Linear regression with random projections

Odalric-Ambrym Maillard, Rémi Munos

Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

Linear regression with random projections

Odalric-Ambrym Maillard ODALRIC.MAILLARD@INRIA.FR
Rémi Munos REMI.MUNOS@INRIA.FR
SequeL Project, INRIA Lille Nord Europe, France

Editor: Undefined

Abstract

We consider least-squares regression using a randomly generated subspace $\mathcal{G}_P \subset \mathcal{F}$ of finite dimension P, where $\mathcal{F}$ 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 $\mathcal{F}$ weighted by random Gaussian i.i.d. coefficients. In particular, we consider multi-resolution random combinations at all scales of a given mother function, such as a hat function or a wavelet. In this latter case, the resulting Gaussian objects are called *scrambled wavelets* and we show that they enable to approximate functions in Sobolev spaces $H^s([0,1]^d)$. As a result, given N data, the least-squares estimate $\hat{g}$ built from P scrambled wavelets has excess risk $||f^* - \hat{g}||^2_{\mathcal{P}} = O(||f^*||^2_{H^s([0,1]^d)}(\log N)/P + P(\log N)/N)$ for target functions $f^* \in H^s([0,1]^d)$ of smoothness order $s > d/2$. An interesting aspect of the resulting bounds is that they do not depend on the distribution $\mathcal{P}$ from which the data are generated, which is important in a statistical regression setting considered here. Randomisation enables to adapt to any possible distribution.

We conclude by describing an efficient numerical implementation using lazy expansions with numerical complexity $\tilde{O}(2^d N^{3/2} \log N + N^2)$, where d is the dimension of the input space.

1. Problem setting

In this paper, we consider the setting of linear regression with noise. We observe data points $(x_n)_{n \le N} \in \mathcal{X}$ and measurements $(y_n)_{n \le N} \in \mathbb{R}$ assumed to be independently and identically distributed (i.i.d.), from some distribution $\mathcal{P}$. The regression model is given by

$$x_n \sim \mathcal{P}_{\mathcal{X}} \quad \text{and} \quad y_n = f^*(x_n) + \eta_n$$

where f^* is the (unknown) target function we want to learn and η_n is a centred independent noise. We will assume that f^* is bounded by $||f^*||_\infty \le L$ and that the noise η_n has tail distribution controlled in such a way that $||\eta_n||_{\psi_2} \le C$ where ψ_2 is the Orlicz norm of order 2. Moreover, L and C are assumed to be known.

For a given class of functions $\mathcal{F}$, the goal is to return a regression function $\hat{f} \in \mathcal{F}$ that minimises the excess risk $||f^* - \hat{f}||_{\mathcal{P}}$ (where $||g||^2_{\mathcal{P}} = \mathbb{E}_{X \sim \mathcal{P}}[g(X)^2]$), that classically measures the closeness to optimality.

In this paper, we consider infinite dimensional spaces $\mathcal{F}$ that are generated by the span over a denumerable family of functions $\{\phi_i\}_{i \ge 1}$, called *initial features* (such as wavelets). Moreover, we will assume that $f^* \in \mathcal{F}$.

Since $\mathcal{F}$ is an infinite dimensional space, the empirical risk minimiser 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 minimises the empirical error $L_N(f) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{k=1}^N [y_n - f(x_n)]^2$ plus a penalty term, for example $\hat{f} = \arg\min_{f \in \mathcal{F}} L_N(f) + \lambda ||f||_p^p$ 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)).

In this paper we follow an alternative approach introduced in Maillard and Munos (2009), 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 minimiser 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 specific cases of infinite dimensional spaces $\mathcal{F}$ and provide a characterisation of the resulting approximation spaces.

2. Preliminary theory of Gaussian Objects

In this Section we give an interpretation of the random features in terms of random processes using the notion of Gaussian objects (Section 2.1). This enables us to analyse the corresponding limit object when the dimension of the initial feature space F is infinite. We also define the kernel space $\mathcal{K}$ of a gaussian object (Section 2.2), and its expansion (Section 2.5) so as to determine the spaces generated by the random features.

2.1 Gaussian objects

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

Definition 1 (Gaussian objects) *A random $W\in\mathcal{S}$ is called a Gaussian object if for all $\nu\in\mathcal{S}'$, then (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 centred object W, one may naturally define a space $\mathcal{K}\subset\mathcal{S}$ called the **kernel space** of the law $\mathcal{N}(0,K)$. 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 Lifshits (1995) or Janson (1997) 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 we 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 defined to be the closure of the image of $\mathcal{S}'$ by I' (in the L^2 sense). Note that this 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 $\nu_n\in\mathcal{S}'$).

Now, provided that I' is continuous (see Section 2.3 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 $(Iz)(x)=\mathbb{E}_W(W(x)z(W))$.

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

2.3 Application to Hilbert spaces

Let $\mathcal{S}$ be a Hilbert space and $(\phi_i)_i$ an orthonormal basis, then consider $\xi_i \sim \mathcal{N}(0,1)$ i.i.d. and positive coefficients $\sigma_i \geq 0$. Assuming that $\sum_i \sigma_i^2 < \infty$, we define the Gaussian object $W = \sum_i \xi_i \sigma_i \phi_i$, and we want to know what is the kernel of the law of W.

To this aim we identify the functions I' and I. Since S is a Hilbert space, then $\mathcal{S}' = \mathcal{S}$, thus we can consider $f = \sum_i \alpha_i \phi_i \in \mathcal{S}'$. For such an f, the injection mapping is given by $(I'f)(g) = \sum_i \alpha_i (g, \phi_i)$, and we also deduce that

$$||I'f||^2_{\mathcal{S}'_{\mathcal{N}}} = \mathbb{E}((I'f, X)^2) = \mathbb{E}(\sum_i \sigma_i \xi_i \alpha_i)^2 = \sum_i \sigma_i^2 \alpha_i^2$$

Note that since $||f||_{\mathcal{S}} = ||\alpha||_2$, the continuity of I' is insured by assumption on $(\sigma_i)_i$. Now one can easily check that the kernel space of the law of W is

$$\mathcal{K} = \{f = \sum_i \alpha_i \phi_i; \sum_i (\frac{\alpha_i}{\sigma_i})^2 < \infty\}$$

endowed with inner product $(f,g)_{\mathcal{K}} = \sum_i \frac{\alpha_i \beta_i}{\sigma_i^2}$.

Example 2: Scrambled wavelets A direct consequence of this characterisation is the possibility to handle Sobolev spaces.

Indeed, let us consider an orthonormal family of wavelets given by $(\tilde{\phi}_i)_i = (\tilde{\phi}_{\epsilon,j,l}) \in C^q([0,1]^d)$ (where ϵ is a multi-index, j is a scale index, l a multi-index). Then 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 characterisation (independent of the choice of the wavelets Frazier and Jawerth (1985); Bourdaud (1995)): $B_{s,\beta,\gamma} = \{f; ||f||_{s,\beta,\gamma} < \infty\}$ where

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

In particular, one can see that if $f = \sum_i \tilde{\alpha}_i \tilde{\phi}_i$, then $||f||^2_{s,2,2} = \sum_i (\frac{\tilde{\alpha}_i}{\sigma_i})^2$ where $\sigma_i = \sigma_{\epsilon,j,l} = 2^{-js}$. Thus, under the assumption that $\sum_i \sigma_i^2 = (2^d - 1) \sum_j 2^{-2js} 2^{jd} < \infty$, i.e. $s > d/2$, we can define the gaussian object $W = \sum_i \xi_i \phi_i$, where $\phi_{\epsilon,j,l} = 2^{-js} \tilde{\phi}_{\epsilon,j,l}$, and the kernel space of the law of W is exactly the Besov space $B_{s,2,2}$, which is also the Sobolev space $H_2^s([0,1]^d)$. We name the Gaussian object W a scrambled wavelet in order to refer to the disorderly construction of this multi-resolution random process built from wavelets.

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 Sickel and Ullrich (2009)). In this case, if $s_i > 1/2$ for all $1 \leq i \leq d$, and $(\tilde{\phi}_{j,l})_{j,l}$ is an orthonormal 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 $\phi_{j,l} = 2^{-\sum_{i=1}^d j_i s_i} \tilde{\phi}_{j,l}$ is $H_2^{\vec{s}}([0,1]^d)$.

2.4 Equivalent construction of the kernel space.

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

Lemma 2 *Lifshits (1995) 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* Janson (1997), endowed with the norm $||f||_{\mathcal{K}} = ||\frac{\partial^d f}{\partial x_1 \ldots \partial x_d}||_{L^2([0,1]^d)}$. One may interpret this space as the set of functions which have a d-th order crossed (weak) derivative $\frac{\partial^d f}{\partial x_1 \ldots \partial x_d}$ 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.

2.5 Expansion of a Gaussian object:

Let $(\phi_i)_i$ be an orthonormal basis of $\mathcal{K}$ (for the inner product $\langle \cdot, \cdot \rangle_{\mathcal{K}}$). From Lemma 2, 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 $\phi_i = Jh_i$ where $(h_i)_i$ is an orthonormal basis of $\mathcal{H}$.

We now define the expansion of a Gaussian object W (see Lifshits (1995)):

Lemma 3 *Let* $\{\xi_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)\}_{i \geq 1}$. *Then* $\sum_{i=1}^{\infty} \xi_i \phi_i$ *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, but note that orthogonality is actually not required (dictionaries of functions could be handled as well).

3. Regression with random subspaces

In this section, we describe the construction of the random subspace $\mathcal{G}_P \subset \mathcal{F}$ that is generated from the span of features $(\psi_p)_{p \leq P}$ that are randomly generated from the initial features $(\phi_i)_{i \geq 1}$. 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 $\mathcal{F}$. Ensuring that the randomly generated features $(\psi_p)_{p \leq P}$ are well defined in this case is indeed not trivial and makes use of some results of the theory of Gaussian Objects (see Section 2).

The next subsection is devoted to the analysis of approximation power of the the random features. We then describe the algorithm that builds the proposed regression function and then provide excess risk bounds for this algorithm. Eventually, we discuss some asymptotic behaviour of the random subspace $\mathcal{G}_P$, when P tends to $+\infty$, so as to provide more intuition about the terms that appear in the performance bounds.

3.1 Construction of random subspaces

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

$$\sup_{x\in\mathcal{X}} ||\phi(x)||^2 < \infty, \text{ where } ||\phi(x)||^2 \stackrel{\text{def}}{=} \sum_{i\geq 1} \phi_i(x)^2. \tag{1}$$

Examples of feature spaces satisfying this property include rescaled wavelets as described in Section 2.3.

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

$$\psi_p(x) \stackrel{\text{def}}{=} \sum_{i\geq 1} A_{p,i}\phi_i(x), \text{ for } 1 \leq p \leq P \tag{2}$$

where the (infinitely many) coefficients $A_{p,i}$ are drawn i.i.d. from a centred 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 called an expansion of a Gaussian object) and we refer to the Section 2 for elements of theory about Gaussian objects and Lemma 3 for the expansion of a Gaussian object. It is shown that under assumption (1), the random features are well defined. Actually, they are random samples of a centred Gaussian process indexed by the space $\mathcal{X}$ with covariance structure given by $\frac{1}{P}\langle\phi(x), \phi(x')\rangle$, where we used the notation $\langle u, v\rangle = \sum_i u_i v_i$ for two square-summable sequences u and v. Indeed, $\mathbb{E}_{A_p}[\psi_p(x)] = 0$, and

$$\mathrm{Cov}_{A_p}(\psi_p(x), \psi_p(x')) = \mathbb{E}_{A_p}[\psi_p(x)\psi_p(x')] = \frac{1}{P}\sum_{i\geq 1}\phi_i(x)\phi_i(x') = \frac{1}{P}\langle\phi(x), \phi(x')\rangle$$

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.

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

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

3.2 Approximation error of random projections

In this section, we analyse the approximation error of the random projection method. Thus, we are interested in bounding $\inf_{g\in\mathcal{G}} ||f^* - T_L(g)||^2_{\mathcal{P}}$ in high probability with respect to any source of randomness.

Thanks to the properties of random projections, we can derive the following statement, the proof of which is reported in Appendix A.

Theorem 4 (Approximation error) *With probability higher than $1-\delta$ with respect to any source of randomness, the following approximation error bound holds true:*

$$\inf_{g\in\mathcal{G}} ||f^* - T_L(g)||^2_{\mathcal{P}} \leq \inf_{m:\log(8m/\delta)\leq P/15} \frac{8\log(8m/\delta)}{P}||\alpha^*||^2 \sup_x ||\phi(x)||_2^2 + (2L)^2\sqrt{\frac{\log(2/\delta)}{2m}}$$

Moreover, the following bound also holds true, provided that $P \geq \frac{\log(4)}{(1-\frac{4\gamma}{3})\gamma^2}$:

$$\mathbb{E}_{\mathcal{G}_P}[\inf_{g\in\mathcal{G}} ||f^* - T_L(g)||^2_{\mathcal{P}}] \leq \frac{10||\alpha^*||_2^2\mathbb{E}(||\phi(X)||_2^2)}{P}\frac{1}{1-\frac{4\gamma}{3}}.$$

where we introduced the quantity $\gamma = \frac{L}{||\alpha|| \sup_x ||\phi(x)||_2}$

Then we deduce the following corollary, after optimisation:

Corollary 5 *Provided that* $P \geq 30 \log(\frac{P\gamma^2}{2}\sqrt{\log(2/\delta)/\delta})$*, then with probability higher than* $1-\delta$ *with respect to any source of randomness, the following holds true:*

$$\inf_{g \in \mathcal{G}} ||f^* - T_L(g)||_{\mathcal{P}}^2 \leq 16 \frac{||\alpha||^2 \sup_x ||\phi(x)||_2^2}{P} \big(1 + \log(\frac{P\gamma^2}{2}\sqrt{\log(2/\delta)/\delta})\big)$$

Proof Indeed, optimising the right hand term without constrained gives the following optimal value for m:

$$m_{opt} = m = \frac{P^2 L^4 \log(2/\delta)}{2^5 ||\alpha||^4 \sup_x ||\phi(x)||_2^4}$$

We thus deduce that provided that $P \geq 30 \log(\frac{P}{2}\sqrt{\log(2/\delta)/\delta}\frac{L^2}{||\alpha||^2 \sup_x ||\phi(x)||_2^2})$, then with probability higher than $1-\delta$ with respect to any source of randomness, the following holds true:

$$\inf_{g \in \mathcal{G}} ||f^* - T_L(g)||_{\mathcal{P}}^2 \leq 16 \frac{||\alpha||^2 \sup_x ||\phi(x)||_2^2}{P} \big(1 + \log(\frac{P\sqrt{\log(2/\delta)/\delta} L^2}{2||\alpha||^2 \sup_x ||\phi(x)||_2^2})\big)$$

■

Note that the rate $||\alpha||_2^2 \sup_x ||\phi(x)||_2^2$ given here is not tight. We can actually replace it with $||\alpha||_2^2(\mathbb{E}(||\phi(X)||_2^2) + (\sup_x ||\phi(x)||_2)^2\sqrt{\frac{\log(1/\delta')}{2m}})$ using Hoeffding's inequality, which leads to the following improved bound:

Corollary 6 *Provided that* $P \geq 30 \log(P\gamma^2\sqrt{\log(4/\delta)/2\delta})$*, then with probability higher than* $1-\delta$ *with respect to any source of randomness, the following holds true:*

$$\inf_{g \in \mathcal{G}} ||f^* - T_L(g)||_{\mathcal{P}}^2 \leq \big[\frac{2^4||\alpha||^2 \mathbb{E}(||\phi(x)||_2^2)}{P} + \frac{2^7||\alpha||^2 \sup_x ||\phi(x)||_2^2}{\gamma^2 P^2}\big]\big(1 + \log(P\gamma^2\sqrt{\log(2/\delta)/2\delta})\big).$$

Intuitively, when $L \sim ||\alpha||_2 \sup_x ||\phi(x)||_2$, this result says that, provided that P is larger than a constant (depending on δ), then with probability higher than $1-\delta$ the following result holds

$$\inf_{g \in \mathcal{G}} ||f^* - T_L(g)||_{\mathcal{P}}^2 \leq O(\frac{||\alpha||^2 \mathbb{E}||\phi(X)||_2^2 \log(P/\delta)}{P}).$$

Note that even assuming $\sup_x ||\phi(x)||_2 < \infty$ is not needed. Indeed the careful reader may have seen that a weaker assumption still strong enough in order to derive high probability bounds would be to control the tail distribution of the same quantity, by assumption such that $||||\phi(x)||_2^2 - \mathbb{E}(||\phi(x)||_2^2||_{\psi_\alpha} \leq \sum_k ||\phi_k^2(X) - \mathbb{E}(\phi_k^2(X))||_{\psi_a} < \infty$ where ψ_a is the Orlicz norm of order $a \in [1,2]$. Such an assumption will only modify the second order terms (i.e. in P^2) of Corollary 6, at the price of poorer readability.

3.3 Regression algorithm.

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

$$g_{\hat{\beta}} = \arg \min_{g_\beta \in \mathcal{G}_P} L_N(g_\beta), \tag{3}$$

and is the solution of a least-squares regression problem, i.e. $\hat{\beta} = \Psi^{\dagger} Y \in \mathbb{R}^P$, 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 $\hat{g}(x)$ is the truncation (to the threshold $\pm L$) of $g_{\hat{\beta}}$, i.e. $\hat{g}(x) \stackrel{\text{def}}{=} T_L[g_{\hat{\beta}}(x)]$, where $T_L(u) \stackrel{\text{def}}{=} \begin{cases} u & \text{if } |u| \leq L, \\ L \text{ sign}(u) & \text{otherwise.} \end{cases}$

In next subsection 3.4, we provide excess risk bounds w.r.t f^* in $\mathcal{G}_P$.

3.4 Excess risk of the random projection estimator

In this section, we analyse the excess risk of the random projection method. Thus, we are interested in bounding $||f^* - T_L(g_{\hat{\beta}})||^2_{\mathcal{P}}$ in high probability with respect to any source of randomness.

The main result of this section is the following Theorem.

Theorem 7 *Assuming that $P \geq 15 \log(24n/\delta)$, then with probability higher than $1 - \delta$ with respect to any source of randomness, the following holds true:*

$$\begin{aligned} ||f^* - T_L(g_{\hat{\beta}})||^2_{\mathcal{P}} &\leq \frac{64 \log(12n/\delta)}{P} ||\alpha||_2^2 \frac{1}{n} \sum_{i=1}^{n} ||\phi(X_i)||_2^2 \\ &+ \frac{16C^2}{c} \frac{(253P + 145 \log(18/\delta))}{n} \\ &+ (24L)^2 \frac{2P \log(n) + 4 \log(9/\delta)}{n} \Big), \end{aligned}$$

where $||\eta||_{\psi_2} \leq C$ and c is the universal constant of Lemma 19.

Assuming that $P \geq 15 \log(8n/\delta)$, then with probability higher that $1 - \delta_G$ with respect to the choice of the random supspace $\mathcal{G}_P$,

$$\begin{aligned} \mathbb{E}_{X,Y}[||f^* - T_L(g_{\hat{\beta}})||^2_{\mathcal{P}}] &\leq \frac{8 \log(4n/\delta_G)}{P} ||\alpha^*||_2^2 \mathbb{E}_X(||\phi(X)||_2^2) \\ &+ \frac{2C^2}{cn}(253P + 258) \\ &+ \frac{(24L)^2}{n}(2P \log(n) + 6). \end{aligned}$$

Moreover, assuming that $P \geq 12 \log(4n)$, we also have the following upper bound, in expectation w.r.t. any source of randomness,

$$\begin{aligned} \mathbb{E}_{\mathcal{G},X,Y}[||f^* - T_L(g_{\hat{\beta}})||^2_{\mathcal{P}}] &\leq \frac{12(\log(4n) + 1)}{P} ||\alpha^*||_2^2 \mathbb{E}_{X,Y}(||\phi(X)||_2^2) \\ &+ \frac{2C^2}{cn}(253P + 258) \\ &+ \frac{(24L)^2}{n}(2P \log(n) + 6). \end{aligned}$$

Note that this bound can be seen in the more readable way as the sum of three different terms:

$$||f^* - T_L(g_{\hat{\beta}})||^2_{\mathcal{P}} \leq O\Big(\frac{\log(n/\delta)}{P} ||\alpha^*||_2^2 \frac{1}{n} \sum_{i=1}^{n} ||\phi(X_i)||_2^2 + C^2 \frac{P + \log(1/\delta)}{n} + L^2 \frac{P \log(n) + \log(1/\delta)}{n}\Big)$$

In order to prove this Theorem, we first decompose the excess risk $||f^* - T_L(g_{\hat{\beta}})||^2_{\mathcal{P}}$ of the estimate built using random projections into three different terms:

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

Lemma 8 *The following decomposition holds,*

$$||f^* - T_L(g_{\hat{\beta}})||_{\mathcal{P}}^2 \quad \leq \quad ||f^* - g_{\tilde{\beta}}||_n^2 + ||g_{\tilde{\beta}} - g_{\hat{\beta}}||_n^2 + \sup_{g \in \mathcal{G}} \left(||f^* - T_L(g)||_{\mathcal{P}}^2 - ||f^* - T_L(g)||_n^2 \right)$$

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

Remark 9 *The attentive reader aware about classical proof techniques for regression would have noticed that we do not use at this point the standard decomposition of the excess risk in terms of the sum of the approximation error of the class $\mathcal{G}$ in $L_2(\mathcal{P})$ norm plus the estimation and noise error coming from the finiteness of the sample. The reason is that without any further assumption on the features, we cannot bound the approximation error of the class $\mathcal{G}$ in $L_2(\mathcal{P})$ norm but only the approximation error of the truncated class $T_L(\mathcal{G})$ in $L_2(\mathcal{P})$ norm, that unfortunately does not seem to be useful for controlling the excess risk. Thus, the above decomposition makes use of the approximation error of the class $\mathcal{G}$ in empirical l_2^n norm.*

Note also that the same modification of the proof replacing $\sup_x ||\phi(x)||_2^2$ with $\mathbb{E}(||\phi(X)||_2^2)$ plus some small order term can be performed here. We now discuss each of these three terms separately. The proof of Proposition 10,11 and 12 are reported in Appendix B.

3.4.1 Approximation term

The first term, $||f^* - g_{\tilde{\beta}}||_n^2$, is an approximation error term in empirical norm, it makes appear the number of projections as well as the norm of the target function. This terms plays the equivalent of the approximation term that exists for regression with penalization by a factor $\lambda||f||^2$.

We can prove the first bound on the approximation term using the fact that by definition

$$||f^* - g_{\tilde{\beta}}||_n^2 \leq ||f_{\alpha^*} - g_{A\alpha^*}||_n^2$$

and the following proposition:

Proposition 10 *Assuming that $P \geq 15 \log(4n/\delta)$, then with probability higher than $1-\delta$ w.r.t. the gaussian random variables and random sample $(X_i)_{i=1..n}$, we have that:*

$$||f_\alpha - g_{A\alpha}||_n^2 \leq \frac{8 \log(4n/\delta)}{P} ||\alpha||_2^2 \frac{1}{n} \sum_{i=1}^n ||\phi(X_i)||_2^2$$

3.4.2 Noise 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 $\frac{P}{n}$, but note that this dependency with P may change with the knowledge on the functions ψ (for instance, if G is included in a Sobolev space of order s, we would have $P^{1/2s}$ instead).

We may consider different assumptions on the noise term. The assumption that leads to the most straightforward proof is that the noise is bounded $||\eta||_\infty \leq L$. Other classical assumption is that the noise has finite variance σ^2. We here consider and assumption saying that the tail of the distribution of the noise behaves nicely, like for instance that $||\eta||_{\psi_\alpha} \leq C$, where ψ_α is the Orlicz norm or order α. We may consider for instance either that $||\eta||_{\psi_1} \leq C$ or that $||\eta||_{\psi_2} \leq C$ but in order to improve readability of the final bound, we now assume that $||\eta||_{\psi_2} \leq C$.

Proposition 11 *Assuming that* $||\eta||_{\psi_2} \leq C$*, then with probability higher than* $1-\delta$ *w.r.t. any source of randomness, the following holds true:*

$$||g_{\tilde{\beta}} - g_{\hat{\beta}}||_n^2 \leq \frac{2C^2}{c} \frac{(253P + 145\log(6/\delta))}{n}$$

where c *is the universal constant defined in Lemma 19.*

3.4.3 Estimation term

The third term, $\sup_{g\in\mathcal{G}}(||f^* - T_L(g)||_{\mathcal{P}}^2 - ||f^* - T_L(g)||_n^2)$, is an estimation 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. We bound the third term by applying Theorem 11.2 of Györfi et al. (2002) to the class of functions $\mathcal{G}^0 = \{f^* - T_L(g), g \in \mathcal{G}\}$, for fixed random gaussian variables. Note that for all $f \in \mathcal{G}^0$, $||f||_\infty \leq 2L$.

Proposition 12 *Assuming that* $n\log(n) \geq \frac{4}{P}$*, then with probability higher than* $1-\delta$ *w.r.t. all sources of randomness, the following holds true:*

$$\sup_{g\in\mathcal{G}} ||f^* - T_L(g)||_{\mathcal{P}}^2 - 8||f^* - T_L(g)||_n^2 \leq (24L)^2 \frac{4\log(3/\delta) + 2P\log(n)}{n}$$

Note that this bound may be improved further, without loosing the constant 8. This would result in a bound of order $\max(\frac{c}{n}, \sqrt{\frac{c'}{n}})$, for some c, c', that has to be compared to the speed of order $\frac{1}{n}$ one could get on the quantity $\sup_{g\in\mathcal{G}}(||f^* - T_L(g)||_{\mathcal{P}}^2 - 2||f^* - T_L(g)||_n^2)$, at the price of loosing "just" a constant factor 2. Actually, it appears that due to the fact we loose this constant factor, the bound that scales as $\frac{1}{n}$ is indeed *less* tight that the one that scales with $\max(\frac{c}{n}, \sqrt{\frac{c'}{n}})$ (so $\frac{1}{\sqrt{n}}$ for large n). This phenomenon is not a consequence of using random projections and is a general fact in the non asymptotic setting. However, we do not consider such refinements here for clarity.

4. Application to specific function spaces

In this section, we now consider application of Theorem 7 to specific function spaces. We first consider the case of Brownian motions that is linked to Cameron-Martin spaces, and then the case of Scrambled Wavelets that enables to handle Sobolev spaces.

4.1 Cameron-Martin spaces

Regression with Brownian motions When one considers Brownian sheets for regression with a target function $f^* = \sum_i \alpha_i^* \phi_i$ that lies in the Cameron-Martin space $\mathcal{K}$ (defined previously), then we have:

Lemma 13 *Assume that* f^* *lies in the the Cameron-Martin space* $\mathcal{K}$*, then the following bound holds true:*

$$||\alpha^*||^2 \sup_{x\in\mathcal{X}} ||\phi(x)||^2 \leq 2^{-d}||f^*||_{\mathcal{K}}^2.$$

Proof Since the Haar basis $(h_i)_i$ is an orthonormal basis of $L_2([0,1]^d)$, we have $||f^*||_{\mathcal{K}}^2 = ||\frac{\partial^d f^*}{\partial x_1 \ldots \partial x_d}||_{L^2([0,1]^d)}^2$ which by definition is also $\sum_i (\alpha_i^*)^2 ||h_i||^2 = ||\alpha^*||^2$. Thus

$$||\alpha^*||^2 = ||f^*||_{\mathcal{K}}^2.$$

Now remember that the functions $(\phi_i)_i$ 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 $\phi_{j,l}$ 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 $\phi_{j,l}(x) = \Pi_{i=1}^d \Lambda_{j_i,l_i}(x_i) \neq 0$. Thus $||\phi(x)||^2 = \sum_{j,l} \phi_{j,l}^2(x) = \sum_j (\Pi_{i=1}^d \Lambda_{j_i,l_i}(x_i))^2 \leq \sum_j (2^{-\sum_{i=1}^d j_i/2} 2^{-d})^2 = \frac{2^{-2d}}{(1-2^{-1})^d} = 2^{-d}$, and the result follows. ■

Thus, from Theorem 7, ordinary least-squares performed on random subspaces spanned by P Brownian sheets has an expected excess risk

$$\mathbb{E}_{G_P,X,Y} ||f^* - \hat{g}||_{\mathcal{P}}^2 = O\Big(\frac{\log N}{N} P + \frac{\log N}{P} ||f^*||_{\mathcal{K}}^2\Big), \tag{4}$$

(and a similar bound holds in high probability).

4.2 Sobolev spaces

Construction of adapted wavelets When we consider wavelet functions that are $C^q([0,1]^d)$ with at least $q > s$ vanishing moments, we have proved that provided $s > d/2$, the Kernel space of the law of the associated scrambled wavelets is the homogeneous Sobolev space $H^s([0,1]^d)$.

We now provide practical construction of such wavelets. One easy example 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 Mallat (1999)). The extension to dimension d is easy.

However, one may want an easier way to control the Lipschitz order of the wavelets, i.e. having wavelets such that by construction the wavelet of order p is $\mathcal{C}^p$. To do so, one may consider some approximations of the Gabor wavelet. We remind that the Gabor wavelet is given by the mother function $\phi(x) = e^{i\eta x} g(x)$ where g is a gaussian window, and also that the gaussian window can be seen as the limit of an iterated convolution filter. Indeed, let u denote the following function defined by $u(x) = \mathbb{I}_{[-1,1]}(x)$, and let also define, for $h > 0$, $u_h(x) = \frac{1}{h^{d/2}} u(\frac{x}{h})$. Then the following Lemma holds (see for instance Theorem 3.3 page 46, in Guichard et al. (2001)):

Lemma 14 *Let $(h_p)_p$ be a sequence of positive real numbers, and $t > 0$. Then provided that $h_p p \to t$, we have the following (uniform) convergence property*

$$u_{h_p}^{*p} \to_p g_t.$$

where $g_t(x) = \frac{1}{(4\pi t)^{d/2}} e^{\frac{|x|^2}{4t}}$ is the gaussian window that satisfies $g_t * g_s = g_{t+s}$ and $*$ is the symbol for the convolution operation.

Thus, if we set $h_p = \frac{t}{p}$, and define the wavelet of order p to be $\phi^{(p)}(x) = e^{i\eta x} u_{h_{p+1}}^{*(p+1)}(x)$ for some well chosen η, then we have by construction that $\phi^{(p)}$ is $\mathcal{C}^p$, with p vanishing moments. Note that the Gabor wavelet is $\mathcal{C}^\infty$ with ∞ vanishing moments.

Regression with scrambled wavelets

Lemma 15 *Assume that the mother wavelet $\tilde{\phi}$ has compact support $[0,1]^d$ and is bounded by λ, and assume that the target function $f^* = \sum_i \alpha_i^* \phi_i$ lies in the Sobolev space $H^s([0,1]^d)$ with $s > d/2$ (i.e. such that $||\alpha^*|| < \infty$). Then, we have*

$$||\alpha^*||^2 \sup_{x \in \mathcal{X}} ||\phi(x)||^2 \leq \frac{\lambda^{2d}(2^d - 1)}{1 - 2^{-2(s-d/2)}} ||f^*||_{H^s([0,1]^d)}^2.$$

Proof Indeed, we have by definition of the Kernel space of the law of the scrambled wavelets,

$$||\alpha^*||^2 = ||f^*||_{B_{s,2,2}}^2 = ||f^*||_{\mathcal{K}}^2$$

In addition, by definition, the rescaled wavelet are $\phi_{\epsilon,j,l}(x) = 2^{-js}\tilde{\phi}_{\epsilon,j,l}(x) = 2^{-js}2^{jd/2}\tilde{\phi}_{\epsilon}(2^j x - l)$, where $\tilde{\phi}_{\epsilon}(x) = \Pi_{i=1}^{d}\tilde{\phi}_{\epsilon_i}(x)$. Thus for all $x \in [0,1]^d$, by the assumption on the support on $\tilde{\phi}$, $||\phi(x)||^2 = \sum_{\epsilon}\sum_j (2^{-js}2^{jd/2}\tilde{\phi}_{\epsilon}(2^j x_i - l_i))^2$, and by definition of λ, this is bounded by $\sum_{\epsilon}\sum_j (2^{-j(s-d/2)}\lambda^d)^2 \leq (2^d - 1)\frac{\lambda^{2d}}{1-2^{-2(s-d/2)}}$ whenever $s > d/2$. Thus $||\phi(x)||^2 \leq \frac{\lambda^{2d}(2^d-1)}{1-2^{-2(s-d/2)}}$. ■

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

Thus from Theorem 7, ordinary least-squares performed on random subspaces spanned by P scrambled wavelets has an expected excess risk of the following order

$$\mathbb{E}_{\mathcal{G}_P,X,Y}||f^* - \hat{g}||^2_{\mathcal{P}} = O\Big(\frac{\log n}{P}||f^*||^2_{H^s([0,1]^d)} + P\frac{\log n}{n}\Big), \tag{5}$$

(and a similar bound holds in high probability).

4.3 Adaptivity to the law

The bounds on the excess risk obtained in (4) and (5) do not depend on the distribution $\mathcal{P}$ under which the data are generated. This is crucial in our setting since $\mathcal{P}$ is usually unknown. It should be noticed that this property does not hold when one considers non-randomised approximation spaces. Indeed, it is relatively easy to exhibit a particularly well-chosen set of features ϕ_i that will approximate functions in a given class using a particular measure $\mathcal{P}$. For example when $\mathcal{P} = \lambda$, the Lebesgue measure, and $f^* \in H^s([0,1]^d)$ (with $s > d/2$), then linear regression using wavelets (with at least $d/2$ vanishing moments), which form an orthonormal basis of $L_{2,\lambda}([0,1]^d)$, enables to achieve a bound similar to (5). However, this is no more the case when $\mathcal{P}$ is not the Lebesgue measure and it seems very difficult to modify the features ϕ_i in order to recover the same bound, even when $\mathcal{P}$ is known. This seems to be even harder when $\mathcal{P}$ is arbitrary and not known in advance.

Randomisation 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 $\mathcal{P}$ used to assess the performance (even when $\mathcal{P}$ is unknown) is.

Example For illustration, consider a very peaky (a spot) distribution $\mathcal{P}$ in a high-dimensional space $\mathcal{X}$. Regular linear approximation, say with wavelets (see e.g. DeVore (1997)), will most probably miss the specific characteristics of f^* at the spot, since the first wavelets have large support. On the contrary, scrambled wavelets, which 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^* at the spot. This is illustrated in the numerical experiment below.

Here $\mathcal{P}$ is a very peaky Gaussian distribution and f^* is a 1-dimensional periodic function. We consider as initial features $(\phi_i)_{i\geq 1}$ the set of hat functions defined in Section 2.5. Figure 4.3 shows the target function f^*, the distribution $\mathcal{P}$, and the data $(x_n, y_n)_{1\leq n\leq 100}$ (left plots). The middle plots represents the least-squares estimate $\hat{g}$ using $P = 40$ scrambled objects $(\psi_p)_{1\leq p\leq 40}$ (here Brownian motions). The right plots shows the least-squares estimate using the initial features $(\phi_i)_{1\leq i\leq 40}$. The top figures represent a high level view of the whole domain $[0,1]$. No method is able to learn f^* 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^* in terms of the measure $\mathcal{P}$.

Approximately $\mathcal{P}$-smooth functions Let us consider the case when $f^* \in H^s(\mathcal{P})$ but not in $\mathcal{H}^s(\lambda)$. In this case, the bound we have makes appear $||f||_{H^s(\lambda)}$ that is infinite. However, since we only see the function through the points generated by the distribution $\mathcal{P}$, the intuition is that it should be possible to have a better bound.

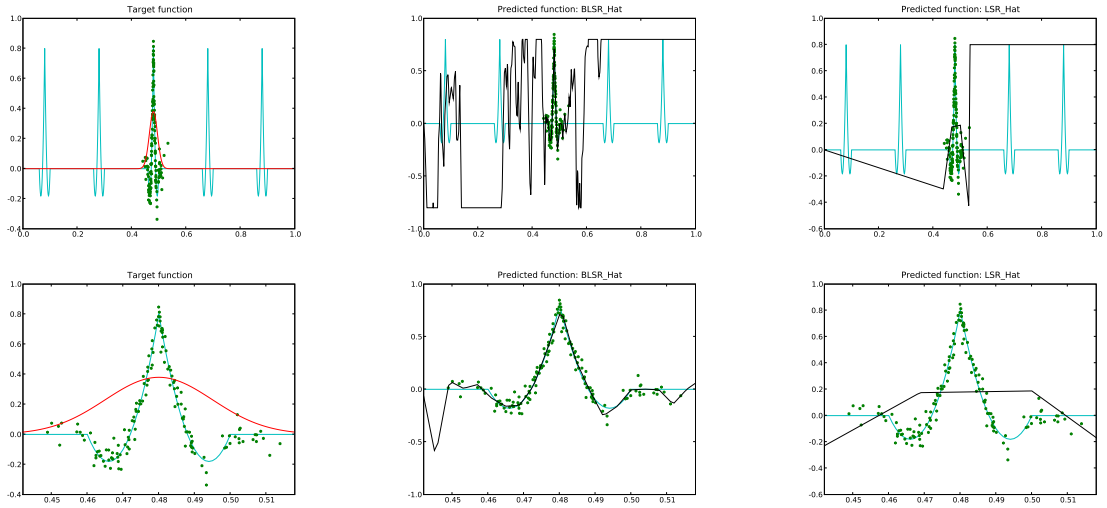

Figure 1: LS estimate of f^* using $N = 100$ data generated from a peaky distribution $\mathcal{P}$ (left plots), using 40 Brownian motions (ψ_p) (middle plots) and 40 hat functions (ϕ_i) (right plots). The bottom row shows a zoom around the data.

We introduce for that purpose $H^s(\lambda, \mathcal{P})$, the space of functions f such that $||f||_{H^s(\lambda)} \leq ||f||_{H^s(\mathcal{P})}$ and $[f^*]_{H^s(\mathcal{P})}$, the space of functions f such that $||f||_{H^s(\mathcal{P})} < \infty$ and $f_{|\,\mathrm{supp}(\mathcal{P})} = f^*_{|\,\mathrm{supp}(\mathcal{P})}$. Intuitively speaking, the space $H^s(\lambda, \mathcal{P})$ is the space of functions that are smooth for the $\mathcal{P}$ Sobolev norm, and that can be embedded in $H^s(\lambda)$ with same smoothness degree, while $[f^*]_{H^s(\mathcal{P})}$ is the set of $\mathcal{P}$-smooth functions that are not distinguishable from f^* via $\mathcal{P}$.

We say that f^* is *approximatively $\mathcal{P}$-smooth* if the projection of f^* in the space $H^s(\lambda, \mathcal{P}) \cap [f^*]_{H^s(\mathcal{P})}$ for the sobolev norm of $H^s(\mathcal{P})$ exists. We write it $\overline{f} = \Pi_{H^s(\mathcal{P})}(f^*, H^s(\lambda, \mathcal{P}) \cap [f^*]_{H^s(\mathcal{P})})$. When this is the case, then we can replace in the bounds the quantity $||f^*||_{H^s(\lambda)}$ with the following finite quantity

$$||\overline{f}||_{H^s(\lambda)} \leq ||\overline{f}||_{H^s(\mathcal{P})} \leq ||f^*||_{H^s(\mathcal{P})} + ||\overline{f} - f^*||_{H^s(\mathcal{P})}$$

5. Efficient implementation

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

Approximation error Using a finite F introduces an additional approximation (squared) error term in the final excess risk bounds. This additional error (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, e.g. $o(N^{-1/2})$, whenever the depth of the wavelet dyadic-tree is bigger than $\frac{\log N}{d}$. Our main concern is thus about efficient computation.

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

In the multi-resolution schemes described here, provided that the mother function has 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 of the random features $(\psi_p)_{p \leq P}$ is built only when needed for the evaluation at the points $(x_n)_n$.

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 $O(2^d HPN)$.

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 $O(2^d N^{3/2} \log N)$. Then in order to solve the least squares system, one has to compute $\Psi^T \Psi$, that has cost $O(P^2 N)$, 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 $O(2^d N^{3/2} \log N + N^2)$.

Eventually, the numerical complexity to make a new prediction is $O(2^d N^{1/2} \log(N))$.

References

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

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

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

D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. In *STOC '87: Proceedings of the nineteenth annual ACM symposium on Theory of computing*, pages 1–6, New York, NY, USA, 1987. ACM.

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

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

Frederic Guichard, Jean michel Morel, and Jean michel Morel. Image analysis and p.d.e.s. *IPAM GBM Tutorial*, 2001.

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

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

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

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

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

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. ISSN 0021-9045. doi: http://dx.doi.org/10.1016/j.jat.2009.01.001.

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

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

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.

Appendix A. Approximation error

In order to prove Theorem 4, we use the fact that random projections of a function f_α approximately preserves its norm. More precisely we firstprove proposition 10.

A.1 Proof of Proposition 10

In this section, we prove Proposition 10. Thus we want to bound $||f_\alpha - g_{A\alpha}||_n^2$ for one given α.

In order to state our result, we first remind the so-called Johnson-Lindenstrauss Lemma concerning the approximate preservation of norms, and then we extend this result to the preservation of inner products of vectors and then of sequences satisfying some conditions satisfied by Gaussian objects.

Johnson-Lindenstrauss Lemma Let A be a $P \times F$ matrix of iid gaussian $\mathcal{N}(0, 1/P)$ entries. Then the following lemma states that the random (with respect to the choice of the matrix A) variable $||Au||^2$ concentrates around its expectation $||u||^2$ when P is large.

Lemma 16 *For any vector u in $\mathbb{R}^F$ and any $\epsilon \in (0,1)$, we have*

$$\begin{aligned}\mathbb{P}\Big(||Au||^2 \geq (1+\epsilon)||u||^2\Big) &\leq e^{-P(\epsilon^2/4-\epsilon^3/6)}\\ \mathbb{P}\Big(||Au||^2 \leq (1-\epsilon)||u||^2\Big) &\leq e^{-P(\epsilon^2/4-\epsilon^3/6)}\end{aligned}$$

The proof directly uses concentration inequalities (1938 Cramer's large deviation Theorem) and may be found e.g. in Achlioptas (2003). Note that the Gaussianity is not needed here, and this is also true for some other distributions, like for instance:

- $\pm$ Bernoulli distributions, i.e. 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 $2/3$.

Thus, we deduce by polarisation of the Euclidean norm that a similar result holds for dot products:

Lemma 17 *Let $(u_k)_{1\leq k\leq K}$ and v be vectors of $\mathbb{R}^F$. Let A be a $P\times F$ matrix of i.i.d. elements drawn from one of the previously defined distributions. For any $\epsilon \in (0,1)$, $\delta > 0$, for $P \geq \frac{1}{\frac{\epsilon^2}{4}-\frac{\epsilon^3}{6}} \log \frac{4K}{\delta}$, we have, with probability at least $1-\delta$, for all $k \leq F$,*

$$|Au_k \cdot Av - u_k \cdot v| \leq \epsilon ||u_k||\,||v||.$$

Proof To prove the Lemma, we apply Lemma 16 to any couple of vectors $u+w$ and $u-w$, where u and w are vectors of norm 1. By polarisation, we have that

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

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

Thus for each $k \leq K$, we have with same probability:

$$Au_k \cdot Av \leq u_k \cdot v + \epsilon ||u_k||\,||v||.$$

Now the symmetric inequality holds with the same probability, and using a union bound for considering all $(u_k)_{k\leq K}$, we have that

$$|Au_k \cdot Av - u_k \cdot v| \leq \epsilon ||u_k||\,||v||,$$

holds for all $k \leq K$, with probability $1 - 4Ke^{-P(\epsilon^2/4-\epsilon^3/6)}$, and the proposition follows. ■

Random projections with Gaussian object In our setting, we deal with infinite dimensional spaces, and thus we have to extend the Johnson-Lindenstrauss Lemma to the case of infinite sequences. Fortunately, thanks to the measurability properties of Gaussian Objects, this extension is a direct consequence of the theory of Gaussian objects. Indeed, Lemma 17 applies to the two truncated sequences $\overline{\alpha}_F = (\alpha_1, \ldots, \alpha_F)$ that plays the role of v and $\overline{\phi}_k^F = (\phi_1(X'_k(\omega_1)), \ldots, \phi_1(X'_k(\omega_1))_F)$ that plays the role of u_k for any finite F. Then the extension to sequences follows due to the converging properties of these two sequences w.r.t. the random gaussian elements $(A_{p,i})_{1\leq p\leq P, i\geq 1}$ and the measurability of the limit objects, ensured by the properties of Gaussian objects.

Thus, if we now set $\epsilon^2 = \frac{8\log(4n/\delta)}{P}$, we deduce that for $P \geq 15\log(4n/\delta)$, then $\epsilon \leq 3/4$, so $\epsilon^2/4 - \epsilon^3/6 \geq \epsilon^2/8$ and thus $P \geq \frac{\log(4n/\delta)}{\epsilon^2/4-\epsilon^3/6}$, which enables us to apply Lemma 17 and state the following result:

Lemma 18 *Provided that $P \geq 15\log(4n/\delta)$, we have with probability higher than $1 - \delta_G$ w.r.t. the gaussian random variables, for fixed variables $(X_i)_{i=1..n}$:*

$$||f_\alpha - g_{A\alpha}||_n^2 \leq \frac{8\log(4n/\delta_G)}{P}||\alpha||_2^2 \frac{1}{n}\sum_{i=1}^n ||\phi(X_i)||_2^2$$

A.2 Proof of Theorem 4

In this section, we prove Theorem 4.

Since $f^* = f_{\alpha^*}$, we have the following first bound

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

Now if we introduce m i.i.d. ghost samples $(X'_j)_{j\leq m}$, since by definition

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

and for all j, $(f_{\alpha^*} - T_L(g_{A\alpha^*}))^2(X'_j) \in [0, (2L)^2]$ a.s. we can apply Hoeffding's inequality, and deduce that there exists an event of probability higher than $1 - \delta'$ such that

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

Now it remains to bound the first term of the right hand of this inequality. Thanks to fact that $||f_{\alpha^*}||_\infty \leq L$, we have

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

and this term is bounded on an event of probability higher than $1 - \delta$ according to Proposition 10. Thus, we deduce the final result by an union bound over the two events.

Bound in expectation In order to deduce the bound in expectation, we first use the fact that

$$\begin{aligned}\mathbb{E}_{\mathcal{G}}[\inf_{g\in\mathcal{G}}(||f^*-T_L(g)||_{\mathcal{P}}^2] &\leq \mathbb{E}_{\mathcal{G}}[||f_{\alpha^*}-T_L(g_{A\alpha^*})||_{\mathcal{P}}^2]\\ &= \mathbb{E}_{\mathcal{G}}\mathbb{E}_{\mathcal{P}}[||f_{\alpha^*}-T_L(g_{A\alpha^*})||_m^2]\\ &= \mathbb{E}_{\mathcal{P}}\mathbb{E}_{\mathcal{G}}(||f_{\alpha^*}-T_L(g_{A\alpha^*})||_m^2).\end{aligned}$$

Then by Lemma 16, we have for all $\epsilon\in(0,1)$,

$$\mathbb{P}_G(||f_\alpha-g_{A\alpha}||_m^2>\epsilon^2u^2)\leq 4m\exp(-P(\epsilon^2/4-\epsilon^3/6))$$

where we introduced the quantity $u^2=||\alpha^*||^2\frac{1}{m}||\phi(X_i)||_m^2$.

Thus, we deduce that

$$\begin{aligned}\mathbb{E}_G(||f_\alpha-T_L[g_{A\alpha}]||_m^2) &\leq \inf_\eta[\eta+4m\int_\eta^{(2L)^2}\exp(-P(\frac{t}{4u^2}-\frac{t^{3/2}}{6u^3}))dt]\\ &\leq \inf_\eta[\eta+4m\int_\eta^{(2L)^2}\exp(-\frac{P}{u^2}(\frac{1}{4}-\frac{\gamma}{3})t)dt].\end{aligned}$$

Now, provided that $P\geq\frac{\log(4m)}{(1-\frac{4\gamma}{3})\gamma^2}$, the optimal point is reached for $\eta=\frac{u^2\log(4m)}{(\frac{1}{4}-\frac{\gamma}{3})P}$, where $\gamma=\frac{L}{||\alpha^*||_2\sup_{x\in\mathcal{X}}||\phi(x)||_2}$ and we get

$$\mathbb{E}_G(||f_\alpha-T_L[g_{A\alpha}]||_m^2)\in[\frac{u^2}{P(\frac{1}{4}-\frac{\gamma}{3})}\log(4m),\frac{u^2}{P(\frac{1}{4}-\frac{\gamma}{3})}(\log(4m)+1)].$$

Thus, by optimizing over m, this enables to deduce that provided $P\geq\frac{\log(4)}{(1-\frac{4\gamma}{3})\gamma^2}$, then

$$\mathbb{E}_{\mathcal{G}}[\inf_{g\in\mathcal{G}}(||f^*-T_L(g)||_{\mathcal{P}}^2]\leq\frac{||\alpha^*||_2^2\mathbb{E}(||\phi(X)||_2^2)}{P}\frac{\log(4)+1}{\frac{1}{4}-\frac{\gamma}{3}}.$$

Appendix B. Excess risk bounds with explicit constants

In this section, we prove Theorem 7. Thus, we bound the three terms identified as an approximation term, a noise term and an estimation term.

We remind the following decomposition:

$$||f^*-T_L(g_{\hat\beta})||_{\mathcal{P}}^2 \leq 8||f^*-g_{\tilde\beta}||_n^2+8||g_{\tilde\beta}-g_{\hat\beta}||_n^2+\sup_{g\in\mathcal{G}}\left(||f^*-T_L(g)||_{\mathcal{P}}^2-8||f^*-T_L(g)||_n^2\right)$$

Proof First we can write the following inequality:

$$||f^*-T_L(g_{\hat\beta})||_{\mathcal{P}}^2\leq 8||f^*-T_L(g_{\hat\beta})||_n^2+\sup_{g\in\mathcal{G}}\left(||f^*-T_L(g)||_{\mathcal{P}}^2-8||f^*-T_L(g)||_n^2\right).$$

Then, we deduce the desired statement using the fact that $||f^*||_\infty\leq L$ and the definition of $g_{\tilde\beta}$:

$$||f^*-T_L(g_{\hat\beta})||_n^2\leq||f^*-g_{\hat\beta}||_n^2\leq||f^*-g_{\tilde\beta}||_n^2+||g_{\tilde\beta}-g_{\hat\beta}||_n^2,$$

■

Thus in order to prove Theorem 7, we combine Proposition 10, Proposition 11 and Proposition 12 using a union bound over the three events. Not that the condition for Proposition 10, $P\geq 15\log(24n/\delta)$, ensures that the condition for Proposition 12, $n\log(n)\geq\frac{4}{P}$, is also satisfied.

B.1 Noise term

We can bound the noise term $||g_{\tilde{\beta}} - g_{\hat{\beta}}||_n^2$ using a simple Chernoff bound together with a chaining argument. Indeed, by definition of $g_{\tilde{\beta}}$ and $g_{\hat{\beta}}$, if we write $Y = f + \eta$ where we introduced the noise vector η, we have

$$\begin{aligned} ||g_{\tilde{\beta}} - g_{\hat{\beta}}||_n^2 &= \langle g_{\tilde{\beta}} - g_{\hat{\beta}}, \eta \rangle_n \\ &= \frac{1}{n} \sum_{i=1}^n \eta_i (g_{\tilde{\beta}} - g_{\hat{\beta}})(X_i) \\ &\leq \Big(\sup_{g \in \mathcal{G}} \frac{\frac{1}{n} \sum_{i=1}^n \eta_i g(X_i)}{||g||_n} \Big) ||g_{\tilde{\beta}} - g_{\hat{\beta}}||_n \\ &\leq \Big(\sup_{g \in \mathcal{G}} \frac{\frac{1}{n} \sum_{i=1}^n \eta_i g(X_i)}{||g||_n} \Big)^2 . \end{aligned}$$

Thus, we focus on the space $\mathcal{G}^1 = \{g \in \mathcal{G}; ||g||_n = 1\}$.

Generic chaining enables to bound the supremum term, once we have introduced the distance d such that the following holds true

$$\mathbb{P}(|\sum_{i=1}^n \eta_i (g_1(X_i) - g_2(X_i))| > t d(g_1, g_2)) \leq 2 \exp(-t^2/2),$$

where the probability is taken over the noise term only, i.e. conditionally on the sampling points $(X_i)_{i=1..n}$ and the gaussian random variables.

Thus, we apply the following Lemma that enables to control the deviations in Orlicz norm:

Lemma 19 *There exists an absolute constant $c > 0$ such that the following holds. Let $\alpha \in [1, 2]$ and $X_1, ..., X_n$ be independent random variables such that for all i, $||X_i||_{\psi_\alpha} \leq C$. Then, for every vector $a \in \mathbb{R}^n$ and every $\epsilon > 0$ we have*

$$\mathbb{P}(|\sum_{i=1}^n a_i X_i| \geq tC) \leq 2 \exp(-c \min(\frac{t^2}{||a||_2^2}, \frac{t^\alpha}{||a||_{\alpha'}^\alpha}))$$

where $\alpha^{-1} + \alpha'^{-1} = 1$.

Thus, since we assumed that the noise satisfies $||\eta_i||_{\psi_2} \leq C$, then we deduce that conditionally on the sampling points and gaussian random variables,

$$\mathbb{P}(|\sum_{i=1}^n (g_1(X_i) - g_2(X_i))\eta_i| \geq tC) \leq 2 \exp(-c \frac{t^2}{\sum_{i=1}^n (g_1(X_i) - g_2(X_i))^2}).$$

Thus, the distance d is given by $d(g_1, g_2) = \frac{C}{\sqrt{2c}} (\sum_{i=1}^n (g_1(X_i) - g_2(X_i))^2)^{1/2} = C \sqrt{\frac{n}{2c}} ||g_1 - g_2||_n$.

Note that under the stronger assumption that the noise is bounded, i.e. $||\eta||_\infty \leq C$, then by Hoeffding's inequality, we would have $d(g_1, g_2) = C (\sum_{i=1}^n (g_1(X_i) - g_2(X_i))^2)^{1/2} = C \sqrt{n} ||g_1 - g_2||_n$.

We can give an intuition of the final result thanks to Generic chaining. Indeed, we have the following property:

$$\begin{aligned}
\mathbb{E}(\sup_{g\in\mathcal{G}^1}\frac{1}{n}\sum_{i=1}^n\eta_i g(X_i)) &\leq O(\frac{1}{n}\int_0^\infty\sqrt{\log(\mathcal{N}(\epsilon,\mathcal{G}^1,d))}d\epsilon)\\
&\leq O(\frac{1}{n}\int_0^\infty\sqrt{\log(\mathcal{N}(\frac{\epsilon\sqrt{2c}}{C\sqrt{n}},\mathcal{G}^1,||.||_n))}d\epsilon)\\
&\leq O(3C\sqrt{\frac{P}{2cn}}\int_0^\infty\sqrt{\log_+(\frac{1}{\epsilon})}d\epsilon)\\
&= O(C\sqrt{\frac{P}{n}}).
\end{aligned}$$

where we used the fact that since $\mathcal{G}^1$ is a linear space of dimension P, its covering number in empirical norm is bounded above by $\mathcal{N}(\epsilon,\mathcal{G}^1,||.||_n))\leq(\frac{2}{\epsilon}+1)^P\leq\max(\frac{3}{\epsilon},1)^P$.

Now we prove Proposition 11, that gives the behaviour of the noise error term in high probability with explicit constants:

Proposition 20 *With probability higher than $1-\delta$ w.r.t all sources of randomness:*

$$||g_{\tilde{\beta}}-g_{\hat{\beta}}||_n^2\leq\frac{2C^2}{c}\frac{(253P+145\log(6/\delta))}{n}.$$

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

$$\rho(t)=\mathbb{P}_Y(\exists g\in\mathcal{G}\frac{\frac{1}{n}\sum_{i=1}^n\eta_i g(X_i)}{||g||_n}>t) \quad = \quad \mathbb{P}_Y(\exists g\in\mathcal{G}^1\frac{1}{n}\sum_{i=1}^n\eta_i g(X_i)>t).$$

Let us consider ϵ_j-covers C_j of $\mathcal{G}^1$, for $j=0...\infty$, with $C_0=g_0$. We moreover assume that C_{j+1} is a refinement of C_j and that $\epsilon_j\leq\epsilon_{j-1}$. 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 $||g-g_0||_n\leq 2$, we need to consider $\epsilon_0\geq 2$.

Thus if we now introduce γ and γ_j such that $\sum_{j=1}^J \gamma_j \leq \gamma$, then,

$$\begin{aligned}
\rho(\gamma t_1 + t_2 + t_3) &\leq \mathbb{P}(\exists g \in \mathcal{G}^1 \frac{1}{n}\sum_{i=1}^n \eta_i (g-g_0)(X_i) > \gamma t_1 + t_2) + 2\exp(-\frac{ct_3^2 n}{C^2}) \\
&\leq \mathbb{P}(g \in \exists \mathcal{G}^1 \frac{1}{n}\sum_{i=1}^n \eta_i (g-g_J)(X_i) + \sum_{j=1}^J \frac{1}{n}\sum_{i=1}^n \eta_i (g_j - g_{j-1})(X_i) \geq \sum_{j=1}^J \gamma_j t_1 + t_2) \\
&\quad +2\exp(-\frac{ct_3^2 n}{C^2}) \\
&\leq \sum_{j=1}^J \mathbb{P}(\exists g \in \mathcal{G}^1 \frac{1}{n}\sum_{i=1}^n \eta_i (g_j - g_{j-1})(X_i) > t_1 \gamma_j) \\
&\quad +2\exp(-\frac{ct_2^2 n}{C^2 \epsilon_J^2}) + 2\exp(-\frac{ct_3^2 n}{C^2}) \\
&\leq \mathbb{E}\sum_{j=1}^J \mathcal{N}(\epsilon_j, G^1, ||.||_n)\mathcal{N}(\epsilon_{j-1}, G^1, ||.||_n)\mathbb{P}(\frac{1}{n}\sum_{i=1}^n \eta_i (g_j - g_{j-1})(X_i) > t_1\gamma_j) \\
&\quad +2\exp(-\frac{ct_2^2 n}{C^2 \epsilon_J^2}) + 2\exp(-\frac{ct_3^2 n}{C^2}).
\end{aligned}$$

Now, note that since $\epsilon_j \leq \epsilon_{j-1}$, then $\mathcal{N}(\epsilon_{j-1}, G^1, ||.||_n) \leq \mathcal{N}(\epsilon_j, G^1, ||.||_n)$. Note also that $||g_j - g_{j-1}||_n \leq \eta_j$ since C_j is a refinement of C_{j-1}. We can bound the entropy number by $\mathcal{N}(\epsilon_j, G^1, ||.||_n) \leq N_j = \max(\frac{3}{\epsilon_j}, 1)^P$ where P is the dimension of $\mathcal{G}$. Thus we deduce that:

$$\rho(\gamma t_1 + t_2 + t_3) \leq 2\sum_{j=1}^J N_j^2 \exp(-\frac{ct_1^2 n\gamma_j^2}{C^2\epsilon_j^2}) + 2\exp(-\frac{ct^2 n}{C^2\epsilon_J^2}) + 2exp(-\frac{ct_3^2 n}{C^2}).$$

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

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

Thus, it remains to define ϵ_j. Since $N_j = \max(\frac{3}{\epsilon_j}, 1)^P$, we define the covering radius ϵ_j to be $\epsilon_j = 2^{-j} 3\delta_1^{1/2P}(2^{2P}-1)^{1/2P}$ for some $\delta_1 \in (0,1]$. Thus $\sum_{j=1}^J \frac{1}{N_j^2} \leq \delta_1$. Now since $\epsilon_j \to 0$ when $j \to \infty$, we can make the sum goes to infinity. We deduce that:

$$\rho(\gamma t_1 + C\sqrt{\frac{\log(2/\delta_3)}{cn}}) \leq \delta_1 + \delta_2 + \delta_3.$$

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

$$\begin{aligned}
\gamma t_1 &= 2\sum_{j=1}^{\infty} \epsilon_j C\sqrt{\frac{\log(N_j)}{cn}} \\
&\leq \frac{12C}{\sqrt{cn}}\sum_{j=1}^{\infty} 2^{-j}\sqrt{jP\log(2) + \frac{1}{2}\log(1/\delta_1) - \frac{1}{2}\log(2^{2P}-1)} \\
&\leq \frac{12C}{\sqrt{cn}}\sum_{j=1}^{\infty} 2^{-j}\sqrt{(j-1)P\log(2) + \frac{1}{2}\log(2/\delta_1)} \\
&\leq \frac{12C}{\sqrt{cn}}\Big(\sum_{j=1}^{\infty} 2^{-j}\sqrt{(j-1)P\log(2)} + \sqrt{\frac{1}{2}\log(2/\delta_1)}\Big) \\
&\leq \frac{12C}{\sqrt{cn}}\Big((1+\sqrt{2})\sqrt{P\log(2)} + \sqrt{\frac{1}{2}\log(2/\delta_1)}\Big).
\end{aligned}$$

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

Using the inequalities $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}$, we thus deduce the following bound:

$$\begin{aligned}
\gamma t_1 + t_2 + t_3 &\leq \frac{C}{\sqrt{cn}}\Big(12(1+\sqrt{2})\sqrt{P\log(2)} + \sqrt{144\log(2/\delta_1) + 2\log(2/\delta_3)}\Big) \\
&\leq \sqrt{\frac{2}{c}}\frac{C}{\sqrt{n}}\sqrt{253P + 144\log(2/\delta_1) + \log(2/\delta_3)}.
\end{aligned}$$

Thus, conditionally on the gaussian random variables and the random sample $(X_i)_{1\ldots n}$ we deduce that

$$\mathbb{P}(\exists g \in \mathcal{G}\, \frac{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(X_i)}{||g||_n} > \sqrt{\frac{2}{c}}\frac{C}{\sqrt{n}}\sqrt{253P + 145\log(6/\delta)}|G, X) \leq \delta.$$

Thus, by independence of the three sources of randomness (noise, random sample and gaussian variables), since we only consider one $\mathcal{G}$ and one sample $(X_i)_{i=1..n}$ at a time, this enables to deduce that the following bound for the second term holds with probability higher than $1-\delta$ w.r.t all source of randomness:

$$||g_{\tilde{\beta}} - g_{\hat{\beta}}||_n^2 \leq \frac{2C^2}{c}\frac{(253P + 145\log(6/\delta))}{n}.$$

B.2 Estimation term

We prove Proposition 12 that we remind here:

Proposition 21 *Assuming that $n\log(n) \geq \frac{4}{P}$, then with probability higher than $1-\delta_X$ w.r.t. all source of randomness, the following holds true:*

$$\sup_{g\in\mathcal{G}} ||f^* - T_L(g)||_{\mathcal{P}}^2 - 8||f^* - T_L(g)||_n^2 \leq (24L)^2\frac{4\log(3/\delta) + 2P\log(n)}{n}.$$

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

$$\mathbb{P}(\sup_{g\in\mathcal{G}} ||f^* - T_L(g)||_{\mathcal{P}} - 2||f^* - T_L(g)||_n > \epsilon) \leq 3\mathbb{E}(\mathcal{N}(\frac{\sqrt{2}}{24}\epsilon, \mathcal{G}^0, ||.||_{2n}))\exp(-\frac{n\epsilon^2}{288(2L)^2}).$$

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

$$\mathcal{N}(\frac{\sqrt{2}}{24}\epsilon, \mathcal{G}^0, ||.||_{2n}) \leq \mathcal{N}(\frac{\sqrt{2}}{24}\epsilon, T_L(\mathcal{G}), ||.||_{2n}) \leq (\frac{2(2L).24}{\sqrt{2}\epsilon} + 1)^P.$$

Thus we deduce that if $\epsilon \geq \frac{24.4L}{\sqrt{2}}u$, then with probability higher than $1-\delta$ w.r.t the law $\mathcal{P}$, for fixed random gaussian variables,

$$\sup_{g\in\mathcal{G}} ||f^* - T_L(g)||_{\mathcal{P}} - 2||f^* - T_L(g)||_n \leq \epsilon = 24L\sqrt{\log(3/\delta) + P\log(\frac{1}{u}+1)}\sqrt{\frac{2}{n}}.$$

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

$$\sup_{g\in\mathcal{G}} ||f^* - T_L(g)||_{\mathcal{P}} - 2||f^* - T_L(g)||_n \leq 24L\sqrt{\frac{2\log(3/\delta) + P\log(n)}{n}}.$$

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

$$\begin{aligned} ||f^* - T_L(g)||_{\mathcal{P}}^2 &\leq (2||f^* - T_L(g)||_n + 24L\sqrt{\frac{2\log(3/\delta) + P\log(n)}{n}})^2 \\ &\leq 8||f^* - T_L(g)||_n^2 + (24L)^2\frac{4\log(3/\delta) + 2P\log(n)}{n}. \end{aligned}$$

This gives the following upper bound, that holds with probability higher than $1-\delta$ w.r.t. all source of randomness (after deconditioning, by independence and since we consider one $\mathcal{G}$ at a time):

$$\sup_{g\in\mathcal{G}} ||f^* - T_L(g)||_{\mathcal{P}}^2 - 8||f^* - T_L(g)||_n^2 \leq (24L)^2\frac{4\log(3/\delta) + 2P\log(n)}{n}$$

■

B.3 Expectation

In order to state the results in expectation over (X, Y), note that we deduce from the previous propositions, that there is an event Ω with probability $1-(\delta_X+\delta_Y+\delta_G)$ with respect to the all sources of randmoness, such that on this event, provided that $P \geq 15\log(8n/\delta_G)$ and $n\log(n) \geq \frac{4}{P}$, then

$$\begin{aligned} ||f^* - T_L(g_{\hat{\beta}})||_{\mathcal{P}}^2 &\leq \frac{64\log(4n/\delta_G)}{P}||\alpha||_2^2\frac{1}{n}\sum_{i=1}^n ||\phi(X_i)||_2^2 \\ &+ \frac{16C^2}{c}\frac{(253P + 145\log(6/\delta_Y))}{n} \\ &+ (24L)^2\frac{2P\log(n) + 4\log(3/\delta_X)}{n}\Big). \end{aligned}$$

Now we also now that $||f^* - T_L(g_{\hat{\beta}})||_{\mathcal{P}}^2 \leq (2L)^2$ on Ω^C. This enables us to deduce a bound on $\mathbb{E}_{X,Y}(||f^* - T_L(g_{\hat{\beta}})||_{\mathcal{P}}^2)$ as well as $\mathbb{E}_{\mathcal{G},X,Y}(||f^* - T_L(g_{\hat{\beta}})||_{\mathcal{P}}^2)$.