HAL
archives-ouvertes.fr

Brownian Motions and Scrambled Wavelets for Least-Squares Regression

Odalric-Ambrym Maillard, Rémi Munos

▶ **To cite this version:**

Odalric-Ambrym Maillard, Rémi Munos. Brownian Motions and Scrambled Wavelets for Least-Squares Regression. [Technical Report] 2010, pp.13. <inria-00483017>

HAL Id: inria-00483017
https://hal.inria.fr/inria-00483017

Submitted on 12 May 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Brownian Motions and Scrambled Wavelets for Least-Squares Regression

Odalric-Ambrym Maillard, Rémi Munos
SequeL Project, INRIA Lille
40 avenue Halley,
59650 Villeneuve d'Ascq, France
{odalric.maillard, remi.munos}@inria.fr

Abstract

We consider ordinary (non penalized) least-squares regression where the regression function is chosen in a randomly generated sub-space $\mathcal{G}_P \subset \mathcal{S}$ of finite dimension P, where $\mathcal{S}$ is a function space of infinite dimension, e.g. $L_2([0,1]^d)$. $\mathcal{G}_P$ is defined as the span of P random features that are linear combinations of the basis functions of $\mathcal{S}$ weighted by random Gaussian i.i.d. coefficients. We characterize the so-called kernel space $\mathcal{K} \subset \mathcal{S}$ of the resulting Gaussian process and derive approximation error bounds of order $O(||f||_{\mathcal{K}}^2 \log(P)/P)$ for functions $f \in \mathcal{K}$ approximated in $\mathcal{G}_P$. We apply this result to derive excess risk bounds for the least-squares estimate in various spaces. For illustration, we consider regression using the so-called *scrambled wavelets* (i.e. random linear combinations of wavelets of $L_2([0,1]^d)$) and derive an excess risk rate $O(||f^*||_{\mathcal{K}}(\log N)/\sqrt{N})$ which is arbitrarily close to the minimax optimal rate (up to a logarithmic factor) for target functions f^* in $\mathcal{K} = H^s([0,1]^d)$, a Sobolev space of smoothness order $s > d/2$. We describe an efficient implementation using lazy expansions with numerical complexity $\tilde{O}(2^d N^{3/2} \log N + N^{5/2})$, where d is the dimension of the input data and N is the number of data.

1 Introduction

We consider ordinary least-squares regression using randomly generated feature spaces. Let us first describe the general regression problem: we observe data $\mathcal{D}_N = (\{x_n, y_n\}_{n \le N})$ (with $x_n \in \mathcal{X}$, $y_n \in \mathbb{R}$), assumed to be independently and identically distributed (i.i.d.) from some distribution P, where $x_n \sim P_{\mathcal{X}}$ and

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

where f^* is some (unknown) target function such that $||f^*||_\infty \le L$ and η_n is a centered, independent noise of variance bounded by σ^2. We assume that L and σ are known.

Now, for a given class of functions $\mathcal{F}$, and $f \in \mathcal{F}$, we define the empirical ℓ_2-error

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

and the generalization error

$$L(f) \stackrel{\text{def}}{=} \mathbb{E}_{(X,Y)\sim P}[(Y - f(X))^2].$$

The goal is to return a regression function $\widehat{f} \in \mathcal{F}$ with lowest possible generalization error $L(\widehat{f})$. The excess risk $L(\widehat{f}) - L(f^*)$ measures the closeness to optimality.

In this paper we consider infinite dimensional spaces $\mathcal{F}$ that are generated by a denumerable family of functions $\{\varphi_i\}_{i \ge 1}$, called *initial features*. We assume that $f^* \in \mathcal{F}$.

Since $\mathcal{F}$ is an infinite dimensional space, the empirical risk minimizer in $\mathcal{F}$ is certainly subject to overfitting. Traditional methods to circumvent this problem have considered penalization, i.e. one searches for a function in $\mathcal{F}$ which minimizes the empirical error plus a penalty term, for example

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} L_N(f) + \lambda ||f||_p^p, \quad \text{for } p = 1 \text{ or } 2.$$

where λ is a parameter and usual choices for the norm are ℓ_2 (ridge-regression [17]) and ℓ_1 (LASSO [16]).

In this paper we follow an alternative approach introduced in [10], called Compressed Least Squares Regression, which considers generating randomly a subspace $\mathcal{G}_P$ (of finite dimension P) of $\mathcal{F}$, and then

returning the empirical risk minimizer in $\mathcal{G}_P$, i.e. $\arg\min_{g\in\mathcal{G}_P} L_N(g)$. Their work considered the case when $\mathcal{F}$ is of finite dimension. Here we consider the case of infinite dimension and provide a characterization of the resulting approximation spaces for which learning is possible.

Regression with random subspaces: Let us briefly recall the method described in [10]. The random subspace $\mathcal{G}_P$ is generated by constructing a set of P *random features* $(\psi_p)_{1\leq p\leq P}$ defined as linear combinations of the initial features $\{\varphi_i\}_{1\leq i\leq F}$ (in their work, they assumed that $\mathcal{F}$ has a finite dimension F) weighted by some random coefficients:

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

where the coefficient $A_{p,i}$ are drawn i.i.d. from a centered distribution with variance $1/P$. Here we explicitly choose a Gaussian distribution $\mathcal{N}(0, 1/P)$. We write A the random $P \times F$ matrix with elements $(A_{p,i})$.

Then $\mathcal{G}_P \subset \mathcal{F}$ is defined as the vector space spanned by those features, i.e.

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

Now, the least-squares estimate $g_{\widehat{\beta}} \in \mathcal{G}_P$ is the function in $\mathcal{G}_P$ with minimal empirical error, i.e.

$$g_{\widehat{\beta}} = \arg\min_{g_\beta\in\mathcal{G}_P} L_N(g_\beta), \tag{1}$$

and is such that $\widehat{\beta} = \Psi^\dagger Y \in \mathbb{R}^P$, where 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 of Ψ[1]. The final prediction function $\widehat{g}(x)$ is the truncation (to the threshold $\pm L$) of $g_{\widehat{\beta}}$, i.e. $\widehat{g}(x) \stackrel{\text{def}}{=} T_L[g_{\widehat{\beta}}(x)]$, where

$$T_L(u) \stackrel{\text{def}}{=} \begin{cases} u & \text{if } |u| \leq L, \\ L \text{ sign}(u) & \text{otherwise.} \end{cases}$$

The result of [10] states that with probability at least $1-\delta$ (on the choice of the random matrix A, thus on $\mathcal{G}_P$), the generalization error of $\widehat{g}$ is bounded as

$$L(\widehat{g}) - \inf_{f\in\mathcal{F}} L(f) = O\Big(\frac{P\log(N)}{N} + ||\alpha^*||^2 \sup_{x\in\mathcal{X}} ||\varphi(x)||^2 \frac{\log(N/\delta)}{P}\Big), \tag{2}$$

where α^* is the vector coefficient of the target function $f^* \in \mathcal{F}$ (i.e. $f^* = \sum_{i=1}^F \alpha_i^*\varphi_i$) and $||u||^2 \stackrel{\text{def}}{=} \sum_{i=1}^F u_i^2$. The bound shows a usual decomposition in terms of estimation error (first term) and approximation error w.r.t. $\mathcal{G}_P$ (second term). Now, the best tradeoff, obtained for P of order $||\alpha^*|| \sup_x ||\varphi(x)||\sqrt{N}$, provides an excess risk of $O(||\alpha^*|| \sup_x ||\varphi(x)|| \frac{\log N/\delta}{\sqrt{N}})$.

Our motivation The result (2) says that if the term $||\alpha^*|| \sup_x ||\varphi(x)||$ is small, then the least-squares estimate in the random subspace $\mathcal{G}_P$ has low excess risk. The questions we wish to address now are: Can we characterize the spaces for which this is the case? And, since the final bound does not involve the dimension F, can we handle the case when $\mathcal{F}$ is of infinite dimension? This paper gives a positive answer to these questions.

An initial motivation for this work lies in the following simple property of the random features $(\psi_p)_{1\leq p\leq P}$:

Lemma 1 *Provided that the initial features $(\varphi_i)_{1\leq i\leq F}$ are continuous, each random feature ψ_p is a Gaussian process indexed by the space $\mathcal{X}$ with covariance structure given by $\frac{1}{P}\langle\varphi(x), \varphi(x')\rangle$ (where $\varphi(x) \in \mathbb{R}^F$ is the vector with components $(\varphi_i(x))_{1\leq i\leq F}$).*

Proof: By definition, $\psi_p(x) = A_p\varphi(x)$ where A_p is the p^{th} row of the random matrix A, which contains i.i.d. $\mathcal{N}(0, \frac{1}{P})$ entries. We may rewrite $\psi_p(x) = \psi_x(A_p)$ to highlight the role of the random variable A_p. Thus $\mathbb{E}_{A_p}(\psi_x(A_p)) = 0$, and by definition:

$$\begin{aligned} \text{Cov}_{A_p}(\psi_x, \psi_{x'}) &= \mathbb{E}_{A_p}(\psi_x(A_p)\psi_{x'}(A_p)) \\ &= \frac{1}{P}\sum_{i=1}^{F} \varphi_i(x)\varphi_i(x') = \frac{1}{P}\langle\varphi(x), \varphi(x')\rangle \end{aligned}$$

[1] In the full rank case when $N \geq P$, $\Psi^\dagger = (\Psi^T\Psi)^{-1}\Psi^T$

The continuity of the initial features (φ_i) guarantees that there exists a continuous version of the process ψ_p which is thus a Gaussian process. ■

Then it is natural to ask what happens when $F \to \infty$. Since $\psi_p(x) = \sum_{i=1}^{F} A_{p,i}\varphi_i(x)$, this means we want to understand and give a meaning to the object $W(x) = \sum_{i=1}^{\infty} \xi_i \varphi_i(x)$ (where we drop the index p for simplicity and introduce the standard normal i.i.d. variables $(\xi_i)_i$). We will use the theory of Gaussian random functions in Section 2.3 (see [9, 8]) for that purpose. But let us start with a motivating example.

Motivating Example 1: Brownian motions indexed by $[0,1]$**.** Consider the following simple example where the initial features are multiscaled hat functions on the space $\mathcal{X} = [0,1]$. The mother hat function is $\Lambda(x) = x\mathbb{I}_{[0,1/2[} + (1-x)\mathbb{I}_{[1/2,1[}$, and the rescaled hat functions are $\Lambda_{j,l}(x) = 2^{-j/2}\Lambda(2^j x - l)$ for any scale $j \geq 1$ and translation index $0 \leq l \leq 2^j - 1$. We also write $\Lambda_0(x) = x$ and $\Lambda_1(x) = 1$. This defines a basis of $\mathcal{C}^0([0,1])$ (introduced by Faber in 1910, and known as the Schauder basis, see [7] for an interesting overview).

Those functions are indexed by the scale j and translation index l, but all functions may be equivalently indexed by a unique index $i \geq 1$.

We will see in Section 2.2 that the random features $\psi_p(x)$, defined as linear combinations of those hat functions weighted by Gaussian i.i.d. random numbers, are Brownian motions. In Section 3 we will show that such Brownian motions are indeed good for regression in the sense that if the target function f^* belongs to the Sobolev space $H^1([0,1])$ of order 1 (space of functions which have a weak derivative in $L_2([0,1])$), then the least-squares estimate in $\mathcal{G}_P$ (defined by P Brownian motions) has an excess risk of $O(\frac{P\log N}{N} + ||f||_{H^1}^2 \frac{\log P}{P})$. Now choosing P of order $||f||_{H^1}\sqrt{N}$, we deduce the excess risk $O(||f||_{H^1}\frac{\log N}{\sqrt{N}})$.

Our contribution In this paper, we analyze least-squares regression with random subspaces and illustrate our analysis using two examples: *Brownian motions* (Example 1) and what we call *scrambled wavelets* (Example 2), defined as linear combination of wavelets weighted by random Gaussian coefficients.

In Section 2, we study the random objects (random features, random subspace) considered in this setting, and precisely characterize a special space of functions related to these objects called the kernel space. We make use of two theories: (1) Gaussian Random Function Theory (that studies precise properties of objects like Brownian bridges, Strassen balls, etc.) that allows for good flexibility in the choice of the basis functions of the target space, and (2) Approximation Theory, and more precisely, Multi-Resolution Analysis (MRA) (which deals for instance with wavelets and Besov spaces).

In Section 3, we apply this analysis to provide generalization bounds which extend the results of [10] to the case when $\mathcal{F}$ is of infinite dimension. The analysis shows that in Example 2, the excess risk of the least-squares estimate built with scrambled wavelets is arbitrarily close to the minimax rates on the associated kernel space (which is a Sobolev space).

Finally, in Section 5, we describe an efficient numerical implementation using a multiresolution tree structure that generates the expansion of the random features (ψ_p) only at the data points (lazy implementation). The resulting algorithm has numerical cost $O(2^d N^{3/2}\log N)$ (where d is the dimension of $\mathcal{X}$) for building the linear system, and $O(N^{5/2})$ for solving it.

2 Elements of theory about Gaussian objects

We now give an interpretation of the random features in terms of random processes and analyze the corresponding limit object when the dimension of the initial feature space F is infinite.

In this Section we will introduce the notion of a Gaussian object W (Section 2.1), define its kernel space $\mathcal{K}$ (Section 2.2), and its expansion (Section 2.3).

2.1 Gaussian objects

Let $\mathcal{S}$ be a vector space and $\mathcal{S}'$ its dual. We write $(\cdot,\cdot)$ its duality product.

Definition 2 (Gaussian objects) *A random $W \in \mathcal{S}$ is called a Gaussian object if for all $\nu \in \mathcal{S}'$, we have that (ν, W) is a Gaussian (real-valued) variable. Now we call*

- *$a \in \mathcal{S}$ an expectation of W if $\forall \nu \in \mathcal{S}'$, $\mathbb{E}(\nu, W) = (\nu, a)$.*
- *$K : \mathcal{S}' \to \mathcal{S}$ a covariance operator of W if $\forall \nu, \nu' \in \mathcal{S}'$, $Cov((\nu, W)(\nu', W)) = (\nu, K\nu')$.*

When a and K exists, we write $W \sim \mathcal{N}(a, K)$.

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

2.2 Definition of the kernel space

Given a Gaussian centered object W, one may naturally define a space $\mathcal{K} \subset \mathcal{S}$ called the **kernel space** of $\mathcal{N}(0,K)$. It is built by first enriching $\mathcal{S}'$ with all measurable linear functionals (w.r.t. W), and then taking the dual of its closure. We now define it precisely by introducing the canonical injection I' of the continuous linear functionals into the space of measurable linear functionals, and its adjoint I. We refer the interested reader to [9] or [8] for refinements.

For any $\nu \in \mathcal{S}'$, we have $(\nu, K\nu) = \mathbb{E}(\nu, W)^2 < \infty$. Thus $(\nu, \cdot) \in L^2(\mathcal{S}, \mathcal{N}(0,K))$ which is the space of square integrable functionals under measure $\mathcal{N}(0,K)$, i.e. $\{z : \mathcal{S} \to \mathbb{R}, \mathbb{E}_{W\sim\mathcal{N}(0,K)}|z(W)|^2 < \infty\}$. Now define the injection $I' : \mathcal{S}' \to L^2(\mathcal{S}, \mathcal{N}(0,K))$ by $I'(\nu) = (\nu, \cdot)$. Then the space of measurable linear functionals $\mathcal{S}'_{\mathcal{N}} = \overline{I'(\mathcal{S}')}$ is the closure of the image of $\mathcal{S}'$ by I' (in the L^2 sense). It is a Hilbert space with inner product inherited from $L^2(\mathcal{S}, \mathcal{N}(0,K))$, i.e. $\langle z_1, z_2 \rangle_{\mathcal{S}'_{\mathcal{N}}} = \mathbb{E}(z_1(W) z_2(W))$ (where z can be written as $z = \lim_n (\nu_n, \cdot)$ with $\nu_n \in \mathcal{S}'$).

Provided that I' is continuous (see Section 2.4 for practical conditions ensuring when this is the case) we define the adjoint $I : \mathcal{S}'_{\mathcal{N}} \to \mathcal{S}$ of I', by duality: For any $\mu \in \mathcal{S}'$, $(\mu, Iz) = \langle I'\mu, z \rangle_{\mathcal{S}'_{\mathcal{N}}} = \mathbb{E}_W((\mu, W) z(W))$, from which we deduce that $(Iz)(x) = \mathbb{E}_W(W(x) z(W))$.

Now the **kernel space of** $\mathcal{N}(0,K)$ is defined as $\mathcal{K} \stackrel{\text{def}}{=} I(\overline{I'(\mathcal{S}')}) \subset \mathcal{S}$.

Equivalent construction of the kernel space. The kernel space can be built alternatively based on a separable Hilbert space $\mathcal{H}$ as follows:

Lemma 3 *[9] 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 $\langle Jh_1, Jh_2 \rangle_H \stackrel{\text{def}}{=} \langle h_1, h_2 \rangle_{\mathcal{H}}$.*

Example 1 (continued) In the case of the Brownian motions already considered, one may build $\mathcal{K}$ by choosing the Hilbert space $\mathcal{H} = L^2([0,1])$ and the mapping $J : \mathcal{H} \mapsto \mathcal{S}$ defined by $(Jh)(t) = \int_{[0,t]} h(s) ds$, which satisfies $(J'\nu)(t) = \nu([t,1])$ and $K = JJ'$. Thus, the kernel space $\mathcal{K}$ is $J(L^2([0,1])) = \{k \in H^1([0,1]); k(0) = 0\}$, the Sobolev space of order 1 with functions being equal to 0 on the left boundary.

Now, for the extension to dimension d, we consider the space $\mathcal{S} = \mathcal{C}([0,1]^d)$ and the covariance operator of the Brownian sheet (Brownian motion in dimension d) $(K\nu)(t) = \int_{[0,1]^d} \Pi_{i=1}^d \min(s_i, t_i) \nu(ds)$. The Hilbert space is $\mathcal{H} = L^2([0,1]^d)$ and we choose J to be the volume integral $(Jh)(t) = \int_{[0,t]} h(s) ds$, which implies that $K = JJ'$.

Thus $\mathcal{K} = J(L^2([0,1]^d))$ is the so-called *Cameron-Martin space* [8], endowed with the norm $||f||_{\mathcal{K}} = ||\frac{\partial^d f}{\partial x_1 ... \partial x_d}||_{L^2([0,1]^d)}$. One may interpret this space as the set of functions which have a d-th order crossed (weak) derivative $\frac{\partial^d f}{\partial x_1 ... \partial x_d}$ in $L^2([0,1]^d)$, vanishing on the "left" boundary (edges containing 0) of the unit d-dimensional cube.

Note that in dimension $d > 1$, this space differs from the Sobolev space H^1.

Extension of Example 1 to smooth functions One may extend the previous example to the space $\mathcal{S} = \mathcal{C}^p([0,1]^d)$ of p-times continuously differentiable functions, by choosing $\mathcal{H} = L^2([0,1]^d)$ and $J : \mathcal{H} \to \mathcal{S}$ defined by $(Jh)(s_0) = \int_{[0,s_0]} \int_{[0,s_1]} ... \int_{[0,s_p]} h(u) du ds_p ... ds_1$.

Then the covariance operator defined by:

$$(K\nu)(t_0) = \int_{[0,t_0],...,\times[0,t_p]} (J'\nu)(u) du dt_p ... dt_1,$$

with $(J'\nu)(u) = \int_{[0,1]^d} \mathbb{I}_{0\le u\le s_p\le ...\le s_0} \nu(ds_0) ds_1 ... ds_p$, corresponds to a Gaussian random object $W \sim \mathcal{N}(0,K)$ whose kernel space is $\mathcal{K} = J(L^2([0,1]^d))$.

Indeed, $\mathcal{S}'$ is the set of signed measures on the unit cube. Thus, for $\nu \in \mathcal{S}'$ and $h \in \mathcal{H}$, we have $(J'\nu, f) = \int_{[0,1]^d} (Jh)(s_0) \nu(ds_0)$, from which we deduce $K = JJ'$.

Note that when $p = 0$, we recover the previous case of Brownian sheets.

2.3 Expansion of a Gaussian object:

Let $(\varphi_i)_i$ be an orthonormal basis of $\mathcal{K}$ (for the inner product $\langle \cdot, \cdot \rangle_{\mathcal{K}}$). From Lemma 3, in the case of the alternative construction via the mapping J and the Hilbert space $\mathcal{H}$, one can build such an orthonormal basis with the functions $\varphi_i = Jh_i$ where $(h_i)_i$ is an orthonormal basis of $\mathcal{H}$.

We now define the expansion of a Gaussian object W (see [9]):

Lemma 4 *Let* $\{\xi_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)\}_{i \geq 1}$. *Then* $\sum_{i=1}^{\infty} \xi_i \varphi_i$ *is a Gaussian object, written* W, *with law* $\mathcal{N}(0, K)$. *It is called an* **expansion** *of* $W \sim \mathcal{N}(0, K)$.

Example 1 (continued) To build an expansion for the Brownian motions, we use the Haar basis of $L^2([0,1])$. It is defined by $h_{j,l}(x) = 2^{j/2} h(2^j x - l)$, where $h(x) = \mathbb{I}_{[0,1/2[} - \mathbb{I}_{[1/2,1[}$, together with $h_0(x) = \mathbb{I}_{[0,1]}(x)$. We deduce that a basis of the kernel space is obtained by the integrals of those functions (since $Jh(t) = \int_0^t h(s)ds$), i.e. which are the hat functions defined in the introduction: $\Lambda_{j,l} = Jh_{j,l}$, and $\Lambda_0 = Jh_0$.

Note that the rescaling factor inside $\Lambda_{j,l}$ naturally appears as $2^{-j/2}$, and not $2^{j/2}$ as usually defined in wavelet-like transformations.

In the sequel, we only consider the case of an orthogonal basis since this corresponds to the Examples 1 and 2 (described below), but note that orthogonality is actually not required (dictionaries of functions could be handled as well).

2.4 Expansion of the kernel space

We now give a characterization of the functions of the kernel space in terms of its basis. When $\mathcal{S}$ is a Hilbert space with *orthogonal* basis $(h_i)_i$ (thus $\varphi_i = h_i$ defines a basis of $\mathcal{K}$), the condition that I' is continuous (see Section 2.1) can be guaranteed by the condition that $\sum_i ||h_i||^2$ is finite. In such case, we have [9]:

$$\mathcal{K} = \{f \overset{\text{def}}{=} \sum_i \alpha_i \varphi_i, \sum_i \alpha_i^2 < \infty\}. \tag{3}$$

In the general case when $\mathcal{S}$ is not a Hilbert space, and we use the construction (via Lemma 3) of $\mathcal{K}$ via a mapping $J : \mathcal{H} \to \mathcal{S}$, where $\mathcal{H}$ is a Hilbert space with *orthonormal* basis $(h_i)_i$, the continuity of I' is ensured provided that $\sum_i ||\varphi_i||^2 < \infty$ where $\varphi_i = Jh_i$, in which case we also recover (3).

Note that we write $f = \sum_{i \geq 1} \alpha_i h_i$, where α_i are real coefficients, with the usual meaning that $\lim_F ||f - \sum_{i=1}^F \alpha_i h_i|| = 0$ in the least-squares sense. When $(h_i)_i$ is an orthogonal basis, the existence of such decomposition corresponds to the separability of the space, otherwise, it corresponds to $(h_i)_i$ being a Riesz basis.

Thus the kernel space $\mathcal{K}$ may be seen from two equivalent points of view: either as a set of functions that are expectations of some random process, or as a set of functions that are random linear combinations of the initial features. We will use both points of view to derive our results.

Example 2: Scrambled wavelets We now introduce a second example built from a family of orthogonal wavelets $(\tilde{\varphi}_{\varepsilon,j,l}) \in C^q([0,1]^d)$ (where ε is a multi-index, j is a scale index, l a multi-index) with at least $q > d/2$ vanishing moments. For $s \in (d/2, q)$, we define the so-called scrambled wavelets $W = \sum_{\varepsilon,j,l} \xi_{\varepsilon,j,l} \varphi_{\varepsilon,j,l}$, where $\xi_{\varepsilon,j,l} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $\varphi_{\varepsilon,j,l} = 2^{-js} \frac{\tilde{\varphi}_{\varepsilon,j,l}}{||\tilde{\varphi}_{\varepsilon,j,l}||^2}$ are the rescaled wavelets.

The term "scrambled wavelets" refers to the disorderly construction of this multi-resolution random process built from wavelets.

Lemma 5 *The Kernel space of the law of the scrambled wavelets is the Sobolev space* $H^s([0,1]^d)$.

Proof: The condition $s > d/2$ guarantees that $\sum_{\varepsilon,j,l} ||\varphi_{\varepsilon,j,l}||^2 = (2^d - 1) \sum_j 2^{-2js} 2^{jd} < \infty$, so that the scrambled wavelets are well defined and the corresponding mapping I' is continuous.

When the wavelet functions are $C^q([0,1]^d)$ with at least $q > s$ vanishing moments, it is known that the homogeneous Besov space $B_{s,\beta,\gamma}([0,1]^d)$ admits the following characterization (independent of the choice of the wavelets [5, 2]): $B_{s,\beta,\gamma} = \{f; ||f||_{s,\beta,\gamma} < \infty\}$ where

$$||f||_{s,\beta,\gamma} \overset{\text{def}}{=} \sum_{\varepsilon} \Big(\sum_j \Big[2^{j(s+d/2-d/\beta)} \big(\sum_{l_1 \cdots l_d = 0}^{2^j - 1} |\langle f, \tilde{\varphi}_{\varepsilon,j,l} \rangle|^\beta \big)^{1/\beta} \Big]^\gamma \Big)^{1/\gamma}$$

When the family of wavelets satisfies this condition, we say that it is adapted to the space $H^s([0,1]^d)$ for $s < q$. One example of such wavelets is given by Daubechies wavelets in dimension 1. For instance,

for $s = 1$, the Daubechies 3 wavelets with $3 > 1$ vanishing moments are Lipschitz of order $1,08 > s$, i.e. $C^1([0,1])$. For $s = 2$, we can consider Daubechies 10 wavelets with 10 vanishing moments (see [11]). The extension to dimension d is easy.

Thus, any $f \in B_{s,\beta,\beta}$ writes $f = \sum_{\varepsilon,j,l} \overline{\alpha}_{\varepsilon,j,l} \overline{\varphi}_{\varepsilon,j,l}$ with $\overline{\alpha}_{\varepsilon,j,l} = (f, \overline{\varphi}_{\varepsilon,j,l})$ and $\overline{\varphi}_{\varepsilon,h,l} = \frac{\tilde{\varphi}_{\varepsilon,h,l}}{||\tilde{\varphi}_{\varepsilon,h,l}||}$. Let us consider $\alpha_{\varepsilon,j,l} = 2^{j(s+d/2-d/\beta)}(f, \tilde{\varphi}_{\varepsilon,j,l})$. Then we deduce that $||\alpha||_\beta = ||f||_{s,\beta,\beta}$.

Thus if we introduce the rescaled wavelet functions $\varphi_{\varepsilon,j,l} = 2^{-j(s+1/2-1/\beta)} \frac{\tilde{\varphi}_{\varepsilon,j,l}}{||\tilde{\varphi}_{\varepsilon,j,l}||^2}$, any $f \in B_{s,\beta,\beta}$ writes $f = \sum_{\varepsilon,j,l} \alpha_{\varepsilon,j,l} \varphi_{\varepsilon,j,l}$ with $||\alpha||_\beta < \infty$. When $\beta = 2$, $B_{s,\beta,\beta}$ is $H^s([0,1]^d)$. Thus we deduce

$$H^s([0,1]^d) = \{f = \sum_{\varepsilon,j,l} \alpha_{\varepsilon,j,l} \varphi_{\varepsilon,j,l}; ||\alpha|| < \infty\},$$

which is also the definition of the kernel space of $W = \sum_{\varepsilon,j,l} \xi_{\varepsilon,j,l} \varphi_{\varepsilon,j,l}$. ■

This example explains how to handle Sobolev spaces $H^s([0,1]^d)$ of any order $s > d/2$.

Note that this result extends similarly to inhomogeneous Sobolev spaces $H_2^{\vec{s}}$, where $\vec{s} \in \mathbb{R}^d$ is a multi-index, via tensorisation of one dimensional Sobolev spaces (see [15]). In this case, if $s_i > 1/2$ for all $1 \leq i \leq d$, and $(\tilde{\varphi}_{j,l})_{j,l}$ is a wavelet basis of $H_2^{\vec{s}}([0,1]^d)$ (adapted to this space), then one can check that the kernel space of the law of Brownian (inhomogeneous) wavelets $\varphi_{j,l} = 2^{-\sum_{i=1}^d j_i s_i} \frac{\tilde{\varphi}_{j,l}}{||\tilde{\varphi}_{j,l}||^2}$ is $H_2^{\vec{s}}([0,1]^d)$.

3 Back to regression

Now we wish to understand when a target function f^* can be well approximated by functions in random subspaces. In this section we show that this is exactly when f^* belongs to the kernel space $\mathcal{K}$.

3.1 When kernel spaces are approximation spaces

We now interpret the functions of the random subspaces $\mathcal{G}_P$ as empirical estimates of functions of the kernel space $\mathcal{K}$ of W. This provides conditions on the coefficients of the linear combinations to guarantee the convergence of the random functions, when the dimension P of the random subspaces increases.

Let us consider P i.i.d. realizations of a Gaussian object W with law $\mathcal{N}(0,K)$, i.e. $W_1, ..., W_P \sim \mathcal{N}(0,K)$, together with their expansion on an orthonormal basis $(\varphi_i)_{i\geq 1}$ of the kernel space $\mathcal{K}$. Thus $W_p = \sum_{i\geq 1} \xi_{p,i}\varphi_i$ where $\xi_{p,i} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$.

We define the random features as $\psi_p \overset{\text{def}}{=} \sum_{i\geq 1} A_{p,i}\varphi_i = \frac{W_p}{\sqrt{P}}$, where $A_{p,i} \overset{\text{def}}{=} \frac{1}{\sqrt{P}}\xi_{p,i}$. Using abbreviated notations, we write $\psi_p(\cdot) = A_p\varphi(\cdot)$. Similarly, for any $\beta \in \mathbb{R}^P$, and $(\alpha_i)_{i\geq 1} \in l_2(\mathbb{N})$ we write $\beta = A\alpha$ to mean $\beta_p = \sum_{i\geq 1} A_{p,i}\alpha_i$, for $p = 1 \dots P$.

Define $\mathcal{G}_P$ as the span of the random features $(\psi_p)_{1\leq p\leq P}$, and $\mathcal{G}_P^0$ as the set of functions $g_{A\alpha} \in \mathcal{G}_P$ such that $f_\alpha \in \mathcal{K}$, i.e.:

$$G_P^0 \overset{\text{def}}{=} \{g_\beta = \sum_{p=1}^P \beta_p\psi_p; \text{ s.t. } \beta = A\alpha, \text{ and } ||\alpha|| < \infty\}.$$

Lemma 6 *$\mathcal{G}_P^0$ is the set of (functional) unbiased empirical estimates of functions in $\mathcal{K}$ w.r.t. the law of W (i.e. for $f_\alpha \in \mathcal{K}$, for all $x \in \mathcal{X}$, $\mathbb{E}g_{A\alpha}(x) = f_\alpha(x)$ and $\lim_{P\to\infty} g_{A\alpha}(x) = f_\alpha(x)$, almost surely).*

Proof: By definition of I, any function f_α of the kernel space $\mathcal{K}$ may be written $f_\alpha(\cdot) = \mathbb{E}(W(\cdot)z(W))$ for some $z \in \mathcal{S}'_\mathcal{N}$ (see Section 2.2) that satisfies $\alpha_i = z(\varphi_i)$. Thus, if we consider the empirical estimates $\widehat{g}_P(\cdot) = \frac{1}{P}\sum_{p=1}^P z(W_p)W_p(\cdot)$ and introduce $\beta_p = \frac{z(W_p)}{\sqrt{P}}$, then $\widehat{g}_P = \sum_{p=1}^P \beta_p\psi_p$. Now, by linearity of z, we have $\beta = A\alpha$ thus $\widehat{g}_P = g_{A\alpha}$. Since $f_\alpha = \sum_i z(\varphi_i)\varphi_i$, we deduce that $\widehat{g}_P \in \mathcal{G}_p^0$. Thus, for each $f_\alpha \in \mathcal{K}$, $\mathbb{E}(\widehat{g}_P)_P = f_\alpha$ and the Law of Large Numbers guarantees that for any $x \in \mathcal{X}$, the sequence $(\widehat{g}_P(x))_P$ converges a.s. to $f_\alpha(x)$ when $P \to \infty$.

Conversely, let $(\alpha_i)_{i\geq 1}$ be a sequence such that $||\alpha|| < \infty$. Then the sequence $(g_{A\alpha})_P$ converges since $z(W)W(\cdot) \in L^1(\mathcal{S}, \mathcal{N}(0,K))$, where the linear functional z is defined by $z(\varphi_i) = \alpha_i$. Indeed, this measurability property is ensured by definition of $\mathcal{S}'_\mathcal{N}$ whenever $\sum_i z^2(\varphi_i) < \infty$, i.e. $||\alpha|| < \infty$. Its limit is $\mathbb{E}g_{A\alpha}(\cdot) = \mathbb{E}[\frac{1}{P}\sum_{p=1}^P \sum_i \xi_{p,i}\alpha_i \sum_j \xi_{p,j}\varphi_j(\cdot)] = \sum_i \alpha_i\varphi_i(\cdot)$ which is in $\mathcal{K}$. ■

Note that this extends the case of finitely many linear combinations to the case of infinitely many linear combinations (the transformation $\beta = A\alpha$ generalizes the case of random matrices A that appears in [10]).

One can go one step further and characterize the variance of the empirical estimates $g_{A\alpha} \in G_P^0$. This provides a nice interpretation (discussed in Examples 1 and 2 after the lemma) of the product $||\alpha^*||^2 \sup_x ||\varphi(x)||^2$ that appears in the excess risk bounds of [10] (see e.g. equation (2) of Section 1).

Lemma 7 *The variance of the empirical estimate $g_{A\alpha} \in G_P^0$ of any function $f_\alpha = \sum_i \alpha_i \varphi_i \in \mathcal{K}$, is, at any $x \in \mathcal{X}$,*

$$\mathbb{V}(g_{A\alpha}(x)) = \frac{1}{P}(f_\alpha^2(x) + ||f_\alpha||_{\mathcal{K}}^2 ||\varphi(x)||^2) \leq \frac{2}{P}||f_\alpha||_{\mathcal{K}}^2 ||\varphi(x)||^2.$$

Proof: By definition, $\mathbb{V}(\widehat{g}_P(x)) = \frac{1}{P}(\mathbb{E}[(W(x)z(W))^2] - f^2(x))$, where $z(\varphi_i) = \alpha_i$. Thus, it is sufficient to consider the term: $\mathbb{E}[(W(x)z(W))^2] = \mathbb{E}[(\sum_i \xi_i \varphi_i(x) \sum_j \xi_j z(\varphi_j))^2] = \mathbb{E}[(\sum_i \sum_j \xi_i \xi_j (\gamma_{i,j}))^2]$, with $\gamma_{i,j} = \varphi_i(x) z(\varphi_j)$. From the fact that ξ_i are i.i.d. $\mathcal{N}(0,1)$ variables (thus $\mathbb{E}(\xi_i^4) = 3$), we deduce

$$\begin{aligned}
\mathbb{E}[(W(x)z(W))^2] &= \sum_i \sum_{k \neq i} \gamma_{i,i}\gamma_{k,k} + \sum_i \sum_{j \neq i} \gamma_{i,j}^2 + \sum_i \sum_{j \neq i} \gamma_{i,j}\gamma_{j,i} + \sum_i 3\gamma_{i,i}^2 \\
&= 2\sum_i \sum_j z(k_j) k_j(\cdot) z(\varphi_i) \varphi_i(\cdot) + \sum_i \sum_j z(k_j)^2 \varphi_i^2(\cdot), \\
&= 2f_\alpha^2(x) + ||f_\alpha||_{\mathcal{K}}^2 ||\varphi(x)||^2,
\end{aligned}$$

from which we conclude by using Cauchy-Schwarz for deriving the inequality. ■

Example 1 (continued) When one considers Brownian sheets for regression with a target function $f^* = \sum_i \alpha_i^* \varphi_i$ that lies in the Cameron-Martin space $\mathcal{K}$ (defined previously), then the term $||\alpha^*||^2 \sup_{x \in \mathcal{X}} ||\varphi(x)||^2$ that appears in (2) is bounded by $2^{-d}||f^*||_{\mathcal{K}}^2$. Indeed:

- Since the Haar basis $(h_i)_i$ is an orthonormal basis of $L_2([0,1]^d)$, we have $||f^*||_{\mathcal{K}}^2 = ||\frac{\partial^d f^*}{\partial x_1 ... \partial x_d}||_{L^2([0,1]^d)}^2$ which by definition is also $\sum_i (\alpha_i^*)^2 ||h_i||^2 = ||\alpha^*||^2$. Thus $||\alpha^*||^2 = ||f^*||_{\mathcal{K}}^2$.
- Remember that the functions $(\varphi_i)_i$ are the hat functions $\Lambda_{j,l}$. The mother hat function Λ satisfies $||\Lambda||_\infty \leq 1/2$. In dimension d, we consider the tensor product $\varphi_{j,l}$ of one-dimensional hat functions (thus j and l are multi-indices). Since the support of Λ is $[0,1]$, then for any $x \in [0,1]^d$, for all j there exists at most one $l(x) = l = (l_1, ..., l_d)$ such that $\varphi_{j,l}(x) = \Pi_{i=1}^d \Lambda_{j_i,l_i}(x_i) \neq 0$. Thus $||\varphi(x)||^2 = \sum_{j,l} \varphi_{j,l}^2(x) = \sum_j (\Pi_{i=1}^d \Lambda_{j_i,l_i}(x_i))^2 \leq \sum_j (2^{-\sum_{i=1}^d j_i/2} 2^{-d})^2 = \frac{2^{-2d}}{(1-2^{-1})^d} = 2^{-d}$. Thus $\sup_{x \in \mathcal{X}} ||\varphi(x)||^2 \leq 2^{-d}$.

Example 2 (continued) We now consider regression with scrambled wavelets. Assume that the mother wavelet $\tilde{\varphi}$ has compact support $[0,1]^d$ and is bounded by λ, and assume that the target function $f^* = \sum_i \alpha_i^* \varphi_i$ lies in the Sobolev space $H^s([0,1]^d)$ with $s > d/2$. Then $||\alpha^*||^2 ||\varphi(\cdot)||^2 \leq \frac{\lambda^{2d}(2^d-1)}{1-2^{-2(s-d/2)}} ||f^*||_{H^s}^2$. Indeed:

- We have $||\alpha^*||^2 = ||f^*||_{B_{s,2,2}}^2 = ||f^*||_{H^s}^2 = ||f^*||_{\mathcal{K}}^2$ (see Example 2 of Section 2.4).
- By definition, the rescaled wavelet are $\varphi_{\varepsilon,j,l}(x) = 2^{-js}\tilde{\varphi}_{\varepsilon,j,l}(x) = 2^{-js}2^{jd/2}\tilde{\varphi}_\varepsilon(2^j x - l)$, where $\tilde{\varphi}_\varepsilon(x) = \Pi_{i=1}^d \tilde{\varphi}_{\varepsilon_i}(x)$. Thus for all $x \in [0,1]^d$, by the assumption on the support on $\tilde{\varphi}$, $||\varphi(x)||^2 = \sum_\varepsilon \sum_j (2^{-js}2^{jd/2}\tilde{\varphi}_\varepsilon(2^j x_i - l_i))^2$, and by definition of λ, this is bounded by $\sum_\varepsilon \sum_j (2^{-j(s-d/2)}\lambda^d)^2 \leq (2^d - 1)\frac{\lambda^{2d}}{1-2^{-2(s-d/2)}}$ whenever $s > d/2$. Thus $||\varphi(x)||^2 \leq \frac{\lambda^{2d}(2^d-1)}{1-2^{-2(s-d/2)}}$.

Note that in the case of inhomogeneous Sobolev spaces, we would have instead: $||\alpha^*||^2 ||\varphi(\cdot)||^2 \leq \frac{\lambda^{2d}}{\Pi_{i=1}^d (1-2^{-2(s_i-1/2)})} ||f^*||_{H_2^{\vec{s}}}^2$.

3.2 Approximation error with random subspaces

We now provide an approximation result that generalizes Theorem 1 of [10] when one considers the approximation of functions $f \in \mathcal{K}$ by the random subspaces $\mathcal{G}_P$.

From the results of the previous paragraph, one knows that whenever $f_\alpha \in \mathcal{K}$, for any $x \in \mathcal{X}$, the empirical estimate $g_{A\alpha}(x)$ concentrates around $f_\alpha(x)$. The next result relies on the additional property that $g_{A\alpha}$ also concentrates around f_α in $||\cdot||_P$-norm (the full proof makes use of Johnson-Lindenstrauss Lemma and is provided in Appendix A).

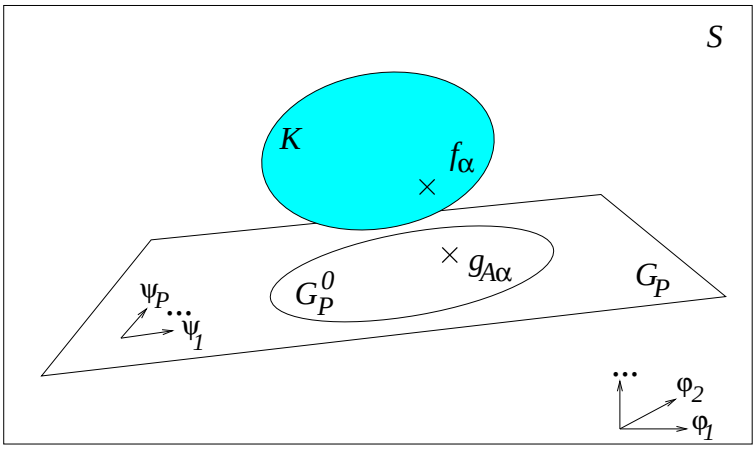

Figure 1: Schematic view of the kernel space $\mathcal{K}$, the space $\mathcal{G}_P$ spanned by the random features $(\psi_p)_{1\leq p\leq P}$, and $\mathcal{G}_P^0 = \{g_{A\alpha} \in \mathcal{G}_P, f_\alpha \in \mathcal{K}\}$. For illustration, in Example 2, $\mathcal{S} = L_2([0,1]^d)$, $\mathcal{K} = H^s([0,1]^d)$, the initial feature $(\varphi_i)_{i\geq 1}$ are the rescaled wavelets, and the random features are the scrambled wavelets.

Let $f_\alpha = \sum_i \alpha_i \varphi_i \in \mathcal{K}$. Write g^* the projection of f_α onto $\mathcal{G}_P$, i.e. $g^* = \arg\min_{g\in\mathcal{G}_P} ||f_\alpha - g||_P$, and $\bar{g}^* = T_L g^*$ its truncation at the threshold $L \geq ||f_\alpha||_\infty$. Notice that due to the randomness of the features $(\psi_p)_{1\leq p\leq P}$ of $\mathcal{G}_P$, the space $\mathcal{G}_P$ is also random, and so is $\bar{g}^*$. The following result provides bounds for the approximation error $||f_\alpha - \bar{g}^*||_P$ both in expectation and in high probability.

Theorem 8 *Whenever $P \geq 15\log P$ (i.e. $P \geq 20$) we have:*

$$E_{\mathcal{G}_P}\big[||f_\alpha - \bar{g}^*||_P^2\big] \leq \frac{8\log P}{P}\big((1+\sqrt{2})||\alpha||^2 \sup_{x\in\mathcal{X}} ||\varphi(x)||^2 + L^2\big).$$

For any $\eta > 0$, whenever $P/\eta \geq 15\log(P/\eta)$, we have with probability $1-\eta$ (w.r.t. the choice of the random subspace $\mathcal{G}_P$),

$$||f_\alpha - \bar{g}^*||_P^2 \leq \frac{8\log P/\eta}{P}\big((1+\sqrt{2})||\alpha||^2 \sup_{x\in\mathcal{X}} ||\varphi(x)||^2 + L^2\big).$$

3.3 Excess risk bounds

We now return to the initial regression problem and derive excess risk bounds for the two specific examples. The idea is to combine the approximation error bound from Theorem 8 with usual estimation error bounds for linear spaces (see e.g. [6]).

Let us consider a target function $f^* = \sum_i \alpha_i^* \varphi_i \in \mathcal{K}$. Remember that our prediction function $\widehat{g}$ is the truncation $\widehat{g} \stackrel{\text{def}}{=} T_L[g_{\widehat{\beta}}]$ of the (ordinary) least-squares estimate $g_{\widehat{\beta}}$ (empirical risk minimizer in the random space $\mathcal{G}_P$) defined by (1).

Note that since we consider an infinite number of initial features, the method considered here is not directly implementable. However, we will address in Section 5 practical considerations concerning the additional approximation error and the numerical complexity of an algorithmic implementation.

We now provide upper bounds (both in expectation and in high probability) on the excess risk for the least-squares estimate using random subspaces.

Theorem 9 *Whenever $P \geq 15\log(P)$, we have the following bound in expectation (w.r.t. all sources of randomness, i.e. input data, noise, and random features):*

$$\mathbb{E}_{G_P,X,Y}||f^* - \widehat{g}||_P^2 \leq c\max(\sigma^2, L^2)\frac{(\log N + 1)P}{N} + \frac{8\log P}{P}\big(8(1+\sqrt{2})||\alpha^*||^2 \sup_{x\in\mathcal{X}} ||\varphi(x)||^2 + L^2\big), \quad (4)$$

where c is a universal constant (see[6]).

Now, for any $\eta > 0$, whenever $P/\eta \geq 15\log(P/\eta)$, we have the following bound in high probability (w.r.t. the choice of the random features):

$$\mathbb{E}_{X,Y}||f^* - \widehat{g}||_P^2 \leq c\max(\sigma^2, L^2)\frac{(\log N + 1)P}{N} + \frac{8\log P/\eta}{P}\big(8(1+\sqrt{2})||\alpha^*||^2 \sup_{x\in\mathcal{X}} ||\varphi(x)||^2 + L^2\big). \quad (5)$$

Example 1: Regression with Brownian sheets. Consider a target function f^* that lies in the Cameron-Martin space $\mathcal{K}$ (equipped with the norm $||f||_{\mathcal{K}} \stackrel{\text{def}}{=} ||\frac{\partial^d f}{\partial x_1 \dots \partial x_d}||^2_{L_2}$). Then ordinary least-squares performed on the random subspaces spanned by P Brownian sheets has an expected excess risk

$$\mathbb{E}||f^* - \widehat{g}||^2_P = O\Big(\frac{\log N}{N} P + \frac{\log P}{P} ||f^*||^2_{\mathcal{K}}\Big),$$

(and a similar bound holds in high probability).

Indeed, this is a direct consequence of the previous Theorem and the results of Section 3.1 (which says that the kernel space of the law of the Brownian sheet is the Cameron-Martin space and that $||\alpha^*|| = ||f^*||_{\mathcal{K}}$).

Example 2: Regression with scrambled wavelets. Now consider a target function $f^* \in H^s([0,1]^d)$, with $s > d/2$. Similarly, ordinary least-squares performed on the random subspaces spanned by P scrambled wavelets (as defined previously) has an excess risk (as a direct consequence of Theorem 9 and the results of previous section):

$$\mathbb{E}||f^* - \widehat{g}||^2_P = O\Big(\frac{\log N}{N} P + \frac{\log P}{P} ||f^*||^2_{H^s}\Big).$$

In both examples, by choosing P of order $\sqrt{N}||f^*||_{\mathcal{K}}$, one deduces the excess risk

$$\mathbb{E}||f^* - \widehat{g}||^2_P = O\Big(\frac{||f^*||_{\mathcal{K}} \log N}{\sqrt{N}}\Big). \tag{6}$$

4 Discussion

Minimax optimality Note that although the rate $\tilde{O}(N^{-1/2})$ deduced from Theorem 9, Equation (6), does not depend on the dimension d of the input data $\mathcal{X}$, it does not contradict the known minimax lower bounds, which are $\Omega(N^{-2s/(2s+d)})$ for functions defined over $[0,1]^d$ that possess s-degrees of smoothness (e.g. that are s-times differentiable), see e.g. Chapter 3 of [6]. Indeed, the kernel space is composed of functions whose order of smoothness may depend on d. For illustration, in Example 2, the kernel space is the Sobolev space $H^s([0,1]^d)$ with $s > d/2$. Thus $2s/(2s+d) > 1/2$.

Notice that if one considers wavelets with q vanishing moments, where $q > d/2$, then one may choose s (such that $q > s > d/2$) arbitrarily close to $d/2$, and deduce that the excess risk rate $\tilde{O}(N^{-1/2})$ deduced from Theorem 9 is arbitrarily close to the minimax lower rate. Thus regression using scrambled wavelets is minimax optimal (up to logarithmic factors).

Now, about Example 1, we are not aware of minimax lower bounds for Cameron-Martin spaces, thus we do not know whether regression using Brownian sheets (Example 1) is minimax optimal or not.

Links with RKHS Theory There are strong links between the kernel space of Gaussian objects and Reproducing Kernel Hilbert Spaces (RKHS). We now remind two properties that illustrate those links:

- There is a bijection between Carleman operators and the set of RKHSs [4, 14]. 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 $\mathcal{H}$. From Lemma 3, a Carleman operator defines the kernel space of the law of the Gaussian object with covariance operator $K = JJ'$. This kernel space is also a RKHS with a kernel $k(s,t) = \langle \Gamma_s, \Gamma_t \rangle$. Now, conversely, from any kernel k, one can construct a Carleman operator J generating the RKHS [4].

- Expansion of a Mercer kernel. The expansion of a Mercer kernel k (i.e. when $\mathcal{X}$ is compact Haussdoff and k is a continuous kernel) is given by $k(x,y) = \sum_{i=1}^{\infty} \lambda_i e_i(x) e_i(y)$, where $(\lambda_i)_i$ and $(e_i)_i$ are the eigenvalues and eigenfunctions of the integral operator $L_k : L^2(\mathcal{X}, \mu) \to L^2(\mathcal{X}, \mu)$ defined by $(L_k(f))(x) = \int_{\mathcal{X}} k(x,y) f(y) d\mu(y)$. The associated RKHS is $\mathcal{K} = \{f = \sum_i \alpha_i \varphi_i; \quad \sum_i \alpha_i^2 < \infty\}$, where $\varphi_i = \sqrt{\lambda_i} e_i$, endowed with the inner product $\langle f_\alpha, f_\beta \rangle = \langle \alpha, \beta \rangle_{l_2}$. This space is thus also the kernel space of the Gaussian object $W = \sum_i \xi_i \varphi_i$.

The expansion of a Mercer kernel gives an explicit construction of the functions of the RKHS. However it may not be straightforward to compute the eigenvalues and eigenfunctions of the integral operator L_k and thus the basis functions φ_i in the general case.

The approach described in this paper (expansion of a Gaussian object via a linear injective mapping) enables to choose explicitly the initial basis functions, and build the corresponding kernel space. For example one may choose a multiresolution basis of a given Hilbert space (which is interesting for numerical implementations as described later), which is not obvious from the Mercer expansion. We believe that this main difference is of practical interest.

Note also that an alternative construction of a RKHS from a Gaussian object is given by [9]: provided $\mathcal{S}'$ contains Dirac measures, the kernel space $\mathcal{K}$ of a Gaussian measure $\mathcal{N}(0, K)$ on the space $\mathcal{S}$ is the RKHS with kernel k induced by K, i.e. defined as $k(s,t) = (\delta_s, K\delta_t)$ (for example the Cameron-Martin space considered in Example 1 is a RKHS).

Related works In [13, 12], 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 $(\theta_p)_{p\leq P} \overset{i.i.d}{\sim} \mu$, there exist coefficients $(c_p)_{p\leq P}$ such that $\widehat{f}(x) = \sum_{p=1}^P c_p\Phi(x,\theta_p)$ satisfies $||\widehat{f} - f||_2^2 \leq O(\frac{||f||_\mu}{\sqrt{P}})$. 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 (see Lemma 6). 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 law μ with the law of this infinite sequence. However, note that the condition $\sup_{x,\theta}\Phi(x,\theta) \leq 1$ does not hold in our setting and the fact that Θ is a set of infinite sequences makes the identification tedious without the Gaussian random functions theory. Anyway, we believe that this link provides a better mutual understanding of both approaches (i.e. [13] and the present paper).

In the work [1], the authors provide excess risk bounds for greedy algorithms (i.e. in a non-linear approximation setting). The bounds derived in their Theorem 3.1 is similar to the result stated in our Theorem 9. The main difference is that their bound makes use of the l_1 norm of the coefficients α^* instead of the l_2 norm in our setting. It would be interesting to further investigate 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.

5 Efficient implementation using a lazy multi-resolution expansion

In practice, in order to build the least-squares estimate, one needs to compute the values of the random features $(\psi_p)_p$ at the data points $(x_n)_n$, i.e. the matrix $\Psi = (\psi_p(x_n))_{p\leq P, n\leq N}$.

Due to finite memory and precision of computers, numerical implementations can only handle a finite number F of initial features $(\varphi_i)_{1\leq i\leq F}$. In [10] it was mentioned that the computation of Ψ, which makes use of the random matrix $A = (A_{p,i})_{p\leq P, i\leq F}$, has a complexity $O(FPN)$. However, thanks to our analysis, one can define the random features $(\psi_p)_{p\leq P}$ as an expansion of Gaussian objects built from a multi-resolution basis $(\varphi_{j,l})_{j,l} = (\varphi_i)_{i\leq F}$ of the underlying kernel space $\mathcal{K}$. This implies that, provided that these functions have compact support (such as the hat functions or the Daubechie wavelets), we can significantly speed up the computation of the matrix Ψ by using a *tree-based lazy expansion*, i.e. where the expansion is built only when needed for the evaluation at the points $(x_n)_n$.

Consider Example 2 (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 $O(2^dHPN)$.

Note that one may alternatively use the so-called sparse-grids instead of wavelet trees, which have been introduced by Griebel and Zenger (see [18, 3]). 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.

Now, using a finite F introduces an additional approximation (squared) error term in the final excess risk bounds or order $O(F^{-\frac{2s}{d}})$ for a wavelet basis adapted to $H^s([0,1]^d)$. This additional error (due to the numerical approximation) can be made arbitrarily small, e.g. $o(N^{-1/2})$, whenever $H \geq \frac{\log N}{d}$.

Thus, using $P = O(\sqrt{N})$ random features, we deduce that the complexity of building the matrix Ψ is $O(2^d N^{3/2}\log N)$ and solving the least squares system has a numerical cost $O(PN^2) = O(N^{5/2})$.

6 Conclusion and future works

We analyzed least-squares regression using sub-spaces $\mathcal{G}_P$ that are generated by P random linear combinations of infinitely many initial features. We showed that the kernel space $\mathcal{K}$ of the related Gaussian object provides an exact characterization of the set of target functions f^* for which this random regression works. We illustrated those results on two examples for which the kernel space is a known functional space. We derived a general approximation error result from which we deduced excess risk bounds of order $O(\frac{\log N}{N}P + \frac{\log P}{P}||f^*||_{\mathcal{K}}^2)$.

Example 2 shows that least-squares regression with scrambled wavelets provides rates that are arbitrarily close to minimax optimality. For Example 1, the relevant space of functions is the Cameron-Martin space, for which we are not aware of minimax lower bounds in dimension $d > 1$. The functions in that space need only to possess one d-th order derivative in L^2, thus this space is certainly larger than a Sobolev space of order d.

A limitation of Example 2 is that, so far, the scrambled wavelets are unable to consider Sobolev spaces of smoothness $s \leq d/2$. Possible directions for handling such spaces may involve developing a similar analysis for Besov spaces with $\beta < 2$ and fractional Brownian motions, which is the object of future works.

A nice property of using multiscale objects like Brownian sheets and scrambled wavelets is the possibility to design efficient numerical implementations. We described a method for computing the regression function with complexity $O(N^{5/2} + 2^d N^{3/2} \log N)$.

References

[1] Andrew Barron, Albert Cohen, Wolfgang Dahmen, and Ronald Devore. Approximation and learning by greedy algorithms. 36:1:64–94, 2008.

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

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

[4] Stéphane Canu, Xavier Mary, and Alain Rakotomamonjy. Functional learning through kernel. *arXiv*, 2009, October.

[5] M Frazier and B Jawerth. Decomposition of besov spaces. *Indiana University Mathematics Journal*, (34), 1985.

[6] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer-Verlag, 2002.

[7] Stéphane Jaffard. Décompositions en ondelettes. In *Development of mathematics 1950–2000*, pages 609–634. Birkhäuser, Basel, 2000.

[8] Svante Janson. *Gaussian Hilbert spaces*. Cambridge Univerity Press, Cambridge, UK, 1997.

[9] Mikhail A. Lifshits. *Gaussian random functions*. Kluwer Academic Publishers, Dordrecht, Boston, 1995.

[10] Odalric-Ambrym Maillard and Rémi Munos. Compressed Least-Squares Regression. In *NIPS 2009*, Vancouver Canada, 2009.

[11] Stephane Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1999.

[12] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In John C. Platt, Daphne Koller, Yoram Singer, Sam T. Roweis, John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *NIPS*. MIT Press, 2007.

[13] Ali Rahimi and Benjamin Recht. Uniform approximation of functions with random bases. 2008.

[14] S. Saitoh. *Theory of reproducing Kernels and its applications*. Longman Scientific & Technical, Harlow, UK, 1988.

[15] Winfried Sickel and Tino Ullrich. Tensor products of sobolev-besov spaces and applications to approximation from the hyperbolic cross. *J. Approx. Theory*, 161(2):748–786, 2009.

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

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

[18] C. 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 Mech., Kiel, 1990. Vieweg-Verlag.

A Proof of Theorem 8

Proof: In order to prove that $g_{A\alpha}$ concentrates around f_α in $||\cdot||_P$-norm, we sample auxiliary states $X'_1, \cdots, X'_J \overset{i.i.d.}{\sim} P$, and prove that on an event of high probability, we simultaneously have that:

- $||g_{A\alpha} - f_\alpha||_P^2$ is close to the empirical estimate $\frac{1}{J}\sum_{j=1}^J |g_{A\alpha}(X'_j) - f_\alpha(X'_j)|^2$ (thanks to Chernoff-Hoeffding inequality),
- each $g_{A\alpha}(X'_j)$ concentrates around $f_\alpha(X'_j)$ when P is large (thanks to a version of Johnson-Lindenstrauss Lemma which approximately preserves the inner-products).

We make explicit the corresponding probability spaces. Consider the probability space defined over the product sample space $\Omega_1 \times \Omega_2$, where Ω_1 consists of all the possible realizations of J states $X'_1, \ldots, X'_J$ drawn i.i.d. from $P_{\mathcal{X}}$, and Ω_2 is the set of all possible realizations of the random elements $(A_{p,i})_{1\leq p\leq P, i\geq 1}$ (which defines the random feature space G_P).

For any $\delta > 0$, set $\varepsilon^2 \stackrel{\text{def}}{=} \frac{8}{P}\log(4J/\delta)$. Thus for $P \geq 15\log(4J/\delta)$ we have $\varepsilon \leq 3/4$ thus $\varepsilon^2/4 - \varepsilon^3/6 \geq \varepsilon^2/8$ and $P \geq \frac{\log(4J/\delta)}{\varepsilon^2/4-\varepsilon^3/6}$.

We may thus apply a version of Johnson-Lindenstrauss (JL) Lemma for inner-products, deduced from JL Lemma for norms by polarization inequality (see Proposition 1 of [10]). This lemma is stated for vectors of finite dimension F, and can be extended in our setting to the case of infinite converging sequences. Indeed, it applies to the two truncated sequences $\overline{\alpha}_F = (\alpha_1, \ldots, \alpha_F)$ and $\overline{\varphi}_F = (\varphi_1(X'_j(\omega_1)), \ldots, \varphi_1(X'_j(\omega_1))_F)$ for any finite F, and the extension follows due to the converging properties of these two sequences w.r.t. the random elements $(A_{p,i})_{1\leq p\leq P, i\geq 1}$ and the measurability of the limit objects. Thus for any $\omega_1 \in \Omega_1$ (which defines $X'_1(\omega_1), \ldots, X'_J(\omega_1)$), there exists an event $\Omega_2(\omega_1) \subset \Omega_2$ of probability $\mathbb{P}(\Omega_2(\omega_1)) \geq 1-\delta$ such that for all $\omega_2 \in \Omega_2(\omega_1)$, for all $1 \leq j \leq J$,

$$|A(\omega_2)\alpha \cdot A(\omega_2)\varphi(X'_j(\omega_1)) - \alpha \cdot \varphi(X'_j(\omega_1))| \leq \varepsilon||\alpha||\,||\varphi(X'_j(\omega_1))|| \leq \varepsilon||\alpha|| \sup_{x\in\mathcal{X}} ||\varphi(x)||. \tag{7}$$

Let $\Omega = \{\{\omega_1\} \times \Omega_2(\omega_1), \omega_1 \in \Omega_1\}$. Obviously, $\mathbb{P}(\Omega) = \int_{\Omega_1} \mathbb{P}(\Omega_2(\omega_1))d\omega_1 \geq 1-\delta$. Now for each $\omega_2 \in \Omega_2$, define $\Omega_1(\omega_2) = \{\omega_1 \in \Omega_1, \{\omega_1\} \times \{\omega_2\} \in \Omega\} \subset \Omega_1$.

Let $\delta' > 0$. By Chernoff-Hoeffding's inequality, we have that for any $\omega_2 \in \Omega_2$ (which defines $A(\omega_2)$ and the random subspace $\mathcal{G}_P(\omega_2)$)), there exists an event $\Omega'_1(\omega_2) \subset \Omega_1(\omega_2)$ with probability $\mathbb{P}(\Omega'_1(\omega_2)|\Omega_1(\omega_2)) \geq 1-\delta'$ such that for all $\omega_1 \in \Omega'_1(\omega_2)$, we have

$$\begin{aligned}
&\mathbb{E}_{X\sim P_{\mathcal{X}}}|A(\omega_2)\alpha \cdot A(\omega_2)\varphi(X) - \alpha \cdot \varphi(X)|^2 \\
&\leq \frac{1}{J}\sum_{j=1}^{J} |A(\omega_2)\alpha \cdot A(\omega_2)\varphi(X'_j(\omega_1)) - \alpha \cdot \varphi(X'_j(\omega_1))|^2 + \varepsilon||\alpha||^2 \sup_{x\in\mathcal{X}} ||\varphi(x)||^2 \sqrt{\frac{\log(1/\delta')}{2J}}.
\end{aligned}$$

We deduce that under the event $\Omega'_1(\omega_2)$,

$$\begin{aligned}
\inf_{g\in\mathcal{G}_P(\omega_2)} ||f_\alpha - g||_P^2 &\leq ||f_\alpha - g_{A(\omega_2)\alpha}||_P^2 = \mathbb{E}_{X\sim P_{\mathcal{X}}}|A(\omega_2)\alpha \cdot A(\omega_2)\varphi(X) - \alpha \cdot \varphi(X)|^2 \\
&\leq \varepsilon^2||\alpha||^2 \sup_{x\in\mathcal{X}} ||\varphi(x)||^2 \Big(1 + \sqrt{\frac{\log(1/\delta')}{2J}}\Big)
\end{aligned} \tag{8}$$

Let us define $\Omega' = \{\Omega'_1(\omega_2) \times \{\omega_2\}, \omega_2 \in \Omega_2\} \subset \Omega$. Note that

$$\begin{aligned}
\mathbb{P}(\Omega') &= \int_{\Omega_2} \mathbb{P}(\Omega'_1(\omega_2))d\omega_2 \\
&= \int_{\Omega_2} \mathbb{P}(\Omega'_1(\omega_2))|\Omega_1(\omega_2))\mathbb{P}(\Omega_1(\omega_2))d\omega_2 \\
&\geq (1-\delta')\int_{\Omega_2} \mathbb{P}(\Omega_1(\omega_2))d\omega_2 = (1-\delta')\mathbb{P}(\Omega) \geq 1-(\delta+\delta')
\end{aligned}$$

Bound in expectation Notice that trivially $|f(x) - \bar{g}(\omega_2)(x)| \leq |f(x) - g(\omega_2)(x)|$. Thus (8) implies

$$\begin{aligned}
\mathbb{E}_{\omega_2}\big[||f - \bar{g}^*_{\omega_2}||_P^2\big] &\leq \int_{\Omega'} ||f - g^*_{\omega_2}||_P^2 d\omega_1 d\omega_2 + \int_{(\Omega')^c} ||f - \bar{g}^*_{\omega_2}||_P^2 d\omega_1 d\omega_2 \\
&\leq \varepsilon^2||\alpha||^2 \sup_{x\in\mathcal{X}} ||\varphi(x)||^2 \Big(1 + \sqrt{\frac{\log(1/\delta')}{2J}}\Big) + (2L)^2(\delta+\delta')
\end{aligned}$$

Now from the definition of ε, by setting $\delta = \delta' = (\log P)/P$ and $J = (\log P)/4$, we deduce that whenever $P \geq 15\log P$ we have:

$$E_{\omega_2}\big[||f - \bar{g}^*_{\omega_2}||_P^2\big] \leq \frac{8\log P}{P}\big((1+\sqrt{2})||\alpha||^2 \sup_{x\in\mathcal{X}} ||\varphi(x)||^2 + L^2\big).$$

Bound in high probability Define $\Lambda \stackrel{\text{def}}{=} \{\omega_2, \mathbb{P}(\Omega'_1(\omega_2)) \geq 1-\gamma\} \subset \Omega_2$. We have

$$\begin{aligned}\mathbb{P}(\Omega') &= \int_{\Omega_2} \mathbb{P}(\Omega'_1(\omega_2))d\omega_2 \\ &\leq \int_{\Omega_2} \mathbb{I}_{\mathbb{P}(\Omega'_1(\omega_2))\geq 1-\gamma} d\omega_2 + (1-\gamma)\int_{\Omega_2} \mathbb{I}_{\mathbb{P}(\Omega'_1(\omega_2))<1-\gamma} d\omega_2 \\ &\leq \mathbb{P}(\Lambda) + (1-\gamma)(1-\mathbb{P}(\Lambda)).\end{aligned}$$

Now, since $\mathbb{P}(\Omega') \geq 1-\delta-\delta'$, we have $\mathbb{P}(\Lambda) \geq 1 - \frac{\delta+\delta'}{\gamma}$.

We deduce that for $\eta = \frac{\delta+\delta'}{\gamma} > 0$, the event Λ has a probability $\mathbb{P}(\Lambda) \geq 1-\eta$, and for all $\omega_2 \in \Lambda$,

$$\begin{aligned}||f - \bar{g}^*_{\omega_2}||_P^2 &\leq \int_{\Omega'_1(\omega_2)} ||f - g^*_{\omega_2}||_P^2 d\omega_1 + \int_{(\Omega'_1(\omega_2))^c} ||f - \bar{g}^*_{\omega_2}||_P^2 d\omega_1 \\ &\leq \varepsilon^2 ||\alpha||^2 \sup_{x\in\mathcal{X}} ||\varphi(x)||^2 \Big(1 + \sqrt{\frac{\log(1/\delta')}{2J}}\Big) + (2L)^2 \frac{\delta+\delta'}{\eta}\end{aligned}$$

from (8). Now, by setting $\delta = \delta' = \eta/P\log(P/\eta)$ and $J = (1/4)\log(P/\eta)$, we deduce that whenever $P/\eta \geq 15\log(P/\eta)$ we have with probability $1-\eta$ (w.r.t. the random choice of A),

$$||f - \bar{g}^*||_P^2 \leq \frac{8\log P/\eta}{P}\big((1+\sqrt{2})||\alpha||^2 \sup_{x\in\mathcal{X}} ||\varphi(x)||^2 + L^2\big).$$

■

B Proof of Theorem 9

Proof: Like in the proof of Theorem 8 in Appendix A, we consider the probability space defined over the sample space $\Omega_1 \times \Omega_2$ where Ω_1 consists of the possible realizations of J states $X'_1, \ldots, X'_J$ drawn i.i.d. from $P_{\mathcal{X}}$, and Ω_2 is the set of all possible realizations of the random features $(\psi_p)_p$. Note that $X'_1, \ldots, X'_J$ are auxiliary random variable that should not be confused with the random training data $(x_n, y_n)_{n\leq N}$. Denote by $\mathbb{E}_{X,Y}$ the expectation w.r.t. the input data and the noise.

Since $\mathbb{P}(\Omega') \geq 1-(\delta+\delta')$ and $||f^* - \widehat{g}||_\infty \leq 2L$, we have

$$\mathbb{E}_{\omega_1,\omega_2,X,Y}||f^* - \widehat{g}(\omega_2)||_P^2 \leq \int_{\Omega'} \mathbb{E}_{X,Y}||f^* - \widehat{g}(\omega_2)||_P^2 + (2L)^2(\delta+\delta'). \tag{9}$$

For any $\omega_2 \in \Omega_2$, Theorem 11.3 of [6] says that

$$\mathbb{E}_{X,Y}||f^* - \widehat{g}(\omega_2)||_P^2 \leq c\max(\sigma^2, L^2)\frac{(\log N+1)P}{N} + 8\inf_{g\in\mathcal{G}_P(\omega_2)} ||f^* - g||_P^2,$$

which, combined with (8) and applied in (9) gives that whenever $P \geq 15\log(4J/\delta)$,

$$\begin{aligned}\mathbb{E}_{\omega_1,\omega_2,X,Y}||f^* - \widehat{g}(\omega_2)||_P^2 &\leq c\max(\sigma^2, L^2)\frac{(\log N+1)P}{N} \\ &\quad + \frac{64\log(4J/\delta)}{P}||\alpha||^2 \sup_{x\in\mathcal{X}} ||\varphi(x)||^2 \Big(1 + \sqrt{\frac{\log(1/\delta')}{2J}}\Big) + (2L)^2(\delta+\delta').\end{aligned}$$

Thus by setting $\delta = \delta' = (\log P)/P$ and $J = (\log P)/4$, we deduce that whenever $P \geq 15\log P$ we have the result in expectation (4).

Similarly, for the result in high probability, we deduce that with probability $1-\eta$,

$$\begin{aligned}\mathbb{E}_{X,Y}||f^* - \widehat{g}(\omega_2)||_P^2 &\leq c\max(\sigma^2, L^2)\frac{(\log N+1)P}{N} \\ &\quad + \frac{64\log(4J/\delta)}{P}||\alpha||^2 \sup_{x\in\mathcal{X}} ||\varphi(x)||^2 \Big(1 + \sqrt{\frac{\log(1/\delta')}{2J}}\Big) + (2L)^2\frac{\delta+\delta'}{\eta}.\end{aligned}$$

The result (5) follows by setting $\delta = \delta' = \eta/P\log(P/\eta)$ and $J = (1/4)\log(P/\eta)$.

■