Greed is Fine: on Finding Sparse Zeros of Hilbert Operators

François-Xavier Dupé

Greed is Fine: on Finding Sparse Zeros of Hilbert Operators

François-Xavier Dupé FRANCOIS-XAVIER.DUPE@LIF.UNIV-MRS.FR
LIF, UMR 7279, Aix Marseille Université, CNRS, 13288 Marseille, France

Abstract

We propose an generalization of the classical Orthogonal Matching Pursuit (OMP) algorithm for finding sparse zeros of Hilbert operator. First we introduce a new condition called the *restricted diagonal deviation property* which allow us to analysis of the consistency of the estimated support and vector. Secondly when using a perturbed version of the operator, we show that a partial recovery of the support is possible and remain possible even if some of the steps of the algorithm are inexact. Finally we discuss about the links between recent works on other version of OMP.

1. Introduction

Consistency of the selection process of features is an important problem when dealing with data. In the linear setting, Zhang (2009) proposes a full analysis of the orthogonal matching pursuit and shows when the selected features are guaranteed. In this paper, we propose to cast the features selection problem in higher dimensional spaces and with non-linear Hilbert operator. We model the selection problem as an optimization problem where one seeks for a sparse zero of an operator. Such point of view generalizes many recent works on greedy methods (Zhang, 2011; Bahmani et al., 2013) and open new directions.

The orthogonal matching pursuit (OMP) (Mallat & Zhang, 1993) is now a very famous method with many extensions (Temlyakov, 2000; 2012; Livshitz & Temlyakov, 2014) to cite a few. In the linear setting, (Zhang, 2009) followed by (Swirszcz et al., 2009) shows the consistency of the algorithm using the *exact recovery condition* (ERC) (Tropp, 2004) (or an extension). Furthermore ERC is almost a necessary condition (Tropp, 2004; Gribonval & Vandergheynst, 2006) to guarantee that OMP select good elements.

Proceedings of the 31^{st} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

On the signal processing side, OMP and its extensions has been analysed in infinite dimensional spaces (Mallat & Zhang, 1993; Gribonval & Vandergheynst, 2006; Foucart, 2013) and also in the case of generalization of the Compressed Sensing to infinite sensing (Hansen & Adcock, 2011). Still, all these works formulate the problem as a linear system of equations with a sparsity constraint.

Recently some well known greedy algorithms have been generalized to function minimization: OMP (Zhang, 2011), CoSaMP (Bahmani et al., 2013; Dupé & Anthoine, 2013), Iterative Hard Thresholding (Beck & Eldar, 2013; Yuan et al., 2014; Jain et al., 2014)... Such generalizations open way to other loss functions than the classical least square. This include recent works around logistic regression (Lozano et al., 2011; Bahmani et al., 2013) or multiple kernel learning (Sindhwani & Lozano, 2011).

Such generalization can been used for M-estimators (Jain et al., 2014) and so open well-known algorithms to robust estimation. However these are recent works and many questions remains open, such as the construction of function that will fulfill the condition for either convergence and consistency. Beck & Hallak (2014) even suggest the possibility of adding other constraints in addition of the sparsity one. The scope of this paper is to prove that the Orthogonal Matching Pursuit is still a good algorithm in general setting and can consistently select features.

Main results The paper's principal contributions include a new sufficient criterion for consistency called the *restricted diagonal deviation property* (RDDP) and an analysis of the consistency of the OMP with non-linear Hilbert operator in both noisy and non-noisy case. Contrarily to Zhang (2009) or Swirszcz et al. (2009), we do not restrict ourself to finite subset of $\mathbb{R}^n$ (with $n > 0$), we even do not require the support (i.e. the set of indice of the non-null coordinate) to be finite. Two extensions are also proposed where the requirement of some parts of the algorithm are relaxed and we give the corresponding consistency analysis. Finally we propose a *realistic* version the OMP.

Motivations OMP is a well known greedy method for least square regression, but little is known when we use

it for finding sparse zeros of Hilbert operators. Such operators include the gradient of convex functions and the classical linear system of equations. The work of Zhang (2011) only considers convex functions and seeks to characterize the distance between the current estimate and the optimal minimizer through the involved function. Furthermore, as computing the gradient of a function may be expensive, we provide with the noisy operator case consistency results. This leads to new methods of feature selection using non-linear operator or M-estimator (DasGupta, 2008; Jain et al., 2014).

Outline of the paper In Section 2 we formulate the optimization problem and propose a generalization of OMP for solving it. The consistency of our version of OMP is analysed in Section 3 in both non-noisy and noisy operator cases. Then in Section 4 we propose two extensions : a weak version (like in (Temlyakov, 2000)) and an inexact step version. Finally in Section 5 we provide a discussion about the different elements of the analysis and the extensions.

Notations Let $\mathcal{H}$ be a separable Hilbert of possibly infinite dimension with a scalar product $\langle .,. \rangle$ and its associated norm $\|.\|$. Let $\{\varphi_i\}_{i\in\mathcal{I}}$ be a orthonormal basis on $\mathcal{H}$ with $\mathcal{I}$ an countable set of indices and $\mathbf{I}$ the identity operator on $\mathcal{H}$. For $x \in \mathcal{H}$, the support of x is denoted by $\mathrm{supp}(x) = \{i \in \mathcal{I} \mid \langle x, \varphi_i \rangle \neq 0\}$. Let $\mathcal{O} \subseteq \mathcal{I}$ a set of indices and $x \in \mathcal{H}$, we denote by $x_{|\mathcal{O}}$ the restriction of x to the support represented by $\mathcal{O}$. Let $\mathcal{S} \subseteq \mathcal{I}$ a support, we denote by $\mathbf{P}_{\mathcal{S}} : x \mapsto \mathrm{argmin}_{z\in\mathcal{H},\mathrm{supp}(z)\subseteq\mathcal{S}} \|z - x\|$ the orthogonal projection to the support $\mathcal{S}$.

We also need to introduce the ℓ_0 sub-norm, $\|x\|_0 = \mathrm{card}\{\mathrm{supp}(x)\}$ (i.e. the cardinality of the support), the inf-norm, $\|x\|_\infty = \sup_{i\in\mathcal{I}} |\langle x, \varphi_i \rangle|$ and the operator norm $|||\mathbf{A}||| = \sup_{x\in\mathcal{H},\|x\|=1} \|\mathbf{A}(x)\|$.

2. Non-linear orthogonal matching pursuit

Given an operator $\mathbf{T} : \mathcal{H} \to \mathcal{H}$, consider the problem of finding a sparse zero of $\mathbf{T}$,

$$\min_{x\in\mathcal{H}} \|x\|_0 \text{ s.t. } \mathbf{T}(x) = 0 \,. \tag{P}$$

Many problems can be cast as instance of (P). For example, if $\mathbf{T}$ is a affine bounded operator, it leads to the classical optimization problem that occurs in Compressed Sensing (Candès et al., 2006) or a feature selection (Zhang, 2009),

$$\min_{x\in\mathbb{R}^n} \|x\|_0 \text{ s.t. } \mathbf{A}^t(\mathbf{A}x - y) = 0 \,,$$

where $y \in \mathbb{R}^d$ and $\mathbf{A} \in \mathbb{R}^{d\times n}$ is a matrix. In a more general setting, (P) can model non-linear eigenproblems as presented by Hein & Bühler (2010), e.g. for computing sparse PCA we have,

$$\min_{x\in\mathbb{R}^n} \|x\|_0 \text{ s.t. } \mathbf{A}x - \left(\frac{x^t\mathbf{A}x}{\|x\|^2}\right) x = 0 \,,$$

with $\mathbf{A} \in \mathbb{R}^{n\times n}$ a symmetric matrix.

Solving (P) is an NP-hard problem and the potential non-linearity of $\mathbf{T}$ make it even harder. In order to find a good approximated solution, we provide a version of the OMP presented by Algo 1. The steps are simple, (i) we select the best coordinate (through a basis of $\mathcal{H}$) in magnitude given the current estimate, (ii) if the magnitude is below some tolerance we stop, (iii) we update the current support and (iv) we compute the next estimate by finding a zero for $\mathbf{T}$ on the current support. The sub-optimization problem in line 9 of the algorithm can be difficult to solve, even impossible. However for some classes of operators, for example if $\mathbf{T}$ is a firmly non-expansive operator (i.e. the gradient of a convex function), standard point-fixed algorithms are available (Cegielski, 2013).

Algorithm 1 Non-linear orthogonal matching pursuit

1: **Input:** $\mathbf{T}$ an operator, $k_{\max}$ the maximal sparsity, ε the tolerance.
2: **Initialization:** $x^0 \leftarrow 0$, $\mathcal{S}^0 \leftarrow \emptyset$.
3: **for** $k = 0$ to $k_{\max}$ **do**
4: $\quad j^{k+1} \leftarrow \mathrm{argmax}_{i\in\mathcal{I}\setminus\mathcal{S}^k} \left|\left\langle \mathbf{T}(x^k), \varphi_i \right\rangle\right|$,
5: $\quad$ **if** $(\left|\left\langle \mathbf{T}(x^k), \varphi_{j^{k+1}} \right\rangle\right| \leq \varepsilon)$ **then**
6: $\quad\quad$ break.
7: $\quad$ **end if**
8: $\quad \mathcal{S}^{k+1} \leftarrow \mathcal{S}^k \cup \{j^{k+1}\}$,
9: $\quad x^{k+1} \in \mathcal{H}$ s.t. $\begin{cases} \mathrm{supp}(x^{k+1}) \subseteq \mathcal{S}^{k+1} \,, \\ \mathbf{T}(x^{k+1})|_{\mathcal{S}^{k+1}} = 0. \end{cases}$
10: **end for**
11: **Output:** the estimate x^k with its support $\mathcal{S}^k$.

Indeed Algo 1 includes many other algorithms based on OMP. For example, if $\mathbf{T}$ is the gradient of some convex function, we retrieve the algorithm proposed by (Zhang, 2011) and if $\mathbf{T}(x) \equiv \mathbf{A}^t(\mathbf{A}x - y)$ with $y \in \mathcal{H}$ and $\mathbf{A}$ a linear bounded operator, we directly retrieve the classical orthogonal matching pursuit (Mallat & Zhang, 1993).

However, $\mathbf{T}$ may be unavailable, either its computation is too costly or it is simply unaccessible (e.g. if $\mathbf{T}$ is only accessible though a black-box), but a *perturbed* (or approximated) version $\mathbf{U}$ is usable and is related to $\mathbf{T}$ through an error term,

$$\mathbf{U} : x \mapsto \mathbf{T}(x) + e(x), \tag{1}$$

with $e : \mathcal{H} \to \mathcal{H}$ a perturbation operator which verifies $\forall x \in \mathcal{H}, \|e(x)\|_\infty \leq \rho$. Then solving (P) is irrelevant because it may have no sparse solution, we instead propose

to solve a relaxed problem,

$$\min_{x\in\mathcal{H}} \|x\|_0 \text{ s.t. } \begin{cases} \mathbf{U}(x)_{|\operatorname{supp}(x)} = 0\,, \\ \|\mathbf{U}(x)\|_\infty \leq \varepsilon\,, \end{cases} \tag{Q}$$

where ε is the deviation. Such formulation generalized the noise model presented in (Zhang, 2009; Swirszcz et al., 2009) in the classical linear setting. Then in the next section, we answer two questions, (i) when solving (P) is possible and when and (ii) how solving (Q) gives information on the optimal solution of (P)?

Existence of a solution While it is clear that a solution of (P) is also a solution of (Q), the existence of a solution for both problems for a given operator $\mathbf{T}$ is an open question. Depending on the properties of the operator, we may be able to state when a solution exist (but not necessarily with a way to find it). In the next section, we will show that the proposed property implies the existence of solutions.

Well-definiteness of Algo 1 Using the algorithm is possible only if the sub-optimization step on line 9 is solvable. This asks for the existence of a solution and a way to find it. These two points are highly dependent on the properties of the operator (see (Cegielski, 2013) for examples). We will discuss in Section 5 about some examples where we know both existence of a solution and a numerical scheme for its computation.

3. Consistency results

Algo 1 is a forward greedy algorithm, but can we guarantee that the selected elements belong to the *true* support? The following analysis use the same kind of arguments as in (Zhang, 2009; Swirszcz et al., 2009). However since we work with non-linear operators, we need a new condition for building our analysis. For this purpose we propose the *restricted diagonal deviation property* (RDDP).

3.1. Restricted diagonal deviation property

First, let us introduce $\mathcal{D}_1^\infty$ the set of diagonal operators bounded away by 1:

$$\begin{aligned}\mathcal{D}_1^\infty = \{&\mathbf{D} : \mathcal{H} \to \mathcal{H}, \mathbf{D} \text{ diagonal, and} \\ &\forall x \in \mathcal{H}, \|\mathbf{D}x\|_\infty \geq 1 \|x\|_\infty\}.\end{aligned} \tag{2}$$

Then we can define the central property for our analysis,

Definition 1 (Restricted Diagonal Deviation Property)**.** *An operator $\mathbf{T}$ is said to have the* Restricted Diagonal Deviation Property *(RDDP) on the support $\mathcal{S}$ if there exists $\alpha > 0$ such that $\forall x, y \in \mathcal{H}$,*

$$\begin{aligned}(\operatorname{supp}(x) \cup \operatorname{supp}(y)) \subseteq \mathcal{S} \Rightarrow \exists \mathbf{D}_{xy} \in \mathcal{D}_1^\infty, \\ \|\mathbf{T}(x) - \mathbf{T}(y) - \mathbf{D}_{xy}(x-y)\|_\infty \leq \alpha \|x-y\|_\infty .\end{aligned} \tag{3}$$

Notice that in finite dimensional spaces, if $\mathbf{T}$ is an affine bounded operator, then RDDP leads to the well-known restricted isometry property or RIP (using equivalence between norms). If $\mathbf{T}$ is the gradient of some function, then this criterion meets the requirement proposed by Zhang (2011) for its analysis of OMP (see Section 5 for a discussion about links with other works).

Assuming that $\mathbf{T}$ fulfills the RDDP on the support $\mathcal{O}$ with constant α, we are able to build two lemmas that highlight the behavior of $\mathbf{T}$ over $\mathcal{O}$.

This first lemma shows the relation between a restriction of $\mathbf{T}$ and restrictions of $x - y$ using elements inside and outside a given support.

Lemma 2. *Let $\mathbf{T} : \mathcal{H} \to \mathcal{H}$ be an operator and $\mathcal{O}$ the support of an optimal solution of (P). Assume that $\mathbf{T}$ fulfills the restricted diagonal deviation property on $\mathcal{O}$ with constant $\alpha > 0$. Then we have $\forall x, y \in \mathcal{H}$ such that $\operatorname{supp}(x) \cup \operatorname{supp}(y) \subseteq \mathcal{O}$,*

$$\begin{aligned}&\forall \mathcal{R} \subseteq \mathcal{O}, \quad \|(\mathbf{T}(x) - \mathbf{T}(y))_{|\mathcal{R}}\|_\infty \geq \\ &(1-\alpha)\|(x-y)_{|\mathcal{R}}\|_\infty - \alpha \left\|(x-y)_{|\mathcal{I}\setminus\mathcal{R}}\right\|_\infty .\end{aligned}$$

Proof. We have as $\mathbf{D}_{xy}$ is a diagonal operator,

$$\begin{aligned}&\left\|(\mathbf{T}(x) - \mathbf{T}(y))_{|\mathcal{R}}\right\|_\infty \\ &= \left\|(\mathbf{T}(x) - \mathbf{T}(y) - \mathbf{D}_{xy}(x-y) + \mathbf{D}_{xy}(x-y))_{|\mathcal{R}}\right\|_\infty , \\ &= \left\|(\mathbf{D}_{xy}(x-y))_{|\mathcal{R}} - (\mathbf{T}(x) - \mathbf{T}(y) - \mathbf{D}_{xy}(x-y))_{|\mathcal{R}}\right\|_\infty , \\ &\geq \left\|\mathbf{D}_{xy}((x-y)_{|\mathcal{R}})\right\|_\infty - \|\mathbf{T}(x) - \mathbf{T}(y) - \mathbf{D}_{xy}(x-y)\|_\infty , \\ &\geq \left\|(x-y)_{|\mathcal{R}}\right\|_\infty - \alpha\|x-y\|_\infty \text{ , since } \|\mathbf{D}_{xy}(z)\|_\infty \geq \|z\|_\infty \\ &\geq (1-\alpha)\left\|(x-y)_{|\mathcal{R}}\right\|_\infty - \alpha\left\|(x-y)_{|\mathcal{I}\setminus\mathcal{R}}\right\|_\infty .\end{aligned}$$

□

Notice that the last term of the inequalities gives information about the missed elements by the current support.

This next lemma gives a bound on the energy of $\mathbf{T}$ restricted to elements outside the supports of x and y.

Lemma 3. *Let $\mathbf{T} : \mathcal{H} \to \mathcal{H}$ be an operator and $\mathcal{O}$ the support of an optimal solution of (P). Assume that $\mathbf{T}$ fulfills the restricted diagonal deviation property on $\mathcal{O}$ with constant $\alpha > 0$. Then $\forall x, y \in \mathcal{H}$ such that $\operatorname{supp}(x) \cup \operatorname{supp}(y) \subseteq \mathcal{O}$, we have,*

$$\begin{aligned}&\forall \mathcal{F} \subseteq \mathcal{I} \text{ s.t. } \mathcal{F} \cap (\operatorname{supp}(x) \cup \operatorname{supp}(y)) = \emptyset, \\ &\left\|(\mathbf{T}(x) - \mathbf{T}(y))_{|\mathcal{F}}\right\|_\infty \leq \alpha\|x-y\|_\infty\end{aligned}$$

Proof. Since $\mathcal{F} \cap (\text{supp}(x) \cup \text{supp}(y)) = \emptyset$ and $\mathbf{D}_{xy}(x-y)_{|\mathcal{F}} = 0$, we have

$$\left\|(\mathbf{T}(x) - \mathbf{T}(y))_{|\mathcal{F}}\right\|_\infty = \left\|(\mathbf{T}(x) - \mathbf{T}(y) - \mathbf{D}_{xy}(x-y))_{|\mathcal{F}}\right\|_\infty \leq \alpha \|x-y\|_\infty \ .$$

□

Using these lemmas we are able to prove the existence and unicity of a solution of (Q) when the operator fulfills the RDDP.

Proposition 4. *Assume that* $\mathbf{T}$ *fulfills the RDDP on support* $\mathcal{O} \subseteq \mathcal{I}$ *with constant* $\alpha < 1$*, then* (Q) *has a unique solution.*

Proof. **Existence** From the RDDP, we have, $\forall x, y \in \mathcal{H}$ such that $(\text{supp}(x) \cup \text{supp}(y)) \subseteq \mathcal{O}$,

$$\begin{aligned} \alpha \|x-y\|_\infty &\geq \|\mathbf{T}(x) - \mathbf{T}(y) - \mathbf{D}_{xy}(x-y)\|_\infty \ , \\ &\geq \|(\mathbf{D}_{xy} - \mathbf{T})(x) - (\mathbf{D}_{xy} - \mathbf{T})(y)\|_\infty \ . \end{aligned} \quad (4)$$

Equation (4) suggests that there exists an operator $\mathbf{R} = \hat{\mathbf{D}} \circ \mathbf{T}$ (with $\hat{\mathbf{D}}$ a diagonal operator) such that,

$$\|(\mathbf{I} - \mathbf{R})(x) - (\mathbf{I} - \mathbf{R})(y)\|_\infty \leq \alpha \|x-y\|_\infty \ ,$$

thus if $\alpha < 1$ the operator $\mathbf{I} - \mathbf{R}$ is a Banach contraction. Then $\forall \mathcal{S} \subseteq \mathcal{O}$, $\mathbf{P}_\mathcal{S}(\mathbf{I} - \mathbf{R})$ is also a Banach contraction ($\mathbf{P}_\mathcal{S}$ is a orthogonal projection), thus we are able to build a iterative scheme that converges to a point fixed of $\mathbf{P}_\mathcal{S}(\mathbf{I} - \mathbf{R})$. Such a scheme implies the existence of $x \in \mathcal{H}$ such that $\text{supp}(x) \subseteq \mathcal{S}$ and $\mathbf{T}(x)_{|\mathcal{S}} = 0$.

Unicity The unicity is directly given by Lemma 2. □

With the two lemmas and the proposition we are now able to analyse the OMP in both non-noisy and noisy settings.

3.2. Non-noisy case

We first focus on the non-noisy case. From definition 1 we are able to build the following theorem.

Theorem 5. *Let* $\mathbf{T} : \mathcal{H} \to \mathcal{H}$ *be an operator and* $\mathcal{O}$ *the support of an optimal solution of* (P) *with* $\mathbf{T}$*. Assume that* $\mathbf{T}$ *fulfills the restricted diagonal deviation property on* $\mathcal{O}$ *with constant* $\alpha > 0$*. If* $\alpha < 0.5$ *then Algo 1 solves* (P)*.*

Proof. Let denote by $\hat{x}$ the solution of (P) and $\mathcal{O} = \text{supp}(\hat{x})$ its support. From Lemma 2 and Lemma 3, we have, $\forall x \in \mathcal{H}$ such that $\text{supp}(x) \subseteq \mathcal{O}$,

$$\frac{\left\|\mathbf{T}(x)_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty}{\left\|\mathbf{T}(x)_{|\mathcal{O}}\right\|_\infty} = \frac{\left\|\mathbf{T}(x)_{|\mathcal{I}\setminus\mathcal{O}} - \mathbf{T}(\hat{x})_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty}{\left\|\mathbf{T}(x)_{|\mathcal{O}} - \mathbf{T}(\hat{x})_{|\mathcal{O}}\right\|_\infty} \ , \leq \frac{\alpha \|x-\hat{x}\|_\infty}{(1-\alpha)\|x-\hat{x}\|_\infty} \ .$$

So if $\alpha < 0.5$, we always select a good coordinate (or atoms). □

This result suggests that the RDDP is a sufficient condition and is not as tight as the ERC (Tropp, 2004). Working with non-linear operators forbids many mathematical tools that are available with the linear algebra, however there exists generalization that may help to build a tighter condition (Appell et al., 2004).

3.3. Noisy case

The non-noisy analysis gives us a condition on the RDDP. Now let $\mathbf{U} : x \mapsto \mathbf{T}(x) + e(x)$ be a perturbed version of an operator $\mathbf{T}$ with e such that $\forall x \in \mathcal{H}, \|e(x)\|_\infty \leq \rho$. Let $\hat{x}$ be the optimal sparse solution of $\mathbf{T}$ with $\text{supp}(x) \equiv \mathcal{O}$ (i.e. a solution of (P)). Let $\tilde{x}$ be the solution of (Q) with $\mathbf{U}$ with the support $\mathcal{O}$. Notice that if $\mathbf{T}$ fulfills the *restricted diagonal deviation property* on $\mathcal{O}$ then $\mathbf{U}$ has exactly the same property (and reciprocally).

We can now state the theorem that links $\hat{x}$ and $\tilde{x}$ and their supports,

Theorem 6. *Let* $\mathbf{T} : \mathcal{H} \to \mathcal{H}$ *be an operator and* $\mathcal{O}$ *the support of* $\hat{x}$*, the optimal solution of* (P)*. Let* $\mathbf{U} : x \mapsto \mathbf{T}(x) + e(x)$ *with* $\forall x \in \mathcal{H}, \|e(x)\|_\infty \leq \rho$ *a perturbed version of* $\mathbf{T}$*. Let* $\tilde{x} \in \mathcal{H}$ *be an optimal solution of* (Q) *with* $\mathbf{U}$ *and support* $\mathcal{O}$*. Assume that* $\mathbf{U}$ *fulfills the restricted diagonal property on support* $\mathcal{O}$ *with constant* $\alpha < 0.5$*. If the stopping criterion of Algo. 1 is such that*

$$\varepsilon > \frac{\rho}{1-2\alpha} \ ,$$

then when the algorithm stops at iteration k*, the following holds,*

(C1) $\mathcal{S}^k \subseteq \mathcal{O}$,

(C2) $\left\|x^k - \tilde{x}\right\|_\infty \leq \frac{\varepsilon}{1-\alpha}$,

(C3) $\|\hat{x} - \tilde{x}\|_\infty \leq \frac{\rho}{1-\alpha}$.

Before proving this theorem, we need to state some useful lemmas (these are reformulation of previously stated lemmas in (Zhang, 2009; Swirszcz et al., 2009)).

The following lemma gives a lower bound on the magnitude of the best element inside the optimal support when we consider an vector living on this support,

Lemma 7. *Let* $\mathcal{S} \subseteq \mathcal{O}$*, i.e.* $\mathcal{S}$ *is a set of good indices. Let* $\tilde{x} \in \mathcal{H}$ *a vector such that* $\mathbf{U}(\tilde{x})_{|\mathcal{S}} = 0$*. Then if* $\mathbf{U}$ *fulfills the RDDP on* $\mathcal{O}$ *with constant* α*, we have* $\forall z \in \mathcal{H}$ *such that* $\text{supp}(z) \subseteq \mathcal{S}$*,*

$$\left\|\mathbf{U}(z)_{|\mathcal{O}}\right\|_\infty \geq (1-\alpha)\|z - \tilde{x}\|_\infty \ .$$

Proof. This is a direct consequence of Lemma 2. □

The following lemma relates the solution of (P) with $\mathbf{T}$ with the solution (Q) using $\mathbf{U}$ and the same support,

Lemma 8. *Let $\hat{x}$ the solution of (P) with $\mathbf{T}$ and $\tilde{x}$ the solution of (Q) on the same support with $\mathbf{U}$. Let $\mathcal{O}$ the support of the optimal solution of (P) with $\mathbf{T}$. If $\mathbf{T}$ fulfills the RDDP on $\mathcal{O}$ with constant α and the perturbation e is such that $\forall x \in \mathcal{H}, \|e(x)\|_\infty < \rho$, we have,*

$$\|\hat{x} - \tilde{x}\|_\infty \leq \frac{\rho}{1-\alpha} .$$

Proof. This result comes from Lemma 2 and the definition of the perturbation. □

This last lemma provides an upper bound on the magnitude of the *best* element outside the optimal support $\mathcal{O}$ when considering a vector living on $\mathcal{O}$,

Lemma 9. *Let $\mathbf{U}$ be a perturbed version of an operator $\mathbf{T}$, $\forall x \in \mathcal{H}$, $\mathbf{U}(x) = \mathbf{T}(x) + e(x)$ with $\|e(x)\|_\infty < \rho$. Let $\mathcal{O}$ the support of the optimal solution $\hat{x}$ of (P) with $\mathbf{T}$. Let $\tilde{x}$ the solution of (Q) on $\mathcal{O}$ with $\mathbf{U}$. If $\mathbf{T}$ fulfill the RDDP on $\mathcal{O}$ with constant α, we have,*

$$\left\|\mathbf{U}(\tilde{x})_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty \leq \frac{\rho}{1-\alpha} .$$

Proof. As $\hat{x}$ is such that $\mathbf{T}(\hat{x}) = 0$ and $\mathrm{supp}(\tilde{x}) = \mathcal{O}$. We have, using Lemma 3 and Lemma 8,

$$\begin{aligned}
\left\|\mathbf{U}(\tilde{x})_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty &= \left\|\mathbf{U}(\tilde{x})_{|\mathcal{I}\setminus\mathcal{O}} - \mathbf{T}(\hat{x})_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty , \\
&\leq \left\|\mathbf{U}(\tilde{x})_{|\mathcal{I}\setminus\mathcal{O}} - \mathbf{U}(\hat{x})_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty + \|e(\hat{x})\|_\infty , \\
&\leq \alpha \|\tilde{x} - \hat{x}\|_\infty + \rho , \\
&\leq \alpha \tfrac{\rho}{1-\alpha} + \rho \leq \tfrac{\rho}{1-\alpha} .
\end{aligned}$$

□

We are now ready to prove the theorem.

Proof of Theorem 6. This proof works by induction. Now, we assume that all claims hold at step k. So at the beginning, we have $\mathcal{S}^k \subseteq \mathcal{O}$. First by combining Lemma 2 and Lemma 3, we have $\forall x, y \in \mathcal{H}$ such that $\mathrm{supp}(x) \cup \mathrm{supp}(y) \subseteq \mathcal{O}$,

$$\begin{aligned}
\left\|\mathbf{U}(x)_{|\mathcal{I}\setminus\mathcal{O}} - \mathbf{U}(y)_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty &\leq \alpha \|x - y\|_\infty , \\
&\leq \tfrac{\alpha}{1-\alpha} \left\|\mathbf{U}(x)_{|\mathcal{O}} - \mathbf{U}(y)_{|\mathcal{O}}\right\|_\infty ,
\end{aligned}$$

Then we yield,

$$\begin{aligned}
&\left\|\mathbf{U}(x^k)_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty \\
&\quad\leq \left\|\mathbf{U}(x^k)_{|\mathcal{I}\setminus\mathcal{O}} - \mathbf{U}(\tilde{x})_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty + \left\|\mathbf{U}(\tilde{x})_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty , \\
&\quad\leq \tfrac{\alpha}{1-\alpha} \left\|\mathbf{U}(x^k)_{|\mathcal{O}}\right\|_\infty + \left\|\mathbf{U}(\tilde{x})_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty . \qquad (5)
\end{aligned}$$

From the condition on ε and Lemma 9, it implies,

$$\left\|\mathbf{U}(\tilde{x})_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty \leq \tfrac{\rho}{1-\alpha} < \tfrac{1-2\alpha}{1-\alpha}\varepsilon . \qquad (6)$$

We now have to deal with 4 cases.

Case 1: $\left\|x^k - \tilde{x}\right\|_\infty > \frac{\varepsilon}{1-\alpha}$. This yields,

$$\left\|\mathbf{U}(x^k)_{|\mathcal{O}}\right\|_\infty > \varepsilon > \tfrac{1-\alpha}{1-2\alpha} \left\|\mathbf{U}(\tilde{x})_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty .$$

Using (5), we have $\left\|\mathbf{U}_{|\mathcal{I}\setminus\mathcal{O}}(\tilde{x})\right\|_\infty < \left\|\mathbf{U}(x^k)_{|\mathcal{O}}\right\|_\infty$ which implies, since $\alpha < 0.5$, the selection of a good element from the basis and the algorithm continues.

Case 2: $\left\|x^k - \tilde{x}\right\|_\infty \leq \frac{\varepsilon}{1-\alpha}$. We have then three choices.

Case 2.1: $j^{k+1} \in \mathcal{O}$ and the algorithm continues.

Case 2.2: $j^{k+1} \in \mathcal{O}$ and the algorithm stops.

Case 2.3: $j^{k+1} \notin \mathcal{O}$ then using (5) we have,

$$\begin{aligned}
\left\|\mathbf{U}(x^k)_{|\mathcal{O}}\right\|_\infty &\leq \left\|\mathbf{U}(x^k)_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty , \\
&\leq \tfrac{\alpha}{1-\alpha} \left\|\mathbf{U}(x^k)_{|\mathcal{O}}\right\|_\infty + \left\|\mathbf{U}(\tilde{x})_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty , \\
&\leq \tfrac{\alpha}{1-\alpha} \left\|\mathbf{U}(x^k)_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty + \left\|\mathbf{U}(\tilde{x})_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty .
\end{aligned}$$

This implies,

$$\left\|\mathbf{U}(x^k)_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty \leq \tfrac{1-\alpha}{1-2\alpha} \left\|\mathbf{U}(\tilde{x})_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty < \varepsilon ,$$

thus the algorithm stops. This leads to Theorem 6. □

This theorem gives lower bounds on the distance between the current estimate and the solutions of both problems (P) and (Q), it also guarantees the consistency of the estimated support. As the support of the optimal solution may be of infinite dimension (but countable), we cannot as in (Zhang, 2009) provide a bound on the number of missing elements. Notice also that both theorems ask for $\alpha < 0.5$, so Proposition 4 implies that Algo 1 is well-defined and the sub-optimization problem of line 9 has always a solution.

4. Extensions

We propose here two extensions of Algo. 1 that describe the behavior of OMP with errors on some steps.

4.1. Weak Non-Linear OMP

If $\mathcal{H}$ is an infinite dimensional space, then computing the inf-norm is troublesome. Instead of searching for the best element, Temlyakov (2000) proposes to make a good guess

by introducing some tolerance β, this implies to replace the selection step by,

$$j^{k+1} \in \mathcal{I} \setminus \mathcal{S}^k \text{ s.t.} \\ \left|\left\langle \mathbf{T}(x^k), \varphi_{j^{k+1}} \right\rangle\right| \geq \beta \sup_{i \in \mathcal{I} \setminus \mathcal{S}^k} \left|\left\langle \mathbf{T}(x^k), \varphi_i \right\rangle\right| , \tag{7}$$

with $0 < \beta \leq 1$. Such a change has of course repercussions on the two theorems. For the non-noisy case, we now have

Theorem 10. *Let $\mathbf{T} : \mathcal{H} \to \mathcal{H}$ be an operator and $\mathcal{O}$ the support of the optimal solution of (P). If $\mathbf{T}$ fulfills the restricted diagonal deviation property on $\mathcal{O}$ with constant $\alpha > 0$. If $\alpha < \frac{\beta}{1+\beta}$ then Algo 1, with (7) as selection step, solves (P).*

Proof. Since we want $\forall x \in \mathcal{H}$, $\text{supp}(x) \subseteq \mathcal{O}$, $\left\|\mathbf{T}(x)_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty < \beta \left\|\mathbf{T}(x)_{|\mathcal{O}}\right\|_\infty$ we just adapt the previous proof. □

For the noisy case, we need to introduce a new variable to strengthen the inequalities, then we have,

Theorem 11. *Let $\mathbf{T} : \mathcal{H} \to \mathcal{H}$ be an operator and $\mathcal{O}$ the support of $\hat{x}$, the optimal solution of (P). Let $\mathbf{U} : x \mapsto \mathbf{T}(x) + e$ with $\|e\|_\infty \leq \rho$ a perturbed version of $\mathbf{T}$. Let $\tilde{x} \in \mathcal{H}$ be an optimal solution of (Q) with support $\mathcal{O}$. Assume that $\mathbf{U}$ fulfills the restricted diagonal deviation property on support $\mathcal{O}$ with constant $\alpha < \frac{a\beta}{1+a\beta}$ (with a such that $0 < a < 1$). If the stopping criterion of Algo. 1, with (7) as selection step, is such that*

$$\varepsilon > \tfrac{\rho}{(1-a)\alpha} ,$$

then when the algorithm stops the following holds,

(C1) $\mathcal{S}^k \subseteq \mathcal{O}$,

(C2) $\left\|x^k - \tilde{x}\right\|_\infty \leq \frac{\varepsilon}{1-\alpha}$,

(C3) $\|\hat{x} - \tilde{x}\|_\infty \leq \frac{\rho}{1-\alpha}$.

Proof. This proof works by induction. Now we assume that all claims hold at step k. So at the beginning, we have $\mathcal{S}^k \subseteq \mathcal{O}$.

From the condition on ε, we imply with Lemma 9,

$$\left\|\mathbf{U}(\tilde{x})_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty \leq \tfrac{\rho}{1-\alpha} < \tfrac{(1-a)\alpha}{1-\alpha}\varepsilon . \tag{8}$$

We now have to deal with 4 cases.

Case 1: $\left\|x^k - \tilde{x}\right\|_\infty > \frac{\varepsilon}{1-\alpha}$. This yields,

$$\left\|\mathbf{U}(x^k)_{|\mathcal{O}}\right\|_\infty > \varepsilon > \tfrac{1-\alpha}{(1-a)\alpha}\left\|\mathbf{U}(\tilde{x})_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty .$$

Since $\frac{\alpha}{1-\alpha} < a\beta$ and using (5), we have $\left\|\mathbf{U}_{|\mathcal{I}\setminus\mathcal{O}}(\tilde{x})\right\|_\infty < \beta\left\|\mathbf{U}(x^k)_{|\mathcal{O}}\right\|_\infty$ which implies the selection of a good element from the basis and the algorithm continues.

Case 2: $\left\|x^k - \tilde{x}\right\|_\infty \leq \frac{\varepsilon}{1-\alpha}$. We have then three choices.

Case 2.1: $j^{k+1} \in \mathcal{O}$ and the algorithm continues.

Case 2.2: $j^{k+1} \in \mathcal{O}$ and the algorithm stops.

Case 2.3: $j^{k+1} \notin \mathcal{O}$ then using (5) we have,

$$\begin{aligned}\left\|\mathbf{U}(x^k)_{|\mathcal{O}}\right\|_\infty &\leq \left\|\mathbf{U}(x^k)_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty / \beta , \\ &\leq \tfrac{\alpha}{\beta(1-\alpha)}\left\|\mathbf{U}(x^k)_{|\mathcal{O}}\right\|_\infty + \left\|\mathbf{U}(\tilde{x})_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty , \\ &\leq \tfrac{\alpha}{\beta^2(1-\alpha)}\left\|\mathbf{U}(x^k)_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty + \left\|\mathbf{U}(\tilde{x})_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty , \\ &\leq \tfrac{a}{\beta}\left\|\mathbf{U}(x^k)_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty + \left\|\mathbf{U}(\tilde{x})_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty .\end{aligned}$$

As we have $\beta > \alpha$, it implies,

$$\left\|\mathbf{U}(x^k)_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty \leq \tfrac{\beta}{1-a}\left\|\mathbf{U}(\tilde{x})_{|\mathcal{I}\setminus\mathcal{O}}\right\|_\infty < \varepsilon ,$$

thus the algorithm stops. This leads to Theorem 11. □

The introduction of the variable a makes the comparison between Theorem 6 and Theorem 11 non trivial. In fact, this variable is here to guarantee $\alpha < 0.5$. In both case choosing α close to the limit leads to an high value for ϵ. So these two formulations lead to the same results and requirements.

4.2. Inexact solution of the sub-optimization problem

With non-linear operators, finding the *true* estimate at line 9 of Algo. 1 can be difficult. Most point fixed algorithms converge at infinity (Cegielski, 2013) so getting the exact solution is almost impossible. Some methods allow a good control on the convergence (e.g. with Banach contraction), this often allows to guess the accuracy of the estimate. As we require operators to fulfill the *restricted diagonal deviation property*, we have some information about convergence when dealing with sparse vectors of $\mathcal{H}$.

The error can be modelled as a deviation toward 0 of the estimate on the support, i.e we replace the line 9 of Algo. 1 by,

$$x^{k+1} \in \mathcal{H} \text{ s.t. } \begin{cases} \text{supp}(x^{k+1}) \subseteq \mathcal{S}^{k+1} , \\ \left\|\mathbf{T}(x^{k+1})|_{\mathcal{S}^{k+1}}\right\|_\infty < \tau , \end{cases} \tag{9}$$

with $\tau \geq 0$ the maximal deviation of the estimation. We assume here that the support constraint is always guaranteed (i.e. in some case a projection on the final estimate may be necessary).

For the non-noisy case, we have the following theorem,

Theorem 12. *Let $\mathbf{T} : \mathcal{H} \to \mathcal{H}$ be an operator and $\mathcal{O}$ the support of the optimal solution of (P). If $\mathbf{T}$ fulfills the restricted diagonal deviation property with $\alpha < 0.5$, then Algo. 1 with (9) recovers the support at first.*

Proof. If we guarantee the estimate it belongs to the support $\mathcal{S}^{k+1}$, i.e. $\operatorname{supp}(x^{k+1}) \subseteq \mathcal{S}^{k+1}$. Then if $\alpha < 0.5$ we still select element of the support at each steps. Of course if the size of the support is finite at one point we will select out-of-support elements. □

This results clearly suggests that the RDDP is only a sufficient condition for the convergence and consistency of OMP.

For the noisy case, we have,

Theorem 13. *Let $\mathbf{T} : \mathcal{H} \to \mathcal{H}$ be an operator and $\mathcal{O}$ the support of $\hat{x}$, the optimal solution of (P). Let $\mathbf{U} : x \mapsto \mathbf{T}(x) + e(x)$ with $\forall x \in \mathcal{H}, \|e(x)\|_\infty \leq \rho$ a perturbed version of $\mathbf{T}$. Let $\tilde{x} \in \mathcal{H}$ be an optimal solution of (Q) with support $\mathcal{O}$. Assume that $\mathbf{U}$ fulfills the restricted diagonal deviation property on support $\mathcal{O}$ with constant $\alpha < 0.5$. If the stopping criterion of Algo. 1 with (9) is such that*

$$\varepsilon > \tfrac{\rho}{1-2\alpha} \ ,$$

then when the algorithm stops the following holds,

(C1) $\mathcal{S}^k \subseteq \mathcal{O}$,

(C2) $\left\|x^k - \tilde{x}\right\|_\infty \leq \tfrac{\varepsilon}{1-\alpha}$,

(C3) $\|\hat{x} - \tilde{x}\|_\infty \leq \tfrac{\tau+\rho}{1-\alpha}$.

Proof. We only need to adapt Lemma 8 to reflect the error, the rest of the proof of Theorem 6 still hold. □

5. Discussion

In this section, we first propose to discuss about the RDDP and its links with others criteria. Then we will propose a *realistic* version of OMP using both extensions of Section 4. Finally, we show the links with the previous work of Zhang (2011) on function minimization.

5.1. The restricted diagonal deviation property

Definition 1 introduces the key of all our analysis. This definition clearly leads to a sufficient condition for the consistency of Algo. 1, finding an expression closer to the *exact recovery criterion* (ERC) asks for deeper analysis around the non-linearity of the operator. Still one may require the assumption of the definition to work only on vectors $x \in \mathcal{H}$ such that, for the involved operator $\mathbf{T}$, we have $\mathbf{T}(x)_{|\operatorname{supp}(x)} = 0$. This reduces the number of vectors that should fulfill the condition of both Theorem 5 and Theorem 6 (and their *weak* versions still hold). However, when dealing with inexact steps like in Section 4.2 we need to make a new analysis.

For the linear setting (i.e. when $\mathbf{T}$ is a linear or affine operator), the famous *restricted isometry property* (RIP) can be used to build the consistency analysis. However RIP uses the ℓ_2-norm while our RDDP used the inf-norm, but if the optimal support $\mathcal{O}$ is finite we can apply equivalence relation between norms. Assume that $\mathbf{T}$ fulfills RDDP on finite support $\mathcal{O}$ with constant α, we have $\forall x, y \in \mathcal{H}$ such that $\operatorname{supp}(x) \cup \operatorname{supp}(y) \subseteq \mathcal{O}$,

$$\begin{aligned}
&\|\mathbf{T}(x) - \mathbf{T}(y) - \mathbf{D}_{xy}(x-y)\| \\
&\quad \leq \|\mathbf{T}(x) - \mathbf{T}(y) - \mathbf{D}_{xy}(x-y)\|_\infty \\
&\quad \leq \alpha \|x-y\|_\infty \\
&\quad \leq \alpha\sqrt{\operatorname{card}\{\mathcal{O}\}}\, \|x-y\| \ .
\end{aligned}$$

Thus this implies,

$$\begin{aligned}
&\left(1 - \alpha\sqrt{\operatorname{card}\{\mathcal{O}\}}\right) \|x-y\| \\
&\quad \leq \|\mathbf{T}(x) - \mathbf{T}(y)\| \\
&\quad \leq \left(\|\mathbf{D}_{xy}\| + \alpha\sqrt{\operatorname{card}\{\mathcal{O}\}}\right) \|x-y\| \ ,
\end{aligned}$$

which directly leads to RIP when $\mathbf{T}$ is affine. As we work with infinite dimensional spaces, it is possible to use the works of Hansen & Adcock (2011) who generalize Compressed Sensing to infinite dimensional spaces.

The RDDP works with a basis of the Hilbert space and the selection process is toward the element of the basis. Extensions are possible, for example, Gribonval & Vandergheynst (2006) use redundant dictionary (that could have an infinite number of atoms). Then guarantees on the selection of atoms are delicate to set and one needs strong hypothesis on the dictionary to fulfill the consistency requirements. However such dictionaries are better models for feature selection task as they are often correlated.

5.2. Dealing with perturbations

From the analysis of both Section 3 and Section 4, we are able to build Algo. 2, a *realistic* version of the non-linear orthogonal matching pursuit. Such algorithm is more reliable when working in very high dimensional spaces. However, while finding values for τ and $k_{\max}$ can be easily done, ε is difficult to set as it depends on β the tolerance on the selection step and ρ the perturbation of the operator. Still in a given statistical context, one may build some procedures that, with an high probability, leads to accurate values (e.g. using concentration inequalities (Massart & Picard, 2007)).

Algo. 2 requires to solve a sub-optimization (line 9), however depending on the properties of $\mathbf{T}$ such problem may be very hard to solve. For example, if $\mathbf{T}$ is continuous and

Algorithm 2 Realistic NL-OMP

Input: $\mathbf{T}$ an operator, $k_{\max}$ the maximal sparsity, ε the tolerance, τ the maximal deviation, β the tolerance for the maximum research step.
Initialization: $x^0 \leftarrow 0, \mathcal{S}^0 \leftarrow \emptyset$.
for $k = 0$ to $k_{\max}$ **do**
  $j^{k+1} \in \mathcal{I} \setminus \mathcal{S}^k$ s. t.
  $\left|\left\langle \mathbf{T}(x^k), \varphi_{j^{k+1}} \right\rangle\right| \geq \beta \sup_{i \in \mathcal{I} \setminus \mathcal{S}^k} \left|\left\langle \mathbf{T}(x^k), \varphi_i \right\rangle\right|$,
  if $\left(\left|\left\langle \mathbf{T}(x^k), \varphi_{j^{k+1}} \right\rangle\right| \leq \varepsilon\right)$ **then**
    break.
  end if
  $\mathcal{S}^{k+1} \leftarrow \mathcal{S}^k \cup \{j^{k+1}\}$,
  $x^{k+1} \in \mathcal{H}$ s.t. $\begin{cases} \operatorname{supp}(x^{k+1}) \subseteq \mathcal{S}^{k+1} \\ \left\|\mathbf{T}(x^{k+1})|_{\mathcal{S}^{k+1}}\right\| \leq \tau. \end{cases}$
end for
Output: the estimate x^k with its support $\mathcal{S}^k$.

restricted to a compact convex set $\mathcal{X} \subseteq \mathbb{R}^n$ (with $n > 0$), then $\mathbf{T}$ has fixed point (see Brouwer theorem (Chow et al., 1978; Cegielski, 2013)), but finding such point can be very difficult. However for some family of operators, many methods are available (see (Bauschke & Combettes, 2011; Cegielski, 2013)) and some provide also some controls on the convergence.

We can finally state a theorem combining all the previous in the noisy case,

Theorem 14. *Let $\mathbf{T} : \mathcal{H} \to \mathcal{H}$ be an operator and $\mathcal{O}$ the support of $\hat{x}$, the optimal solution of (P). Let $\mathbf{U} : x \mapsto \mathbf{T}(x) + e(x)$ with $\forall x \in \mathcal{H}, \|e(x)\|_\infty \leq \rho$ a perturbed version of $\mathbf{T}$. Let $\tilde{x} \in \mathcal{H}$ be an optimal solution of (Q) with support $\mathcal{O}$. Assume that $\mathbf{U}$ fulfills the restricted diagonal deviation property on support $\mathcal{O}$ with constant $\alpha < \frac{a\beta}{1+a\beta}$ ($0 < a < 1$). If the stopping criterion of Algo. 2 is such that*

$$\varepsilon > \tfrac{\rho}{(1-a)\alpha} ,$$

then when the algorithm stops the following holds,

(C1) $\mathcal{S}^k \subseteq \mathcal{O}$,

(C2) $\left\|x^k - \tilde{x}\right\|_\infty \leq \frac{\varepsilon}{1-\alpha}$,

(C3) $\|\hat{x} - \tilde{x}\|_\infty \leq \frac{\tau+\rho}{1-\alpha}$.

Such theorem implies that feature selection in noisy context is still possible but with some prices.

5.3. Links with (Zhang, 2011)

In this paper, the author proposes a generalization of the orthogonal matching pursuit to the minimization of convex functions with a sparsity constraint, i.e.

$$\min_{x \in \mathcal{H}} f(x) \text{ s.t. } \|x\|_0 \leq k , \tag{10}$$

with k the maximum number of coefficients and $f : \mathcal{H} \to \mathbb{R}$ a convex differentiable function. In order to build the analysis of the algorithm, the author use the *restricted strong smoothness property* (RSSP) and the *restricted strong convexity property* (RSCP) (see also (Blumensath, 2013; Bahmani et al., 2013; Jain et al., 2014) for other use of both properties) as an alternative to RIP for non-linear setting.

However these properties use the Bregman divergence which requires convex functions (at least on sparse sets in this case). Such functions enter in our framework as a special case by taking $\mathbf{T}$ as the gradient operator (thanks to the diagonal operator in RDDP). Furthermore, if $\mathcal{H}$ is a real Hilbert space, take $f : \mathcal{H} \to \mathbb{R}$ a smooth convex function then, we have $\forall x, y \in \mathcal{H}$ (Bauschke et al., 2012),

$$\|\nabla f(x) - \nabla f(y)\|^2 + \|x - y - \nabla f(x) + \nabla f(y)\|^2 \leq \|x - y\|^2 ,$$

so if f is such that for some support $\mathcal{O} \subseteq \mathcal{I}$, we have $\forall x, y \in \mathcal{H}, \operatorname{supp}(x) \cup \operatorname{supp}(y) \subseteq \mathcal{O}$,

$$\|\nabla f(x) - \nabla f(y)\| \geq m \|x - y\| , \tag{11}$$

with $m > 0$, then,

$$\|x - y - \nabla f(x) + \nabla f(y)\| \leq \sqrt{1 - m^2} \|x - y\| . \tag{12}$$

(11) implies the *restricted strong smoothness property* as we can lower bound f using a strictly convex and smooth function. So with (12) we have both RSSP and RSCP. Equation (12) also implies with Proposition 4 the existence of solutions for the sub-optimization problems and for (Q), furthermore as $\mathbf{I} - \nabla f$ is a Banach contraction making a iterative scheme for solving the sub-optimization problems is easy (Cegielski, 2013).

6. Conclusion

In this paper we present a full analysis of the orthogonal matching pursuit algorithm as a solver for finding sparse zero of Hilbert operator. The proposed algorithms include some of the previous generalizations (e.g. function minimization and the weak OMP) with new results on the consistency of the estimated support. Such guarantees should be very helpful when dealing with M-estimators (as in (Jain et al., 2014)). Future works includes generalization to Banach space like in (Temlyakov, 2001), working with structured sparsity (Huang et al., 2011) or links with nonlinear eigenproblems (Hein & Bühler, 2010).

References

Appell, Jürgen, De Pascale, Espedito, and Vignoli, Alfonso. *Nonlinear spectral theory*, volume 10. Walter de Gruyter, 2004.

Bahmani, Sohaul, Raj, Bhiksha, and Boufounos, Petros T. Greedy sparsity-constrained optimization. *J. of Machine Learning Research*, 14(3):807–841, 2013.

Bauschke, Heinz H and Combettes, Patrick L. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011.

Bauschke, Heinz H, Moffat, Sarah M, and Wang, Xianfu. Firmly nonexpansive mappings and maximally monotone operators: correspondence and duality. *Set-Valued and Variational Analysis*, 20(1):131–153, 2012.

Beck, Amir and Eldar, Yonina C. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *SIAM Journal on Optimization*, 23(3):1480–1509, 2013.

Beck, Amir and Hallak, Nadav. On the minimization over sparse symmetric sets. Technical report, Technion, 2014.

Blumensath, Thomas. Compressed sensing with nonlinear observations and related nonlinear optimization problems. *IEEE Transactions on Information Theory*, 59(6): 3466–3474, 2013.

Candès, Emmanuel J., Romberg, Justin, and Tao, Terence. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Trans. on*, 52(2):489–509, 2006.

Cegielski, Andrzej. *Iterative methods for fixed point problems in Hilbert spaces*. Springer, 2013.

Chow, Shui Nee, Mallet-Paret, John, and Yorke, James A. Finding zeroes of maps: homotopy methods that are constructive with probability one. *Math. Comp.*, 32:887–899, 1978.

DasGupta, Anirban. *Asymptotic theory of statistics and probability*. Springer Science & Business Media, 2008.

Dupé, François-Xavier and Anthoine, Sandrine. A greedy approach to sparse poisson denoising. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2013.

Foucart, Simon. Stability and robustness of weak orthogonal matching pursuits. In *Recent advances in harmonic analysis and applications*, pp. 395–405. Springer, 2013.

Gribonval, Rémi and Vandergheynst, Pierre. On the exponential convergence of matching pursuits in quasi-incoherent dictionaries. *Information Theory, IEEE Transactions on*, 52(1):255–261, 2006.

Hansen, Anders C and Adcock, Ben. Generalized sampling and infinite dimensional compressed sensing. *Magnetic Resonance Imaging*, pp. 1, 2011.

Hein, Matthias and Bühler, Thomas. An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse pca. In *Advances in Neural Information Processing Systems*, pp. 847–855, 2010.

Huang, Junzhou, Zhang, Tong, and Metaxas, Dimitris. Learning with structured sparsity. *The Journal of Machine Learning Research*, 12:3371–3412, 2011.

Jain, Prateek, Tewari, Ambuj, and Kar, Purushottam. On iterative hard thresholding methods for high-dimensional m-estimation. In *Advances in Neural Information Processing Systems*, pp. 685–693, 2014.

Livshitz, Eugene D and Temlyakov, Vladimir N. Sparse approximation and recovery by greedy algorithms. *Information Theory, IEEE Transactions on*, 60(7):3989–4000, 2014.

Lozano, Aurélie C, Swirszcz, Grzegorz, and Abe, Naoki. Group orthogonal matching pursuit for logistic regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 452–460, 2011.

Mallat, Stéphane G. and Zhang, Zhifeng. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Trans. on*, 41(12):3397–3415, 1993.

Massart, Pascal and Picard, Jean. *Concentration inequalities and model selection*, volume 1896. Springer, 2007.

Sindhwani, Vikas and Lozano, Aurélie C. Non-parametric group orthogonal matching pursuit for sparse learning with multiple kernels. In *Advances in Neural Information Processing Systems*, pp. 2519–2527, 2011.

Swirszcz, Grzegorz, Abe, Naoki, and Lozano, Aurelie C. Grouped orthogonal matching pursuit for variable selection and prediction. In *Advances in Neural Information Processing Systems*, pp. 1150–1158, 2009.

Temlyakov, Vladimir N. Weak greedy algorithms. *Advances in Computational Mathematics*, 12(2-3):213–227, 2000.

Temlyakov, Vladimir N. Greedy algorithms in banach spaces. *Advances in Computational Mathematics*, 14(3): 277–292, 2001.

Temlyakov, Vladimir N. Greedy approximation in convex optimization. *Constructive Approximation*, pp. 1–28, 2012.

Tropp, Joel A. Greed is good: Algorithmic results for sparse approximation. *Information Theory, IEEE Transactions on*, 50(10):2231–2242, 2004.

Yuan, Xiaotong, Li, Ping, and Zhang, Tong. Gradient hard thresholding pursuit for sparsity-constrained optimization. In *The 31st International Conference on Machine Learning*, pp. 127–135, 2014.

Zhang, T. Sparse recovery with orthogonal matching pursuit under RIP. *Information Theory, IEEE Trans. on*, 57 (9):6215–6221, 2011.

Zhang, Tong. On the consistency of feature selection using greedy least squares regression. *The Journal of Machine Learning Research*, 10:555–568, 2009.