Generalized Greedy Alternatives

François-Xavier Dupé, Sandrine Anthoine

Generalized Greedy Alternatives[☆]

François-Xavier Dupé[a], Sandrine Anthoine[b]

[a]*Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France*
[b]*Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France*

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, zeros-finding, non-convex optimization, Poisson denoising.

1. Introduction

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

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

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 broad categories of methods have been proposed to tackle such problems. *Greedy methods* on one side, focus on the original combinatorial problem and try to solve it by taking local decisions. Relaxation methods, on the other side, relax (P0) by replacing the ℓ_0 pseudo-norm ($\|\cdot\|_0$) with a more tractable norm like the convex ℓ_1 norm.

These *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

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

Email addresses: francois-xavier.dupe@lis-lab.fr (François-Xavier Dupé), sandrine.anthoine@univ-amu.fr (Sandrine Anthoine)

considering other penalty terms (Kullback-Leibler divergence, hinge-loss) (Combettes and Pesquet, 2007), by modifying the sparse structure to group-sparsity (Peyré and Fadili, 2011)... 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 be 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 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?

1.1. A quick review of greedy methods in the linear setting

To find the best k-sparse approximation of u on $\mathbf{\Phi}$, that is to solve (P0), *greedy methods* take local decisions. In fact, they look for 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 corresponding coefficients is of course estimated in the process as well. *Matching Pursuit* (MP) and *Orthogonal Matching Pursuit* (OMP) (Mallat and Zhang, 1993) were the first greedy techniques employed to solve Problem (P0). These are forward methods in the sense that they add one element in the support at each iteration and never question this choice. Later, two-stage methods were proposed that work on the whole support of size k globally. At each iteration, the first stage widens the candidate support to allow for new directions to be explored and the second stage prunes it to select the most appropriate k directions. These two-stages methods have the advantage of keeping the computationnal burden fixed since the support size is fixed. In this paper, we will focus on these methods and more specifically on the four following procedures: *Compressed Sampling Matching Pursuit* (CoSaMP) (Needell and Tropp, 2009), *Subspace Pursuit* (SP) (Dai and Milenkovic, 2009), *Hard Thresholding Pursuit* (HTP) (Foucart, 2011) and *Iterative Hard Thresholding* (IHT) (Blumensath and Davies, 2008).

Whether one-stage or two-stage methods, the greedy procedures mentioned above come with guarantees of convergence and convergence rates that rely on the good behavior of the dictionary $\mathbf{\Phi}$ with respect to the subspaces of dimension k, such as the *Restricted Isometry Property* (RIP), *spark* or *mutual (in)coherence* (Tropp, 2006).

1.2. Greedy methods for finding a k-sparse minimizer

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 $\frac{1}{2}\|\mathbf{\Phi}x - u\|_2^2$ in Problem (P0) with a more generic function). Let us extend (P0) to this more general setting. Let $\mathcal{H}$ be a real Hilbert space, given an objective function $f : \mathcal{H} \to \mathbb{R}$, we want to solve

$$\text{find } \hat{x} \in \underset{x \in \mathcal{H}}{\operatorname{argmin}} f(x) \text{ s.t. } \|x\|_0 \leq k, \tag{P1}$$

i.e. we seek for a k-sparse minimizer of f.

The previous tools such as *Restricted Isometry Property* can not guarantee the behavior of the algorithms in this setting because the linearity is lost. New tools had to be developed. Among the first works in this spirit, one can find results on OMP (Shalev-Shwartz et al., 2010; 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 function f with the two following restrictions: i) f has to have some convexity property on the k-dimensional subspaces (and must generally be smooth), ii) $\mathcal{H}$ is a finite dimensional Hilbert space. One of our contribution is to break these two barriers.

1.3. Greedy methods for finding a k-sparse solution of non-linear equations

To go a step further, note that if f is a convex differentiable function, then using the duality between fixed-point theory and convex analysis (Bauschke et al., 2005, 2017), we can reformulate (P1) as

$$\text{find } x \in \mathcal{H} \text{ s.t. } \nabla f(x)_{|\operatorname{supp}(x)} = 0 \text{ and } \|x\|_0 \leq k \,.$$

In other words, we seek a k-sparse vector x such that the gradient of f at x is null on the support of x. With this formulation (P1) is seen through the light of fixed-point theory. Now, we generalize this problem by replacing the gradient ∇f, which is an operator from $\mathcal{H}$ to itself by a generic operator $\mathbf{T} : \mathcal{H} \to \mathcal{H}$. For commodity, we also drop the restriction to the support. This leads to the problem we wish to solve.

$$\text{Find } x \in \mathcal{H} \text{ s.t. } \mathbf{T}(x) = 0 \text{ and } \|x\|_0 \leq k \,. \tag{P2}$$

In words, Problem (P2) aims at finding a k-sparse vector that satisfies the non-linear set of equation $\mathbf{T}(x) = 0$. Note that this problem may not always have a solution. For example, if $\mathbf{T}$ is the differential of a convex function, Problem (P2) amounts to finding a minimizer, both global and k-sparse. Such a minimizer may not exist. We will come back to this. Let us first stress that compared to the literature, that focuses on minimization problems as seen in the previous section, 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.

In this paper, we adapt four classical *greedy methods*, namely CoSaMP, *Subspace Pursuit*, *Iterative Hard thresholding* and *Hard Thresholding Pursuit* to solve Problem (P2). We develop a property on $\mathbf{T}$, the *Restricted Diagonal Property*, that is quite inspired by the Restricted Isometry Property and guarantees the good behavior of these generalizations. We show in particular that if Problem (P2) 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 (P2). 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 postponed to the Appendix.) Examples of applications are given in Section 5 and numerical experiments in Section 6.

1.5. Notation

Before describing the algorithms, 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).
- $x_{|\mathcal{R}}$ 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

$$x_{|k} = \sum_{i=1}^{k} x_{\psi(i)} e_{\psi(i)} . \tag{3}$$

2. Four Generalized Greedy Algorithms

Let us start with explaining how to adapt what we call here *greedy methods* to solve Problem (P2). The seminal methods that were designed to find the best k-sparse approximation on a dictionary $\mathbf{\Phi} = (\phi_1, \ldots, \phi_M)$ 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 that does so is CoSaMP (Needell and Tropp, 2009), other have followed such as *Subspace Pursuit* (SP) (Dai and Milenkovic, 2009), *Hard Thresholding Pursuit* (HTP) (Foucart, 2011) or *Iterative Hard Thresholding* (IHT) (Blumensath and Davies, 2008) - which may be seen as a simpler version (less focused on the support).

In the following, we first focus on SP and explain in Section 2.1 how to extend it to solve Problem (P1) then (P2). In Section 2.2, we then explain more concisely the similar extension for CoSaMP, HTP and IHT.

2.1. Extending Subspace Pursuit from (P0) *to* (P2)

Let us detail the example of *Subspace Pursuit* for (P0). Given a dictionary $\mathbf{\Phi}$, the sparsity k and the observation u, *Subspace Pursuit* (Algorithm 1) looks for the support $\mathcal{T}$ of the solution x of Problem (P0) or in other words the corresponding columns of $\mathbf{\Phi}$. To do so, it first doubles the number of considered columns (line 6) compared to the targeted sparsity by adding columns corresponding to the largest correlation with the residual u_r ($\mathcal{G}$). It then computes an estimate (line 7) which belongs to this subspace ($\mathbf{\Phi}_{|\mathcal{S}}$ is the dictionary with only the columns whose indexes are in $\mathcal{S}$, and $\mathbf{\Phi}_{|\mathcal{S}}^{\dagger}$ the pseudo-inverse of $\mathbf{\Phi}_{|\mathcal{S}}$). Then it halves the number of columns (line 8) by considering the columns involved by the biggest coefficients (in magnitude) of the estimate and computes a new estimate (line 9). One stage of exploration where the number of columns is twice the sparsity ($\mathcal{S}$), one stage of pruning where a new estimate is computed using only the most important columns in the previous stage ($\mathcal{T}$).

Algorithm 1 Subspace Pursuit (Dai and Milenkovic, 2009)

1: **Require:** the dictionary $\mathbf{\Phi}$, the sparsity k, the observation u .
2: **Initialization:** $\mathcal{S} \leftarrow \emptyset$, $\mathcal{T} \leftarrow \emptyset$, $u_r \leftarrow u$, $x \leftarrow 0$.
3: **repeat**
4: $\quad \mathcal{G} \leftarrow \{k \text{ indexes corresponding to the largest absolute values of } \mathbf{\Phi}^* u_r\}$.
5: $\quad \mathcal{S}_{old} \leftarrow \mathcal{S}$.
6: $\quad \mathcal{S} \leftarrow \mathcal{G} \cup \mathcal{T}$.
7: $\quad b \leftarrow \mathbf{\Phi}_{|\mathcal{S}}^{\dagger} u$.
8: $\quad \mathcal{T} \leftarrow \{k \text{ indexes corresponding to the largest absolute values of } b\}$.
9: $\quad x \leftarrow \mathbf{\Phi}_{|\mathcal{T}}^{\dagger} u$.
10: $\quad u_r \leftarrow u - \mathbf{\Phi}_{|\mathcal{T}} \mathbf{\Phi}_{|\mathcal{T}}^{\dagger} u = u - \mathbf{\Phi}_{|\mathcal{T}} x$.
11: **until** $\mathcal{S}_{old} = \mathcal{S}$ or $u_r = 0$.
12: **Output:** x.

To extend Subspace Pursuit from (P0) to (P1), let us start by noting that in Algorithm 1, one produces at each iteration, two sparse approximations, one of size $2k$: $b = \mathbf{\Phi}_{|\mathcal{S}}^{\dagger} u$, and one of size k: $x = \mathbf{\Phi}_{|\mathcal{T}}^{\dagger} u$. Both can be seen as the minimizers of a restriction of Problem (P0) to the case where the support is fixed:

$$x = \underset{z}{\operatorname{argmin}} \tfrac{1}{2} \|\mathbf{\Phi} z - u\|_2^2 \text{ s.t. } \operatorname{supp}(z) \subseteq \mathcal{T} \ ,$$

and

$$b = \underset{z}{\operatorname{argmin}} \tfrac{1}{2} \|\mathbf{\Phi} z - u\|_2^2 \text{ s.t. } \operatorname{supp}(z) \subseteq \mathcal{S}.$$

We have $u_r = u - \mathbf{\Phi} x$, one can thus rewrite Algorithm 1 in terms of the solutions x and b, and have u_r disappear. Additionally, noting $f(x) = \frac{1}{2} \|\mathbf{\Phi} x - u\|_2^2$, we deduce that $\mathbf{\Phi}^* u_r = \mathbf{\Phi}^*(u - \mathbf{\Phi}_{|\mathcal{T}} x) = -\nabla f(x)$. Thus the directions added to $\mathcal{S}$ at line 6 of Algorithm 1 correspond to those of largest amplitude in $\nabla f(x)$. Putting these remarks together, we obtain that Algorithm 1 can be rewritten as Algorithm 2 for the particular case of $f(x) = \frac{1}{2} \|\mathbf{\Phi} x - u\|_2^2$. This first extension is based on the same ideas as the extension of CoSaMP to Problem (P1) described in (Bahmani et al., 2013).

Algorithm 2 Subspace Pursuit for (P1)

1: **Require:** f, k .
2: **Initialization:** $x \leftarrow 0, \mathcal{S} \leftarrow \emptyset$.
3: **repeat**
4: $\quad \mathcal{G} \leftarrow \text{supp}(|\nabla f(x)|_{|k})$,
5: $\quad \mathcal{S}_{old} \leftarrow \mathcal{S}$,
6: $\quad \mathcal{S} \leftarrow \mathcal{G} \cup \text{supp}(x)$,
7: $\quad b \in \text{argmin}_{\{x \text{ s.t. } \text{supp}(x) \subseteq \mathcal{S}\}} f(x)$
8: $\quad \mathcal{T} \leftarrow \text{supp}(b_{|k})$,
9: $\quad x \in \text{argmin}_{\{x \text{ s.t. } \text{supp}(x) \subseteq \mathcal{T}\}} f(x)$
10: **until** $\mathcal{S}_{old} = \mathcal{S}$.
11: **Output:** x.

Algorithm 3 Generalized Subspace Pursuit

1: **Require:** $\mathbf{T}, k$.
2: **Initialization:** $x \leftarrow 0, \mathcal{S} \leftarrow \emptyset$.
3: **repeat**
4: $\quad \mathcal{G} \leftarrow \text{supp}(\mathbf{T}(x)_{|k})$,
5: $\quad \mathcal{S}_{old} \leftarrow \mathcal{S}$,
6: $\quad \mathcal{S} \leftarrow \mathcal{G} \cup \text{supp}(x)$,
7: $\quad b \in \mathcal{H}$ s.t. $\begin{cases} \text{supp}(b) \subseteq \mathcal{S} \\ \mathbf{T}(b)_{|\mathcal{S}} = 0. \end{cases}$
8: $\quad \mathcal{T} \leftarrow \text{supp}(b_{|k})$,
9: $\quad x \in \mathcal{H}$ s.t. $\begin{cases} \text{supp}(x) \subseteq \mathcal{T} \\ \mathbf{T}(x)_{|\mathcal{T}} = 0. \end{cases}$
10: **until** $\mathcal{S}_{old} = \mathcal{S}$.
11: **Output:** x.

Now to extend SP furthermore to Problem (P2), we observe that the minimizer of f over a specific set $\mathcal{S}$ is the solution of a gradient equation:

$$x \in \underset{\{x \text{ s.t. } \text{supp}(x) \subseteq \mathcal{T}\}}{\text{argmin}} f(x) \Leftrightarrow \begin{cases} \text{supp}(x) \subseteq \mathcal{T} \\ \nabla f(x)_{|\mathcal{T}} = 0. \end{cases}$$

Finally, one formally identifies ∇f with the operator $\mathbf{T}$ and obtains the *generalized Subspace Pursuit Algorithm* described in Algorithm 3, which is a candidate to solve Problem (P2).

We now have a formal algorithm, heuristically designed to solve Problem (P2). First, one needs to ensure that each step of the algorithm is well-defined. This raises the question of the existence of b and x, which are both solutions of optimization problems of the form

$$\text{find } z \in \mathcal{H} \text{ s.t. } \text{supp}(z) \subseteq \mathcal{R} \text{ and } \mathbf{T}(z)_{|\mathcal{R}} = 0 \,, \tag{P3}$$

for a particular set $\mathcal{R}$ of finite size. These problems do not have a solution for all $\mathbf{T}$ and $\mathcal{R}$. However, when $\mathbf{T}$ is the differential of a convex function, the problem corresponds to a minimization among the vectors supported in $\mathcal{R}$ which does exist under mild conditions. More generally, a solution exists when $\mathbf{I} - \mathbf{T}$ is a contraction for example. We will see in Section 4.1 that we need a less strong property on $\mathbf{T}$ to ensure that this algorithm is well-defined and guarantee its convergence at the same time. This property is called the *Restricted Diagonal Property*.

Before going into these details, we describe in the next section, how to generalize CoSaMP, HTP and IHT in a similar manner.

2.2. Three other generalized algorithms

CoSaMP differs from Subspace Pursuit for solving (P0) only in two aspects:

- one adds $2k$ directions from $\mathbf{\Phi}^* u_r = \mathbf{\Phi}^*(u - \mathbf{\Phi}x)$ to $\mathcal{T}$ to obtain a set $\mathcal{S}$ of size $3k$ (line 6 of Algorithm 1).
- the solution $x = \mathbf{\Phi}^{\dagger}_{|\mathcal{T}} u$ is not computed, but a computationally cheaper approximate is used: one only keeps the best k-sparse approximation of b: $x = b_{|k}$ (line 9 of Algorithm 1).

Doing so, we obtain the *Generalized CoSaMP* algorithm described in Algorithm 4, where we have marked by (*) the steps that differ from Algorithm 3.

Similarly HTP differs from *Subspace Pursuit* for solving (P0) only in one aspect. The solution $b = \mathbf{\Phi}_{|\mathcal{S}}^{\dagger} u$ is not computed, but a cheaper approximate is computed using a gradient step: $b = (x - \eta \mathbf{\Phi}^*(\mathbf{\Phi} x - u))_{|\mathcal{S}}$. This expression translates using the same arguments as previously in $b = (\mathbf{I} - \eta \mathbf{T})(x)_{|\mathcal{S}}$. The *Generalized HTP* algorithm obtained is described in Algorithm 5, where we have marked by (*) the step that differs from Algorithm 3.

Finally IHT makes the same approximation as HTP for b using a gradient step, and the same approximation as CoSaMP for x using the best k-term approximation. This leads to the *Generalized IHT* algorithm described in Algorithm 6.

Algorithm 4 Generalized CoSaMP

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

Algorithm 5 Generalized HTP

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

Algorithm 6 Generalized IHT

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

As noted in the previous Section, the presented generalized Algorithms (Algorithm 3 to 6) are heuristic, and we need to ensure that they are well-defined. The problem is posed for lines 7 and 9 of GSP, line 7 of GCoSaMP and line 9 of GHTP which are all instances of Problem (P3) with a finite set $\mathcal{R}$ of size k, $2k$ or $3k$. We show in the next Section how to guarantee the existence

of solutions to these problems.

3. The Restricted Diagonal Property

To ensure that Algorithms 3 to 6 are well-defined, but also to guarantee error bounds on the iterates, we will ask the operator $\mathbf{T}$ to fulfill the *Restricted Diagonal Property* that we introduce here and which was inspired by the *Restricted Isometry Property*.

3.1. Behind the Restricted Isometry Property

The *Restricted Isometry Property* (RIP) is a powerful tool to analyze algorithms for solving (P0) (Candès et al., 2006). Let $\mathbf{A}$ be a linear operator, $\mathbf{A}$ fulfills the δ_k-RIP when

$$(1-\delta_k)\,\|x\|^2 \leq \|\mathbf{A}(x)\|^2 \leq (1+\delta_k)\,\|x\|^2\,, \quad \forall\ x\ \text{ s.t. }\ \|x\|_0 \leq k\,. \tag{4}$$

In essence, RIP ensures that $\mathbf{A}$ behaves like an isometry on the set of k-sparse vectors, and thus that one can identify without errors k-sparse vectors from their image by $\mathbf{A}$. If $\mathbf{A}$ is δ_k-RIP then $\mathbf{A}^*\mathbf{A}$ is close to the identity on k-sparse vectors (Needell and Tropp, 2009):

$$\|\mathbf{A}^*\mathbf{A}(x) - x\| \leq \delta_k\,\|x\|\,, \quad \forall\ x\ \text{ s.t. }\ \|x\|_0 \leq k\,. \tag{5}$$

or equivalently

$$\|\mathbf{A}^*\mathbf{A}(x) - \mathbf{A}^*\mathbf{A}(y) - \mathbf{I}(x-y)\| \leq \delta_k\,\|x-y\|\,, \quad \forall\ x,\ y\ \text{ s.t. }\ \mathrm{card}(\mathrm{supp}(x,y)) \leq k\,. \tag{6}$$

In our generalizations, we essentially want to replace $\mathbf{A}^*\mathbf{A}$ by a generic operator $\mathbf{T}$ and thus lose the linearity and invariance to scaling. To account for the non-linearity, we start from Eq. (6) rather than Eq. (5). To account for possible scalings, we will replace the identity by diagonal operators. This leads us to introduce the following *restricted diagonal property* (RDP).

3.2. The Restricted Diagonal Property

Let us introduce $\mathcal{D}_1$ the set of diagonal operators bounded away by 1:

$$\mathcal{D}_1 = \left\{ \mathbf{D} : \begin{array}{ccc} \mathcal{H} & \to & \mathcal{H}, \\ x & \mapsto & \sum_i d_i x_i e_i \end{array} \quad \text{s.t. } \forall x\ \|\mathbf{D}x\| \geq \|x\| \right\} \tag{7}$$

$\mathbf{T}$ has the *Restricted Diagonal Property* of order k if it looks like a diagonal operator bounded away from zero (an operator of $\mathcal{D}_1$) locally on k-sparse vectors.

Definition 1 (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*

$$\forall (x,y) \in \mathcal{H}^2,\ \mathrm{supp}(x,y) \subseteq \mathcal{S} \Rightarrow \|\mathbf{T}(x) - \mathbf{T}(y) - \mathbf{D}_\mathcal{S}(x-y)\| \leq \alpha_k\,\|x-y\|\,. \tag{8}$$

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

Remark 2. *When $\mathbf{T}$ has the RDP of order $2k$ with $\alpha_{2k} < 1$, 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 $\|x\|_0 \leq k$ and $\|y\|_0 \leq k$.*

In the restricted case where these diagonal operators $\mathbf{D}_{\mathcal{S}}$ are all equal to the identity, one recovers properties that exists in the literature. Firstly, for $\mathbf{A}$ linear and δ_k-RIP, then $\mathbf{T} = \mathbf{A}^*\mathbf{A}$ has the RDP with the operators $\mathbf{D}_{\mathcal{S}}$ all equal to the identity. Secondly, for $\mathbf{T} = \nabla f$ with $f : \mathcal{H} \to \mathbb{R}^+$ smooth, if f has the *Restricted Strong Smoothness* and *Restricted Strong Convexity* properties developed in (Bahmani et al., 2013; Yuan et al., 2014) for the resolution of Problem (P1), then $\mathbf{T}$ is RDP with the operators $\mathbf{D}_{\mathcal{S}}$ all equal to the identity. RDP thus provides a generalization of these notions by allowing the $\mathbf{D}_{\mathcal{S}}$ to differ from identity. To show this, it is useful to characterize the RDP property as done in the next section.

3.3. Characterization of the Restricted Diagonal Property

Definition 3. *If $\mathbf{D} \in \mathcal{D}_1$ is bounded, we define $|||\mathbf{D}|||_k$ by*

$$|||\mathbf{D}|||_k \stackrel{def}{=} \sup_{\{x\neq y,\ \mathrm{card}(\mathrm{supp}(x,y))\leq k\}} \frac{||\mathbf{D}(x-y)||}{||x-y||} = \sup_{\{x\neq 0,\ \mathrm{card}(\mathrm{supp}(x))\leq k\}} \frac{||\mathbf{D}(x)||}{||x||} < \infty. \tag{9}$$

Theorem 4. *We have:*

1. *If $\beta\mathbf{T}$ is RDP of order k with $\mathbf{D}_{\mathcal{S}} = \mathbf{D}$ for all $\mathcal{S}$, $\alpha_k < 1$ and $\beta > 0$ then*

$$\forall (x,y) \in \mathcal{H}^2,\ \mathrm{card}(\mathrm{supp}(x,y)) \leq k \Rightarrow \begin{cases} ||\mathbf{T}(x) - \mathbf{T}(y)|| \leq \frac{|||\mathbf{D}|||_k + \alpha_k}{\beta} ||x-y|| \\ \langle \mathbf{T}(x) - \mathbf{T}(y), \mathbf{D}(x-y)\rangle \geq \frac{1-\alpha_k}{\beta} ||x-y||^2 . \end{cases} \tag{10}$$

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

$$\forall (x,y) \in \mathcal{H}^2,\ \mathrm{card}(\mathrm{supp}(x,y)) \leq k \Rightarrow \begin{cases} ||\mathbf{T}(x) - \mathbf{T}(y)|| \leq L ||x-y|| \\ \langle \mathbf{T}(x) - \mathbf{T}(y), \mathbf{D}(x-y)\rangle \geq m ||x-y||^2 , \end{cases} \tag{11}$$

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

Remark 5. *Consider the assumption $\frac{m^2}{L^2} \leq |||\mathbf{D}|||_k^2 \leq \frac{m^2}{L^2} + 1$ made in Theorem 4. Notice that Eq. (11) implies that $m \leq L|||\mathbf{D}|||_k$ so that the left inequality is always true. We also pinpoint that $|||\mathbf{D}|||_k^2 \leq \frac{m^2}{L^2} + 1$ always holds for $\mathbf{D} = \mathbf{I}$.*

Remark 6. *A similar theorem holds when $\mathbf{T}$ has the RDP and there exists a uniform bound of the type of Eq. (9) on the matrices $\mathbf{D}_{\mathcal{S}}$ in Definition 1.*

The proof is postponed to Appendix A. Theorem 4 shows that being RDP of order k for a constant and bounded diagonal operator $\mathbf{D}$ 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 $\langle \mathbf{T}(x) - \mathbf{T}(y), \mathbf{D}(x-y)\rangle$, both properties holding on the couples (x, y) such that $\mathrm{card}(\mathrm{supp}(x, y)) \leq k$.

For $\mathbf{T} = \nabla f$ with $f : \mathcal{H} \to \mathbb{R}^+$ smooth. Theorem 4 shows that a scaled version of ∇f is RDP 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 $\mathrm{card}(\mathrm{supp}(x, y)) \leq k$. This is exactly the *Restricted Strong Smoothness* and *Restricted Strong Convexity* developed in (Bahmani et al., 2013; Yuan et al., 2014) for the resolution of Problem (P1).

Let us also emphasize that Theorem 4 shows that the set of operators $\mathbf{T} = \nabla f$ which have the RDP with $\mathbf{D}_{\mathcal{S}} = \mathbf{D}$ but different from identity encompasses a greater set of functions than the classical notions seen above. Indeed as soon as $\mathbf{D}$ has at least one negative eigenvalue the second inequality in Eq. (10) does not yield convexity of f 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 3 to 6. 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 (P2) *and* (P3)

Proposition 7.

1. *Problem* (P2) *has at most one solution when* $\mathbf{T}$ *has the* Restricted Diagonal Property *of order* $2k$ *with* $\alpha_{2k} < 1$.
2. *Problem* (P3) *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 (P3) appears at step 3 or 5 of the algorithm, one deduces that

Corollary 8.

- *GSP (Algo. 3) is well-defined when* $\mathbf{T}$ *is an operator having the* Restricted Diagonal Property *of order* $2k$ *with* $\alpha_{2k} < 1$.
- *GCoSaMP (Algo 4) is well-defined when* $\mathbf{T}$ *is an operator having the* Restricted Diagonal Property *of order* $3k$ *with* $\alpha_{3k} < 1$.
- *GHTP (Algo 5) 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 7:

Proof.

1. As noted in Remark 2, 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 (P2) 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. (8) holds. Noting $\mathbf{D}_{\mathcal{R}}(z) = \sum_{i\in\mathcal{R}} d_i z_i e_i$, we have

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

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. (8), we obtain:

$$\forall (x, y) \in \operatorname{span}(e_i, i \in \mathcal{R})^2,\ \left\|\mathbf{T}^{\mathcal{R}}(x) - \mathbf{T}^{\mathcal{R}}(y) - (x - y)\right\| \leqslant \alpha_k \|x - y\| . \tag{12}$$

In other words $\mathbf{I} - \mathbf{T}^{\mathcal{R}}$ 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 $\mathbf{I}(x_0) - \mathbf{T}^{\mathcal{R}}(x_0) = x_0$ 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 (P3).

□

4.2. Error bounds

The 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 Subspace Pursuit:

Theorem 9. *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. 3), verifies*

$$\left\|x^N - x^\star\right\| \leqslant \frac{1}{2^N} \left\|x^0 - x^\star\right\| + 12\rho \left\|\mathbf{T}(x^\star)_{|2k}\right\| . \tag{13}$$

The first term $\frac{1}{2^N} \left\|x^0 - x^\star\right\|$ vanishes exponentially fast to zero. What is left is the *incompressible* error $\left\|\mathbf{T}(x^\star)_{|2k}\right\|$. Note that under the hypotheses of the theorem, and by Proposition 7, Problem (P2) has at most one solution. If this solution exists, let us note it x^{P2}, then $\mathbf{T}(x^{P2}) = 0$ and thus GSP converges exponentially fast to x^{P2}. 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 $\mathbf{T}(x)$ is the smallest.

A similar bound holds for GCoSaMP:

Theorem 10. *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. 4), verifies*

$$\left\|x^N - x^\star\right\| \leqslant \frac{1}{2^N} \left\|x^0 - x^\star\right\| + 12\rho \left\|\mathbf{T}(x^\star)_{|3k}\right\| . \tag{14}$$

Notice that a similar theorem was proposed in (Bahmani et al., 2013) for the special case where $\mathbf{T} = \nabla f$ is RDP with $\mathbf{D}_{\mathcal{S}} = \mathbf{I}$ for all $\mathcal{S}$ of size at most k.

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

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

$$\left\|x^N - x^\star\right\| \leqslant \frac{1}{2^N} \left\|x^0 - x^\star\right\| + 2\frac{(1+2\eta)(1-\alpha_{2k})+4}{(1-\alpha_{2k})^2} \left\|\mathbf{T}(x^\star)_{|2k}\right\| . \tag{15}$$

A similar bound has been shown by 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. 5), which has been proved to be beneficial by Jain et al. (2011).

For GIHT, the error bound reads:

Theorem 12. *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* Restricted Diagonal Property *of order $2k$ with $\mathbf{D}_\mathcal{S} = \mathbf{I}$ for all $\mathcal{S}$ of size at most $2k$ and $\alpha_{2k} \leqslant \alpha^\eta$. Then x^N, the N-th iterate of Generalized IHT (Algo. 6), verifies*

$$\left\|x^N - x^\star\right\| \leqslant \frac{1}{2^N}\left\|x^0 - x^\star\right\| + 4\eta\left\|\mathbf{T}(x^\star)_{|3k}\right\| . \tag{16}$$

The guarantees derived for the algorithms differ for the RDP bounds (α^S, α^C, α^H, α^η) and the 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 generic diagonal operators (for GSP and GCoSaMP) compared to the identity (for GIHT and GHTP).

Indeed, as we have seen in Section 3, 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 10 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 5 for details). Thus, Theorems 9 and 10 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 9 to 12 extend the scope of greedy algorithms to finding k-sparse zeros of operators (Problem (P2)).

5. 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 5.2), finding sparse approximations of minimizers of a non-differentiable convex function (Section 5.3), and finding sparse approximations of stationary points of a function that is neither concave nor convex but differentiable (Section 5.4). Let us emphasize that in Sections 5.1 and 5.3 $\mathcal{H}$ may be infinite dimensional.

5.1. Restricted Diagonal Property and second order differentiability

Let us first further characterize the RDP 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 $\mathcal{H}$)

$$\lambda_k(\mathbf{A}) \stackrel{\text{def}}{=} \inf_{\left\{\substack{(z,u),\ u\neq 0 \\ \text{card}(\text{supp}(z,u))\leq k}\right\}} \frac{\langle u, \mathbf{A}(z)(u)\rangle}{\|u\|^2}. \tag{17}$$

$$\Lambda_k(\mathbf{A}) \stackrel{\text{def}}{=} \sup_{\left\{\substack{(z,u),\ u\neq 0 \\ \text{card}(\text{supp}(z,u))\leq k}\right\}} \frac{\|\mathbf{A}(z)(u)\|}{\|u\|}. \tag{18}$$

The following characterization follows from Theorem 4

Theorem 13. *Assume $f : \mathcal{H} \to \mathbb{R}$ is twice differentiable on $\{x/\ \text{card}(\text{supp}(x)) \leq k\}$. Assume that $\mathbf{D}$ is in $\mathcal{D}_1$. If $0 < \lambda_k(\mathbf{D}^T\nabla^2 f)$, $\Lambda_k(\nabla^2 f) < \infty$, and $0 \leq |||D|||_k^2 - \frac{\lambda_k(\mathbf{D}^T\nabla^2 f)^2}{\Lambda_k(\nabla^2 f)^2} < 1$ then $(\beta\nabla f)$ is RDP of order k for $\mathbf{D}_\mathcal{S} = \mathbf{D}$ for all $\mathcal{S}$ of size at most k, 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. It also bears similarities with the "Sparse eigenvalue" criterion in (Yang et al., 2016). We give the example of logistic regression in the next section.

Note that this theorem also shows that RDP could also be used as an alternative to the new notions of coherence defined in (Jones et al., 2016) and (Adcock and Hansen, 2016), where the authors proposed an extension of compressed sensing (linear setting) to continuous spaces and in the infinite dimensional case.

5.2. 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 $\mu > 0$

$$f(x) = \tfrac{1}{N}\sum_{n=1}^{N} -\log(\mathbb{P}(l_n|y_n, x)) + \tfrac{1}{2}\mu\,\|x\|_2^2\ . \tag{19}$$

We have:

$$f(x) = \tfrac{1}{N}\sum_{n=1}^{N} (\log(1+\exp(\langle y_n, x\rangle)) - l_n\,\langle y_n, x\rangle) + \tfrac{1}{2}\mu\,\|x\|_2^2\ . \tag{20}$$

$$\nabla f(x) = \tfrac{1}{N}\sum_{n=1}^{N} \left(\frac{1}{1+\exp(\langle y_n, x\rangle)} - l_n\right) y_n + \mu x\,. \tag{21}$$

$$\nabla^2 f(x) = \tfrac{1}{N}\sum_{n=1}^{N} \frac{1}{(1+\exp(\langle y_n, x\rangle))(1+\exp(-\langle y_n, x\rangle))} y_n y_n^{\mathrm{T}} + \mu\mathbf{I}\,. \tag{22}$$

We thus have:

$$\forall x,\ \mu\mathbf{I} \preceq \nabla^2 f(x) \preceq \mu\mathbf{I} + \tfrac{1}{4N}\sum_{n=1}^{N} y_n y_n^{\mathrm{T}}\ . \tag{23}$$

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 $\mathrm{card}(\mathrm{supp}(u)) \leq k \Rightarrow u^{\mathrm{T}} y_n y_n^{\mathrm{T}} u \leq R\,\|u\|^2$. Hence $\Lambda_k(\nabla^2 f) \leq \mu + \frac{R}{4}$ so that $\beta\nabla f$ is RDP 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).

5.3. Application to Non-Smooth Convex Functions

In the last subsection, we considered a smooth function, but most of the functions in $\Gamma_0(\mathcal{H})$ (the set of functions from $\mathcal{H}$ to $]-\infty, +\infty]$ which are lower semicontinuous, convex, and proper) are not smooth. One way to deal with this 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:

$$\mathcal{M}_{\lambda,f}(x) = \inf_{y\in\mathcal{H}} \left[\tfrac{1}{2\lambda}\,\|x-y\|^2 + f(y)\right]\ . \tag{24}$$

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 functions, but it can also be used to regularize smooth functions (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,

$$\nabla\mathcal{M}_{\lambda,f} = (\mathbf{I} - \mathrm{prox}_{\lambda f})/\lambda\ , \tag{25}$$

with the proximal operator of f defined as,

$$\mathrm{prox}_{\lambda f} : x \mapsto \operatorname*{argmin}_{z\in\mathcal{H}} \lambda f(z) + \frac{1}{2}\,\|z-x\|\ . \tag{26}$$

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 RDP is shown (the natural Lipschitz property of $\mathbf{T}$ is a step toward it). In the experimental section (Section 6), we show an example in finite dimension where f denotes the Poisson likelihood.

5.4. 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

$$f(x) = f(x_1, x_2) = \tfrac{1}{2} \|A_1 x_1 - z_1\|_2^2 - \tfrac{\gamma}{2} \|A_2 x_2 - z_2\|_2^2 \tag{27}$$

Notice that f is a difference of convex functions so it is neither convex nor concave. We have: $\nabla f(x) = \begin{pmatrix} A_1^T A_1 x_1 - A_1^T z_1 \\ -\gamma A_2^T A_2 x_2 + \gamma A_2^T z_2 \end{pmatrix}$. We define $\mathbf{D} = \begin{pmatrix} \mathbf{I}_{N_1} & 0 \\ 0 & -\gamma \mathbf{I}_{N_2} \end{pmatrix}$. We have:

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

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 RDP with $\mathbf{D}$ and so that GSP and GCoSaMP may be used in that case.

6. Experiments

6.1. 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,

$$\forall i \in \{1, \dots, n\}, \quad y_i \sim \mathcal{P}(x_i) , \tag{28}$$

where $\mathcal{P}(\lambda)$ stands for the Poisson distribution of parameter λ.

We also assume that x has a k-sparse representation on the dictionary $\mathbf{\Phi} = (\varphi_1, \dots, \varphi_m) \in \mathbb{R}^{n \times m}$:

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

where the atoms are normalized ($\left\|\varphi_j\right\| = 1$), 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:

$$\hat{x} = \mathbf{\Phi} \hat{c}, \quad \text{where} \quad \hat{c} = \underset{c \in \mathbb{R}^m \text{ s.t. } \|c\|_0 \leqslant k}{\operatorname{argmin}} F_y(\mathbf{\Phi} c) , \tag{P4}$$

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

$$F_y(x) = -\log \mathbb{P}(y|x) = \sum_{i=1}^{n} f(x_i, y_i), \quad \text{with}$$
$$f(\xi, \eta) = \begin{cases} -\eta \log(\xi) + \xi & \text{if } \eta > 0 \text{ and } \xi > 0, \\ \xi & \text{if } \eta = 0 \text{ and } \xi \geqslant 0, \\ +\infty & \text{otherwise.} \end{cases} \tag{29}$$

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 5.3, we seek $\hat{x}$ using our four Generalized Algorithms on $\mathbf{T} = \nabla \mathcal{M}_{\lambda, F_y \circ \mathbf{\Phi}}$.

Proposition 14 (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:*

$$\nabla \mathcal{M}_{\lambda, F_y \circ \mathbf{\Phi}} = \tfrac{1}{\nu\lambda} \mathbf{\Phi}^* \circ (\mathbf{I} - \operatorname{prox}_{\nu\lambda F_y}) \circ \mathbf{\Phi} \qquad \textit{with}$$
$$\operatorname{prox}_{\nu\lambda F_y}(x)_i = \frac{x_i - \nu\lambda + \sqrt{|x_i - \nu\lambda|^2 + 4\nu\lambda y_i}}{2} \,. \tag{30}$$

6.2. 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). We use 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)), i.e.

$$\hat{x} = \mathbf{\Phi}\hat{c}, \quad \text{where} \quad \hat{c} = \underset{c \in \mathbb{R}^m \text{ s.t. } \|c\|_1 \leqslant \rho}{\operatorname{argmin}} F_y(\mathbf{\Phi} c) \,, \tag{P5}$$

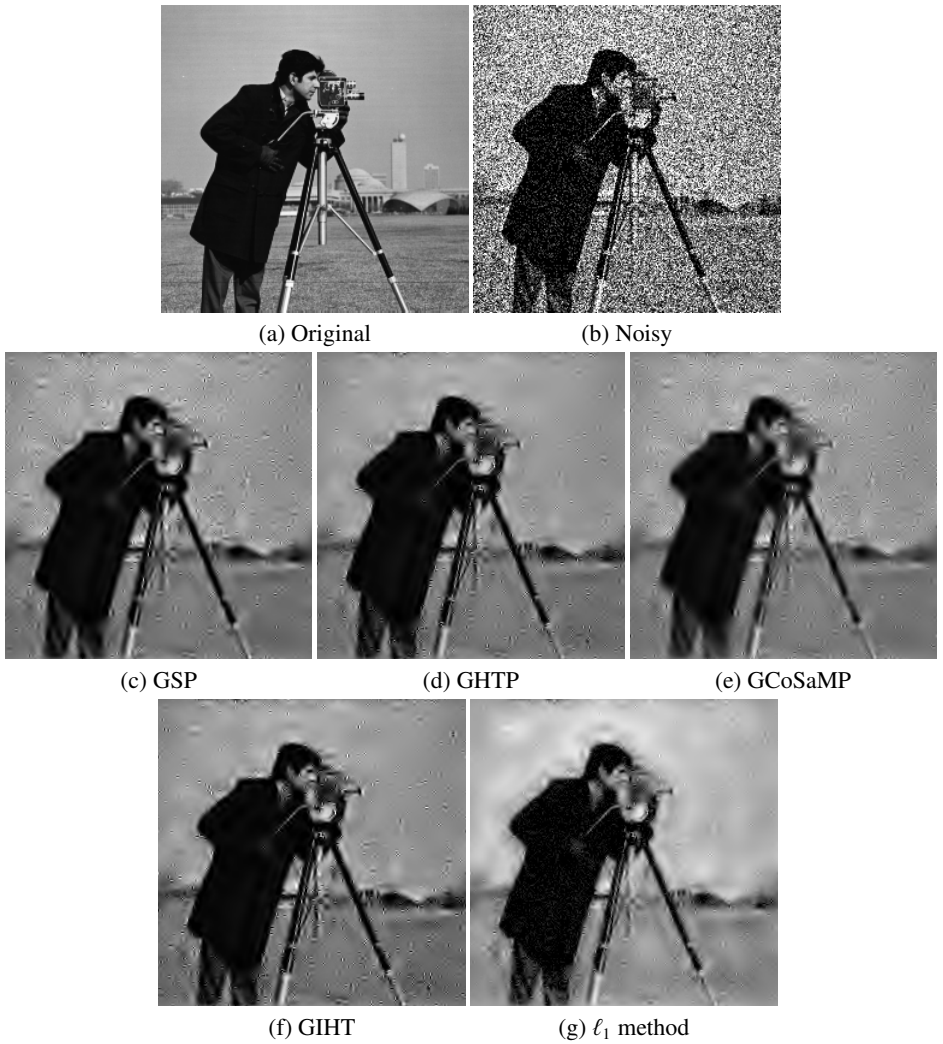
where $\rho > 0$ is the equilibrium parameter.

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 a classical image (*Cameraman*) 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).

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 1: Denoising *Cameraman* with a maximal intensity of 5 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).

(a) Original (b) Noisy

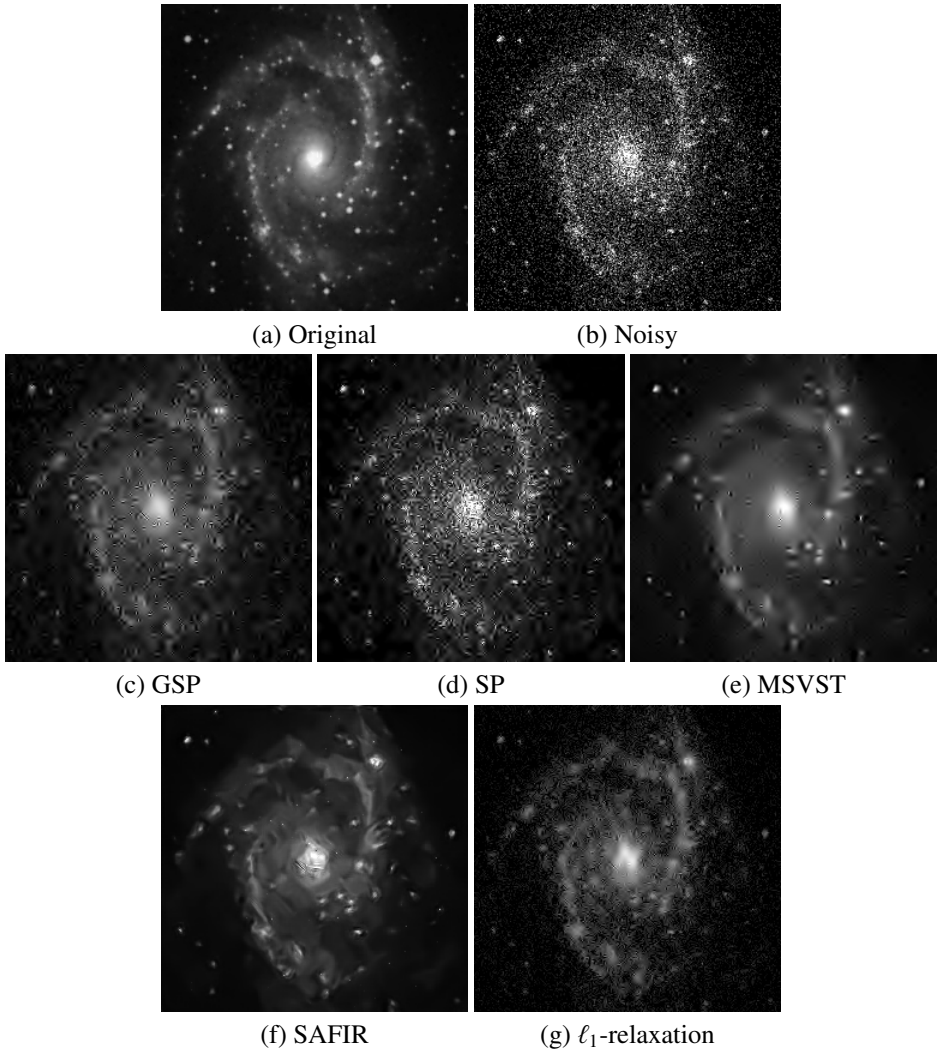
(c) GSP (d) GHTP (e) GCoSaMP

(f) GIHT (g) ℓ_1 method

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

	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.



(a) Original (b) Noisy

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

(f) SAFIR (g) ℓ_1-relaxation

Figure 3: NGC 2997 Galaxy image (a), a noisy version (b) and several denoising results (c-g).

6.3. 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 the convex ℓ_1-relaxation in (P5).

We apply these methods on two images, a sparse version of the Cameraman and the NGC 2997 galaxy (peak intensity at 5, see Fig. 3). 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.

7. 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, restricted strong property, restricted strong smoothness) 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 around these generalizations are of interest. Firstly, one could generalize other algorithms such as OMP and family. Secondly, 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). Thirdly, one could extend the setting from Hilbert to Banach spaces, as has already been proposed for OMP 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.

Appendix A. Restricted Diagonal Property

Proof of Theorem 4. Assume that $\mathbf{D}$ is in $\mathcal{D}_1$ and that $|||\mathbf{D}|||_k$ is finite.

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

$$\begin{aligned} \|\beta\mathbf{T}(x) - \beta\mathbf{T}(y) - \mathbf{D}(x-y)\| \leq \alpha_k \|x-y\| &\Rightarrow \|\beta\mathbf{T}(x) - \beta\mathbf{T}(y)\| - \|\mathbf{D}(x-y)\| \leq \alpha_k \|x-y\| \\ &\Rightarrow \|\beta\mathbf{T}(x) - \beta\mathbf{T}(y)\| \leq \alpha_k \|x-y\| + \|\mathbf{D}(x-y)\| \\ &\Rightarrow \|\mathbf{T}(x) - \mathbf{T}(y)\| \leq \frac{|||\mathbf{D}|||_k + \alpha_k}{\beta} \|x-y\| \end{aligned} \tag{A.1}$$

and

$$\begin{aligned}\|\beta\mathbf{T}(x)-\beta\mathbf{T}(y)-\mathbf{D}(x-y)\|\leq\alpha_k\|x-y\| &\Rightarrow \|\mathbf{D}(x-y)\|-\|\beta\mathbf{T}(x)-\beta\mathbf{T}(y)\|\leq\alpha_k\|x-y\| \\ &\Rightarrow \|\beta\mathbf{T}(x)-\beta\mathbf{T}(y)\|\geq\|\mathbf{D}(x-y)\|-\alpha_k\|x-y\| \\ &\Rightarrow \|\mathbf{T}(x)-\mathbf{T}(y)\|\geq\frac{1-\alpha_k}{\beta}\|x-y\| \text{ (since } \mathbf{D}\in\mathcal{D}_1)\end{aligned} \tag{A.2}$$

so that

$$\begin{aligned}\langle\mathbf{T}(x)-\mathbf{T}(y),\mathbf{D}(x-y)\rangle =& \tfrac{1}{\beta}\langle\beta\mathbf{T}(x)-\beta\mathbf{T}(y),\mathbf{D}(x-y)\rangle \\ =&\frac{1}{2\beta}\left(\|\beta\mathbf{T}(x)-\beta\mathbf{T}(y)\|^2+\|\mathbf{D}(x-y)\|^2-\|\beta\mathbf{T}(x)-\beta\mathbf{T}(y)-\mathbf{D}(x-y)\|^2\right) \\ \geq&\frac{1}{2\beta}\left((1-\alpha_k)^2\|x-y\|^2+\|x-y\|^2-\alpha_k^2\|x-y\|^2\right) \text{ (using RDP, } \mathbf{D}\in\mathcal{D}_1 \text{ and Eq. (A.2))} \\ \geq&\frac{(1-\alpha_k)^2+1-\alpha_k^2}{2\beta}\|x-y\|^2 \\ \geq&\frac{1-\alpha_k}{\beta}\|(x-y)\|^2\end{aligned} \tag{A.3}$$

Proof of 2: Assume that f verifies Eq. (11). Pick $(x,y)\in\mathcal{H}^2$, such that $\text{card}(\text{supp}(x,y))\leq k$; for any $\beta>0$ we have:

$$\begin{aligned}\|\beta\mathbf{T}(x)-\beta\mathbf{T}(y)-\mathbf{D}(x-y)\|^2 &= \|\beta\mathbf{T}(x)-\beta\mathbf{T}(y)\|^2+\|\mathbf{D}(x-y)\|^2-2\beta\langle\mathbf{T}(x)-\mathbf{T}(y),\mathbf{D}(x-y)\rangle \\ &\leq \beta^2L^2\|x-y\|^2+|||\mathbf{D}|||_k^2\|x-y\|^2-2\beta m\|x-y\|^2 \quad \text{(using Eq. (11))} \\ &\leq (\beta^2L^2+|||\mathbf{D}|||_k^2-2\beta m)\|x-y\|^2\end{aligned}$$

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

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

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

Appendix B. Proof of GSP's error bound (Theorem 9)

In this Appendix, we show how to derive the error bound for GSP (Theorem 9). 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 Appendix B.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 Appendix B.2. We then proceed to the complete proof of Theorem 9 by successively bounding the error made at each step of the algorithm in Appendix B.3.

The proofs of the bounds for GCoSaMP, GHTP and GIHT (Theorems 10 to 12) are similar and postponed to Appendix C to Appendix E.

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

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

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

Appendix B.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:*

$$\forall \mathcal{R}' \subseteq \mathbb{N}, \quad \left\|(\mathbf{T}(x) - \mathbf{T}(y))_{|\mathcal{R}'}\right\| \geqslant (1 - \alpha_k) \left\|(x - y)_{|\mathcal{R}'}\right\| - \alpha_k \left\|(x - y)_{|\operatorname{supp}(x,y)\setminus\mathcal{R}'}\right\| \tag{B.2}$$

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

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

□

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:*

$$\forall \mathcal{F} \subseteq \mathbb{N} \; s.t \; \mathcal{F} \cap \operatorname{supp}(x, y) = \emptyset, \quad \left\|(\mathbf{T}(x) - \mathbf{T}(y))_{|\mathcal{F}}\right\| \leqslant \alpha_k \|x - y\| . \tag{B.3}$$

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

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

□

Appendix B.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 $\mathbf{T}(x)$.

For $\mathcal{R}$ a subset of $\mathbb{N}$, we denote by $\mathcal{R}^c$ its complement ($\mathcal{R}^c = \{i \in \mathbb{N} / i \notin \mathcal{R}\}$).

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*

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

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

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

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,

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

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

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

Combining these two inequalities, we obtain

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

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

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

□

The second lemma controls the distance from a k-sparse vector to a solution of the subproblem (P3).

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* (P3)*, 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 $\left\|x - y_{|k}\right\| \leqslant 2\left\|x_{|\mathcal{S}} - y\right\| + \left\|x_{|\mathcal{S}^c}\right\| = 2\left\|(x-y)_{|\mathcal{S}}\right\| + \left\|x_{|\mathcal{S}^c}\right\|$. □

Appendix B.3. Proof of Theorem 9 (error bound on GSP)

With these five lemmas at hand, one can prove Theorems 9 to 12. We detail here the proof for GSP and postpone the proofs for the other algorithms in Appendix 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}.

Appendix B.3.1. Analysis of the main steps

First, let us fix the iteration t of Algorithm 3, 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 $\mathcal{R} = \text{supp}(x^t, x^\star)$, we have $\text{card}(\mathcal{R}) \leqslant 2k$. Remember that $\mathcal{S} = \text{supp}(x^t) \cup \text{supp}(T(x^t)_{|k})$. Note that $\text{supp}(x^t) \subseteq \mathcal{S}$ and $\left\|T(x^t)_{|\mathcal{R}}\right\| \leqslant \left\|T(x^t)_{|\mathcal{S}}\right\|$ since $\text{supp}(x^t) \cap \text{supp}(T(x^t)) = \emptyset$. 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 $\text{card}(\mathcal{R} \setminus \mathcal{S}) \leqslant k$ and $\text{card}(\mathcal{S} \setminus \mathcal{R}) \leqslant k$, we conclude

$$\left\|x^\star_{|\mathcal{S}^c}\right\| \leqslant \tfrac{2\alpha_{2k}}{1-\alpha_{2k}} \left\|x^t - x^\star\right\| + \tfrac{2}{1-\alpha_{2k}} \left\|\mathbf{T}(x^\star)_{|k}\right\| . \tag{B.4}$$

Optimization over the extended support (Step 3)

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

$$\left\|(b - x^\star)_{|\mathcal{S}}\right\| \leqslant \tfrac{1}{1-\alpha_{3k}} \left\|\mathbf{T}(x^\star)_{|\mathcal{S}}\right\| + \tfrac{\alpha_{3k}}{1-\alpha_{3k}} \left\|x^\star_{|\mathcal{S}^c}\right\| . \tag{B.5}$$

Updating the support set (Step 4)

Lemma 19 proves that

$$\left\|b_{|k} - x^\star\right\| \leqslant 2 \left\|(b - x^\star)_{|\mathcal{S}}\right\| + \left\|x^\star_{|\mathcal{S}^c}\right\| . \tag{B.6}$$

Combining this with Eq. (B.5) yields:

$$\left\|b_{|k} - x^\star\right\| \leqslant \tfrac{2}{1-\alpha_{3k}} \left\|\mathbf{T}(x^\star)_{|\mathcal{S}}\right\| + \tfrac{1+\alpha_{3k}}{1-\alpha_{3k}} \left\|x^\star_{|\mathcal{S}^c}\right\| . \tag{B.7}$$

Since $\mathcal{T} = \text{supp}(b_{|k})$, 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

$$\left\|x^\star_{|\mathcal{T}^c}\right\| \leqslant \tfrac{2}{1-\alpha_{3k}} \left\|\mathbf{T}(x^\star)_{|\mathcal{S}}\right\| + \tfrac{1+\alpha_{3k}}{1-\alpha_{3k}} \left\|x^\star_{|\mathcal{S}^c}\right\| . \tag{B.8}$$

Optimization over the updated support (Step 5)

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

$$\left\|(x^{t+1} - x^\star)_{|\mathcal{T}}\right\| \leqslant \tfrac{1}{1-\alpha_{2k}} \left\|\mathbf{T}(x^\star)_{|\mathcal{T}}\right\| + \tfrac{\alpha_{2k}}{1-\alpha_{2k}} \left\|x^\star_{|\mathcal{T}^c}\right\| . \tag{B.9}$$

So

$$\begin{aligned} \left\|x^{t+1} - x^\star\right\| &\leqslant \left\|(x^{t+1} - x^\star)_{|\mathcal{T}}\right\| + \left\|(x^{t+1} - x^\star)_{|\mathcal{T}^c}\right\| \\ &\leqslant \left\|(x^{t+1} - x^\star)_{|\mathcal{T}}\right\| + \left\|x^\star_{|\mathcal{T}^c}\right\| \\ &\leqslant \tfrac{1}{1-\alpha_{2k}} \left\|\mathbf{T}(x^\star)_{|\mathcal{T}}\right\| + \tfrac{1}{1-\alpha_{2k}} \left\|x^\star_{|\mathcal{T}^c}\right\| . \end{aligned} \tag{B.10}$$

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

Appendix B.3.2. Recursion

We start from Eq. (B.10), insert Eq (B.8) 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. (B.4) 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

$$\left\|x^{t+1}-x^\star\right\| \leqslant \tfrac{2\alpha_{3k}(1+\alpha_{3k})}{(1-\alpha_{3k})^3}\left\|x^t-x^\star\right\| + \tfrac{\alpha_{3k}^2-2\alpha_{3k}+5}{(1-\alpha_{3k})^3}\left\|\mathbf{T}(x^\star)_{|2k}\right\| . \tag{B.11}$$

Appendix B.3.3. Error bound

From Eq.(B.11), we deduce

$$\left\|x^t-x^\star\right\| \leqslant \left(\tfrac{2\alpha_{3k}(1+\alpha_{3k})}{(1-\alpha_{3k})^3}\right)^t\left\|x^0-x^\star\right\| + \tfrac{\alpha_{3k}^2-2\alpha_{3k}+5}{(1-\alpha_{3k})^3}\sum_{i=0}^{t-1}\left(\tfrac{2\alpha_{3k}(1+\alpha_{3k})}{(1-\alpha_{3k})^3}\right)^i\left\|\mathbf{T}(x^\star)_{|2k}\right\| . \tag{B.12}$$

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 9 the sufficient condition

$$\tfrac{2\alpha_{3k}(1+\alpha_{3k})}{(1-\alpha_{3k})^3} \leqslant \frac{1}{2} \Leftrightarrow \alpha_{3k} \leqslant \alpha^S , \tag{B.13}$$

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 9.

Appendix C. Proof of GCoSaMP's error bound (Theorem 10)

Proof of Theorem 10. $\mathbf{T}$ 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. 4. Let $\mathcal{G} = \operatorname{supp}(\mathbf{T}(x^t)_{|2k})$ and $\mathcal{S} = \mathcal{G} \cup \operatorname{supp}(x^t)$, b such that $\operatorname{supp}(b) \subseteq \mathcal{S}$ and $\mathbf{T}(b)_{|\mathcal{S}} = 0$ and $\mathcal{T} = \operatorname{supp}(b_{|k})$.

Influence of the guessed support (Step 1 and 2)

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

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

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

$$\left\|(b - x^\star)_{|\mathcal{S}}\right\| \leqslant \tfrac{1}{1-\alpha_{4k}} \left\|\mathbf{T}(x^\star)_{|\mathcal{S}}\right\| + \tfrac{\alpha_{4k}}{1-\alpha_{4k}} \left\|x^\star_{|\mathcal{S}^c}\right\| . \tag{C.2}$$

Updating the support set (Step 4 and 5)

Lemma 19 proves that

$$\left\|b_{|k} - x^\star\right\| \leqslant 2\left\|(b - x^\star)_{|\mathcal{S}}\right\| + \left\|x^\star_{|\mathcal{S}^c}\right\| . \tag{C.3}$$

Since $x^{t+1} = b_{|k}$, we combine Eq. (C.1), (C.2) and (C.3)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

$$\tfrac{2\alpha(1+\alpha)}{(1-\alpha)^2} \leqslant \tfrac{1}{2} \Leftrightarrow 3\alpha^2 - 6\alpha - 1 \leqslant 0 \Leftrightarrow -1 - \tfrac{2}{\sqrt{3}} \leqslant \alpha \leqslant -1 + \tfrac{2}{\sqrt{3}} = \alpha^C \tag{C.4}$$

and

$$\tfrac{4}{(1-\alpha_{4k})^2} \leqslant \tfrac{4}{(1-\alpha^C)^2} = \tfrac{3}{(\sqrt{3}-1)^2} \leqslant 6 , \tag{C.5}$$

which completes the proof. □

Appendix D. Proof of GHTP's error bound (Theorem 11)

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

Proof of Theorem 11. Assume that $\mathbf{T}$ has the *Restricted Diagonal Property* of order $2k$ with $\mathbf{D}_{\mathcal{S}} = \mathbf{I}$ 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. 5. Let $\mathcal{G} = \mathrm{supp}(\mathbf{T}(x^t)_{|k})$ and $\mathcal{S} = \mathcal{G} \cup \mathrm{supp}(x^t)$, $b = (x^t - \eta\mathbf{T}(x^t))_{|\mathcal{S}}$ and $\mathcal{T} = \mathrm{supp}(b_{|k})$.

Influence of the guessed support (Step 1 and 2)

Notice that x^t solves Problem (P3) and $\mathcal{G}$ and $\mathcal{S}$ are the same as in GSP so the same arguments as in the proof in Appendix B.3.1 apply and Eq. (B.4) holds:

$$\left\|x^\star_{|\mathcal{S}^c}\right\| \leqslant \tfrac{2\alpha_{2k}}{1-\alpha_{2k}} \left\|x^t - x^\star\right\| + \tfrac{2}{1-\alpha_{2k}} \left\|\mathbf{T}(x^\star)_{|k}\right\| . \tag{B.4}$$

Solution on the extended support (Step 3)

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

where Eq. (D.1) holds because $\mathbf{T}$ is RDP of order $2k$ with $\mathbf{D}_{\mathcal{S}} = \mathbf{I}$, and $\mathrm{card}(\mathcal{S}) \leqslant 2k$.

Updating the support set (Step 4)

Notice that

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

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

$$\left\|x^\star_{|\mathcal{T}^c}\right\| \leqslant 2\left\|b - x^\star\right\| . \tag{D.3}$$

Optimization over the updated support (Step 5)

We have

$$\begin{aligned}
\left\|x^{t+1} - x^\star\right\| &\leqslant \left\|(x^{t+1} - x^\star)_{|\mathcal{T}}\right\| + \left\|x^\star_{|\mathcal{T}^c}\right\| \\
&\leqslant \tfrac{1}{1-\alpha_{2k}} \left\|\mathbf{T}(x^\star)_{|\mathcal{T}}\right\| + \tfrac{\alpha_{2k}}{1-\alpha_{2k}} \left\|x^\star_{|\mathcal{T}^c}\right\| + \left\|x^\star_{|\mathcal{T}^c}\right\| \qquad \text{(D.4)} \\
&\leqslant \tfrac{1}{1-\alpha_{2k}} \left\|\mathbf{T}(x^\star)_{|\mathcal{T}}\right\| + \tfrac{1}{1-\alpha_{2k}} \left\|x^\star_{|\mathcal{T}^c}\right\| \qquad \text{(D.5)}
\end{aligned}$$

where Eq. (D.4) holds by applying Lemma 18 using that x^{t+1} solves Problem (P3) on $\mathcal{T}$, that $x^\star$ is k-sparse and $\mathbf{T}$ is RDP of order $2k$.

Let us finally combine Eq. (B.4), (D.2), (D.3) and (D.5) to obtain

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

We have

$$2\tfrac{|1-\eta|\left(1-\alpha^2\right)+(3-\alpha)\alpha}{(1-\alpha)^2} \leqslant \tfrac{1}{2} \Leftrightarrow (5+4\,|1-\eta|)\alpha^2 - 14\alpha + 1 - 4\,|1-\eta| \geqslant 0 \tag{D.6}$$

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:

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

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. □

Appendix E. Proof of GIHT's error bound (Theorem 12)

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

Proof of Theorem 12. Let $x^\star \in \mathcal{H}$ be a k-sparse vector, $x^t \in \mathcal{H}$ the t-th iterate of Algo. 6 and $\mathcal{S} = \operatorname{supp}(x^t) \cup \operatorname{supp} x^\star$. Assume $\mathbf{T}$ has the *Restricted Diagonal Property* of order $2k$ with $\mathbf{D}_\mathcal{S} = \mathbf{I}$ 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

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

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

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. □

Adcock, B., Hansen, A. C., 2016. Generalized sampling and infinite-dimensional compressed sensing. Foundations of Computational Mathematics 16 (5), 1263–1323.

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

Bauschke, H., Combettes, P., Reich, S., 2005. The asymptotic behavior of the composition of two resolvents. Nonlinear Analysis: Theory, Methods, and Applications 5 (56), 283–301.

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

Bauschke, H. H., Combettes, P. L., et al., 2017. Convex analysis and monotone operator theory in Hilbert spaces. Vol. 2011. Springer.

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.

Blumensath, T., Davies, M. E., 2008. Iterative hard thresholding for compressed sensing. CoRR abs/0805.0510.

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., 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.

Jones, A., Tamtögl, A., Calvo-Almazán, I., Hansen, A., 2016. Continuous compressed sensing for surface dynamical processes with helium atom scattering. Scientific reports 6, 27776.

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.

Peyré, G., Fadili, J., 2011. Group sparsity with overlapping partition functions. In: Signal Processing Conference, 2011 19th European. IEEE, pp. 303–307.

Shalev-Shwartz, S., Srebro, N., Zhang, T., 2010. Trading accuracy for sparsity in optimization problems with sparsity constraints. SIAM Journal on Optimization 20 (6), 2807–2832.

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

Tropp, J. A., Sep. 2006. Greed is good: Algorithmic results for sparse approximation. IEEE Trans. Inf. Theor. 50 (10), 2231–2242.

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.