Generalized Greedy Alternatives *

François-Xavier Dupé, Sandrine Anthoine

Generalized Greedy Alternatives*

François-Xavier Dupé [†1] and Sandrine Anthoine [‡2]

[1]Aix Marseille Univ, CNRS, Centrale Marseille, LIF, Marseille, France
[2]Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France

January 10, 2017

Abstract

In this paper, we develop greedy algorithms to tackle the problem of finding sparse approximations of zeros of operators in a Hilbert space of possibly infinite dimension. Four greedy algorithms, *Subspace Pursuit*, *CoSaMP*, *HTP* and *IHT*, which are classically used for sparse approximation, are extended. A criterion, the *Restricted Diagonal Property*, is developed. It guarantees the definition of the extended algorithms and is used to derive error bounds. We also provide examples and experiments that illustrate these theoretical results.

Keywords: Sparse representation; greedy algorithm; Hilbert space; zero-finding; non-convex optimization; Poisson denoising.

1 Introduction

1.1 Motivations

The demand for high-dimensional data analysis has been significantly boosted over the past decade by, on the one hand, the emergence of large-size data such as found in medical or sonar imaging, social networking, bioinformatics... and on the other hand, the democratization of portable devices required to perform complex task with limited resources. When formalizing these problems, the large dimensionality or amount of data available usually leads to a complex high-dimensional set of solutions, while the number of collected samples to evaluate these solutions is significantly smaller. This renders any inference from the data ill-posed. Sparse modeling has proved to be quite efficient to alleviate this problem by capturing the intrinsic low-dimensional structure in the data that can not be revealed on the raw problem. Having at hand computational methods aiming at modeling high dimensional data with sparse representations is therefore of the utmost importance.

The seminal problem in sparse modeling is that of finding the best k-sparse approximation of a signal y on a redundant dictionary $\mathbf{\Phi}$ or more formally, to solve the problem:

$$\text{Find } \hat{x} \in \underset{x\in\mathbb{R}^K}{\operatorname{argmin}} \tfrac{1}{2}\left\|\mathbf{\Phi}x - y\right\|_2^2 \text{ s.t. } \left\|x\right\|_0 \leq k. \tag{P0}$$

Here K represents the dimension of the raw data and may thus be very large, while k denotes the intrinsic low dimension and is small. Since the problem involves the number of non-zeros entries, $\|x\|_0$, it is combinatorial by nature and thus hard to solve. Two main categories of methods have been proposed to tackle such problems. On the one hand, *greedy methods* focus on the original combinatorial problem and try to solve it by taking local decisions. In fact, they aim at finding the so-called support of x, which is the location of the non-zero entries, by taking the best decision locally. The value of the

*This work is partially supported by the French GIP ANR under contract ANR GRETA 12-BS02-004-01 GREediness: Theory and Algorithms.

†francois-xavier.dupe@lif.univ-mrs.fr
‡sandrine.anthoine@univ-amu.fr

corresponding coefficients is of course estimated in the process as well. Although *greedy methods* such as *Matching Pursuit* (MP) or *Orthogonal Matching Pursuit* (OMP) (Mallat and Zhang, 1993) were the first techniques employed to solve Problem (P0), the second category, namely the *relaxation methods*, has enjoyed a lot of attention in the literature. Relaxation methods use a proxy for the l_0-pseudo norm $\|.\|_0$, thus rendering the problem easier to solve. The most popular relaxation is the l_1-norm, which leads to a convex optimization problem, but other techniques involving non-convex norms approaching the l_0-pseudo norm are used as well. Relaxation methods, although not designed to solved the original problem, may provably solve it, as is the case for the l_1 relaxation under the RIP property (Candès et al., 2006). Moreover they have been studied quite thoroughly in several aspects since they also pertain to a large class of research areas such as convex optimization, control, etc. As a result it has been possible to extend their use quite further from the original setting of Problem (P0), for example by considering other penalty terms (Kullback-Liebler divergence, hinge-loss), by modifying the sparse structure to group-sparsity... or even tackling inverse problems instead of the approximation problem.

The scope of *greedy methods* has not been extended so far for now. The goal of this paper is to go one step further in this direction. It tackles in particular the following questions: 1) Can *greedy methods* be used in the inverse problem setting, where the penalty may not by differentiable and more precisely, can the proximal tools, quite useful in convex optimization, be of help in this framework? 2) Can *greedy methods* be employed with non-linear operators, or solve for the more general questions than approximation such as best k-sparse solution to a non-linear problem? 3) Can *greedy methods* work in infinite dimension, knowing that so far, most guarantees in the matter actually rely on the fact that the ambient space is of finite dimension?

1.2 State-of-the-art

Let us make a quick review of the literature concerning what we call here *greedy methods*.

In the context of finding the best k-sparse approximation of linear system of equations (i.e. Problem (P0)), many methods has been proposed. Besides the original Matching Pursuit (MP), Orthogonal Matching Pursuit (OMP) (Mallat and Zhang, 1993), one can cite Iterative Hard Thresholding (IHT), Compressed Sampling Matching Pursuit (CoSaMP) (Needell and Tropp, 2009) or Subspace Pursuit (SP) (Dai and Milenkovic, 2009). These methods have been analyzed using linear algebra tools, in particular notions like the Restricted Isometry Property (RIP), spark or mutual (in)coherence played an essential role to analyze them. Moving away from the linear setting, that is, trying to exploit *greedy methods* to find a k-sparse approximation to a non-linear system of equation thus requires new tools.

In the literature, several generalizations of the previously mentioned *greedy methods* have been proposed, they all attempt to find the k-sparse element that minimizes a loss, (i.e. they replace the $\frac{1}{2}\|\mathbf{\Phi}x - y\|_2^2$ Problem (P0) with a more generic function):

$$\text{Find } \hat{x} \in \underset{x\in\mathbb{R}^K}{\operatorname{argmin}} f(x) \text{ s.t. } \|x\|_0 \leq k. \tag{P1}$$

The first works in this direction are on OMP by (Zhang, 2011) and IHT (Blumensath, 2013) where the authors used the Bregman divergence to build a criterion for convergence. This criterion has then been used to generalize CoSaMP with the Gradient Support Pursuit (GRASP) (Bahmani et al., 2013), Hard Thresholding Pursuit (Yuan et al., 2014) and the greedy forward-backward (Jalali et al., 2011). Recently (Jain et al., 2014) provided a generalization of many existing methods in the context of function minimizing and M-estimation. All these methods aim at finding the k-sparse minimizer of a convex (generally smooth) function and in a finite dimensional Hilbert space.

In this paper, we propose a new point of view: instead of generalizing Problem (P0) to more complex minimization problem, we rather ask whether *greedy methods* can solve a non-linear set of equation denoted by an operator $\mathbf{T}$ under a sparsity constraint. We formulate the problem in a Hilbert space that is not necessarily of finite dimension. We adapt four classical *greedy methods*, namely CoSaMP, Subspace Pursuit, Iterative Hard thresholding and Hard Thresholding Pursuit (Foucart, 2011) to this setting. We propose a criterion that is quite inspired by the Restricted Isometry Property and guarantees the behavior of these generalizations. We finally exhibit examples, both theoretical and numerical to illustrate these results.

1.3 Problem Formulation

Before expressing the problem we want to tackle, let us first detail the notation used throughout the paper:

- $\mathcal{H}$ is a real separable Hilbert space, $\{e_i\}_{i\in\mathbb{N}}$ is an orthonormal basis of $\mathcal{H}$ (note that $\mathcal{H}$ need not be finite-dimensional) and $\langle .,. \rangle$ and $\|.\|$ the corresponding scalar product and endowed norm. For x in $\mathcal{H}$, $x_i = \langle e_i, x \rangle$.
- $\mathbf{T} : \mathcal{H} \to \mathcal{H}$ is an operator on $\mathcal{H}$ with full domain: $\mathrm{dom}(\mathbf{T}) = \mathcal{H}$. $\mathbf{I}$ denotes the identity.
- $\mathrm{supp}(x)$ is the support of x:
$$\mathrm{supp}(x) = \{i \in \mathbb{N} : x_i \neq 0\} \ . \tag{1}$$
- For $(x, y) \in \mathcal{H}^2$, the union of their supports is $\mathrm{supp}(x, y) = \mathrm{supp}(x) \cup \mathrm{supp}(y)$.
- $\|x\|_0 = \mathrm{card}(\mathrm{supp}(x))$ is the l_0-pseudo norm of x. i.e. the number of non-zero entries in the sequence $\{x_i\}_{i\in\mathbb{N}}$ (card denotes the cardinal).
- If $\mathcal{R}$ is a subset of $\mathbb{N}$, $\mathcal{R}^c$ is its complement $\mathcal{R}^c = \{i \in \mathbb{N}/i \notin \mathcal{R}\}$.
- $x_{|\mathcal{R}}$ denotes the orthogonal linear projection of x onto $\mathrm{span}\{e_i,\ i \in \mathcal{R}\}$ for $\mathcal{R}$ a finite subset of $\mathbb{N}$:
$$x_{|\mathcal{R}} = \sum_{i\in\mathcal{R}} x_i e_i \ . \tag{2}$$
- $x_{|k}$ denotes the restriction of x to its k largest entries when k is an integer: given ψ a permutation such that $|x_{\psi(i)}| \geq |x_{\psi(i+1)}|$ for all i, then
$$x_{|k} = \sum_{i=1}^{k} x_{\psi(i)} e_{\psi(i)} \ . \tag{3}$$
- $\mathbf{P}_{\mathcal{R}}$ is the orthogonal linear projection $\mathbf{P}_{\mathcal{R}} : x \mapsto x_{|\mathcal{R}}$ when $\mathcal{R}$ a finite subset of $\mathbb{N}$.
- $\Gamma_0(\mathcal{H})$ is the set of functions from $\mathcal{H}$ to $]-\infty, +\infty]$ which are lower semicontinuous, convex, and proper.

In this paper we tackle the problem of finding a k-sparse solution of non-linear system of equation, in other words:

$$\text{Given } \mathbf{T} : \mathcal{H} \to \mathcal{H} \text{ and } d \in \mathcal{H},\ \text{find } x \in \mathcal{H} \text{ s.t. } \mathbf{T}(x) = d \text{ and } \|x\|_0 \leq k \ . \tag{P2}$$

As is clear by replacing $\mathbf{T}$ with $\mathbf{T}' :\ x \mapsto \mathbf{T}(x) - d$, it is equivalent to finding the zero of a non-linear operator that is k-sparse:

$$\text{Find } x \in \mathcal{H} \text{ s.t. } \mathbf{T}'(x) = 0 \text{ and } \|x\|_0 \leq k \ . \tag{P3}$$

This problem may not always have a solution: for example, if $\mathbf{T}$ is the differential of a convex function, Problem (P3) amounts to finding a global minimizer of that function which is k-sparse: that may not exist ! The criterion we propose in Section 3.1 and the subsequent theorems show that if Problem (P3) has a solution, we guarantee to converge to it, and otherwise we guarantee to find a "good" guess in a sense that will be made clear by the theorems themselves.

1.4 Paper Organization

The rest of the paper is organized as follows. Section 2 details four greedy algorithms generalized to the non-linear setting of Problem (P3). Section 3 introduces our new criterion, which is consequently used in Section 4 to derive theoretical convergence guarantees for these four algorithms. The proofs are given in Section 5. Examples of applications are given in Section 6 and numerical experiments in Section 7.

2 Four Generalized Greedy Algorithms

2.1 Algorithms

Let us start with explaining how to adapt what we call here *greedy methods* to solve Problem (P3). The seminal methods that were designed to find the best k-sparse approximation on a dictionary $\mathbf{\Phi} = (\phi_1, \dots, \phi_M)$ in Problem (P0), namely MP and OMP, were termed greedy because of the way they aggregate information. Indeed, starting from the null approximation or equivalently the empty set for the support of the solution, they select at every iteration a new atom ϕ_i that is added to the support and update the approximation accordingly. Here we will focus on the second generation of so-called greedy methods designed for Problem (P0), where instead of starting from scratch and selecting one atom at a time, one rather starts with a set of k atoms and have this support set evolve. The first algorithm proposed that does so is CoSaMP (Needell and Tropp, 2009), other have followed such as Subspace Pursuit (Dai and Milenkovic, 2009) or HTP (Foucart, 2011). We will generalize those three which share the same structure, as well as Iterative Hard Thresholding which may be seen as a simpler version (less focussed on the support).

The common structure of Subpace Pursuit, CoSaMP and HTP for Problem (P0) is the following. Given a candidate support $\mathcal{T}$ of size k and a candidate solution x:

Step 1: One identifies new candidate atoms ϕ_i by examining the residual $\mathbf{\Phi}x - y$.

Step 2: One defines an extended support $\mathcal{S}$ of size $2k$ or $3k$ or $4k$, by aggregating $\mathcal{T}$ and the labels of the atoms found at step 1.

Step 3: One designs b, a -possibly approximate- solution of Pb. (P0) on the extended support $\mathcal{S}$.

Step 4: One identifies in b the k "best" atoms and update the support $\mathcal{T}$ of size k accordingly.

Step 5: One designs x, a -possibly approximate- solution of Pb. (P0) on the support $\mathcal{T}$.

CoSaMP, HTP and Subspace Pursuit all proceed through these five steps and iterate. The new directions chosen at step 1 are a set of k or $2k$ or $3k$ atoms yielding the largest coefficients of $\mathbf{\Phi}^*(\mathbf{\Phi}x-y)$. Step 4 consists in selecting the k atoms with largest coefficients of b (i.e. $\mathcal{T} = \text{supp}(b_{|k})$). Step 1 and 4 being essentially the same, the main differences lie in the way they design the solutions b and x at step 3 and 5. The choices made in CoSaMP, HTP and Subspace Pursuit are: i) b or x is the exact solution to the subproblem at stake, ii) b is an approximate solution found after a gradient step: $b = x - \eta\mathbf{\Phi}^*(\mathbf{\Phi}x - y)$, iii) x is $b_{|k}$.

To adapt these algorithms to the minimization of a generic differentiable loss in Problem (P1), one notices that $\mathbf{\Phi}^*(\mathbf{\Phi}x - y)$ is the gradient of $\frac{1}{2}\|\mathbf{\Phi}x - y\|_2^2$ and thus simply use ∇f instead. To tackle our problem of finding the zero of an operator, we naturally further replace ∇f with the operator $\mathbf{T}$. The common structure thus becomes:

Step 1: One identifies new candidate atoms ϕ_i by examining $\mathbf{T}(x)$.

Step 2: One defines an extended support $\mathcal{S}$ of size $2k$ or $3k$ or $4k$, by aggregating $\mathcal{T}$ and the labels of the atoms found at step 1.

Step 3: One designs b, a -possibly approximate- solution of Pb. (P3) on the extended support $\mathcal{S}$.

Step 4: One identifies in b the k "best" atoms and update the support $\mathcal{T}$ of size k accordingly.

Step 5: One designs x, a -possibly approximate- solution of Pb. (P3) on the support $\mathcal{T}$.

Algorithm 1 Generalized Subspace Pursuit

Require: $\mathbf{T}$, k .
Initialization: $x \leftarrow 0$, $\mathcal{S} \leftarrow \emptyset$.
repeat
  $\mathcal{G} \leftarrow \text{supp}(\mathbf{T}(x)_{|k})$, ▷ Step 1: select new directions.
  $\mathcal{S}_{old} \leftarrow \mathcal{S}$,
  $\mathcal{S} \leftarrow \mathcal{G} \cup \text{supp}(x)$, ▷ Step 2: set extended support.
  $b \in \mathcal{H}$ s.t. $\begin{cases} \text{supp}(b) \subseteq \mathcal{S} \\ \mathbf{T}(b)_{|\mathcal{S}} = 0. \end{cases}$ ▷ Step 3: solve on extended support.
  $\mathcal{T} \leftarrow \text{supp}(b_{|k})$, ▷ Step 4: set support.
  $x \in \mathcal{H}$ s.t. $\begin{cases} \text{supp}(x) \subseteq \mathcal{T} \\ \mathbf{T}(x)_{|\mathcal{T}} = 0. \end{cases}$ ▷ Step 5: solve on the support.
until $\mathcal{S}_{old} = \mathcal{S}$.
Output: x.

Algorithm 2 Generalized CoSaMP

Require: $\mathbf{T}$, k .
Initialization: $x \leftarrow 0$, $\mathcal{S} \leftarrow \emptyset$.
repeat
  $\mathcal{G} \leftarrow \text{supp}(\mathbf{T}(x)_{|2k})$, $(*)$
  $\mathcal{S}_{old} \leftarrow \mathcal{S}$,
  $\mathcal{S} \leftarrow \mathcal{G} \cup \text{supp}(x)$,
  $b \in \mathcal{H}$ s.t. $\begin{cases} \text{supp}(b) \subseteq \mathcal{S} \\ \mathbf{T}(b)_{|\mathcal{S}} = 0. \end{cases}$
  $\mathcal{T} \leftarrow \text{supp}(b_{|k})$,
  $x \leftarrow b_{|k}$. $(*)$
until $\mathcal{S}_{old} = \mathcal{S}$.
Output: x.

Algorithm 3 Generalized HTP

Require: $\mathbf{T}$, k, η . $(*)$
Initialization: $x \leftarrow 0$, $\mathcal{S} \leftarrow \emptyset$.
repeat
  $\mathcal{G} \leftarrow \text{supp}(\mathbf{T}(x)_{|k})$, $(*)$
  $\mathcal{S}_{old} \leftarrow \mathcal{S}$,
  $\mathcal{S} \leftarrow \mathcal{G} \cup \text{supp}(x)$,
  $b \leftarrow (\mathbf{I} - \eta\mathbf{T})(x)_{|\mathcal{S}}$, $(*)$
  $\mathcal{T} \leftarrow \text{supp}(b_{|k})$,
  $x \in \mathcal{H}$ s.t. $\begin{cases} \text{supp}(x) \subseteq \mathcal{T} \\ \mathbf{T}(x)_{|\mathcal{T}} = 0. \end{cases}$
until $\mathcal{S}_{old} = \mathcal{S}$.
Output: x.

Algorithm 4 Generalized Iterative Hard Thresholding

Require: $\mathbf{T}$, k, η, ε .
Initialization: $x \leftarrow 0$.
repeat
  $b \leftarrow (\mathbf{I} - \eta\mathbf{T})(x)$, $(*)$
  $\mathcal{T} \leftarrow \text{supp}(b_{|k})$,
  $x_{old} \leftarrow x$,
  $x \leftarrow b_{|k}$. $(*)$
until $\|x_{old} - x\| \leq \varepsilon$.
Output: x.

Generalized Subspace Pursuit (GSP) described in Algo. 1, *Generalized CoSaMP* (GCoSaMP) described in Algo. 2 and *Generalized HTP* (GHTP) described in Algo. 3 follow the structure just described. We pinpoint exactly steps one to five for *Generalized Subspace Pursuit* in Algo. 1. For the two subsequent algorithms, we only marked with a star the differences with *Generalized Subspace Pursuit*. As noted before, the difference besides the number of new directions selected lies in the choice for b and x. While *Generalized Subspace Pursuit* chooses the exact solution for both - which might be computationally costly -, *Generalized CoSaMP* and *Generalized HTP* both choose to make some approximations: *Generalized CoSaMP* does so in the last step $x = b_{|k}$, while *Generalized HTP* does

it in the third step: $b = (\mathbf{I} - \eta\mathbf{T})(x)_{|\mathcal{S}}$.

Let us discuss briefly the "exact solution" mentioned here. Given a finite set $\mathcal{R}$ the goal here is to find the best element of $\mathcal{H}$ with support in $\mathcal{R}$ complying with Problem (P3). As mentioned in Section 1, Problem (P3) may not have a solution, and naturally its restriction to elements supported in $\mathcal{R}$ might not as well. However, one can guarantee the existence of a solution for the slightly relaxed problem (used in GSP, GCoSaMP and GHTP):

$$\text{Find } z \in \mathcal{H} \text{ s.t. } \operatorname{supp}(z) \subseteq \mathcal{R} \text{ and } \mathbf{T}(z)_{|\mathcal{R}} = 0 \ . \tag{P4}$$

Note that when $\mathbf{T}$ is the differential of a convex function, this corresponds to a minimizer among the vectors supported in $\mathcal{R}$ which does exist under mild conditions.

Let us finally mention a fourth generalized greedy algorithm: *Generalized Iterative Hard Thresholding* (GIHT) described in Algo. 4. This simpler algorithm does not focus on finding the support but rather directly x. It does so by selecting the k largest entries of $b = (\mathbf{I} - \eta\mathbf{T})(x)$, thus essentially bypassing the concept of extended support used above. Its name stems from the fact that selecting the k largest entries of an element z of $\mathcal{H}$ i.e. $z_{|k}$ is also called hard-thresholding. Although simpler in essence, we will show that the same convergence analysis may be conducted for *Generalized Iterative Hard Thresholding* and the three other algorithms described above.

2.2 When are these algorithms well-defined ?

Before discussing the convergence of these algorithms, one must first ensure that they are well-defined. Since $\mathbf{T}$ is an operator on $\mathcal{H}$ with full domain, GIHT as well as steps 1, 2 and 4 in GSP, GCoSaMP and GHTP are well-defined. So are the approximations used in step 5 of GCoSaMP ($x = b_{|k}$) or step 3 of GHTP ($b = (\mathbf{I} - \eta\mathbf{T})(x)_{|\mathcal{S}}$). The remaining question is thus to ensure the existence of the solution of the subproblems in steps 3 and 5 of GSP. More precisely, for GSP one needs to ensure for all sets $\mathcal{R}$ of size smaller or equal than $2k$, there exists an element z in $\mathcal{H}$, verifying i) $\mathcal{R}$ contains the support of z and ii) $\mathcal{R}$ is disjoint from the support of $\mathbf{T}(z)$ (i.e. $\operatorname{supp}(z) \subseteq \mathcal{R}$ and $\mathbf{T}(z)_{|\mathcal{R}} = 0$). Although this is not the case in general, this happens for example when $\mathbf{I} - \mathbf{T}$ is a contraction. We will see in Section 4.1 that we need a less strong property to ensure that the algorithms are well-defined and guarantee their convergence at the same time. This property is called the *Restricted Diagonal Property*.

3 The Restricted Diagonal Property

In this section, we present the *Restricted Diagonal Property* and its links to the criteria used in the literature.

3.1 The Restricted Diagonal Property

To ensure the convergence of algorithms 1 to 4, we ask that the operator $\mathbf{T}$ has a specific property, namely it looks like a diagonal operator bounded away from zero locally on k-sparse vectors. To this end, we introduce $\mathcal{D}_1$ the set of diagonal operators bounded away by 1:

$$\mathcal{D}_1 = \left\{ \mathbf{D} : \begin{array}{ccc} \mathcal{H} & \to & \mathcal{H}, \\ x & \mapsto & \sum_i d_i x_i e_i \end{array} \quad \text{s.t. } \forall x \, \|\mathbf{D}x\| \geq \|x\| \right\}. \tag{4}$$

Definition 1 (Uniform Restricted Diagonal Property)**.** $\mathbf{T}$ *is said to have the* Uniform Restricted Diagonal Property *(URDP) of order k if there exists $\alpha_k > 0$ and a diagonal operator $\mathbf{D}_k$ in $\mathcal{D}_1$ such that*

$$\forall (x, y) \in \mathcal{H}^2, \ \operatorname{card}(\operatorname{supp}(x, y)) \leqslant k \Rightarrow \|\mathbf{T}(x) - \mathbf{T}(y) - \mathbf{D}_k(x - y)\| \leqslant \alpha_k \|x - y\| \, . \tag{5}$$

In other words, the $\mathbf{T}$ has the *Uniform Restricted Diagonal Property* of order k when its k-sparse increments look diagonal. Depending on the algorithm, the existence and convergence may be proved with a less stringent version, the *Restricted Diagonal Property*, that allows the diagonal operator to vary on different subspaces of size k:

Definition 2 (Restricted Diagonal Property)**.** $\mathbf{T}$ *is said to have the* Restricted Diagonal Property *(RDP) of order* k *if there exists* $\alpha_k > 0$ *such that for all subsets* $\mathcal{S}$ *of* $\mathbb{N}$ *of cardinal at most* k*, there exists a diagonal operator* $\mathbf{D}_{\mathcal{S}} \in \mathcal{D}_1$ *such that*

$$\forall (x,y) \in \mathcal{H}^2,\ \mathrm{supp}(x,y) \subseteq \mathcal{S} \Rightarrow \|\mathbf{T}(x) - \mathbf{T}(y) - \mathbf{D}_{\mathcal{S}}(x-y)\| \leqslant \alpha_k \|x-y\| . \tag{6}$$

Note that by definition $\alpha_k \leqslant \alpha_{k+1}$.

Remark 3. *When* $\mathbf{T}$ *has the Restricted Diagonal Property (resp. Uniform RDP) of order* $2k$ *then for all* k*-sparse vectors i.e. when* $\|x\|_0 \leqslant k$ *and* $\|y\|_0 \leqslant k$*, we have* $\|\mathbf{T}(x) - \mathbf{T}(y) - \mathbf{D}(x-y)\| \leqslant \alpha_{2k} \|x-y\|$*, with* $\mathbf{D} = \mathbf{D}_{\mathrm{supp}(x,y)}$ *(resp.* $\mathbf{D} = \mathbf{D}_{2k}$*).*

Remark 4. *If in addition* $\alpha_{2k} < 1$*, then* $\mathbf{T}$ *is injective on the set of* k*-sparse vectors. Indeed, since the diagonal operators involved are bounded away from zero by one, we have in this case:* $\|\mathbf{T}(x) - \mathbf{T}(y)\| \geq (1-\alpha_{2k}) \|x-y\|$ *when* $\|x\|_0 \leqslant k$ *and* $\|y\|_0 \leqslant k$*.*

Remark 5. *When* $\mathbf{T}$ *has the* Uniform Restricted Diagonal Property *with* $\alpha_{2k} < 1$ *and* $\mathbf{D}$ *bounded, one also has*

$$(1-\alpha_{2k}) \|x-y\| \leqslant \|\mathbf{T}(x) - \mathbf{T}(y)\| \leqslant (\|\mathbf{D}_{2k}\| + \alpha_{2k}) \|x-y\| \ ,$$

when $\|x\|_0 \leqslant k$ *and* $\|y\|_0 \leqslant k$*. If* $\mathbf{T}$ *is linear, this means that* $\mathbf{T}$ *is RIP with constant* $\max(\alpha_{2k}, \|\mathbf{D}_{2k}\| - 1 + \alpha_{2k})$*. Thus for linear operators, the operators having the* Uniform Restricted Diagonal Property *of order* $2k$ *with a bounded* $\mathbf{D}$ *form a subset of the RIP operators of order* k*. Note that this is not the case any more for the operators having* Restricted Diagonal Property *which thus concerns a different set of operators.*

3.2 Characterization of the Uniform Restricted Diagonal Property

To further link the URDP to properties used in the literature, we first give a characterization expressing the differences $\mathbf{T}(x) - \mathbf{T}(y)$ rather than the differences against diagonal $\mathbf{T}(x) - \mathbf{T}(y) - \mathbf{D}(x-y)$:

Theorem 6. *Assume that* $D \in \mathcal{D}_1$ *is bounded in the following sense*

$$|||D|||_k \stackrel{def}{=} \sup_{\{x \neq y,\ \mathrm{card}(\mathrm{supp}(x,y)) \leq k\}} \frac{\|D(x-y)\|}{\|x-y\|} = \sup_{\{x \neq 0,\ \mathrm{card}(\mathrm{supp}(x)) \leq k\}} \frac{\|D(x)\|}{\|x\|} < \infty. \tag{7}$$

we have:

1. *If* $\beta\mathbf{T}$ *is URDP of order* k *for* D*, with* $\alpha_k < 1$ *and* $\beta > 0$ *then*

$$\forall (x,y) \in \mathcal{H}^2,\ \mathrm{card}(\mathrm{supp}(x,y)) \leq k \Rightarrow \left\{ \begin{array}{l} \|\mathbf{T}(x) - \mathbf{T}(y)\| \leq \frac{|||D|||_k + \alpha_k}{\beta} \|x-y\| \\ \langle \mathbf{T}(x) - \mathbf{T}(y), D(x-y) \rangle \geq \frac{1-\alpha_k}{\beta} \|x-y\|^2 . \end{array} \right. \tag{8}$$

2. *If there exists* $(m, L) \in \mathbb{R}^2$ *such that* $0 < m$ *and* $0 \leq |||D|||_k^2 - \frac{m^2}{L^2} < 1$ *and*

$$\forall (x,y) \in \mathcal{H}^2,\ \mathrm{card}(\mathrm{supp}(x,y)) \leq k \Rightarrow \left\{ \begin{array}{l} \|\mathbf{T}(x) - \mathbf{T}(y)\| \leq L \|x-y\| \\ \langle \mathbf{T}(x) - \mathbf{T}(y), D(x-y) \rangle \geq m \|x-y\|^2 . \end{array} \right. \tag{9}$$

then $(\beta\mathbf{T})$ *is URDP of order* k *for* D*, with* $\alpha_k = |||D|||_k^2 - \frac{m^2}{L^2}$ *and* $\beta = \frac{m}{L^2}$*.*

Remark 7. *Consider the assumption* $\frac{m^2}{L^2} \leq |||D|||_k^2 \leq \frac{m^2}{L^2} + 1$ *made in Theorem 6. Notice that Eq.* (9) *implies that* $m \leq L|||D|||_k$ *so that the left inequality is always true. We also pinpoint that* $|||D|||_k^2 \leq \frac{m^2}{L^2} + 1$ *always holds for* $D = \mathbf{I}$*.*

The proof is postponed to A. The conclusion is that being URDP of order k for a bounded D and $\alpha_k < 1$ is up to a scaling equivalent to $\mathbf{T}$ having a Lipschitz property and a lower bound on the scalar product $\langle \mathbf{T}(x) - \mathbf{T}(y), D(x-y) \rangle$, both properties holding on the couples (x,y) such that $\mathrm{card}(\mathrm{supp}(x,y)) \leqslant k$.

Remark 8. *A similar theorem hold for* $\mathbf{T}$ *having the RDP and further assuming that there exist a uniform bounded of the type of Eq.* (7) *on the matrices* $\mathbf{D}_{\mathcal{S}}$ *in Definition 2.*

3.3 Uniform Restricted Diagonal Property with the identity, RIP and other criteria

Let us consider $\mathbf{T} = \nabla f$ with $f : \mathcal{H} \to \mathbb{R}^+$ smooth. Theorem 6 shows that a scaled version of ∇f is URDP of order k, with $\alpha_k < 1$ and $\mathbf{D}$ the identity if and only if ∇f is Lipschitz and f has a strong convexity property on the couples (x, y) such that $\text{card}(\text{supp}(x, y)) \leqslant k$. This is the same content as the *Restricted Strong Smoothness* and *Restricted Strong Convexity* developed in (Bahmani et al., 2013; Yuan et al., 2014) for the resolution of Problem (P1).

Furthermore, for the classical case of the quadratic loss i.e. $f(x) = \|Ax - z\|_2^2$, one also verifies easily that Theorem 6 shows that a scaled version of ∇f is URDP of order k, with $\alpha_k < 1$ and $\mathbf{D}$ the identity if and only if A is RIP (Candès et al., 2006).

Let us also emphasize that Theorem 6 shows that Uniform RDP with $\mathbf{D}$ different from identity encompasses a greater set of functions than the classical notions seen above. Indeed as soon as D has at least one negative eigenvalue the second inequality in Eq. (8) does not yield convexity anymore, and the Lipschitz characteristic is preserved only if $\mathbf{D}$ is bounded.

4 Theoretical Guarantees

In this section, we present the theoretical guarantees we obtain for Algorithms 1 to 4. They rely on $\mathbf{T}$ having the *Restricted Diagonal Property*. We show how this ensures that the algorithms are well-defined and derive the error bounds.

4.1 The Restricted Diagonal Property and existence of solutions to Problems (P3) and (P4)

Proposition 9.

1. *Problem* (P3) *has at most one solution when* $\mathbf{T}$ *has the* Restricted Diagonal Property *of order* $2k$ *with* $\alpha_{2k} < 1$.
2. *Problem* (P4) *has at least one solution for all sets* $\mathcal{R}$ *of cardinal at most* k *when* $\mathbf{T}$ *has the* Restricted Diagonal Property *of order* k *with* $\alpha_k < 1$.

Examining the size of the support set when Problem (P4) appears at step 3 or 5 of the algorithm, one deduces that

Corollary 10.

- *GSP (Algo. 1) is well-defined when* $\mathbf{T}$ *is an operator having the* Restricted Diagonal Property *of order* $2k$ *with* $\alpha_{2k} < 1$.
- *GCoSaMP (Algo 2) is well-defined when* $\mathbf{T}$ *is an operator having the* Restricted Diagonal Property *of order* $3k$ *with* $\alpha_{3k} < 1$.
- *GHTP (Algo 3) is well-defined when* $\mathbf{T}$ *is an operator having the* Restricted Diagonal Property *of order* k *with* $\alpha_k < 1$.

Let us now prove Proposition 9:

Proof. 1. As noted in Remark 4, if $\mathbf{T}$ has the RDP of order $2k$ with $\alpha_{2k} < 1$, then $\mathbf{T}$ is injective on the set of k-sparse vector and thus Problem (P3) has at most one solution.

2. Let $\mathcal{R}$ be of cardinal at most k. Since $\mathbf{T}$ has the RDP of order k with $\alpha_k < 1$, there exists $\mathbf{D}_\mathcal{R}$ in $\mathcal{D}_1$ such that Eq. (6) holds. Noting $\mathbf{D}_\mathcal{R}(z) = \sum_{i\in\mathcal{R}} d_i z_i e_i$, we have

$$\begin{aligned}
\|\mathbf{T}(x) - \mathbf{T}(y) - \mathbf{D}_\mathcal{R}(x-y)\|^2 &= \sum_{i\notin\mathcal{R}} [\mathbf{T}(x)_i - \mathbf{T}(y)_i]^2 + \sum_{i\in\mathcal{R}} [\mathbf{T}(x)_i - \mathbf{T}(y)_i - d_i(x_i - y_i)]^2 \,, \\
&\geq \sum_{i\in\mathcal{R}} [\mathbf{T}(x)_i - \mathbf{T}(y)_i - d_i(x_i - y_i)]^2 \,, \\
&\geq \sum_{i\in\mathcal{R}} d_i^2 \left[\frac{\mathbf{T}(x)_i}{d_i} - \frac{\mathbf{T}(y)_i}{d_i} - (x_i - y_i)\right]^2 \,, \\
&\geq \min_{j\in\mathcal{R}} d_j^2 \sum_{i\in\mathcal{R}} \left[\frac{\mathbf{T}(x)_i}{d_i} - \frac{\mathbf{T}(y)_i}{d_i} - (x_i - y_i)\right]^2 \,, \\
&\geq \left\|\mathbf{T}^\mathcal{R}(x)_i - \mathbf{T}^\mathcal{R}(y)_i - \mathbf{P}_\mathcal{R}(x-y)\right\|^2 \,,
\end{aligned}$$

where we have defined $\mathbf{T}^\mathcal{R} : \mathcal{H} \to \mathcal{H}, x \mapsto \sum_{i\in\mathcal{R}} \frac{\mathbf{T}(x)_i}{d_i} e_i$ and we have used that $\min_{j\in\mathcal{R}} d_j^2 \geq 1$ (since $\mathbf{D}_\mathcal{R}$ in $\mathcal{D}_1$). Combining this with Eq. (6), we obtain:

$$\forall (x,y) \in \operatorname{span}(e_i, i\in\mathcal{R})^2,\ \left\|\mathbf{T}^\mathcal{R}(x) - \mathbf{T}^\mathcal{R}(y) - (x-y)\right\| \leqslant \alpha_k \|x-y\| \,. \tag{10}$$

In other words $\mathbf{T}^\mathcal{R} - \mathbf{I}$ is a contraction from $\operatorname{span}(e_i, i\in\mathcal{R})$ into itself. $\operatorname{span}(e_i, i\in\mathcal{R})$ being a subspace of $\mathcal{H}$ of finite dimension, it is also a Banach space. The contraction thus has a fixed point (Cegielski, 2013). There exists x_0 in $\operatorname{span}(e_i, i\in\mathcal{R})$ such that $\mathbf{T}^\mathcal{R}(x_0) - \mathbf{I}(x_0) = x_0$ i.e. $\mathbf{T}^\mathcal{R}(x_0) = \sum_{i\in\mathcal{R}} \frac{\mathbf{T}(x_0)_i}{d_i} e_i = 0$. Using again that $|d_i| \geq 1$, we have: $\mathbf{T}(x_0)_i = 0$, for i in $\mathcal{R}$ and thus x_0 solves Problem (P4).

□

4.2 Error bounds

The (Uniform-)Restricted Diagonal Property allows to guarantee the good behavior of the different algorithms presented in Section 2. As is the case for the original versions such as CoSaMP, the guarantee is an error bound divided into two parts: one is vanishing exponentially fast while the second refers to an *incompressible* error as seen in (Needell and Tropp, 2009).

Let us state first the error bound for Generalized Support Pursuit:

Theorem 11. *Denote by $x^\star$ any k-sparse vector and α^S the unique real root of $g(x) = x^3 + x^2 + 7x - 1$ ($\alpha^S < 1$). If there exists $\rho > 0$ such that $\rho\mathbf{T}$ has the* Restricted Diagonal Property *of order $3k$ with $\alpha_{3k} \leqslant \alpha^S$. Then x^N, the N-th iterate of GSP (Algo. 1), verifies*

$$\left\|x^N - x^\star\right\| \leqslant \frac{1}{2^N} \left\|x^0 - x^\star\right\| + 12\rho \left\|\mathbf{T}(x^\star)_{|2k}\right\| \,. \tag{11}$$

Note that the *incompressible* error $\left\|\mathbf{T}(x^\star)_{|2k}\right\|$ vanishes if $x^\star$ is a solution of Problem (P3), assuring that GSP converges exponentially fast to it. Otherwise, the iterates are guaranteed to approach the "best" k-sparse vector in the sense that it is the one for which the best $2k$-sparse approximation of $T(x)$ is the smallest.

A similar bound holds for GCoSaMP:

Theorem 12. *Denote by $x^\star$ any k-sparse vector and $\alpha^C = \frac{2}{\sqrt{3}} - 1$. If there exists $\rho > 0$ such that $\rho\mathbf{T}$ has the* Restricted Diagonal Property *of order $4k$ with $\alpha_{4k} \leqslant \alpha^C$. Then x^N, the N-th iterate of Generalized CoSaMP (Algo. 2), verifies*

$$\left\|x^N - x^\star\right\| \leqslant \frac{1}{2^N} \left\|x^0 - x^\star\right\| + 12\rho \left\|\mathbf{T}(x^\star)_{|3k}\right\| \,. \tag{12}$$

Notice that a similar theorem was proposed in Bahmani et al. (2013) for the special case where $\mathbf{T} = \nabla f$ is URDP ($\mathbf{D} = \mathbf{I}$).

By contrast with GSP and GCoSaMP, we require the *Uniform Restricted Diagonal Property* to hold with the identity matrix for GHTP and GIHT:

Theorem 13. *Denote by $x^\star$ any k-sparse vector. Assume that $\frac{3}{4} < \eta < \frac{5}{4}$. If $\mathbf{T}$ has the* Uniform Restricted Diagonal Property *of order $2k$ with $\mathbf{D}_{2k} = \mathbf{I}$ and $\alpha_{2k} \leqslant \alpha^H = 7 - 2\sqrt{11}$. Then x^N, the N-th iterate of Generalized HTP (Algo. 3), verifies*

$$\left\|x^N - x^\star\right\| \leqslant \frac{1}{2^N} \left\|x^0 - x^\star\right\| + 2\frac{(1+2\eta)(1-\alpha_{2k})+4}{(1-\alpha_{2k})^2} \left\|\mathbf{T}(x^\star)_{|2k}\right\| \ . \tag{13}$$

A similar bound has been shown in (Yuan et al., 2014). Notice that there exists a variant of HTP where one can select $l < k$ new directions at each iteration instead of k (line 4 of Algo. 3), which has been proved to be beneficial in Jain et al. (2011).

For GIHT, the error bound reads:

Theorem 14. *Denote by $x^\star$ any k-sparse vector and $\alpha^\eta = \frac{1-4|\eta-1|}{4(1+|\eta-1|)}$. Assume that $\frac{3}{4} < \eta < \frac{5}{4}$ so that $\alpha^\eta > 0$. If $\mathbf{T}$ has the* Uniform Restricted Diagonal Property *of order $2k$ with $\mathbf{D}_{2k} = \mathbf{I}$ and $\alpha_{2k} \leqslant \alpha^\eta$. Then x^N, the N-th iterate of Generalized IHT (Algo. 4), verifies*

$$\left\|x^N - x^\star\right\| \leqslant \frac{1}{2^N} \left\|x^0 - x^\star\right\| + 4\eta \left\|\mathbf{T}(x^\star)_{|3k}\right\| \ . \tag{14}$$

The guarantees derived for the algorithms differ for the RDP bounds (α^S, α^C, α^H, α^η) and factor in front of the *incompressible* error term. However, the fundamental difference between the algorithms does not lie there, but rather in the possibility to consider RDP operators (for GSP and GCoSaMP) rather than the Uniform-RDP with the identity (for GIHT and GHTP).

Indeed, as we have seen in Section 3, Uniform-RDP with the identity relates to the properties previously developed in the literature to control greedy algorithms for minimizing functions i.e. when $\mathbf{T} = \nabla f$. More precisely it is equivalent to a Lipschitz property on ∇f combined with a strong convexity property on f on couples of sparse vectors. This is precisely what is used in (Bahmani et al., 2013) for example to prove a bound similar to Theorem 12 for GCoSaMP for minimizing a function.

By contrast, GSP and GCoSaMP only require the *Restricted Diagonal Property* which, for $\mathbf{T} = \nabla f$, does not imply Lipschitz properties nor convexity: the diagonal operators involved may change with subspaces, and include negative diagonal values... (see also Section 6 for details). Thus, Theorems 11 and 12 also show that GSP and GCoSaMP can naturally handle a larger class of Problems than GIHT and GHTP, namely they can be used for finding k-sparse extrema of functions with either convex, or concave or neither of both properties on k-sparse vectors.

Additionally Theorems 11 to 14 extend the scope of greedy algorithms to finding k-sparse zeros of operators (Problem P3).

5 Proof

In this section, we show how to derive the error bound for GSP (Theorem 11). The proof has a similar structure as the ones in (Needell and Tropp, 2009) or (Bahmani et al., 2013). It relies on the two lemmas described in Section 5.1 that use the RDP to control the differences $\mathbf{T}(x) - \mathbf{T}(y)$ on $\text{supp}(x, y)$ and its complement when x and y are sparse. With those, we are able to control the error made in the three main actions taken by the algorithms. We state the corresponding three Lemmas in Section 5.2. We then proceed to the complete proof of Theorem 11 by successively bounding the error made at each step of the algorithm in Section 5.3.

The proofs of the bounds for GCoSaMP, GHTP and GIHT (Theorems 12 to 14) are similar and postponed to B.

Note that running GSP with $\rho\mathbf{T}$ or $\mathbf{T}$ yields the same iterates, therefore the error bound may be proved by setting $\mathbf{T}' = \rho\mathbf{T}$ and showing that if $\mathbf{T}'$ has the *Restricted Diagonal Property* of order $3k$ with $\alpha_{3k} \leqslant \alpha^S$ then

$$\left\|x^N - x^\star\right\| \leqslant \tfrac{1}{2^N} \left\|x^0 - x^\star\right\| + 12 \left\|\mathbf{T}'(x^\star)_{|2k}\right\| \ . \tag{15}$$

Here we assume that $x^\star$ is a k-sparse vector in $\mathcal{H}$.

5.1 The Two Pillars

Let us show that the RDP allows to control the difference $\mathbf{T}(x) - \mathbf{T}(y)$ of any two sparse vectors both on the union of their support $\operatorname{supp}(x, y)$ (Lemma 15) and on its complement (Lemma 16). These two lemmas are the pillar of the analysis of the error bounds.

Lemma 15 characterizes how the energy of $\mathbf{T}(x) - \mathbf{T}(y)$ spreads on subsets of the support $\operatorname{supp}(x, y)$.

Lemma 15. *Assume that* $\mathbf{T}$ *has the* Restricted Diagonal Property *of order* k *with* α_k*. Let* $x, y \in \mathcal{H}$ *be such that* $\operatorname{card}(\operatorname{supp}(x, y)) \leqslant k$*, we have:*

$$\forall \mathcal{R}' \subseteq \mathbb{N}, \quad \left\|(\mathbf{T}(x) - \mathbf{T}(y))_{|\mathcal{R}'}\right\| \geqslant (1 - \alpha_k) \left\|(x - y)_{|\mathcal{R}'}\right\| - \alpha_k \left\|(x - y)_{|\operatorname{supp}(x,y)\setminus\mathcal{R}'}\right\| \tag{16}$$

Proof. Let $\mathcal{S} = \operatorname{supp}(x, y)$ then we have,

$$\begin{aligned}
\left\|(\mathbf{T}(x) - \mathbf{T}(y))_{|\mathcal{R}'}\right\| &= \left\|(\mathbf{T}(x) - \mathbf{T}(y) - \mathbf{D}_\mathcal{S}(x - y) + \mathbf{D}_\mathcal{S}(x - y))_{|\mathcal{R}'}\right\| \\
&= \left\|(\mathbf{D}_\mathcal{S}(x - y))_{|\mathcal{R}'} - (\mathbf{T}(x) - \mathbf{T}(y) - \mathbf{D}_\mathcal{S}(x - y))_{|\mathcal{R}'}\right\| \\
&\geqslant \left\|(\mathbf{D}_\mathcal{S}(x - y))_{|\mathcal{R}'}\right\| - \left\|(\mathbf{T}(x) - \mathbf{T}(y) - \mathbf{D}_\mathcal{S}(x - y))_{|\mathcal{R}'}\right\| \\
&\geqslant \left\|\mathbf{D}_\mathcal{S}((x - y)_{|\mathcal{R}'})\right\| - \left\|(\mathbf{T}(x) - \mathbf{T}(y) - \mathbf{D}_\mathcal{S}(x - y))\right\| \\
&\geqslant \left\|(x - y)_{|\mathcal{R}'}\right\| - \alpha_k \left\|x - y\right\| \ , \quad \text{since } \left\|\mathbf{D}_\mathcal{S}(z)\right\| \geq \left\|z\right\| \\
&\geqslant (1 - \alpha_k) \left\|(x - y)_{|\mathcal{R}'}\right\| - \alpha_k \left\|(x - y)_{|\mathcal{S}\setminus\mathcal{R}'}\right\| \ .
\end{aligned}$$

□

The following lemma controls the energy of $\mathbf{T}(x) - \mathbf{T}(y)$ outside of the supports of x and y.

Lemma 16. *Assume that* $\mathbf{T}$ *has the* Restricted Diagonal Property *of order* k *with* α_k*. Let* $x, y \in \mathcal{H}$ *be such that* $\operatorname{card}(\operatorname{supp}(x, y)) \leqslant k$*, we have:*

$$\forall \mathcal{F} \subseteq \mathbb{N} \ s.t \ \mathcal{F} \cap \operatorname{supp}(x, y) = \emptyset, \quad \left\|(\mathbf{T}(x) - \mathbf{T}(y))_{|\mathcal{F}}\right\| \leqslant \alpha_k \left\|x - y\right\| \ . \tag{17}$$

Proof. Let $\mathcal{S} = \operatorname{supp}(x, y)$, since $\mathcal{F} \cap \mathcal{S} = \emptyset$ then $\mathbf{D}_\mathcal{S}(x - y))_{|\mathcal{F}} = 0$ and

$$\left\|(\mathbf{T}(x) - \mathbf{T}(y))_{|\mathcal{F}}\right\| = \left\|(\mathbf{T}(x) - \mathbf{T}(y) - \mathbf{D}_\mathcal{S}(x - y))_{|\mathcal{F}}\right\| \leqslant \left\|\mathbf{T}(x) - \mathbf{T}(y) - \mathbf{D}_\mathcal{S}(x - y)\right\| \leqslant \alpha_k \left\|x - y\right\| \ .$$

□

5.2 Error Bounds for the Main Steps

Here we derive error bounds for the following actions

1. changing support (steps 1 and 2 of the algorithms),
2. computing an exact solution of a subproblem (steps 3 and 5 of GSP, step 3 of GCoSaMP, step 5 of GHTP),
3. computing a k-sparse approximation (steps 5 of GCoSaMP, step 3 of GHTP).

In the first lemma, we consider the influence of merging two sets when seeking for new directions. One set is usually the current support set $\mathcal{T}$ while the other is derived from the support of $\mathbf{T}(x)$.

Lemma 17 (Changing supports)**.** *Assume that* $\mathbf{T}$ *has the* Restricted Diagonal Property *of order* k *with* $\alpha_k < 1$*. Let* $\mathcal{R}$ *and* $\mathcal{S}$ *be subsets of* $\mathbb{N}$ *and* $x, y \in \mathcal{H}$*. Assume that* $\operatorname{supp}(x,y) \subseteq \mathcal{R}$*,* $\operatorname{card}(\mathcal{R}) \leqslant k$*, and* $\left\|\mathbf{T}(x)_{|\mathcal{R}}\right\| \leqslant \left\|\mathbf{T}(x)_{|\mathcal{S}}\right\|$*, then*

$$\left\|(x-y)_{|\mathcal{S}^c}\right\| \leqslant \tfrac{2\alpha_k}{1-\alpha_k}\left\|x-y\right\| + \tfrac{1}{1-\alpha_k}\left(\left\|\mathbf{T}(y)_{|\mathcal{R}\setminus\mathcal{S}}\right\| + \left\|\mathbf{T}(y)_{|\mathcal{S}\setminus\mathcal{R}}\right\|\right) \ .$$

If additionally $\operatorname{supp}(x) \subseteq \mathcal{S}$*, we have*

$$\left\|y_{|\mathcal{S}^c}\right\| \leqslant \tfrac{2\alpha_k}{1-\alpha_k}\left\|x-y\right\| + \tfrac{1}{1-\alpha_k}\left(\left\|\mathbf{T}(y)_{|\mathcal{R}\setminus\mathcal{S}}\right\| + \left\|\mathbf{T}(y)_{|\mathcal{S}\setminus\mathcal{R}}\right\|\right) \ .$$

Proof. $\left\|\mathbf{T}(x)_{|\mathcal{R}}\right\| \leqslant \left\|\mathbf{T}(x)_{|\mathcal{S}}\right\|$ implies $\left\|\mathbf{T}(x)_{|\mathcal{R}\setminus\mathcal{S}}\right\| \leqslant \left\|\mathbf{T}(x)_{|\mathcal{S}\setminus\mathcal{R}}\right\|$.

Using the triangle inequality and Lemma 15, we have,

$$\begin{aligned}\left\|\mathbf{T}(x)_{|\mathcal{R}\setminus\mathcal{S}}\right\| &\geqslant \left\|(\mathbf{T}(x)-\mathbf{T}(y))_{|\mathcal{R}\setminus\mathcal{S}}\right\| - \left\|\mathbf{T}(y)_{|\mathcal{R}\setminus\mathcal{S}}\right\| \\ &\geqslant (1-\alpha_k)\left\|(x-y)_{|\mathcal{R}\setminus\mathcal{S}}\right\| - \alpha_k\left\|(x-y)_{|\operatorname{supp}(x,y)\setminus(\mathcal{R}\setminus\mathcal{S})}\right\| - \left\|\mathbf{T}(y)_{|\mathcal{R}\setminus\mathcal{S}}\right\| \\ &\geqslant (1-\alpha_k)\left\|(x-y)_{|\mathcal{R}\setminus\mathcal{S}}\right\| - \alpha_k\left\|x-y\right\| - \left\|\mathbf{T}(y)_{|\mathcal{R}\setminus\mathcal{S}}\right\| \ .\end{aligned}$$

Moreover since $\operatorname{supp}(x,y) \cap (\mathcal{S}\setminus\mathcal{R}) = \emptyset$, Lemma 16 yields

$$\begin{aligned}\left\|\mathbf{T}(x)_{|\mathcal{S}\setminus\mathcal{R}}\right\| &\leqslant \left\|(\mathbf{T}(x)-\mathbf{T}(y))_{|\mathcal{S}\setminus\mathcal{R}}\right\| + \left\|\mathbf{T}(y)_{|\mathcal{S}\setminus\mathcal{R}}\right\| \\ &\leqslant \alpha_k\left\|x-y\right\| + \left\|\mathbf{T}(y)_{|\mathcal{S}\setminus\mathcal{R}}\right\| \ .\end{aligned}$$

Combining these two inequalities, we obtain

$$\alpha_k\left\|x-y\right\| + \left\|\mathbf{T}(y)_{|\mathcal{S}\setminus\mathcal{R}}\right\| \geqslant (1-\alpha_k)\left\|(x-y)_{|\mathcal{R}\setminus\mathcal{S}}\right\| - \alpha_k\left\|x-y\right\| - \left\|\mathbf{T}(y)_{|\mathcal{R}\setminus\mathcal{S}}\right\| \ .$$

Noting that $(x-y)_{|\mathcal{R}\setminus\mathcal{S}} = (x-y)_{|\mathcal{S}^c}$ because $\operatorname{supp}(x,y) \subseteq \mathcal{R}$ and $\alpha_k < 1$, we conclude

$$\left\|(x-y)_{|\mathcal{S}^c}\right\| \leqslant \tfrac{2\alpha_k}{1-\alpha_k}\left\|x-y\right\| + \tfrac{1}{1-\alpha_k}\left(\left\|\mathbf{T}(y)_{|\mathcal{R}\setminus\mathcal{S}}\right\| + \left\|\mathbf{T}(y)_{|\mathcal{S}\setminus\mathcal{R}}\right\|\right) \ .$$

□

The second lemma controls the distance from a k-sparse vector to a solution of the subproblem (P4).

Lemma 18 (Control of the distance to an exact solution of a subproblem)**.** *Assume* $\mathcal{R}$ *is a subset of* $\mathbb{N}$ *of cardinal at most* l*,* b *solves Problem* (P4)*, and* $\mathbf{T}$ *has the* Restricted Diagonal Property *of order* $k+l$ *with* $\alpha_{k+l} < 1$*.* $\forall x \in \mathcal{H}$ *k-sparse, we have*

$$\left\|(x-b)_{|\mathcal{R}}\right\| \leqslant \tfrac{1}{1-\alpha_{k+l}}\left\|\mathbf{T}(x)_{|\mathcal{R}}\right\| + \tfrac{\alpha_{k+l}}{1-\alpha_{k+l}}\left\|x_{|\mathcal{R}^c}\right\| \ .$$

Proof. Lemma 15 shows

$$\begin{aligned}\left\|(\mathbf{T}(x)-\mathbf{T}(b))_{|\mathcal{R}}\right\| &\geqslant (1-\alpha_{k+l})\left\|(x-b)_{|\mathcal{R}}\right\| - \alpha_{k+l}\left\|(x-b)_{|\operatorname{supp}(x,b)\setminus\mathcal{R}}\right\| \\ \left\|(\mathbf{T}(x)-\mathbf{T}(b))_{|\mathcal{R}}\right\| &\geqslant (1-\alpha_{k+l})\left\|(x-b)_{|\mathcal{R}}\right\| - \alpha_{k+l}\left\|(x-b)_{|\mathcal{R}^c}\right\|\end{aligned}$$

Using $\operatorname{supp}(b) \subseteq \mathcal{R}$ and $\mathbf{T}(b)_{|\mathcal{R}} = 0$, we obtain

$$\left\|\mathbf{T}(x)_{|\mathcal{R}}\right\| \geqslant (1-\alpha_{k+l})\left\|(x-b)_{|\mathcal{R}}\right\| - \alpha_{k+l}\left\|x_{|\mathcal{R}^c}\right\| \ .$$

The result follows from $\alpha_{k+l} < 1$. □

The two previous lemmas use the RDP. By contrast, the next lemma does not, it is a result of linear algebra.

Lemma 19 (k-sparse approximation)**.** *Assume that* x *is* k*-sparse and* $\operatorname{supp}(y) \subseteq \mathcal{S}$*, we have*

$$\left\|x-y_{|k}\right\| \leqslant 2\left\|(x-y)_{|\mathcal{S}}\right\| + \left\|x_{|\mathcal{S}^c}\right\| \ .$$

Proof. The proof is a direct application of the k-sparse approximation,

$$\begin{aligned}\left\|x-y_{|k}\right\| &\leqslant \left\|(x-y_{|k})_{|\mathcal{S}}\right\| + \left\|(x-y_{|k})_{|\mathcal{S}^c}\right\| \\ &\leqslant \left\|x_{|\mathcal{S}}-y_{|k}\right\| + \left\|x_{|\mathcal{S}^c}\right\| \\ &\leqslant \left\|x_{|\mathcal{S}}-y\right\| + \left\|y-y_{|k}\right\| + \left\|x_{|\mathcal{S}^c}\right\| \ .\end{aligned}$$

Since $y_{|k}$ is the best k-sparse approximation of y and x is k-sparse, we have $\left\|y-y_{|k}\right\| \leqslant \left\|x_{|\mathcal{S}}-y\right\|$. So $\left\|x-y_{|k}\right\| \leqslant 2\left\|x_{|\mathcal{S}}-y\right\| + \left\|x_{|\mathcal{S}^c}\right\| = 2\left\|(x-y)_{|\mathcal{S}}\right\| + \left\|x_{|\mathcal{S}^c}\right\|$. □

5.3 Proof of Theorem 11 (error bound on GSP)

With these five lemmas at hand, one can prove Theorems 11 to 14. We detail here the proof for GSP and postpone the proofs for the other algorithms in B.

The proof proceeds by proving an error bound at each step of the algorithm, to obtain a recursive bound on the error $\left\|x^t - x^*\right\|$. This in turn gives sufficient conditions on α_{3k}.

5.3.1 Analysis of the main steps

First, let us fix the iteration t of Algorithm 1, we denote by x^t the current estimate and x^{t+1} the one obtained at end of the iteration, the sets $\mathcal{G}$, $\mathcal{S}$, $\mathcal{T}$ and the element b are the ones computed during this iteration (i.e. with $x = x^t$). $x^\star$ stands for a fixed k-sparse vector in $\mathcal{H}$.

Influence of the guessed support (Step 1 and 2)

Note that $\mathbf{T}$ has RDP of order $3k$ with $\alpha_{3k} \leqslant \alpha^S < 1$ implies that $\mathbf{T}$ has RDP of order $2k$ with $\alpha_{2k} \leqslant \alpha_{3k} < 1$.

Define $\mathcal{R} = \mathrm{supp}(x^t, x^\star)$, we have $\mathrm{card}(\mathcal{R}) \leqslant 2k$. Remember that $\mathcal{S} = \mathrm{supp}(x^t) \cup \mathrm{supp}(T(x^t)_{|k})$. Note that $\mathrm{supp}(x^t) \subseteq \mathcal{S}$ and $\left\|T(x^t)_{|\mathcal{R}}\right\| \leqslant \left\|T(x^t)_{|\mathcal{S}}\right\|$ since $\mathrm{supp}(x^t) \cap \mathrm{supp}(T(x^t)) = \emptyset$. Thus lemma 17 yields

$$\left\|x^\star_{|\mathcal{S}^c}\right\| \leqslant \tfrac{2\alpha_{2k}}{1-\alpha_{2k}} \left\|x^t - x^\star\right\| + \tfrac{1}{1-\alpha_{2k}} \left(\left\|\mathbf{T}(x^\star)_{|\mathcal{R}\setminus\mathcal{S}}\right\| + \left\|\mathbf{T}(x^\star)_{|\mathcal{S}\setminus\mathcal{R}}\right\|\right) .$$

Noting that $\mathrm{card}(\mathcal{R} \setminus \mathcal{S}) \leqslant k$ and $\mathrm{card}(\mathcal{S} \setminus \mathcal{R}) \leqslant k$, we conclude

$$\left\|x^\star_{|\mathcal{S}^c}\right\| \leqslant \tfrac{2\alpha_{2k}}{1-\alpha_{2k}} \left\|x^t - x^\star\right\| + \tfrac{2}{1-\alpha_{2k}} \left\|\mathbf{T}(x^\star)_{|k}\right\| . \tag{18}$$

Optimization over the extended support (Step 3)

Apply Lemma 18 with $l = 2k$ and $\mathcal{R} = \mathcal{S}$ to obtain

$$\left\|(b - x^\star)_{|\mathcal{S}}\right\| \leqslant \tfrac{1}{1-\alpha_{3k}} \left\|\mathbf{T}(x^\star)_{|\mathcal{S}}\right\| + \tfrac{\alpha_{3k}}{1-\alpha_{3k}} \left\|x^\star_{|\mathcal{S}^c}\right\| . \tag{19}$$

Updating the support set (Step 4)

Lemma 19 proves that

$$\left\|b_{|k} - x^\star\right\| \leqslant 2 \left\|(b - x^\star)_{|\mathcal{S}}\right\| + \left\|x^\star_{|\mathcal{S}^c}\right\| . \tag{20}$$

Combining this with Eq. (19) yields:

$$\left\|b_{|k} - x^\star\right\| \leqslant \tfrac{2}{1-\alpha_{3k}} \left\|\mathbf{T}(x^\star)_{|\mathcal{S}}\right\| + \tfrac{1+\alpha_{3k}}{1-\alpha_{3k}} \left\|x^\star_{|\mathcal{S}^c}\right\| . \tag{21}$$

Since $\mathcal{T} = \mathrm{supp}(b_{|k})$, we have,

$$\left\|x^\star_{|\mathcal{T}^c}\right\| = \left\|(b_{|k} - x^\star)_{|\mathcal{T}^c}\right\| \leqslant \left\|b_{|k} - x^\star\right\| .$$

So that

$$\left\|x^\star_{|\mathcal{T}^c}\right\| \leqslant \tfrac{2}{1-\alpha_{3k}} \left\|\mathbf{T}(x^\star)_{|\mathcal{S}}\right\| + \tfrac{1+\alpha_{3k}}{1-\alpha_{3k}} \left\|x^\star_{|\mathcal{S}^c}\right\| . \tag{22}$$

Optimization over the updated support (Step 5)

Pick $l = k$ and $\mathcal{R} = \mathcal{T}$ and apply Lemma 18 to obtain

$$\left\|(x^{t+1} - x^\star)_{|\mathcal{T}}\right\| \leqslant \tfrac{1}{1-\alpha_{2k}} \left\|\mathbf{T}(x^\star)_{|\mathcal{T}}\right\| + \tfrac{\alpha_{2k}}{1-\alpha_{2k}} \left\|x^\star_{|\mathcal{T}^c}\right\| . \tag{23}$$

So

$$\begin{aligned}
\left\|x^{t+1} - x^\star\right\| &\leqslant \left\|(x^{t+1} - x^\star)_{|\mathcal{T}}\right\| + \left\|(x^{t+1} - x^\star)_{|\mathcal{T}^c}\right\| \\
&\leqslant \left\|(x^{t+1} - x^\star)_{|\mathcal{T}}\right\| + \left\|x^\star_{|\mathcal{T}^c}\right\| \\
&\leqslant \tfrac{1}{1-\alpha_{2k}} \left\|\mathbf{T}(x^\star)_{|\mathcal{T}}\right\| + \tfrac{1}{1-\alpha_{2k}} \left\|x^\star_{|\mathcal{T}^c}\right\| .
\end{aligned} \tag{24}$$

Let us now combine these inequalities to bound recursively $\left\|x^{t+1} - x^\star\right\|$.

5.3.2 Recursion

We start from Eq. (24), insert Eq (22) and obtain

$$\begin{aligned}\left\|x^{t+1}-x^\star\right\| &\leqslant \tfrac{1}{1-\alpha_{2k}}\left\|\mathbf{T}(x^\star)_{|\mathcal{T}}\right\| + \tfrac{1}{1-\alpha_{2k}}\left\|x^\star_{|\mathcal{T}^c}\right\| \\ &\leqslant \tfrac{1}{1-\alpha_{2k}}\left\|\mathbf{T}(x^\star)_{|\mathcal{T}}\right\| + \tfrac{1}{1-\alpha_{2k}}\tfrac{2}{1-\alpha_{3k}}\left\|\mathbf{T}(x^\star)_{|\mathcal{S}}\right\| + \tfrac{1}{1-\alpha_{2k}}\tfrac{1+\alpha_{3k}}{1-\alpha_{3k}}\left\|x^\star_{|\mathcal{S}^c}\right\| \ .\end{aligned}$$

Since $\alpha_{2k} \leqslant \alpha_{3k}$, we can simplify the constants and since $\mathrm{card}(\mathcal{T}) \leq \mathrm{card}(\mathcal{S}) \leq 2k$ we have $\left\|\mathbf{T}(x^\star)_{|\mathcal{T}}\right\| \leq \left\|\mathbf{T}(x^\star)_{|2k}\right\|$ and $\left\|\mathbf{T}(x^\star)_{|\mathcal{S}}\right\| \leq \left\|\mathbf{T}(x^\star)_{|2k}\right\|$, so

$$\left\|x^{t+1}-x^\star\right\| \leqslant \left(\tfrac{1}{1-\alpha_{3k}}+\tfrac{2}{(1-\alpha_{3k})^2}\right)\left\|\mathbf{T}(x^\star)_{|2k}\right\| + \tfrac{1+\alpha_{3k}}{(1-\alpha_{3k})^2}\left\|x^\star_{|\mathcal{S}^c}\right\| \ .$$

Then inserting Eq. (18) yields

$$\begin{aligned}\left\|x^{t+1}-x^\star\right\| &\leqslant \tfrac{3-\alpha_{3k}}{(1-\alpha_{3k})^2}\left\|\mathbf{T}(x^\star)_{|2k}\right\| + \tfrac{1+\alpha_{3k}}{(1-\alpha_{3k})^2}\left(\tfrac{2\alpha_{2k}}{1-\alpha_{2k}}\left\|x^t-x^\star\right\| + \tfrac{2}{1-\alpha_{2k}}\left\|\mathbf{T}(x^\star)_{|k}\right\|\right) \\ &\leqslant \tfrac{3-\alpha_{3k}}{(1-\alpha_{3k})^2}\left\|\mathbf{T}(x^\star)_{|2k}\right\| + \tfrac{1+\alpha_{3k}}{(1-\alpha_{3k})^2}\tfrac{2\alpha_{3k}}{1-\alpha_{3k}}\left\|x^t-x^\star\right\| + \tfrac{1+\alpha_{3k}}{(1-\alpha_{3k})^2}\tfrac{2}{1-\alpha_{3k}}\left\|\mathbf{T}(x^\star)_{|2k}\right\| \\ &\leqslant \left(\tfrac{3-\alpha_{3k}}{(1-\alpha_{3k})^2}+\tfrac{1+\alpha_{3k}}{(1-\alpha_{3k})^2}\tfrac{2}{1-\alpha_{3k}}\right)\left\|\mathbf{T}(x^\star)_{|2k}\right\| + \tfrac{1+\alpha_{3k}}{(1-\alpha_{3k})^2}\tfrac{2\alpha_{3k}}{1-\alpha_{3k}}\left\|x^t-x^\star\right\| \ , \\ &\leqslant \tfrac{2\alpha_{3k}(1+\alpha_{3k})}{(1-\alpha_{3k})^3}\left\|x^t-x^\star\right\| + \tfrac{\alpha_{3k}^2-2\alpha_{3k}+5}{(1-\alpha_{3k})^3}\left\|\mathbf{T}(x^\star)_{|2k}\right\| \ .\end{aligned}$$

The sequence $\left\{\left\|x^\star-x^t\right\|\right\}_t$ thus verifies

$$\left\|x^{t+1}-x^\star\right\| \leqslant \tfrac{2\alpha_{3k}(1+\alpha_{3k})}{(1-\alpha_{3k})^3}\left\|x^t-x^\star\right\| + \tfrac{\alpha_{3k}^2-2\alpha_{3k}+5}{(1-\alpha_{3k})^3}\left\|\mathbf{T}(x^\star)_{|2k}\right\| \ . \tag{25}$$

5.3.3 Error bound

From Eq.(25), we deduce

$$\left\|x^t-x^\star\right\| \leqslant \left(\tfrac{2\alpha_{3k}(1+\alpha_{3k})}{(1-\alpha_{3k})^3}\right)^t\left\|x^0-x^\star\right\| + \tfrac{\alpha_{3k}^2-2\alpha_{3k}+5}{(1-\alpha_{3k})^3}\sum_{i=0}^{t-1}\left(\tfrac{2\alpha_{3k}(1+\alpha_{3k})}{(1-\alpha_{3k})^3}\right)^i\left\|\mathbf{T}(x^\star)_{|2k}\right\| \ . \tag{26}$$

The geometric sequence converges if only if $\frac{2\alpha_{3k}(1+\alpha_{3k})}{(1-\alpha_{3k})^3} < 1$ which is equivalent to $\alpha_{3k} < \alpha^1$, where α^1 is the unique real root of $g(x) = x^3 - x^2 + 5x - 1$ (note that $\alpha^1 < 1$).

For more clarity, we gave in Theorem 11 the sufficient condition

$$\tfrac{2\alpha_{3k}(1+\alpha_{3k})}{(1-\alpha_{3k})^3} \leqslant \frac{1}{2} \Leftrightarrow \alpha_{3k} \leqslant \alpha^S \ , \tag{27}$$

where α^S is the unique real root of $h(x) = x^3 + x^2 + 7x - 1$ (note that $\alpha^S < 1$). If $\alpha_{3k} < \alpha^S$, we also have

$$\tfrac{\alpha_{3k}^2-2\alpha_{3k}+5}{(1-\alpha_{3k})^3}\sum_{i=0}^{t-1}\left(\tfrac{2\alpha_{3k}(1+\alpha_{3k})}{(1-\alpha_{3k})^3}\right)^i \leqslant \tfrac{\alpha_{3k}^2-2\alpha_{3k}+5}{(1-\alpha_{3k})^3}\sum_{i=0}^{t-1}\tfrac{1}{2^i} \leqslant 2\tfrac{\alpha_{3k}^2-2\alpha_{3k}+5}{(1-\alpha_{3k})^3} \leqslant 2\tfrac{(\alpha^S)^2-2\alpha^S+5}{(1-\alpha^S)^3} \leqslant 12 \ .$$

We conclude

$$\left\|x^t-x^\star\right\| \leqslant 2^{-t}\left\|x^0-x^\star\right\| + 12\left\|\mathbf{T}(x^\star)_{|2k}\right\| \ .$$

This finishes the proof of Theorem 11.

6 Examples

In this section, we show examples of applications where $\mathbf{T}$ is related to a function $f : \mathcal{H} \to \mathbb{R}$. To show the versatility of the Generalized Greedy algorithms, the examples cover: finding sparse approximations of minimizers of a twice differentiable convex function (Section 6.2), finding sparse approximations of minimizers of a non-differentiable convex function (Section 6.3), and finding sparse approximations of stationary points of a function that is neither concave nor convex but differentiable (Section 6.4).

6.1 Uniform Restricted Diagonal Property and second order differentiability

Let us first further characterize the URDP for $\mathbf{T} = \nabla f$ and f twice differentiable on sparse sets.

Let us define for $\mathbf{A} : \mathcal{H} \to \mathcal{L}(\mathcal{H}, \mathcal{H})$ ($\mathcal{L}(\mathcal{H}, \mathcal{H})$ is the set of linear bounded operators on $\mathcal{H}$)

$$\lambda_k(\mathbf{A}) \stackrel{\text{def}}{=} \inf_{\left\{\substack{(z,u),\ u\neq 0 \\ \text{card}(\text{supp}(z,u))\leq k}\right\}} \frac{\langle u, \mathbf{A}(z)(u)\rangle}{\|u\|^2}. \tag{28}$$

$$\Lambda_k(\mathbf{A}) \stackrel{\text{def}}{=} \sup_{\left\{\substack{(z,u),\ u\neq 0 \\ \text{card}(\text{supp}(z,u))\leq k}\right\}} \frac{\|\mathbf{A}(z)(u)\|}{\|u\|}. \tag{29}$$

The following characterization follows from Theorem 6

Theorem 20. *Assume $f : \mathcal{H} \to \mathbb{R}$ is twice differentiable on $\{x/\,\text{card}(\text{supp}(x)) \leq k\}$. Assume that D is in $\mathcal{D}_1$. If $0 < \lambda_k(D^T\nabla^2 f)$, $\Lambda_k(\nabla^2 f) < \infty$, and $0 \leq |||D|||_k^2 - \frac{\lambda_k(D^T\nabla^2 f)^2}{\Lambda_k(\nabla^2 f)^2} < 1$ then $(\beta\nabla f)$ is URDP of order k for D, with $\alpha_k = |||D|||_k^2 - \frac{\lambda_k(\nabla^2 f)^2}{\Lambda_k(\nabla^2 f)^2}$ and $\beta = \frac{\lambda_k(\nabla^2 f)}{\Lambda_k(\nabla^2 f)^2}$.*

Again, for $\mathbf{D} = \mathbf{I}$, we recover the second order criteria defined along with the *Restricted Strong Smoothness* and *Restricted Strong Convexity* developed in (Bahmani et al., 2013; Yuan et al., 2014), confirming that our results encompass those of the present literature. We give the example of logistic regression in the next section. It also bear similarities with the "Sparse eigenvalue" criterion in (Yang et al., 2016).

6.2 Example of the Logistic Regression

Let us consider the case of supervised learning with a learning set $\{(y_n, l_n)\}_{n=1\ldots N}$, where $y_n \in \mathcal{H} = \mathbb{R}^d$ is the n-th training vector and $l_n \in \{0, 1\}$ its label. We assume that the y_n are independent identically distributed with the same law as the random vector $\mathcal{Y}$ and that the labels l_n follows a logistic law of parameter x knowing y_n that is:

$$\mathbb{P}(l_n|y_n, x) = \frac{1}{(1 + \exp(-\langle y_n, x\rangle))^{l_n}(1 + \exp(\langle y_n, x\rangle))^{1-l_n}} .$$

We wish to estimate the parameter $x \in \mathcal{H}$ to classify new instances of $\mathcal{Y}$. Here we choose to estimate x as the minimizer of the negative log-likelihood penalized by a Tikhonov term i.e. we minimize for a given $\mu > 0$

$$f(x) = \tfrac{1}{N}\sum_{n=1}^{N} -\log(\mathbb{P}(l_n|y_n, x)) + \tfrac{1}{2}\mu\,\|x\|_2^2 \ . \tag{30}$$

We have:

$$f(x) = \tfrac{1}{N}\sum_{n=1}^{N} (\log(1 + \exp(\langle y_n, x\rangle)) - l_n\,\langle y_n, x\rangle) + \tfrac{1}{2}\mu\,\|x\|_2^2 \ . \tag{31}$$

$$\nabla f(x) = \tfrac{1}{N}\sum_{n=1}^{N} \left(\frac{1}{1 + \exp(\langle y_n, x\rangle)} - l_n\right) y_n + \mu x \ . \tag{32}$$

$$\nabla^2 f(x) = \tfrac{1}{N}\sum_{n=1}^{N} \frac{1}{(1 + \exp(\langle y_n, x\rangle))(1 + \exp(-\langle y_n, x\rangle))} y_n y_n^{\mathrm{T}} + \mu\mathbf{I} \ . \tag{33}$$

We thus have:

$$\forall x,\ \mu\mathbf{I} \preceq \nabla^2 f(x) \preceq \mu\mathbf{I} + \tfrac{1}{4N}\sum_{n=1}^{N} y_n y_n^{\mathrm{T}} \ . \tag{34}$$

Hence f is strongly convex on $\mathcal{H}$ and $\lambda_k(\nabla^2 f) \geq \mu$. Assuming furthermore that the k-sparse projections of $\mathcal{Y}$ are bounded almost surely i.e.

$$\exists\ R,\ \text{such that } \mathbb{P}(\left\|\mathcal{Y}_{|k}\right\|^2 \leq R) = 1,$$

we have $\text{card}(\text{supp}(u)) \leq k \Rightarrow u^{\mathrm{T}} y_n y_n^{\mathrm{T}} u \leq R\left\|u\right\|^2$. Hence $\Lambda_k(\nabla^2 f) \leq \mu + \frac{R}{4}$ so that $\beta\nabla f$ is URDP of order k for $D_k = \mathbf{I}$, $\alpha_k = 1 - \frac{1}{\left(1+\frac{R}{4\mu}\right)^2}$ and $\beta = \frac{\mu}{\left(\mu+\frac{R}{4}\right)^2}$. For μ large enough, one can then guarantee the convergence of all four algorithms for f. (Note that similar bounds were obtained in Bahmani et al. (2013) for λ_k and Λ_k).

6.3 Application to Non-Smooth Convex Functions

In the last subsection, we considered a smooth function, but most of functions in $\Gamma_0(\mathcal{H})$ are not smooth. One way to deal with such issue is to regularize the function itself using an infimal convolution. Here we use the Moreau-Yosida regularization which is an infimal convolution with an ℓ_2-norm.

The Moreau-Yosida regularization of a function $f \in \Gamma_0(\mathcal{H})$ (Lemaréchal and Sagastizábal, 1997) of parameter λ is defined by:

$$\mathcal{M}_{\lambda,f}(x) = \inf_{y\in\mathcal{H}} \left[\tfrac{1}{2\lambda}\left\|x-y\right\|^2 + f(y)\right] \ . \tag{35}$$

The Moreau-Yosida envelope of a convex lower semi-continuous function has full domain and its gradient is Lipschitz. Moreover, its minimizer is also a minimizer of the original function. These properties make this regularization useful when dealing with non-smooth function, but it can also be useful to regularize smooth function (e.g.the exponential as it does not have a Lipschitz gradient).

One interesting fact about the Moreau-Yosida regularization, is its link with proximal operator, we have,

$$\nabla\mathcal{M}_{\lambda,f} = (\mathbf{I} - \text{prox}_{\lambda f})/\lambda \ , \tag{36}$$

with the proximal operator of f defined as,

$$\text{prox}_{\lambda f} : x \mapsto \underset{z\in\mathcal{H}}{\text{argmin}}\ \lambda f(z) + \frac{1}{2}\left\|z-x\right\| \ . \tag{37}$$

Proximal operators can be viewed as generalization of orthogonal projections and are easily computable for a large set of functions (l_2-, l_1-, or mixed-norms, TV-norm....).

We propose to use $\mathbf{T} = \nabla\mathcal{M}_{\lambda,f}$ in our Generalized Greedy algorithms. All the iterates will be well-defined because $\mathcal{M}_{\lambda,f}$ is convex and differentiable on $\mathcal{H}$. The theoretical error bounds will hold however only for the cases where the (U)RDP is shown (the natural Lipschitz property of $\mathbf{T}$ is a step toward it). In the experimental section (Section 7), we show an example where f denotes the Poisson likelihood.

6.4 Sparse approximation of stationary points of a differentiable function that is neither convex nor concave

Here $\mathcal{H} = \mathbb{R}^N$ is split into $\mathcal{H} = \mathbb{R}^{N_1} \times \mathbb{R}^{N_2}$ with $N_1 + N_2 = N$. $x \in \mathcal{H}$ is written accordingly $x = (x_1, x_2)$. Let us assume that $A_1 \in \mathbb{R}^{P\times N_1}$ and $A_2 \in \mathbb{R}^{P\times N_2}$ have the RIP property of order k, with constants δ_k^1 and δ_k^2. Pick $z_1 \in \mathbb{R}^P$, $z_2 \in \mathbb{R}^P$ and γ such that $1 \leqslant \gamma < \frac{1}{\sqrt{\delta_k^2}}$. We wish to find the best k-sparse approximation of the stationary points of

$$f(x) = f(x_1, x_2) = \tfrac{1}{2}\left\|A_1x_1 - z_1\right\|_2^2 - \tfrac{\gamma}{2}\left\|A_2x_2 - z_2\right\|_2^2 \tag{38}$$

Notice that f is a difference of convex functions so it is neither convex nor concave. We have: $\nabla f(x) = \begin{pmatrix} A_1^TA_1x_1 - A_1^Tz_1 \\ -\gamma A_2^TA_2x_2 + \gamma A_2^Tz_2 \end{pmatrix}$. We define $\mathbf{D} = \begin{pmatrix} \mathbf{I}_{N_1} & 0 \\ 0 & -\gamma\mathbf{I}_{N_2} \end{pmatrix}$. We have:

$$\nabla f(x) - \nabla f(y) - \mathbf{D}(x-y) = \begin{pmatrix} (A_1^TA_1 - \mathbf{I}_{N_1})(x_1-y_1) \\ (-\gamma A_2^TA_2 + \gamma\mathbf{I}_{N_2})(x_2-y_2) \end{pmatrix}$$

so that:

$$\begin{aligned}\|\nabla f(x) - \nabla f(y) - \mathbf{D}(x-y)\|_2^2 &= \left\|A_1^T A_1 - \mathbf{I}_{N_1})(x_1 - y_1)\right\|_2^2 + \gamma^2 \left\|A_2^T A_2 - \mathbf{I}_{N_2})(x_2 - y_2)\right\|_2^2 \\ &\leqslant \delta_k^1 \|x_1 - y_1\|_2^2 + \delta_k^2 \gamma^2 \|x_2 - y_2\|_2^2 \\ &\leqslant \max(\delta_k^1, \delta_k^2\gamma^2)(\|x_1 - y_1\|_2^2 + \|x_2 - y_2\|_2^2) \\ &\leqslant \max(\delta_k^1, \delta_k^2\gamma^2) \|x - y\|_2^2 .\end{aligned}$$

Since $\max(\delta_k^1, \delta_k^2\gamma^2) < 1$ We conclude that $\mathbf{T} = \nabla f(x)$ has the URDP with $\mathbf{D}$ and so that GSP and GCoSaMP may be used in that case.

7 Experiments

7.1 Poisson-sparsity

Let us assume that we observe $y \in \mathbb{R}^n$, a Poisson corrupted version of the true image $x \in \mathbb{R}^n$, both containing n pixels. We also assume that x has a k-sparse representation on the dictionary $\mathbf{\Phi} = (\varphi_1, \ldots, \varphi_m) \in \mathbb{R}^{n \times m}$:

$$x = \mathbf{\Phi} c = \sum c_j \varphi_j \text{ with } ||c||_0 = k \ll n \ ,$$

where the atoms are normalized ($\|\varphi_j\| = 1$), and that $\mathbf{\Phi}$ is a tight frame with constant ν.

Our goal is to reconstruct x given the data y, the sparsity k and the dictionary $\mathbf{\Phi}$, which may be done by solving:

$$\hat{x} = \mathbf{\Phi}\hat{c}, \quad \text{where} \quad \hat{c} = \underset{c \in \mathbb{R}^m \text{ s.t. } ||c||_0 \leqslant k}{\operatorname{argmin}} F_y(\mathbf{\Phi} c) \ , \tag{P5}$$

where $F_y(\hat{x})$ is a data fidelity term that quantifies how well an estimated image $\hat{x}$ fits the observed data y. A natural fidelity term is the negative-log-likelihood $F_y(x) = -\log \mathbb{P}(y|x)$ which reads in the case of Poisson noise

$$\begin{aligned} F_y(x) = -\log \mathbb{P}(y|x) &= \sum_{i=1}^{n} f(x_i, y_i), \quad \text{with} \\ f(\xi, \eta) &= \begin{cases} -\eta \log(\xi) + \xi & \text{if } \eta > 0 \text{ and } \xi > 0, \\ \xi & \text{if } \eta = 0 \text{ and } \xi \geqslant 0, \\ +\infty & \text{otherwise.} \end{cases} \end{aligned} \tag{39}$$

Notice that $F_y(x)$ is finite only when x complies with the data, which implies $x \in \mathbb{R}_+^n$ and $x_i > 0$ if $y_i > 0$. Moreover, due to the logarithm, it gradient is not defined on its all domain. As proposed in Section 6.3, we seek $\hat{x}$ using our four Generalized Algorithms on $\mathbf{T} = \nabla \mathcal{M}_{\lambda, F_y \circ \mathbf{\Phi}}$.

Proposition 21 (Gradient of the Moreau-Yosida regularization of the Poisson neg-log-likelihood (Combettes and Pesquet, 2007))**.** *If $\mathbf{\Phi}$ is a tight frame of constant $\nu > 0$, then the gradient of $\mathcal{M}_{\lambda, F_y \circ \mathbf{\Phi}}$ is:*

$$\begin{aligned} \nabla \mathcal{M}_{\lambda, F_y \circ \mathbf{\Phi}} &= \tfrac{1}{\nu\lambda} \mathbf{\Phi}^* \circ (\mathbf{I} - \operatorname{prox}_{\nu\lambda F_y}) \circ \mathbf{\Phi} \qquad \textit{with} \\ \operatorname{prox}_{\nu\lambda F_y}(x)_i &= \frac{x_i - \nu\lambda + \sqrt{|x_i - \nu\lambda|^2 + 4\nu\lambda y_i}}{2} \ . \end{aligned} \tag{40}$$

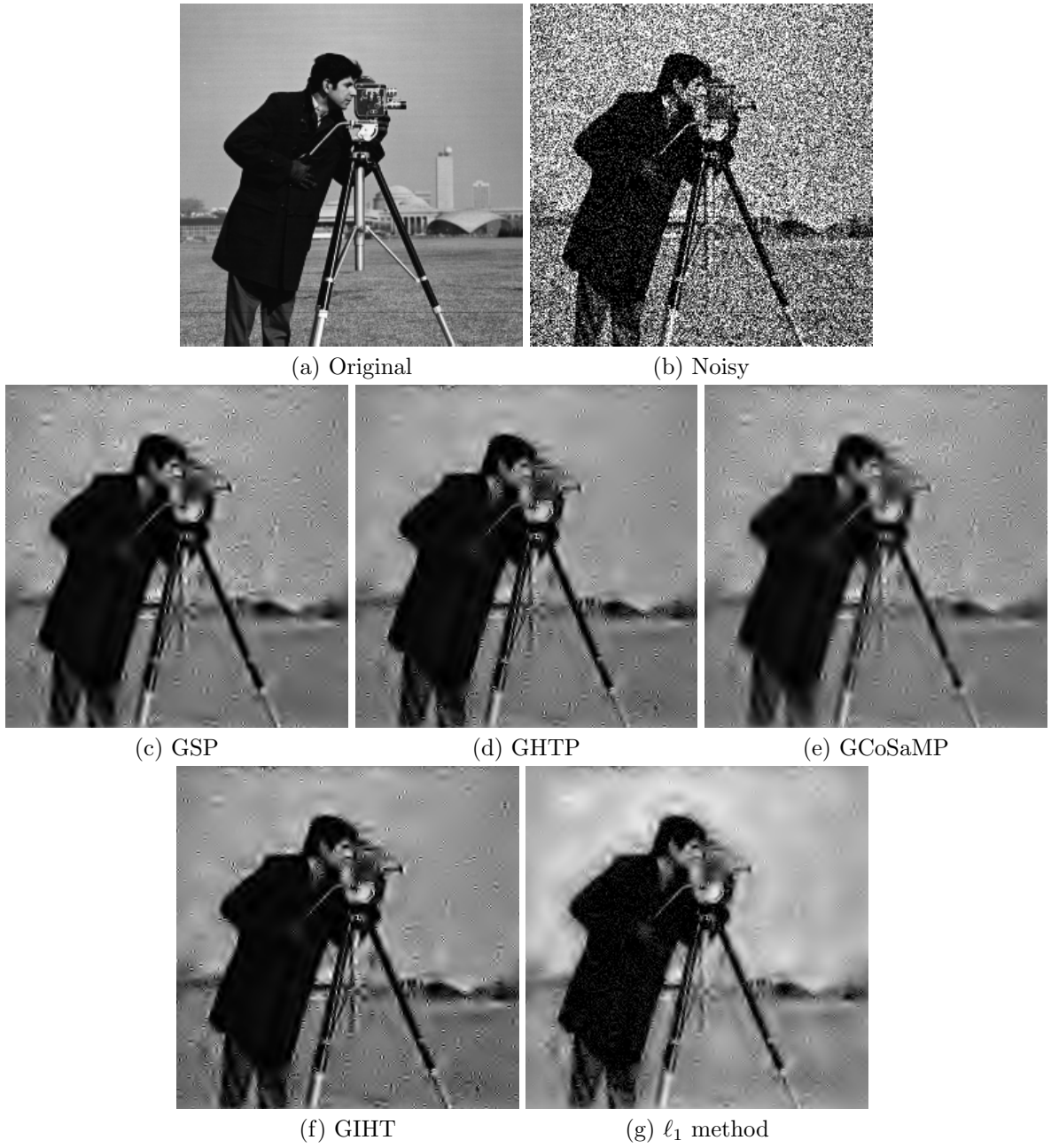
7.2 Visual comparison

In this section, we evaluate the performance of our the Generalized Greedy alternatives (named ℓ_0 methods) and compare them to the classical convex relaxation using the ℓ_1-norm instead of the ℓ_0-pseudo-norm (hereafter named ℓ_1 method). The estimation is implemented using the model in (Combettes and Pesquet, 2008). The experiments shed light on the effects of using the ℓ_0-pseudo-norm

instead the ℓ_1-norm, the consequences of using the Moreau-Yosida regularization and the difference between the different sparse methods.

Two experiments are proposed using two classical images (*Cameraman* and *Barbara*) with two different dictionaries, the undecimated wavelet transform (with the symlet 6) and the curvelet transform. Notice that both transforms are redundant and so well fit for the denoising task. The noise level, which is parameterized by the maximal intensity of the original image x (higher maximal intensity means less noise) is set to either high or medium.

For all the experiments, we set the Moreau-Yosida regularization parameter to 1. This value may lead to a non-negligible bias but allows for a better convergence rate. The sparsity parameter of the ℓ_0 and ℓ_1 methods have been fixed to give comparable sparsity levels. Note that finding *good* (or optimal) parameters is an open problem in both cases (see (Vaiter et al., 2012) for an example for the Gaussian noise case).



(a) Original (b) Noisy

(c) GSP (d) GHTP (e) GCoSaMP

(f) GIHT (g) ℓ_1 method

Figure 1: Denoising *Cameraman* with a maximal intensity of 5 with the undecimated wavelet transform.

Figure 1 shows the results for the *Cameraman* with a maximal intensity of 5 (high noise). To show the differences of photometry, the images in a same figure are always displayed using the same

grayscale colormap. Assuming that the image is sparse in the undecimated wavelet domain (which is partly true), we apply the ℓ_0 methods (Fig. 1(c)-(f)) and compare them to the ℓ_1 method (Fig. 1(g)). Notice that both GSP and GCoSaMP leads a smoother image, while most of the details are preserved with GHTP, GIHT and the ℓ_1-norm. Because the *Cameraman* is not truly sparse in the chosen domain, enforcing the sparsity for the reconstruction is not relevant. However, using the ℓ_0 preserves the photometry better: for example the coat of the *Cameraman* is darker in Fig. 1(f) than in Fig. 1(g).

(a) Original (b) Noisy

(c) GSP (d) GHTP (e) GCoSaMP

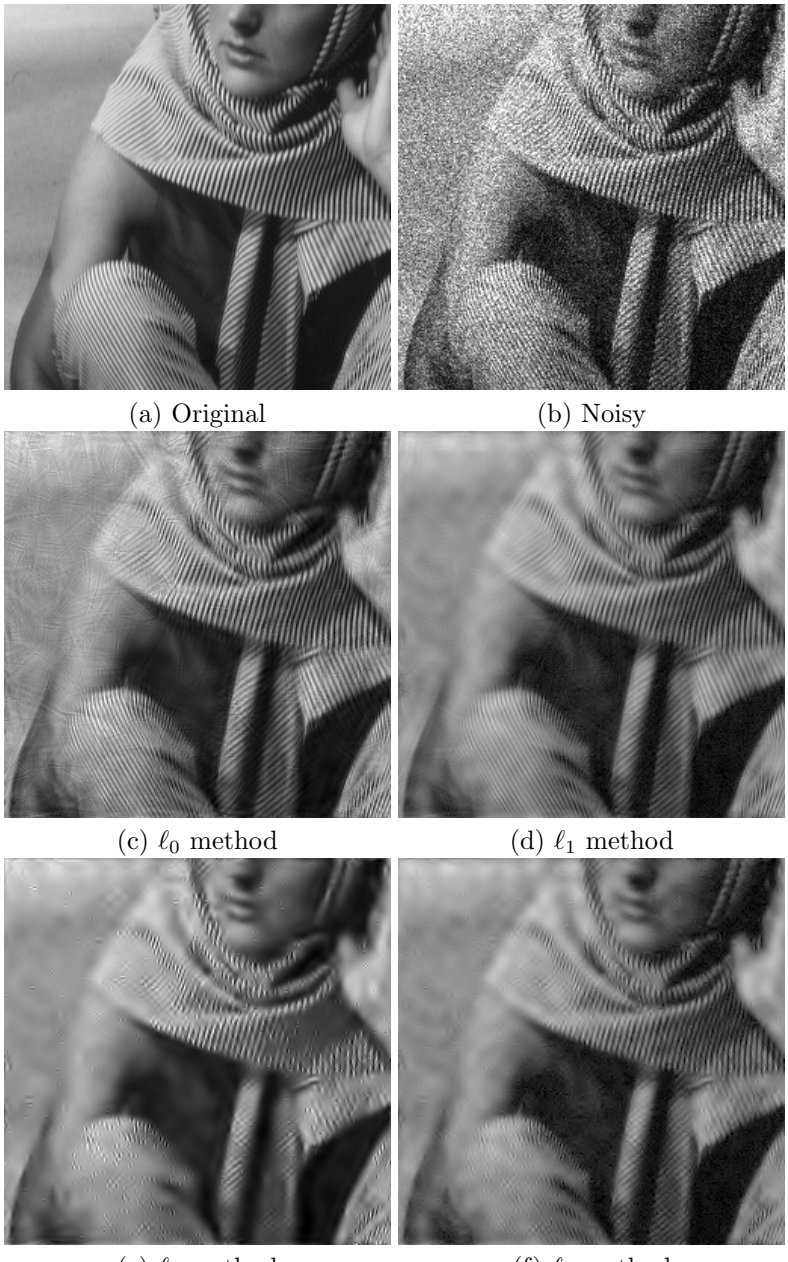
(f) GIHT (g) ℓ_1 method

Figure 2: Denoising *Cameraman* with a maximal intensity of 30 with the undecimated wavelet transform.

We repeat the experiment with a maximal intensity of 30 (medium noise). As the noise is weaker, more details should be recovered. Figure 2 shows the results with both methods. The ℓ_0 methods (Fig. 2(c)-(f)) preserves the details as well as the ℓ_1 method (Fig. 2(g)). Furthermore, both GHTP and GIHT lead to pretty good reconstruction. As with the previous experiment, the most important difference between Fig. 2(f) and Fig. 2(g) is the photometry. For example, the camera is brighter in Fig. 2(f) (like in the original) than in Fig. 2(g). We believe that the difference of reconstruction quality between GHTP and GIHT on one side, and GSP and GCoSaMP on the other side, is the diagonal fixed to identity. As the Poisson negative-log-likelihood is a convex function, as a consequence of the

Baillon-Haddad theorem (Bauschke and Combettes, 2010), the diagonal should be close to the identity and so the algorithms that make such assumption (GHTP and GIHT) are better than scale-insensitive methods (GSP and GCoSaMP).

To show the effect of the choice of the dictionary, we repeat the experiment with the *Barbara* image (Fig. 3(a)) because of the curve-like textures on the pants, at a maximal intensity of 30 (medium noise). Figure 3 shows the results for each method using two different dictionaries, Fig. 3(c) and Fig. 3(d) for the curvelet transform and Fig. 3(e) and Fig. 3(f) for undecimated wavelet transform. As expected the ℓ_0 method also shows a better photometry. Moreover, for both methods, the curvelet transform is better at restoring the textures. With the undecimated wavelet transform, part of the textures is lost and the ℓ_0 method is less efficient than the ℓ_1 method (see specifically the shawl). This shows the importance of the selection of the dictionary while using sparse method. Using the wrong one may lead to artifacts and loss of some structures (like textures).



(a) Original (b) Noisy

(c) ℓ_0 method (d) ℓ_1 method

(e) ℓ_0 method (f) ℓ_1 method

Figure 3: Denoising *Barbara* with a maximal intensity of 30. (c) and (d) using the curvelet transform. (e) and (f) using the undecimated wavelet transform.

	Sparse Cameraman		Galaxy	
	MAE	SSIM	MAE	SSIM
Noisy	1.57	0.32	0.63	0.19
GSP	**0.32**	**0.87**	0.17	0.71
SP (Dai and Milenkovic, 2009)	0.55	0.63	0.28	0.55
SAFIR (Boulanger et al., 2010)	0.36	**0.86**	0.15	**0.84**
MSVST (Zhang et al., 2008)	**0.31**	0.84	**0.12**	**0.83**
ℓ_1-relaxation	0.64	0.73	0.32	0.50

Table 1: Comparison of denoising methods on a sparse version of Cameraman ($k/n = 0.15$) and the NGC 2997 Galaxy.

7.3 Comparison with state of art methods

Finally, we compare one of the proposed method (GSP) with other states-of-art methods: Subspace Pursuit (SP) (Dai and Milenkovic, 2009) (denoising with the Gaussian negative-log-likelihood), SAFIR (Boulanger et al., 2010) (with the parameters from (Makitalo and Foi, 2011)), MSVST (Zhang et al., 2008) (a variance-stabilizing approach) and a convex ℓ_1-relaxation of (P5) i.e. a procedure minimizing the Poisson negative-log-likelihood on a ℓ_1-ball (using a projection onto the ℓ_1-ball (Combettes and Pesquet, 2012; Chierchia et al., 2012)).

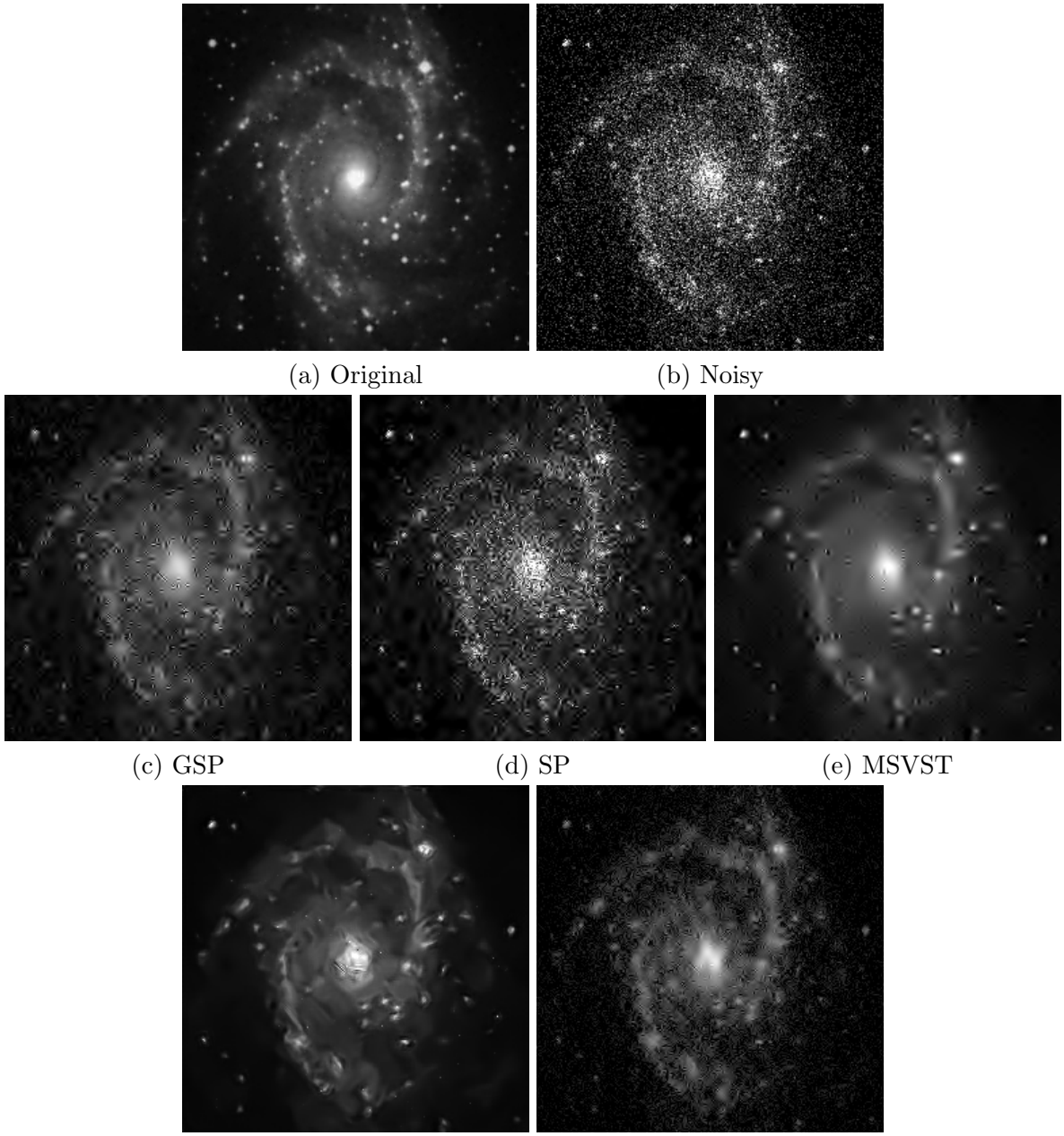
We apply these methods on two images, a sparse version of the Cameraman and the NGC 2997 galaxy (peak intensity at 5, see Fig. 4). We use the exact parameters (sparsity for GSP and SP and ℓ_1-norm for the ℓ_1-relaxation) for the sparse cameraman and tune them for the Galaxy. For each method we compute the mean absolute deviation (MAE) and the SSIM (Wang et al., 2004) and display them in Table 1.

The MSVST gives the best results in terms of MAE and SAFIR does so for the SSIM. GSP is competitive with both these state-of-the-arts methods in Poisson denoising. Moreover GSP is in both cases better than SP and the ℓ_1-relaxation. While the performance over SP shows the benefits of using non-quadratic cost in the case of high Poisson noise level, the performance over the ℓ_1-relaxation emphasizes that using the exact sparse ℓ_0 a-priori enables to recover a better dynamic than its convex relaxation counterpart. These remarks are visually confirmed for the galaxy (Fig. 4).

8 Conclusion

In this paper, we have extended the scope of four common greedy methods from sparse approximation or sparse constrained minimization to the more general problem of sparse approximation of zeros of operators in a Hilbert space of possibly infinite dimension. This enables to run these algorithms with operators that are not gradient (and so not related to a function). We introduced a convergence criterion, the *Restricted Diagonal Property*, that generalizes the previous proposed criteria (RIP, RSC, RSS) and bounds the error after N steps. We have shown that RDP enables the generalized versions of *Subspace Pursuit* and *CoSaMP* to handle neither convex nor concave optimization problems. This suggests that both algorithms are not "corrected versions" (i.e. with corrected steps) of the more classical GIHT or GHTP, but belong to another class of methods. We plan to study what kind of algorithms (or schemes) show such an invariance property.

Several perspectives toward further generalizations of these algorithms are of interest. First, introducing additional constraints (e.g. positivity, unit simplex) as it has been done for IHT by Beck and Hallak (2015) should lead to very interesting applications like sparse support vector machine (with the Hinge loss function). Secondly, one could extend the setting from Hilbert to Banach spaces, as has already been proposed for the Orthogonal Matching Pursuit by Temlyakov (2008). Finally, we would like to broaden our *Restricted Diagonal Property* by comparing the increments to the action of an isometry rather than a diagonal operator. Such an extension is linked with Hyers-Ulam stability analysis (Jung, 2011) for isometry, and would help to build a less restrictive criterion.



(a) Original (b) Noisy

(c) GSP (d) SP (e) MSVST

(f) SAFIR (g) ℓ_1-relaxation

Figure 4: NGC 2997 Galaxy image (a), a noisy version (b) and several denoising results (c-g).

A Uniform Restricted Diagonal Property

Proof of Theorem 6. Assume that D is in $\mathcal{D}_1$ and that $|||D|||_k$ is finite.

Proof of 1: Assume $(\beta\mathbf{T})$ is URDP of order k for D, $\alpha_k < 1$ and $\beta > 0$. Pick $(x, y) \in \mathcal{H}^2$, such that $\text{card}(\text{supp}(x, y)) \leq k$, we have:

$$\begin{aligned}\|\beta\mathbf{T}(x) - \beta\mathbf{T}(y) - D(x-y)\| \leq \alpha_k \|x-y\| &\Rightarrow \|\beta\mathbf{T}(x) - \beta\mathbf{T}(y)\| - \|D(x-y)\| \leq \alpha_k \|x-y\| \\ &\Rightarrow \|\beta\mathbf{T}(x) - \beta\mathbf{T}(y)\| \leq \alpha_k \|x-y\| + \|D(x-y)\| \\ &\Rightarrow \|\mathbf{T}(x) - \mathbf{T}(y)\| \leq \frac{|||D|||_k + \alpha_k}{\beta} \|x-y\| \end{aligned} \tag{41}$$

and

$$\begin{aligned}\|\beta\mathbf{T}(x) - \beta\mathbf{T}(y) - D(x-y)\| \leq \alpha_k \|x-y\| &\Rightarrow \|D(x-y)\| - \|\beta\mathbf{T}(x) - \beta\mathbf{T}(y)\| \leq \alpha_k \|x-y\| \\ &\Rightarrow \|\beta\mathbf{T}(x) - \beta\mathbf{T}(y)\| \geq \|D(x-y)\| - \alpha_k \|x-y\| \\ &\Rightarrow \|\mathbf{T}(x) - \mathbf{T}(y)\| \geq \frac{1-\alpha_k}{\beta} \|x-y\| \ \text{ (since } D \in \mathcal{D}_1) \end{aligned} \tag{42}$$

so that

$$\begin{aligned}\langle \mathbf{T}(x) - \mathbf{T}(y), D(x-y)\rangle =& \tfrac{1}{\beta}\langle \beta\mathbf{T}(x) - \beta\mathbf{T}(y), D(x-y)\rangle \\ =& \frac{1}{2\beta}\left(\|\beta\mathbf{T}(x) - \beta\mathbf{T}(y)\|^2 + \|D(x-y)\|^2 - \|\beta\mathbf{T}(x) - \beta\mathbf{T}(y) - D(x-y)\|^2\right) \\ \geq& \frac{1}{2\beta}\left((1-\alpha_k)^2 \|x-y\|^2 + \|x-y\|^2 - \alpha_k^2 \|x-y\|^2\right) \ \text{(using URDP, } D \in \mathcal{D}_1 \text{ and Eq. (42} \\ \geq& \frac{(1-\alpha_k)^2 + 1 - \alpha_k^2}{2\beta} \|x-y\|^2 \\ \geq& \frac{1-\alpha_k}{\beta} \|(x-y)\|^2 \end{aligned} \tag{43}$$

Proof of 2: Assume that f verifies Eq. (9). Pick $(x, y) \in \mathcal{H}^2$, such that $\text{card}(\text{supp}(x, y)) \leq k$; for any $\beta > 0$ we have:

$$\begin{aligned}\|\beta\mathbf{T}(x) - \beta\mathbf{T}(y) - D(x-y)\|^2 &= \|\beta\mathbf{T}(x) - \beta\mathbf{T}(y)\|^2 + \|D(x-y)\|^2 - 2\beta \langle \mathbf{T}(x) - \mathbf{T}(y), D(x-y)\rangle \\ &\leq \beta^2 L^2 \|x-y\|^2 + |||D|||_k^2 \|x-y\|^2 - 2\beta m \|x-y\|^2 \quad \text{(using Eq. (9))} \\ &\leq (\beta^2 L^2 + |||D|||_k^2 - 2\beta m) \|x-y\|^2 \end{aligned}$$

Pick $\beta = \frac{m}{L^2}$. Note that $\beta > 0$ and $\beta^2 L^2 + |||D|||_k^2 - 2\beta m = |||D|||_k^2 - \frac{m^2}{L^2}$ thus if $(x, y) \in \mathcal{H}^2$, such that $\text{card}(\text{supp}(x, y)) \leq k$, we obtain:

$$\|\beta\mathbf{T}(x) - \beta\mathbf{T}(y) - (x-y)\|^2 \leq \left(|||D|||_k^2 - \frac{m^2}{L^2}\right) \|x-y\|^2$$

which shows that $(\beta\mathbf{T})$ is URDP of order k for D, $\alpha_k = |||D|||_k^2 - \frac{m^2}{L^2}$ and $\beta = \frac{m}{L^2}$. Note that $\alpha_k = |||D|||_k^2 - \frac{m^2}{L^2} < 1$ by assumption. □

B Error bounds

B.0.1 Proof of GCoSaMP's error bound (Theorem 12)

Proof of Theorem 12. $\mathbf{T}$ and $\rho\mathbf{T}$ yield the same iterates, so we assume that $\mathbf{T}$ has the *Restricted Diagonal Property* of order $4k$ with $\alpha_{4k} \leqslant \alpha^C = \frac{2}{\sqrt{3}} - 1 < 1$. Let $x^\star$ in $\mathcal{H}$ be any k-sparse vector, x^t be the t-th iterate of Algo. 2. Let $\mathcal{G} = \text{supp}(\mathbf{T}(x^t)_{|2k})$ and $\mathcal{S} = \mathcal{G} \cup \text{supp}(x^t)$, b such that $\text{supp}(b) \subseteq \mathcal{S}$ and $\mathbf{T}(b)_{|\mathcal{S}} = 0$ and $\mathcal{T} = \text{supp}(b_{|k})$.

Influence of the guessed support (Step 1 and 2)

Define $\mathcal{R} = \text{supp}(x^t, x^\star)$, we have $\text{card}(\mathcal{R}) \leqslant 2k$. Note that $\text{supp}(x^t) \subseteq \mathcal{S}$. We have $\left\|T(x^t)_{|\mathcal{R}}\right\| \leqslant \left\|T(x^t)_{|\mathcal{S}}\right\|$ because $\text{card}(\mathcal{R}) \leqslant 2k$ and $\mathcal{S} = \mathcal{G} \cup \text{supp}(x^t)\,\text{supp}(\mathbf{T}(x^t)_{|2k}) \cup \text{supp}(x^t)$. Thus lemma 17 yields

$$\left\|x^\star_{|\mathcal{S}^c}\right\| \leqslant \tfrac{2\alpha_{2k}}{1-\alpha_{2k}} \left\|x^t - x^\star\right\| + \tfrac{1}{1-\alpha_{2k}} \left(\left\|\mathbf{T}(x^\star)_{|\mathcal{R}\setminus\mathcal{S}}\right\| + \left\|\mathbf{T}(x^\star)_{|\mathcal{S}\setminus\mathcal{R}}\right\|\right) \ .$$

Noting that $\text{card}(\mathcal{R} \setminus \mathcal{S}) \leqslant k$ and $\text{card}(\mathcal{S} \setminus \mathcal{R}) \leqslant 2k$, we conclude

$$\left\|x^\star_{|\mathcal{S}^c}\right\| \leqslant \tfrac{2\alpha_{2k}}{1-\alpha_{2k}} \left\|x^t - x^\star\right\| + \tfrac{2}{1-\alpha_{2k}} \left\|\mathbf{T}(x^\star)_{|2k}\right\| \ . \tag{44}$$

Optimization over the extended support (Step 3)

Apply Lemma 18 with $l = 3k$ and $\mathcal{R} = \mathcal{S}$ and using that $\mathbf{T}$ has the RDP of order $l + k = 4k$ with $\alpha_{4k} \leqslant \alpha^C < 1$, we obtain

$$\left\|(b - x^\star)_{|\mathcal{S}}\right\| \leqslant \tfrac{1}{1-\alpha_{4k}} \left\|\mathbf{T}(x^\star)_{|\mathcal{S}}\right\| + \tfrac{\alpha_{4k}}{1-\alpha_{4k}} \left\|x^\star_{|\mathcal{S}^c}\right\| \ . \tag{45}$$

Updating the support set (Step 4 and 5)

Lemma 19 proves that

$$\left\|b_{|k} - x^\star\right\| \leqslant 2\left\|(b - x^\star)_{|\mathcal{S}}\right\| + \left\|x^\star_{|\mathcal{S}^c}\right\| \ . \tag{46}$$

□

Since $x^{t+1} = b_{|k}$, we combine Eq. (44), (45) and (46)to obtain

$$\begin{aligned}
\left\|x^{t+1} - x^\star\right\| &\leqslant 2\left\|(b - x^\star)_{|\mathcal{S}}\right\| + \left\|x^\star_{|\mathcal{S}^c}\right\| \\
&\leqslant \tfrac{2}{1-\alpha_{4k}} \left\|\mathbf{T}(x^\star)_{|\mathcal{S}}\right\| + \tfrac{2\alpha_{4k}}{1-\alpha_{4k}} \left\|x^\star_{|\mathcal{S}^c}\right\| + \left\|x^\star_{|\mathcal{S}^c}\right\| \\
&\leqslant \tfrac{2}{1-\alpha_{4k}} \left\|\mathbf{T}(x^\star)_{|3k}\right\| + \tfrac{1+\alpha_{4k}}{1-\alpha_{4k}} \left\|x^\star_{|\mathcal{S}^c}\right\| \\
&\leqslant \tfrac{2}{1-\alpha_{4k}} \left\|\mathbf{T}(x^\star)_{|3k}\right\| + \tfrac{1+\alpha_{4k}}{1-\alpha_{4k}} \left(\tfrac{2\alpha_{2k}}{1-\alpha_{2k}} \left\|x^t - x^\star\right\| + \tfrac{2}{1-\alpha_{2k}} \left\|\mathbf{T}(x^\star)_{|2k}\right\|\right) \\
&\leqslant \tfrac{2}{1-\alpha_{4k}} \left\|\mathbf{T}(x^\star)_{|3k}\right\| + \tfrac{1+\alpha_{4k}}{1-\alpha_{4k}} \left(\tfrac{2\alpha_{4k}}{1-\alpha_{4k}} \left\|x^t - x^\star\right\| + \tfrac{2}{1-\alpha_{4k}} \left\|\mathbf{T}(x^\star)_{|3k}\right\|\right) \\
&\leqslant \tfrac{4}{(1-\alpha_{4k})^2} \left\|\mathbf{T}(x^\star)_{|3k}\right\| + \tfrac{2\alpha_{4k}(1+\alpha_{4k})}{(1-\alpha_{4k})^2} \left\|x^t - x^\star\right\|
\end{aligned}$$

We have

$$\tfrac{2\alpha(1+\alpha)}{(1-\alpha)^2} \leqslant \tfrac{1}{2} \Leftrightarrow 3\alpha^2 - 6\alpha - 1 \leqslant 0 \Leftrightarrow -1 - \tfrac{2}{\sqrt{3}} \leqslant \alpha \leqslant -1 + \tfrac{2}{\sqrt{3}} = \alpha^C \tag{47}$$

and

$$\tfrac{4}{(1-\alpha_{4k})^2} \leqslant \tfrac{4}{(1-\alpha^C)^2} = \tfrac{3}{(\sqrt{3}-1)^2} \leqslant 6 \ , \tag{48}$$

which completes the proof.

B.0.2 Proof of GHTP's error bound (Theorem 13)

The core of GHTP is a descent step followed by an optimization on the estimated support.

Proof of Theorem 13. Assume that $\mathbf{T}$ has the *Uniform Restricted Diagonal Property* of order $2k$ with $\mathbf{D}_{2k} = \mathbf{I}$ and $\alpha_{2k} \leqslant \alpha^H = 7 - 2\sqrt{11}$, and that $\frac{3}{4} < \eta < \frac{5}{4}$. Let $x^\star$ in $\mathcal{H}$ be any k-sparse vector, x^t be the t-th iterate of Algo. 3. Let $\mathcal{G} = \text{supp}(\mathbf{T}(x^t)_{|k})$ and $\mathcal{S} = \mathcal{G} \cup \text{supp}(x^t)$, $b = (x^t - \eta\mathbf{T}(x^t))_{|\mathcal{S}}$ and $\mathcal{T} = \text{supp}(b_{|k})$.

Influence of the guessed support (Step 1 and 2)

Notice that x^t solves Problem (P4) and $\mathcal{G}$ and $\mathcal{S}$ are the same as in GSP so the same arguments as in the proof in Section 5.3.1 apply and Eq. (18) holds:

$$\left\|x^\star_{|\mathcal{S}^c}\right\| \leqslant \tfrac{2\alpha_{2k}}{1-\alpha_{2k}} \left\|x^t - x^\star\right\| + \tfrac{2}{1-\alpha_{2k}} \left\|\mathbf{T}(x^\star)_{|k}\right\| \ . \tag{18}$$

Solution on the extended support (Step 3)

$$\begin{aligned}
\|b - x^\star\| &\leqslant \left\|(b - x^\star)_{|\mathcal{S}}\right\| + \left\|x^\star_{|\mathcal{S}^c}\right\| \\
&\leqslant \left\|(x^t - \eta\mathbf{T}(x^t) - x^\star)_{|\mathcal{S}}\right\| + \left\|x^\star_{|\mathcal{S}^c}\right\| \\
&\leqslant \left\|\left[x^t - x^\star - \mathbf{T}(x^t) + \mathbf{T}(x^\star) + (1-\eta)\left(\mathbf{T}(x^t) - \mathbf{T}(x^\star)\right) - \eta\mathbf{T}(x^\star)\right]_{|\mathcal{S}}\right\| + \left\|x^\star_{|\mathcal{S}^c}\right\| \\
&\leqslant \left\|(x^t - x^\star - \mathbf{T}(x^t) + \mathbf{T}(x^\star))_{|\mathcal{S}}\right\| + |1-\eta| \left\|(\mathbf{T}(x^t) - \mathbf{T}(x^\star))_{|\mathcal{S}}\right\| + \eta \left\|\mathbf{T}(x^\star)_{|\mathcal{S}}\right\| + \left\|x^\star_{|\mathcal{S}^c}\right\| \\
&\leqslant \alpha_{2k} \left\|x^t - x^\star\right\| + |1-\eta| \left(1 + \alpha_{2k}\right) \left\|x^t - x^\star\right\| + \eta \left\|\mathbf{T}(x^\star)_{|2k}\right\| + \left\|x^\star_{|\mathcal{S}^c}\right\| \qquad (49) \\
&\leqslant \left(\alpha_{2k} + |1-\eta| \left(1 + \alpha_{2k}\right)\right) \left\|x^t - x^\star\right\| + \eta \left\|\mathbf{T}(x^\star)_{|2k}\right\| + \left\|x^\star_{|\mathcal{S}^c}\right\| \qquad (50)
\end{aligned}$$

where Eq. (49) holds because $\mathbf{T}$ is URDP of order $2k$ with $D_{2k} = \mathbf{I}$, and $\text{card}(\mathcal{S}) \leqslant 2k$.

Updating the support set (Step 4)

Notice that

$$\left\|x^\star_{|\mathcal{T}^c}\right\| = \left\|(b_{|k} - x^\star)_{|\mathcal{T}^c}\right\| \leqslant \left\|b_{|k} - x^\star\right\| \leqslant \left\|b_{|k} - b\right\| + \|b - x^\star\| \ .$$

But $\|b - x^\star\| \leqslant \left\|b_{|k} - b\right\|$ since $x^\star$ is k-sparse. Hence

$$\left\|x^\star_{|\mathcal{T}^c}\right\| \leqslant 2 \|b - x^\star\| \ . \tag{51}$$

Optimization over the updated support (Step 5)

We have

$$\begin{aligned}
\left\|x^{t+1} - x^\star\right\| &\leqslant \left\|(x^{t+1} - x^\star)_{|\mathcal{T}}\right\| + \left\|x^\star_{|\mathcal{T}^c}\right\| \\
&\leqslant \tfrac{1}{1-\alpha_{2k}} \left\|\mathbf{T}(x^\star)_{|\mathcal{T}}\right\| + \tfrac{\alpha_{2k}}{1-\alpha_{2k}} \left\|x^\star_{|\mathcal{T}^c}\right\| + \left\|x^\star_{|\mathcal{T}^c}\right\| \qquad (52) \\
&\leqslant \tfrac{1}{1-\alpha_{2k}} \left\|\mathbf{T}(x^\star)_{|\mathcal{T}}\right\| + \tfrac{1}{1-\alpha_{2k}} \left\|x^\star_{|\mathcal{T}^c}\right\| \qquad (53)
\end{aligned}$$

where Eq. (52) holds by applying Lemma 18 using that x^{t+1} solves Problem (P4) on $\mathcal{T}$, that $x^\star$ is k-sparse and $\mathbf{T}$ is URDP of order $2k$.

Let us finally combine Eq. (18), (50), (51) and (53) to obtain

$$\begin{aligned}
\left\|x^{t+1} - x^\star\right\| &\leqslant \tfrac{1}{1-\alpha_{2k}} \left\|\mathbf{T}(x^\star)_{|\mathcal{T}}\right\| + \tfrac{1}{1-\alpha_{2k}} \left\|x^\star_{|\mathcal{T}^c}\right\| \\
&\leqslant \tfrac{1}{1-\alpha_{2k}} \left\|\mathbf{T}(x^\star)_{|\mathcal{T}}\right\| + \tfrac{2}{1-\alpha_{2k}} \|b - x^\star\| \\
&\leqslant \tfrac{1}{1-\alpha_{2k}} \left\|\mathbf{T}(x^\star)_{|\mathcal{T}}\right\| + \tfrac{2}{1-\alpha_{2k}} \left[\left(\alpha_{2k} + |1-\eta| \left(1 + \alpha_{2k}\right)\right) \left\|x^t - x^\star\right\| + \eta \left\|\mathbf{T}(x^\star)_{|2k}\right\| + \left\|x^\star_{|\mathcal{S}^c}\right\|\right] \\
&\leqslant \tfrac{1+2\eta}{1-\alpha_{2k}} \left\|\mathbf{T}(x^\star)_{|2k}\right\| + \tfrac{2(\alpha_{2k} + |1-\eta|(1+\alpha_{2k}))}{1-\alpha_{2k}} \left\|x^t - x^\star\right\| + \tfrac{2}{1-\alpha_{2k}} \left\|x^\star_{|\mathcal{S}^c}\right\| \\
&\leqslant \tfrac{1+2\eta}{1-\alpha_{2k}} \left\|\mathbf{T}(x^\star)_{|2k}\right\| + \tfrac{2(\alpha_{2k} + |1-\eta|(1+\alpha_{2k}))}{1-\alpha_{2k}} \left\|x^t - x^\star\right\| + \tfrac{2}{1-\alpha_{2k}} \left(\tfrac{2\alpha_{2k}}{(1-\alpha_{2k})} \left\|x^t - x^\star\right\| + \tfrac{2}{1-\alpha_{2k}} \left\|\mathbf{T}(x^\star)_{|2k}\right\|\right) \\
&\leqslant \tfrac{(1+2\eta)(1-\alpha_{2k})+4}{(1-\alpha_{2k})^2} \left\|\mathbf{T}(x^\star)_{|2k}\right\| + 2\tfrac{(\alpha_{2k} + |1-\eta|(1+\alpha_{2k}))(1-\alpha_{2k})+2\alpha_{2k}}{(1-\alpha_{2k})^2} \left\|x^t - x^\star\right\| \\
&\leqslant 2\tfrac{|1-\eta|\left(1-\alpha_{2k}^2\right)+(3-\alpha_{2k})\alpha_{2k}}{(1-\alpha_{2k})^2} \left\|x^t - x^\star\right\| + \tfrac{(1+2\eta)(1-\alpha_{2k})+4}{(1-\alpha_{2k})^2} \left\|\mathbf{T}(x^\star)_{|2k}\right\|
\end{aligned}$$

We have

$$2\frac{|1-\eta|\left(1-\alpha^2\right)+(3-\alpha)\alpha}{(1-\alpha)^2} \leqslant \frac{1}{2} \Leftrightarrow (5+4\,|1-\eta|)\alpha^2 - 14\alpha + 1 - 4\,|1-\eta| \geqslant 0 \tag{54}$$

Notice that $h^\eta \mapsto h^\eta(x) = (5+4\,|1-\eta|)x^2 - 14x + 1 - 4\,|1-\eta|$ verifies

$$\exists\ \alpha^\eta > 0 \text{ s.t. } h^\eta(x) \geqslant 0,\ \forall x \in [0, \alpha^\eta] \Leftrightarrow |1-\eta| < \tfrac{1}{4}.$$

And in this case, α^η is the smallest root of h^η:

$$\alpha^\eta = \frac{14-\sqrt{14^2-4(5+4|1-\eta|)(1-4|1-\eta|)}}{2(1-4|1-\eta|)}.$$

If η verifies $|1-\eta| < \frac{1}{4}$ then $7-2\sqrt{11} = \alpha^1 \leqslant \alpha^\eta \leqslant \lim_{|1-\eta|\to\frac{1}{4}} \alpha^\eta = \frac{3}{7}$ which finishes the proof. □

B.0.3 Proof of GIHT's error bound (Theorem 14)

The proof relies on the Uniform Restricted Diagonal Property of $\mathbf{T}$.

Proof of Theorem 14. Let $x^\star \in \mathcal{H}$ be a k-sparse vector, $x^t \in \mathcal{H}$ the t-th iterate of Algo. 4 and $\mathcal{S} = \operatorname{supp}(x^t) \cup \operatorname{supp} x^\star$. Assume $\mathbf{T}$ has the *Uniform Restricted Diagonal Property* of order $2k$ with $\mathbf{D}_{2k} = \mathbf{I}$ and $\alpha_{2k} \leqslant \alpha^\eta = \frac{1-4|\eta-1|}{4(1+|\eta-1|)}$ and $\frac{3}{4} < \eta < \frac{5}{4}$. Define $\mathcal{R} = \operatorname{supp}(x^t) \cup \operatorname{supp}(x^{t+1}) \cup \operatorname{supp}(x^\star)$.

Since $x^{t+1} = b_{|k}$ with $b = x^t - \eta\mathbf{T}(x^t)$, we have

$$\begin{aligned}
\left\|x^{t+1} - x^\star\right\| &= \left\|b_{|k} - x^\star\right\| \\
&\leqslant \left\|b_{|k} - b_{|\mathcal{R}}\right\| + \left\|b_{|\mathcal{R}} - x^\star\right\| \\
&\leqslant 2\left\|b_{|\mathcal{R}} - x^\star\right\| \qquad (55)\\
&\leqslant 2\left\|\left(x^t - x^\star - \eta\mathbf{T}(x^t)\right)_{|\mathcal{R}}\right\| \\
&\leqslant 2\left\|\left(x^t - x^\star - \left(\mathbf{T}(x^t) - \mathbf{T}(x^\star)\right) + (1-\eta)\left(\mathbf{T}(x^t) - \mathbf{T}(x^\star)\right) - \eta\mathbf{T}(x^\star)\right)_{|\mathcal{R}}\right\| \\
&\leqslant 2\left\|x^t - x^\star - \mathbf{T}(x^t) + \mathbf{T}(x^\star)\right\| + 2\,|\eta-1|\left\|\mathbf{T}(x^t) - \mathbf{T}(x^\star)\right\| + 2\eta\left\|\mathbf{T}(x^\star)_{|\mathcal{R}}\right\| \\
&\leqslant 2\alpha_{2k}\left\|x^t - x^\star\right\| + 2\,|\eta-1|\,(1+\alpha_{2k})\left\|x^t - x^\star\right\| + 2\eta\left\|\mathbf{T}(x^\star)_{|3k}\right\| \qquad (56)\\
&\leqslant 2(\alpha_{2k} + |\eta-1|\,(1+\alpha_{2k}))\left\|x^t - x^\star\right\| + 2\eta\left\|\mathbf{T}(x^\star)_{|3k}\right\|\ ,
\end{aligned}$$

where Eq. (55) holds because $x^\star$ is k-sparse and $b_{|k}$ is the best k-sparse approximation of $b_{|\mathcal{R}}$ since $\operatorname{supp}(b_{|k}) = \operatorname{supp}(x^{t+1}) \subseteq \mathcal{R}$; and Eq. (56) holds because $\mathbf{T}$ has the *Uniform Restricted Diagonal Property* of order $2k$ with $\mathbf{D}_{2k} = \mathbf{I}$.

Noticing that $2(\alpha_{2k} + |\eta-1|\,(1+\alpha_{2k})) \leqslant \frac{1}{2} \Leftrightarrow \alpha_{2k} \leqslant \alpha^\eta = \frac{1-4|\eta-1|}{4(1+|\eta-1|)}$ finishes the proof. □

References

Bahmani, S., Raj, B., Boufounos, P., 2013. Greedy sparsity-constrained optimization. J. of Machine Learning Research 14 (3), 807–841.

Bauschke, H. H., Combettes, P. L., 2010. The baillon-haddad theorem revisited. J. of Convex Analysis 17 (4), 781–787.

Beck, A., Hallak, N., 2015. On the minimization over sparse symmetric sets: Projections, optimality conditions, and algorithms. Mathematics of Operations Research.

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

Boulanger, J., Kervrann, C., Bouthemy, P., Elbau, P., Sibarita, J.-B., Salamero, J., 2010. Patch-based nonlocal functional for denoising fluorescence microscopy image sequences. IEEE Trans. Med. Imaging 29 (2), 442–454.

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

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

Chierchia, G., Pustelnik, N., Pesquet, J.-C., Pesquet-Popescu, B., 2012. Epigraphical projection and proximal tools for solving constrained convex optimization problems: Part i. arXiv preprint arXiv:1210.5844.

Combettes, P. L., Pesquet, J.-C., 2007. A Douglas-Rachford splittting approach to nonsmooth convex variational signal recovery. IEEE J. Selec. Top. Sig. Pro. 1 (4), 564–574.

Combettes, P. L., Pesquet, J.-C., 2008. A proximal decomposition method for solving convex variational inverse problems. Inv. Prob. 24 (6).

Combettes, P. L., Pesquet, J.-C., 2012. Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators. Set-Valued and variational analysis 20 (2), 307–330.

Dai, W., Milenkovic, O., 2009. Subspace pursuit for compressive sensing signal reconstruction. IEEE Transactions on Information Theory 55 (5), 2230–2249.

Foucart, S., 2011. Hard thresholding pursuit: an algorithm for compressive sensing. SIAM Journal on Numerical Analysis 49 (6), 2543–2563.

Jain, P., Tewari, A., Dhillon, I. S., 2011. Orthogonal matching pursuit with replacement. In: Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., Weinberger, K. (Eds.), Advances in Neural Information Processing Systems 24. pp. 1215–1223.

Jain, P., Tewari, A., Kar, P., 2014. On iterative hard thresholding methods for high-dimensional m-estimation. In: Advances in Neural Information Processing Systems. pp. 685–693.

Jalali, A., Johnson, C. C., Ravikumar, P. K., 2011. On learning discrete graphical models using greedy methods. In: Advances in Neural Information Processing Systems 24. pp. 1935–1943.

Jung, S.-M., 2011. Hyers-Ulam-Rassias stability of functional equations in nonlinear analysis. Vol. 48. Springer Science & Business Media.

Lemaréchal, C., Sagastizábal, C., 1997. Practical Aspects of the Moreau–Yosida Regularization: Theoretical Preliminaries. SIAM J. on Optimization 7 (2), 367–385.

Makitalo, M., Foi, A., 2011. Optimal inversion of the anscombe transformation in low-count poisson image denoising. Image Processing, IEEE Trans. on 20 (1), 99–109.

Mallat, S. G., Zhang, Z., 1993. Matching pursuits with time-frequency dictionaries. Signal Processing, IEEE Trans. on 41 (12), 3397–3415.

Needell, D., Tropp, J. A., 2009. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. Applied and Computational Harmonic Analysis 26 (3), 301–321.

Temlyakov, V. N., 2008. Greedy approximation. Acta Numerica 17, 235–409.

Vaiter, S., Deledalle, C.-A., Peyré, G., Dossal, C., Fadili, J., 2012. Local behavior of sparse analysis regularization: Applications to risk estimation. Applied and Computational Harmonic Analysis.

Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E., 2004. Image quality assessment: From error visibility to structural similarity. Image Processing, IEEE Trans. on 13 (4).

Yang, Z., Wang, Z., Liu, H., Eldar, Y. C., Zhang, T., 2016. Sparse nonlinear regression: Parameter estimation under nonconvexity. In: Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016. pp. 2472–2481.

Yuan, X., Li, P., Zhang, T., 2014. Gradient hard thresholding pursuit for sparsity-constrained optimization. In: The 31st International Conference on Machine Learning. pp. 127–135.

Zhang, B., Fadili, J. M., Starck, J.-L., 2008. Wavelets, ridgelets, and curvelets for Poisson noise removal. Image Processing, IEEE Trans. on 17 (7), 1093–1108.

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