Approximation and learning by greedy algorithms

Andrew Barron, Albert Cohen, Wolfgang Dahmen, and Ronald DeVore *

August 1, 2006

Abstract

We consider the problem of approximating a given element f from a Hilbert space $\mathcal{H}$ by means of greedy algorithms and the application of such procedures to the regression problem in statistical learning theory. We improve on the existing theory of convergence rates for both the orthogonal greedy algorithm and the relaxed greedy algorithm, as well as for the forward stepwise projection algorithm. For all these algorithms, we prove convergence results for a variety of function classes and not simply those that are related to the convex hull of the dictionary. We then show how these bounds for convergence rates leads to a new theory for the performance of greedy algorithms in learning. In particular, we build upon the results in [18] to construct learning algorithms based on greedy approximations which are universally consistent and provide provable convergence rates for large classes of functions. The use of greedy algorithms in the context of learning is very appealing since it greatly reduces the computational burden when compared with standard model selection using general dictionaries.

1 Introduction

We consider the problem of approximating a function f from a Hilbert space $\mathcal{H}$ by a finite linear combination $\hat{f}$ of elements of a given dictionary $\mathcal{D} = (g)_{g\in\mathcal{D}}$. Here by a *dictionary* we mean any family of functions from $\mathcal{H}$. In this paper, this problem is adressed in two different contexts, namely

(i) Deterministic approximation: f is a known function in a Hilbert space $\mathcal{H}$. The approximation error is naturally measured by $\|f-\hat{f}\|$ where $\|\cdot\|$ is the corresponding norm generated from the inner product $\langle\cdot,\cdot\rangle$ on $\mathcal{H}$, i.e., $\|g\|^2 := \|g\|_{\mathcal{H}}^2 := \langle g, g\rangle$.

(ii) Statistical learning: $f = f_\rho$ where $f_\rho(x) = E(y|x)$ is the regression function of an unknown distribution ρ on $X \times Y$ with $x \in X$ and $y \in Y$ respectively representing the feature and output variables, from which we observe independent realizations

*This research was supported by the Office of Naval Resarch Contracts ONR-N00014-03-1-0051, ONR/DEPSCoR N00014-03-1-0675 and ONR/DEPSCoR N00014-00-1-0470; the Army Research Office Contract DAAD 19-02-1-0028; the AFOSR Contract UF/USAF F49620-03-1-0381; the French-German PROCOPE contract 11418YB; and NSF contracts DMS-0221642 and DMS-0200187

$(z_i) = (x_i, y_i)$ for $i = 1, \cdots, n$. The approximation error is now measured in the Hilbertian norm $\|u\|^2 := E(|u(x)|^2)$.

In either of these situations, we may introduce the set Σ_N of all possible linear combinations of elements of $\mathcal{D}$ with at most N terms and define the best N-term approximation error $\sigma_N(f)$ as the infimum of $\|f - \hat{f}\|$ over all $\hat{f}$ of this type,

$$\sigma_N(f) = \inf_{\#(\Lambda)\le N} \inf_{(c_g)} \|f - \sum_{g\in\Lambda} c_g g\|. \tag{1.1}$$

In the case where $\mathcal{D}$ is an orthonormal basis, the minimum is attained by

$$\hat{f} = \sum_{g\in\Lambda_N(f)} c_g g, \tag{1.2}$$

where $\Lambda_N(f)$ corresponds to the coordinates $c_g := \langle f, g\rangle$ which are the N-largest in absolute value. The approximation properties of this process are well understood, see e.g. [9]. In particular one can easily check that the convergence rate $\|f - \hat{f}\|_{\mathcal{H}} \lesssim N^{-s}$ is equivalent to the property that the sequence $(c_g)_{g\in\mathcal{D}}$ belongs to the weak space $w\ell_p$ with $1/p = 1/2 + s$ (see e.g. the survey [9] or standard books on functional analysis for the definition of weak ℓ_p spaces). Here and later in this paper we use the notation $A \lesssim B$ to mean $A \le CB$ for some absolute constant C that does not depend on the parameters which define A and B.

One of the motivations for utilizing general dictionaries rather than orthonormal systems is that in many applications, such as signal processing or statistical estimation, it is not clear which orthonormal system, if any, is best for representing or approximating f. Thus, dictionaries which are a union of several bases or collections of general waveforms are preferred. Some well known examples are the use of Gabor sytems, curvelets, and wavepackets in signal processing and neural networks in learning theory.

When working with dictionaries $\mathcal{D}$ which are not orthonormal bases, the realization of a best N-term approximation is usually out of reach from a computational point of view since it would require minimizing $\|f - \hat{f}\|$ over all $\hat{f}$ in an infinite or huge number of N dimensional subspaces. *Greedy algorithms* or matching pursuit aim to build "sub-optimal yet good" N-term approximations through a greedy selection of elements g_k, $k = 1, 2, \cdots$, within the dictionary $\mathcal{D}$, and to do so with a more manageable number of computations.

1.1 Greedy algorithms

Greedy algorithms have been introduced in the context of statistical estimation. They have also been considered for applications in signal processing [1]. Their approximation properties have been explored in [4, 14, 18] in relation with neural network estimation, and in [10, 15, 19] for general dictionaries. A recent survey of the approximation properties of such algorithms is given in [21].

There exists several versions of these algorithms. The four most commonly used are the *pure greedy*, the *orthogonal greedy*, the *relaxed greedy* and the *stepwise projection* algorithms, which we respectively denote by the acronyms PGA, OGA, RGA and SPA. We describe these algorithms in the deterministic setting. We shall assume here and later that the elements of the dictionary are normalized according to $\|g\| = 1$ for all $g \in \mathcal{D}$ unless it is explicitly stated otherwise.

All four of these algorithms begin by setting $f_0 := 0$. We then define recursively the approximant f_k based on f_{k-1} and its residual $r_{k-1} := f - f_{k-1}$.

In the PGA and the OGA, we select a member of the dictionary as

$$g_k := \operatorname*{Argmax}_{g \in \mathcal{D}} |\langle r_{k-1}, g\rangle|. \tag{1.3}$$

The new approximation is then defined as

$$f_k := f_{k-1} + \langle r_{k-1}, g_k\rangle g_k, \tag{1.4}$$

in the PGA, and as

$$f_k = P_k f, \tag{1.5}$$

in the OGA, where P_k is the orthogonal projection onto $V_k := \mathrm{Span}\{g_1, \cdots, g_k\}$. It should be noted that when $\mathcal{D}$ is an orthonormal basis both algorithms coincide with the computation of the best k-term approximation.

In the RGA, the new approximation is defined as

$$f_k = \alpha_k f_{k-1} + \beta_k g_k, \tag{1.6}$$

where (α_k, β_k) are real numbers and g_k is a member of the dictionary. There exists many possibilities for the choice of (α_k, β_k, g_k), the most greedy being to select them according to

$$(\alpha_k, \beta_k, g_k) := \operatorname*{Argmin}_{(\alpha,\beta,g) \in \mathbb{R}^2 \times \mathcal{D}} \|f - \alpha f_{k-1} - \beta g\|. \tag{1.7}$$

Other choices specify one or several of these parameters, for example by taking g_k as in (1.3) or by setting in advance the value of α_k and β_k, see e.g. [14] and [4]. Note that the RGA coincides with the PGA when the parameter α_k is set to 1.

In the SPA, the approximation f_k is defined by (1.5) as in the OGA, but the choice of g_k is made so as to minimize over all $g \in \mathcal{D}$ the error between f and its orthogonal projection onto $\mathrm{Span}\{g_1, \cdots, g_{k-1}, g\}$.

Note that, from a computational point of view, the OGA and SPA are more expensive to implement since at each step they require the evaluation of the orthogonal projection $P_k f$ (and in the case of SPA a renormalization). Such projection updates are computed preferrably using Gramm-Schmidt orthogonalization (e.g. via the QR algorithm) or by solving the normal equations

$$G_k a_k = b_k, \tag{1.8}$$

where $G_k := (\langle g_i, g_j\rangle)_{i,j=1,\cdots,k}$ is the Grammian matrix, $b_k := (\langle f, g_i\rangle)_{i=1,\cdots,k}$, and $a_k := (\alpha_j)_{j=1,\cdots,k}$ is the vector such that $f_k = \sum_{j=1}^k \alpha_j g_j$.

In order to describe the known results concerning the approximation properties of these algorithms, we introduce the class $\mathcal{L}_1 := \mathcal{L}_1(\mathcal{D})$ consisting of those functions f which admit an expansion $f = \sum_{g\in\mathcal{D}} c_g g$ where the coefficient sequence (c_g) is absolutely summable. We define the norm

$$\|f\|_{\mathcal{L}_1} := \inf\{\sum_{g\in\mathcal{D}} |c_g| \ : \ f = \sum_{g\in\mathcal{D}} c_g g\} \tag{1.9}$$

for this space. This norm may be thought of as an ℓ_1 norm on the coefficients in representation of the function f by elements of the dictionary; it is emphasized that it is not to be confused with the L_1 norm of f. An alternate and closely related way of defining the $\mathcal{L}_1$ norm is by the infimum of numbers V for which f/V is in the closure of the convex hull of $\mathcal{D} \cup (-\mathcal{D})$. This is know as the "variation" of f with respect to $\mathcal{D}$, and was used in [16, 17], building on the earlier terminology in [3].

In the case where $\mathcal{D}$ is an orthonormal basis, we know that

$$\sigma_N(f) \leq \|f\|_{\mathcal{L}_1} N^{-1/2}, \quad f \in \mathcal{L}_1. \tag{1.10}$$

For the PGA, it was proved [10] that $f \in \mathcal{L}_1$ implies that

$$\|f - f_N\| \lesssim N^{-1/6}. \tag{1.11}$$

This rate was improved to $N^{-\frac{11}{62}}$ in [15], but on the other hand it was shown [19] that for a particular dictionary there exists $f \in \mathcal{L}_1$ such that

$$\|f - f_N\| \gtrsim N^{-0.27}. \tag{1.12}$$

When compared with (1.10), we see that the PGA is far from being optimal.

The RGA, OGA and SPA behave somewhat better: it was proved respectively in [14] for the RGA and SPA, and in [10] for the OGA, that one has

$$\|f - f_N\| \lesssim \|f\|_{\mathcal{L}_1} N^{-1/2}, \tag{1.13}$$

for all $f \in \mathcal{L}_1$.

For each of these algorithms, it is known that the convergence rate $N^{-1/2}$ cannot in general be improved even for functions which admit a very sparse expansion in the dictionary $\mathcal{D}$ (see [10] for such a result with a function being the sum of two elements of $\mathcal{D}$).

At this point, some remarks are in order regarding the meaning of the condition $f \in \mathcal{L}_1$ for some concrete dictionaries. A commonly made statement is that greedy algorithms break the *curse of dimensionality* in that the rate $N^{-1/2}$ is independent of the dimension d of the variable space for f, and only relies on the assumption that $f \in \mathcal{L}_1$. This is not exactly true since in practice the condition that $f \in \mathcal{L}_1$ becomes more and more stringent as d grows. For instance, in the case where we work in the Hilbert space $\mathcal{H} := L_2([0,1]^d)$ and when $\mathcal{D}$ is a *wavelet basis* (ψ_λ), it is known that the smoothness property which ensures that $f \in \mathcal{L}_1$ is that f should belong to the Besov space $B^s_1(L_1)$ with $s = d/2$, which roughly means that f has all its derivatives of order less or equal to $d/2$ in L_1 (see [9] for the characterization of Besov spaces by the properties wavelet coefficients). Moreover, for this to hold it is required that the dual wavelets $\tilde{\psi}_\lambda$ have at least $d/2 - 1$ vanishing moments. Another instance is the case where $\mathcal{D}$ consists of sigmoidal functions of the type $\sigma(v \cdot x - w)$ where σ is a fixed function and v and w are arbitrary vectors in $\mathbb{R}^d$. For such dictionaries, it was proved in [4] that a sufficient condition to have $f \in \mathcal{L}_1$ is the convergence of $\int |\omega||\mathcal{F}f(\omega)|d\omega$ where $\mathcal{F}$ is the Fourier operator. This integrability condition requires a larger amount of decay on the Fourier transform $\mathcal{F}f$ as d grows. Assuming that $f \in \mathcal{L}_1$ is therefore more and more restrictive as d grows. Similar remarks also hold for other dictionaries (hyperbolic wavelets, Gabor functions etc.).

1.2 Results of this paper

The discussion of the previous section points to a significant weakness in the present theory of greedy algorithms in that there are no viable bounds for the performance of greedy algorithms for general functions $f \in \mathcal{H}$. This is a severe impediment in some application domains (such as learning theory) where there is no a priori knowledge that would indicate that the target function is in $\mathcal{L}_1$. One of the main contributions of the present paper is to provide error bounds for the performance of greedy algorithms for general functions $f \in \mathcal{H}$. We shall focus our attention on the OGA and RGA, which as explained above have better convergence properties in $\mathcal{L}_1$ than the PGA. We shall consider the specific version of the RGA in which α_k is fixed to $1 - 1/k$ and (β_k, g_k) are optimized.

Inspection of the proofs in our paper show that all further approximation results proved for this version of the RGA also hold for *any* greedy algorithm such that

$$\|f - f_k\| \leq \min_{\beta,g} \|f - \alpha_k f_{k-1} + \beta g\|, \tag{1.14}$$

irrespective of how f_k is defined. In particular, they hold for the more general version of the RGA defined by (1.7) as well as for the SPA.

In §2, we introduce both algorithms and recall the optimal approximation rate $N^{-1/2}$ when the target function f is in $\mathcal{L}_1$. Later in this section, we develop a technique based on interpolation of operators that provides convergence rates N^{-s}, $0 < s < 1/2$, whenever f belongs to a certain intermediate space between $\mathcal{L}_1$ and the Hilbert space $\mathcal{H}$. Namely, we use the spaces

$$\mathcal{B}_p := [\mathcal{H}, \mathcal{L}_1]_{\theta,\infty}, \quad \theta := 2/p - 1, \quad 1 < p < 2, \tag{1.15}$$

which are the real interpolation spaces between $\mathcal{H}$ and $\mathcal{L}_1$. We show that if $f \in \mathcal{B}_p$, then the OGA and RGA, when applied to f, provide approximation rates CN^{-s} with $s := \theta/2 = 1/p - 1/2$. Thus, if we set $\mathcal{B}_1 = \mathcal{L}_1$, then these spaces provide a full range of approximation rates for greedy algorithms. Recall, as discussed previously, for general dictionaries, greedy algorithms will not provide convergence rates better than $N^{-1/2}$ for even the simplest of functions. The results we obtain are optimal in the sense that we recover the best possible convergence rate in the case where the dictionary is an orthonormal basis. For an arbitrary target function $f \in \mathcal{H}$, convergence of the OGA and RGA holds without rate. We also discuss in that section how the OGA can be monitored by a threshold parameter. Finally, we conclude the section by discussing several issues related to numerical implementation of these greedy algorithms. In particular, we consider the effect of reducing the dictionary $\mathcal{D}$ to a finite sub-dictionary.

In §3, we consider the learning problem under the assumption that the data $\mathbf{y} := (y_1, \dots, y_n)$ are bounded in absolute value by some fixed constant B. Our estimator is built on the application of the OGA or RGA to the noisy data $\mathbf{y}$ in the Hilbert space defined by the empirical norm

$$\|f\|_n := \frac{1}{n}\sum_{i=1}^{n} |f(x_i)|^2, \tag{1.16}$$

and its associated inner product. At each step k, the algorithm generates an approximation $\hat{f}_k$ to the data. Our estimator is defined by

$$\hat{f} := T\hat{f}_{k^*} \tag{1.17}$$

where

$$Tx := T_B x := \min\{B, |x|\}\text{sgn}(x) \tag{1.18}$$

is the truncation operator at level B and the value of k^* is selected by a complexity regularization procedure. The main result for this estimator is (roughly) that when the regression function f_ρ is in $\mathcal{B}_p$ (where this space is defined with respect to the norm $\|u\|^2 := E(|u(x)|^2)$), the estimator has convergence rate

$$E(\|\hat{f} - f_\rho\|^2) \lesssim \left(\frac{n}{\log n}\right)^{-\frac{2s}{1+2s}}, \tag{1.19}$$

again with $s := 1/p - 1/2$. In the case where $f_\rho \in \mathcal{L}_1$, we obtain the same result with $p = 1$ and $s = 1/2$. We also show that this estimator is universally consistent.

In order to place these results into the current state of the art of statistical learning theory, let us first remark that similar convergence rate for the denoising and the learning problem could be obtained by a more "brute force" approach which would consist in selecting a proper subset of $\mathcal{D}$ by complexity regularization with techniques such as those in [2] or Chapter 12 of [13]. Following for instance the general approach of [13], this would typically first require restricting the size of the dictionary $\mathcal{D}$ (usually to be of size $O(n^a)$ for some $a > 1$) and then considering all possible subsets $\Lambda \subset \mathcal{D}$ and spaces $\mathcal{G}_\Lambda := \text{Span}\{g \in \Lambda\}$, each of them defining an estimator

$$\hat{f}_\Lambda := T_B\Big(\text{Argmin}_{f\in\mathcal{G}_\Lambda}\|y - f\|_n^2\Big) \tag{1.20}$$

The estimator $\hat{f}$ is then defined as the $\hat{f}_\Lambda$ which minimizes

$$\min_{\Lambda\subset\mathcal{D}}\{\|y - \hat{f}_\Lambda\|_n^2 + \text{Pen}(\Lambda, n)\} \tag{1.21}$$

with $\text{Pen}(\Lambda, n)$ a complexity penalty term. The penalty term usually restricts the size of Λ to be at most $\mathcal{O}(n)$ but even then the search is over $O(n^{an})$ subsets. In some other approaches, the sets $\mathcal{G}_\Lambda$ might also be discretized, transforming the subproblem of selecting $\hat{f}_\Lambda$ into a discrete optimization problem.

The main advantage of using the greedy algorithm in place of (1.21) for constructing the estimator is a dramatic reduction of the computational cost. Indeed, instead of considering all possible subsets $\Lambda \subset \mathcal{D}$ the algorithm only considers the sets $\Lambda_k := \{g_1, \cdots, g_k\}$, $k = 1, \ldots, n$, generated by the empirical greedy algorithm.

This approach was proposed and analyzed in [18] using a version of the RGA in which

$$\alpha_k + \beta_k = 1 \tag{1.22}$$

which implies that the approximation f_k at each iteration stays in the convex hull $\mathcal{C}_1$ of $\mathcal{D}$. The authors established that if f does not belong to $\mathcal{C}_1$, the RGA converges to its projection onto $\mathcal{C}_1$, In turn, the estimator was proved to converge in the sense of (1.19) to f_ρ, with rate $(n/\log n)^{-1/2}$, if f_ρ lies in $\mathcal{C}_1$, and otherwise to its projection onto $\mathcal{C}_1$. In that sense, this procedure is not universally consistent.

One of the main contributions of the present paper is to remove requirements of the type $f_\rho \in \mathcal{L}_1$ when obtaining convergence rates. In the learning context, there is indeed typically no advanced information that would guarantee such restrictions on f_ρ. The estimators that we construct for learning are now universally consistent and have provable convergence rates for more general regression functions described by means of interpolation spaces. One of the main ingredient in our analysis of the performance of our greedy algorithms in learning is a powerful exponential concentration inequality which was introduced in [18]. Let us mention that a closely related analysis, which however does not involve interpolation spaces, was developed in [6, 5].

The most studied dictionaries in learning theory are in the context of neural networks. In §4 we interpret our results in this setting and in particular describe the smoothness conditions on a function f which ensure that it belongs to the spaces $\mathcal{L}_1$ or $\mathcal{B}_p$.

2 Approximation properties

Let $\mathcal{D}$ be a dictionary in some Hilbert space $\mathcal{H}$, with $\|g\| = 1$ for all $g \in \mathcal{D}$. We recall that, for a given $f \in \mathcal{H}$, the OGA builds embedded approximation spaces

$$V_k := \operatorname{Span}\{g_1, \cdots, g_k\}, \quad k = 1, 2, \ldots, \tag{2.1}$$

and approximations

$$f_k := P_k f \tag{2.2}$$

where P_k is the orthogonal projection onto V_k. The rule for generating the g_k is as follows. We set $V_0 := \{0\}$, $f_0 := 0$ and $r_0 := f$, and given V_{k-1}, $f_{k-1} = P_{k-1} f$ and $r_{k-1} := f - f_{k-1}$, we define g_k by

$$g_k := \operatorname*{Argmax}_{g \in G} |\langle r_{k-1}, g \rangle|, \tag{2.3}$$

which defines the new V_k, f_k and r_k.

In its most general form, the RGA sets

$$f_k = \alpha_k f_{k-1} + \beta_k g_k, \tag{2.4}$$

where (α_k, β_k, g_k) are defined according to (1.7). We shall consider a simpler version in which the first parameter is a fixed sequence. Two choices will be considered, namely

$$\alpha_k = 1 - \frac{1}{k}, \tag{2.5}$$

and

$$\alpha_k = 1 - \frac{2}{k} \ \text{ if } \ k > 1, \ \ \alpha_1 = 0. \tag{2.6}$$

The two other parameters are optimized according to

$$(\beta_k, g_k) := \operatorname*{Argmin}_{(\beta, g) \in \mathbb{R} \times \mathcal{D}} \|f - \alpha_k f_{k-1} - \beta g\|. \tag{2.7}$$

Since

$$\|f-\alpha_k f_{k-1}-\beta g\|^2=\beta^2-2\beta\langle f-\alpha_k f_{k-1},g\rangle+\|f-\alpha_k f_{k-1}\|^2, \tag{2.8}$$

it is readily seen that (β_k, g_k) are given explicitely by

$$\beta_k=\langle f-\alpha_k f_{k-1},g_k\rangle, \tag{2.9}$$

and

$$g_k:=\operatorname*{Argmax}_{g\in\mathcal{D}}|\langle f-\alpha_k f_{k-1},g\rangle|. \tag{2.10}$$

Therefore, from a computational point of view this RGA is very similar to the PGA.

We denote by $\mathcal{L}_p$ the functions f which admit a converging expansion $f=\sum c_g g$ with $\sum |c_g|^p<+\infty$, and we write $\|f\|_{\mathcal{L}_p}=\inf\|(c_g)\|_{\ell_p}$ where the infimum is taken over all such expansions. In a similar way, we consider the spaces $w\mathcal{L}_p$ corresponding to expansions which are in the weak space $w\ell_p$. We denote by $\sigma_N(f)$ the best N-term approximation error in the $\mathcal{H}$ norm for f and for any $s>0$ define the approximation space

$$\mathcal{A}^s:=\{f\in\mathcal{H}:\ \sigma_N(f)\le MN^{-s},\ N=1,2,\ldots\}. \tag{2.11}$$

Finally, we denote by $\mathcal{G}^s$ the set of functions f such that the greedy algorithm under consideration converges with rate $\|r_N\|\lesssim N^{-s}$, so that obviously $\mathcal{G}^s\subset\mathcal{A}^s$.

In the case $\mathcal{D}$ is an orthonormal basis, the space $\mathcal{A}^s$ contains the space $\mathcal{L}_p$ with $1/p=1/2+s$, and in fact actually coincides with the weak versions $w\mathcal{L}_p$ of these spaces. In those cases, an algorithm for building a best (or near best) N-term approximation is simply to keep the N largest coefficients of f and discard the others. The best N-term approximation is also obtained by the orthogonal greedy algorithm so that obviously $\mathcal{A}^s=\mathcal{G}^s$.

2.1 Approximation of $\mathcal{L}_1$ functions

In this section, we recall for convenience the approximation properties of the OGA and RGA for functions $f\in\mathcal{L}_1$. We first recall the result obtained in [10] for the OGA. We shall make use of the following fact: if $f,g\in\mathcal{H}$ with $\|g\|=1$, then $\langle f,g\rangle g$ is the best approximation to f from the one dimensional space generated by g and

$$\|f-\langle f,g\rangle g\|^2=\|f\|^2-|\langle f,g\rangle|^2. \tag{2.12}$$

Theorem 2.1 *For all $f\in\mathcal{L}_1$ the error of the OGA satisfies*

$$\|r_N\|\le\|f\|_{\mathcal{L}_1}(N+1)^{-1/2},\quad N=1,2,\ldots. \tag{2.13}$$

Proof: Since f_k is the best approximation to f from V_k, we have from (2.12)

$$\|r_k\|^2\le\|r_{k-1}-\langle r_{k-1},g_k\rangle g_k\|^2=\|r_{k-1}\|^2-|\langle r_{k-1},g_k\rangle|^2, \tag{2.14}$$

with equality in the case where g_k is orthogonal to V_{k-1}. Since r_{k-1} is orthogonal to f_{k-1}, we have

$$\|r_{k-1}\|^2=\langle r_{k-1},f\rangle\le\|f\|_{\mathcal{L}_1}\sup_{g\in\mathcal{D}}|\langle r_{k-1},g\rangle|=\|f\|_{\mathcal{L}_1}|\langle r_{k-1},g_k\rangle|, \tag{2.15}$$

which combined with (2.14) gives the reduction property

$$\|r_k\|^2 \leq \|r_{k-1}\|^2(1 - \|r_{k-1}\|^2\|f\|_{\mathcal{L}_1}^{-2}). \tag{2.16}$$

We also know that $\|r_1\| \leq \|r_0\| = \|f\| \leq \|f\|_{\mathcal{L}_1}$.

We then check by induction that a decreasing sequence $(a_n)_{n\geq 0}$ of nonnegative numbers which satisfy $a_0 \leq M$ and $a_k \leq a_{k-1}(1-\frac{a_{k-1}}{M})$ for all $k > 0$ has the decay property $a_n \leq \frac{M}{n+1}$ for all $n \geq 0$. Indeed, assuming $a_{n-1} \leq \frac{M}{n}$ for some $n > 0$, then either $a_{n-1} \leq \frac{M}{n+1}$ so that $a_n \leq \frac{M}{n+1}$, or else $a_{n-1} \geq \frac{M}{n+1}$ so that

$$a_n \leq \frac{M}{n}(1 - \frac{1}{n+1}) = \frac{M}{n+1}. \tag{2.17}$$

The result follows by applying this to $a_k = \|r_k\|^2$ and $M := \|f\|_{\mathcal{L}_1}^2$, since we indeed have

$$a_0 = \|f\|^2 \leq \|f\|_{\mathcal{L}_1}^2. \tag{2.18}$$

□

We now turn to the RGA for which we shall prove a slightly stronger property.

Theorem 2.2 *For all $f \in \mathcal{L}_1$ the error of the RGA using* (2.5) *satisfies*

$$\|r_N\| \leq (\|f\|_{\mathcal{L}_1}^2 - \|f\|^2)^{1/2} N^{-1/2}, \quad N = 1, 2, \ldots. \tag{2.19}$$

Proof: From the definition of the RGA, we see that the sequence f_k remains unchanged if the dictionary $\mathcal{D}$ is symmetrized by including the sequence $(-g)_{g\in\mathcal{D}}$. Under this assumption, since $f \in \mathcal{L}_1$, for any $\epsilon > 0$ we can expand f according to

$$f = \sum_{g\in\mathcal{D}} b_g g, \tag{2.20}$$

where all the b_g are non-negative and satisfy

$$\sum_{g\in\mathcal{D}} b_g = \|f\|_{\mathcal{L}_1} + \delta, \tag{2.21}$$

with $0 \leq \delta \leq \epsilon$. According to (2.7), we have for all $\beta \in \mathbb{R}$ and all $g \in \mathcal{D}$

$$\begin{aligned} \|r_k\|^2 &\leq \|f - \alpha_k f_{k-1} - \beta g\|^2 \\ &= \|\alpha_k r_{k-1} + \tfrac{1}{k} f - \beta g\|^2 \\ &= \alpha_k^2 \|r_{k-1}\|^2 - 2\alpha_k \langle r_{k-1}, \tfrac{1}{k} f - \beta g\rangle + \|\tfrac{1}{k} f - \beta g\|^2 \\ &= \alpha_k^2 \|r_{k-1}\|^2 - 2\alpha_k \langle r_{k-1}, \tfrac{1}{k} f - \beta g\rangle + \tfrac{1}{k^2}\|f\|^2 + \beta^2 - \tfrac{2\beta}{k}\langle f, g\rangle. \end{aligned}$$

This inequality holds for all $g \in \mathcal{D}$, so it also holds on the average with weights $\frac{b_g}{\sum_{g\in\mathcal{D}} b_g}$, which gives for the particular value $\beta = \frac{1}{k}(\|f\|_{\mathcal{L}_1} + \delta)$,

$$\|r_k\|^2 \leq \alpha_k^2 \|r_{k-1}\|^2 - \frac{1}{k^2}\|f\|^2 + \beta^2. \tag{2.22}$$

Therefore, letting ϵ go to 0, we obtain

$$\|r_k\|^2 \le (1-\frac{1}{k})^2\|r_{k-1}\|^2 + \frac{1}{k^2}(\|f\|_{\mathcal{L}_1}^2 - \|f\|^2). \tag{2.23}$$

We now check by induction that a sequence $(a_k)_{k>0}$ of positive numbers such that $a_1 \le M$ and $a_k \le (1-\frac{1}{k})^2 a_{k-1} + \frac{1}{k^2}M$ for all $k > 0$ has the decay property $a_n \le \frac{M}{n}$ for all $n > 0$. Indeed, assuming that $a_{n-1} \le \frac{M}{n-1}$, we write

$$\begin{aligned} a_n - \tfrac{M}{n} &\le (1-\tfrac{1}{n})^2\tfrac{M}{n-1} + \tfrac{1}{n^2}M - \tfrac{M}{n} \\ &= M(\tfrac{n-1}{n^2} + \tfrac{1}{n^2} - \tfrac{1}{n}) = 0. \end{aligned}$$

The result follow by applying this to $a_k = \|r_k\|^2$ and $M := \|f\|_{\mathcal{L}_1}^2 - \|f\|^2$, since (2.23) also implies that $a_1 \le M$. □

The above results show that for both OGA and RGA we have

$$\mathcal{L}_1 \subset \mathcal{G}^{1/2} \subset \mathcal{A}^{1/2}. \tag{2.24}$$

From this, it also follows that $w\mathcal{L}_p \subset \mathcal{A}^s$ with $s = 1/p - 1/2$ when $p < 1$. Indeed, from the definition of $w\mathcal{L}_p$, any function f in this space can be written as $f = \sum_{j=1}^{\infty} c_j g_j$ with each $g_j \in \mathcal{D}$ and the coefficients c_j decreasing in absolute value and satisfying $|c_j| \le Mj^{-1/p}$, $j \ge 1$ with $M := \|f\|_{w\mathcal{L}_p}$. Therefore, $f = f_a + f_b$ with $f_a := \sum_{j=1}^{N} c_j g_j$ and $\|f_b\|_{\mathcal{L}_1} \le C_p M N^{1-1/p}$. It follows from Theorem 2.1 for example, that f_b can be approximated by an N-term expansion obtained by the greedy algorithm with accuracy $\|f_b - P_N f_b\| \le C_p M N^{1/2-1/p} = N^{-s}$, and therefore by taking $f_a + P_N f_b$ as a $2N$ term approximant of f, we obtain that $f \in \mathcal{A}^s$. Observe, however, that this does not mean that $f \in \mathcal{G}^s$ in the sense that we have not proved that the greedy algorithm converges with rate N^{-s} when applied to f. It is actually shown in [10] that there exists simple dictionaries such that the greedy algorithm does not converge faster than $N^{-1/2}$, even for functions f which are in $w\mathcal{L}_p$ for all $p > 0$.

2.2 Approximation of general functions

We now want to study the behaviour of the OGA and RGA when the function $f \in \mathcal{H}$ is more general in the sense that it is less sparse than being in $\mathcal{L}_1$. The simplest way of expressing this would seem to be by considering the spaces $\mathcal{L}_p$ or $w\mathcal{L}_p$ with $1 < p < 2$. However for general dictionaries, these spaces are not well adapted, since $\|f\|_{\mathcal{L}_p}$ does not control the Hilbert norm $\|f\|$.

Instead, we shall consider the real interpolation spaces

$$\mathcal{B}_p = [\mathcal{H}, \mathcal{L}_1]_{\theta,\infty}, \quad 0 < \theta < 1, \tag{2.25}$$

with again p defined by $1/p = \theta + (1-\theta)/2 = (1+\theta)/2$. Recall that $f \in [X, Y]_{\theta,\infty}$ if and only if for all $t > 0$, we have

$$K(f,t) \le Ct^{\theta}, \tag{2.26}$$

where

$$K(f,t) := K(f,t,X,Y) := \inf_{h \in Y} \{\|f-h\|_X + t\|h\|_Y\}, \tag{2.27}$$

is the so-called K-functional. In other words, f can be decomposed into $f = f_X + f_Y$ with

$$\|f_X\|_X + t\|f_Y\|_Y \le Ct^\theta. \tag{2.28}$$

The smallest C such that the above holds defines a norm for $Z = [X,Y]_{\theta,\infty}$. We refer to [8] or [7] for an introduction to interpolation spaces. The space $\mathcal{B}_p$ coincides with $w\mathcal{L}_p$ in the case when $\mathcal{D}$ is an orthonormal system but may differ from it for a more general dictionary.

The main result of this section is that for both the OGA and the RGA,

$$\|r_N\| \le C_0 K(f, N^{-1/2}, \mathcal{H}, \mathcal{L}_1), \quad N = 1,2,\dots, \tag{2.29}$$

so that, according to (2.26), $f \in \mathcal{B}_p$ implies the rate of decay $\|r_N\| \lesssim N^{-\theta/2}$. Note that if f_N were obtained as the action on f of a continuous linear operator L_N from $\mathcal{H}$ onto itself such that $\|L_N\| \le C$ with C independent of k, then we could write for any $h \in \mathcal{L}_1$

$$\|f - f_N\| \le \|(I - L_N)[f-h]\| + \|h - L_N h\| \lesssim \|f-h\| + \|h\|_{\mathcal{L}_1} N^{-1/2}, \tag{2.30}$$

so that (2.29) would follow by minimizing over $h \in \mathcal{L}_1$. However, f_N is obtained by a highly nonlinear algorithm and it is therefore quite remarkable that (2.29) still holds. We first prove this for the OGA.

Theorem 2.3 *For all $f \in \mathcal{H}$ and any $h \in \mathcal{L}_1$ the error of the OGA satisfies*

$$\|r_N\|^2 \le \|f-h\|^2 + 4\|h\|_{\mathcal{L}_1}^2 N^{-1}, \quad N = 1,2,\dots, \tag{2.31}$$

and therefore

$$\|r_N\| \le K(f, 2N^{-1/2}, \mathcal{H}, \mathcal{L}_1) \le 2K(f, N^{-1/2}, \mathcal{H}, \mathcal{L}_1), \quad N = 1,2,\dots. \tag{2.32}$$

Proof: Fix an arbitrary $f \in \mathcal{H}$. For any $h \in \mathcal{L}_1$, we write

$$\|r_{k-1}\|^2 = \langle r_{k-1}, h + f - h\rangle \le \|h\|_{\mathcal{L}_1} |\langle r_{k-1}, g_k\rangle| + \|r_{k-1}\|\,\|f-h\| \tag{2.33}$$

from which it follows that

$$\|r_{k-1}\|^2 \le \|h\|_{\mathcal{L}_1} |\langle r_{k-1}, g_k\rangle| + \frac{1}{2}(\|r_{k-1}\|^2 + \|f-h\|^2). \tag{2.34}$$

Therefore, using the shorthand notation $a_k := \|r_k\|^2 - \|f-h\|^2$, we have

$$|\langle r_{k-1}, g_k\rangle| \ge \frac{a_{k-1}}{2\|h\|_{\mathcal{L}_1}}. \tag{2.35}$$

Note that if for some k_0 we have $\|r_{k_0-1}\| \le \|f-h\|$, then the theorem holds trivially for all $N \ge k_0 - 1$. We therefore assume that a_{k-1} is positive, so that we can write

$$|\langle r_{k-1}, g_k\rangle|^2 \ge \frac{a_{k-1}^2}{4\|h\|_{\mathcal{L}_1}^2}. \tag{2.36}$$

From (2.14), we therefore obtain

$$\|r_k\|^2 \leq \|r_{k-1}\|^2 - \frac{a_{k-1}^2}{4\|h\|_{\mathcal{L}_1}^2}, \tag{2.37}$$

which by substracting $\|f-h\|^2$ gives

$$a_k \leq a_{k-1}\left(1 - \frac{a_{k-1}}{4\|h\|_{\mathcal{L}_1}^2}\right). \tag{2.38}$$

As in the proof of Theorem 2.1, we can conclude that $a_N \leq 4\|h\|_{\mathcal{L}_1}^2 N^{-1}$ provided that we have initially $a_1 \leq 4\|h\|_{\mathcal{L}_1}^2$. In order to check this initial condition, we remark that either $a_0 \leq 4\|h\|_{\mathcal{L}_1}^2$ so that the same holds for a_1, or $a_0 \geq 4\|h\|_{\mathcal{L}_1}^2$, in which case $a_1 \leq 0$ according to (2.38), which means that we are already in the trivial case $\|r_1\| \leq \|f-h\|$ for which there is nothing to prove. We have therefore obtained (2.31) and (2.32) follows by taking the square root. □

We next treat the case of the RGA for which we have a slightly different result. In this result, we use the second choice (2.6) for the sequence α_k in order to obtain a multiplicative constant equal to 1 in the term $\|f-h\|^2$ appearing on the right side the quadratic bound. This will be important in the learning application. We also give the non-quadratic bound with the first choice (2.5), since it gives a slightly better result than by taking the square root of the quadratic bound based on (2.6).

Theorem 2.4 *For all $f \in \mathcal{H}$ and any $h \in \mathcal{L}_1$ the error of the RGA using* (2.6) *satisfies*

$$\|r_N\|^2 \leq \|f-h\|^2 + 4(\|h\|_{\mathcal{L}_1}^2 - \|h\|^2)N^{-1}, \quad N = 1, 2, \ldots. \tag{2.39}$$

and therefore

$$\|r_N\| \leq K(f, 2N^{-1/2}, \mathcal{H}, \mathcal{L}_1) \leq 2K(f, N^{-1/2}, \mathcal{H}, \mathcal{L}_1), \quad N = 1, 2, \ldots. \tag{2.40}$$

Using the first choice (2.5)*, the error satisfies*

$$\|r_N\| \leq \|f-h\| + (\|h\|_{\mathcal{L}_1}^2 - \|h\|^2)^{1/2} N^{-1/2}, \quad N = 1, 2, \ldots, \tag{2.41}$$

and therefore

$$\|r_N\| \leq K(f, N^{-1/2}, \mathcal{H}, \mathcal{L}_1), \quad N = 1, 2, \ldots. \tag{2.42}$$

Proof: Fix $f \in \mathcal{H}$ and let $h \in \mathcal{L}_1$ be arbitrary. Similar to the proof of Theorem 2.2, for any $\epsilon > 0$, we can expand h

$$h = \sum_{g \in \mathcal{D}} b_g g, \tag{2.43}$$

where all the b_g are non-negative and satisfy

$$\sum_{g \in \mathcal{D}} b_g = \|h\|_{\mathcal{L}_1} + \delta, \tag{2.44}$$

with $0 \le \delta \le \epsilon$. Using the notation $\bar\alpha_k = 1-\alpha_k$, we have for all $\beta \in \mathbb{R}$ and all $g \in \mathcal{D}$

$$\begin{array}{rcl} \|r_k\|^2 & \le & \|f-\alpha_k f_{k-1} - \beta g\|^2 \\ & = & \|\alpha_k r_{k-1} + \bar\alpha_k f - \beta g\|^2 \\ & = & \alpha_k^2\|r_{k-1}\|^2 - 2\alpha_k\langle r_{k-1}, \bar\alpha_k f - \beta g\rangle + \|\bar\alpha_k f - \beta g\|^2 \\ & = & \alpha_k^2\|r_{k-1}\|^2 - 2\alpha_k\langle r_{k-1}, \bar\alpha_k f - \beta g\rangle + \|\bar\alpha_k (f-h) + \bar\alpha_k h - \beta g\|^2 \\ & = & \alpha_k^2\|r_{k-1}\|^2 - 2\alpha_k\langle r_{k-1}, \bar\alpha_k f - \beta g\rangle + \bar\alpha_k^2\|f-h\|^2 + 2\bar\alpha_k\langle f-h, \bar\alpha_k h - \beta g\rangle \\ & & +\bar\alpha_k^2\|h\|^2 - 2\beta\bar\alpha_k\langle h, g\rangle + \beta^2. \end{array}$$

This inequality holds for all $g \in \mathcal{D}$, so it also holds on the average with weights $\frac{b_g}{\sum_{g\in\mathcal{D}} b_g}$, which gives for the particular value $\beta = \bar\alpha_k(\|h\|_{\mathcal{L}_1} + \delta)$,

$$\begin{array}{rl} \|r_k\|^2 & \le \alpha_k^2\|r_{k-1}\|^2 - 2\alpha_k\bar\alpha_k\langle r_{k-1}, f-h\rangle + \bar\alpha_k^2\|f-h\|^2 - \bar\alpha_k^2\|h\|^2 + \beta^2 \\ & = \|\alpha_k r_{k-1} - \bar\alpha_k(f-h)\|^2 - \bar\alpha_k^2\|h\|^2 + \beta^2. \end{array}$$

Letting ϵ tend to 0 and using the notation $M := \|h\|_{\mathcal{L}_1}^2 - \|h\|^2$, we thus obtain

$$\|r_k\|^2 \le (\alpha_k\|r_{k-1}\| + \bar\alpha_k\|f-h\|)^2 + \bar\alpha_k^2 M. \tag{2.45}$$

Note that this immediately implies the validity of (2.39) and (2.41) at $N=1$, using that $\alpha_1 = 0$ for both choices (2.5) and (2.6). We next proceed by induction, assuming that these bounds hold at $k-1$.

For the proof of (2.41) we derive from (2.45)

$$\begin{array}{rl} \|r_k\|^2 & \le (\alpha_k(\|f-h\| + (\frac{M}{k-1})^{1/2}) + \bar\alpha_k\|f-h\|)^2 + \bar\alpha_k^2 M \\ & = (\|f-h\| + \alpha_k(\frac{M}{k-1})^{1/2})^2 + \bar\alpha_k^2 M \\ & = \|f-h\|^2 + 2M^{1/2}\|f-h\|\frac{1-\frac{1}{k}}{\sqrt{k-1}} + M(\frac{(1-\frac{1}{k})^2}{k-1} + \frac{1}{k^2}) \\ & = \|f-h\|^2 + 2M^{1/2}\|f-h\|\frac{\sqrt{k-1}}{k} + \frac{M}{k} \\ & \le \|f-h\|^2 + 2(\frac{M}{k})^{1/2}\|f-h\| + \frac{M}{k} \\ & = (\|f-h\| + (\frac{M}{k})^{1/2})^2, \end{array}$$

which is the desired bound at k.

For the proof of (2.39), we derive from (2.45)

$$\|r_k\|^2 \le \alpha_k^2\|r_{k-1}\|^2 + 2\alpha_k\bar\alpha_k\|r_{k-1}\|\,\|f-h\| + \bar\alpha_k^2\|f-h\|^2 + \bar\alpha_k^2 M. \tag{2.46}$$

Remarking that $2\alpha_k\bar\alpha_k\|r_{k-1}\|\,\|f-h\| \le \alpha_k\bar\alpha_k(\|r_{k-1}\|^2 + \|f-h\|^2)$, we obtain

$$\|r_k\|^2 \le \alpha_k\|r_{k-1}\|^2 + \bar\alpha_k\|f-h\|^2 + \bar\alpha_k^2 M, \tag{2.47}$$

and therefore for

$$\|r_k\|^2 - \|f-h\|^2 \le \alpha_k(\|r_{k-1}\|^2 - \|f-h\|^2) + \bar\alpha_k^2 M. \tag{2.48}$$

We now check by induction that a sequence $(a_k)_{k>0}$ of positive numbers such that $a_1 \le 4M$ and $a_k \le (1-\frac{2}{k})a_{k-1} + \frac{4}{k^2}M$ for all $k>1$ has the decay property $a_n \le \frac{4M}{n}$ for all $n>0$. Indeed, assuming that $a_{n-1} \le \frac{4M}{n-1}$, we obtain

$$\begin{array}{rl} a_n - \frac{4M}{n} & \le (1-\frac{2}{n})\frac{4M}{n-1} + \frac{4}{n^2}M - \frac{4M}{n} \\ & = M(\frac{4(n-2)-4(n-1)}{n(n-1)} + \frac{4}{n^2}) \\ & = M(\frac{4}{n^2} - \frac{4}{n(n-1)}) \le 0. \end{array}$$

Applying this with $a_k = \|r_k\|^2 - \|f-h\|^2$, we thus obtain (2.39) and (2.40) follows by taking the square root. □

An immediate consequence of Theorems 2.3 and 2.4 combined with the definition of the $\mathcal{B}_p$ spaces (see (2.26)) is a rate of convergence of the OGA and RGA for the functions in $\mathcal{B}_p$.

Corollary 2.5 *For all $f \in \mathcal{B}_p$, the approximation error for both the OGA and RGA satisfy the decay bound*

$$\|r_N\| \ \lesssim\ \|f\|_{\mathcal{B}_p} N^{-s}, \tag{2.49}$$

with $s = 1/p - 1/2$. Therefore we have $\mathcal{B}_p \subset \mathcal{G}^s \subset \mathcal{A}^s$.

In addition, when $\mathcal{D}$ is a complete family in $\mathcal{H}$ we know that $\mathcal{L}_1$ is dense in $\mathcal{H}$ so that

$$\lim_{t\to 0} K(f,t,\mathcal{H},\mathcal{L}_1) = 0, \tag{2.50}$$

for any $f \in \mathcal{H}$. This implies the following corollary.

Corollary 2.6 *For any $f \in \mathcal{H}$, the approximation error $\|r_N\|$ goes to zero as $N \to +\infty$ for both the OGA and RGA.*

2.3 Orthogonal greedy algorithm and thresholding

We shall show in this section that the OGA can be monitored by thresholding in a similar way to an orthonormal basis. We set a threshold $t > 0$ and iterate the orthogonal greedy algorithm as long as $|\langle r_{k-1}, g_k\rangle| > t$. We define k_t as the smallest k such that $|\langle r_k, g_{k+1}\rangle| \le t$.

In the case where $\mathcal{D}$ is an orthonormal basis, we know that whenever the coefficients of f are in $w\ell_p$, then

$$k_t \le \|f\|^p_{w\mathcal{L}_p} t^{-p}, \tag{2.51}$$

and

$$\|r_{k_t}\| \ \lesssim\ \|f\|^{p/2}_{w\mathcal{L}_p} t^{1-p/2}. \tag{2.52}$$

The following result for a general dictionary $\mathcal{D}$ shows that when $f \in \mathcal{L}_1$ or $f \in \mathcal{B}_p$ with $1 < p < 2$, both the complexity k_t and the error $\|r_{k_t}\|$ can be controlled in a similar way by the threshold t.

Theorem 2.7 *Let $\mathcal{D}$ be an arbitrary dictionary. If $f \in \mathcal{B}_p$, $1 < p < 2$, then*

$$k_t \ \lesssim\ \|f\|^p_{\mathcal{B}_p} t^{-p}, \tag{2.53}$$

and

$$\|r_{k_t}\| \ \lesssim\ \|f\|^{p/2}_{\mathcal{B}_p} t^{1-p/2}. \tag{2.54}$$

If $f \in \mathcal{L}_1$, the same result holds with $p = 1$.

Remark 2.8 *Notice that the bounds in the theorem show that for k_t iterations of the greedy algorithm we obtain $\|r_{k_t}\| \le Ck_t^{-s}$ with $s := 1/p - 1/2$. Thus, thresholding produces the same approximation rates as given in Corollary 2.5.*

Proof of Theorem: We first prove (2.53). Since

$$|\langle r_{k-1}, g_k\rangle|^2 \le \|r_{k-1}\|^2 - \|r_k\|^2, \tag{2.55}$$

it follows that

$$\sum_{l>k} |\langle r_{l-1}, g_l\rangle|^2 \le \|r_k\|^2. \tag{2.56}$$

Using Corollary 2.5, we have for $s = 1/p - 1/2$,

$$\frac{k_t}{2} t^2 < \sum_{\frac{k_t}{2} < l \le k_t} |\langle r_{l-1}, g_l\rangle|^2 \le \|r_{k_t/2}\|^2 \lesssim \|f\|_{\mathcal{B}_p}^2 k_t^{-2s}. \tag{2.57}$$

From this, we derive

$$k_t^{1+2s} \lesssim \|f\|_{\mathcal{B}_p}^2 t^{-2}, \tag{2.58}$$

which is equivalent to (2.53). The same argument applies in the case $p = 1$.

We next prove (2.54). In the case $p = 1$, we already know from (2.15) that

$$\|r_{k_t}\|^2 \le \|f\|_{\mathcal{L}_1} |\langle r_{k_t}, g_{k_t+1}\rangle|. \tag{2.59}$$

Since $|\langle r_{k_t}, g_{k_t+1}\rangle| \le t$, it follows that

$$\|r_{k_t}\| \le \|f\|_{\mathcal{L}_1}^{1/2} t^{1/2}. \tag{2.60}$$

In the case $1 < p < 2$, we know $f \in \mathcal{B}_p$ implies that

$$K(f, u) \le \|f\|_{\mathcal{B}_p} u^\theta, \quad u \ge 0, \tag{2.61}$$

with $\theta = 2/p - 1$. Therefore, taking u such that $\|f\|_{\mathcal{B}_p} u^\theta = \|r_{k_t}\|/2$, we know that there exists $h \in \mathcal{L}_1$ such that $\|h - f\| \le \|r_{k_t}\|/2$ and $u\|h\|_{\mathcal{L}_1} \le \|r_{k_t}\|/2$. From (2.33), we obtain

$$\|r_{k_t}\|^2 \le 2\|h\|_{\mathcal{L}_1} |\langle r_{k_t}, g_{k_t+1}\rangle| \lesssim \|r_{k_t}\|^{1-1/\theta} \|f\|_{\mathcal{B}_p}^{1/\theta} |\langle r_{k_t}, g_{k_t+1}\rangle|, \tag{2.62}$$

and therefore

$$\|r_{k_t}\|^{1+1/\theta} \lesssim \|f\|_{\mathcal{B}_p}^{1/\theta} |\langle r_{k_t}, g_{k_t+1}\rangle|, \tag{2.63}$$

from which (2.54) follows since $|\langle r_{k_t}, g_{k_t+1}\rangle| \le t$. □

2.4 Greedy algorithms with a truncated dictionary

In concrete applications it is not possible to evaluate the supremum of $|\langle r_{k-1}, g\rangle|$ over the whole dictionary $\mathcal{D}$, but only over a finite subset of it. For applications in learning theory, it will also be useful that the size of this subset has at most polynomial growth in the number of samples n. We therefore introduce a fixed exhaustion of $\mathcal{D}$,

$$\mathcal{D}_1 \subset \mathcal{D}_2 \subset \cdots \subset \mathcal{D} \tag{2.64}$$

with $\#(\mathcal{D}_m) = m$. The analysis we present in this section is similar to that in [23]. We are now interested in the functions f which can be approximated at a certain accuracy by application of the OGA only using the elements of $\mathcal{D}_m$. For this purpose, we first introduce the space $\mathcal{L}_1(\mathcal{D}_m)$ of those functions in $\mathrm{Span}(\mathcal{D}_m)$ equipped with the (minimal) ℓ_1 norm of the coefficients. We next define for $r > 0$ the space $\mathcal{L}_{1,r}$ as the set of all functions f such that for all m, there exists h (depending on m) such that

$$\|h\|_{\mathcal{L}_1(\mathcal{D}_m)} \leq C, \tag{2.65}$$

and

$$\|f - h\| \leq Cm^{-r}. \tag{2.66}$$

The smallest constant C such that this holds defines a norm for $\mathcal{L}_{1,r}$. In order to understand how these spaces are related to the space $\mathcal{L}_1$ for the whole dictionary consider the example where $\mathcal{D}$ is a Schauder basis, and consider the decomposition of f into

$$f = \sum_{g\in\mathcal{D}_m} c_g g + \sum_{g\notin\mathcal{D}_m} c_g g = h + f - h. \tag{2.67}$$

Then it is obvious that $\|h\|_{\mathcal{L}_1(\mathcal{D}_m)} \leq \|f\|_{\mathcal{L}_1}$. Therefore, a sufficient condition for f to be in $\mathcal{L}_{1,r}$ is $f \in \mathcal{L}_1$ and its tail $\|\sum_{g\notin\mathcal{D}_m} c_g g\|$ decays like m^{-r}.

Application of Theorems 2.3 and 2.4 shows us that if we apply the OGA or RGA with the restricted dictionary and if the target function f is in $\mathcal{L}_{1,r}$ we have

$$\|r_k\| \leq C_0\|f\|_{\mathcal{L}_{1,r}}(k^{-1/2} + m^{-r}), \tag{2.68}$$

where C_0 is an absolute constant ($C_0 = 2$ for OGA and $C_0 = 1$ for RGA with choice (2.5)).

In a similar manner, we can introduce the interpolation space

$$\mathcal{B}_{p,r} := [\mathcal{H}, \mathcal{L}_{1,r}]_{\theta,\infty}, \tag{2.69}$$

with again $1/p = (1 + \theta)/2$. From the definition of interpolation spaces, if $f \in \mathcal{B}_{p,r}$, then for all $t > 0$ there exists $\tilde{f} \in \mathcal{L}_{1,r}$ such that

$$\|\tilde{f}\|_{\mathcal{L}_{1,r}} \leq \|f\|_{\mathcal{B}_{p,r}} t^{\theta-1}, \tag{2.70}$$

and

$$\|f - \tilde{f}\| \leq \|f\|_{\mathcal{B}_{p,r}} t^{\theta}. \tag{2.71}$$

We also know that for all m, there exists h (depending on m) such that

$$\|h\|_{\mathcal{L}_1(\mathcal{D}_m)} \leq \|\tilde{f}\|_{\mathcal{L}_{1,r}} \leq \|f\|_{\mathcal{B}_{p,r}} t^{\theta-1} \tag{2.72}$$

and

$$\|\tilde{f} - h\| \leq \|\tilde{f}\|_{\mathcal{L}_{1,r}} m^{-r} \leq \|f\|_{\mathcal{B}_{p,r}} t^{\theta-1} m^{-r}, \tag{2.73}$$

so that by the triangle inequality

$$\|f - h\| \leq \|f\|_{\mathcal{B}_{p,r}} (t^{\theta} + t^{\theta-1} m^{-r}). \tag{2.74}$$

Application of Theorems 2.3 and 2.4 shows us that if we apply the OGA or RGA with the restricted dictionary and if the target function f is in $\mathcal{B}_{p,r}$ we have for any $t > 0$,

$$\|r_k\| \leq C_0 \|f\|_{\mathcal{B}_{p,r}} (t^{\theta-1} k^{-1/2} + t^{\theta} + t^{\theta-1} m^{-r}). \tag{2.75}$$

In particular, taking $t = k^{-1/2}$ and noting that $\theta = 2s$ gives

$$\|r_k\| \leq C_0 \|f\|_{\mathcal{B}_{p,r}} (k^{-s} + k^{1/2-s} m^{-r}). \tag{2.76}$$

We therefore recover the rate of Corollary 2.5 up to an additive perturbation which tends to 0 as $m \to +\infty$.

Let us close this section by making some remarks on the spaces $\mathcal{B}_{p,r}$. These spaces should be viewed as being slightly smaller than the spaces $\mathcal{B}_p$. The smaller the value of $r > 0$ the smaller the distinction between $\mathcal{B}_p$ and $\mathcal{B}_{p,r}$. Also note that the classes $\mathcal{B}_{p,r}$ depend very much on how we exhaust the dictionary $\mathcal{D}$. For example, if $\mathcal{D} = B_0 \cup B_1$ is the union of two bases B_0 and B_1, then exhausting the elements of B_0 faster than those of B_1 will result in different classes than if we exhaust those of B_1 faster than those of B_0. However, in concrete settings there is usually a natural order in which to exhaust the dictionary.

2.5 Selecting g_k

In each of the two algorithms OGA and RGA, each iteration updates the current approximation by using the function g_k which maximizes $|\langle \tilde{r}_{k-1}, g_k \rangle|$, with $\tilde{r}_{k-1} := r_{k-1}$ in the OGA and $\tilde{r}_{k-1} := f - \alpha_k f_{k-1}$ in the RGA. The choice of g_k can be modified by selecting instead a function g_k^* such that

$$|\langle \tilde{r}_{k-1}, g \rangle| \geq \gamma \max |\langle \tilde{r}_{k-1}, g \rangle| \tag{2.77}$$

for some fixed $0 < \gamma < 1$, and using g_k^* in place of g_k in defining f_k. The results we have presented thus far would remain valid with this change however the proofs would require modification, as well as the multiplicative constants in the bounds. For example in the proof of Theorem 2.3 the key inequality (2.35) would be modified and involve γ. In practice it might be easier to implement the search for a g_k^* rather than g_k. For a general treatment of these ideas see [22] and [24].

3 Application to statistical learning

3.1 Notation and definition of the estimator

We consider the classical bounded regression problem. We observe n independent realizations $(z_i) = (x_i, y_i)$, $i = 1, \cdots, n$, of an unknown distribution ρ on $Z = X \times Y$. We assume here that the output variable satisfies almost surely

$$|y| \leq B, \tag{3.1}$$

where the bound B is known to us. We denote by

$$f_\rho(x) = E(y|x), \tag{3.2}$$

the regression function which minimizes the quadratic risk

$$R(f) := E(|f(x) - y|^2), \tag{3.3}$$

over all functions f. For any f we have

$$R(f) - R(f_\rho) = \|f - f_\rho\|^2 \tag{3.4}$$

where we use the notation

$$\|u\|^2 := E(|u(x)|^2) = \|u\|^2_{L_2(\rho_X)}, \tag{3.5}$$

with ρ_X the marginal probabilty measure defined on X. We are therefore interested in constructing from our data an estimator $\hat{f}$ such that $\|\hat{f} - f_\rho\|^2$ is small. Since $\hat{f}$ depends on the realization of the training sample $\mathbf{z} := (z_i) \in Z^n$, we shall measure the estimation error by the expectation $E(\|\hat{f} - f_\rho\|^2)$ taken with respect to ρ^n.

Given our training sample $\mathbf{z}$, we define the empirical norm

$$\|f\|_n^2 := \frac{1}{n}\sum_{i=1}^{n} |f(x_i)|^2. \tag{3.6}$$

Note that $\|\cdot\|_n$ is the L_2 norm with respect to the discrete measure $\nu_{\mathbf{x}} := \sum_{i=1}^{N} \delta_{x_i}$ with δ_u the Dirac measure at u. As such the norm depends on $\mathbf{x} := (x_1, \dots, x_n)$ and not just n but we adopt the notation (3.6) to conform with other major works in learning. We view the vector $\mathbf{y} := (y_1 \dots, y_n)$ as a function y defined on the design $\mathbf{x} := (x_1, \dots, x_n)$ with $y(x_i) = y_i$. Then, for any f defined on $\mathbf{x}$,

$$\|y - f\|_n^2 := \frac{1}{n}\sum_{i=1}^{n} |y_i - f(x_i)|^2, \tag{3.7}$$

is the empirical risk for f.

In order to bound f_ρ from the given data $\mathbf{z}$ we shall use the greedy algorithms OGA and RGA described in the previous section. We choose an arbitrary value of $a \geq 1$ and then fix it. We consider a dictionary $\mathcal{D}$ and truncations of this dictionary $\mathcal{D}_1, \mathcal{D}_2, \dots$ as described in §2.4. Given our data size n, we choose

$$m := m(n) := \lfloor n^a \rfloor. \tag{3.8}$$

We will use approximation from the span of the dictionary $\mathcal{D}_m$ in our algorithm.

Our estimator is defined as follows.

(i) Given a data set $\mathbf{z}$ of size n, we apply the OGA, SPA or the RGA for the dictionary $\mathcal{D}_m$ to the function y using the empirical inner product associated to the norm $\|\cdot\|_n$. In the case of the RGA, we use the second choice (2.6) for the parameter α_k. This gives a sequence of functions $(\hat{f}_k)_{k=0}^{\infty}$ defined on $\mathbf{x}$.

(ii) We define the estimator $\hat{f} := T\hat{f}_{k^*}$, where $Tu := T_B \min\{B, |u|\}\mathrm{sgn}(u)$ is the truncation operator at level B and k^* is chosen to minimize (over all $k > 0$) the penalized empirical risk

$$\|y - T\hat{f}_k\|_n^2 + \kappa \frac{k \log n}{n}, \tag{3.9}$$

with $\kappa > 0$ a constant to be fixed later.

We make some remarks about this algorithm. First note that for $k = 0$ the penalized risk is bounded by B^2 since $\hat{f}_0 = 0$ and $|y| \leq B$. This means that we need not run the greedy algorithm for values of k larger than Bn/κ. Second, our notation $\hat{f}$ suppresses the dependence of the estimator on $\mathbf{z}$ which is again customary notation in statistics. The application of the k-th step of the greedy algorithms requires the evaluation of $O(n^a)$ inner products. In the case of the OGA we also need to compute the projection of y onto a k dimensional space. This could be done by doing Gramm-Schmidt orthogonalization. Assuming that we already had computed an orthonormal system for step $k-1$ this would require an additional evaluation of $k-1$ inner products and then a normalization step. Finally, the truncation of the dictionary $\mathcal{D}$ is not strictly needed in some more specific cases, such as neural networks (see §4).

In the following, we want to analyze the performance of our algorithm. For this analysis, we need to assume something about f_ρ. To impose conditions on f_ρ, we shall also view the elements of the dictionary normalized in the $L_2(\rho_X)$ norm $\|\cdot\|$. With this normalization, we denote by $\mathcal{L}_1$, $\mathcal{B}_p$, $\mathcal{L}_{1,r}$ and $\mathcal{B}_{p,r}$ the space of functions that have been previously introduced for a general Hilbert space $\mathcal{H}$. Here, we have $\mathcal{H} = L_2(\rho_X)$.

Finally, we denote by $\mathcal{L}_1^n$ the space of functions which admit an ℓ_1 expansion in the dictionary when the elements are normalized in the empirical norm $\|\cdot\|_n$. This space is again equipped with a norm defined as the smallest ℓ_1 norm among every admissible expansion. Similarly to $\|\cdot\|_n$ this norm depends on the realization of the design $\mathbf{x}$.

3.2 Error analysis

In this section, we establish our main result which will allow us in the next section to analyze the performance of the estimator under various smoothness assumptions on f_ρ.

Theorem 3.1 *There exists κ_0 depending only on B and a such that if $\kappa \geq \kappa_0$, then for all $k > 0$ and for all functions h in* $\mathrm{Span}(\mathcal{D}_m)$*, the estimator satisfies*

$$E(\|\hat{f} - f_\rho\|^2) \leq 8 \frac{\|h\|_{\mathcal{L}_1(\mathcal{D}_m)}^2}{k} + 2\|f_\rho - h\|^2 + C \frac{k \log n}{n}, \tag{3.10}$$

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

The proof of Theorem 3.1 relies on a few preliminary results that we collect below. The first one is a direct application of Theorem 3 from [18] or Theorem 11.4 from [13].

Lemma 3.2 *Let $\mathcal{F}$ be a class of functions which are all bounded by B. For all n and $\alpha, \beta > 0$, we have*

$$\begin{aligned} \Pr\{\exists f \in \mathcal{F} \quad &: \quad \|f - f_\rho\|^2 \geq 2(\|y - f\|_n^2 - \|y - f_\rho\|_n^2) + \alpha + \beta\} \\ &\leq \quad 14 \sup_{\mathbf{x}} \mathcal{N}\Big(\frac{\beta}{40B}, \mathcal{F}, L_1(\nu_{\mathbf{x}})\Big) \exp\Big(-\frac{\alpha n}{2568 B^4}\Big), \end{aligned} \tag{3.11}$$

where $\mathbf{x} = (x_1, \dots, x_n) \in X^n$, and $\mathcal{N}(t, \mathcal{F}, L_1(\nu_{\mathbf{x}}))$ is the covering number for the class $\mathcal{F}$ by balls of radius t in $L_1(\nu_{\mathbf{x}})$ with $\nu_{\mathbf{x}} := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ the empirical discrete measure.

Proof: This follows from Theorem 11.4 of [13] by taking $\epsilon = 1/2$ in that theorem. □

We shall apply this result in the the following setting. Given any set $\Lambda \subset \mathcal{D}$, we denote by $\mathcal{G}_\Lambda := \text{span}\{g : \; g \in \Lambda\}$ and we denote by $T\mathcal{G}_\Lambda := \{Tf : \; f \in \mathcal{G}_\Lambda\}$ the set of all truncations of the elements of $\mathcal{G}_\Lambda$ where $T = T_B$ as before. We then define

$$\mathcal{F}_k := \bigcup_{\Lambda \subset \mathcal{D}_m, \#(\Lambda) \leq k} T\mathcal{G}_\Lambda. \tag{3.12}$$

The following result gives an upper bound for the entropy numbers $\mathcal{N}(t, \mathcal{F}_k, L_1(\nu_{\mathbf{x}}))$.

Lemma 3.3 *For any probability measure ν, for any $t > 0$, and for any Λ with cardinality k we have the bound*

$$\mathcal{N}(t, T\mathcal{G}_\Lambda, L_1(\nu)) \leq 3\Big(\frac{2eB}{t} \log \frac{3eB}{t}\Big)^{k+1}. \tag{3.13}$$

Additionally,

$$\mathcal{N}(t, \mathcal{F}_k, L_1(\nu)) \leq 3n^{ak}\Big(\frac{2eB}{t} \log \frac{3eB}{t}\Big)^{k+1}. \tag{3.14}$$

Proof: For each Λ with cardinality k, Theorem 9.4 in [13] gives

$$\mathcal{N}(t, T\mathcal{G}_\Lambda, L_1(\nu)) \leq 3\Big(\frac{2eB}{t} \log \frac{3eB}{t}\Big)^{V_\Lambda} \tag{3.15}$$

with V_Λ the VC dimension of the set of all subgraphs of $T\mathcal{G}_\Lambda$. It is easily seen that V_Λ is smaller than the VC dimension of the set of all subgraphs of $\mathcal{G}_\Lambda$, which by Theorem 9.5 in [13] is less than $k+1$. This establishes (3.13). Since there are less than n^{ak} possible sets Λ, the result (3.14) follows by taking the union of the coverings for all $T\mathcal{G}_\Lambda$ as a covering for $\mathcal{F}_k$. □

Finally, we shall need a result that relates the $\mathcal{L}_1$ and $\mathcal{L}_1^n$ norms.

Lemma 3.4 *Given any dictionary $\mathcal{D}$, for all functions h in* $\text{Span}(\mathcal{D})$, *we have*

$$E(\|h\|_{\mathcal{L}_1^n}^2) \leq \|h\|_{\mathcal{L}_1}^2. \tag{3.16}$$

Proof: We normalize the elements of the dictionary in $\|\cdot\| = \|\cdot\|_{\mathcal{H}}$. Given any $h = \sum_{g\in\mathcal{D}} c_g g$ and any $\mathbf{z}$ of length n, we write

$$h = \sum_{g\in\mathcal{D}} c_g g = \sum_{g\in\mathcal{D}} c_g^n \frac{g}{\|g\|_n}, \tag{3.17}$$

with $c_g^n := c_g \|g\|_n$. We then observe that

$$\begin{aligned} E((\textstyle\sum_{g\in\mathcal{D}} |c_g^n|)^2) &= \textstyle\sum_{(g,g')\in\mathcal{D}\times\mathcal{D}} |c_g c_{g'}| E(\|g\|_n \|g'\|_n) \\ &\leq \textstyle\sum_{(g,g')\in\mathcal{D}\times\mathcal{D}} |c_g c_{g'}| \Big(E(\|g\|_n^2) E(\|g'\|_n^2)\Big)^{1/2} \\ &= \textstyle\sum_{(g,g')\in\mathcal{D}\times\mathcal{D}} |c_g c_{g'}| \Big(\|g\|^2 \|g'\|^2\Big)^{1/2} \\ &= \textstyle\sum_{(g,g')\in\mathcal{D}\times\mathcal{D}} |c_g c_{g'}| \\ &= (\textstyle\sum_{g\in\mathcal{D}} |c_g|)^2. \end{aligned}$$

The result follows by taking the infimum over all possible admissible (c_g) and using the fact that

$$E(\inf[\sum_{g\in\mathcal{D}} |c_g^n|]^2) \leq \inf E([\sum_{g\in\mathcal{D}} |c_g^n|]^2). \tag{3.18}$$

□

Proof of Theorem 3.1: We write

$$\|\hat{f} - f_\rho\|^2 = T_1 + T_2, \tag{3.19}$$

where

$$T_1 := \|\hat{f} - f_\rho\|^2 - 2(\|y - \hat{f}\|_n^2 - \|y - f_\rho\|_n^2 + \kappa \frac{k^* \log n}{n}), \tag{3.20}$$

and

$$T_2 := 2(\|y - \hat{f}\|_n^2 - \|y - f_\rho\|_n^2 + \kappa \frac{k^* \log n}{n}). \tag{3.21}$$

From the definition of our estimator, we know that for all $k > 0$, we have

$$T_2 \leq 2(\|y - \hat{f}_k\|_n^2 - \|y - f_\rho\|_n^2 + \kappa \frac{k \log n}{n}). \tag{3.22}$$

Therefore, for all $k > 0$ and $h \in L_2(\rho_X)$, we have $T_2 \leq T_3 + T_4$ with

$$T_3 := 2(\|y - \hat{f}_k\|_n^2 - \|y - h\|_n^2), \tag{3.23}$$

and

$$T_4 := 2(\|y - h\|_n^2 - \|y - f_\rho\|_n^2) + 2\kappa \frac{k \log n}{n}. \tag{3.24}$$

We now bound the expectations of T_1, T_3 and T_4. For the last term, we have

$$E(T_4) = 2E(|y - h(x)|^2 - |y - f_\rho(x)|^2) + 2\kappa \frac{k \log n}{n} = 2\|f_\rho - h\|^2 + 2\kappa \frac{k \log n}{n}. \tag{3.25}$$

For T_3, we know from Theorems 2.3 and 2.4, that we have

$$\|y - \hat{f}_k\|_n^2 - \|y - h\|_n^2 \leq 4\frac{\|h\|_{\mathcal{L}_1^n}^2}{k}. \tag{3.26}$$

Using in addition Lemma 3.4, we thus obtain

$$E(T_3) \leq 8\frac{\|h\|_{\mathcal{L}_1}^2}{k}. \tag{3.27}$$

For T_1, we introduce Ω, the set of $\mathbf{z} \in Z^n$ for which

$$\|\hat{f} - f_\rho\|^2 \geq 2(\|y - \hat{f}\|_n^2 - \|y - f_\rho\|_n^2 + \kappa\frac{k^* \log n}{n}). \tag{3.28}$$

Since $T_1 \leq \|\hat{f} - f_\rho\|^2 + 2\|y - f_\rho\|_n^2 \leq 6B^2$, we have

$$E(T_1) \leq 6B^2 \Pr(\Omega). \tag{3.29}$$

We thus obtain that for all $k > 0$ and for all $h \in L_2(\rho_X)$, we have

$$E(\|\hat{f} - f_\rho\|^2) \leq 8\frac{\|h\|_{\mathcal{L}_1}^2}{k} + 2\|f_\rho - h\|^2 + 2\kappa\frac{k \log n}{n} + 6B^2 \Pr(\Omega). \tag{3.30}$$

It remains to bound $\Pr(\Omega)$. Since k^* can take an arbitrary value depending on the sample realization, we simply control this quantity by the union bound

$$\sum_{1 \leq k \leq Bn/\kappa} \Pr\{\exists f \in \mathcal{F}_k \ : \|f - f_\rho\|^2 \geq 2(\|y - f\|_n^2 - \|y - f_\rho\|_n^2) + 2\kappa\frac{k \log n}{n}\}. \tag{3.31}$$

Denoting by p_k each term of this sum, we obtain by Lemma 3.2

$$p_k \leq 14 \sup_{\mathbf{x}} \mathcal{N}\Big(\frac{\beta}{40B}, \mathcal{F}_k, L_1(\nu_{\mathbf{x}})\Big) \exp\Big(-\frac{\alpha n}{2568B^4}\Big) \tag{3.32}$$

provided $\alpha + \beta \leq 2\kappa\frac{k \log n}{n}$. Assuming without loss of generality that $\kappa > 1$, we can take $\alpha := \kappa\frac{k \log n}{n}$ and $\beta = 1/n$, from which it follows that

$$p_k \leq 14 \sup_{\mathbf{x}} \mathcal{N}\Big(\frac{1}{40Bn}, \mathcal{F}_k, L_1(\nu_{\mathbf{x}})\Big) n^{-\frac{\kappa k}{2568B^4}}. \tag{3.33}$$

Using Lemma 3.3, we finally obtain

$$p_k \leq Cn^{ak} n^{2(k+1)} n^{-\frac{\kappa}{2568B^4}}, \tag{3.34}$$

so that by choosing $\kappa \geq \kappa_0$ large enough, we always have $p_k \leq Cn^{-2}$. It follows that

$$\Pr(\Omega) \leq \sum_{k \leq \frac{Bn}{\kappa}} p_k \leq \frac{C}{n}. \tag{3.35}$$

This contribution is therefore absorbed in the term $2\kappa\frac{k \log n}{n}$ in the main bound and this concludes our proof. $\square$

3.3 Rates of convergence and universal consistency

In this section, we apply Theorem 3.1 in several situations which correspond to different prior assumptions on f_ρ. We first consider the case where $f_\rho \in \mathcal{L}_{1,r}$. In that case, we know that for all m there exists $h \in \mathrm{Span}(\mathcal{D}_m)$ such that $\|h\|_{\mathcal{L}_1(\mathcal{D}_m)} \le M$ and $\|f_\rho - h\| \le M m^{-r}$ with $M := \|f_\rho\|_{\mathcal{L}_{1,r}}$. Therefore Theorem 3.1 yields

$$E(\|\hat{f} - f_\rho\|^2) \le C \min_{k>0}\Big(\frac{M^2}{k} + M^2 n^{-2ar} + \frac{k\log n}{n}\Big). \tag{3.36}$$

In order that the effect of truncating the dictionary does not affect the estimation bound, we make the assumption that $2ar \ge 1$. This allows us to delete the middle term in (3.36). Note that this is not a strong additional restriction over $f_\rho \in \mathcal{L}_1$ since a can be fixed arbitrarily large.

Corollary 3.5 *If $f_\rho \in \mathcal{L}_{1,r}$ with $r > 1/2a$, then*

$$E(\|\hat{f} - f_\rho\|^2) \le C(1 + \|f\|_{\mathcal{L}_{1,r}})\Big(\frac{n}{\log n}\Big)^{-1/2}. \tag{3.37}$$

Proof: We take $k := \lceil (M+1)^2 \frac{n}{\log n}\rceil^{1/2}$ in (3.36) and obtain the desired result. □

We next consider the case where $f_\rho \in \mathcal{B}_{p,r}$. In that case, we know that for all m and for all $t > 0$, there exists $h \in \mathrm{Span}(\mathcal{D}_m)$ such that $\|h\|_{\mathcal{L}_1} \le M t^{\theta-1}$ (see (2.72)) and $\|f_\rho - h\| \le M(t^\theta + t^{\theta-1}m^{-r})$ (see (2.74)), with $1/p = (1+\theta)/2$ and with $M = \|f\|_{B_{p,r}}$. Taking $t = k^{-1/2}$ and applying Theorem 3.1, we obtain

$$E(\|\hat{f} - f_\rho\|^2) \le C \min_{k>0}\Big(M^2 k^{-2s} + M^2(k^{-s} + k^{-s+1/2}n^{-ar})^2 + \frac{k\log n}{n}\Big), \tag{3.38}$$

with $s = 1/p - 1/2$. We now impose that $ar \ge 1/2$ which allows us to delete the term involving n^{-ar}.

Corollary 3.6 *If $f_\rho \in \mathcal{B}_{p,r}$ with $r \ge 1/(2a)$,*

$$E(\|\hat{f} - f_\rho\|^2) \le C(1 + \|f_\rho\|_{\mathcal{B}_{p,r}})^{\frac{2}{2s+1}}\Big(\frac{n}{\log n}\Big)^{-\frac{2s}{1+2s}}. \tag{3.39}$$

with C a constant depending only on κ, B and a,

Proof: We take $k := \lceil (M+1)^2 \frac{n}{\log n}\rceil^{\frac{1}{1+2s}}$ in (3.38) and obtain the desired result. □

Let us finally prove that the estimator is universally consistent. For this we need to assume that the dictionary $\mathcal{D}$ is complete in $L_2(\rho_X)$.

Theorem 3.7 *For an arbitrary regression function, we have*

$$\lim_{n\to+\infty} E(\|\hat{f} - f_\rho\|^2) = 0. \tag{3.40}$$

Proof For any $\epsilon > 0$ and n sufficiently large there exists $h \in \text{Span}(\mathcal{D}_m)$ which satisfies $\|f - h\| \le \epsilon$. According to Theorem 3.1, we thus have

$$E(\|\hat{f} - f_\rho\|^2) \le C \min_{k>0} \Big(\frac{\|h\|_{\mathcal{L}_1}^2}{k} + \epsilon^2 + \frac{k \log n}{n} \Big). \tag{3.41}$$

Taking $k = n^{1/2}$, this gives

$$E(\|\hat{f} - f_\rho\|^2) \le C(\epsilon^2 + n^{-1/2} \log n \Big), \tag{3.42}$$

which is smaller than $2C\epsilon^2$ for n large enough. □

Remark 3.8 *We have seen in §2.4 that the OGA can be monitored by a thresholding procedure, similar to an orthonormal basis. It is therefore tempting to adopt another approach for selecting the proper value* k^* *in the case where we are using the OGA. take for* k^* *the largest value such that the coefficient* $|\langle \hat{r}_{k-1}, g_k \rangle_n|$ *is larger than a threshold* t_n *of the order* $(\frac{\log n}{n})^{1/2}$*, which is the universal threshold proposed e.g. in wavelet shrinkage methods [11, 12]. Although we were not able to prove it so far, we conjecture that this approach yields the same convergence rates that we have proved for our estimator.*

Finally, we remark that although our results for the learning problem are stated and proved for the OGA and RGA, they hold equally well when the SPA is employed.

4 Neural networks

Neural networks have been one of the main motivations for the use of greedy algorithms in statistics [2, 4, 14, 18]. They correspond to a particular type of dictionary. One begins with a univariate positive and increasing function σ that satisfies $\sigma(-\infty) = 0$ and $\sigma(+\infty) = 1$ and defines the dictionary consisting of all functions

$$x \mapsto \sigma(\langle v, x \rangle + w) \tag{4.1}$$

for all vectors $v \in I\!R^D$ and scalar $w \in I\!R$ where D is the dimension of the feature variable x. Typical choices for σ are the Heaviside function $h = \chi_{x>0}$ or more general sigmoidal functions which are regularized version of h.

In [18], the authors consider neural networks where σ is the Heaviside function and the vectors v are restricted to have at most d non-zero coordinates (d-bounded fan-in) for some fixed $d \le D$. We denote this dictionary by $\tilde{\mathcal{D}}$. With this choice of dictionary and using the standard relaxed greedy algorithm, they establish the convergence rate

$$E(\|\hat{f}_k - f_\rho\|^2 - \|f_\rho - f_a\|^2) \lesssim \frac{1}{k} + kd \frac{\log(Dn)}{n}, \tag{4.2}$$

where f_a is the projection of f_ρ onto the convex hull of the dictionary $\tilde{\mathcal{D}}$. This can also be expressed by

$$E(\|\hat{f}_k - f_a\|^2) \lesssim \frac{1}{k} + kd \frac{\log(Dn)}{n}, \tag{4.3}$$

which shows that with the choice $k = n^{1/2}$, the estimator converges to f_a with rate $n^{-1/2}$ up to a logarithmic factor. In particular, the algorithm is not universally consistent since it only converges to the regression function when it belongs to this convex hull.

4.1 Convergence results

Let us apply our results to this particular setting. We want to first note that in this case it is not necessary to truncate the dictionary $\tilde{\mathcal{D}}$ into a finite dictionary in order to achieve our theoretical results. The truncation of dictionaries was used in the proof of Theorem 3.1 to bound the covering numbers of the sets $\mathcal{F}_k$ through the bound established in Lemma 3.3. In the specific case of $\tilde{\mathcal{D}}$, one can bound these covering numbers without truncation. Let us note that in this case

$$\mathcal{F}_k := \bigcup_{\Lambda \subset \tilde{\mathcal{D}}, \#(\Lambda) \leq k} T\mathcal{G}_\Lambda \tag{4.4}$$

where we no longer have the restriction that Λ is in $\tilde{\mathcal{D}}_m$.

Lemma 4.1 *For the dictionary $\tilde{\mathcal{D}}$ and for any probability measure ν of the type $\nu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, and for any $k > 0$ and $t > 0$ we have the bound*

$$\mathcal{N}(t, \mathcal{F}_k, L_1(\nu)) \leq 3(n+1)^{k(D+1)} \Big(\frac{2eB}{t} \log \frac{3eB}{t}\Big)^{k+1} \tag{4.5}$$

where the sets $\mathcal{F}_k$ are defined as in (4.4)

Proof: As in the proof of Lemma 3.3, we first remark that

$$\mathcal{N}(t, T\mathcal{G}_\Lambda, L_1(\nu)) \leq 3\Big(\frac{2eB}{t} \log \frac{3eB}{t}\Big)^{k+1}. \tag{4.6}$$

We next use two facts from Vapnik-Chervonenkis theory (see for example Chapter 9 in [13]) : (i) if $\mathcal{A}$ is a collection of sets with VC dimension λ, there are at most $(n+1)^\lambda$ sets of $\mathcal{A}$ separating the points $(x_1, \cdots, x_n)$ in different ways, and (ii) the VC dimension of the collection of half-hyperplanes in $I\!R^D$ has VC dimension $D+1$. It follows that there are at most $(n+1)^{D+1}$ hyperplanes separating the points $(x_1, \cdots, x_n)$ in different ways, and therefore there are at most $(n+1)^{k(D+1)}$ ways of picking $(g_1, \cdots, g_k)$ in $\mathcal{D}$ which will give different functions on the sample $(x_1, \cdots, x_n)$. The result follows by taking the union of the coverings on all possible k-dimensional subspaces. $\Box$.

Remark 4.2 *Under the d-bounded fan-in assumption, the factor $(n+1)^{k(D+1)}$ can be reduced to $D^{kd}\Big(\frac{en}{d+1}\Big)^{k(d+1)}$, see the proof of Lemma 3 in [18].*

Based on this bound, a brief inspection of the proof of Theorem 3.1 show that its conclusion still holds, now with κ_0 depending on B and D. It follows that the rates of convergence in Corollaries 3.5 and 3.6 now hold under the sole assumptions that $f \in \mathcal{L}_1$ and $f \in \mathcal{B}_p$ respectively. On the other hand, the universal consistency result in Theorem 3.7 requires that the dictionary is complete in $L_2(\rho_X)$, which only holds when $d = D$, i.e. all direction vectors are considered.

Theorem 3.1 improves in two ways the bound (4.2) of [18]: on the one hand f_a is replaced by an arbitrary function h which can be optimized, and on the other hand the value of k can also be optimized.

4.2 Smoothness conditions

Let us finally briefly discuss the meaning of the conditions $f \in \mathcal{L}_1$ and $f \in \mathcal{B}_p$ in the case of a dictionary $\mathcal{D}$ consisting of the functions (4.1) for a fixed σ and for all $v \in I\!R^D$ and $w \in I\!R$. The following can be deduced from a classical result obtained in [4] : assuming that the marginal distribution ρ_X is supported in a ball $B_r := \{|x| \le r\}$, for any function f having a converging Fourier representation

$$f(x) = \int \mathcal{F}f(\omega) e^{i\langle \omega, x\rangle} d\omega, \tag{4.7}$$

the smoothness condition

$$\int |\omega||\mathcal{F}f(\omega)| d\omega < +\infty, \tag{4.8}$$

ensures that

$$\|f\|_{\mathcal{L}_1} \le (2rC_f + |f(0)|) \le 2rC_f + \|f\|_{L_\infty}, \tag{4.9}$$

with $C_f := \int |\omega||\mathcal{F}f(\omega)|d\omega$. Barron actually proves that condition (4.8) ensures that $f(x) - f(0)$ lies in the closure of the convex hull of $\mathcal{D}$ multiplied by $2rC_f$, the closure being taken in $L_2(\rho_X)$ and the elements of the dictionary being normalized in L_∞. The bound (4.9) follows by remarking that the L_∞ norm controls the $L_2(\rho_X)$ norm.

We can therefore obtain smoothness conditions which ensure that $f \in \mathcal{B}_p$ by interpolating between the condition $\omega\mathcal{F}f(\omega) \in L_1$ and $f \in L_2(\rho_X)$.

In the particular case where ρ_X is continuous with respect to the Lebesgue measure, i.e. $\rho_X(A) \le c|A|$, we know that a sufficient condition to have $f \in L_2(\rho_X)$ is given by $\mathcal{F}f \in L_2$.

We can then rewrite the two conditions that we want to interpolate as $|\omega|^{-1}|\mathcal{F}f(\omega)| \in L_1(|\omega|^2 d\omega)$ and $|\omega|^{-1}|\mathcal{F}f(\omega)| \in L_2(|\omega|^2 d\omega)$. Therefore by standard interpolation arguments, we obtain that a sufficient condition for a bounded function f to be in $\mathcal{B}_p$ is given by

$$|\omega|^{-1}|\mathcal{F}f(\omega)| \in wL_p(|\omega|^2 d\omega), \tag{4.10}$$

or in other words

$$\sup_{\eta>0} \eta^p \int \chi_{\{|\mathcal{F}f(\omega)| \ge \eta|\omega|\}} |\omega|^2 d\omega < +\infty. \tag{4.11}$$

A slightly stronger, but simpler condition is

$$|\omega|^{-1}|\mathcal{F}f(\omega)| \in L_p(|\omega|^2 d\omega), \tag{4.12}$$

which reads

$$\int |\omega|^{2-p}|\mathcal{F}f(\omega)|^p d\omega < +\infty. \tag{4.13}$$

When ρ_X is arbitrary, a sufficient condition for $f \in L_2(\rho_X)$ is $\mathcal{F}f \in L_1$, which actually also ensures that $f \in L_\infty$. In that case, we can again apply standard interpolation arguments and obtain that a sufficient condition for a bounded function f to be in $\mathcal{B}_p$ is given by

$$\sup_{A>0} A^{1-2/p} \int_{|\omega|\ge A} |\mathcal{F}f(\omega)| d\omega < +\infty. \tag{4.14}$$

A slightly stronger, but simpler condition is

$$\int |\omega|^{2/p-1} |\mathcal{F}f(\omega)| d\omega < +\infty. \tag{4.15}$$

Remark 4.3 *Note that truncating the dictionary might still be of practical importance in order to limit the computational complexity of the algorithm. Such a truncation can be done by restricting to a finite number m of direction vectors v, which typically grows together with sample size n. In this case, we need to consider the spaces $\mathcal{L}_{1,r}$ and $\mathcal{B}_{p,r}$ which contain an additional smoothness assumption compared to $\mathcal{L}_1$ and $\mathcal{B}_p$. This additional amount of smoothness is meant to control the error resulting from the discretization of the direction vectors. We refer to [20] for general results on this problem.*

4.3 Thanks:

The authors wish to thank Professor Rich Baraniuk and Rice University who hosted two of us (A.C. and R.D) during the Fall of 2005 when much of this research was completed.

References

[1] Avellaneda, M., G. Davis and S. Mallat (1997) *Adaptive greedy approximations*, Constructive Approximation **13**, 57–98.

[2] Barron, A. (1990) *Complexity regularization with application to artificial neural network*, in *Nonparametric functional estimation and related topics*, G. Roussas (ed.), 561–576, Kluwer Academic Publishers.

[3] Barron, A. (1992) *Neural net approximation*, Proc 7th Yale Workshop on Adaptive and Learning Systems, K.S. Narendra Ed, New Haven, CT, 1 69–72.

[4] Barron, A. (1993) *Universal approximation bounds for superposition of n sigmoidal functions*, IEEE Trans. Inf. Theory **39**, 930–945

[5] Barron, A. and G.H.L. Cheang (2005) *Risk bounds and greedy computations for penalized least squares, model selection, and neural networks* , Preprint, Department of Statistics, Yale University.

[6] Barron, A. and G.H.L. Cheang (2001) *Penalized least squares, model selection, convex hull classes, and neural nets*, in Verleysen, M. (editor). Proceedings of the 9th ESANN, Brugge, Belgium, De-Facto press, 371-376.

[7] Bennett, C. and R. Sharpley (1988) *Interpolation of Operators*, in Pure and Applied Mathematics, Academic Press, N.Y.

[8] Bergh J. and J. Löfström (1976) *Interpolation spaces*, Springer Verlag, Berlin.

[9] DeVore, R. (1998) *Nonlinear approximation*, Acta Numerica **7**, 51–150.

[10] DeVore, R. and V. Temlyakov (1996) *Some remarks on greedy algorithms*, Advances in Computational Mathematics **5**, 173-187.

[11] Donoho, D.L. and I. M. Johnstone (1998) *Minimax Estimation via Wavelet shrinkage*, Annals of Statistics **26**, no. 3, 879–921.

[12] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1996) *Wavelet Shrinkage: Asymptopia?* , J. Royal Statistical Soc. **57**, 301–369.

[13] Györfy, L., M. Kohler, A. Krzyzak, A. and H. Walk (2002) *A distribution-free theory of nonparametric regression*, Springer Verlag, Berlin.

[14] Jones, L.K. (1992) *A simple lemma on greedy approximation in Hilbert spaces and convergence rates for projection pursuit regression and neural network training*, Ann. Stat. **20**, 608–613.

[15] Konyagin, S.V. and V.N. Temlyakov (1999) *Rate of convergence of Pure greedy Algorithm*, East J. Approx. **5**, 493–499.

[16] Kurkova, V. and M. Sanguineti (2001) *Bounds on Rates of Variable-Basis and Neural-Network Approximation*, IEEE Transactions on Information Theory 47-6, 2659-2665.

[17] Kurkova, V. and M. Sanguineti (2002) *Comparison of Worst Case Errors in Linear and Neural Network Approximation*, IEEE Transactions on Information Theory 48-1, 264-275.

[18] Lee, W.S., P. Bartlett and R.C. Williamson (1996) *Efficient agnostic learning of neural networks with bounded fan-in*, IEEE Trans. Inf. Theory **42** 2118–2132.

[19] Livshitz, E.D. and V.N. Temlyakov (2003) *Two lower estimates in greedy approximation*, Constr. Approx. **19**, 509–524.

[20] Petrushev (1998) Approximation by ridge functions and neural networks, SIAM J. Math. Anal., **30**, 155–189.

[21] Temlyakov, V., *Nonlinear methods of approximation* (2003) Journal of FOCM, **3**, 33–107.

[22] Temlyakov, V., *Greedy algorithms in Banach spaces* (2001) Advances in Comp. Math. **14**, 277–292.

[23] Temlyakov, V., *Greedy algorithms with restricted depth search* (2005) Proc. of the Steklov Inst. Math., **248**, 255–267.

[24] Temlyakov, V., *Weak greedy algorithms* (2000) Advances in Comp. Math., **12**, 213–227.

Andrew Barron, Department of Statistics, Yale University, P.O. Box 208290, New Haven CT 06520-8290, andrew.barron@yale.edu

Albert Cohen, Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie 175, rue du Chevaleret, 75013 Paris, France, cohen@ann.jussieu.fr

Wolfgang Dahmen, Institut für Geometrie und Praktische Mathematik, RWTH Aachen, Templergraben 55, D-52056 Aachen Germany, dahmen@igpm.rwth-aachen,de

Ronald DeVore, Industrial Mathematics Institute, University of South Carolina, Columbia, SC 29208, devore@math.sc.edu