Function Approximation in Hilbert Spaces: A General Sequential Method and a Particular Implementation with Neural Networks

Enrique Romero *
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
eromero@lsi.upc.es

February 14, 2000

Abstract

A sequential method for approximating vectors in Hilbert spaces, called *Sequential Approximation with Optimal Coefficients (SAOC)*, is presented. Most of the existing sequential methods choose the new term so that it matches the previous residue as best as possible. Although this strategy leads to approximations convergent towards the target function, it may be far from being the best strategy with regard to the number of terms of the approximation. *SAOC* combines two key ideas. The first is the optimization of the coefficients (the linear part of the approximation). The second is the flexibility to choose the frequencies (the nonlinear part). The only relation with the residue has to do with its approximation capability of the target vector f. *SAOC* maintains orthogonal-like properties. The theoretical results obtained proof that, under reasonable conditions, the construction of the approximation is always possible and, in the limit, the residue of the approximation obtained with *SAOC* is the best one that can be obtained with any subset of the given set of vectors. In addition, it seems that it should achieve the same accuracy that other existent sequential methods with fewer terms. In the particular case of L^2, it can be applied to polynomials, Fourier series, wavelets and neural networks, among others. Also, a particular implementation using neural networks is presented. In fact, the profit is reciprocal, because *SAOC* can be used as an inspiration to construct and train a neural network.

*This work was supported by Consejo Interministerial de Ciencia y Tecnología (CICYT), under project TAP1999-0747

1 Introduction

The main problem in approximation theory can be stated as follows ([Achieser 1956], [Lorentz 1966]): "Let f be an element of a space M. Given a space of parameters Φ and a function $F : 2^{\Phi} \rightarrow M$, determine the subset of parameters $\phi \subseteq \Phi$ such that the deviation between f and $F(\phi)$ is minimum". This statement leads, in a natural way, to define the concept of distance. The spaces where a distance can be defined are called metric spaces. Usually, the approximation is linear with regard to a subset of parameters called *coefficients*. In this case, it makes sense working in vector spaces and using the concepts of vector norm and normed space. Hilbert spaces are a particular case of normed spaces in which an inner product can be defined. Hence, in Hilbert spaces we have the concepts of projection and orthogonality between vectors.

Vector approximation in Hilbert spaces is present in different areas, such as polynomial approximation [Weierstrass 1885], Fourier series [Young 1980], statistics [Huber 1985], signal processing [Mallat 1998] or neural networks [Bishop 1995]. In most cases, the Hilbert space of interest is L^2, where the vector f is a square integrable function defined on a subset of $\mathbb{R}^I$, that we want to approximate by linear combinations of simpler functions. These functions depend on a finite number of parameters, that we will call *frequencies*. The approximations are usually non-linear with regard to the frequencies. Due to this fact the problem of finding the best approximation with a finite number of terms is extremely complex [Horst & Tuy 1993].

The techniques developed up to date are strongly dependent on whether an analytical expression is available or not. In the former case, we probably can compute exactly the inner products or at least evaluate the function on any desired set of points. Without an analytical expression, all we usually have is the function value on a finite set of points (dataset), and perhaps some kind of information about the function behaviour in some regions of the space (the function to be approximated may be, for example, a probability function or a finite signal). In this latter case, the concept of interpolation or generalization is specially important, since the basic aim is to predict the function behaviour on points that do not belong to the dataset. Because of the fact that the dataset could be approximated, in principle, by the vectors associated with many different sets of parameters, finding a method that picks the best one becomes a fundamental problem. It is very easy to verify that, if we have a finite dataset, the problem of approximating a function in L^2 is equivalent to the problem of approximating (by Least Squares) a vector

in $\mathbb{C}^T$, where T is the number of elements in the dataset.

In theory, the approximation in L^2 may consist of an infinite number of terms. In practical applications, however, this is not possible. Suppose we consider that an approximation is valid if its deviation is less than a fixed $\varepsilon > 0$. For every $\varepsilon > 0$ and every basis in L^2 it is possible to find a function f so that we need a very large number of basis terms to approximate f with deviation less than ε. Linear expansions in a single basis are not flexible enough. The information can be diluted across the whole basis [Mallat & Zhang 1993]. This happens even with an orthogonal basis.

An attractive way to construct an approximation is, starting from scratch, adding terms one at a time to the partial approximations, until the desired approximation accuracy is achieved. This is the aim of sequential methods. The goodness of the added terms is essential to yield the desired accuracy. Most of the existing methods choose the new term so that it matches the previous residue as best as possible (see Section 6). Although this strategy leads to approximations convergent towards the target function, it may be far from being the best strategy, as can be observed in the example in Figure 1. When approximating the vector f with v_1 and v_2 we obtain X_2. Clearly, this is not the best possible approximation, since v_1 and v_2 form a basis of $\mathbb{R}^2$. In this case, recalculating the coefficients of the previous added terms would lead to a much better approximation (exact, in fact) of the target function. This recalculation would work as follows. Once we have selected the vector that matches the residue as best as possible (v_2), find the best approximation of f with the whole set of vectors selected up to the moment (v_1 and v_2). But recalculating the coefficients is not enough, as illustrated in the example in Figure 2. The vector s (not lying on the plane that contains X_1 and f), which best matches the residue r_1, is not necessarily such that, together with the previous terms, best approximates the target vector $f \in \mathbb{R}^3$. In this case any vector lying on the plane that contains X_1 and f (g, for example) allows an exact approximation of the target vector. Trying to approximate the residue does not take into account the interactions with the previous selected terms. The vector that best matches the residue has nothing to do, in principle, with all the planes that contain f and the vectors of the subspace spanned by the previuos terms. Any vector lying on the plane that contains f and a vector of the subspace spanned by the previuos terms allows an exact approximation of the target vector. In infinite dimensional spaces, this problem can lead to approximations with a very large number of terms.

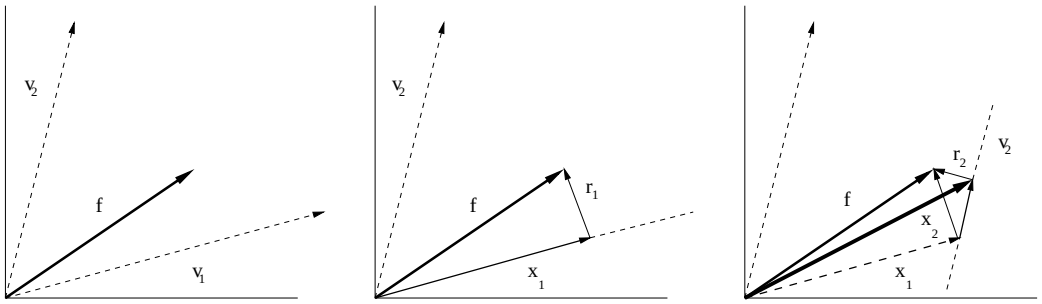

Figure 1: Approximation of a vector f in $\mathbb{R}^2$ maximizing the approximation to the previous residue. The best approximation is orthogonal to the residue. The resulting vector (X_2) is not the best approximation that can be achieved with v_1 and v_2.

In this paper we present a general sequential method for function approximation, named *SAOC*, that takes into account these problems. On the one hand, the vectors can be selected at every step in a flexible manner. On the other, it optimizes the coefficients, so that we always achieve the best approximation with the selected vectors. The only relation with the residue has to do with its approximation capability of the target vector f. The method is general in the sense that it can be constructed independently of the concrete Hilbert space. Under very mild conditions, it is possible to guarantee that the approximation given by *SAOC* can always be constructed (that is, there always exists a vector satisfying the conditions in the definition that can be chosen in the next step). In the limit, the residue of the approximation obtained with *SAOC* is the best one that can be obtained with any subset of the given set of vectors. In the particular case of L^2, *SAOC* can be applied to polynomials, Fourier series, wavelets and neural networks, among others. In fact, the universal approximation capability of a family of functions is enough to apply the *SAOC* method with guarantee of convergence (whenever the *SAOC* construction is feasible).

A particular implementation with neural networks is also presented. Neural networks are a suitable approach to deal with function approximation problems when we only have a dataset. A feedforward neural network architecture with a non-linear hidden layer and a linear output layer leads to very similar approximations to those provided by *SAOC*. Therefore, there are reasons to think that the method can be implemented in a neural network. In fact, the profit is reciprocal, because *SAOC* can be used as a guide to construct and

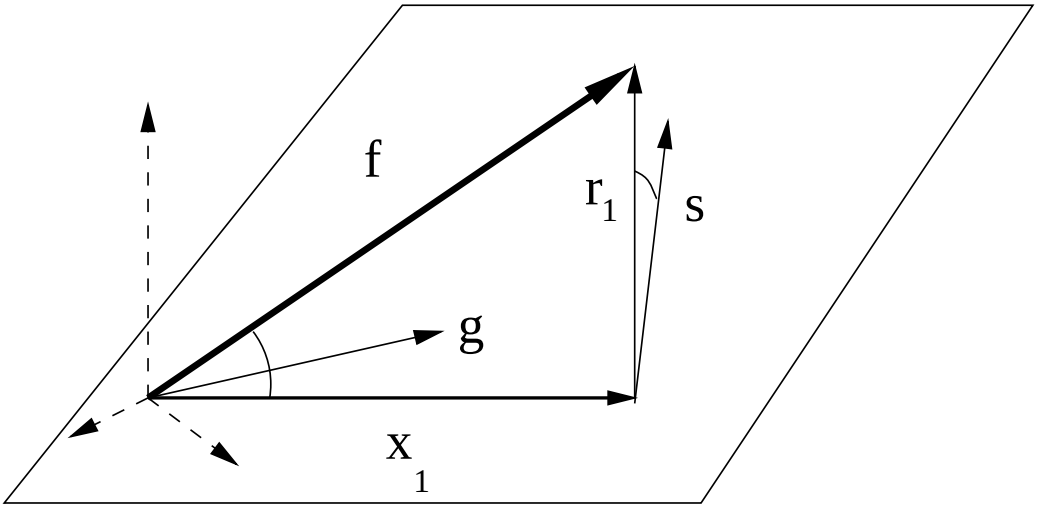

Figure 2: Approximation of a vector f in $\mathbb{R}^3$ maximizing the approximation to the previous residue and recalculating the coefficients. The vector g, which is lying on the plane that contains f and X_1 allows a better approximation to f that the vector s (not on this plane), which is the vector that best matches the residue r_1.

train a neural network.

Some preliminaries can be found in Section 2. The general definition and some basic properties of $SAOC$ are explained in Section 3. The main results of existence and convergence are proved is Section 4. The application to specific vectors in L^2 is described in Section 5. An overview of the related work and a comparison of the previous methods against $SAOC$ in terms of basic features is presented in Section 6. Some practical properties of $SAOC$ are discussed in Section 7. Finally, a particular implementation using neural networks is presented in Section 8.

2 Preliminaries

First, we refresh the concepts of metric space, vector space, normed space and Hilbert space. More detailed explanations can be found in [Achieser 1956], [Berberian 1961], [Yosida 1965], [Kolmogorov & Fomin 1975] or [Reddy 1998].

A *metric space* M is a set with a distance $D : M \times M \to \mathbb{R}^+$ such that $\forall x, y, z \in M$

(a) $D(x,x) = 0$.

(b) $D(x,y) = D(y,x)$.

(c) $D(x,z) \leqslant D(x,y) + D(y,z)$ (triangular inequality).

In metric spaces it is possible to define the concepts of continuity and sequence convergence. A function $f : M_1 \to M_2$ between two metric spaces (M_1, D_1) and (M_2, D_2) is said to be *continuous* at $x_0 \in M_1$ if

$$\forall \varepsilon > 0 \quad \exists \delta(\varepsilon, x_0) > 0 \quad \forall x \in M_1 \qquad D_1(x_0, x) < \delta \Longrightarrow D_2(f(x_0), f(x)) < \varepsilon.$$

If the function is continuous at every point, we say that it is continuous on M_1. A sequence $\{x_k\}_{k \geqslant 1}$ of points in a metric space (M, D) is said to be *convergent* towards a point x_0 if

$$\forall \varepsilon > 0 \quad \exists N(\varepsilon) > 0 \quad \forall n \geqslant N \qquad D(x_0, x_n) < \varepsilon.$$

The point x_0 is called the *sequence limit*. As an equivalent definition, x_0 is the limit of the sequence $\{x_k\}_{k \geqslant 1}$ if

$$\lim_{n \to \infty} D(x_0, x_n) = 0.$$

In fact, the definition of continuous function and convergence can be made in topological spaces. Metric spaces are a particular case of topological spaces, where the topology is defined in terms of the distance.

A *vector space* V is a set satisfying

(a) $(V, +)$ is an abelian group with an addition $+ : V \times V \to V$.

(b) A scalar multiplication $\cdot : \mathbb{K} \times V \to V$ is defined such that $\forall \alpha, \beta \in \mathbb{K} \ \ \forall x, y \in V$

(1) $\alpha \cdot (x + y) = \alpha \cdot x + \alpha \cdot y$.

(2) $(\alpha + \beta) \cdot x = \alpha \cdot x + \beta \cdot x$.

(3) $\alpha \cdot (\beta \cdot x) = (\alpha \ \beta) \cdot x$.

(4) $1 \cdot x = x$.

As a consequence of the definition, $\forall x \in V$ we have

- $0 \cdot x = 0$.
- $(-1) \cdot x = -x$.

Every element $v \in V$ is called a vector. If the set of scalars is $\mathbb{C}$ ($\mathbb{R}$), V is named a complex (real) vector space. In this paper we will deal with complex vector spaces.

A *normed space* E is a vector space with a norm $\|\| : E \to \mathbb{R}^+$ such that $\forall x, y \in E$

(a) $\|x\| = 0$ if and only if $x = 0$.

(b) $\|\alpha \cdot x\| = |\alpha| \cdot \|x\|$.

(c) $\|x + y\| \leqslant \|x\| + \|y\|$.

Every normed space is a metric space by defining the distance between vectors as $D(x, y) = \|x - y\|$.

A *pre-Hilbert space* (or inner product space) is a vector space with an *inner product* $\langle , \rangle : H \times H \to \mathbb{C}$ such that $\forall \alpha_1, \alpha_2 \in \mathbb{C} \quad \forall x_1, x_2, x, y \in H$

(a) $\langle x, y \rangle = \overline{\langle y, x \rangle}$.

(b) $\langle \alpha_1 \cdot x_1 + \alpha_2 \cdot x_2, y \rangle = \alpha_1 \langle x_1, y \rangle + \alpha_2 \langle x_2, y \rangle$.

(c) $\langle x, x \rangle \geqslant 0$ (in particular $\langle x, x \rangle \in \mathbb{R}$).

(d) $\langle x, x \rangle = 0$ if and only if $x = 0$.

As a consequence of the definition, we have

- $\langle x, \alpha_1 \cdot y_1 + \alpha_2 \cdot y_2 \rangle = \overline{\alpha_1} \cdot \langle x, y_1 \rangle + \overline{\alpha_2} \cdot \langle x, y_2 \rangle$.
- $\langle \alpha \cdot x, \alpha \cdot x \rangle = |\alpha|^2 \langle x, x \rangle$.
- $\langle x, 0 \rangle = \langle 0, x \rangle = 0$.
- If $\forall z \in H \quad \langle x, z \rangle = \langle y, z \rangle$, then $x = y$.
- $|\langle x, y \rangle|^2 \leqslant \langle x, x \rangle \langle y, y \rangle$ (Schwartz inequality).
- $\langle x + y, x + y \rangle^{1/2} \leqslant \langle x, x \rangle^{1/2} + \langle y, y \rangle^{1/2}$.

With these properties, a pre-Hilbert space is a normed space, defining $\|x\| = \langle x, x \rangle^{1/2}$. In this case, Schwartz inequality says

$$|\langle x, y \rangle| \leqslant \|x\| \, \|y\| \,, \tag{1}$$

from which we can derive the following properties:

- The inner product is a continuous function with regard to every one of its arguments. That is, for every $y_0 \in H$, the functions $p_1, p_2 : H \to \mathbb{C}$ defined as $p_1(x) = \langle x, y_0 \rangle$, $p_2(x) = \langle y_0, x \rangle$ are continuous.
- $\|x - y\|^2 \geqslant (\|x\| - \|y\|)^2$.
- The norm is a continuous function. That is, the function $n : H \to \mathbb{C}$ defined as $n(x) = \|x\|$ is continuous.

Hence, a pre-Hilbert space is in particular a normed space and a metric space. Another interesting property of pre-Hilbert spaces is the parallelogram law:

$$\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2.$$

In fact, it can be proved that if a normed space satisfies the parallelogram law, then it is possible to define an inner product [Yosida 1965] by means of

$$\langle x, y \rangle = \frac{1}{4}(\|x + y\|^2 - \|x - y\|^2)$$

In a pre-Hilbert space H, two vectors x, y are *orthogonal* if $\langle x, y \rangle = 0$. A vector system is orthonormal if the vectors are mutually orthogonal, and all of them have norm 1. A set A of vectors is said to be *closed* in H if every vector $x \in H$ can be approximated to any degree of accuracy by a linear combination of vectors from A, that is

$$\forall \varepsilon > 0 \quad \exists n \geqslant 1 \quad \exists x_1, \cdots, x_n, \in A \quad \exists \alpha_1, \cdots, \alpha_n \in \mathbb{C} \quad \left\| x - \sum_{k=1}^{n} \alpha_k x_k \right\| < \varepsilon.$$

Equivalently, it is said to be complete, total, fundamental, or that its linear span is dense in H.

In a metric space M, a sequence $\{x_n\}_{n \in \mathbb{N}}$ converges towards a point x_0 if $\lim_{n \to \infty} D(x_n, x_0) = 0$. By triangular inequality, this definition implies $\lim_{n,m \to \infty} D(x_n, x_m) = 0$ (*Cauchy sequence*). However, the converse is not necessarily true. A metric space M is said to be *complete* if all Cauchy sequences converge towards a certain element $x \in M$, that is

$$\lim_{n,m \to \infty} D(x_n, x_m) = 0 \iff \exists x \in M \quad \lim_{n \to \infty} D(x_n, x) = 0.$$

A *Banach space* is a complete normed space. A *Hilbert space* is a complete pre-Hilbert space. So, a Hilbert space is, in particular, a Banach space.

Some example of Hilbert spaces are:

- $\mathbb{R}^N$, with $\langle u, v \rangle = \sum_{k=1}^{N} u_k v_k$.
- $\mathbb{C}^N$, with $\langle u, v \rangle = \sum_{k=1}^{N} u_k \overline{v_k}$.
- l^2, the set of all series such that $\sum_{k=1}^{\infty} |u_k|^2 < \infty$, with

$$\langle u, v \rangle = \sum_{k=1}^{\infty} u_k \overline{v_k}.$$

- $L^2(\Delta)$ (or simply L^2), the space of measurable functions $f : \Delta \to \mathbb{C}$, with $\Delta \subseteq \mathbb{R}^N$ measurable, such that the Lebesgue integral

$$\int_{\Delta} |f(x)|^2 dx = \int_{\Delta} f(x)\overline{f(x)} dx$$

  is finite. The inner product is defined as $\langle u, v \rangle = \int_{\Delta} u(x)\overline{v(x)} dx$ (also finite by Hölder inequality), and the norm is denoted as $\|u\|_2$. Clearly, with this definition there are functions different from 0 such that $\langle x, x \rangle = 0$. So it would not be a proper Hilbert space. To avoid this problem, it is necessary to regard the elements of this space not as functions, but rather as the equivalence classes resultant of considering two functions equivalent if they are equal almost everywhere (a.e), that is, equal except in a measure zero set. Abusing of notation, the quotient space is denoted $L^2(\Delta)$ again. With this definition, $f = 0$ if and only if $\|f\|_2 = 0$.

3 Definition of SAOC and basic properties

The problem of approximation in Hilbert spaces that we will deal with in this paper can be defined as follows: "Let H be a Hilbert space, Ω a space of parameters, and $f \in H$ a vector to approximate with vectors $v_\omega = v(\omega)$, $v : \Omega \to H$, $\omega \in \Omega$, such that $\forall \omega \in \Omega \quad \|v_\omega\| \neq 0$. We want to find $\omega_1, \omega_2, \cdots \in \Omega$ and $\lambda_1, \lambda_2, \cdots \in \mathbb{C}$ such that

$$\lim_{N \to \infty} \left\| f - \sum_{k=1}^{N} \lambda_k v_{\omega_k} \right\| = 0.$$

We will call frequencies to the elements $\omega_1, \omega_2, \cdots \in \Omega$, and coefficients to $\lambda_1, \lambda_2, \cdots \in \mathbb{C}$".

This definition is, in essence, the usual one in approximation of vectors in

Hilbert spaces. Observe that every vector $v_\omega \in H$ depends on a parameter $\omega \in \Omega$. Once we fix the parameter, we have a vector in the Hilbert space. The fact that with every frecuency ω_0 we only have a vector $v(\omega_0)$ is not a real restriction. We could have defined $v : \Omega \to H^Q$ with $Q \geqslant 1$ and pick one of its components. For the sake of simplicity, we will consider this notation.

Definition. Let H be a Hilbert space, Ω a space of parameters, and $f \in H$ a vector to approximate with vectors $v_\omega = v(\omega)$, $v : \Omega \to H$, $\omega \in \Omega$, such that $\forall \omega \in \Omega \quad \|v_\omega\| \neq 0$. A *Sequential Approximation of f with Optimal Coefficients (SAOC)* is a sequence of vectors $\{X_N\}_{N \geqslant 0}$, which terms are defined as:

- $X_0 = 0$.
- $X_N = \sum_{k=1}^{N-1} \lambda_k^N v_{\omega_k} + \lambda_N^N v_{\omega_N}$, so that
  (a) The coefficients are optimal. That is, X_N is the best approximation of f with the vectors $v_{\omega_1}, \cdots, v_{\omega_{N-1}}, v_{\omega_N}$.
  (b) $\forall \mu \in \mathbb{C} \quad \forall \omega_0 \in \Omega \quad \|f - X_N\|^2 \leqslant \|f - (X_{N-1} + \mu v_{\omega_0})\|^2$. That is, the approximation of f with X_N is better than the best approximation of the residue $f - X_{N-1}$ that one could achieve with only one vector $v_{\omega_0} \in v(\Omega)$ (or, equivalently, keeping fixed the coefficients of X_{N-1}).

Remarks.

- We can suppose that $\|f\| \neq 0$. If not, the approximation is trivial with $N = 0$. The condition $\|v_\omega\| \neq 0$ is equivalent, by inner product's properties, to that of $v_\omega \neq 0$. In fact, this condition is not a restriction, because the vector 0 cannot approximate any vector.

- At step N, a new frequency (ω_N) is considered, the number of terms of the approximation is increased by one ($\lambda_N^N v_{\omega_N}$), and the coefficients $\lambda_1^N, \lambda_2^N, \cdots, \lambda_{N-1}^N$ are recalculated in order to obtain the best approximation of f with $v_{\omega_1}, \cdots, v_{\omega_{N-1}}, v_{\omega_N}$. The frequencies $\omega_1, \omega_2, \cdots, \omega_{N-1}$ are kept fixed. Observe that $\forall N \geqslant 0 \quad X_N \in H$.

- Since X_N is the best approximation of f with $v_{\omega_1}, \cdots, v_{\omega_{N-1}}, v_{\omega_N}$, it holds that [Achieser 1956]

$$\forall k : 1 \leqslant k \leqslant N \quad \langle f - X_N, v_{\omega_k} \rangle = 0. \tag{2}$$

  That is, the residue $f - X_N$ is orthogonal to the space generated by $v_{\omega_1}, \cdots, v_{\omega_{N-1}}, v_{\omega_N}$. Equivalently, X_N is the orthogonal projection of

f onto the space spanned by $v_{\omega_1}, \cdots, v_{\omega_N}$. As we will see, this is one of the keys of *SAOC*'s properties. By inner product's definition, (2) is equivalent to the following linear equations system:

$$\begin{pmatrix} \langle v_{\omega_1}, v_{\omega_1} \rangle & \langle v_{\omega_2}, v_{\omega_1} \rangle & \cdots & \langle v_{\omega_N}, v_{\omega_1} \rangle \\ \langle v_{\omega_1}, v_{\omega_2} \rangle & \langle v_{\omega_2}, v_{\omega_2} \rangle & \cdots & \langle v_{\omega_N}, v_{\omega_2} \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle v_{\omega_1}, v_{\omega_N} \rangle & \langle v_{\omega_2}, v_{\omega_N} \rangle & \cdots & \langle v_{\omega_N}, v_{\omega_N} \rangle \end{pmatrix} \begin{pmatrix} \lambda_1^N \\ \lambda_2^N \\ \vdots \\ \lambda_N^N \end{pmatrix} = \begin{pmatrix} \langle f, v_{\omega_1} \rangle \\ \langle f, v_{\omega_2} \rangle \\ \vdots \\ \langle f, v_{\omega_N} \rangle \end{pmatrix}. \quad (3)$$

Consequently, once the frequencies $\omega_1, \cdots, \omega_{N-1}, \omega_N \in \Omega$ are fixed, the optimal coefficients $\lambda_1^N, \cdots, \lambda_{N-1}^N, \lambda_N^N \in \mathbb{C}$ can be calculated by solving (3). It can be proved easily that the system has only one solution if and only if $v_{\omega_1}, \cdots, v_{\omega_{N-1}}, v_{\omega_N}$ are linearly independent. Otherwise, the system has more than one solution. Since the frequencies $\omega_1, \omega_2, \cdots, \omega_{N-1}$ are kept fixed, the proposed system at step N is equal to the system at step $N-1$, but with a new row and a new column. The system solution is a continuous function of the matrix and the independent vector elements at any point where the matrix is nonsingular [Ortega 1972]. By inner product's definition, the matrix in (3) is hermitian. If the inner product is real, then it is symmetric.

- If $H = L^2$ and we only have a dataset X, the inner products can be approximated with a pass through X:

$$\langle v_{\omega_i}, v_{\omega_j} \rangle \cong \frac{1}{|X|} \sum_{x \in X} v_{\omega_i}(x) \overline{v_{\omega_j}(x)}$$

$$\langle f, v_{\omega_j} \rangle \cong \frac{1}{|X|} \sum_{x \in X} f(x) \overline{v_{\omega_j}(x)}.$$

  In this case we will suppose that the integral is defined with regard to the probability measure of the problem represented by the dataset. This is similar to approximate the expectation of a random variable by the arithmetic mean. In addition, solving (3) is equivalent to solving the Least Squares (LS) problem associated with the dataset.

- Observe that, in principle, there can be more than one frequency ω_N such that, together with its optimal coefficients, satisfy the property (b) of *SAOC*'s definition. On one end there would be those frequencies that are on the upper limit of the inequality, that is, those satisfying

$$\|f - X_N\|^2 = \inf_{\mu \in \mathbb{C}\ \omega_0 \in \Omega} \|f - (X_{N-1} + \mu v_{\omega_0})\|^2.$$

In this case, the residue norm obtained is the same that the one that would be obtained by approximating in an optimal manner the residue $f - X_{N-1}$ with only one vector of $v(\Omega)$. On the other end there would be the optimal frequencies, those satisfying

$$\forall \mu_1, \cdots, \mu_{N-1}, \mu_N \in \mathbb{C} \quad \forall \omega_0 \in \Omega$$
$$\|f - X_N\|^2 \leqslant \left\| f - \left(\sum_{k=1}^{N-1} \mu_k v_{\omega_k} + \mu_N v_{\omega_0} \right) \right\|^2 .$$

Clearly, the residue norm in the second case is less than in the first one, and therefore the approximation is better. On the contrary, the difficulty of finding the optimal frequency is probably greater than that of finding any other frequency satisfying the aforementioned property (b). In practice, there will be a trade-off between these two matters.

- In order to be well-defined, at every step there must be at least one frequency ω_N that satisfy the $SAOC$ definition. This existence may, in principle, depend on the Hilbert space H, the vector f and the set of vectors $v(\Omega)$. In the next section we present sufficient conditions to guarantee the existence of this frequency.

- The vectors $v_{\omega_1}, \cdots, v_{\omega_{N-1}}, v_{\omega_N}$ are not necessarily mutually orthogonal. The approximation with orthogonal vectors has been widely studied [Achieser 1956]. The coefficients of the best approximation of $f \in H$ by means of an orthogonal system $g_1, \cdots, g_N$ only depend on the projections of f onto the vectors of the system:

$$Y_N = \sum_{k=1}^{N} \lambda_k \, g_k = \sum_{k=1}^{N} \frac{\langle f, g_k \rangle}{\|g_k\|^2} \, g_k .$$

In this case, we have

$$\begin{aligned}
\|f - Y_N\|^2 &= \left\| f - \sum_{k=1}^{N} \lambda_k \, g_k \right\|^2 \\
&= \|f\|^2 - 2\mathrm{Re}\left(\left\langle f, \sum_{k=1}^{N} \lambda_k \, g_k \right\rangle \right) + \left\| \sum_{k=1}^{N} \lambda_k \, g_k \right\|^2 \\
&= \|f\|^2 - 2\mathrm{Re}\left(\sum_{k=1}^{N} \overline{\lambda_k} \, \langle f, g_k \rangle \right) + \sum_{k=1}^{N} |\lambda_k|^2 \|g_k\|^2 \\
&= \|f\|^2 - 2\mathrm{Re}\left(\sum_{k=1}^{N} \frac{\overline{\langle f, g_k \rangle}}{\|g_k\|^2} \, \langle f, g_k \rangle \right) + \sum_{k=1}^{N} |\lambda_k|^2 \|g_k\|^2 \\
&= \|f\|^2 - \sum_{k=1}^{N} |\lambda_k|^2 \|g_k\|^2 \\
&= \|f\|^2 - \left\| \sum_{k=1}^{N} \lambda_k \, g_k \right\|^2 = \|f\|^2 - \|Y_N\|^2.
\end{aligned}$$

In particular, $\forall N \geqslant 1 \quad \|f - Y_{N+1}\|^2 \leqslant \|f - Y_N\|^2$. As can be easily seen, the information provided by every term is independent of the others. This allows, for instance, the construction of the approximation in a sequential manner, where the terms are added one at a time until the approximation is satisfactory. If the orthogonal system is infinite, we may wonder about the behaviour of the series

$$g = \sum_{k=1}^{\infty} \frac{\langle f, g_k \rangle}{\|g_k\|^2} \, g_k.$$

If the system is closed in H, then the series is always convergent, and $\|f - g\| = 0$ (Parseval equation). As we will see below, the *SAOC* holds most of these properties. The main problem of approximating with a fixed system resides on the lack of flexibility. Linear expansions in a single basis are not flexible enough. The information can be diluted across the whole basis [Mallat & Zhang 1993]. For this reason, to achieve a good approximation, a very large number of vectors may be needed, even if we order them by $|\langle f, g_k \rangle|$. The *SAOC* keeps the idea of adding terms one at a time, but the residue can be reduced in a flexible and (in some sense) optimal manner. So we can expect to reduce the necessary number of terms to achieve the same degree of approximation.

As a first result, the approximations that satify (2) are characterized.

Lemma 1. Let H be a Hilbert space, $f \in H$ and $X_N = \sum_{k=1}^{N} \lambda_k^N v_{\omega_k}$, such that its vectors and coefficients satisfy (2). Then,

(L1a) $\forall j : 1 \leqslant j \leqslant N \quad \lambda_j^N = \frac{\left\langle f - \sum_{k=1,k\neq j}^{N} \lambda_k^N v_{\omega_k}, v_{\omega_j} \right\rangle}{\left\| v_{\omega_j} \right\|^2}$.

(L1b) $\forall j : 1 \leqslant j \leqslant N \quad \|f - X_N\|^2 = \left\| f - \sum_{k=1,k\neq j}^{N} \lambda_k^N v_{\omega_k} \right\|^2 - \left|\lambda_j^N\right|^2 \left\| v_{\omega_j} \right\|^2$.

(L1c) $\|f - X_N\|^2 = \|f\|^2 - \|X_N\|^2$ (energy conservation).

(L1d) $\|X_N\|^2 = \sum_{k=1}^{N} \lambda_k^N \overline{\langle f, v_{\omega_k} \rangle}$.

(L1e) $\langle f - X_N, f \rangle = \|f - X_N\|^2$.

Remarks.

- As an immediate consequence, every element X_N of the *SAOC* satisfies these properties.
- There is a great parallelism between these properties and those satisfied by an approximation with orthogonal vectors (see above). The only differences are in (L1a) and (L1b), and both are a generalization. Using orthogonal vectors, (L1a) and (L1b) would be converted, respectively, into

  $$\lambda_j^N = \frac{\left\langle f, v_{\omega_j} \right\rangle}{\left\| v_{\omega_j} \right\|^2}$$

  and

  $$\left\| f - \sum_{k=1,k\neq j}^{N} \lambda_k^N v_{\omega_k} \right\|^2 = \|f\|^2 - \sum_{k=1,k\neq j}^{N} \left|\lambda_k^N\right|^2 \|v_{\omega_k}\|^2.$$

  Both are known properties for orthogonal vectors. Hence, an important part of the good properties of the approximations with orthogonal vectors are actually a consequence of the fact that its coefficients are optimal, more than a consequence of the orthogonality itself.

Proof.

(L1a) By (2) we have $\left\langle f - \left(\sum_{k=1,k\neq j}^{N} \lambda_k^N v_{\omega_k} + \lambda_j^N v_{\omega_j} \right), v_{\omega_j} \right\rangle = 0$, and as a consequence

$$\left\langle f - \sum_{k=1,k\neq j}^{N} \lambda_k^N v_{\omega_k}, v_{\omega_j} \right\rangle = \lambda_j^N \left\langle v_{\omega_j}, v_{\omega_j} \right\rangle = \lambda_j^N \left\| v_{\omega_j} \right\|^2.$$

(L1b) Descomposing $X_N = \sum_{k=1,k\neq j}^{N} \lambda_k^N v_{\omega_k} + \lambda_j^N v_{\omega_j}$ we have

$$\|f - X_N\|^2 = \left\| f - \sum_{k=1,k\neq j}^{N} \lambda_k^N v_{\omega_k} \right\|^2 + \left|\lambda_j^N\right|^2 \left\|v_{\omega_j}\right\|^2 - 2\mathrm{Re}\left(\left\langle f - \sum_{k=1,k\neq j}^{N} \lambda_k^N v_{\omega_k}, \lambda_j^N v_{\omega_j} \right\rangle \right).$$

By (2) we can state

$$\begin{aligned}
&\|f - X_N\|^2 = \\
&= \left\| f - \sum_{k=1,k\neq j}^{N} \lambda_k^N v_{\omega_k} \right\|^2 + \left|\lambda_j^N\right|^2 \left\|v_{\omega_j}\right\|^2 - \\
&\qquad\qquad 2\mathrm{Re}\left(\left\langle X_N - \sum_{k=1,k\neq j}^{N} \lambda_k^N v_{\omega_k}, \lambda_j^N v_{\omega_j} \right\rangle \right) \\
&= \left\| f - \sum_{k=1,k\neq j}^{N} \lambda_k^N v_{\omega_k} \right\|^2 + \left|\lambda_j^N\right|^2 \left\|v_{\omega_j}\right\|^2 - 2\mathrm{Re}\left(\left\langle \lambda_j^N v_{\omega_j}, \lambda_j^N v_{\omega_j} \right\rangle \right) \\
&= \left\| f - \sum_{k=1,k\neq j}^{N} \lambda_k^N v_{\omega_k} \right\|^2 + \left|\lambda_j^N\right|^2 \left\|v_{\omega_j}\right\|^2 - 2\left|\lambda_j^N\right|^2 \left\|v_{\omega_j}\right\|^2 \\
&= \left\| f - \sum_{k=1,k\neq j}^{N} \lambda_k^N v_{\omega_k} \right\|^2 - \left|\lambda_j^N\right|^2 \left\|v_{\omega_j}\right\|^2.
\end{aligned}$$

(L1c) Expressing f as $(f - X_N) + X_N$ we have

$$\|f\|^2 = \|f - X_N\|^2 + \|X_N\|^2 + 2\mathrm{Re}\left(\langle f - X_N, X_N\rangle\right).$$

By (2), $2\mathrm{Re}\left(\langle f - X_N, X_N\rangle\right) = 0$ holds.

(L1d) By the definition of X_N we have

$$\|X_N\|^2 = \langle X_N, X_N\rangle = \left\langle \sum_{k=1}^{N} \lambda_k^N v_{\omega_k}, X_N \right\rangle = \sum_{k=1}^{N} \lambda_k^N \langle v_{\omega_k}, X_N\rangle.$$

Since X_N satisfies (2) we have

$$\forall k : 1 \leqslant k \leqslant N \quad \langle v_{\omega_k}, X_N\rangle = \langle v_{\omega_k}, f\rangle,$$

Hence,

$$\|X_N\|^2 = \sum_{k=1}^{N} \lambda_k^N \langle v_{\omega_k}, f\rangle = \sum_{k=1}^{N} \lambda_k^N \overline{\langle f, v_{\omega_k}\rangle}.$$

(L1e) $\|f - X_N\|^2 = \langle f - X_N, f - X_N\rangle = \langle f - X_N, f\rangle - \langle f - X_N, X_N\rangle$.

By (2), $\langle f - X_N, X_N\rangle = 0$ holds.

□

4 Main results

4.1 Existence

First, we prove some results that establish sufficient conditions to assure the existence of the frequency ω_N in *SAOC*'s definition.

Lemma 2. Let H be a Hilbert space, $f \in H$, and suppose that the element X_{N-1} of the *SAOC* exists for some $N \geqslant 1$. Then,

(L2a) The function $R_N : v(\Omega) \to \mathbb{R}$ defined as

$$R_N(v_\omega) = \inf_{\mu_1,\cdots,\mu_{N-1},\mu_N \in \mathbb{C}} \left\| f - \left(\sum_{k=1}^{N-1} \mu_k v_{\omega_k} + \mu_N v_\omega \right) \right\|^2$$

is always well defined and is continuous at every point v_ω such that $\{v_{\omega_1}, \cdots, v_{\omega_{N-1}}, v_\omega\}$ is linearly independent.

(L2b) The function $P_N : v(\Omega) \to \mathbb{R}$ defined as

$$P_N(v_\omega) = \inf_{\mu \in \mathbb{C}} \|f - (X_{N-1} + \mu v_\omega)\|^2$$

is always well defined, can be computed as

$$P_N(v_\omega) = \|f - X_{N-1}\|^2 - \frac{|\langle f - X_{N-1}, v_\omega\rangle|^2}{\|v_\omega\|^2}, \tag{4}$$

and is continuous at $v(\Omega)$.

Remarks.

- Observe that $\forall \omega \in \Omega \quad R_N(v_\omega) \leqslant P_N(v_\omega)$. The functions R_N and P_N are those that, in some way, point out the frontier of the existence of the frequency ω_N in *SAOC*'s definition: the frequency ω_N exists if and only if $\exists \omega_0 \in \Omega$ such that $R_N(v_{\omega_0}) \leqslant \inf_{\omega \in \Omega} P_N(v_\omega)$. Observe that, in principle, it can be false (we can think, for example, at $R_N(x) = \frac{1}{x^4}$ and $P_N(x) = \frac{1}{x^2}$, with $x \in [1, \infty)$).

- In general, it is not possible to assure R_N to be continuous at the vectors v_ω such that $\{v_{\omega_1}, \cdots, v_{\omega_{N-1}}, v_\omega\}$ is linearly dependent. For example, suppose that $H = \mathbb{R}^3$ and X_1 satisfies $\|f - X_1\|^2 > 0$. Every vector $v_{\omega_2} \in v(\Omega)$ linearly independent with regard to v_{ω_1} makes possible an exact approximation of any vector lying on the plane generated by $\{v_{\omega_1}, v_{\omega_2}\}$. As a consequence, $R_N(v_{\omega_2}) = 0$ if and only if f lies on the plane generated by $\{v_{\omega_1}, v_{\omega_2}\}$. So there are vectors v_{ω_2} arbitrarily near to v_{ω_1} such that $R_N(v_{\omega_2}) = 0$. In contrast, $R_N(v_{\omega_1}) = \|f - X_1\|^2 > 0$. Observe that, in order to make the residue zero, the coefficient modulus grows uncontrolably in vectors very near to v_{ω_1}. If R_N were continuous, there would be possible to find conditions to assure that its infimum is attained at $v(\Omega)$ (see Proposition 1).

Proof.

(L2a) If $\{v_{\omega_1}, \cdots, v_{\omega_{N-1}}, v_\omega\}$ is linearly dependent, then $R_N(v_\omega) = \|f - X_{N-1}\|^2$ holds. Let v_ω be such that $\{v_{\omega_1}, \cdots, v_{\omega_{N-1}}, v_\omega\}$ is linearly independent. The value of $R_N(v_\omega)$ can be obtained as follows. First, obtain $\mu_1, \cdots, \mu_{N-1}, \mu_N \in \mathbb{C}$ imposing (2), that is, as the solution of the linear equations system

$$\begin{pmatrix} \langle v_{\omega_1}, v_{\omega_1} \rangle & \cdots & \langle v_{\omega_{N-1}}, v_{\omega_1} \rangle & \langle v_\omega, v_{\omega_1} \rangle \\ \vdots & \ddots & \vdots & \vdots \\ \langle v_{\omega_1}, v_{\omega_{N-1}} \rangle & \cdots & \langle v_{\omega_{N-1}}, v_{\omega_{N-1}} \rangle & \langle v_\omega, v_{\omega_{N-1}} \rangle \\ \langle v_{\omega_1}, v_\omega \rangle & \cdots & \langle v_{\omega_{N-1}}, v_\omega \rangle & \langle v_\omega, v_\omega \rangle \end{pmatrix} \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_{N-1} \\ \mu_N \end{pmatrix} = \begin{pmatrix} \langle f, v_{\omega_1} \rangle \\ \vdots \\ \langle f, v_{\omega_{N-1}} \rangle \\ \langle f, v_\omega \rangle \end{pmatrix}.$$

Next, compute $R_N(v_\omega) = \left\| f - \left(\sum_{k=1}^{N-1} \mu_k v_{\omega_k} + \mu_N v_\omega \right) \right\|^2$. The previous system has only one solution, since $\{v_{\omega_1}, \cdots, v_{\omega_{N-1}}, v_\omega\}$ is linearly independent. Therefore, R_N is always well defined.
To prove the continuity, let $v_\omega \in v(\Omega)$ such that $\{v_{\omega_1}, \cdots, v_{\omega_{N-1}}, v_\omega\}$ is linearly independent. Since the norm and the inner product are continuous functions with regard to every one of its arguments, there exists $\delta_N > 0$ small enough such that for every vector v'_ω satisfying $\left\| v_\omega - v'_\omega \right\| < \delta_N$, their associated linear equations systems are as near

as desired. As the matrix is nonsingular, the system solution is continuous with respect to the matrix and the independent vector elements. Using (L1c) and (L1d), we can compute $R_N(v_\omega)$ as

$$R_N(v_\omega) = \|f\|^2 - \left(\sum_{k=1}^{N-1} \mu_k \overline{\langle f, v_{\omega_k} \rangle} + \mu_N \overline{\langle f, v_\omega \rangle} \right),$$

a composition of continuous functions and, therefore, continuous.

(L2b) Since $P_N(v_\omega)$ is the residue of the best approximation of $f - X_{N-1}$ with the vector v_ω, the value of $P_N(v_\omega)$ can be obtained as follows. First, impose (2) in order to obtain the optimal coefficient $\mu_\omega \in \mathbb{C}$. In this case, μ_ω is such that $\langle f - X_{N-1} - \mu_\omega v_\omega, v_\omega \rangle = 0$. By inner product's properties, $\mu_\omega = \frac{\langle f - X_{N-1}, v_\omega \rangle}{\|v_\omega\|^2}$. This is always well defined, since the vectors norms do not vanish in $v(\Omega)$. Using (L1c) and (L1d) with $f = f - X_{N-1}$ and $X_N = \mu_\omega v_\omega$, we have

$$\begin{aligned} P_N(v_\omega) &= \|f - X_{N-1} - \mu_\omega v_\omega\|^2 \\ &= \|f - X_{N-1}\|^2 - \|\mu_\omega v_\omega\|^2 \\ &= \|f - X_{N-1}\|^2 - \mu_\omega \overline{\langle f - X_{N-1}, v_\omega \rangle} \\ &= \|f - X_{N-1}\|^2 - \frac{|\langle f - X_{N-1}, v_\omega \rangle|^2}{\|v_\omega\|^2}. \end{aligned}$$

In particular, P_N is always well defined. Since the norm and the inner product are continuous functions with regard to every one of its arguments, P_N is continuous at $v(\Omega)$.

□

Proposition 1. Let H be a Hilbert space, $f \in H$, and suppose that the element X_{N-1} of the *SAOC* exists for some $N \geqslant 1$. Suppose also that exists $C \subseteq v(\Omega)$ such that

(a) $\exists v_{\omega_C} \in C \quad \forall \omega \in \Omega \quad v_\omega \notin C \Longrightarrow P_N(v_{\omega_C}) \leqslant P_N(v_\omega)$. That is, if P_N attains the infimum, it is attained at a vector v_{ω_C} belonging to C.

(b) Every sequence of vectors in C has a subsequence that converges towards an element in $v(\Omega)$.

Then, the infimum of P_N is attained at $v(\Omega)$ and the frequency ω_N of the *SAOC* does exist.

Proof.
By means of (L1c) we know that $\mathrm{Im}(P_N) \subseteq [0, \|f - X_{N-1}\|^2]$, and therefore, it is bounded. Thus, the infimum of P_N does exist. Let

$$p_N = \inf_{\omega \in \Omega} P_N(v_\omega).$$

By hypothesis (a) we have

$$p_N = \inf_{v_\omega \in C} P_N(v_\omega).$$

By infimum's definition, there exists a vector sequence $\{v_{\omega'_k}\}_{k \geqslant 1}$ in C such that

$$p_N = \lim_{k \to \infty} P_N(v_{\omega'_k}).$$

By hypothesis (b), the sequence $\{v_{\omega'_k}\}_{k \geqslant 1}$ has a subsequence that converges towards an element in $v(\Omega)$. Abusing of notation, we will denote again this subsequence as $\{v_{\omega'_k}\}_{k \geqslant 1}$. That is,

$$\exists \omega'_0 \in \Omega \quad \lim_{k \to \infty} v_{\omega'_k} = v_{\omega'_0}.$$

The continuity of P_N finishes the proof: by Lemma 2, P_N is continuous at $v_{\omega'_0}$. Therefore,

$$\forall \varepsilon > 0 \quad \exists \delta(\varepsilon) > 0 \quad \left\| v_{\omega'_0} - v_\omega \right\| < \delta \Longrightarrow \left| P_N(v_{\omega'_0}) - P_N(v_\omega) \right| < \varepsilon.$$

For every $\varepsilon > 0$, let $\delta = \delta(\varepsilon/2)$ as stated in the previous affirmation. By definition of ω'_0 there exists $k_1 \in \mathbb{N}$ such that $\forall k \geqslant k_1 \quad \left\| v_{\omega'_0} - v_{\omega'_k} \right\| < \delta$. The previous affirmation implies that

$$\forall k \geqslant k_1 \quad \left| P_N(v_{\omega'_0}) - P_N(v_{\omega'_k}) \right| < \varepsilon/2.$$

Since p_N is the limit of the sequence $\{v_{\omega'_k}\}_{k \geqslant 1}$, there exists $k_2 \in \mathbb{N}$ such that

$$\forall k \geqslant k_2 \quad \left| P_N(v_{\omega'_k}) - p_N \right| < \varepsilon/2.$$

Let $k_0 = max(k_1, k_2)$. Therefore we have

$$\left| P_N(v_{\omega'_0}) - p_N \right| \leqslant \left| P_N(v_{\omega'_0}) - P_N(v_{\omega'_{k_0}}) \right| + \left| P_N(v_{\omega'_{k_0}}) - p_N \right| < \varepsilon.$$

Hence, $P_N(v_{\omega'_0}) = p_N$ holds. That is, the infimum of P_N is attained at $v(\Omega)$. Since R_N and P_N are always well defined, and $\forall \omega \in \Omega \;\; R_N(v_\omega) \leqslant P_N(v_\omega)$, we have $R_N(v_{\omega'_0}) \leqslant P_N(v_{\omega'_0}) = p_N$. Therefore, the frequency ω_N of the $SAOC$ does exist.

□

Corollary 1. Under the same conditions of Proposition 1, suppose that every sequence of vectors in $v(\Omega)$ has a subsequence that converges towards an element in $v(\Omega)$. Then, the infimum of P_N is attained at $v(\Omega)$ and the frequency ω_N of the *SAOC* does exist.

Proof.
It is derived immediatly from Proposition 1 with $C = v(\Omega)$. □

Corollary 2. Under the same conditions of Proposition 1, for every $\delta > 0$, define P_δ as

$$P_\delta = \left\{ v_\omega \in v(\Omega) \ : \ \frac{|\langle f - X_{N-1}, v_\omega \rangle|^2}{\|v_\omega\|^2} \geqslant \delta \right\}.$$

Suppose that there exist $\delta_N \geqslant 0$ and $C_{\delta_N} \subseteq v(\Omega)$ such that

(a) $P_{\delta_N} \neq \emptyset$ and $P_{\delta_N} \subseteq C_{\delta_N}$.

(b) Every subsequence of vectors in C_δ has a subsequence that converges towards an element in $v(\Omega)$.

Then, the infimum of P_N is attained at $v(\Omega)$ and the frequency ω_N of the *SAOC* does exist.

Proof.
By (L1c) we know that $\mathrm{Im}(P_N) \subseteq [0, \|f - X_{N-1}\|^2]$, and therefore, it is bounded. Thus, the infimum of P_N does exist. Let

$$p_N = \inf_{\omega \in \Omega} P_N(v_\omega) \leqslant \|f - X_{N-1}\|^2.$$

If $p_N = \|f - X_{N-1}\|^2$, then we can assign $\omega_N = \omega_{N-1}$ finishing the proof. Suppose that $p_N < \|f - X_{N-1}\|^2$. Then it is enough to prove that C_{δ_N} satisfies the hypothesis of Proposition 1.
By definition of P_δ we have

$$\forall \omega \in \Omega \quad v_\omega \notin C_{\delta_N} \Longrightarrow v_\omega \notin P_{\delta_N} \Longrightarrow \frac{|\langle f - X_{N-1}, v_\omega \rangle|^2}{\|v_\omega\|^2} < \delta_N.$$

As Lemma 2 states, for every $\omega \in \Omega$

$$P_N(v_\omega) = \|f - X_{N-1}\|^2 - \frac{|\langle f - X_{N-1}, v_\omega \rangle|^2}{\|v_\omega\|^2}.$$

Let $\omega_C \in P_{\delta_N}$. Every frequency $\omega \in \Omega$ such that $v_\omega \notin C_{\delta_N}$ satisfies

$$P_N(v_\omega) > \|f - X_{N-1}\|^2 - \delta_N \geqslant P_N(v_{\omega_C}).$$

□

It seems clear that the existence of the frequency will depend on f. But, especially, it will depend on $v(\Omega)$. Regarding to the problem of choosing Ω and $v(\Omega)$, there are two possible problems that must be taken into account in order to assure that the frequency does always exist:

1. For every convergent sequence of vectors there must exist a frequency which associated vector is the limit of that sequence. This must happen at least in a subset of $v(\Omega)$ where we can be sure that the infimum of P_N is attained. This "compactness" condition is necessary in order to apply Proposition 1 and its corolaries. Suppose, for example, that we want to approximate a vector that is the limit of a specific sequence. If it does not exist a frequency for that vector, the infimum of P_N will not be attained at $v(\Omega)$.

2. The norms of the vectors in $v(\Omega)$ should always be larger than some $\delta > 0$. At least, this must be true for the vectors that matches the residue $f - X_{N-1}$ as best as possible. In this way, the function $\frac{|\langle f - X_{N-1}, v_\omega \rangle|^2}{\|v_\omega\|^2}$ does never suffer from lack of definition when the norms of the vectors tend to 0, and the application of Corollary 2 will be surely simpler.

Specifically, one must take special care with the vector 0. Suppose that $H = L^2([-\pi, \pi])$, we are approximating by real Fourier series ($\Omega \subseteq \mathbb{R}$), and $f(x) = x$. A simple calculation allows us to assert that

$$\lim_{\omega \to 0} \left\| x - \frac{\langle x, \sin \omega x \rangle}{\|\sin \omega x\|^2} \sin \omega x \right\| = 0 \quad \text{and} \quad \lim_{\omega \to 0} \frac{\langle x, \sin \omega x \rangle}{\|\sin \omega x\|^2} = \infty.$$

Hence, the frequency that minimizes P_N should be, by continuity, the frequency 0. But the vector associated with this frequency is the vector 0, that cannot belong to $v(\Omega)$ because its norm is 0. To avoid this problem, Ω should be a closed subset without any element in a neighbourhood of 0. In this way, the existence of the frequency is assured at every step. As will be shown (see Theorem 2), this restriction does not reduce the approximation capability (assuming $\mathbb{N}^+ \subseteq \Omega$).

This kind of reflections lead us to enunciate the following result.

Theorem 1. Let H be a Hilbert space, and $f \in H$. Suppose that it is possible to define a topology in Ω such that

(a) Ω is compact.

(b) The function $v : \Omega \to H$ at $SAOC$'s definition is continuous on Ω, with the topology in H induced by the distance.

Then, for every $N \geqslant 1$ the infimum of P_N is attained at $v(\Omega)$ and the frequencies of the $SAOC$ do always exist.

Proof.
For a given sequence of vectors in $v(\Omega)$, consider its sequence of associated frequencies in Ω. Since Ω is compact, it contains a subsequence that converges towards an element ω_N (that satisfies $v_{\omega_N} \neq 0$). Since $v : \Omega \to H$ is continuous, the limit of the vectors associated with the frequencies of the subsequence is v_{ω_N}. Hence, every sequence of vectors in $v(\Omega)$ has a subsequence that converges towards an element in $v(\Omega)$. Corollary 1 finishes the proof. □

As we will see, these results can be applied to a number of vector families that are very usual in the literature. In the rest of the section we will suppose that the frequencies of the $SAOC$ always exist.

4.2 Convergence

Since the $SAOC$ is a sequence of vectors, wondering whether it converges or not is a natural question. If so, it would be of interest knowing whether it converges towards the target vector f or not. In the sequel these questions are answered.

Proposition 2. Let H be a Hilbert space, and $f \in H$. The $SAOC$ $\{X_N\}_{N \geqslant 0}$ satisfies the following properties:

(P2a) $\forall N \geqslant 0 \quad \|f - X_{N+1}\|^2 \leqslant \|f - X_N\|^2$.

(P2b) If $M \geqslant N$, then

(P2b1) $\|X_M\|^2 \geqslant \|X_N\|^2$.

(P2b2) $\langle f - X_M, f - X_N \rangle = \|f - X_M\|^2$.

(P2b3) $\langle X_M, f - X_N \rangle \in \mathbb{R}$ and $\langle X_M, f - X_N \rangle \geqslant 0$.

(P2c) Suppose that $\lambda_N^N \neq 0$. The vector v_{ω_N} is orthogonal to the space spanned by $\{v_{\omega_1}, \cdots, v_{\omega_{N-1}}\}$ if and only if the previous existing coefficients do not change between the steps $N-1$ and N. That is, if

$$\forall j : 1 \leqslant j \leqslant N-1 \quad \lambda_j^N = \lambda_j^{N-1}.$$

Remarks.

- Again there is a great parallelism between these properties and those satisfied by an approximation with orthogonal vectors (see above).
- (P2a) implies that the sequence $\{\|f - X_N\|^2\}_{N \geqslant 0}$ is decreasing and positive. Therefore, it is convergent. In principle, it does not imply $\{X_N\}_{N \geqslant 0}$ to be convergent (except if $\lim_{N\to\infty} \|f - X_N\|^2 = 0$). The convergence could depend, in principle, on H, f and $v(\Omega)$. We will see that the *SAOC*'s convergence is independent of these matters.
- By (P2c), the only directions that guarantee that, without recalculating the coefficients, the approximation is optimal are the orthogonal directions. Hence, if the approximation vectors are not mutually orthogonal, the coefficients must be recalculated.

Proof.

(P2a) Evident, by definition:
$\|f - X_{N+1}\|^2 \leqslant \left\|f - X_N + 0 \cdot v_{\omega_{N+1}}\right\|^2 = \|f - X_N\|^2$.

(P2b1) Evident, combining (P2a) and (L1c).

(P2b2) Expressing $f - X_N$ as $f - X_M + X_M - X_N$ we have

$$\langle f - X_M, f - X_N \rangle = \|f - X_M\|^2 + \langle f - X_M, X_M - X_N \rangle .$$

By (2), $\langle f - X_M, X_M - X_N \rangle = 0$ holds.

(P2b3) By (P2b2) and (L1e) we have

$$\begin{aligned} \|f - X_M\|^2 &= \langle f - X_M, f - X_N \rangle = \langle f, f - X_N \rangle - \langle X_M, f - X_N \rangle \\ &= \|f - X_N\|^2 - \langle X_M, f - X_N \rangle . \end{aligned}$$

The proof finishes with (P2a).

(P2c) The necessity is clear by (2). To prove the sufficiency, suppose that $\forall j : 1 \leqslant j \leqslant N-1 \quad \lambda_j^N = \lambda_j^{N-1}$. Hence $X_N = X_{N-1} + \lambda_N^N v_{\omega_N}$ holds. By (2) we have

$$\forall j : 1 \leqslant j \leqslant N \quad \left\langle f - X_N, v_{\omega_j} \right\rangle = 0,$$

$$\forall j : 1 \leqslant j \leqslant N-1 \quad \left\langle f - X_{N-1}, v_{\omega_j} \right\rangle = 0.$$

Therefore, $\forall j : 1 \leqslant j \leqslant N-1$ we have

$$0 = \left\langle f - X_N, v_{\omega_j} \right\rangle = \left\langle f - X_{N-1} - \lambda_N^N v_{\omega_N}, v_{\omega_j} \right\rangle = \left\langle \lambda_N^N v_{\omega_N}, v_{\omega_j} \right\rangle .$$

Since $\lambda_N^N \neq 0$, the vectors v_{ω_N} and v_{ω_j} are orthogonal for every j between 0 and $N-1$.

□

Theorem 2. Let H be a Hilbert space, and $f \in H$. Suppose that the element X_N of the *SAOC* does always exist for every $N \geqslant 0$. Then:

(T2a) The *SAOC* $\{X_N\}_{N \geqslant 0}$ is convergent in H. That is,

$$\exists g \in H \quad \lim_{N \to \infty} \|g - X_N\| = 0.$$

(T2b) In addition, g satisfies:

(T2b1) $\forall \omega_0 \in \Omega \quad \lim_{N \to \infty} \langle g - X_N, v_{\omega_0} \rangle = 0.$

(T2b2) $\forall \omega_0 \in \Omega \quad \langle f, v_{\omega_0} \rangle = \langle g, v_{\omega_0} \rangle$. In particular, we have

(T2b21) $\forall N \geqslant 0 \quad \langle f - g, X_N \rangle = 0.$

(T2b22) $\forall N \geqslant 1 \quad \forall j : 1 \leqslant j \leqslant N \quad \forall M \geqslant N \quad \left\langle g - X_M, v_{\omega_j} \right\rangle = 0.$

(T2b3) $\langle f - g, g \rangle = 0.$

(T2b4) There is no subset of vectors in $v(\Omega)$ that approximate f more than g. That is,

$$\|f - g\| = \inf_{\substack{\mu_k \in \mathbb{C} \\ \psi_k \in \Omega}} \left\| f - \sum_k \mu_k v_{\psi_k} \right\| .$$

(T2c) If there exists $A \subseteq \Omega$ such that the set of vectors $\{v_\psi : \psi \in A\}$ is closed in H, then $\{X_N\}_{N \geqslant 0}$ converges towards f. That is,

$$\lim_{N \to \infty} \|f - X_N\| = 0.$$

Remarks.

- These results are very little restrictive. In order to assure the convergence towards f, the family of vectors used in the construction must have the capability of approximating any vector. Hence, the *SAOC* allows us to choose any of the multiple vector families satisfying this property.
- In order to satisfy (T2a), it is enough for $\|f - X_N\|^2$ to be decreasing and positive, and X_N to satisfy (2). Clearly, this latter condition is the most important to assure the convergence. Observe, again, the importance of working with optimal coefficients.

Proof.

(T2a) Since H is complete, it is enough to prove that $\lim_{N,M\to\infty} \|X_M - X_N\|^2 = 0$. Suppose that $M > N$. Expressing $X_M - X_N$ as $(X_M - f) + (f - X_N)$, and using (P2b2) we have

$$\begin{aligned}\|X_M - X_N\|^2 &= \|f - X_M\|^2 + \|f - X_N\|^2 - 2\mathrm{Re}\left(\langle f - X_M, f - X_N\rangle\right) \\ &= \|f - X_M\|^2 + \|f - X_N\|^2 - 2\|f - X_M\|^2 \\ &= \|f - X_N\|^2 - \|f - X_M\|^2.\end{aligned}$$

Since the sequence $\{\|f - X_N\|^2\}_{N\geqslant 0}$ is decreasing and positive, it is convergent. Hence,

$$\lim_{N,M\to\infty} \|X_M - X_N\|^2 = \lim_{N,M\to\infty} \left(\|f - X_N\|^2 - \|f - X_M\|^2\right) = 0.$$

(T2b1) By Schwartz inequality (1) we have

$$\forall \omega_0 \in \Omega \quad |\langle g - X_N, v_{\omega_0}\rangle| \leqslant \|g - X_N\| \, \|v_{\omega_0}\| \, .$$

Using (T2a),

$$\forall \omega_0 \in \Omega \quad \lim_{N\to\infty} |\langle g - X_N, v_{\omega_0}\rangle| \leqslant \lim_{N\to\infty} \|g - X_N\| \, \|v_{\omega_0}\| = 0.$$

(T2b2) Let $\omega_0 \in \Omega$. By definition of X_N, for every $N \geqslant 0$ and every $\mu \in \mathbb{C}$

$$\begin{aligned}\|f - X_{N+1}\|^2 &\leqslant \|f - (X_N + \mu v_{\omega_0})\|^2 \\ &= \|f - X_N\|^2 - 2\mathrm{Re}\left(\langle f - X_N, \mu v_{\omega_0}\rangle\right) + |\mu|^2 \|v_{\omega_0}\|^2\end{aligned}$$

hold. Expressing $f - X_N$ as $f - g + g - X_N$ we have

$$\begin{aligned}&\|f - X_{N+1}\|^2 - \|f - X_N\|^2 \leqslant \\ &\quad |\mu|^2\|v_{\omega_0}\|^2 - 2\mathrm{Re}\left(\langle f - g, \mu v_{\omega_0}\rangle\right) - 2\mathrm{Re}\left(\langle g - X_N, \mu v_{\omega_0}\rangle\right).\end{aligned}$$

Hence,

$$\begin{aligned}&2\mathrm{Re}\left(\langle f - g, \mu v_{\omega_0}\rangle\right) - |\mu|^2\|v_{\omega_0}\|^2 \leqslant \\ &\quad \leqslant \|f - X_N\|^2 - \|f - X_{N+1}\|^2 - 2\mathrm{Re}\left(\langle g - X_N, \mu v_{\omega_0}\rangle\right) \\ &\quad \leqslant \|f - X_N\|^2 - \|f - X_{N+1}\|^2 + 2\left|\langle g - X_N, \mu v_{\omega_0}\rangle\right| \\ &\quad = \|f - X_N\|^2 - \|f - X_{N+1}\|^2 + 2\left|\mu\right|\left|\langle g - X_N, v_{\omega_0}\rangle\right|.\end{aligned}$$

for every $\mu \in \mathbb{C}$. Let $\mu_0 = \frac{\langle f-g, v_{\omega_0}\rangle}{\|v_{\omega_0}\|^2}$, and $\varepsilon > 0$. Since the sequence $\{\|f - X_N\|^2\}_{N\geqslant 0}$ is decreasing and positive, and using (T2b1), there exists N_0 such that $\forall N \geqslant N_0$,

$$0 \leqslant \|f - X_N\|^2 - \|f - X_{N+1}\|^2 \leqslant \varepsilon/2$$

and, in addition,

$$2\left|\mu_0\right|\left|\langle g - X_N, v_{\omega_0}\rangle\right| \leqslant \varepsilon/2.$$

Thus we have

$$2\mathrm{Re}\left(\langle f - g, \mu_0 v_{\omega_0}\rangle\right) - |\mu_0|^2\|v_{\omega_0}\|^2 \leqslant \varepsilon.$$

Since

$$\langle f - g, \mu_0 v_{\omega_0}\rangle = \overline{\mu_0}\langle f - g, v_{\omega_0}\rangle = \overline{\mu_0}\mu_0\|v_{\omega_0}\|^2 = |\mu_0|^2\|v_{\omega_0}\|^2,$$

$2\mathrm{Re}\left(\langle f - g, \mu_0 v_{\omega_0}\rangle\right) = 2|\mu_0|^2\|v_{\omega_0}\|^2$ holds, and therefore

$$\forall \varepsilon \geqslant 0 \qquad |\mu_0|^2\|v_{\omega_0}\|^2 = 2\mathrm{Re}\left(\langle f - g, \mu_0 v_{\omega_0}\rangle\right) - |\mu_0|^2\|v_{\omega_0}\|^2 \leqslant \varepsilon.$$

Hence, $|\mu_0|^2\|v_{\omega_0}\|^2 = 0$. Since $\|v_{\omega_0}\|^2 \neq 0$, we have $\mu_0 = 0$. Thus, by definition of μ_0, $\langle f - g, v_{\omega_0}\rangle = 0$ for every $\omega_0 \in \Omega$. In particular we have

(T2b21) $\forall N \geqslant 1 \quad \langle f - g, X_N\rangle = 0$

(T2b22) Using (2), $\forall j : 1 \leqslant j \leqslant N \quad \forall M \geqslant N$

$$\left\langle g - X_M, v_{\omega_j}\right\rangle = \left\langle g, v_{\omega_j}\right\rangle - \left\langle X_M, v_{\omega_j}\right\rangle = \left\langle g, v_{\omega_j}\right\rangle - \left\langle f, v_{\omega_j}\right\rangle = 0.$$

(T2b3) Expressing g as $g - X_N + X_N$, and using (T2b21), we can derive

$$\langle f-g, g\rangle = \langle f-g, g-X_N\rangle + \langle f-g, X_N\rangle = \langle f-g, g-X_N\rangle\,.$$

By Schwartz inequality (1) we have

$$|\langle f-g, g-X_N\rangle| \leqslant \|f-g\|\,\|g-X_N\|\,.$$

Using (T2a), $\lim_{N\to\infty} |\langle f-g, g-X_N\rangle| = 0$ holds. Therefore, $\langle f-g, g\rangle = 0$.

(T2b4) By (T2b2), any vector combination $\sum_k \mu_k v_{\psi_k}$ in $v(\Omega)$ satisfies $\langle f-g, \sum_k \mu_k v_{\psi_k}\rangle = 0$. Hence we have

$$\|f-g\|^2 = \langle f-g, f-g\rangle = \left\langle f-g, f-\sum_k \mu_k v_{\psi_k}\right\rangle - \langle f-g, g\rangle\,.$$

Using (T2b3) and Schwartz inequality (1) we have

$$\|f-g\|^2 = \left|\left\langle f-g, f-\sum_k \mu_k v_{\psi_k}\right\rangle\right| \leqslant \|f-g\|\left\|f-\sum_k \mu_k v_{\psi_k}\right\|.$$

Therefore, $\|f-g\| \leqslant \|f-\sum_k \mu_k v_{\psi_k}\|$. The other inequality is clear, since for every $N \geqslant 0$

$$\inf_{\substack{\mu_k \in \mathbb{C}\\ \psi_k \in \Omega}} \left\|f-\sum_k \mu_k v_{\psi_k}\right\| \leqslant \|f-X_N\| \leqslant \|f-g\| + \|g-X_N\|\,.$$

The proof finishes using (T2a).

(T2c) It is derived immediately from (T2b4).

□

4.3 Methodology

A possible methodology to approximate any function $f \in H$ with the *SAOC* approach could be the following:

1. Choose the desired approximation accuracy $\varepsilon \geqslant 0$.
2. Select Ω and $v(\Omega)$ such that

(a) The frequencies ω_N of the *SAOC* do always exist. (Theorem 1).

(b) There exists a set of frequencies $A \subseteq \Omega$ such that it is possible to approximate f less than ε with the set of vectors $\{v_\psi : \psi \in A\}$ (Theorem 2).

3. Construct the *SAOC* $\{X_N\}_{N \geqslant 0}$.

After these steps, the deviation between f and the obtained approximation (in the limit) is less than ε.

5 Specific vectors in $H = L^2$

From now on we will work in the space L^2. The vector to approximate is a square integrable function $f(\vec{t})$. In *SAOC*'s definition there are hardly restrictions about the vectors $v_{\omega_k}(\vec{t}) = v(\omega_k, \vec{t}) \in H$ used to approximate f. The only required condition is to have a norm different from 0. Thus, the method can be applied to a number of vector families that are very usual in the literature. The vectors discussed here are not, surely, the only ones that can be used within the method. In fact, the universal approximation capability of a family of functions is enough to apply the *SAOC* method with guarantee of convergence (whenever the *SAOC* construction is feasible).
Before proceeding to enumerate some of these families, it is convenient to remember some results about denseness in L^2 (a subset M of a topological space X is said to be dense in X if $\overline{M} = X$, where $\overline{M}$ is the topological closure of M):

1. If K is compact, the set of continuous functions in K is dense in $L^2(K)$ [Reddy 1998], and $L^2(K) \subsetneq L^1(K)$.

2. The set of continuous functions in $\mathbb{R}^I$ with compact support is dense in $L^2(\mathbb{R}^I)$ [Rudin 1987]. The support of a function is the closure of the set where the function is different from 0. Since a continuous function with compact support belongs to $L^2(K)$, where K is its support, every function in $L^2(\mathbb{R}^I)$ may be approximated by square integrable functions on a compact.

5.1 Approximation by polynomials

The celebrated Weierstrass theorem states that any real continuous function on $[a, b] \subseteq \mathbb{R}$ can be approximated uniformly by an algebraic polynomial [Weierstrass 1885]. By the Stone-Weierstrass theorem for real functions, this

result can be extended to real continuous functions on a compact of $\mathbb{R}^I$ [Lang 1989], with the usual definition of a polynomial in several variables. By [Reddy 1998], any real function in $L^2(K)$ can be approximated by polynomials

$$P(t_1, \cdots, t_I) = \sum \lambda_{n_1 \cdots n_I} t_1^{n_1} \cdots t_I^{n_I},$$

where the coefficients are real, and the sum is taken over a finite number of I-tuples $(n_1, \cdots, n_I) \in \mathbb{N}^I$. That is, the set of polynomials $\{t_1^{n_1} \cdots t_I^{n_I} : (n_1, \cdots, n_I) \in \mathbb{N}^I\}$ is closed in the space of real functions of $L^2(K)$. Hence, the *SAOC* can be applied to any function of this space considering vectors $v_\omega(t_1, \cdots, t_I) = t_1^{\omega_1} \cdots t_I^{\omega_I}$ $(\Omega \subseteq \mathbb{R}^I)$. Observe that, in this case, the coefficients of the *SAOC* can be supposed to be real. If Ω is compact, the existence is guaranteed (see Theorem 1).

5.2 Approximation by nonharmonic Fourier series

Suppose that $H = L^2([-\pi, \pi])$ and we are approximating by nonharmonic Fourier series with vectors $v_\omega(t) = e^{i\omega t}$ $(\Omega \subseteq \mathbb{R})$. Since $\{e^{ikt} : k \in \mathbb{Z}\}$ is closed in $L^2([-\pi, \pi])$ [Young 1980], the *SAOC* can be applied to any function in this space. In this case we have $\forall \omega_0 \in \Omega \quad \|v_{\omega_0}\| = \sqrt{2\pi}$, and $\langle f - X_N, v_\omega \rangle$ is the value of the Fourier transform of the residue in the vector v_ω. By *SAOC*'s definition, together with (4), the vector associated with the new frequency ω_N allows a better approximation of f than the vector that maximizes the modulus of the Fourier transform of the residue. If the function is real and is directly approximated by sines and cosines, ($v_\omega(t) = \sin(\omega t)$ or $v_\omega(t) = \cos(\omega t)$, $\omega \in \mathbb{R}$), the results are the same [Achieser 1956]. In this case the Fourier transform is normalized by the vector norm (see (4)), that may be nonconstant for the different frequencies. A linear change of variable allows the obtaining of the same results in $L^2([a, b])$. By the Stone-Weierstrass theorem for complex functions, the continuous functions on $[-\pi, \pi]^I$ can be approximated by trigonometric polynomials

$$P(\vec{t}) = \sum \lambda_{\vec{k}_n} e^{i\vec{k}_n \cdot \vec{t}},$$

where the sum is taken over a finite number of I-tuples $\vec{k}_n \in \mathbb{Z}^I$, and $\vec{k}_n \cdot \vec{t}$ represents the inner product in $\mathbb{R}^I$. By [Reddy 1998], any function in $L^2([-\pi, \pi]^I)$ can be approximated by the trigonometric polynomials aforementioned. Hence, the *SAOC* can be applied to any function of this space with vectors $v_\omega(\vec{t}) = e^{i\vec{\omega} \cdot \vec{t}}$, $(\Omega \subseteq \mathbb{R}^I)$. As in the one-dimensional case, a linear change of variable allows the obtaining of the same results in any hypercube of $\mathbb{R}^I$, and therefore in any compact. In this case the existence is

guaranteed both if Ω is compact (see Theorem 1) and if $\Omega = \mathbb{R}^I$: A known result states that the Fourier transform of a function in $L^1(K)$ tends to 0 as the frequency modulus tends to infinite [Stein & Weiss 1971]. Since every function in $L^2(K)$ belongs to $L^1(K)$, this also applies to $f - X_{N-1}$. Hence, $|\langle f - X_{N-1}, v_\omega \rangle|$ tends to 0 if $|\omega|$ tends to infinite. The vector norm $\|v_\omega\|$ is constant. We can apply Corollary 2: for every $\delta > 0$ there exists a hypercube $C_\Delta \subseteq \mathbb{R}^I$ such that $P_\delta \subseteq v(C_\Delta) = C_\delta$. Since C_Δ is compact, every sequence of vectors in C_δ has a subsequence convergent towards an element in C_δ.
If the function is real, it can be approximated by sines and cosines ($v_\omega(\vec{t}) = \sin(\vec{\omega} \cdot \vec{t})$ or $v_\omega(\vec{t}) = \cos(\vec{\omega} \cdot \vec{t})$, $\vec{\omega} \in \Omega \subseteq \mathbb{R}^I$). If we use sines, Ω cannot contain frequencies arbitrarily near to 0.

5.3 Approximation by wavelets

A similar result can be found in the field of signal processing. A signal is a function $f \in L^2(\mathbb{R})$. A wavelet is a function $\phi \in L^2(\mathbb{R})$ such that $\|\phi\| = 1$, $\int_{-\infty}^{\infty} \phi(t)dt = 0$ and is centered in the neighborhood of $t = 0$. The wavelets are translated and scaled to build a family of time-frequency atoms

$$\phi_{u,s}(t) = \frac{1}{\sqrt{s}} \, \phi\left(\frac{t-u}{s}\right) \quad u \in \mathbb{R} \;\; s \in \mathbb{R}^+,$$

also satisfying $\|\phi_{u,s}\| = 1$. In practice, the most common wavelets are continuous (except, maybe, in a finite set) and bounded, such as *Mexican hats* (second derivative of a gaussian) or Gabor wavelets ($\phi(t) = g(t)e^{i\eta t}$, with $g(t)$ a gaussian). There are wavelets such that a subset of its family of time-frequency atoms is an orthonormal basis and, therefore, closed in $L^2(\mathbb{R})$ [Mallat 1998]. With the same definitions, and considering $v_\omega(t) = \phi_{u,s}(t)$ ($\Omega \subseteq \mathbb{R} \times \mathbb{R}^+$), the *SAOC* can be applied to any signal $f \in L^2(\mathbb{R})$. To guarantee the existence it is enough for s to belong to a compact in $\mathbb{R}^+$ not containing 0 (if u tends to infinite, $|\langle f - X_{N-1}, v_\omega \rangle|$ tends to 0, whereas the norm of v_ω keeps constant). With an analogous reasoning to that of Fourier series, the vector associated with the new frequency ω_N allows a better approximation of f than the vector that maximizes the modulus of the wavelet transform of the residue. The wavelet transform is, in this case, the projection of the residue onto the vector.

5.4 Approximation by neural networks

In Artificial Intelligence, neural networks have been shown to be a very suitable mechanism to approximate functions. A feed-forward neural network

(FNN) with a single hidden layer and a linear output unit approximates a function $f : \mathbb{R}^I \to \mathbb{R}$ as follows:

$$f_N(\vec{t}) = b_0 + \sum_{k=1}^{N} \lambda_k \, \varphi\left(\vec{\omega_k} \cdot \vec{t} + b_k\right) \quad \vec{\omega_k} \in \mathbb{R}^I \;\; b_k \in \mathbb{R}, \tag{5}$$

where $\vec{\omega_k} \cdot \vec{t}$ represents the inner product in $\mathbb{R}^I$, N is the number of units in the hidden layer, λ_k is the weight of the connection between the unit k in the hidden layer and the output unit, $\vec{\omega_k}$ is the weight vector associated with the connections between the input layer units and the unit k of the hidden layer and $\varphi : \mathbb{R} \to \mathbb{R}$ is the so called activation function. The biases (b_k) are external parameters for each unit, and can be considered as weights with a simple transformation. The most usual activation functions are sigmoidal-like (continuous, nonconstant, increasing and bounded), such as the logistic function $\varphi(x) = \frac{1}{1+e^{-x}}$ or the hyperbolic tangent function $\varphi(x) = \frac{1-e^{-x}}{1+e^{-x}}$. In [Funahashi 1989] it is proved that FNNs with a hidden layer of sigmoidal units and a linear output layer can approximate any function $f : K \to \mathbb{R}^O$ in $L^2(K)$, with K compact. Hence, the $SAOC$ with sigmoidal functions $(\Omega \subseteq \mathbb{R}^I \times \mathbb{R})$ can be applied to any real function in this space. In this case, the existence is guaranteed as follows. If the modulus of $\vec{\omega}$ can be arbitrarily large, there will be necessary to consider the accummulation points of the selected sigmoidal functions (that will be step functions) to be able to guarantee the "compactness" of the vector space. If any of the asymptotes of φ is 0, the biases cannot be arbitrarily large (since in the limit we would have the function 0).
More recent results prove that with any activation functions different from an algebraic polynomial (a.e.), neural networks are universal approximators [Leshno et al. 1993].

Other kind of neural networks, the radial basis function networks (RBFN), approximate a function $f : \mathbb{R}^I \to \mathbb{R}$ as follows:

$$f_N(\vec{t}) = \sum_{k=1}^{N} \lambda_k \, g\left(\frac{\vec{t} - \vec{u_k}}{s_k}\right) \quad \vec{u_k} \in \mathbb{R}^I \;\; s_k \in \mathbb{R}, \tag{6}$$

where $g : \mathbb{R}^I \to \mathbb{R}$ is the radial basis function (RBF). The more frequent RBFs are radially symmetric (only depend on $\left\|\vec{t} - \vec{u_k}\right\|$), such as the gaussian function $g(\vec{x}) = e^{-\|\vec{x}\|^2}$. In [Park & Sandberg 1993] it is proved, among other things, that RBFNs can approximate any function $f : \mathbb{R}^I \to \mathbb{R}^O$ in $L^2(\mathbb{R}^I)$ with a function g integrable in $\mathbb{R}^I$ satisfying

(a) $\int_{\mathbb{R}^I} g(\vec{x})d\vec{x} \neq 0$.

(b) $\int_{\mathbb{R}^I} |g(\vec{x})|^2 d\vec{x} < \infty$.

In the same paper it is proved that a necessary and sufficient condition for the RBF is to be square integrable and pointable. In particular, there may be functions with $\int_{\mathbb{R}^I} g(\vec{x})d\vec{x} = 0$. Observe that it is not necessary for the function to be radially symmetric. Consequently, the $SAOC$ with functions g satisfying the previous conditions ($\Omega \subseteq \mathbb{R}^I \times \mathbb{R}$) can be applied to approximate any function in $L^2(\mathbb{R}^I)$. As for wavelets, to guarantee the existence it is enough for s_k to belong to a compact in $\mathbb{R}$ not containing 0 (if $|\vec{u_k}|$ tends to infinite, $|\langle f - X_{N-1}, v_\omega \rangle|$ tends to 0, whereas the norm of v_ω, for s_k fixed, is constant).

6 Related work

6.1 Projection Pursuit

Projection Pursuit (PP) is a family of optimization methods appeared in the statistics literature. Its name is derived from the fact that the data are projected onto several well-chosen directions, which are selected to maximize a certain objective function. In general, given a random variable X, the methods based in PP search for a linear projection A optimizing an objective function $Q(F_A)$, where F_A is the distribution of the random variable $A \cdot X$ [Huber 1985]. By changing the objective function $Q(F_A)$, the particular PP methods are obtained. PP is able of bypassing the curse of dimensionality caused by the fact that a high dimensional space is mostly empty, and it is also able to ignore irrelevant variables (noisy and information-poor variables). In addition, PP generalizes classical methods in multivariate analysis, such as principal components and discriminant analysis, and in factor analysis (the quartimax and oblimax methods). As a drawback, they use to be high-demanding on computation time. Due to this computational cost, and to the interest in getting an ordered set of projections, the stepwise methods are very attractive. As a particular case of function approximation, and historically the first one, Projection Pursuit Regression (PPR) [Friedman & Stuetzle 1981] estimates the conditional expectation of a random variable $Y \in \mathbb{R}$ given $X \in \mathbb{R}^I$ by means of a sum of ridge functions

$$E[Y \mid X = \vec{x}] = f(\vec{x}) \cong \sum_{j=1}^{N} g_j(\vec{a_j}^t \cdot \vec{x})$$

as follows (the $\vec{a_j}$'s act as the frequencies). Suppose that the first $n-1$ terms of the approximation have been determined. That is, the vectors $\vec{a_j}$ and the functions g_j, $1 \leqslant j \leqslant n-1$ have been calculated. Let

$$r_{n-1}(\vec{x}) = f(\vec{x}) - f_{n-1}(\vec{x}) = f(\vec{x}) - \sum_{j=1}^{n-1} g_j(\vec{a_j}^t \cdot \vec{x})$$

be the residue at step $n-1$. Find $\vec{a_n}$ and g_n such that $\left\| r_{n-1}(\vec{x}) - g_n(\vec{a_n}^t \cdot \vec{x}) \right\|$ is the minimum. This process is repeated until the residue is smaller than a user-defined threshold. In the original definition, the approximation is defined from a set of observations [Friedman & Stuetzle 1981]. For a given $\vec{a_n}$, the function g_n is constructed from the scatterplot of r_{n-1} against $\vec{a_n}^t \cdot X$, so that g_n is smooth and fits the scatterplot. In the abstract version, the function itself is available instead of just a set of observations [Huber 1985]. In this case it is possible to prove that, for a fixed $\vec{a_n}$, the function g_n minimizing $\left\| r_{n-1}(\vec{x}) - g_n(\vec{a_n}^t \cdot \vec{x}) \right\|$ is $g_n(z) = E[r_{n-1}(X) \mid \vec{a_n}^t \cdot X = z]$. In order to be well-defined, $f \in L^2$ and the integral is defined with regard to a probability measure. The problem of finding $\vec{a_n}$ is much more difficult, and many times there is no guarantee that the minimum is global, regardless of its existence. The process may be improved by backfitting: omit some of the earlier summands g_j, determine its best possible replacement, and then iterate. Usually, the directions $\vec{a_j}$ are kept fixed. In [Huber 1985] it is conjectured that $\lim_{n\to\infty} E[r_n] = 0$ under mild smoothness conditions. In [Jones 1987] those conditions are pointed out: for a fixed ρ, $0 < \rho < 1$, $\vec{a_n}$ must be such that

$$E[g_n^2(\vec{a_n}^t \cdot X)] > \rho \sup_{b^t \cdot b = 1} E[g_n^2(\vec{b}^t \cdot X)].$$

Later, [Jones 1992] proved that the convergence may be accelerated approximating by an optimal convex combination. Given a function set P_n, find $0 \leqslant \alpha_n \leqslant 1$, $\vec{a_n}$ and $g_n \in P_n$ so that $\left\| f - ((1-\alpha_n) f_{n-1} + \alpha_n g_n(\vec{a_n}^t \cdot \vec{x})) \right\|$ is the minimum. Defining *relaxed* PPR as

$$f_n(\vec{x}) = (1-\alpha_n) f_{n-1}(\vec{x}) + \alpha_n g_n(\vec{a_n}^t \cdot \vec{x}),$$

its approximation error is $O(1/\sqrt{n})$.

In philosophic terms, the *SAOC* could be understood as a function approximation method based on PP and very similar to PPR. However, there are at least two important differences with respect to PPR:

1. The functions g_j, while calculated in PPR, are fixed in the *SAOC*. Although the calculation of g_j leads to a more flexible model, there can

be considerable technical difficulties. In particular, the choice of the bandwidth of the smoother used to find g_j is very critical [Huber 1985].

2. When calculating the new term in PPR, the coefficients of the previous terms are kept constant (in PPR, the coefficients are incorporated in the functions g_j). The possibility that a modification of the coefficients can lead to a better reduction of the total error is not foreseen in PPR (see Figure 1). This is a consequence of trying to approximate the residue at the previous step with only one term, regardeless of its interaction with the other terms (observe that backfitting is not enough). In *relaxed* PPR, the optimization is carried out on the segments connecting f_{n-1} and g_n, whereas in *SAOC* the optimization is performed on the hyperplane that generate $g_1, \cdots, g_{n-1}, g_n$.

The idea of Projection Pursuit has been applied in different areas.

6.1.1 Projection Pursuit in Neural Networks

The two layer architecture of a neural network is well suited to implement PPR [Hwang et al. 1992], [Hwang et al. 1994]. The Projection Pursuit Learning Network (PPLN) is modeled as a two-layer (one-hidden layer) feedforward neural network

$$\hat{y}_i = \sum_{k=1}^{m} \beta_{ik} f_k(\sum_{j=1}^{p} \alpha_{kj} x_j),$$

where $\{\beta_{ik} : i = 1, \cdots, q\}$ are the output-layer weights connecting the kth hidden neuron to all the output units, f_k is the unknown (trainable) "smooth" activation function of the kth hidden neuron, and $\{\alpha_{kj} : j = 1, \cdots, p\}$ denote the hidden-layer weights connecting all the input units to the kth hidden neuron. The training of all the parameters is based on the criterion of minimizing the error function $L_2 = \sum_{i=1}^{q} W_i E[(y_i - \hat{y}_i)^2]$. A PPLN learns neuron by neuron, and layer by layer cyclically after all the training patterns are presented. All the parameters to be estimated are hierarchically divided into m groups (each associated with one hidden neuron), and each group, say the kth group, is further divided into three subgroups: the output-layer weights $\{\beta_{ik} : i = 1, \cdots, q\}$, the smooth nonparametric function f_k of the kth hidden neuron, and the input-layer weights $\{\alpha_{kj} : j = 1, \cdots, p\}$ connected to the kth hidden neuron. The PPL starts from updating the parameters associated with the first hidden neuron (group) by updating each subgroup, $\{\beta_{i1}\}$, f_1 and $\{\alpha_{1j}\}$ consecutively (layer by layer) to minimize the error function L_2. It then updates the parameters associated with the second hidden

neuron by consecutively updating $\{\beta_{i2}\}$, f_2 and $\{\alpha_{2j}\}$. A complete updating pass ends at the updating of the parameters associated with the mth (the last) hidden neuron. Repeated updating passes are made over all the groups until convergence (*backfitting*). The kth group parameters are estimated as follows. Least Squares (LS) is applied to estimate $\{\beta_{ik}\}$, (given f_k and $\{\alpha_{kj} : j = 1, \cdots, p\}$, L_2 is quadratic in the $\{\beta_{ik}\}$), f_k is estimated by a one-dimensional data smoother and a nonlinear optimization algorithm (Gauss-Newton originally) estimates $\{\alpha_{kj}\}$.

Note that in order to compute $\{\beta_{ik}\}$, the residue resulting of consider all the units except the kth unit is minimized. Again, this is the idea of keeping the coefficients of the previous terms fixed.

In comparative studies between PPLN and Backpropagation Learning Networks (BPLN), both have quite comparable training speed and achieve comparable accuracy for independent test data, but PPLN are considerably more parsimonious in that fewer units are required to approximate the desired function [Hwang et al. 1994]. In addition, in BPLN the weights (directions) in the first layer may be very different between different simulations with different number of hidden units, whereas PPLN is more consistent in that sense.

6.1.2 Projection Pursuit in Signal Processing

Some methods with the same underlying ideas than PP have appeared in the area of Signal Processing. In [Mallat & Zhang 1993] Matching Pursuit (MP) is described, an algorithm that decomposes any signal into a linear expansion of waveforms that are selected from a redundant dictionary of functions. The theoretical results in [Mallat & Zhang 1993] are general in the sense that can be applied to any vector f in any Hilbert space H. They define a dictionary as a family $D = (g_\gamma)_{\gamma \in \Gamma}$ of vectors in H, such that $\|g_\gamma\| = 1$ for all $\gamma \in \Gamma$. The MP works roughly as follows. Let $R^0 f = f$, and suppose that the nth order residue $R^n f$ is computed, for $n \geqslant 0$. Choose an element $g_{\gamma_n} \in D$ which closely matches the residue $R^n f$, that is

$$|\langle R^n f, g_{\gamma_n} \rangle| \geqslant \alpha \sup_{\gamma \in \Gamma} |\langle R^n f, g_\gamma \rangle| ,$$

where α is an optimality factor that satisfies $0 < \alpha \leqslant 1$. The residue $R^n f$ is subdecomposed into

$$R^n f = \langle R^n f, g_{\gamma_n} \rangle \, g_{\gamma_n} + R^{n+1} f$$

which defines the residue at the order $n+1$, so that

$$f = \sum_{i=0}^{n} \left\langle R^i f, g_{\gamma_i} \right\rangle g_{\gamma_i} + R^{n+1} f.$$

In [Mallat & Zhang 1993] it is proved that, if the dictionary D is closed in H, then $f = \sum_{i=0}^{\infty} \langle R^i f, g_{\gamma_i} \rangle g_{\gamma_i}$. A final recalculation of the coefficients is made, called back-projection, to approximate f at best with the finally selected vectors.

Although MP was developed independently from PP and in a very different context, the underlying ideas are similar. We can observe that, again, the coefficients of the previous terms remain fixed when calculating the new term.

6.2 Cascade Correlation

In the field of Neural Networks, the most used constructive method is Cascade-Correlation (CC) [Fahlman & Lebiere 1990]. Constructive (or additive) methods start out from a small network and then insert additional units and connections (weights) until the network can represent the required function. CC combines two key ideas. The former is the cascade architecture, in which hidden units are added one at a time. The newly added hidden neuron receives inputs from the input layer as well as from the previously added hidden neurons. The latter is the learning algorithm. For each new hidden unit, the algorithm tries to maximize the magnitude of the correlation between the new unit's output and the residual error signal of the network. The hidden unit's input weights are frozen at the time the unit is added to the network. Only the output connections are trained repeatedly. The learning algorithm works as follows. It begins with a network without hidden units, and the direct input-output connections are trained as well as possible over the training set. To create a new hidden unit, it begins with a candidate unit that receives trainable input connections from all of the network's inputs and from all the pre-existing hidden units. The output of this candidate is not yet connected to the active network. The candidate unit's input weights are adjusted to maximize the correlation (or, more precisely, the covariance)

$$S = \sum_{o} \left| \sum_{p} (V_p - \overline{V})(E_{p,o} - \overline{E_o}) \right|$$

where o are all the output units, p are the training patterns, V_p is the activation of the candidate, and $E_{p,o}$ is the residual error observed at unit o

(without the candidate unit, which is not yet connected). The quantities $\overline{V}$ and $\overline{E_o}$ are the values of V_p and $E_{p,o}$ averaged over all patterns. In order to maximize S, a gradient ascent is performed. Once again, only a single layer of weights is trained. When S stops improving, the candidate is installed as a hidden unit. Its input weights are frozen, and all the output layer connections are trained as well as possible over the training set. The cycle continues until the network's performance is satisfactory. Instead of a single candidate unit, it is possible to use a pool of candidate units, each with a different set of random initial weights. Alternatively, the candidates might have different nonlinear activation functions, and let them compete to be chosen for addition to the active network.

At the original definition [Fahlman & Lebiere 1990], CC is only designed to approximate datasets, and there is no result of convergence. Even so, there is nothing that prevents to think at an abstract version, where a function is approximated instead of a dataset, and where the numerical optimizations are the zeros of the derivative of the objective function. Some previous studies have shown that the idea of maximizing the correlation tends to produce saturate units [Hwang et al. 1996]. Moreover, the decision boundary may be very zigzag and unsmooth. This makes the CC method more suitable, in principle, for classification problem than for regression problems.

CC has two main problems [Prechelt 1997]:

1. In principle, the covariance is an ill-suited objective function for training the candidates. Maximizing covariance trains candidates to have a large activation (large deviation from average activation) whenever the error at their output is not equal to the average error.

2. Cascading the hidden units results in a network that can represent very strong nonlinearities. Although this power is in principle useful, it can be a disadvantage if such strong nonlinearity is not required to solve the problem.

To remedy the first problem, one can change the learning rule and train directly for minimization of the outputs errors instead of for maximization of covariance. Virtual output connections must be created for the candidate units. These connections do not propagate an activation to the output units, but they receive an error signal during the backward pass. This signal is corrected by the *would-be* contribution of the candidate unit and is then handled like in normal backpropagation. Only the candidate units are being trained while the rest of the network is fixed (see [Prechelt 1997] for details). For the

second problem, not cascading hidden units is better than cascading them for some problems and worse for others. In [Prechelt 1997], the former case occurs more often.

The differences between the original CC and the $SAOC$ are quite evident. Neither the architecture nor the optimization function allow to establish a direct relation. However, and opposite to PPR, the idea of optimizing the coefficients of the other units, once the new hidden unit has been added, appears in CC. By minimizing the error instead of maximizing the correlation and removing the cascaded connections ("Cand" in [Prechelt 1997], which is neither cascaded nor deals with the correlation), the differences with the $SAOC$ are reduced. Even so, the frequencies obtained by these methods are the result of

1. Approximating the residue at the previous step by only one term (only one frequency), exactly the same as in PPR.

2. Finding the optimal coefficients of the terms associated with the existing frequencies.

But the frequencies such that their associated vectors minimize the residue are not always the best, even though the coefficients are optimized (see Figure 2). The $SAOC$ method can find frequencies much more suited to approximate the function than those minimizing the residue. In fact, it would be possible to construct a $SAOC$ with the frequencies found by the "Cand" method. So, in this sense, we may guarantee the convergence of "Cand" (abstract version) to any function in L^2, if the hypotheses of Theorem 2 are satisfied.

6.3 Projection Pursuit and Cascade Correlation

Some hybrid models have appeared in the NN literature that attempt to profit from the advantages of PP and CC. In [You et al. 1994] Cascaded Projection Pursuit Network (CPPN) is defined, which implements PPLN with cascaded connections among the hidden units, to allow high-order nonlinearities. With the aim of optimizing the nonlinearity degree and relaxing the necessity of predefining it, the Pooling Projection Pursuit Networks [Lay et al. 1994] use a pool of Hermite polynomials of several degrees during the training of a new candidate hidden unit.

6.4 Other methods in Neural Networks

In [Zhang & Morris 1998] a sequential orthogonal approach to the building and training of single hidden layer neural networks is described. In the pro-

posed method, hidden neurons are added one at a time. The procedure starts with a single hidden neuron and sequentially increases the number of hidden neurons until the model error is sufficiently small. When adding a neuron, the new information introduced by this neuron is caused by that part of its output vector which is orthogonal to space spanned by the output vectors of previously added hidden neurons. In this context, a vector is an element of $\mathbb{R}^T$, where T is the number of patterns. The classical Gram-Schmidt orthogonalization method is used at each step to form a set of orthogonal bases for the space spanned by the output vectors of the hidden neurons. Hidden layer weights c_n are found through optimization (gradient descent) of $\|E_{n-1} - \omega_n R_n(c_n)\|$, where E_{n-1} is the network error with the previously added hidden neurons. Output layer weights ω_n are obtained from the LS regression. When the training procedure is terminated at the nth step, output layer weights need to be recalculated in order to accommodate the effects of the non-orthogonal part of the output vectors of the hidden neurons. Output layer weights are determined through orthogonal LS using the Gram-Schmidt orthogonalization results obtained at each step.
Although the coefficients are recalculated when the training finishes, the frequencies are obtained again approximating the residue at the previous step with only one term (one frequency), exactly the same as in PPR.

7 SAOC's practical properties

The *SAOC* satisfies a number of interesting properties to implement it in an efficient and reasonably simple fashion.

First, observe that the problem of finding the frequencies and the coefficients of X_N is reduced to the only problem of finding the frequencies: once the frequencies are selected, the optimal coefficients can be obtained solving the linear equations system proposed at (3). From a geometrical point of view, it is equivalent to say that the main problem is to find good approximation directions. By the definition of X_N, at every step it is only necessary to find a new frequency (a new direction), since the frequencies found at previous steps remain fixed. This fact reduces considerably the difficulty of solving the optimization problem proposed: For a fixed N, a method trying to find the frequencies and the coefficients at the same time should search for $2N$ parameters. The N coefficients can be obtained from the N frequencies. With the *SAOC*, $N + 1$ parameters must be found at every step, but the N coefficients are a function of the selected frequency. The difficulty of finding the coefficients is almost null. In contrast, the difficulty of finding N

frequencies at a time seems to be very superior to that of finding N times one frequency ([Huber 1985], [Jones 1992]). Not to forget that, in general, it is impossible to know *a priori* the number of necessary terms N to achieve a satisfactory approximation.

Second, if we only have a dataset, it would be convenient to have an efficient mechanism to compute the required operations. Suppose we are at step N. We have $\omega_1, \omega_2, \cdots, \omega_{N-1} \in \Omega$, and we want to find the new frequency ω_N. Suppose that we have two candidates ω_N^1 and ω_N^2. How can we decide which is the best one? In a first approximation, we could obtain the coefficients, solving (3), and compute $\|f - X_N\|^2$ for both frequencies. The best minimizes the error. This strategy forces to make two passes through the dataset: the first one to propose the linear equations system, and the second one to compute the residue. The first one seems unavoidable, because we need to know the projections of the vector associated with the new frequency onto the vector f and the vectors associated with the previous frequencies (the projections involving the previous frequencies keep constant, and we do not need to recalculate them). In contrast, the second pass is avoidable. By (L1c), we have $\|f - X_N\|^2 = \|f\|^2 - \|X_N\|^2$. The frequency that minimizes the error is such that maximizes $\|X_N\|^2$. By (L1d), we know that

$$\|X_N\|^2 = \sum_{k=1}^{N} \lambda_k^N \overline{\langle f, v_{\omega_k} \rangle}. \tag{7}$$

Hence, to compute $\|X_N\|^2$ it is not necessary to make a new pass through the dataset. The values of $\{\langle f, v_{\omega_k} \rangle\}_{1 \leqslant k \leqslant N}$ are the independent vector of the linear equations system (3) just solved to obtain $\{\lambda_k^N\}_{1 \leqslant k \leqslant N}$. Thus, we will select the frequency that maximizes $\sum_{k=1}^{N} \lambda_k^N \overline{\langle f, v_{\omega_k} \rangle}$. The cost of computing this value is $O(N)$. Note that the cost of computing $\|X_N\|^2$ or $\|f - X_N\|^2$ directly from the dataset is $O(T \cdot N)$, where T is the number of elements in the dataset.

Finally, we may wonder whether the vector norm affects or not in the approximation. It could happen systematically that vectors with large norms were better to approximate than vectors with small norms, or vice versa. If the norm of the vector v_{ω_0} depended on the frequency ω_0, we would have the undesirable property that there are priviligiated frequencies only by the norm of its associated vector and independently of the target vector. But, fortunately, this is not the case. Suppose we are at the step N, and we have just selected ω_N and calculated $\lambda_1^N, \cdots, \lambda_{N-1}^N, \lambda_N^N$. Suppose that we modify the norm of the vector v_{ω_N} by defining $v'_{\omega_N} = h \cdot v_{\omega_N} \quad h \in \mathbb{R}, h \neq 0$

($\left\|v'_{\omega_N}\right\| = |h| \cdot \|v_{\omega_N}\|$). Proposing again the linear equations system (3), we must find $\lambda_1'^N, \cdots, \lambda_{N-1}'^N, \lambda_N'^N$ such that

$$\begin{pmatrix} \langle v_{\omega_1}, v_{\omega_1} \rangle & \langle v_{\omega_2}, v_{\omega_1} \rangle & \cdots & \langle v'_{\omega_N}, v_{\omega_1} \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle v_{\omega_1}, v_{\omega_{N-1}} \rangle & \langle v_{\omega_2}, v_{\omega_{N-1}} \rangle & \cdots & \langle v'_{\omega_N}, v_{\omega_{N-1}} \rangle \\ \langle v_{\omega_1}, v'_{\omega_N} \rangle & \langle v_{\omega_2}, v'_{\omega_N} \rangle & \cdots & \langle v'_{\omega_N}, v'_{\omega_N} \rangle \end{pmatrix} \begin{pmatrix} \lambda_1'^N \\ \vdots \\ \lambda_{N-1}'^N \\ \lambda_N'^N \end{pmatrix} = \begin{pmatrix} \langle f, v_{\omega_1} \rangle \\ \vdots \\ \langle f, v_{\omega_{N-1}} \rangle \\ \langle f, v'_{\omega_N} \rangle \end{pmatrix}.$$

This system solution is $(\lambda_1'^N, \cdots, \lambda_{N-1}'^N, \lambda_N'^N) = (\lambda_1^N, \cdots, \lambda_{N-1}^N, \lambda_N^N/h)$. Computing the norm of the new $X'_N = \sum_{k=1}^{N-1} \lambda_k^N v_{\omega_k} + \lambda_N'^N v'_{\omega_N}$ using (7) we have

$$\begin{aligned} \|X'_N\|^2 &= \sum_{k=1}^{N-1} \lambda_k^N \overline{\langle f, v_{\omega_k} \rangle} + \lambda_N'^N \overline{\langle f, v'_{\omega_N} \rangle} \\ &= \sum_{k=1}^{N-1} \lambda_k^N \overline{\langle f, v_{\omega_k} \rangle} + \frac{\lambda_N^N}{h} \overline{\langle f, h \cdot v_{\omega_N} \rangle} = \|X_N\|^2. \end{aligned}$$

Therefore, $\|f - X_N\|^2 = \|f - X'_N\|^2$. That means that the goodness of the new frequency ω_0 with regard to its approximation capability does not depend on the norm of the vector v_{ω_0}. In particular, we could define $v'_{\omega_N} = v_{\omega_N} / \|v_{\omega_N}\|$, so that their norm would always be 1.

8 SAOC and Neural Networks

From now on, we will focus on Artificial Intelligence, and more precisely in the field of feed-forward neural networks.

Let K be a compact in $\mathbb{R}^I$. We want to approximate a function $f : K \to \mathbb{R}^O$ in $H = L^2(K)$ by functions f_N as defined in (5), where the activation functions are nonlinear (otherwise, the whole problem would be resolved by solving a linear equations system). We only have the value of the function in a dataset, and the main objective is to achieve a successful interpolation or generalization. The dimension of the input space may be very large (of the order of hundreds) depending on the problem at hand. Although $\Omega = \mathbb{R}^{I+1}$, the number of weights needed to approximate the dataset may be much larger $(I \cdot N)$. Since we only have a dataset, the restriction of K being compact is not a real restriction. It is very easy to verify that, if we only have a finite dataset, the problem of approximating a function in L^2 is equivalent to that of approximating (by Least Squares) a vector $\vec{t} \in \mathbb{C}^T$, where T is the number

of elements in the dataset. The vector component t_i is associated with the element i in the dataset. Every frequency is associated with a specific vector in $\mathbb{C}^T$, which components depend on the frequency and the point in the dataset.

In practice, the *SAOC* method presents a problem. To find a valid frequency, we must verify that $\|f - X_N\|^2 \leqslant \inf_{\mu\in\mathbb{C}\ \omega\in\Omega} \|f - (X_{N-1} + \mu v_\omega)\|^2 = \inf_{\omega\in\Omega} P_N(v_\omega)$. Usually, we will be able to compute $\inf_{\omega\in\Omega} P_N(v_\omega)$ if we already know the frequency ω_0 that minimizes P_N. But this minimum must be global in the space of parameters. Global optimization techniques are very expensive computationally. In a high-dimensional space without any kind of convexity, it becames an almost intractable problem [Horst & Tuy 1993]. Anyway, since we only have a dataset, and we seek a good generalization performance, it may happen that approximating too much the points in the dataset leads to a performance degradation. This can happen, for example, if the data is noisy, or few data are available with respect to the input dimension (curse of dimensionality) or if the model is too complex [Bishop 1995]. Many times, finding a local minumum is enough to achieve a good performance.

Neural Networks are a suitable approach to deal with function approximation problems when only a dataset is available. On the one hand, they allow the approximation of the value of the function at the points in the dataset. On the other, they offer a number of techniques to deal with the generalization problem. In addition, neural networks work in high-dimensional spaces with nonlinear approximations without (theoretically) problems derived from the dimension, since the error only depends on the complexity of the function and the number of units in the hidden layer [Barron 1993]. The existing theory can guarantee that the global minimum exists, but it cannot be found (in general) with a reasonable computational cost. The methods are either non-constructive or computationally prohibitive in the general case [Scarselli & Tsoi 1998]. Most of the times, a local minimum is found. An example is offered by the celebrated Kolmogorov's theorem [Kolmogorov 1957], which states that any continuous function of several variables on any compact can be represented as a superposition of one variable functions. Although the proof is constructive, its practical applications are limited ([Girosi & Poggio 1989], [Kůrková 1991], [Kůrková 1992]). Another similar situation is the use of heuristics (both in Neural Networks and in other Artificial Intelligence areas), to solve problems for which we do not even have an existence theorem. As has been said before, this is not a major problem if we are dealing with a dataset and our main aim is the generalization.

Hence, there are reasons to think that the $SAOC$ method can be implemented in a neural network. In general, not every theoretical result can be put into practice in a reasonable manner [Scarselli & Tsoi 1998]. In fact, the profit is reciprocal. The $SAOC$ can serve as an inspiration to train a neural network: adding hidden units one at a time, choosing the initial weights in an optimal manner, so as to train the network until we have a satisfactory model (sequential training). This idea, in addition, offers a number of advantages to train the network. First, it allows to estimate in a reliable way the optimal number of hidden units that is necessary to approximate the function (that is, to be parsimonious). This is one of the basic problems met in constructing a neural network. If there are few hidden units, the dataset will not be approximated properly. If there are too many, the generalization may not be satisfactory, since the model may be too complex. In addition, the most promising activation function can be chosen at every step, so that the network adapts its architecture to the specific target function. Other advantage is related to the training monitoring. With a sequential training, it is possible to display some training aspects that one could consider outstanding (error decreasing, performance, weights, etc), together with the evolution of such aspects. It is possible, for example, to save the parameters of the intermediate steps of the training and recover them if desired.
Concerning the neural network architecture needed to implement the $SAOC$ method, it must present the following characteristics:

- It must be a feedforward architecture with a hidden layer of units (including both two-layer perceptrons and RBFNs).
- There are no restrictions about the dimension of the input and the output. There will be so many as the target function have. If there are several outputs, the total inner products must be calculated as the summation of the individual inner products of every output.
- There is no restriction about the biases in the hidden units. The biases can be treated as part of the frequencies.
- The output units cannot have biases.
- There is no restriction about the activation functions in the hidden units, provided that the hypotheses in Theorem 2 are satisfed. In particular, they can be sines, cosines, sigmoidal functions, gaussian functions, wavelets, etc. Obviously, different units may have different activation funcitons.

- The output units must have a linear activation function.

As we can see, the restrictions only refer to the output units. The biases are not a real problem, since they can be considered as frequencies with a simple transformation. Another solution consists of adding a new hidden unit with constant activation function. Hence, the only real restriction in the output units is the linear activation function.

8.1 Algorithm

An algorithm to implement the $SAOC$ method using a neural network, which is based on the previously discussed ideas, could be the following:

Algorithm
  $N := 0$;
  while the network is not valid
    Increase by 1 the number of hidden units N
    $Outputmax := 0$;
    for $t := 1$ **upto** $Nattempts$ **do**
      Assign randomly a candidate frequency $\omega_{(t)}$ (weights in the first layer) to the last active hidden unit
      Pick an activation function for the new hidden unit
      Compute the coefficients $\{\mu_k(t)\}_{1 \leqslant k \leqslant N}$ (weights in the second layer), by solving the linear equations system (3)
      Compute the $Output$ $\|X_N(t)\|^2 = \sum_{k=1}^{N-1} \mu_k(t)\overline{\langle f, v_{\omega_k}\rangle} + \mu_N(t)\overline{\left\langle f, v_{\omega_{(t)}}\right\rangle}$.
      if $Output > Outputmax$ **then**
        $Outputmax := Output$;
        $\omega_N := \omega_{(t)}$;
      end if
    end for
    Optionally, train the network, to tune the frequency ω_N
    Fix the frequency ω_N in the network, so that it cannot be modified at later trainings
    Validate the network
  end while
end Algorithm

Since $\|f - X_N\|^2 = \|f\|^2 - \|X_N\|^2$, to find out the minimum error $\|f - X_N\|^2$ we only need to calculate the maximum output $\|X_N\|^2$. The random selection of weights for the candidate units is not the best possible strategy for sure, but it can be justified if we consider the selected frequencies as the initial

weights to start the network training (usually, the initial weights in a neural networks are selected randomly within a particular finite interval). Since the frequency goodness does not depend on the norm of its associated vector, the range of weights to look for candidate frequencies may be as large as desired. However, if the range is too large, the number of attempts needed to find an acceptable candidate frequency may also be very large. In practice, there will be a trade-off between these two matters.

9 Future work

There are a number of questions that can be posed after this work:

- According to *SAOC*'s definition, if we want to know whether a frequency is valid or not (i.e. whether it satisfies condition (b) in the definition), we need to know the best approximation of the residue with only one frequency. As already mentioned, this problem may be very difficult to solve. A way of relaxing this condition would be allowing an optimality factor (as in [Jones 1987] or [Mallat & Zhang 1993]). Probably, the convergence property expressed in Theorem 2 would be maintained, and the construction would be easier. Anyway, it is not clear how to profit this definition in practice, since to be able to fix the optimality factor, one will probably need to know the best approximation of the residue.

- A very interesting problem that appears in a sequential approximation deals with studying the error bound as a function of the number of the approximation terms (see [Barron 1993], [Jones 1992] or [Mallat 1998]). If this function were known or estimated, it could provide an estimate of the minimum number of terms needed to construct the approximation, although this would surely depend on the complexity of the target function.

- In the presented particular implementation with neural networks, there are also a lot of matters to study:

  - To choose the candidate frequencies with different heuristics from current random selection. In principle, a more intelligent selection could lead to better approximations. Another strategy could be based on genetic algorithms. In the same way, the selection of the activation function for the new hidden unit admits any number of heuristics.

- Study the optimal criteria to stop adding units. In addition of validating the approximation by computing the error, it could be interesting to know, for example, if the error is being diminished very slowly. In this case we will probably have to change some of the current strategies, if we want a small number of terms. In this sense, the work of [Zhang & Morris 1998], [Hwang et al. 1994], [Mallat & Zhang 1993] or [Prechelt 1997] may be an interesting starting point.
- The approximation interpretation may be of great interest in several problems, especially if we are dealing with real world problems. The selected frequencies, together with the projections of their associated vectors onto the target function may give "understanding" information about the function behaviour [Huber 1985]. In the neural model used, it is possible to plot these projections, since the activation functions in the hidden layer are functions of one variable.

References

[Achieser 1956] Achieser, N.I. (1956). Theory of Approximation. Frederick Ungar Pub. Co., New York.

[Barron 1993] Barron, A.R. (1993). Universal Approximation Bounds for Superposition of a Sigmoidal Function. *IEEE Transactions on Information Theory* 39 (3), 930-945.

[Berberian 1961] Berberian, S.K. (1961). Introduction to Hilbert space. Oxford University Press, Inc.

[Bishop 1995] Bishop, C.M. (1995). Neural Networks for Pattern Recognition. Oxford University Press Inc., New York.

[Cybenko 1989] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems* 2, 303-314.

[Fahlman & Lebiere 1990] Fahlman, S.E. and Lebiere, C. (1990). The Cascade-Correlation Learning Architecture. In *Advances in Neural Information Prcessing Systems* 2, 524-532. Morgan Kaufmann Publishers Inc.

[Friedman & Stuetzle 1981] Friedman, J.H. and Stuetzle, W. (1981). Projection Pursuit Regression. *Journal of the American Statistical Association* 76 (376), 817-823.

[Funahashi 1989] Funahashi, K. (1989). On the Approximate Realization of Continuous Mappings by Neural Networks. *Neural Networks* 2 (3), 183-192.

[Girosi & Poggio 1989] Girosi, F. and Poggio, T. (1989). Representation Properties of Networks: Kolmogorov's Theorem is Irrelevant. *Neural Computation* 1, 465-469.

[Hornik et al. 1989] Hornik, K., Stinchcombe, M. and White, H. (1989). Multilayer Feedforward Networks are Universal Approximators. *Neural Networks* 2 (5), 359-366.

[Horst & Tuy 1993] Horst, R. and Tuy, H. (1993). Global Optimization: Deterministic Approaches. Springer-Verlag, Berlin.

[Huber 1985] Huber, P.J. (1985). Projection Pursuit. *The Annals of Statistics* 13 (2), 435-475.

[Hwang et al. 1992] Hwang, J.N., Li, H., Maechler, M., Martin, D. and Schimert, J. (1992). Projection pursuit learning networks for regression. *Engineering Applications Artificial Intelligence* 5 (3), 193-204.

[Hwang et al. 1994] Hwang, J.N., Ray, S.R., Maechler, M., Martin, D. and Schimert, J. (1994). Regression Modeling in Back-Propagation and Projection Pursuit Learning. *IEEE Tansactions on Neural Networks* 5 (3), 342-353.

[Hwang et al. 1996] Hwang, J.N., You, S.S, Lay, S.R. and Jou, I.C. (1996). The Cascade-Correlation Learning: A Projection Pursuit Learning Perspective. *IEEE Tansactions on Neural Networks* 7 (2), 278-289.

[Jones 1987] Jones, L.K. (1987). On a conjecture of Huber concerning the convergence of projection pursuit regression. *The Annals of Statistics* 15 (2), 880-882.

[Jones 1992] Jones, L.K. (1992). A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *The Annals of Statistics* 20 (1), 608-613.

[Kolmogorov 1957] Kolmogorov, A.N. (1957). On the representation of continuous functions of many variables by superpositions of continuous functions of one variable and addition. *Doklady Akademii Nauk USSR* 114(5), 953-956.

[Kolmogorov & Fomin 1975] Kolmogorov, A.N. and Fomin, S.V. (1975). Elements of the Theory of Functions and Functional Analysis. Ed. Mir, Moscow.

[Kürková 1991] Kürková, V. (1991). Kolmogorov's Theorem is Relevant. *Neural Computation* 3, 617-622.

[Kürková 1992] Kürková, V. (1992). Kolmogorov's Theorem and Multilayer Neural Networks. *Neural Networks* 5 (3), 501-506.

[Lang 1989] Lang, S. (1989). Undergraduate Analysis. Springer-Verlag, New York.

[Lay et al. 1994] Lay, S.R., Hwang, J.N., You, S.S. (1994). Extensions to projection pursuit learning networks with parametric smoothers. *Proc. Int. Conf. Neural Networks*, Orlando FL, 1325-1330.

[Leshno et al. 1993] Leshno, M., Lin, V.Y., Pinkus, A and Schocken, S. (1993). Multilayer Feedforward Networks With a Nonpolynomial Activation Function Can Approximate Any Function. *Neural Networks* 6, 861-867.

[Lorentz 1966] Lorentz, G.G. (1966). Approximation of Functions. Chelsea Pub. Co., New York.

[Mallat 1998] Mallat, S.G. (1998). A wavelet tour of signal processing. Academic Press, New York.

[Mallat & Zhang 1993] Mallat, S.G. and Zhang, Z. (1993). Matching Pursuits with Time-Frequency Dictionaries. *IEEE Transactions on Signal Processing* 41 (12), 3397-3415.

[Ortega 1972] Ortega, J.M. (1972). Numerical Analysis: A second course. Society for Industrial and Applied Mathematics, Philadelphia.

[Park & Sandberg 1993] Park, J. and Sandberg, I.W. (1993). Approximation and Radial-Basis-Function Networks. *Neural Computation* 5, 305-316.

[Prechelt 1997] Prechelt, L. (1997). Investigation of the CasCor Family of Learning Algorithms. *Neural Networks* 10 (5), 888-896.

[Reddy 1998] Reddy, B.D. (1998). Introductory Functional Analysis with Applications to Boundary Value Problems and Finite Elements. Springer-Verlag, New York.

[Rudin 1987] Rudin, W (1987). Real and Complex Analysis. McGraw-Hill, New York.

[Scarselli & Tsoi 1998] Scarselli, F. and Tsoi, A.C. (1998). Universal Approximation Using Feedforward Neural Networks: A Survey of Some Existing Methods and Some New Results. *Neural Networks* 11 (1), 15-37.

[Stein & Weiss 1971] Stein, E.M. and Weiss, G. (1971). Introduction to Fourier Analysis on Euclidean Spaces. Princeton University, New Jersey.

[Weierstrass 1885] Weierstrass, K. (1885). Über die analytische Darstellbarkeit sogennanter willkürlicher Funktionen reeller Argumente. *Sitzungsberichte der Acad. Berlin*, 633-639, 789-805.

[Yosida 1965] Yosida, K. (1965). Functional Analysis. Springer-Verlag, New York.

[You et al. 1994] You, S.S., Hwang, J.N., Jou, I.C. and Lay, S.R. (1994). A new cascaded projection pursuit network for nonlinear regression. *Proc. Int. Conf. Acoustics, Speech and Signal Processing* vol. 2, Adelaida SA, Australia, 585-588.

[Young 1980] Young, R.M. (1980). An Introduction to Nonharmonic Fourier Series. Academic Press, New York.

[Zhang & Morris 1998] Zhang, J., Morris, A.J. (1998). A Sequential Learning Approach for Single Hidden Layer Neural Networks. *Neural Networks* 11 (1), 65-80.