)}{P}\|\alpha^\star\|^2 \frac{1}{N}\sum_{n=1}^N \|\varphi(x_n)\|^2 + \kappa B^2 \frac{P+\log(2/\delta)}{N}\,, \tag{2}$$

for some numerical constant $\kappa > 0$.

Example: Let us consider as an example the features $\{\varphi_i\}_{i\geqslant 1}$ to be a set of functions defined by rescaling and translation of a mother one-dimensional hat function (illustrated in Figure 1, middle column) and defined precisely in paragraph 3.2.2. Then in this case we can show that

$$\|\alpha^\star\|^2 \frac{1}{N}\sum_{n=1}^N \|\varphi(x_n)\|^2 \leqslant \frac{1}{2}\|f^\star\|_{H^1}^2\,,$$

where $H^1 = H^1([0,1])$ is the Sobolev space of order 1. Thus we deduce that the excess risk is bounded as $\|g_{\widehat{\beta}} - f^\star\|_N^2 = O(\frac{B\|f^\star\|_{H^1}\log(N/\delta)}{\sqrt{N}})$ for P of the order $\sqrt{N}$.

Similarly, the analysis given in paragraph 3.2.1 below shows that when the features $\{\varphi_i\}_{i\geqslant 1}$ are wavelets rescaled by a factor $\sigma_i = \sigma_{j,l} = 2^{-js}$ for some real number $s > 1/2$, where j,l are the scale and position index corresponding to the ith element of the family, and that the mother wavelet enables to generate the Besov space $\mathcal{B}_{s,2,2}([0,1])$ (see paragraph 3.2.1), then for some constant c, it holds that

$$\|\alpha^\star\|^2 \frac{1}{N}\sum_{n=1}^N \|\varphi(x_n)\|^2 \leqslant \frac{c}{1-2^{-2s+1}}\|f^\star\|_{s,2,2}^2\,.$$

Thus the excess risk in this case is bounded as $\|g_{\widehat{\beta}} - f^\star\|_N^2 = O(\frac{B\|f^\star\|_{s,2,2}\log(N/\delta)}{\sqrt{N}})$.

2.1 Comments

The second term in the bound (2) is a usual estimation error term in regression, while the first term comes from the additional approximation error of the space $\mathcal{G}_P$ w.r.t. $\mathcal{F}$. It involves the norm of the parameter $\alpha^\star$, and also the norm $\|\varphi(x)\|$ at the sample points.

The nice aspects of this result:

- The weak dependency of this bound with the dimension of the initial space $\mathcal{F}$. This appears implicitly in the terms $\|\alpha^\star\|^2$ and $\frac{1}{N}\sum_{n=1}^N \|\varphi(x_n)\|^2$, and we will show that for a large class of function spaces, these terms can be bounded by a function of the norm of $f^\star$ only.

- The result does not require any specific smoothness assumptions on the initial features $\{\varphi_i\}_{i\geqslant 1}$; by optimizing over P, we get a rate of order $N^{-1/2}$ that corresponds to the *minimax* rates under such assumptions up to logarithmic factors.

- Because the choice of the subspace $\mathcal{G}_P$ within which we perform the least-squares estimate is random, we avoid (with high probability) degenerated situations where the target function $f^\star$ cannot be well approximated with functions in $\mathcal{G}_P$. Indeed, in methods that consider a given deterministic finite-dimensional subspace $\mathcal{G}$ of the big space $\mathcal{F}$ (such as linear approximation using a predefined set of wavelets), it is often possible to find a target function $f^\star$ such that

$\inf_{g\in\mathcal{G}_P} \|f^\star - g\|_N$ is large. On the other hand when we use the random projection method, the random choice of $\mathcal{G}_P$ implies that for any $f^\star \in \mathcal{F}$, the approximation error $\inf_{g\in\mathcal{G}_P} \|f^\star - g\|_N$ can be controlled (by the first term of the bound (2)) in high probability. See section 5.2 for an illustration of this property. Thus the results we obtain is able to compete with non-linear approximation (Barron et al., 2008) or kernel ridge regression (Caponnetto and De Vito, 2007).

- In terms of numerical complexity, this approach is more efficient than non-linear regression and kernel ridge regression. Indeed, once the random space has been generated, we simply solve a least squares estimate in a low-dimensional space. The computation of the Gram matrix involves performing random projections (which can be computed efficiently for several choices of the random coefficients $A_{p,i}$, see Liberty et al. 2008; Ailon and Chazelle 2006; Sarlos 2006 and many other references therein). Numerical aspects of the algorithms are described in Section 5.4.

Possible improvements. As mentioned previously we do not make specific assumptions about the initial features $\{\varphi_i\}_{i\geqslant 1}$. However, considering smoothness assumptions on the features would enable to derive a better approximation error term (first term of the bound (2)); typically with a Sobolev assumption or order s, we would get a term of order P^{-2s} instead of P^{-1}. For simplicity of the presentation, we do not consider such assumptions here and report the general results only.

The $\log(N)$ factor may be seen as unwanted and one would like to remove it. However, this term comes from a variant of the Johnson-Lindenstrauss lemma combined with a union bound, and it seems difficult to remove it, unless the dimension of $\mathcal{F}$ is small (e.g., we can then use covers) but this case is not interesting for our purpose.

Possible extensions of the random projection method. It seems natural to consider other constructions than the use of i.i.d. Gaussian random coefficients. For instance we may consider Gaussian variables with variance σ_i^2/P different for each i instead of homeoscedastic variables, which is actually equivalent to considering the features $\{\varphi'_i\}_{i\geqslant 1}$ with $\varphi'_i = \sigma_i\varphi_i$ instead.

Although in the paper we develop results using Gaussian random variables, such method will essentially work similarly for matrices with sub-Gaussian entries as well.

A more important modification of the method would be to consider, like for data-driven penalization techniques, a data-dependent construction of the random space $\mathcal{G}_P$, that is, using a data-driven distribution for the random variable $A_{p,i}$ instead of a Gaussian distribution. However the analysis developed in this paper *will not* work for such modification, due to the fact we longer have independent variables, and thus a different analysis is required.

Illustration. In order to illustrate the method, we show in figure 1 three examples of initial features $\{\varphi_i\}$ (top row) and random features $\{\psi_p\}$ (bottom row). The first family of features is the basis of wavelet Haar functions. The second one consists of multi-resolution hat functions (see paragraph 3.2.2) and the last one shows multi-resolution Gaussian functions. For example, in the case of multi-resolution hat functions (middle column), the corresponding random features are Brownian motions. The linear regression with random projections approach described here simply consists in performing least-squares regression using the set of randomly generated features $\{\psi_p\}_{1\leqslant p\leqslant P}$ (e.g., Brownian motions).

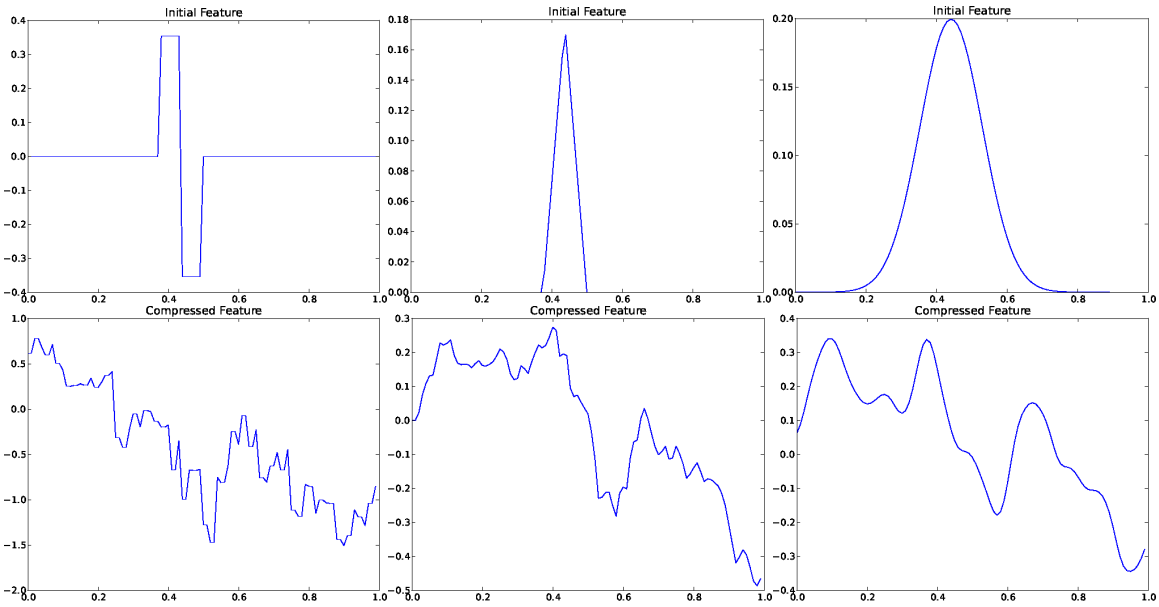

Figure 1: Three representative of initial features φ (top row) and a sample of a corresponding random feature ψ (bottom row). The initial set of features are (respectively) Haar functions (left), multi-resolution hat functions (middle) and multi-resolution Gaussian functions (right).

2.2 Motivation From Practice

We conclude this introduction with some additional motivation to study such objects coming from practical applications. Let us remind that the use of random projections is well-known in many domains and applications, with different names according to the corresponding field, and that the corresponding random objects are widely studied and used. Our contribution is to provide an analysis of this method in a regression setting.

For instance, in Sutton and Whitehead (1993) the authors mentioned such constructions under the name *random representation* as a tool for performing value function approximation in practical implementations of reinforcement learning algorithms, and provided experiments demonstrating the benefit of such methods. They also pointed out that such representations were already used in 1962 in Rosenblatt's perceptron as a preprocessing layer. See also Sutton (1996) for other comments concerning the practical benefit of "random collapsing" methods.

Another example is in image processing, when the initial features are chosen to be a wavelet (rescaled) system, in which case the corresponding random features $\{\psi_p\}_{1\leqslant p\leqslant P}$ are special cases of *random wavelet series*, objects that are well studied in signal processing and mathematical physics (see Aubry and Jaffard 2002; Durand 2008 for a study of the law of the spectrum of singularities of these series).

Noise model and texture generation. The construction of Gaussian objects (see paragraph 3.2.1) is highly flexible and enables to do automatic noise-texture generation easily, as explained in Deguy

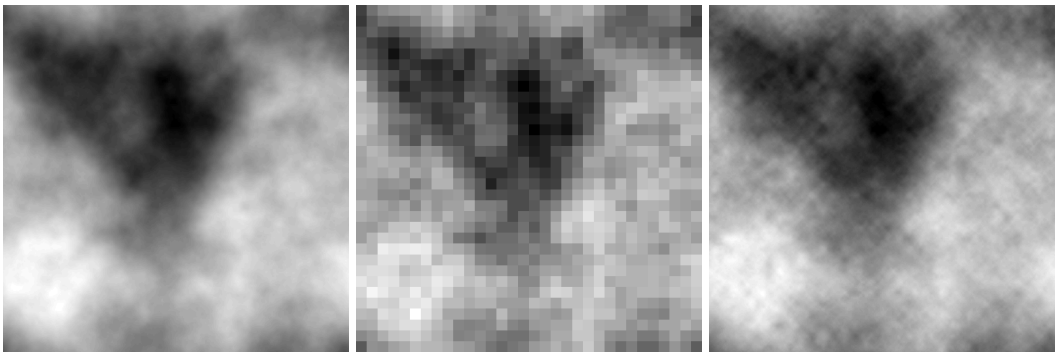

Figure 2: Example of an initial large texture (left), subsampled (middle), and possible recovery using regression with random projections (right)

and Benassi (2001). In their paper, the authors show that with the appropriate choice of the wavelet functions and when using rescaling coefficients of the form $\sigma_{j,l} = 2^{-js}$ with scale index j an position index l (see paragraph 3.2.1), where s is not a constant but is now a function of j and l, we can generate fractional Brownian motions, multi-scale fractional Brownian motions, and more generally what is called intermittent locally self-similar Gaussian processes.

In particular, for image texture generation they introduce a class of functions called *morphlets* that enables to perform approximations of intermittent locally self-similar Gaussian processes. These approximations are both numerically very efficient and have visually imperceptible differences to the targeted images, which make them very suitable for texture generation. The authors also allow other distributions than the Gaussian for the random variables ξ (which thus does not fit the theory presented here), and use this additional flexibility to produce an impressive texture generator.

Figure 2 illustrates an example performed on some simple texture model[1] where an image of size 512×512 is generated (two-dimensional Brownian sheet with Hurst index $H = 1.1$) (left) and then subsampled at 32×32 (middle), which provides the data samples for generating a regression function (right) using random features (generated from the symlets as initial features, in the simplest model when s is constant).

3. Gaussian Objects

We now describe some tools of Gaussian object theory that would be useful in later analysis of the method. Each random feature ψ_p built from Equation (1), when the coefficients are Gaussian, qualifies as a Gaussian object. It is thus natural to study some important features of Gaussian objects.

1. The authors wish to thank Pierre Chainais for performing experimental study of random projection methods applied to image processing, and for providing us with interesting pointers to related works.

3.1 Reminder of Gaussian Objects Theory

In all this section, $\mathcal{S}$ will refer to a vector space, $\mathcal{S}'$ to its topological dual, and $(\cdot,\cdot)$ to its duality product. The reader mostly interested in application of the random projection method may skip this section and directly go to Subsection 3.2 that provides examples of function spaces together with explicit construction of the abstract objects considered here.

Definition 2 (Gaussian objects) *A random variable $W \in \mathcal{S}$ is called a Gaussian object if for all $\nu \in \mathcal{S}'$, (ν, W) is a Gaussian (real-valued) variable. We further call any $a \in \mathcal{S}$ to be an* expectation *of W if*

$$\forall \nu \in \mathcal{S}', \mathbb{E}(\nu, W) = (\nu, a) < \infty,$$

and any $K : \mathcal{S}' \to \mathcal{S}$ to be a covariance operator *of W if*

$$\forall \nu, \nu' \in \mathcal{S}', Cov((\nu, W)(\nu', W)) = (\nu, K\nu') < \infty,$$

where Cov refers to the correlation between two real-valued random variables.

Whenever there exist such a and K, we say that W follows the law $\mathcal{N}(a,K)$. Moreover, W is called a centered Gaussian object if $a = 0$.

Kernel space. We only provide a brief introduction to this notion and refer the interested reader to Lifshits (1995) or Janson (1997) for refinements.

Let $I' : \mathcal{S}' \to L^2(\mathcal{S}, \mathcal{N}(0,K))$ be the canonical injection from the space of continuous linear functionals $\mathcal{S}'$ to the space of measurable linear functionals

$$L_2(\mathcal{S}; \mathbb{R}, \mathcal{N}(0,K)) = \left\{ z : \mathcal{S} \to \mathbb{R}, \mathbb{E}_{W \sim \mathcal{N}(0,K)} |z(W)|^2 < \infty \right\},$$

endowed with inner product $\langle z_1, z_2 \rangle = \mathbb{E}(z_1(W) z_2(W))$, that is, for any $\nu \in \mathcal{S}'$, I' is defined by $I'(\nu) = (\nu, \cdot)$. It belongs to $L_2(\mathcal{S}; \mathbb{R}, \mathcal{N}(0,K))$ since by definition of K we have $(\nu, K\nu) = \mathbb{E}(\nu, W)^2 < \infty$.

Then note that the space defined by $\mathcal{S}'_{\mathcal{N}} \stackrel{\text{def}}{=} \overline{I'(\mathcal{S}')}$, that is, the closure of the image of $\mathcal{S}'$ by I' in the sense of $L_2(\mathcal{S}; \mathbb{R}, \mathcal{N}(0,K))$, is a Hilbert space with inner product inherited from $L_2(\mathcal{S}; \mathbb{R}, \mathcal{N}(0,K))$.

Now under the assumption that I' is continuous (see Section 4.1 for practical conditions ensuring that this is the case), we can define the adjoint $I : \mathcal{S}'_{\mathcal{N}} \to \mathcal{S}$ of I', by duality. Indeed for any $\mu \in \mathcal{S}'$ and $z \in I'(\mathcal{S}')$, we have by definition that

$$(\mu, Iz) = \left\langle I'\mu, z \right\rangle_{\mathcal{S}'_{\mathcal{N}}} = \mathbb{E}_W((\mu, W) z(W)),$$

from which we deduce by continuity that $Iz = \mathbb{E}_W(W z(W))$. For the sake of clarity, this specifies for instance in the case when $\mathcal{S} = L_2(\mathcal{X}; \mathbb{R})$, for all $x \in \mathcal{X}$ as

$$(Iz)(x) = \mathbb{E}_W(W(x) z(W)).$$

Now that the two injection mappings I, I' have been defined, we are ready to provide the formal (though slightly abstract) definition for our main object of interest:

Definition 3 (Kernel space) *Provided that the mapping I' is continuous, then we define the* **kernel space** *of a centered Gaussian object W as $\mathcal{K} \stackrel{\text{def}}{=} I(\overline{I'(\mathcal{S}')}) \subset \mathcal{S}$.*

A more practical way of dealing with kernels is given by the two following lemmas that we use extensively in Section 3.2. First, the kernel space can be built alternatively based on a separable Hilbert space $\mathcal{H}$ as follows (Lifshits, 1995):

Lemma 4 (Construction of the Kernel space.) *Let $J : \mathcal{H} \to \mathcal{S}$ be an injective linear mapping such that $K = JJ'$, where J' is the adjoint operator of J. Then the kernel space of $\mathcal{N}(0,K)$ is $\mathcal{K} = J(\mathcal{H})$, endowed with inner product $\langle Jh_1, Jh_2 \rangle_H \stackrel{\text{def}}{=} \langle h_1, h_2 \rangle_{\mathcal{H}}$.*

We then conclude this section with the following Lemma from Lifshits (1995) that enables to define the expansion of a Gaussian object W.

Lemma 5 (Expansion of a Gaussian object) *Let $\{\varphi_i\}_{i\geqslant 1}$ be an orthonormal basis of $\mathcal{K}$ for the inner product $\langle \cdot, \cdot \rangle_{\mathcal{K}}$ and $\{\xi_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0,1)\}_{i\geqslant 1}$. Then $\sum_{i=1}^{\infty} \xi_i \varphi_i$ is a Gaussian object following the law $\mathcal{N}(0,K)$. It is called an* **expansion** *for $\mathcal{N}(0,K)$.*

Note that from Lemma 4, one can build an orthonormal basis $\{\varphi_i\}_{i\geqslant 1}$ by defining, for all $i \geqslant 1$, $\varphi_i = Jh_i$ where $\{h_i\}_{i\geqslant 1}$ is an orthonormal basis of $\mathcal{H}$ and J satisfies conditions of Lemma 4.

3.2 Interpretation of Some Function Spaces with Gaussian Objects Theory

In this section, we precise the link between Gaussian objects theory and reproducing kernel Hilbert spaces (RKHS) in order to provide more intuition about such objects. Indeed in many cases, the kernel space of a Gaussian object is a RKHS. Note, however, that in general, depending on the Gaussian object we consider, the former space may also be a more general space for instance when the Hilbert assumption is dropped (see Canu et al. 2002 about RKS). Therefore, there is no one-to-one correspondence between RKHS and kernel spaces of Gaussian objects and it is worth explaining when the two notions coincide. More importantly, this section shows various examples of classical function spaces, related to the construction of the space $\mathcal{G}_P$ for different choices of initial features $\{\varphi_i\}_{i\geqslant 1}$, and that can be useful for applications.

3.2.1 GAUSSIAN OBJECTS WITH A SUPPORTING HILBERT SPACE

In this subsection only, we make the assumption that $\mathcal{S} = \mathcal{H}$ is a Hilbert space and we introduce $\{e_i\}_{i\geqslant 1}$ an orthonormal basis of $\mathcal{H}$. Let us now consider $\xi_i \sim \mathcal{N}(0,1)$ i.i.d., and positive coefficients $\sigma_i \geqslant 0$ such that $\sum_i \sigma_i^2 < \infty$. Since $\sum_i \sigma_i^2 < \infty$, the Gaussian object $W = \sum_i \xi_i \sigma_i e_i$ is well defined and our goal is to identify the kernel of the law of W.

To this aim we first identify the function I'. Since $\mathcal{S}$ is a Hilbert space, then its dual is $\mathcal{S}' = \mathcal{S}$, thus we consider $f = \sum_i c_i e_i \in \mathcal{S}'$ for some $c \in \ell_2$. For such an f, we deduce by the previous section that the injection mapping is given by $(I'f)(g) = \sum_i c_i (g, e_i)$, and that we also have

$$\|I'f\|^2_{\mathcal{S}'_{\mathcal{N}}} = \mathbb{E}\big((I'f, W)^2\big) = \mathbb{E}\Big(\big(\sum_{i\geqslant 1} \sigma_i \xi_i c_i\big)^2\Big) = \sum_{i\geqslant 1} \sigma_i^2 c_i^2 .$$

Now, since $\|f\|_{\mathcal{S}} = \|c\|_{\ell_2}$, the continuity of I' is insured by the assumption that $\sum_i \sigma_i^2 < \infty$, and thus I is defined as in the previous section. Therefore, a function in the space $\mathcal{K}$ corresponding to f is of the form $\sum_i \sigma_i c_i e_i$, and one can easily check that the kernel space of the law of W is thus given by

$$\mathcal{K} = \Big\{ f_c = \sum_{i\geqslant 1} c_i e_i \ ; \sum_{i\geqslant 1} \Big(\frac{c_i}{\sigma_i}\Big)^2 < \infty \Big\},$$

endowed with inner product $(f_c, f_d)_{\mathcal{K}} = \sum_{i\geqslant 1} \frac{c_i d_i}{\sigma_i^2}$.

Reproducing Kernel Hilbert Spaces (RKHS). Note that if we now introduce the functions $\{\varphi_i\}_{i\geqslant 1}$ defined by $\varphi_i \stackrel{\text{def}}{=} \sigma_i e_i \in \mathcal{H}$, then we get

$$\mathcal{K} = \Big\{ f_\alpha = \sum_{i\geqslant 1} \alpha_i \varphi_i \; ; \|\alpha\|_{l_2} < \infty \Big\},$$

endowed with inner product $(f_\alpha, f_\beta)_{\mathcal{K}} = \langle \alpha, \beta \rangle_{l_2}$. For instance, if we consider that $\mathcal{H} \subset L_{2,\mu}(\mathcal{X};\mathbb{R})$ for some reference measure μ, and that $\{e_i\}_{i\geqslant 1}$ are orthonormal w.r.t. $L_{2,\mu}(\mathcal{X};\mathbb{R})$, then $\mathcal{K}$ appears to be a RKHS that can be made fully explicit; its kernel is defined by $k(x,y) = \sum_{i=1}^\infty \sigma_i^2 e_i(x) e_i(y)$, and $\{\sigma_i\}_{i\geqslant 1}$ and $\{e_i\}_{i\geqslant 1}$ are trivially the eigenvalues and eigenfunctions of the integral operator $T_k : L_{2,\mu}(\mathcal{X}) \to L_{2,\mu}(\mathcal{X})$ defined by $(T_k(f))(x) = \int_{\mathcal{X}} k(x,y) f(y) d\mu(y)$.

Wavelet basis and Besov spaces. In this paragraph, we now apply the previous construction to the case when the $\{e_i\}_{i\geqslant 1}$ are chosen to be a wavelet basis of functions defined on $\mathcal{X} = [0,1]$ with reference measure μ being the Lebesgue measure. Let e denote the mother wavelet function, and let us write $e_{j,l}$ the ith element of the basis, with $j \in \mathbb{N}$ a scale index and $l \in \{0, \dots, 2^j - 1\}$ a position index, where we re-index all families indexed by i with the indice j,l. Let us define the coefficients $\{\sigma_i\}_{i\geqslant 1}$ to be exponentially decreasing with the scale index:

$$\sigma_{j,l} \stackrel{\text{def}}{=} 2^{-js} \text{ for all } j \geqslant 0 \text{ and } l \in \{0, \dots, 2^j - 1\},$$

where we introduced some positive real number s.

Now assume that for some $q \in \mathbb{N} \setminus \{0\}$ such that $q > s$, the mother wavelet function e belongs to $\mathcal{C}^q(\mathcal{X})$, the set of q-times continuously differentiable functions on $\mathcal{X}$, and admits q vanishing moments. The reason to consider such case is that the (homogeneous) Besov space $\mathcal{B}_{s,2,2}([0,1]^d)$ then admits the following known characterization (independent of the choice of the wavelets, see Frazier and Jawerth 1985; Bourdaud 1995):

$$\mathcal{B}_{s,2,2}(\mathcal{X};\mu) = \Big\{ f \in L_{2,\mu}(\mathcal{X}) ; \; \|f\|_{s,2,2}^2 \stackrel{\text{def}}{=} \sum_{j=1}^\infty \Big[2^{2js} \sum_{l=0}^{2^j-1} |\langle f, e_{j,l} \rangle|^2 \Big] < \infty \Big\}.$$

On the other hand, with the notations above, where in particular $\varphi_{j,l} = \sigma_{j,l} \varepsilon_{j,l}$, we deduce that the kernel space of the Gaussian object $W = \sum_{j,l} \xi_{j,l} \varphi_{j,l}$ (that we call a Scrambled wavelet), is simply the space

$$\mathcal{K} = \Big\{ f_\alpha = \sum_{j,l} \alpha_{j,l} \varphi_{j,l} \; ; \sum_{j,l} \alpha_{j,l}^2 < \infty \Big\},$$

and a straightforward computation shows that $\|\alpha\|_{l_2}^2 = \|f_\alpha\|_{s,2,2}^2$, so that $\mathcal{K} = \mathcal{B}_{s,2,2}(\mathcal{X};\mu)$. Moreover, assuming that the mother wavelet is bounded by a constant λ and has compact support $[0,1]$, then we have the property that is useful in view of our main Theorem

$$\sup_{x\in\mathcal{X}} \|\varphi(x)\|^2 \leqslant \frac{\lambda^2}{1 - 2^{-2s+1}}.$$

Note that a similar construction applies to the case when the orthonormal basis $\{e_i\}_{i\geqslant 1}$ is chosen to be a Fourier basis of functions, and the coeficients $\{\sigma_i\}_{i\geqslant 1}$ are chosen to be of the form $\sigma_i = i^{-s}$.

3.2.2 GAUSSIAN OBJECTS DEFINED BY A CARLEMAN EXPANSION

We now no longer assume that the supporting space $\mathcal{S}$ is a Hilbert space. In this case, it is still possible to generate a Gaussian object with kernel space being a RKHS by resorting to Carleman operators.

A Carleman operator is a linear injective mapping $J : \mathcal{H} \mapsto \mathcal{S}$ (where $\mathcal{H}$ is a Hilbert space) such that $J(h)(t) = \int \Gamma_t(s)h(s)ds$ where $(\Gamma_t)_t$ is a collection of functions of $\mathcal{H}$. As shown for instance in Canu et al. (2002); Saitoh (1988), there is a bijection between Carleman operators and the set of RKHSs. In particular, $J(\mathcal{H})$ is a RKHS.

A Gaussian object admitting $J(\mathcal{H})$ as a kernel space can be built as follows. By application of Lemma 5, we have that $\mathcal{K} = J(\mathcal{H})$ endowed with the inner product $\langle Jh_1, Jh_2\rangle_{\mathcal{K}} \stackrel{\text{def}}{=} \langle h_1, h_2\rangle_{\mathcal{H}}$ is the kernel space of $\mathcal{N}(0, JJ')$. Now, if we consider an orthonormal basis $\{e_i\}_{i\geqslant 1}$ of $\mathcal{H}$, an application of Lemma 5 shows that the functions $\{\varphi_i\}_{i\geqslant 1}$ defined by $\varphi_i \stackrel{\text{def}}{=} J(e_i)$ form an orthonormal basis of $J(\mathcal{H})$ and are such that the object $W = \sum_{i\geqslant 1} \xi_i \varphi$ is first a well-defined Gaussian object and then an expansion for the law $\mathcal{N}(0, JJ')$. We call this expansion a Carleman expansion. Note that this expansion is bottom-up whereas the Mercer expansion of a kernel via the spectral Theorem is top-down, see, for example, Zaanen (1960).

Cameron-Martin space. We apply as an example this construction to the case of the Brownian motion and the Cameron-Martin space.

Let $\mathcal{S} = \mathcal{C}([0,1])$ be the space of continuous real-valued functions of the unit interval. Then $\mathcal{S}'$ is the set of signed measures and we can define the dual product by $(\nu, f) = \int_{[0,1]} f d\nu$. It is straightforward to check that the Brownian motion indexed by $[0,1]$ is a Gaussian object $W \in \mathcal{S}$, with $a \equiv 0$ and K defined by $(K\nu)(t) = \int_{[0,1]} \min(s,t)\nu(ds)$.

Kernel space. We consider the Hilbert space $\mathcal{H} = L_2([0,1])$ and define the mapping $J : \mathcal{H} \mapsto \mathcal{S}$ by

$$(Jh)(t) = \int_{[0,t]} h(s)ds;$$

simple computations show that $(J'\nu)(t) = \nu([t,1])$, $K = JJ'$ and that J is a Carleman operator. Therefore, the kernel space $\mathcal{K}$ is equal to $J(L_2([0,1]))$, or more explicitly

$$\mathcal{K} = \left\{k \in H^1([0,1]); k(0) = 0\right\},$$

where $H^1([0,1])$ is the Sobolev space of order 1.

Expansion of the Brownian motion. We build a Carleman expansion for the Brownian motion thanks to the Haar basis of $L^2([0,1])$, whose image by J defines an orthonormal basis of $\mathcal{K}$; the Haar basis $(e_0, \{e_{j,l}\}_{j,l\in\mathbb{N}})$ is defined in a wavelet-way via a mother function $e(x) = \mathbb{I}_{[0,1/2[} - \mathbb{I}_{[1/2,1[}$ and father function $e_0(x) = \mathbb{I}_{[0,1]}(x)$ with functions $\{e_{j,l}\}_{j,l\in\mathbb{N}}$ defined for any scale $j \geqslant 1$ and translation index $0 \leqslant l \leqslant 2^j - 1$ by

$$e_{j,l}(x) \stackrel{\text{def}}{=} 2^{j/2} e(2^j x - l).$$

An orthonormal basis of the kernel space of the Brownian motion W and an expansion of W is thus obtained by

$$W = \sum_{j,l\geqslant 1} \xi_{j,l}\varphi_{j,l} + \xi_0\varphi_0,$$

$$\text{with } \varphi_{j,l}(x) = Je_{j,l}(x) = 2^{-j/2}\Lambda(2^j x - l) \text{ and } \varphi_0(x) = Je_0(x) = x,$$

where $\Lambda(x) = x\mathbb{I}_{[0,1/2[} + (1-x)\mathbb{I}_{[1/2,1[}$ is the mother hat function.

Bounded energy. Note that the rescaling factor inside $\varphi_{j,l}$ naturally appears as $2^{-j/2}$, and not as $2^{j/2}$ as usually defined in wavelet-like transformations. Note also that since the support of the mother function Λ is $[0,1]$, and also $\|\Lambda\|_\infty \leqslant 1/2$, then for any $x \in [0,1]^d$, for all j there exists at most one $l = l(x)$ such that $\varphi_{j,l}(x) \neq 0$, and we have the property that

$$\|\varphi(x)\|^2 = \sum_{j \geqslant 1} \varphi_{j,l(x)}(x)^2 \leqslant \sum_{j \geqslant 1} (2^{-j/2}\|\Lambda\|_\infty)^2 \leqslant \frac{1}{2}.$$

Remark 6 *This construction can be extended to the dimension $d > 1$ in at least two ways. Consider the space $\mathcal{S} = \mathcal{C}([0,1]^d)$, and the Hilbert space $\mathcal{H} = L_2([0,1]^d)$. Then if we define J to be the volume integral $(Jh)(t) = \int_{[0,t]} h(s)ds$ where $[0,t] \subset [0,1]^d$, this corresponds to the covariance operator defined by $(K\nu)(t) = \int_{[0,1]^d} \Pi_{i=1}^d \min(s_i,t_i)\nu(ds)$, that is, to the Brownian sheet defined by tensorization of the Brownian motion. The corresponding kernel space in this case is thus $\mathcal{K} = J(L^2([0,1]^d))$, endowed with the norm $\|f\|_\mathcal{K} = \|\frac{\partial^d f}{\partial x_1 \ldots \partial x_d}\|_{L^2([0,1]^d)}$. It corresponds to the* Cameron-Martin space *(Janson, 1997) of functions having a d-th order crossed (weak) derivative $\frac{\partial^d f}{\partial x_1 \ldots \partial x_d}$ that belongs to $L^2([0,1]^d)$, vanishing on the "left" boundary (edges containing 0) of the unit d-dimensional cube. A second possible extension that is not detailed here would be to consider the isotropic Brownian sheet.*

3.3 A Johnson-Lindenstrauss Lemma for Gaussian Objects

In this section, we derive a version of the Johnson-Lindenstrauss' lemma that applies to the case of Gaussian objects.

The original Johnson-Lindenstrauss' lemma can be stated as follows; its proof directly uses concentration inequalities (Cramer's large deviation Theorem from 1938) and may be found, for example, in Achlioptas (2003).

Lemma 7 *Let A be a $P \times F$ matrix of i.i.d. Gaussian $\mathcal{N}(0,1/P)$ entries. Then for any vector α in $\mathbb{R}^F$, the random (with respect to the choice of the matrix A) variable $\|A\alpha\|^2$ concentrates around its expectation $\|\alpha\|^2$ when P is large: for $\varepsilon \in (0,1)$, we have*

$$\mathbb{P}\Big(\|A\alpha\|^2 \geqslant (1+\varepsilon)\|\alpha\|^2\Big) \leqslant e^{-P(\varepsilon^2/4 - \varepsilon^3/6)},$$
$$\text{and } \mathbb{P}\Big(\|A\alpha\|^2 \leqslant (1-\varepsilon)\|\alpha\|^2\Big) \leqslant e^{-P(\varepsilon^2/4 - \varepsilon^3/6)}.$$

Remark 8 *Note the Gaussianity is not mandatory here, and this is also true for other distributions, such as:*

- *Rademacher distributions, that is, which takes values $\pm 1/\sqrt{P}$ with equal probability $1/2$,*
- *Distribution taking values $\pm\sqrt{3/P}$ with probability $1/6$ and 0 with probability $2/3$.*

What is very important is the scaling factor $1/P$ appearing in the variance of $\mathcal{N}(0,1/P)$.

This Lemma together with the measurability properties of Gaussian objects enable us to derive the following statement.

Lemma 9 *Let $\{x_n\}_{n\leqslant N}$ be N (deterministic) points of $\mathcal{X}$. Let $A : \ell_2(\mathbb{R}) \mapsto \mathbb{R}^P$ be the operator defined with i.i.d. Gaussian $\mathcal{N}(0,1/P)$ variables $(A_{i,p})_{i\geqslant 1, p\leqslant P}$, such that for all $\alpha \in \ell_2(\mathbb{R})$, then*

$$(A\alpha)_p = \sum_{i\geqslant 1} \alpha_i A_{i,p}\,.$$

Let us also define $\psi_p = \sum_{i\geqslant 1} A_{i,p}\varphi_i$, $f_\alpha = \sum_{i\geqslant 1}\alpha_i\varphi_i$ and $g_\beta = \sum_{p=1}^{P}\beta_p\psi_p$.

Then, A is well-defined and for all $P \geqslant 1$, for all $\varepsilon \in (0,1)$, with probability larger than $1 - 4Ne^{-P(\varepsilon^2/4-\varepsilon^3/6)}$ w.r.t. the Gaussian random variables,

$$\|f_\alpha - g_{A\alpha}\|_N^2 \leqslant \varepsilon^2\|\alpha\|^2\frac{1}{N}\sum_{n=1}^{N}\|\varphi(x_n)\|^2\,,$$

where we recall that by assumption, for any x, $\varphi(x) \stackrel{\text{def}}{=} (\varphi_i(x))_{i\geqslant 1}$ is in ℓ_2.

This result is natural in view of concentration inequalities, since for all $x \in \mathcal{X}$, the expectation satisfies $\mathbb{E}_{\mathcal{P}_\mathcal{G}}(g_{A\alpha}(x)) = f_\alpha(x)$ and the variance $\mathbb{V}_{\mathcal{P}_\mathcal{G}}(g_{A\alpha}(x)) = \frac{1}{P}(f_\alpha^2(x) + \|\alpha\|^2\|\varphi(x)\|^2)$. See Appendix A.1 for the full proof.

Note also that a natural idea in order to derive generalization bounds would be to derive a similar result uniformly over $\mathcal{X}$ instead of a union bound over the samples. However, while such extension would be possible for finite dimensional spaces $\mathcal{F}$ (by resorting to covers) these kind of results are not possible in the general case, since $\mathcal{F}$ is typically big.

More intuition. Let us now provide some more intuition about when such a result is interesting. In interesting situations described in Section 4 we consider a number of projections P lower than the number of data samples N, typically P is of order $\sqrt{N}$. Thus, it may seem counter-intuitive that we can approximate—at a set of N points—a function f_α that lies in a high (possibly infinite) dimensional space $\mathcal{F}$ by a function $g_{A\alpha}$ in a space $\mathcal{G}$ of dimension $P < N$.

Of course in general this is not possible. To illustrate this case, let us consider that there is no noise, assume that all points $(x_n)_{n\leqslant N}$ belong to the unit sphere, and that φ is the identity of $\mathcal{X} = \mathbb{R}^D$. Thus a target function f is specified by some $\alpha \in \mathbb{R}^D$ (where D is assumed to be large, that is, $D > N$) and the response values are $y_n = f_\alpha(x_n) = \alpha^T x_n$. Write $\widehat{y} \in \mathbb{R}^D$ the estimate $g_{A\alpha}$ at the points, that is, such that $\widehat{y}_n = g_{A\alpha}(x_n)$. In that case, the bound of Lemma 9 provides an average quadratic estimation error $\frac{1}{N}\|y-\widehat{y}\|^2$ of order $\dfrac{\log(N/\delta)}{P}||\alpha||^2$, with probability $1-\delta$.

On the other hand the zero-value regressor has an estimation error of

$$\frac{1}{N}\|y\|^2 = \frac{1}{N}\sum_{n=1}^{N}(\alpha^T x_n)^2 = \alpha^T S\alpha\,,\ \text{where } S \stackrel{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N} x_n x_n^T \in \mathbb{R}^{D\times D}\,.$$

This shows that the result of Lemma 9 is essentially interesting when $\dfrac{\alpha^T S\alpha}{\|\alpha\|^2} \gg \dfrac{\log(N/\delta)}{P}$, which may not happen in certain cases: Indeed if we specifically choose $x_n = e_n \in \mathbb{R}^D$, for $n \leqslant N \leqslant D$, where $(e_1,\dots,e_D)$ denotes the Euclidean basis of $\mathbb{R}^D$, then for such a choice, we have

$$\frac{\alpha^T S\alpha}{||\alpha||^2} = \frac{\sum_{d=1}^{N}\alpha_d^2}{N\sum_{d=1}^{D}\alpha_d^2} \leqslant \frac{1}{N} \leqslant \frac{\log(N/\delta)}{P}\,,$$

which means that the random projection method fails to recover a better solution than a trivial one. The reason why it fails is that in that case the points $\{x_n\}_{n\leqslant N}$ lie in a subspace of $\mathbb{R}^D$ of *high-dimension N*, that is, such that the information at any set of points does not help us to predict the value at any other point. Essentially, what Lemma 9 tells us is that the random projection method will work when the points $\{x_n\}_{n\leqslant N}$ lie in a vector subspace of smaller dimension $d_0 < N$ and that the d_0 corresponding coefficients of α contain most information about α (i.e., the other $D - d_0$ coordinates are small). Let us illustrate this case by considering the case where $x_n = e_{1+(n \bmod d_0)}$ for all $n \leqslant N$. In that case, we have (for N multiple of d_0),

$$\frac{\alpha^T S \alpha}{||\alpha||^2} = \frac{\sum_{d=1}^{d_0} \alpha_d^2}{d_0 \sum_{d=1}^{D} \alpha_d^2},$$

which is larger than $\frac{\log(N/\delta)}{P}$ whenever the components $\{\alpha_d\}_{d>d_0}$ decrease fast and P is large enough, in which case, the random projection method will work well.

Now introducing features, the condition says that the number of relevant features should be relatively small, in the sense that the parameter should mostly contain information at the corresponding coordinates, which is the case in many functional spaces, such as the Sobolev and Besov spaces (for which $D = \infty$) described in Section 2 and Section 3.2.1, paragraph "Wavelet basis and Besov spaces", for which $\|\alpha\|$ equals the norm of the function f_α in the corresponding space. Thus a "smooth" function f_α (in the sense of having a low functional norm) has a low norm of the parameter $\|\alpha\|$, and is thus well approximated with a small number of wavelets coefficients. Therefore, Lemma 9 is interesting and the random projection method will work in such cases (i.e., the additional projection error is controlled by a term of order $\|\alpha\|^2 \frac{\log(N/\delta)}{P}$).

4. Regression With Random Subspaces

In this section, we describe the construction of the random subspace $\mathcal{G}_P \subset \mathcal{F}$ defined as the span of the random features $\{\psi_p\}_{p\leqslant P}$ generated from the initial features $\{\varphi_i\}_{i\geqslant 1}$. This method was originally described in Maillard and Munos (2009) for the case when $\mathcal{F}$ is of finite dimension, and we extend it here to the non-obvious case of infinite dimensional spaces $\mathcal{F}$, which relies on the fact that the randomly generated features $\{\psi_p\}_{p\leqslant P}$ are well-defined Gaussian objects.

The next subsection is devoted to the analysis of the approximation power of the random features space. We first give a survey of existing results on regression together with the standard hypothesis under which they hold in section 4.2, then we describe in section 4.4 an algorithm that builds the proposed regression function and provide excess risk bounds for this algorithm.

4.1 Construction of Random Subspaces

Assumption on initial features. In this paper we assume that the set of features $\{\varphi_i\}_{i\geqslant 1}$ are continuous and satisfy the assumption that,

$$\sup_{x\in\mathcal{X}} \|\varphi(x)\|^2 < \infty, \text{ where } \|\varphi(x)\|^2 \stackrel{\text{def}}{=} \sum_{i\geqslant 1} \varphi_i(x)^2. \tag{3}$$

Note that all examples in Section 3 satisfy this condition.

Random features. The random subspace $\mathcal{G}_P$ is generated by building a set of P *random features* $\{\psi_p\}_{1\leqslant p\leqslant P}$ defined as linear combinations of the initial features $\{\varphi_i\}_{1\geqslant 1}$ weighted by random coefficients:

$$\psi_p(x) \stackrel{\text{def}}{=} \sum_{i\geqslant 1} A_{p,i}\varphi_i(x), \text{ for } 1 \leqslant p \leqslant P,$$

where the (infinitely many) coefficients $A_{p,i}$ are drawn i.i.d. from a centered distribution with variance $1/P$. Here we explicitly choose a Gaussian distribution $\mathcal{N}(0,1/P)$. Such a definition of the features ψ_p as an infinite sum of random variable is not obvious (this is an expansion of a Gaussian object) and we refer to the Section 3 for elements of theory about Gaussian objects and Lemma 5 for the expansion of a Gaussian object. It is shown that under Assumption (3), the random features are well defined. Actually, they are random samples of a centered Gaussian process indexed by the space $\mathcal{X}$ with covariance structure given by $\frac{1}{P}\langle\varphi(x),\varphi(x')\rangle$, where we use the notation $\langle u,v\rangle \stackrel{\text{def}}{=} \sum_i u_i v_i$ for two square-summable sequences u and v. Indeed, $\mathbb{E}_{A_p}[\psi_p(x)] = 0$, and

$$\mathrm{Cov}_{A_p}(\psi_p(x),\psi_p(x')) = \mathbb{E}_{A_p}[\psi_p(x)\psi_p(x')] = \frac{1}{P}\sum_{i\geqslant 1}\varphi_i(x)\varphi_i(x') = \frac{1}{P}\left\langle\varphi(x),\varphi(x')\right\rangle .$$

The continuity of each of the initial features $\{\varphi_i\}_{i\geqslant 1}$ guarantees that there exists a continuous version of the process ψ_p that is thus a Gaussian process.

Random subspace. We finally define $\mathcal{G}_P \subset \mathcal{F}$ to be the (random) vector space spanned by those features, that is,

$$\mathcal{G}_P \stackrel{\text{def}}{=} \Big\{ g_\beta(x) \stackrel{\text{def}}{=} \sum_{p=1}^{P} \beta_p\psi_p(x), \beta\in\mathbb{R}^P \Big\}.$$

We now want to compute a high probability bound on the excess risk of an estimator built using the random space $\mathcal{G}_P$. To this aim, we first quickly review known results in regression and see what kind of estimator can be considered and what results can be applied. Then we compute a high probability bound on the approximation error of the considered random space w.r.t. to initial space $\mathcal{F}$. Finally, we combine both bounds in order to derive a bound on the excess risk of the proposed estimate.

4.2 Reminder of Results on Regression

Short review of existing results. For the sake of completeness, we now review other existing results in regression that may or may not apply to our setting. Indeed it seems natural to apply existing results for regression to the space $\mathcal{G}_P$. For that purpose, we focus on the randomness coming from the data points only, and not from the Gaussian entries. We will thus consider in this subsection only a space $\mathcal{G}$ that is the span over a *deterministic* set of P functions $\{\psi_p\}_{p\leqslant P}$, and we will write, for a convex subset $\Theta\subset\mathbb{R}^P$,

$$\mathcal{G}_\Theta \stackrel{\text{def}}{=} \left\{ g_\theta\in\mathcal{G}; \theta\in\Theta \right\}.$$

Similarly, we write $g^\star \stackrel{\text{def}}{=} \operatorname*{argmin}_{g\in\mathcal{G}} R(g)$ and $g^\star_\Theta \stackrel{\text{def}}{=} \operatorname*{argmin}_{g\in\mathcal{G}_\Theta} R(g)$. Examples of well studied estimates are:

- $\widehat{g}^{ols} \stackrel{\text{def}}{=} \text{argmin}_{g\in\mathcal{G}} R_N(g)$, the ordinary least-squares (ols) estimate.
- $\widehat{g}^{erm} \stackrel{\text{def}}{=} \text{argmin}_{g\in\mathcal{G}_\Theta} R_N(g)$ the empirical risk minimizer (erm) that coincides with the ols when $\Theta = \mathbb{R}^P$.
- $\widehat{g}^{ridge} \stackrel{\text{def}}{=} \text{argmin}_{g\in\mathcal{G}} R_N(g) + \lambda\|\theta\|$, $\widehat{g}^{lasso} \stackrel{\text{def}}{=} \text{argmin}_{g\in\mathcal{G}} R_N(g) + \lambda\|\theta\|_1$.

We also introduce for convenience g_B, the truncation at level $\pm B$ of some $g \in \mathcal{G}$, defined by $g_B(x) \stackrel{\text{def}}{=} T_B[g(x)]$, where $T_B(u) \stackrel{\text{def}}{=} \begin{cases} u & \text{if } |u| \leqslant B, \\ B\,\text{sign}(u) & \text{otherwise.} \end{cases}$

There are at least 9 different theorems that one may want to apply in our setting. Since those theorems hold under some assumptions, we list them now. Unfortunately, as we will see, these assumptions are usually slightly too strong to apply in our setting, and thus we will need to build our own analysis instead.

Assumptions Let us list the following assumptions.

- Noise assumptions: (for some constants B, B_1, σ, ξ)
  (N_1) $|Y| \leqslant B_1$,
  (N_2) $\sup_{x\in\mathcal{X}} \mathbb{E}(Y|X=x) \leqslant B$,
  (N_3) $\sup_{x\in\mathcal{X}} \mathbb{V}(Y|X=x) \leqslant \sigma^2$,
  (N_4) $\forall k \geqslant 3$ $\sup_{x\in\mathcal{X}} \mathbb{E}(|Y|^k|X=x) \leqslant \sigma^2 k!\xi^{k-2}$.
- Moment assumptions: (for some constants σ, a, M)
  (M_1) $\sup_{x\in\mathcal{X}} \mathbb{E}([Y - g^\star_\Theta(X)]^2|X=x) \leqslant \sigma^2$,
  (M_2) $\sup_{x\in\mathcal{X}} \mathbb{E}(\exp[a|Y - g^\star_\Theta(X)|]|X=x) \leqslant M$,
  (M_3) $\exists g_0 \in \mathcal{G}_\Theta$ $\sup_{x\in\mathcal{X}} \mathbb{E}(\exp[a|Y - g_0(X)|]|X=x) \leqslant M$.
- Function space assumptions for $\mathcal{G}$: (for some constant D)
  (G_1) $\sup_{g_1,g_2\in\mathcal{G}_\Theta} \|g_1 - g_2\|_\infty \leqslant D$,
  (G_2) $\exists g_0 \in \mathcal{G}_\Theta$, known, such that $\|g_0 - g^\star_\Theta\|_\infty \leqslant D$.
- Dictionary assumptions:
  (D_1) $L = \max_{1\leqslant p\leqslant P} \|\psi_p\|_\infty < \infty$,
  (D_2) $L = \sup_{x\in\mathcal{X}} \|\psi(x)\|_2 < \infty$,
  (D_3) esssup$\|\psi(X)\|_2 \leqslant L$,
  (D_4) $L = \inf_{\{\psi'_p\}_{p\leqslant P}} \sup_{\theta\in\mathbb{R}^d-\{0\}} \frac{\|\sum_{p=1}^P \theta_p\psi'_p\|_\infty}{\|\theta\|_\infty} < \infty$ where the infimum is over all orthonormal basis of $\mathcal{G}$ w.r.t. to $L_{2,\mathcal{P}_\mathcal{X}}(\mathcal{X};\mathbb{R})$.
- Orthogonality assumptions:
  (O_1) $\{\psi_p\}_{p\leqslant P}$ is an orthonormal basis of $\mathcal{G}$ w.r.t. to $L_{2,\mathcal{P}_\mathcal{X}}(\mathcal{X};\mathbb{R})$,
  (O_2) $\det(\Psi) > 0$, where $\Psi = \mathbb{E}(\psi(X)\psi(X)^T)$ is the Gram matrix.
- Parameter space assumptions:
  (P_1) $\sup_{\theta\in\Theta} \|\theta\|_\infty < \infty$,
  (P_2) $\|\theta^\star\|_1 \leqslant S$ where $\theta^\star$ is such that $g_{\theta^\star} = g^\star_\Theta$ and S is known,
  (P_3) $\sup_{\theta\in\Theta} \|\theta\|_2 \leqslant 1$.

Theorem 10 (Györfi et al. 2002) *Let $\Theta = \mathbb{R}^P$. Under assumption (N_2) and (N_3), the truncated estimator $\widehat{g}_L = T_L(\widehat{g}^{ols})$ satisfies*

$$\mathbb{E}R(\widehat{g}_L) - R(f^{(reg)}) \leqslant 8[R(g^*) - R(f^{(reg)})] + \kappa \frac{(\sigma^2 \vee B^2) P \log(N)}{N},$$

where κ is some numerical constant and $f^{(reg)}(x) \stackrel{\text{def}}{=} \mathbb{E}(Y|X=x)$.

Theorem 11 (Catoni 2004) *Let $\Theta \subset \mathbb{R}^P$. Under assumption (M_3), (G_1) and (O_2), there exists constants $C_1, C_2 > 0$ (depending only on a, M and D) such that with probability $1-\delta$, provided that*

$$\left\{ g \in \mathcal{G} \,;\, R_N(g) \leqslant R_N(\widehat{g}^{ols}) + C_1 \frac{P}{N} \right\} \subset \mathcal{G}_\Theta,$$

then the ordinary least squares estimate satisfies

$$R(\widehat{g}^{ols}) - R(g^\star_\Theta) \leqslant C_2 \frac{P + \log(\delta^{-1}) + \log(\frac{det\widehat{\Psi}}{det\Psi})}{N},$$

where $\widehat{\Psi} = \frac{1}{N}\sum_{i=1}^N \psi(X_i)\psi(X_i)^T$ is the empirical Gram matrix.

Theorem 12 (Audibert and Catoni 2010 from Alquier 2008) *Let $\Theta = \mathbb{R}^P$. Under assumption (N_1) and (G_2), there exists a randomized estimate $\widehat{g}$ that only depends on g_0, L, C, such that for all $\delta > 0$, with probability larger than $1-\delta$ w.r.t. all sources of randomness,*

$$R(\widehat{g}) - R(g^\star) \leqslant \kappa(B_1^2 + D^2) \frac{P \log(3\nu_{\min}^{-1}) + \log(\log(N)\delta^{-1})}{N},$$

where κ does not depend on P and N, and $\nu_{\min}$ is the smallest eigenvalue of Ψ.

Theorem 13 (Koltchinskii 2006) *Let $\Theta \subset \mathbb{R}^P$. Under assumption (N_1), (D_3) and (P_3), $\widehat{g}^{erm}$ satisfies, for any $\delta > 0$ with probability higher than $1-\delta$,*

$$R(\widehat{g}^{erm}) - R(g^\star_\Theta) \leqslant \kappa(B_1 + L)^2 \frac{rank(\Psi) + \log(\delta^{-1})}{N},$$

where κ is some constant.

Theorem 14 (Birgé and Massart 1998) *Let $\Theta \subset \mathbb{R}^P$. Under assumption (M_3), (G_1) and (D_4), for all $\delta > 0$ with probability higher than $1-\delta$,*

$$R(\widehat{g}^{erm}) - R(g^\star_\Theta) \leqslant \kappa(a^{-2} + D^2) \frac{P \log(2 + (L^2/N) \wedge (N/P)) + \log(\delta^{-1})}{N},$$

where κ is some constant depending only on M.

Theorem 15 (Tsybakov 2003) *Let $\Theta = \mathbb{R}^P$. Under assumption (N_2), (N_3) and (O_1), the projection estimate $\widehat{g}^{proj}$ satisfies*

$$\mathbb{E}(R(\widehat{g}^{proj})) - R(g^\star) \leqslant \frac{(\sigma^2 + B^2)P}{N}.$$

Theorem 16 (Caponnetto and De Vito 2007) *Under assumption (M_2) and (D_2), for all $\delta > 0$ for $\lambda = PL^2 \log^2(\delta^{-1})/N \leqslant \nu_{\min}$, with probability higher than $1-\delta$,*

$$R(\widehat{g}^{ridge}) - R(g^\star_\Theta) \leqslant \kappa(a^{-2} + \frac{\lambda L^2 \|\theta^\star\|^2 \log^2(\delta^{-1})}{\nu_{\min}}) \frac{P \log^2(\delta^{-1})}{N},$$

where κ is some constant depending only on M.

Theorem 17 (Alquier and Lounici 2011) *Let $\Theta = \mathbb{R}^P$ and define for all $\alpha \in (0,1)$ the prior $\pi_\alpha(J) = \frac{\alpha^{|J|}}{\sum_{i=0}^N \alpha^i} \binom{P}{|J|}^{-1}$ for all $J \subset 2^P$. Under assumption $(N_2), (N_3)$, (N_4), (D_1) and (P_2), by setting $\lambda = \frac{N}{2C}$ where*

$$C \stackrel{\text{def}}{=} \max\{64\sigma^2 + (2B + L(2S + \frac{1}{N}))^2, 64[\xi + 2B + L(2S + \frac{1}{N})]L(2S + \frac{1}{N})\},$$

the randomized aggregate estimator $\widehat{g}$ defined in Alquier and Lounici (2011) based on prior π_α satisfies, for any $\delta > 0$ with probability higher than $1-\delta$,

$$R(\widehat{g}) - R(g^\star_\Theta) \leqslant C \frac{S^\star \log(\frac{(S+c)eNP}{\alpha S^\star}) + \log(2\delta^{-1}/(1-\alpha))}{N} + \frac{3L^2}{N^2},$$

where $S^\star = \|\theta^\star\|_0$.

Theorem 18 (Audibert and Catoni 2010) *Let $\Theta \subset \mathbb{R}^P$. Under assumption (M_1), (G_1) and (P_1) so that one can define the uniform probability distribution over Θ, there exists a random estimator $\widehat{g}$ (drawn according to a Gibbs distribution $\widehat{\pi}$) that satisfies, with probability higher than $1-\delta$ w.r.t. all source of randomness,*

$$R(\widehat{g}) - R(g^\star_\Theta) \leqslant (2\sigma + D)^2 \frac{16.6P + 12.5\log(2\delta^{-1})}{N}.$$

Note that Theorem 10 and Theorem 15 provide a result in expectation only, which is not enough for our purpose, since we need high probability bounds on the excess risk in order to be able to handle the randomness of the space $\mathcal{G}_P$.

Assumptions satisfied by the random space $\mathcal{G}_P$

We now discuss the assumptions that are satisfied in our setting where $\mathcal{G}$ is a random space $\mathcal{G}_P$ built from the random features $\{\psi_p\}_{p \leqslant P}$, in terms of assumptions on the underlying initial space $\mathcal{F}$.

- The noise assumptions (N) do not concern $\mathcal{G}$.
- The moment assumptions (M) are not restrictive. By combining similar assumptions on $\mathcal{F}$, the results on approximation error of Section 4.3 can be shown to hold (with different constants).
- Assumptions (P) are generally too strong. For (P_1), the reason is that there is no high probability link between $\|A\alpha\|_\infty$ and $\|\alpha\|$ for usual norms. Now even if $\alpha^\star$ is sparse or has low l_1-norm, this does not imply this is the case for $\beta^\star = \text{argmin}_{\beta \in \mathbb{R}^P} R(g_\beta)$ or $A\alpha^\star$ in general, thus (P_2) cannot be assumed either. Finally (P_3) may be assumed in some cases: Let us assume that we know that $\|\alpha^\star\|_2 \leqslant 1$. Then $\|A\alpha^\star\|_2 \leqslant 1+\varepsilon$ with high probability, thus it is enough to consider the space $\mathcal{G}_P(\Theta)$ with parameter space $\Theta = \{\beta; \|\beta\|_2 \leqslant (1+\varepsilon)\}$, and thus $A\alpha^\star \in \Theta$ with high probability.

- Assumptions (G) are strong assumptions. The reason is that it is difficult to relate the vector coefficient $\beta^\star$ or even $A\alpha^\star$ to the vector coefficient $\alpha^\star$ of $f^\star = f_{\alpha^\star}$ in l_∞ norm. Thus even if we know some f_0 close to $f^\star$ in ℓ_∞-norm, this does not imply that we can build a function g_0 close to $g^\star = g_{\beta^\star}$.

- Assumptions (D) will not be valid a.s. w.r.t. the law of the Gaussian variables. The assumptions (D_1) and (D_4) are difficult to satisfy since they concern $\|.\|_\infty$. For assumption (D_2) and (D_3), we have the property that for each x, $\|\psi(x)\|_2^2$ is close to $\|\varphi(x)\|_2^2$ with high probability. However, we need here a uniform result over $x \in \mathcal{X}$ which seems difficult to get since the space $\mathcal{F}$ is actually big (not of finite dimension).

- Assumptions (O), which are typically strong assumptions for specific features φ appear to be almost satisfied. The reason is due to the covariance structure of the random features. Indeed whatever the distribution $\mathcal{P}_\mathcal{X}$ (independent of $\mathcal{P}_\mathcal{G}$), we have that $\langle \psi_p, \psi_q \rangle$ concentrates around

$$\mathbb{E}_{\mathcal{P}_\mathcal{G}} \langle \psi_p, \psi_q \rangle = \frac{1}{P} \| \sum_{i \geqslant 1} \varphi_i \|_{\mathcal{P}_\mathcal{X}}^2 \delta_{p,q} \,,$$

  where $\delta_{p,q}$ is the Kronecker symbol between p and q. Thus the orthogonality assumption is satisfied with high probability. Note that the knowledge of $\mathcal{P}_\mathcal{X}$ is still needed in order to rescale the features and obtain orthonormality. Similar argument shows that (O_2) is also valid.

As a consequence, only Theorems 10 and 15 would apply safely, but unfortunately these Theorems do not give results in high probability.

In the next two sections, we derive similar results but in high probability with assumptions that correspond to our setting. We provide a hand-made Theorem that makes use of the technique introduced in Györfi et al. (2002) and that can be applied without too restrictive assumptions, although not being optimal in terms of constant and logarithmic factors.

4.3 Approximation Power of Random Spaces

We assume from now on that we are in the case when $f^\star = f_{\alpha^\star} \in \mathcal{F}$.

Theorem 19 (Approximation error with deterministic design) *For all $P \geqslant 1$, for all $\delta \in (0,1)$ there exists an event of $\mathcal{P}_\mathcal{G}$-probability higher than $1-\delta$ such that on this event,*

$$\inf_{g \in \mathcal{G}_P} \|f^\star - g\|_N^2 \leqslant 12 \frac{\log(4N/\delta)}{P} \|\alpha^\star\|^2 \frac{1}{N} \sum_{n=1}^N \|\varphi(x_n)\|^2 \,.$$

Theorem 20 (Approximation error with random design) *Under assumption (N_2), then for all $P \geqslant 1$, for all $\delta \in (0,1)$, the following bound holds with $\mathcal{P}_\mathcal{G}$-probability higher than $1-\delta$:*

$$\inf_{g \in \mathcal{G}_P} \|f^\star - T_B(g)\|_{\mathcal{P}_\mathcal{X}}^2 \leqslant 25 \frac{\|\alpha^\star\|^2 \sup_x \|\varphi(x)\|^2}{P} \Big(1 + \frac{1}{2} \log \big(\frac{P \log(8P/\gamma^2\delta)}{18 \gamma^2 \delta} \big) \Big) \,,$$

where $\gamma \stackrel{\text{def}}{=} \frac{1}{B} \|\alpha^\star\| \sup_x \|\varphi(x)\|$ and T_B is the truncation operator at level B.

The result is not trivial because of the randomness of the space $\mathcal{G}_P$. Thus in order to keep the explanation simple, the proof (detailed in the Appendix) makes use of Hoeffding's Lemma only, which relies on the bounded assumption of the features (which can be seen either as a nice assumption, since it is simple and easy to check, or as a too strong assumption for some cases). Note that this result can be further refined by making use, for instance, of moment assumptions on the feature space instead.

4.4 Excess Risk of Random Spaces

In this section, we analyze the excess risk of the random projection method. Thus for a proposed random estimate $\widehat{g}$, we are interested in bounding $R(\widehat{g}) - R(f^\star)$ in high probability with respect to any source of randomness.

4.4.1 Regression Algorithm

From now on we consider the estimate $\widehat{g}$ to be the least-squares estimate $g_{\widehat{\beta}} \in \mathcal{G}_P$ that is the function in $\mathcal{G}_P$ with minimal empirical error, that is,

$$g_{\widehat{\beta}} \stackrel{\text{def}}{=} \arg\min_{g_\beta \in \mathcal{G}_P} R_N(g_\beta), \tag{4}$$

and that is the solution of a least-squares regression problem, that is, $\widehat{\beta} = \Psi^\dagger Y \in \mathbb{R}^P$ with matrix-wise notations, where $Y \in \mathbb{R}^N$ is here the vector of observations (not to be confused with the random variable Y that shares the same notation), Ψ is the $N \times P$-matrix composed of the elements: $\Psi_{n,p} \stackrel{\text{def}}{=} \psi_p(x_n)$, and $\Psi^\dagger$ is the Moore-Penrose pseudo-inverse[2] of Ψ. The final prediction function $\widehat{g}(x)$ is the truncation (at level $\pm B$) of $g_{\widehat{\beta}}$, that is, $\widehat{g}(x) \stackrel{\text{def}}{=} T_B[g_{\widehat{\beta}}(x)]$.

In the next subsection, we provide excess risk bounds w.r.t. $f^\star$ in $\mathcal{G}_P$.

4.4.2 Regression with Deterministic Design

Theorem 21 *Under assumption (N_1), then for all $P \geqslant 1$, for all $\delta \in (0,1)$ there exists an event of $\mathcal{P}_\mathcal{Y} \times \mathcal{P}_\mathcal{G}$-probability higher than $1-\delta$ such that on this event, the excess risk of the estimator $g_{\widehat{\beta}}$ is bounded as*

$$\|f^\star - g_{\widehat{\beta}}\|_N^2 \leqslant \frac{12\log(8N/\delta)}{P}\|\alpha^\star\|^2 \frac{1}{N}\sum_{n=1}^N \|\psi(x_n)\|^2 + \kappa B_1^2 \frac{P + \log(2/\delta)}{N},$$

for some numerical constant $\kappa > 0$.

Note that from this theorem, we deduce (without further assumptions on the features $\{\psi_i\}_{i\geqslant 1}$) that for instance for the choice $P = \frac{\sqrt{N}}{\log(N/\delta)}$ then

$$\|f^\star - g_{\widehat{\beta}}\|_N^2 \leqslant \kappa' \Big[\|\alpha^\star\|^2 \frac{1}{N}\sum_{n=1}^N \|\psi(x_n)\|^2 \sqrt{\frac{\log(N/\delta)}{N}} + \frac{\log(1/\delta)}{N} \Big],$$

for some positive constant κ'. Note also that whenever an upper-bound on the square terms $\|\alpha^\star\|^2 \frac{1}{N}\sum_{n=1}^N \|\psi(x_n)\|^2$ is known, this can be used in the definition of P in order to improve this bound.

2. In the full rank case when $N \geqslant P$, $\Psi^\dagger = (\Psi^T\Psi)^{-1}\Psi^T$.

4.4.3 REGRESSION WITH RANDOM DESIGN

In the regression problem with random design, the analysis of the excess risk of a given method is not straightforward, since the assumptions to apply standard techniques may not be satisfied without further knowledge on the structure of the features. In a general case, we can use the techniques introduced in Györfi et al. (2002), which yields to the following (not optimal) result:

Theorem 22 *Under assumption* (N_1) *and* (N_2)*, provided that* $N\log(N) \geqslant \frac{4}{P}$ *(thus whenever* $\min(N,P) \geqslant 2$*), then with* $\mathcal{P}_{\mathcal{G}} \times \mathcal{P}$*-probability at least* $1-\delta$*,*

$$R(T_B(g_{\widehat{\beta}})) - R(f^\star) \leqslant \kappa\Big[\frac{\log(12N/\delta)}{P}\|\alpha^\star\|^2 \sup_{x\in\mathcal{X}}\|\varphi(x)\|^2 + \max\{B_1^2, B^2\}\frac{P + P\log(N) + \log(3/\delta)}{N}\Big],$$

for some positive constant κ*.*

Let us now provide some intuition about the proof of this result. We first start by explaining what does not work. A natural idea in order to derive this result would be to consider the following decomposition:

$$R(T_B(g_{\widehat{\beta}})) - R(f^\star) \leqslant [R(T_B(g_B^\star)) - R(f^\star)] + [R(T_B(g_{\widehat{\beta}})) - R(T_B(g_B^\star))],$$

where $g_B^\star \in \operatorname*{argmin}_{g\in\mathcal{G}} R(T_B(g)) - R(f^\star)$.

Indeed the first term is controlled on an event $\Omega_{\mathcal{G}}$ of high $\mathcal{P}_{\mathcal{G}}$-probability by Theorem 20, and since $R(g_{\widehat{\beta}}) - R(g_B^\star) \leqslant R(g_{\widehat{\beta}}) - R(g^\star)$, the second term is controlled for each fixed $\omega_{\mathcal{G}} \in \Omega_{\mathcal{G}}$ with high $\mathcal{P}$-probability by standard Theorems for regression, provided that we can relate $R(T_B(g_{\widehat{\beta}})) - R(T_B(g_B^\star))$ to $R(g_{\widehat{\beta}}) - R(g_B^\star)$. Thus by doing the same careful analysis of the events involved, this should lead to the desired result.

However, the difficulty lies first in ensuring that the conditions of application of standard Theorems are satisfied with high $\mathcal{P}_{\mathcal{G}}$-probability and then in relating the excess risk of the truncated function to that of the non-truncated ones, since it is not true in general that $R(T_B(g_{\widehat{\beta}})) - R(T_B(g_B^\star)) \leqslant R(g_{\widehat{\beta}}) - R(g_B^\star)$. Thus we resort to a different decomposition in order to derive our results. The sketch of proof of Theorem 22 actually consists in applying the following lemma.

Lemma 23 *The following decomposition holds for all* $C > 0$

$$\|T_B(g_{\widehat{\beta}}) - f^\star\|_{\mathcal{P}_X}^2 \leqslant C\|f^\star - g_{\tilde{\beta}}\|_N^2 + C\|g_{\tilde{\beta}} - g_{\widehat{\beta}}\|_N^2 + \sup_{g\in\mathcal{G}}\left(\|f^\star - T_B(g)\|_{\mathcal{P}_X}^2 - C\|f^\star - T_B(g)\|_N^2\right),$$

where $g_{\tilde{\beta}} = \Pi_{\|.\|_N}(f^\star, \mathcal{G})$ *and* $g_{\widehat{\beta}} = \Pi_{\|.\|_N}(Y, \mathcal{G})$ *are the projections of the target function* $f^\star$ *and observation* Y *onto the random linear space* $\mathcal{G}$ *with respect to the empirical norm* $\|.\|_N$*.*

We then call the first term $\|f^\star - g_{\tilde{\beta}}\|_N^2$ an approximation error term, the second $\|g_{\tilde{\beta}} - g_{\widehat{\beta}}\|_N^2$ a noise error term and the third one $\sup_{g\in\mathcal{G}}\left(\|f^\star - T_B(g)\|_{\mathcal{P}_X}^2 - C\|f^\star - T_B(g)\|_N^2\right)$ an estimation of the error term.

In order to prove Theorem 22, we then control each of these terms: We apply Lemma 19 to the first term, Lemma 24 below to the second term and finally Theorem 11.2 of Györfi et al. (2002) to the last term with $C = 8$, and the result follows by gathering all the bounds.

Let us now explain the contribution to each of the three terms in details.

Approximation error term The first term, $\|f^\star - g_{\tilde{\beta}}\|_N^2$, is an approximation error term in empirical norm, it contains the number of projections as well as the norm of the target function. This term plays the role of the approximation term that exists for regression with penalization by a factor $\lambda\|f\|^2$. This term is controlled by application of Theorem 19 conditionally on the random samples, and then w.r.t. all source of randomness by independence of the Gaussian random variables with the random samples.

Noise error term The second term, $\|g_{\tilde{\beta}} - g_{\hat{\beta}}\|_N^2$, is an error term due to the observation noise η. This term classically decreases at speed $\frac{D\sigma^2}{N}$ where σ^2 is the variance of the noise and D is related to the log entropy of the space of function $\mathcal{G}$ considered. Without any more assumption, we only know that this is a linear space of dimension P, so this term finally behaves like $\frac{P\sigma^2}{N}$, but note that this dependency with P may be improved depending on the knowledge about the functions ψ (for instance, if $\mathcal{G}$ is included in a Sobolev space of order s, we would have $P^{1/2s}$ instead of P).

Lemma 24 *Under assumption (N_1), then for each realization of the Gaussian variables, with $\mathcal{P}$-probability higher than $1-\delta$, the following holds true:*

$$\|g_{\tilde{\beta}} - g_{\hat{\beta}}\|_N^2 \leqslant 6B_1^2 \frac{1616P + 200\log(6/\delta) + \log(3/\delta)}{N}.$$

Note that we may consider different assumptions on the noise term. Here we considered only that the noise is upper-bounded as $\|\eta\|_\infty \leqslant B_1$, but another possible assumption is that the noise has finite variance σ^2 or that the tail of the distribution of the noise behaves nicely, for example, that $\|\eta\|_{\psi_\alpha} \leqslant B$, where ψ_α is the Orlicz norm or order α, with $\alpha = 1$ or 2.

Estimation error term The third term, $\sup_{g\in\mathcal{G}_P}(\|f^\star - T_B(g)\|_{\mathcal{P}_X}^2 - C\|f^\star - T_B(g)\|_N^2)$, is an estimation of the error term due to finiteness of the data. This term also depends on the log entropy of the space of functions, thus the same remark applies to the dependency with P as for the noise error term. We bound the third term by applying Theorem 11.2 of Györfi et al. (2002) to the class of functions $\mathcal{G}^0 = \{f^\star - T_B(g), g \in \mathcal{G}_P\}$, for fixed random Gaussian variables. Note that for all $f \in \mathcal{G}^0$, $\|f\|_\infty \leqslant 2B$. The precise result of Györfi et al. (2002) is the following :

Theorem 25 *Let $\mathcal{F}$ be a class of functions $f : \mathbb{R}^d \to \mathbb{R}$ bounded in absolute value by B. Let $\varepsilon > 0$. Then*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \|f\|_{\mathcal{P}_X} - 2\|f\|_N > \varepsilon\right) \leqslant 3\mathbb{E}\left(\mathcal{N}\left(\frac{\sqrt{2}}{24}\varepsilon, \mathcal{F}, \|.\|_{2N}\right)\right)\exp\left(-\frac{N\varepsilon^2}{288B^2}\right).$$

We now have the following lemma whose proof is given in the Appendix:

Lemma 26 *Assuming that $N\log(N) \geqslant \frac{4}{P}$, then for each realization of the Gaussian variables, with $\mathcal{P}$-probability higher than $1-\delta$, the following holds true:*

$$\sup_{g\in\mathcal{G}_P} \|f^\star - T_B(g)\|_{\mathcal{P}_X}^2 - 8\|f^\star - T_B(g)\|_N^2 \leqslant (24B)^2 \frac{4\log(3/\delta) + 2P\log(N)}{N}.$$

5. Discussion

In this section, we now provide more insights about the main results of this paper by reminding some closely related existing works, showing some numerical illustration of the method and discussing some numerical issues.

5.1 Non-linear Approximation

In the work of Barron et al. (2008), the authors provide excess risk bounds for greedy algorithms (i.e., in a non-linear approximation setting). The precise result they derive in their Theorem 3.1 is reported now, using the notations of section 4.2:

Theorem 27 (Barron et al. 2008) *Consider spaces $\{\mathcal{G}_P\}_{P\geqslant 1}$ generated respectively by the span of features $\{e_p\}_{p\leqslant P}$ with increasing dimension P (thus $\Theta=\mathbb{R}^P$ for each P). For each $\mathcal{G}_P$ we compute a corresponding greedy empirical estimate $\widehat{g}_P\in\mathcal{G}_P$ provided by some algorithm (see Barron et al., 2008), then we define $\widehat{P}=\operatorname{argmin}\|y-T_{B_1}\widehat{f}_P\|_N^2+\kappa\frac{P\log(N)}{N}$ for some constant κ, and finally define $\widehat{g}=T_{B_1}(\widehat{g}_{\widehat{P}})$, and fix some P_0.*

Under assumption (N_1), there exists κ_0 depending only on B_1 and a where $P_0=\llcorner N^a\lrcorner$ such that if $\kappa\geqslant\kappa_0$, then for all $P>0$ and for all functions g_θ in $\mathcal{G}_{P_0}$, the estimator $\widehat{g}$ satisfies

$$\mathbb{E}R(\widehat{g})-R(f^{(reg)})\leqslant 2[R(g_\theta)-R(f^{(reg)})]+8\frac{\|\theta\|_1^2}{P}+C\frac{P\log N}{N},$$

where the constant C only depends on κ, B_1 and a.

The bound is thus similar to that of Theorem 22 in Section 4.4. One difference is that this bound contains the l_1 norm of the coefficients θ^* while the ℓ_2 norm of the coefficients $\alpha^\star$ appears in our setting. We leave as an open question to understand whether this difference is a consequence of the non-linear aspect of their approximation or if it results from the different assumptions made about the approximation spaces, in terms of rate of decrease of the coefficients.

The main difference is actually about the tractability of the proposed estimator, since the result of Theorem 27 relies on greedy estimation that is computationally heavy while on the other hand, random projection is cheap (see Subsection 5.4).

5.2 Adaptivity

Randomization enables to define approximation spaces such that the approximation error, either in expectation or in high probability on the choice of the random space, is controlled, whatever the measure $\mathcal{P}$ that is used to assess the performance. This is specially interesting in the regression setting where $\mathcal{P}$ is unknown. As mentioned in the introduction, because the choice of the subspace $\mathcal{G}_P$ within which we perform the least-squares estimate is *random*, we avoid (with high probability) degenerated situations where the target function $f^\star$ cannot be well approximated with functions in $\mathcal{G}_P$. Indeed, in methods that consider a given *deterministic* finite-dimensional subspace $\mathcal{G}$ of the big space $\mathcal{F}$ (such as linear approximation using a predefined set of wavelets), it is often possible to find a target function $f^\star$ such that $\inf_{g\in\mathcal{G}_P}\|f^\star-g\|_N$ is large, whereas using the random projection method, the random choice of $\mathcal{G}_P$ implies that for any $f^\star\in\mathcal{F}$, the approximation error $\inf_{g\in\mathcal{G}_P}\|f^\star-g\|_N$ can be controlled (by the first term of the bound (2)) in high probability. We now illustrate this property on a simple example.

Example Let us consider a very peaky (a spot) distribution $\mathcal{P}$. Regular linear approximation, say with wavelets (see, e.g., DeVore, 1997), will most probably miss the specific characteristics of $f^\star$ at the spot, since the first wavelets have large support. On the contrary, the random features $\{\psi_p\}_{p\leqslant P}$ that are functions that contain (random combinations of) all wavelets, will be able to detect correlations between the data and some high frequency wavelets, and thus discover relevant features of $f^\star$ at the spot. This is illustrated in the numerical experiment below.

Here $\mathcal{P}$ is a very peaky Gaussian distribution and $f^\star$ is a 1-dimensional periodic function. We consider as initial features $\{\varphi_i\}_{i\geqslant 1}$ the set of hat functions defined in Section 3.2.2. Figure 3 shows the target function $f^\star$, the distribution $\mathcal{P}$, and the data $(x_n,y_n)_{1\leqslant n\leqslant 100}$ (left plots). The middle plots represents the least-squares estimate $\widehat{g}$ using $P=40$ scrambled objects $\{\psi_p\}_{1\leqslant p\leqslant 40}$ that are here Brownian motions. The right plots shows the least-squares estimate using the initial features $\{\varphi_i\}_{1\leqslant i\leqslant 40}$. The top figures represent a high level view of the whole domain $[0,1]$. No method is able to learn $f^\star$ on the whole space (this is normal since the available data are only generated from a peaky distribution). The bottom figures shows a zoom $[0.45,0.51]$ around the data. Least-squares regression using scrambled objects is able to learn the structure of $f^\star$ in terms of the measure $\mathcal{P}$, while least-squares regression with the initial features completely fails.

5.3 Other Related Work

In Rahimi and Recht (2008, 2007), the authors consider, for a given parameterized function $\Phi : \mathcal{X}\times\Theta\to\mathbb{R}$ bounded by 1, and a probability measure μ over Θ, the space $\mathcal{F}$ of functions $f(x)=\int_\Theta \alpha(\theta)\Phi(x,\theta)d\theta$ such that $\|f\|_\mu = \sup_\theta|\frac{\alpha(\theta)}{\mu(\theta)}|<\infty$. They show that this is a dense subset of the RKHS with kernel $k(x,y)=\int_\Theta\mu(\theta)\Phi(x,\theta)\Phi(y,\theta)d\theta$, and that if $f\in\mathcal{F}$, then with high probability over $\{\theta_p\}_{p\leqslant P}\overset{i.i.d}{\sim}\mu$, there exist coefficients $\{c_p\}_{p\leqslant P}$ such that $\widehat{f}(x)=\sum_{p=1}^P c_p\Phi(x,\theta_p)$ satisfies $\|\widehat{f}-f\|_2^2\leqslant O(\frac{\|f\|_\mu}{\sqrt{P}})$. The method is analogous to the construction of the empirical estimates $g_{A\alpha}\in\mathcal{G}_P$ of function $f_\alpha\in\mathcal{K}$ in our setting. Indeed we may formally identify $\Phi(x,\theta_p)$ with $\psi_p(x)=\sum_i A_{p,i}\varphi_i(x)$, θ_p with the sequence $(A_{p,i})_i$, and the distribution μ with the distribution of this infinite sequence. However, in our setting we do not require the condition $\sup_{x,\theta}\Phi(x,\theta)\leqslant 1$ to hold and the fact that Θ is a set of infinite sequences makes the identification tedious without the Gaussian random functions theory used here. Anyway, we believe that this link provides a better mutual understanding of both approaches (i.e., Rahimi and Recht 2008 and this paper).

5.4 Tractability

In practice, in order to build the least-squares estimate, one needs to compute the values of the random features $\{\psi_p\}_{1\leqslant p\leqslant P}$ at the data points $\{x_n\}_{1\leqslant n\leqslant N}$, that is, the matrix $\Psi=(\psi_p(x_n))_{p\leqslant P,n\leqslant N}$. Moreover, due to finite memory and precision of computers, numerical implementations can only handle a finite number F of initial features $\{\varphi_i\}_{1\leqslant i\leqslant F}$.

Approximation error Using a finite F introduces an additional approximation (squared) error term in the final excess risk bounds. This additional error that is due to the numerical approximation is of order $O(F^{-\frac{2s}{d}})$ for a wavelet basis adapted to $H^s([0,1]^d)$ and can be made arbitrarily small, for example, $o(N^{-1/2})$, whenever the depth of the wavelet dyadic-tree is bigger than $\frac{\log N}{d}$. Our main concern is thus about efficient computation.

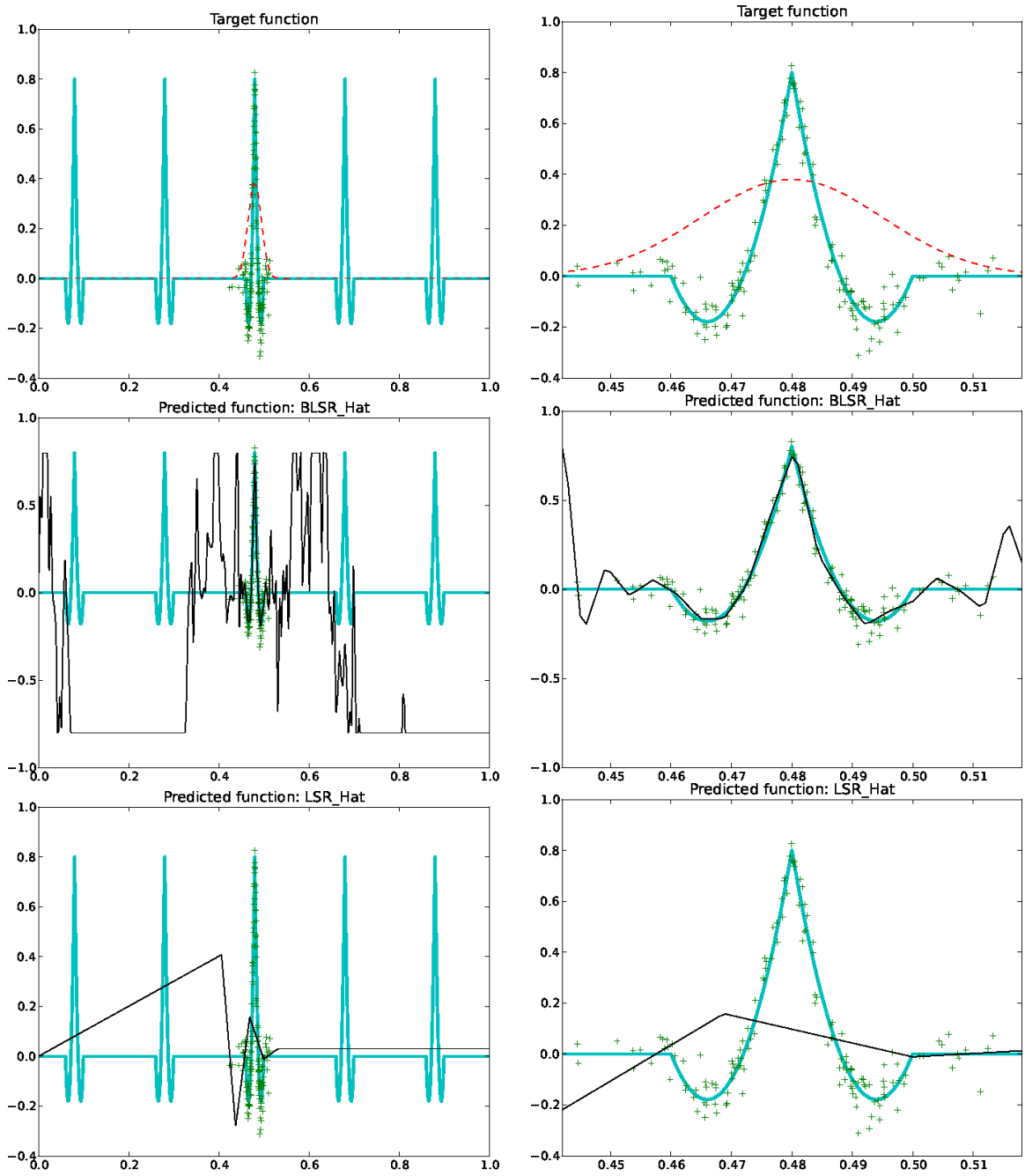

Figure 3: Least squares estimates of $f^\star$, using $N = 100$ data generated from a peaky distribution $\mathcal{P}$ (dashed line in top plots), using 40 Brownian motions $\{\psi_p\}$ (middle plots) and using 40 hat functions $\{\varphi_i\}$ (bottom plots). The target function $f^\star$ is plotted with thick line while the two estimates are plotted with thin line. The right column shows a zoom around the data.

Numerical complexity In Maillard and Munos (2009) it was mentioned that the computation of Ψ, which makes use of the random matrix $A = (A_{p,i})_{p \leqslant P, i \leqslant F}$, has a complexity $O(FPN)$.

In the multi-resolution schemes described now, provided that the mother function has compact support (such as the hat functions), we can significantly speed up the computation of the matrix Ψ by resorting to a *tree-based lazy expansion*, that is, where the expansion of the random features $\{\psi_p\}_{p \leqslant P}$ is built only when needed for the evaluation at the points $\{x_n\}_n$. Note that in the specific case of wavelets, we can even think to combine random projection with tools like fast wavelet transform which would be even faster (which we do not do here for simplicity and generality purpose).

Example: Consider the example of the scrambled wavelets. In dimension 1, using a wavelet dyadic-tree of depth H (i.e., $F = 2^{H+1}$), the numerical cost for computing Ψ is $O(HPN)$ (using one tree per random feature). Now, in dimension d the classical extension of one-dimensional wavelets uses a family of $2^d - 1$ wavelets, thus requires $2^d - 1$ trees each one having 2^{dH} nodes. While the resulting number of initial features F is of order $2^{d(H+1)}$, thanks to the lazy evaluation (notice that one never computes all the initial features), one needs to expand at most one path of length H per training point, and the resulting complexity to compute Ψ is $O(2^d HPN)$. Thus the method is linear with N and reduces the amount of computation by an exponential factor (from 2^{dH} to $2^d H$).

Note that one may alternatively use the so-called sparse-grids instead of wavelet trees, which have been introduced by Griebel and Zenger (see Zenger, 1990; Bungartz and Griebel, 2004). The main result is that one can reduce significantly the total number of features to $F = O(2^H H^d)$ (while preserving a good approximation for sufficiently smooth functions). Similar lazy evaluation techniques can be applied to sparse-grids.

Thus, using $P = O(\sqrt{N})$ random features, we deduce that the complexity of building the matrix Ψ is at most $O(2^d N^{3/2} \log N)$. Then in order to solve the least squares system, one has to compute $\Psi^T \Psi$, that has cost at most $O(P^2 N)$, and then solve the system by inversion, which has numerical cost $O(P^{2.376})$ by Coppersmith and Winograd (1987). Thus, with $P = O(\sqrt{N})$, the overall cost of the algorithm is at most $O(2^d N^{3/2} \log N + N^2)$, without using any fancy computations designed for random matrices, and the numerical complexity to make a new prediction is at most $O(2^d N^{1/2} \log(N))$.

Acknowledgments

The authors want to thank *Pierre Chainais* and *Olivier Degris* for interesting pointers to the literature in image processing and applied reinforcement learning.
This work was supported by the French National Research Agency through the EXPLO-RA project No ANR-08-COSI-004 and the CompLACS project by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No 270327.

Appendix A. Technical Details

In this technical section we gathered the proofs of the important Lemmas and of the main Theorems 19, 20, 21 and 22.

A.1 Proof of Lemma 9

Proof *Step 1.* First, we derive a result similar to Lemma 7 that holds for dot products, by polarisation of the Euclidean norm. The precise statement for our purpose is the following one.

Lemma 28 *Let A be a $P \times F$ matrix of i.i.d. elements drawn from one of the previously defined distributions. Let $(u_n)_{1\leqslant n\leqslant N}$ and v be $N+1$ vectors of $\mathbb{R}^F$.*

Then for any $\varepsilon \in (0,1)$, with probability at least $1-4Ne^{-P(\varepsilon^2/4-\varepsilon^3/6)}$, simultaneoulsy for all $n \leqslant N$,

$$|Au_n \cdot Av - u_n \cdot v| \leqslant \varepsilon \|u_n\| \, \|v\| \, .$$

We apply Lemma 7 to any couple of vectors $u+w$ and $u-w$, where u and w are vectors of norm 1. By polarisation, we have that

$$\begin{aligned} 4Au \cdot Aw &= \|Au+Aw\|^2 - \|Au-Aw\|^2 \\ &\leqslant (1+\varepsilon)\|u+w\|^2 - (1-\varepsilon)\|u-w\|^2 \\ &= 4u \cdot w + \varepsilon(\|u+w\|^2 + \|u-w\|^2) \\ &= 4u \cdot w + 2\varepsilon(\|u\|^2 + \|w\|^2) = 4u \cdot w + 4\varepsilon \, , \end{aligned}$$

fails with probability $2e^{-P(\varepsilon^2/4-\varepsilon^3/6)}$ (we applied the previous lemma twice at line 2).

Thus for each $n \leqslant N$, we have with same probability:

$$Au_n \cdot Av \leqslant u_n \cdot v + \varepsilon \|u_n\| \, \|v\| .$$

Now the symmetric inequality holds with the same probability, and using a union bound for considering all $(u_n)_{n\leqslant N}$, we have that

$$|Au_n \cdot Av - u_n \cdot v| \leqslant \varepsilon \|u_n\| \, \|v\| ,$$

holds for all $n \leqslant N$, with probability $1-4Ne^{-P(\varepsilon^2/4-\varepsilon^3/6)}$.

Step 2. We now extend this Lemma to the case of infinite sequences. This is made possible thanks to the measurability properties of Gaussian Objects. Indeed, for any given F, Lemma 28 applies to the two truncated sequences $\overline{\alpha}_F = (\alpha_1, \ldots, \alpha_F)$ and $\overline{\varphi}_F(x_n) = (\varphi_1(x_n), \ldots, \varphi_F(x_n))$; this gives that for all n simultaneoulsy,

$$|\sum_{i=1}^{F} \alpha_i \varphi_i(x_n) - \frac{1}{P} \sum_{p=1}^{P} \big(\sum_{i=1}^{F} \xi_{i,p} \alpha_i\big) \big(\sum_{i=1}^{F} \xi_{i,p} \varphi_i(x_n)\big)| \leqslant \varepsilon \, \|\overline{\alpha}_F\| \, \|\overline{\varphi}_F(x_n)\| \, ,$$

happens with probability higher than $1-4Ne^{-P(\varepsilon^2/4-\varepsilon^3/6)}$, where we introduced $\xi_{i,p} \stackrel{\text{def}}{=} \sqrt{P} A_{i,p} \sim \mathcal{N}(0,1)$ in order to avoid confusion with the section on Gaussian objects. Now by the assumption that $\alpha \in \ell_2(\mathbb{R})$ and $\varphi(x) \in \ell_2(\mathbb{R})$ for all x, then the Gaussian objects $\sum_{i=1}^{\infty} \xi_{i,p} \alpha_i$ and $\sum_{i=1}^{\infty} \xi_{i,p} \varphi_i(x_n)$ are well-defined square integrable random variables. Thus, taking the limit of the above inequality when F tends to ∞ yields that with same probability, for all $n \leqslant N$

$$|f_\alpha(x_n) - g_{A\alpha}(x_n)| \leqslant \varepsilon \, \|\alpha\| \, \|\varphi(x_n)\| \, .$$

■

A.2 Proof of Lemma 24

Proof We can bound the noise term $\|g_{\tilde{\beta}} - g_{\widehat{\beta}}\|_N^2$ using a simple Chernoff bound together with a chaining argument. Indeed, by definition of $g_{\tilde{\beta}}$ and $g_{\widehat{\beta}}$, if we introduce the noise vector η defined by $\eta = Y - f$, we have

$$\begin{aligned}
\|g_{\tilde{\beta}} - g_{\widehat{\beta}}\|_N^2 &= \langle g_{\tilde{\beta}} - g_{\widehat{\beta}}, \eta \rangle_N \\
&= \frac{1}{N}\sum_{i=1}^N \eta_i (g_{\tilde{\beta}} - g_{\widehat{\beta}})(X_i) \\
&\leqslant \Big(\sup_{g \in \mathcal{G}} \frac{\frac{1}{N}\sum_{i=1}^N \eta_i g(X_i)}{\|g\|_N} \Big) \|g_{\tilde{\beta}} - g_{\widehat{\beta}}\|_N \\
&\leqslant \Big(\sup_{g \in \mathcal{G}} \frac{\frac{1}{N}\sum_{i=1}^N \eta_i g(X_i)}{\|g\|_N} \Big)^2 .
\end{aligned}$$

Thus, we focus on the set $\mathcal{G}^1 = \{g \in \mathcal{G}; \|g\|_N = 1\}$. Note that since $\mathcal{G}^1$ is a sphere in a space of dimension P, its ε-packing number in empirical norm is bounded above by $\mathcal{M}(\varepsilon, \mathcal{G}^1, \|.\|_N) \leqslant \mathcal{N}(\varepsilon/2, \mathcal{G}^1, \|.\|_N) \leqslant \mathcal{N}(\varepsilon/2, \{g \in \mathcal{G}; \|g\|_N \leqslant 1\}, \|.\|_N) \leqslant (\frac{4}{\varepsilon} + 1)^P \leqslant \max(\frac{5}{\varepsilon}, 5)^P$, where $\mathcal{N}$ refers here to the covering number.

We now introduce for convenience the following notation, for fixed Gaussian random variables and data points $(X_i)_{i=1..n}$:

$$\rho(t) \stackrel{\text{def}}{=} \mathbb{P}_Y\Big(\exists g \in \mathcal{G} \frac{\frac{1}{N}\sum_{i=1}^N \eta_i g(X_i)}{\|g\|_N} > t\Big) \quad = \quad \mathbb{P}_Y\Big(\exists g \in \mathcal{G}^1 \frac{1}{N}\sum_{i=1}^N \eta_i g(X_i) > t\Big).$$

For $j = 0 \ldots \infty$, let us consider ε_j-packings C_j of $\mathcal{G}^1$ for the empirical norm $\|.\|_N$, with $C_0 = g_0$, such that C_{j+1} is a refinement of C_j and $\varepsilon_j \leqslant \varepsilon_{j-1}$. Then for a given $g \in \mathcal{G}^1$, we define $g_j = \Pi(g, C_j)$ the projection of g into C_j, for the norm $\|g\|_N$. Thus, $g - g_0 = (g - g_J) + \sum_{j=1}^J (g_j - g_{j-1})$. Note that since by definition of $\mathcal{G}^1$ we have $\|g - g_0\|_N \leqslant 2$, we need to consider $\varepsilon_0 \geqslant 2$.

Thus if we now introduce real numbers γ and $(\gamma_j)_{j \geqslant 1}$ such that $\sum_{j=1}^J \gamma_j \leqslant \gamma$, then we have

$$\begin{aligned}
\rho(\gamma t_1 + t_2 + t_3) &\leqslant \mathbb{P}\Big(\exists g \in \mathcal{G}^1 \frac{1}{N}\sum_{i=1}^N \eta_i (g - g_0)(X_i) > \gamma t_1 + t_2\Big) + \exp(-\frac{t_3^2 N}{2B_1^2}) \\
&\leqslant \mathbb{P}\Big(g \in \exists \mathcal{G}^1 \frac{1}{N}\sum_{i=1}^N \eta_i (g - g_J)(X_i) + \\
&\qquad \sum_{j=1}^J \frac{1}{N}\sum_{i=1}^N \eta_i (g_j - g_{j-1})(X_i) \geqslant \sum_{j=1}^J \gamma_j t_1 + t_2\Big) + \exp(-\frac{t_3^2 N}{2B_1^2}),
\end{aligned}$$

where we applied Hoeffding's inequality in the first line. We further have:

$$\begin{aligned}\rho(\gamma t_1+t_2+t_3) &\leqslant \sum_{j=1}^{J}\mathbb{P}\Big(\exists g\in\mathcal{G}^1\frac{1}{N}\sum_{i=1}^{N}\eta_i(g_j-g_{j-1})(X_i)>t_1\gamma_j\Big)\\ &\quad+\exp(-\frac{t_2^2N}{2B_1^2\varepsilon_J^2})+\exp(-t_3^2N2B_1^2)\\ &\leqslant \mathbb{E}\sum_{j=1}^{J}\mathcal{M}_j\mathcal{M}_{j-1}\mathbb{P}\Big(\frac{1}{N}\sum_{i=1}^{N}\eta_i(g_j-g_{j-1})(X_i)>t_1\gamma_j\Big)\\ &\quad+\exp(-\frac{t_2^2N}{2B_1^2\varepsilon_J^2})+\exp(-\frac{t_3^2N}{2B_1^2}),\end{aligned}$$

where we introduced for convenience the notation $\mathcal{M}_j \stackrel{\text{def}}{=} \mathcal{M}(\varepsilon_j,\mathcal{G}^1,\|.\|_N)$. Now, note that since $\varepsilon_j\leqslant\varepsilon_{j-1}$, then $\mathcal{M}_{j-1}\leqslant\mathcal{M}_j$. Note also that $\|g_j-g_{j-1}\|_N\leqslant\eta_j$ since C_j is a refinement of C_{j-1}. Finally, we can bound the packing number by $\mathcal{M}_j\leqslant N_j=\max(\frac{5}{\varepsilon_j},5)^P$ where P is the dimension of $\mathcal{G}$. Thus we deduce that:

$$\rho(\gamma t_1+t_2+t_3) \leqslant \sum_{j=1}^{J}N_j^2\exp(-\frac{t_1^2N\gamma_j^2}{2B_1^2\varepsilon_j^2})+\exp(-\frac{t^2N}{2B_1^2\varepsilon_J^2})+\exp(-\frac{t_3^2N}{2B_1^2}).$$

Now, we define $\gamma_j=\frac{2\varepsilon_jB_1}{t_1}\sqrt{\frac{2\log(N_j)}{N}}$, $t_2=B_1\varepsilon_J\sqrt{\frac{2\log(1/\delta_2)}{N}}$ and $t_3=B_1\sqrt{\frac{2\log(1/\delta_3)}{N}}$, for some $\delta_2,\delta_3\in(0,1]$. Thus, we get:

$$\rho(\eta t_1+t_2+t_3) \leqslant \sum_{j=1}^{J}\frac{1}{N_j^2}+\delta_2+\delta_3.$$

Thus, it remains to define ε_j. Since $N_j=\max(\frac{5}{\varepsilon_j},5)^P$, we define the covering radius ε_j to be $\varepsilon_j=2^{-j}5\delta_1^{1/2P}(2^{2P}-1)^{1/2P}$ for some $\delta_1\in(0,1]$, which entails that $\sum_{j=1}^{J}\frac{1}{N_j^2}\leqslant\delta_1$. Now since $\varepsilon_j\to0$ when $j\to\infty$, we can make the sum goes to infinity. We deduce that:

$$\rho(\gamma t_1+B_1\sqrt{\frac{2\log(1/\delta_3)}{N}})\leqslant\delta_1+\delta_2+\delta_3.$$

Now, in order to bound the term $\gamma t_1 + t_2 + t_3$,we look at the following term:

$$\begin{aligned}
\gamma t_1 &= 2\sum_{j=1}^{\infty} \varepsilon_j B_1 \sqrt{\frac{2\log(N_j)}{N}} \\
&\leqslant \frac{20B_1}{\sqrt{N}} \sum_{j=1}^{\infty} 2^{-j} \sqrt{2jP\log(2) + \log(1/\delta_1) - \log(2^{2P}-1)} \\
&\leqslant \frac{20B_1}{\sqrt{N}} \sum_{j=1}^{\infty} 2^{-j} \sqrt{2(j-1)P\log(2) + \log(2/\delta_1)} \\
&\leqslant \frac{20B_1}{\sqrt{N}} \Big(\sum_{j=1}^{\infty} 2^{-j} \sqrt{2(j-1)P\log(2)} + \sqrt{\log(2/\delta_1)} \Big) \\
&\leqslant \frac{20B_1}{\sqrt{N}} \Big((1+\sqrt{2}) \sqrt{2P\log(2)} + \sqrt{\log(2/\delta_1)} \Big).
\end{aligned}$$

where we use the fact that $\sum_{j=1}^{\infty} 2^{-j} \leqslant 1$, and that $\sum_{j=1}^{\infty} 2^{-j}\sqrt{(j-1)} \leqslant 1+\sqrt{2}$.

Using the inequalities $\sqrt{a}+\sqrt{b}+\sqrt{c} \leqslant \sqrt{3(a+b+c)}$, we thus deduce the following bound:

$$\begin{aligned}
\gamma t_1 + t_2 + t_3 &\leqslant \frac{B_1}{\sqrt{N}} \Big(20(1+\sqrt{2})\sqrt{2P\log(2)} + 20\sqrt{\log(2/\delta_1)} + \sqrt{2\log(1/\delta_3)} \Big) \\
&\leqslant \frac{\sqrt{6}B_1}{\sqrt{N}} \sqrt{400\log(2)(1+\sqrt{2})^2 P + 200\log(2/\delta_1) + \log(1/\delta_3)}.
\end{aligned}$$

Thus, by setting $\delta_1 = \delta_2 = \delta_3 = \delta/3$, we deduce that with $\mathcal{P}$-probability higher than $1-\delta$,

$$\sup_{g \in \mathcal{G}_P} \frac{\frac{1}{N}\sum_{i=1}^N \varepsilon_i g(X_i)}{\|g\|_N} \leqslant \frac{B_1\sqrt{6}}{\sqrt{N}} \sqrt{400\log(2)(1+\sqrt{2})^2 P + 200\log(6/\delta) + \log(3/\delta)}\,.$$

■

A.3 Proof of Lemma 26

Proof Indeed, let us introduce the space of functions $\mathcal{G}^0 = \{f^\star - T_B(g), g \in \mathcal{G}_P\}$. Then we have for $g \in \mathcal{G}^0$, $\|g\|_N \leqslant \|g\|_\infty \leqslant 2B$. Thus Theorem 11.2 of Györfi et al. (2002) gives the following bound:

$$\mathbb{P}\big(\sup_{g\in\mathcal{G}_P} \|f^\star - T_B(g)\|_{\mathcal{P}_X} - 2\|f^\star - T_B(g)\|_N > \varepsilon\big) \leqslant 3\mathbb{E}\big(\mathcal{N}\big(\frac{\sqrt{2}}{24}\varepsilon, \mathcal{G}^0, \|.\|_{2N}\big)\big) \exp\Big(-\frac{N\varepsilon^2}{288(2B)^2}\Big).$$

Then, since $\mathcal{G}^0 = f^\star + T_B(\mathcal{G}_P)$, we bound the entropy number by:

$$\mathcal{N}\big(\frac{\sqrt{2}}{24}\varepsilon, \mathcal{G}^0, \|.\|_{2N}\big) \leqslant \mathcal{N}\big(\frac{\sqrt{2}}{24}\varepsilon, T_B(\mathcal{G}_P), \|.\|_{2N}\big) \leqslant \big(\frac{2(2B).24}{\sqrt{2}\varepsilon} + 1\big)^P.$$

Thus we deduce that if $\varepsilon \geqslant \frac{24.4B}{\sqrt{2}}u$, then with probability higher than $1-\delta$ w.r.t $\mathbb{P}$, for fixed random Gaussian variables,

$$\sup_{g\in\mathcal{G}_P} \|f^\star - T_B(g)\|_{\mathcal{P}_X} - 2\|f^\star - T_B(g)\|_N \leqslant \varepsilon = 24B\sqrt{\log(3/\delta) + P\log\big(\frac{1}{u}+1\big)}\sqrt{\frac{2}{N}}.$$

Thus,we consider $u = \frac{1}{N-1}$, and deduce that, provided that $N\log(N) \geqslant \frac{4}{P}$, then with probability higher than $1-\delta$ w.r.t $\mathbb{P}$, for fixed random Gaussian variables (i.e., conditionally on them),

$$\sup_{g\in\mathcal{G}_P} \|f^\star - T_B(g)\|_{\mathcal{P}_X} - 2\|f^\star - T_B(g)\|_N \leqslant 24B\sqrt{\frac{2\log(3/\delta)+P\log(N)}{N}}.$$

Thus, we deduce that on this event, for all $g \in \mathcal{G}_P$

$$\begin{aligned}\|f^\star - T_B(g)\|^2_{\mathcal{P}_X} &\leqslant (2\|f^\star - T_B(g)\|_N + 24B\sqrt{\frac{2\log(3/\delta)+P\log(N)}{N}})^2 \\ &\leqslant 8\|f^\star - T_B(g)\|^2_N + (24B)^2\frac{4\log(3/\delta)+2P\log(N)}{N}.\end{aligned}$$

This gives the following upper bound, that holds with probability higher than $1-\delta$:

$$\sup_{g\in\mathcal{G}_P} \|f^\star - T_B(g)\|^2_{\mathcal{P}_X} - 8\|f^\star - T_B(g)\|^2_N \leqslant (24B)^2\frac{4\log(3/\delta)+2P\log(N)}{N}.$$

■

A.4 Proof of Theorem 19

Proof Since by assumption $f^\star = f_{\alpha^\star}$ for some $\alpha^\star$, we have by direct application of Lemma 9

$$\inf_{g\in\mathcal{G}} \|f^\star - g\|^2_N \leqslant \|f_{\alpha^\star} - g_{A\alpha^\star}\|^2_N.$$

Now let us define for some $N \geqslant 1$ the quantity $\varepsilon = \varepsilon_N(\delta)$ that appears in Lemma 9, such that

$$\frac{\log(4N/\delta)}{P} = \frac{\varepsilon^2}{4} - \frac{\varepsilon^3}{6}.$$

Thus, since $\varepsilon \in (0,1)$, this means in particular that we have

$$\frac{\varepsilon^2}{3} \leqslant 4\frac{\log(4N/\delta)}{P} \leqslant \varepsilon^2.$$

■

A.5 Proof of Theorem 20

Proof By assumption, we consider that $f^\star \in \mathcal{F}$. Thus there exists a sequence $\alpha^\star \in \mathbb{R}^{\mathbb{N}}$ such that one can write:

$$f^\star = f_{\alpha^\star} = \sum_{i\geqslant 1} \alpha^\star_i \varphi_i,$$

Thus we consider in the sequel one such $\alpha^\star$. This enables to derive the following upper bound:

$$\inf_{g\in\mathcal{G}} \|f^\star - T_L(g)\|^2_{\mathcal{P}_X} \leqslant \|f_{\alpha^*} - T_L(g_{A\alpha^*})\|^2_{\mathcal{P}_X}.$$

where we applied the gaussian operator A to the sequence $\alpha^\star$.

Step 1. Applying Johnson-Lindenstrauss' Lemma. Let us introduce m ghost samples $(X'_j)_{j\leqslant m}$ i.i.d. according to $\mathcal{P}_\mathcal{X}$, and thus consider the following associated norm

$$\|f_{\alpha^*} - T_L(g_{A\alpha^*})\|_m^2 = \frac{1}{m}\sum_{j=1}^m (f_{\alpha^*} - T_L(g_{A\alpha^*}))^2(X'_j)\,.$$

We now make explicit the probability spaces corresponding to the different sources of randomness. Consider the probability space defined over the product sample space $\Omega_\mathcal{X}\times\Omega_\mathcal{G}$, where $\Omega_\mathcal{X}$ consists of all the possible realizations of J states $X'_1,\ldots,X'_m$ drawn i.i.d. from $\mathcal{P}_\mathcal{X}$, and $\Omega_\mathcal{G}$ is the set of all possible realizations of the random elements $(A_{p,i})_{1\leqslant p\leqslant P, i\geqslant 1}$ (which define the random feature space $\mathcal{G}_P$).

Let us fix some $\omega_\mathcal{G}\in\Omega_\mathcal{G}$ (which defines the random subspace $\mathcal{G}_P(\omega_\mathcal{G})$). Since for all j, we have that $(f_{\alpha^*} - T_L(g_{A\alpha^*}))^2(X'_j)\in[0,4L^2]$ $\mathcal{P}_\mathcal{X}$-a.s., then Hoeffding's inequality applies; we deduce that there exists an event $\Omega_\mathcal{X}(\omega_\mathcal{G})$ of $\mathcal{P}_\mathcal{X}$-probability higher than $1-\delta_\mathcal{X}$ such that on this event

$$\|f_{\alpha^*} - T_L(g_{A\alpha^*})\|_{\mathcal{P}_\mathcal{X}}^2 \leqslant \|f_{\alpha^*} - T_L(g_{A\alpha^*})\|_m^2 + (2L)^2\sqrt{\frac{\log(1/\delta)}{2m}}.$$

Now by independence between the Gaussian random variables and the sample, the same inequality is valid on the event

$$\Omega_1 = \{\omega_\mathcal{X}\times\omega_\mathcal{G}; \omega_\mathcal{G}\in\Omega_\mathcal{G}, \omega_\mathcal{X}\in\Omega_\mathcal{X}(\omega_\mathcal{G})\}\,,$$

and this event has $\mathcal{P}_\mathcal{X}\times\mathcal{P}_\mathcal{G}$-probability higher than $1-\delta_\mathcal{X}$.

In order to bound the first term of the right hand side of this inequality, we first notice that since $\|f_{\alpha^*}\|_\infty\leqslant L$, then

$$\|f_{\alpha^*} - T_L(g_{A\alpha^*})\|_m^2 \quad\leqslant\quad \|f_{\alpha^*} - g_{A\alpha^*}\|_m^2\,,$$

then for some fixed $\omega_\mathcal{X}\in\Omega_\mathcal{X}$, that last term is bounded by $\varepsilon^2\|\alpha^\star\|^2\sup_x\|\varphi(x)\|^2$ on an event $\Omega_\mathcal{G}(\omega_\mathcal{X})$ of $\mathcal{P}_\mathcal{G}$-probability higher than $1-4me^{-P(\varepsilon^2/4-\varepsilon^3/6)}$ by application of Lemma 9.

Thus still by independence, the same inequality is valid on the event

$$\Omega_2 = \{(\omega_\mathcal{X},\omega_\mathcal{G}); \omega_\mathcal{X}\in\Omega_\mathcal{X}, \omega_\mathcal{G}\in\Omega_\mathcal{G}(\omega_\mathcal{X})\}\,,$$

and this event has $\mathcal{P}_\mathcal{X}\times\mathcal{P}_\mathcal{G}$-probability higher than $1-4me^{-P(\varepsilon^2/4-\varepsilon^3/6)}$.

Thus, we deduce, by a union bound that for all $\varepsilon\in(0,1)$ and $m\geqslant 1$ there exists an event $\Omega_1\cap\Omega_2$ of $\mathcal{P}_\mathcal{X}\times\mathcal{P}_\mathcal{G}$-probability higher than $1-\delta_\mathcal{X}-4me^{-P(\varepsilon^2/4-\varepsilon^3/6)}$ such that on this event,

$$\inf_{g\in\mathcal{G}}\|f^\star - T_L(g)\|_{\mathcal{P}_\mathcal{X}}^2 \leqslant \varepsilon^2\|\alpha^\star\|^2\sup_x\|\varphi(x)\|^2 + (2L)^2\sqrt{\frac{\log(1/\delta)}{2m}}\,.$$

Finally in order to get a bound in high $\mathcal{P}_\mathcal{G}$-probability only, we introduce for any $\omega_\mathcal{G}\in\Omega_\mathcal{G}$ the event $\Omega'_\mathcal{X}(\omega_\mathcal{G})\stackrel{\text{def}}{=}\{\omega_\mathcal{X}\in\Omega_\mathcal{X}; (\omega_\mathcal{X},\omega_\mathcal{G})\in\Omega_1\times\Omega_2\}$ and then define for all $\lambda>0$ the event

$$\Lambda\stackrel{\text{def}}{=}\{\omega_\mathcal{G}\in\Omega_\mathcal{G}; \mathbb{P}_\mathcal{X}(\Omega'_\mathcal{X}(\omega_\mathcal{G}))\geqslant 1-\lambda\}\,.$$

Using this notation, we deduce that for all $\omega_{\mathcal{G}} \in \Lambda$, the following bound holds

$$\begin{aligned}\inf_{g\in\mathcal{G}_P(\omega_{\mathcal{G}})} \|f^\star - T_L(g)\|^2_{\mathcal{P}_X} &\leqslant \int_{\Omega'_X(\omega_{\mathcal{G}})} \inf_{g\in\mathcal{G}_P(\omega_{\mathcal{G}})} \|f^\star - T_L(g)\|^2_{\mathcal{P}_X} d\omega_X \\ &\quad + \int_{\Omega'_X(\omega_{\mathcal{G}})^c} \inf_{g\in\mathcal{G}_P(\omega_{\mathcal{G}})} \|f^\star - T_L(g)\|^2_{\mathcal{P}_X} d\omega_X \\ &\leqslant \varepsilon^2 \|\alpha^\star\|^2 \sup_x \|\varphi(x)\|^2 + (2L)^2 \sqrt{\frac{\log(1/\delta)}{2m}} + (2L)^2\lambda .\end{aligned}$$

Moreover, since $\mathbb{P}_{\mathcal{X}\times\mathcal{G}}(\Omega_1 \cap \Omega_2) \geqslant 1 - \delta_{\mathcal{X}} - 4me^{-P(\varepsilon^2/4-\varepsilon^3/6)}$ and on the other side

$$\begin{aligned}\mathbb{P}_{\mathcal{X}\times\mathcal{G}}(\Omega_1 \cap \Omega_2) &= \int_{\Omega_{\mathcal{G}}} \mathbb{P}_{\mathcal{X}}(\Omega'_{\mathcal{X}}(\omega_{\mathcal{G}})) d\omega_{\mathcal{G}} \\ &\leqslant \int_{\Omega_{\mathcal{G}}} \mathbb{I}_{\mathbb{P}_{\mathcal{X}}(\Omega'_{\mathcal{X}}(\omega_{\mathcal{G}}))\geqslant 1-\lambda} d\omega_{\mathcal{G}} + (1-\lambda) \int_{\Omega_{\mathcal{G}}} \mathbb{I}_{\mathbb{P}_{\mathcal{X}}(\Omega'_{\mathcal{X}}(\omega_{\mathcal{G}}))<1-\lambda} d\omega_{\mathcal{G}} \\ &\leqslant \mathbb{P}_{\mathcal{G}}(\Lambda) + (1-\lambda)(1 - \mathbb{P}_{\mathcal{G}}(\Lambda)) ,\end{aligned}$$

then we deduce that $\mathbb{P}_{\mathcal{G}}(\Lambda)) \geqslant 1 - \frac{\delta_{\mathcal{X}} + 4me^{-P(\varepsilon^2/4-\varepsilon^3/6)}}{\lambda}$.

Step 2. Tuning the parameters ε. Now let us introduce $\delta_{\mathcal{G}}$ and define for some $m \geqslant 1$ the quantity $\varepsilon = \varepsilon_m(\delta_{\mathcal{G}})$ such that

$$\frac{\log(4m/\delta\mathcal{G})}{P} = \frac{\varepsilon^2}{4} - \frac{\varepsilon^3}{6} .$$

Thus, since $\varepsilon \in (0,1)$, this means in particular that we have

$$\frac{\varepsilon^2}{3} \leqslant 4\frac{\log(4m/\delta_{\mathcal{G}})}{P} \leqslant \varepsilon^2 .$$

Now by rewriting the bound using $\delta = \frac{\delta_{\mathcal{X}} + \delta_{\mathcal{G}}}{\lambda}$, we deduce that for all δ, for all m and λ, there exists an event of $\mathcal{P}_{\mathcal{G}}$-probability higher than $1-\delta$ such that

$$\inf_{g\in\mathcal{G}} \|f^\star - T_L(g)\|^2_{\mathcal{P}_X} \leqslant 12\frac{\log(\frac{8m}{\lambda\delta})}{P} \|\alpha^\star\|^2 \sup_x \|\varphi(x)\|^2 + (2L)^2 \Big(\sqrt{\frac{\log(\frac{2}{\lambda\delta})}{2m}} + \lambda\Big) .$$

Step 3. Optimizing over λ *and* m. Now, it remains to optimize the free parameter m and λ in this last bound; the optimal value for m is given by

$$m_{opt} = \frac{P^2 L^4 \log(\frac{2}{\lambda\delta})}{72\|\alpha\|^4 \sup_x \|\varphi(x)\|^4} ,$$

and the corresponding bound is thus

$$\inf_{g\in\mathcal{G}} \|f^\star - T_L(g)\|^2_{\mathcal{P}_X} \leqslant 24\frac{\|\alpha^\star\|^2 \sup_x \|\varphi(x)\|^2}{P} \Big(1 + \log\big(\frac{PL^2\sqrt{\log(2/\lambda\delta)/\lambda\delta}}{3\|\alpha^\star\|^2 \sup_x \|\varphi(x)\|^2}\big)\Big) + (2L)^2\lambda .$$

Now one can take $\lambda \stackrel{\text{def}}{=} \frac{\|\alpha^\star\|^2 \sup_x \|\varphi(x)\|^2}{(2L)^2 P}$ and deduce the final bound. ■

A.6 Proof of Theorem 21

Proof We make use of the following decomposition:

$$\|f^\star - g_{\widehat{\beta}}\|_N^2 \leqslant \|f^\star - g_{\tilde{\beta}}\|_N^2 + \|g_{\tilde{\beta}} - g_{\widehat{\beta}}\|_N^2 ,$$

and introduce the sets $\Omega_{\mathcal{G}}$ that consists of all possible realizations of the random elements $(A_{p,i})_{1\leqslant p\leqslant P, i\geqslant 1}$, and $\Omega_{\mathcal{Y}}$ that corresponds to the observation variables Y.

High $\mathcal{P}_{\mathcal{Y}} \times \mathcal{P}_{\mathcal{G}}$-probability bound. We again make explicit the probability spaces. For the first term on right hand side, an application of Theorem 19 ensures that there exists an event $\Omega'_{\mathcal{G}} \subset \Omega_{\mathcal{G}}$ of $\mathcal{P}_{\mathcal{G}}$-probability higher than $1-\delta$ such that for all $\omega_{\mathcal{G}} \in \Omega'_{\mathcal{G}}$,

$$\|f^\star - g_{\tilde{\beta}}\|_N^2 \leqslant 12 \frac{\log(4N/\delta)}{P} \|\alpha^\star\|^2 \frac{1}{N} \sum_{n=1}^{N} \|\varphi(x_n)\|^2 .$$

Since no random variable Y appears in this term, this is also true on the event

$$\Omega_1 \stackrel{\text{def}}{=} \{(\omega_{\mathcal{Y}}, \omega_{\mathcal{G}}) \in \Omega_{\mathcal{Y}} \times \Omega_{\mathcal{G}}; \omega_{\mathcal{G}} \in \Omega'_{\mathcal{G}}\} ,$$

and Ω_1 has $\mathcal{P}_{\mathcal{Y}} \times \mathcal{P}_{\mathcal{G}}$-probability higher than $1-\delta$.

For the second term, let us fix some $\omega_{\mathcal{G}} \in \Omega_{\mathcal{G}}$. Then Lemma 24 below shows that there exists an event $\Omega_{\mathcal{Y}}(\omega_{\mathcal{G}}) \subset \Omega_{\mathcal{G}}$ of $\mathcal{P}_{\mathcal{Y}}$-probability higher than $1-\delta'$ such that for all $\omega_{\mathcal{Y}} \in \Omega_{\mathcal{Y}}(\omega_{\mathcal{G}})$,

$$\|g_{\tilde{\beta}} - g_{\widehat{\beta}}\|_N^2 \leqslant \kappa B^2 \frac{P + \log(1/\delta')}{N} ,$$

for some numerical constant $\kappa > 0$. Thus by independence of the noise term with the Gaussian variables, we deduce that a similar bound holds on the event

$$\Omega_2 \stackrel{\text{def}}{=} \{(\omega_{\mathcal{Y}}, \omega_{\mathcal{G}}) \in \Omega_{\mathcal{Y}} \times \Omega_{\mathcal{G}}; \omega_{\mathcal{Y}} \in \Omega_{\mathcal{Y}}(\omega_{\mathcal{G}})\} ,$$

and that Ω_2 has $\mathcal{P}_{\mathcal{Y}} \times \mathcal{P}_{\mathcal{G}}$-probability higher than $1-\delta'$. Thus, we conclude by a simple union bound in order te get a result in high $\mathcal{P}_{\mathcal{Y}} \times \mathcal{P}_{\mathcal{G}}$-probability. ■

A.7 Proof of Theorem 22

Proof

Similarly to the proof of Theorem 20, we introduce the sets $\Omega_{\mathcal{X}}, \Omega_\eta$ and $\Omega_{\mathcal{G}}$ that consist of all possible realizations of the input, noise and Gaussian random variables. We then define $\Omega \stackrel{\text{def}}{=} \Omega_{\mathcal{X}} \times \Omega_\eta \times \Omega_{\mathcal{G}}$.

Step 1. High $\mathcal{P} \times \mathcal{P}_{\mathcal{G}}$-probability bound. In order to get a high probability bound, we use the decomposition given by Lemma 23. Now let us consider some fixed $\omega_{\mathcal{G}} \in \Omega_{\mathcal{G}}$. One can apply Lemma 24 and Lemma 26 below for the noise and estimation term.

Thus when $N\log(N) \geqslant \frac{4}{P}$, there exists an event $\Omega_1(\omega_{\mathcal{G}})$ of $\mathcal{P}$-probability higher than $1-\delta_1$ and an event $\Omega_2(\omega_{\mathcal{G}})$ of $\mathcal{P}$-probability higher than $1-\delta_2$ such that for all $(\omega_{\mathcal{X}}, \omega_\eta) \in \Omega_1(\omega_{\mathcal{G}})$ we have

$$\|g_{\tilde{\beta}} - g_{\widehat{\beta}}\|_N^2 \leqslant 6B^2 \frac{(1616P + 200\log(6/\delta) + \log(3/\delta))}{N} ,$$

and for all $(\omega_X, \omega_\eta) \in \Omega_2(\omega_G)$ we have

$$\sup_{g \in \mathcal{G}_P} \|f^\star - T_L(g)\|_{\mathcal{P}_X}^2 - 8\|f^\star - T_L(g)\|_N^2 \leqslant (24L)^2 \frac{4\log(3/\delta) + 2P\log(N)}{N}.$$

On the other hand, by application of Theorem 19, for any given (ω_X, ω_η), there exists an event $\Omega_G(\omega_X, \omega_\eta) \subset \Omega_G$ of $\mathcal{P}_G$-probability higher than $1 - \delta_3$ such that on this event

$$\|f^\star - g_{\tilde{\beta}}\|_N^2 \leqslant 12 \frac{\log(4N/\delta_3)}{P} \|\alpha^\star\|^2 \sup_{x \in X} \|\varphi(x)\|^2.$$

Thus by independence of the noise, data points and Gaussian variables, the three previous inequalities are valid respectively on the events

$$\Omega_1 = \{(\omega_X, \omega_\eta, \omega_G) \in \Omega; (\omega_X, \omega_\eta) \in \Omega_1(\omega_G)\},$$
$$\Omega_2 = \{(\omega_X, \omega_\eta, \omega_G) \in \Omega; (\omega_X, \omega_\eta) \in \Omega_2(\omega_G)\},$$
$$\Omega_3 = \{(\omega_X, \omega_\eta, \omega_G) \in \Omega; \omega_G \in \Omega_G(\omega_X)\}.$$

Moreover Ω_1 has $\mathcal{P} \times \mathcal{P}_G$-probability higher than $1 - \delta_1$, Ω_2 has $\mathcal{P} \times \mathcal{P}_G$-probability higher than $1 - \delta_2$, and Ω_1 has $\mathcal{P} \times \mathcal{P}_G$-probability higher than $1 - \delta_3$. We thus conclude by a simple union bound, and then by some cosmetic simplifications introducing some constant κ. ■

References

Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal Of Computer And System Sciences*, 66(4):671–687, June 2003.

Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *Proceedings Of The 38th Annual ACM Symposium On Theory Of Computing*, STOC '06, pages 557–563, New York, NY, USA, 2006. ACM. ISBN 1-59593-134-1.

Pierre Alquier. PAC-Bayesian bounds for randomized empirical risk minimizers. *Mathematical Methods Of Statistics*, 17(4):279–304, dec 2008.

Pierre Alquier and Karim Lounici. PAC-Bayesian Bounds for Sparse Regression Estimation with Exponential Weights. *Electronic Journal Of Statistics*, 5:127–145, 2011.

Jean-Marie Aubry and Stéphane Jaffard. Random wavelet series. *Communications In Mathematical Physics*, 227:483–514, 2002.

Jean-Yves Audibert and Olivier Catoni. Robust linear regression through PAC-Bayesian truncation. arXiv, 2010.

Andrew Barron, Albert Cohen, Wolfgang Dahmen, and Ronald DeVore. Approximation and learning by greedy algorithms. *Annals Of Statistics*, 36:1:64–94, 2008.

Lucien Birgé and Pascal Massart. Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, sep 1998.

Gerard Bourdaud. Ondelettes et espaces de besov. *Rev. Mat. Iberoamericana*, 11:3:477–512, 1995.

Hans-Joachim Bungartz and Michaël Griebel. Sparse grids. In Arieh Iserles, editor, *Acta Numerica*, volume 13. University of Cambridge, 2004.

Stphane Canu, Xavier Mary, and Alain Rakotomamonjy. Functional learning through kernel. In *Advances In Learning Theory: Methods, Models And Applications NATO Science Series III: Computer And Systems Sciences*, pages 89–110. IOS Press, 2002.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations Of Computational Mathematics*, 7:331–368, jul 2007. ISSN 1615-3375.

Olivier Catoni. *Statistical Learning Theory And Stochastic Optimization*. Springer-Verlag, 2004.

Don Coppersmith and Shmuel Winograd. Matrix multiplication via arithmetic progressions. In *Proceedings Of The 19th Annual ACM Symposium On Theory Of Computing*, STOC '87, pages 1–6, New York, NY, USA, 1987. ACM.

Sanjoy Dasgupta and Yoav Freund. Random projection trees and low dimensional manifolds. In *Proceedings Of The 40th Annual ACM Symposium On Theory Of Computing*, STOC '08, pages 537–546, New York, NY, USA, 2008. ACM.

Sébastien Deguy and Albert Benassi. A flexible noise model for designing maps. In *Proceedings Of The Vision Modeling And Visualization Conference 2001*, VMV '01, pages 299–308. Aka GmbH, 2001. ISBN 3-89838-028-9.

Ronald DeVore. *Nonlinear Approximation*. Acta Numerica, 1997.

Richard M. Dudley. *Real Analysis And Probability*. Wadsworth, Belmont, Calif, 1989.

Arnaud Durand. Random wavelet series based on a tree-indexed Markov chain. *Communications In Mathematical Physics*, 283:451–477, 2008.

Michaël Frazier and Björn Jawerth. Decomposition of Besov spaces. *Indiana University Mathematics Journal*, (34), 1985.

László Györfi, Michaël Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-free Theory Of Nonparametric Regression*. Springer-Verlag, New York, 2002.

Svante Janson. *Gaussian Hilbert Spaces*. Cambridge Univerity Press, Cambridge, UK, 1997.

Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

Edo Liberty, Nir Ailon, and Amit Singer. Dense fast random projections and lean Walsh transforms. In *Proceedings Of The 11th International Workshop, APPROX 2008, And 12th International Workshop, RANDOM 2008 On Approximation, Randomization And Combinatorial Optimization: Algorithms And Techniques*, APPROX '08 / RANDOM '08, pages 512–522, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-85362-6.

Mikhail A. Lifshits. *Gaussian Random Functions*. Kluwer Academic Publishers, Dordrecht, Boston, 1995.

Odalric-Ambrym Maillard and Rémi Munos. Compressed least-squares regression. In Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Chris K. I. Williams, and Aron Culotta, editors, *Proceedings Of The 23rd Conference On Advances In Neural Information Processing Systems*, NIPS '09, pages 1213–1221, Vancouver, British Columbia, Canada, dec 2009.

Odalric-Ambrym Maillard and Rémi Munos. Scrambled objects for least-squares regression. In John D. Lafferty, Chris K. I. Williams, John Shawe-Taylor, Richard S. Zemel, and Aron Culotta, editors, *Proceedings Of The 24th Conference On Advances In Neural Information Processing Systems*, NIPS '10, pages 1549–1557, Vancouver, British Columbia, Canada, dec 2010.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *Proceedings Of The 21st Conference On Advances In Neural Information Processing Systems*, NIPS '07, Vancouver, British Columbia, Canada, dec 2007. MIT Press.

Ali Rahimi and Benjamin Recht. Uniform approximation of functions with random bases. *Proceedings Of The 46th Annual Allerton Conference*, 2008.

Saburou Saitoh. *Theory Of Reproducing Kernels And Its Applications*. Longman Scientific & Technical, Harlow, UK, 1988.

Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *Proceedings Of The 47th Annual IEEE Symposium On Foundations Of Computer Science.*, FOCS '06, pages 143–152, 2006.

Richard S. Sutton. Generalization in reinforcement learning: Successful examples using sparse coarse coding. In *Advances In Neural Information Processing Systems 8*, pages 1038–1044. MIT Press, 1996.

Richard S. Sutton and Steven D. Whitehead. Online learning with random representations. In *In Proceedings Of The 10th International Conference On Machine Learning*, ICML '93, pages 314–321. Morgan Kaufmann, 1993.

Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal Of The Royal Statistical Society, Series B*, 58:267–288, 1994.

Andrei N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math Dokl 4*, pages 1035–1038, 1963.

Alexandre B. Tsybakov. Optimal rates of aggregation. In *Proceedings Of The 16th Annual Conference On Learning Theory*, pages 303–313, 2003.

Adriaan Zaanen. *Linear Analysis*. North Holland Publishing, 1960.

Christoph Zenger. Sparse grids. In W. Hackbusch, editor, *Parallel Algorithms For Partial Differential Equations, Proceedings Of The Sixth GAMM-Seminar*, volume 31 of Notes on Num. Fluid Mechanics, Kiel, 1990. Vieweg-Verlag.

Bin Zhao and Changshui Zhang. Compressed spectral clustering. In *Proceedings Of The 2009 IEEE International Conference On Data Mining Workshops*, ICDMW '09, pages 344–349, Washington, DC, USA, 2009. IEEE Computer Society.