A continuous relaxation of the constrained $\ell_2 - \ell_0$ problem

Arne Bechensteen * Laure Blanc-Féraud * Gilles Aubert †

January 4, 2021

Abstract

We focus on the minimization of the least square loss function under a k-sparse constraint encoded by a ℓ_0 pseudo-norm. This is a non-convex, non-continuous and NP-hard problem. Recently, for the penalized form (sum of the least square loss function and a ℓ_0 penalty term) a relaxation has been introduced which has strong results in terms of minimizers. This relaxation is continuous and does not change the global minimizers, among other favorable properties. The question that has driven this paper is the following: can a continuous relaxation of the k-sparse *constraint* problem be developed following the same idea and same steps as for the *penalized* $\ell_2 - \ell_0$ problem? We calculate the convex envelope of the constrained problem when the observation matrix is orthogonal and propose a continuous non-smooth, non-convex relaxation of the k-sparse constraint functional. We give some equivalence of minimizers between the original and the relaxed problems. The subgradient is calculated as well as the proximal operator of the new regularization term, and we propose an algorithm that ensures convergence to a critical point of the k-sparse constraint problem. We apply the algorithm to the problem of single-molecule localization microscopy and compare the results with well-known sparse minimization schemes. The results of the proposed algorithm are as good as the state-of-the-art results for the penalized form, while fixing the constraint constant is usually more intuitive than fixing the penalty parameter.

*Université Côte d'Azur, CNRS, Inria, Laboratoire I3S UMR 7271, 06903 Sophia Antipolis, France. (arne-henrik.bechensteen@inria.fr, blancf@i3s.unice.fr)

†Université Côte d'Azur, UNS, Laboratoire J. A. Dieudonné UMR 7351, 06100 Nice, France. gaubert@unice.fr

keywords inverse problems, ℓ_0 problem, sparse modelling, non-convex, non-smooth, relaxation

1 Introduction

In this paper, we consider the constrained $\ell_2 - \ell_0$ problem:

$$\min_{x \in \mathbb{R}^N} \frac{1}{2}\|Ax - d\|^2 \text{ such that } \|x\|_0 \leq k \tag{1}$$

where $A \in \mathbb{R}^{M \times N}$ is an observation matrix, $d \in \mathbb{R}^M$ is the data, and $\|\cdot\|_0$ is, by abuse of terminology, referred to as the ℓ_0-norm:

$$\|x\|_0 = \#\{x_i, i = 1, \cdots, N : x_i \neq 0\}$$

with $\#S$ defined as the number of elements in S. This formulation ensures that the solution $\hat{x}$ has at maximum k non-zero entries. This type of problem appears in many applications, such as source separation, machine learning, and single-molecule localization microscopy. These problems are often underdetermined, i.e., problems where $M \ll N$. A more studied sparse problem is the penalized $\ell_2 - \ell_0$ problem:

$$\min_{x \in \mathbb{R}^N} \frac{1}{2}\|Ax - d\|^2 + \lambda\|x\|_0 \tag{2}$$

where $\lambda \in \mathbb{R}_{\geq 0}$ is a trade-off parameter. Even though the formulations (1) and (2) are similar, they are not equivalent (see for example [25] for a theoretical comparison). These problems also differ in their sparsity parameter. With the λ parameter, it is not possible to know the sparsity of the solution without testing it. The constrained problem does not have this problem as k fixes the number of non-zero components. However, the problems are both non-convex, non-smooth and NP-hard. In the following paragraph we will outline the different methods to solve (1) and (2).

Greedy algorithms Greedy algorithms are designed to solve problems of the form (1). These algorithms start with a zero initialization and add one component to the signal x at each iteration until the wished-for sparsity is obtained. Among them, we find the Matching Pursuit (MP) algorithm [23], and the Orthogonal Matching Pursuit (OMP) [26]. Newer algorithms add and *subtract* components at each iteration, among them are the algorithm Greedy sparse simplex [3] or Single Best Replacement (SBR) [37].

Mathematical program with equilibrium constraint Another method to solve a sparse optimization problem is to introduce auxiliary variables to simulate the nature of the ℓ_0-norm and add a constraint between primaries and auxiliaries, and thus called a mathematical program with equilibrium constraint. Mixed Integer reformulations [8] and Boolean relaxation [28] are two among the many algorithms based on this method. Algorithms of this method have been proposed to solve sparse problems (see [6, 22], for example). A recent paper [2] proved the exactness of a reformulation of the constrained $\ell_2 - \ell_0$ problem and showed its abilities on single-molecule localization microscopy.

Relaxations An alternative to working with the non-convex ℓ_0-norm is to replace it with the convex ℓ_1-norm. This is called convex relaxation, but only under strict assumptions such as the RIP conditions, the original and the convex relaxed problems are equivalent in terms of minimizers [11]. Furthermore, $\|x\|_1$ penalizes not only the number of components in x but also their magnitude. Thus, the ℓ_0-norm and ℓ_1-norm are very different when x contains large values. Non-smooth, non-convex but continuous relaxations were primarily introduced to avoid this difference. These relaxations are still non-convex, and the convergence of the algorithms to a global minimum is not assured. Some of the non-convex continuous relaxations are the NonNegative Garrote [9], the Log-Sum penalty [12] or Capped-ℓ_1 [27] to mention some. The continuous Exact ℓ_0 penalty introduced in [35] proposes an exact relaxation for problem (2) and a unified view of these functions is given in [36]. A recent convex relaxation has been proposed in [33], which replace the ℓ_0-norm with a non-convex term, but where the sum of the data-fitting term and the relaxation is convex. Relaxation of the constrained $\ell_2 - \ell_0$ problem is less studied. However, the fixed rank problem and its convex envelope has been presented in [1], and the problem has certain similarities with the constrained $\ell_2 - \ell_0$ problem.

Contributions and outline The paper presents and studies a non-smooth and non-convex relaxation of the constrained problem (1). Following the procedure used to design the $CEL0$-relaxation of problem (2) [35], we want to explore if an equivalent continuous relaxation can be found for (1). The next section shows the computation of the convex hull of the constrained $\ell_2 - \ell_0$ formulation in the case of orthogonal matrices. The convex hull yields the square norm plus a penalty term that we name $Q(x)$. Note that the expression of $Q(x)$ could be obtained by applying the quadratic envelope presented in [14], choosing the right parameters. However, in [14] the author uses an additional parameter that depends on the observation matrix A. This will be discussed later in Section 3. In the same section, the relaxed formulation is investigated as a continuous relaxation of the initial problem for any matrix A. We prove some basic properties of $Q(x)$ to show that the relaxation

favors k-sparse vectors. The relaxation does not always ensure a k-sparse solution, but it promotes sparsity. We show that if a minimizer of the relaxed expression is k-sparse, then the minimizer of the relaxed problem is a minimizer of the initial one. We propose an algorithm to minimize the relaxed formulation using an accelerated FBS method, and we add a "fail-safe" strategy which ensures convergence to a critical point of the initial problem. The relaxation and its associated algorithm is applied to the problem of single-molecule localization microscopy and compared to other state-of-the-art algorithms in $\ell_2 - \ell_0$ minimization.

Notations and Assumption

- $A \in \mathbb{R}^{M\times N}$ is an $M \times N$ matrix.
- The vector $x^{\downarrow} \in \mathbb{R}^N$ is the vector x where its components are sorted by their magnitude i.e. $|x_1^{\downarrow}| \geq |x_2^{\downarrow}| \geq \cdots \geq |x_N^{\downarrow}|$.
- Let $P^{(y)} \in \mathbb{R}^{N\times N}$ a permutation matrix such that $P^{(y)}y = y^{\downarrow}$, we denote the vector $x^{\downarrow y} = P^{(y)}x$.
- a_i is the ith column of A. We suppose $\|a_i\| \neq 0\, \forall i$.
- The indicator function χ_X is defined for $X \subset \mathbb{R}^N$ as

$$\chi_X(x) = \begin{cases} +\infty \text{ if } x \notin X \\ 0 \text{ if } x \in X. \end{cases}$$

- $\text{sign}^*(x)$ is the function sign for $x \neq 0$ and $\text{sign}^*(0) = \{-1, 1\}$.
- $\mathbb{R}^N_{\geq 0}$ denotes the space $\{x \in \mathbb{R}^N | x_i \geq 0, \forall i\}$.

Proposition 1. *We can suppose that $\|a_i\|_2 = 1,\ \forall\ i$, without loss of generality.*

Proof. The proof is based on the fact that ℓ_0-norm is invariant to a multiplication factor. Let $\Lambda_{\|a_i\|}$ and $\Lambda_{\frac{1}{\|a_i\|}}$ be diagonal matrices with the norm of a_i (respectively $1/||a_i||$) on its diagonal, and let $z = \Lambda_{\|a_i\|}x$, then $\|\Lambda_{\frac{1}{\|a_i\|}}z\|_0 = \|z\|_0 = \|x\|_0$, and thus

$$\arg\min_x \frac{1}{2}\|Ax - d\|_2^2 + \chi_{\|\cdot\|_0\leq k}(x) = \Lambda_{\frac{1}{\|a_i\|}} \arg\min_z \frac{1}{2}\|A_n z - d\|_2^2 + \chi_{\|\cdot\|_0\leq k}(z)$$

where A_n is a matrix deduced from A where the norm of each column is 1. □

We assume therefore that A has normalized columns throughout this paper.

2 The convex envelope of the constrained $\ell_2 - \ell_0$ problem when A is orthogonal

In this section, we are interested in the case where A is an orthogonal matrix, i.e. $< a_j, a_i >= 0, \forall\, i \neq j$. In contrast to the penalized form (2), the functional with A orthogonal is not separable so the computation of the convex envelope in the N dimensional case cannot be reduced to the sum of N *one* dimensional cases (as in [35]). The problem (1) can be written as the minimization of

$$G_k(x) = \frac{1}{2}\|Ax - d\|^2 + \chi_{\|\cdot\|_0 \leq k}(x) \tag{3}$$

where χ is the indicator function defined in Notations. Before calculating the convex envelope, we need some preliminary results.

Proposition 2. *Let $x \in \mathbb{R}^N$. There exists $T_k(x) \in \mathbb{N}$ such that $0 < T_k(x) \leq k$ and*

$$|x^{\downarrow}_{k-T_k(x)+1}| \leq \frac{1}{T_k(x)} \sum_{i=k-T_k(x)+1}^{N} |x^{\downarrow}_i| \leq |x^{\downarrow}_{k-T_k(x)}| \tag{4}$$

where the left inequality is strict if $T_k(x) \neq 1$, and where $x_0 = +\infty$. Furthermore, $T_k(x)$ is defined as the smallest integer that verifies the double inequality.

The proof of existence is given in the Appendix A.1. We will also use the Legendre-Fenchel transformation which is essential in the calculation of the convex envelope.

Definition 1. *The Legendre-Fenchel transformation of a function $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is defined as*

$$f^*(u^*) = \sup_{u \in \mathbb{R}^N} < u, u^* > - f(u).$$

The biconjugate of a function, that is applying the Legendre-Fenchel transformation twice, is the convex envelope of the function.

Following [35], we present the convex envelope of G_k (3) when A is orthogonal.

Theorem 1. *Let $A \in \mathbb{R}^{M \times N}$ be such that $A^T A = I$. The convex envelope of $G_k(x)$ is*

$$G_k^{**}(x) = \frac{1}{2}\|Ax - d\|_2^2 + Q(x) \tag{5}$$

where

$$Q(x) = -\frac{1}{2} \sum_{i=k-T_k(x)+1}^{N} x_i^{\downarrow 2} + \frac{1}{2T_k(x)} \left(\sum_{i=k-T_k(x)+1}^{N} |x_i^{\downarrow}| \right)^2 \tag{6}$$

and where $T_k(x)$ is defined as in Proposition 2.

Proof. Since $A^T A = I$, the function G_k (3) can be rewritten as

$$G_k(x) = \chi_{\|\cdot\|_0 \leq k}(x) + \frac{1}{2} \|d - b\|_2^2 + \frac{1}{2} \|x - z\|_2^2 \tag{7}$$

where $b = AA^T d$ and $z = A^T d$. This reformulation allows us to decompose the data-fitting term into a sum of 1-dimensional functions. We apply the Legendre transformation on the functional (7):

$$G_k^*(y) = \sup_{x \in \mathbb{R}^N} < x, y > -\chi_{\|\cdot\|_0 \leq k}(x) - \frac{1}{2} \|d - b\|_2^2 - \frac{1}{2} \|x - z\|_2^2 .$$

We leave out the terms that are not depending on x.

$$G_k^*(y) = -\frac{1}{2} \|d - b\|_2^2 + \left(\sup_{x \in \mathbb{R}^N} < x, y > -\chi_{\|\cdot\|_0 \leq k}(x) - \frac{1}{2} \|x - z\|_2^2 \right).$$

Writing differently the expression inside the supremum we get

$$G_k^*(y) = -\frac{1}{2} \|d - b\|_2^2 + \left(\sup_{x \in \mathbb{R}^N} -\chi_{\|\cdot\|_0 \leq k}(x) - \frac{1}{2} \|x - (z + y)\|_2^2 + \frac{1}{2}\|z + y\|_2^2 - \frac{1}{2}\|z\|_2^2 \right).$$

We develop further

$$G_k^*(y) = -\frac{1}{2} \|d - b\|_2^2 - \frac{1}{2}\|z\|_2^2 + \frac{1}{2}\|z + y\|_2^2 + \left(\sup_{x \in \mathbb{R}^N} -\chi_{\|\cdot\|_0 \leq k}(x) - \frac{1}{2} \|x - (z + y)\|_2^2 \right).$$

The supremum is reached when $x_i = (z + y)_i^{\downarrow}$, $i \leq k$, and $x_i = 0$, $\forall i > k$. The Legendre transformation of G_k is therefore

$$G_k^*(y) = -\frac{1}{2} \|d - b\|_2^2 - \frac{1}{2}\|z\|_2^2 + \frac{1}{2} \sum_{i=1}^{k} (z + y)_i^{\downarrow 2}.$$

To obtain the convex envelope of the function G_k, we compute the Legendre transformation of G_k^*.

$$G_k^{**}(x) = \sup_y < x, y > + \frac{1}{2}\|d-b\|_2^2 + \frac{1}{2}\|z\|_2^2 - \frac{1}{2}\sum_{i=1}^{k}(z+y)_i^{\downarrow 2}.$$

We add and subtract $\frac{1}{2}\|x\|^2$ and $< x, z >$ in order to obtain an expression that is easier to work with.

$$\begin{aligned} G_k^{**}(x) = \sup_y < x, y > + \frac{1}{2}\|d-b\|_2^2 + \frac{1}{2}\|z\|_2^2 + \frac{1}{2}\|x\|^2 - \frac{1}{2}\|x\|^2 \\ + < x, z > - < x, z > - \frac{1}{2}\sum_{i=1}^{k}(z+y)_i^{\downarrow 2} \end{aligned}$$

$$G_k^{**}(x) = \sup_y < x, z+y > + \frac{1}{2}\|d-b\|_2^2 + \frac{1}{2}\|x-z\|_2^2 - \frac{1}{2}\|x\|^2 - \frac{1}{2}\sum_{i=1}^{k}(z+y)_i^{\downarrow 2}.$$

Noticing that $\frac{1}{2}\|d-b\|_2^2 + \frac{1}{2}\|x-z\|_2^2 = \frac{1}{2}\|Ax-d\|_2^2$, using the notation $w = z+y$, and given the definition of $w^{\downarrow}$, this is equivalent to

$$G_k^{**}(x) = \frac{1}{2}\|Ax-d\|_2^2 - \frac{1}{2}\|x\|^2 + \sup_{w\in\mathbb{R}^N} < x, w > - \frac{1}{2}\sum_{i=1}^{k} w_i^{\downarrow 2}. \tag{8}$$

The above supremum problem can be solved by using Lemma 1, which is presented after this proof. This yields

$$G_k^{**}(x) = \frac{1}{2}\|Ax-d\|_2^2 - \frac{1}{2}\sum_{i=k-T_k(x)+1}^{N} x_i^{\downarrow 2} + \frac{1}{2T_k(x)}\left(\sum_{i=k-T_k(x)+1}^{N} |x_i^{\downarrow}|\right)^2 \tag{9}$$

□

The following Lemma is necessary in the proof of the convex envelope.

Lemma 1. *Let $x \in \mathbb{R}^N$. Consider the following supremum problem*

$$\sup_{y\in\mathbb{R}^N} -\frac{1}{2}\sum_{i=1}^{k} y_i^{\downarrow 2} + < y, x > . \tag{10}$$

This problem is concave and the value of the supremum problem (10) is

$$\frac{1}{2}\sum_{i=1}^{k-T_k(x)} x_i^{\downarrow 2} + \frac{1}{2T_k(x)}\left(\sum_{i=k-T_k(x)+1}^{N} |x_i^{\downarrow}|\right)^2.$$

$T_k(x)$ is defined in Proposition 2. The supremum argument is given by

$$y = P^{(x)^{-1}}\hat{y}$$

where $\hat{y}$ is

$$\hat{y}_j(x) = \begin{cases} \operatorname{sign}(x_j^{\downarrow})\frac{1}{T_k(x)}\sum_{i=k-T_k(x)+1}^{N}|x_i^{\downarrow}| & \text{if } k \geq j \geq k - T_k(x) + 1 \\ & \text{or if } j > k \text{ and } x_j^{\downarrow} \neq 0 \\ [-1,1]\frac{1}{T_k(x)}\sum_{i=k-T_k(x)+1}^{N}|x_i^{\downarrow}| & \text{if } j > k \text{ and } x_j^{\downarrow} = 0 \\ x_j^{\downarrow} & \text{if } j < k - T_k(x) + 1. \end{cases} \tag{11}$$

The proof can be found in Appendix A.2.

Remark 1. *$\hat{y}$ is such that $\hat{y} = \hat{y}^{\downarrow}$.*

This expression of the convex envelope may be hard to grasp since the expression is on a non-closed form. To understand better $Q(x)$ we have the following properties.

Property 1. $Q(x) : \mathbb{R}^n \to [0, \infty[$.

Proof. Let us show that $Q(x) \geq 0$, $\forall x$. We use equation (6) as starting point.

$$\begin{aligned} Q(x) &= -\frac{1}{2}\sum_{i=k-T_k(x)+1}^{N} x_i^{\downarrow 2} + \frac{1}{2T_k(x)}\left(\sum_{i=k-T_k(x)+1}^{N}|x_i^{\downarrow}|\right)^2 \\ &\geq -\frac{1}{2}|x_{k-T_k(x)+1}^{\downarrow}|\sum_{i=k-T_k(x)+1}^{N}|x_i^{\downarrow}| + \frac{1}{2T_k(x)}\left(\sum_{i=k-T_k(x)+1}^{N}|x_i^{\downarrow}|\right)^2 \\ &\geq -\frac{1}{2}|x_{k-T_k(x)+1}^{\downarrow}|\sum_{i=k-T_k(x)+1}^{N}|x_i^{\downarrow}| + \frac{1}{2}|x_{k-T_k(x)+1}^{\downarrow}|\sum_{i=k-T_k(x)+1}^{N}|x_i^{\downarrow}| = 0. \end{aligned}$$

We used the fact that $|x^{\downarrow}_{k-T_k(x)+1}| \geq |x^{\downarrow}_i|$, $\forall i \geq k - T_k(x) + 1$ for the first inequality. For the second inequality, we used the inequality in the definition of $T_k(x)$ (see Proposition 2) to go from the second to third line. Note that for $T_k(x) > 1$ the last inequality is strict. □

Property 2. *The function $Q(x)$ is continuous on $\mathbb{R}^N$.*

Proof. By definition we have that $G_k^{**}(x) = \frac{1}{2}\|Ax-d\|^2+Q(x)$ when A is orthogonal, and G_k^{**} is lower semi-continuous, and continuous in the interior of its domain. From [29, Corollary 3.47] for coercive functions, $dom(co(f)) = co(dom(f))$, where co is the convex envelope of a function and dom is the domain of the function. First, G_k is coercive when A is orthogonal since we have $\|Ax\|^2 = (Ax)^T Ax = x^T A^T Ax = \|x\|^2$. G_k^{**} is continuous on $\mathbb{R}^N$. Since $dom(G_k)$ is made up of all different supports where $\|x\|_0 \leq k$, so its convex envelope is $\mathbb{R}^N$. Thus $dom(G_k^{**}) = \mathbb{R}^N$, and G_k^{**} is continuous on $\mathbb{R}^N$. Moreover, $Q(x) = G_k^{**}(x) - \frac{1}{2}\|Ax - d\|^2$, so $Q(x)$ is the difference between a continuous function and a continuous function, and is independent of A, and thus continuous. □

Property 3. *Let $\|x\|_0 \leq k$. Then $T_k(x)$ as defined in Proposition 2 is such that $T_k(x) = 1$. The inverse it not necessarily true.*

Proof. From Proposition 2 we know that $T_k(x)$ satisfies

$$|x^{\downarrow}_{k-T_k(x)+1}| \leq \frac{1}{T_k(x)} \sum_{i=k-T_k(x)+1}^{N} |x^{\downarrow}_i| \leq |x^{\downarrow}_{k-T_k(x)}|.$$

First, note that for all x such that $\|x\|_0 \leq k$, we have $\forall j > k$, $x^{\downarrow}_j = 0$, and in this case the inequalities are clearly satisfied for $T_k(x) = 1$. Furthermore, $T_k(x)$ is defined as the smallest possible integer, and thus $T_k(x) = 1$.

An example to prove the inverse is not true: Let $x = (6, 3, 2, 1)^T$. Let $k = 2$, then

$$\sum_{i=k}^{N} |x^{\downarrow}_i| = 6 \leq |x^{\downarrow}_{k-1}| = 6.$$

$T_k(x) = 1$ but the constraint $\|x\|_0 \leq 2$ is clearly not satisfied. □

Property 4. $Q(x) = 0$ ***if and only if*** $\|x\|_0 \leq k$.

Proof. From Property 1, $Q(x) \geq 0$ and the inequality is strict if $T_k(x) > 1$. Thus, it suffices to investigate $T_k(x) = 1$. The expression is thus reduced to:

$$Q(x) = \sum_{j=k+1}^{N} \sum_{i=k}^{j-1} |x_i^{\downarrow}||x_j^{\downarrow}|$$

which is equal to 0 only if at least $\forall j,\ j > k,\ x_j^{\downarrow} = 0$.

□

In the next section, we will investigate the use of $Q(x)$ when A is not orthogonal.

3 A new relaxation

From now on, we suppose $A \in \mathbb{R}^{M \times N}$ with A not necessarily orthogonal.

We are interested in a continuous relaxation of G_k defined as

$$G_k = \frac{1}{2}\|Ax - d\|^2 + \chi_{\|\cdot\|_0 \leq k}(x).$$

Following the $CEL0$ approach, we propose the following relaxation of G_k:

$$G_Q(x) = \frac{1}{2}\|Ax - d\|^2 + Q(x) \tag{12}$$

with

$$Q(x) = -\frac{1}{2} \sum_{i=k-T_k(x)+1}^{N} x_i^{\downarrow 2} + \frac{1}{2T_k(x)} \left(\sum_{i=k-T_k(x)+1}^{N} |x_i^{\downarrow}| \right)^2 \tag{13}$$

where $T_k(x)$ is the function defined in Proposition 2:

$$|x_{k-T_k(x)+1}^{\downarrow}| \leq \frac{1}{T_k(x)} \sum_{i=k-T_k(x)+1}^{N} |x_i^{\downarrow}| \leq |x_{k-T_k(x)}^{\downarrow}| \tag{14}$$

where, by definition, the inequality is strict if $T_k(x) > 1$.

Remark that, from its definition (see Eq. (8)), $Q(x)$ can be written as

$$Q(x) = -\frac{1}{2} \sum_{i=1}^{N} x_i^2 + \sup_{w \in \mathbb{R}^N} -\frac{1}{2} \sum_{i=1}^{k} w_i^{\downarrow 2} + < w, x > . \tag{15}$$

Note that the properties of $Q(x)$ proved in Section 2 are valid for any A.

The exactness of a relaxation means that the relaxation has the same global minimizers as the initial function. Furthermore, it does not add any minimizers that are not minimizers of the initial function. The $CEL0$ relaxation [35] is an exact relaxation of the penalized functional (2). The proposed relaxation G_Q of the constraint functional G_k (3) is not exact as a counterexample later in the paper shows. We can prove, however, some partial results.

Remark 2. *From Property 4, we have $Q(x) = 0 \; \forall x$ such that $\|x\|_0 \leq k$. Thus $G_Q(x) = G_k(x) \; \forall x$ such that $\|x\|_0 \leq k$.*

Theorem 2. *Let $\hat{x}$ be a local (respectively global) minimizer of G_Q. If $\|\hat{x}\|_0 \leq k$, then $\hat{x}$ is a local (respectively global) minimizer of G_k.*

Proof. Let $\mathscr{S} := \{x : \|x\|_0 \leq k\}$. Let $\hat{x}$ be a local minimizer of G_Q, such that $\|\hat{x}\|_0 \leq k$ and let $\mathscr{N}(\hat{x}, \gamma)$ denote the γ-neighborhood of $\hat{x}$. By contradiction assume that $\exists \bar{x} \in \mathscr{N}(\hat{x}, \gamma) \bigcup \mathscr{S}$ s.t. $G_k(\bar{x}) < G_k(\hat{x})$. From Remark 2, $G_Q(\bar{x}) = G_k(\bar{x})$ and $G_Q(\hat{x}) = G_k(\hat{x})$, which means $\exists \bar{x} \in \mathscr{N}(\hat{x}, \gamma) \cup \mathscr{S} s.t. G_Q(\bar{x}) < G_Q(\hat{x})$ which is a contradiction since $\hat{x}$ is a minimizer of G_Q. The same reasoning can be applied in the case of global minimizers. □

Thus, if a minimizer of the relaxed functional satisfies the sparsity constraint, then it is a minimizer of the initial problem. Furthermore, the relaxation is a mix of absolute values and squares and promotes therefore sparsity. The subgradient, as can be seen in the next section, promotes a k-sparse solution.

Further note that we could have applied the quadratic envelope [14] to obtain the relaxation Q. The quadratic envelope can be defined as applying twice the S_γ transformation on a function f. The S_γ transformation is defined as:

$$S_\gamma(f)(y) := \sup_x -f(x) - \frac{\gamma}{2}\|x - y\|^2.$$

If we apply the quadratic envelope to the constrained ℓ_0 indicator function, we obtain γQ. Further, the author proposes to either choose $\gamma I \prec A^T A$, where I is the identity matrix, or $\gamma I \succ A^T A$. It is important to note that if we have a γ such that $\gamma I \not\succ A^T A$, does not mean that $\gamma I \prec A^T A$. When γ is such that $\gamma I \succ A^T A$ the relaxation is *exact*. However, numerically, we found this condition far too strong, and it did not perform better than minimizing the initial hard constraint function G_k (3). For a normalized matrix A, Q can be found by taking $\gamma = 1$ in $S_\gamma(S_\gamma(\chi_{\|\cdot\|_0 \leq k}))$. However, we do not have necessarily $I \succ A^T A$. Nevertheless we show in this paper,

some exact relaxation properties for G_Q.

Furthermore, what is hidden in our proposed method is the fact that each column of A is normalized. Without this assumption, *each* element x_i would be weighted by $\|a_i\|^2$, which is finer than multiplying the same constant to the whole the regularization term. Again, we can compare with the $CEL0$ relaxation. When applying the quadratic envelope to the ℓ_0 penalization term, we obtain $CEL0$, but instead of $\|a_i\|^2$ in the expression, there is a γ.

However, we are obliged to normalize A to calculate the proximal operator of the regularization term.

3.1 The subgradient

In this section, we calculate the subgradient of G_Q. Since G_Q is neither smooth nor convex, we cannot calculate the gradient nor the subgradient in the sense of convex analysis. We calculate the generalized subgradient (or Clarke subgradient). The obtained expression shows the difficulties to give optimal necessary conditions for the relaxation.

To calculate the generalized subgradient, we must first prove that $Q(x)$ is locally Lipschitz.

Definition 2. *A function $f : \mathbb{R}^N \to \mathbb{R}$ is locally Lipschitz at point x if*

$$\exists (L, \epsilon), \forall (y, y') \in \mathscr{N}(x, \epsilon)^2, |f(y) - f(y')| \leq L\|y - y'\|$$

where $L \in \mathbb{R}_{\geq 0}$, and $\mathscr{N}(x, \epsilon)$ is a ϵ neighborhood of x.

Lemma 2. *$Q(x)$ is locally Lipschitz, $\forall x \in \mathbb{R}^N$.*

Proof. First, it is well-known that the supremum of locally Lipschitz functions is locally Lipschitz. Let us use the definition of $Q(x)$ from (15). The function defined as $x \to \sup_w -\frac{1}{2}\sum_{i=1}^k {w_i^{\downarrow}}^2 + < w, x >$ is locally Lipschitz since $\forall i$ the functions $x \to -\frac{1}{2}\sum_{i=1}^k {w_i^{\downarrow}}^2 + < w, x >$ are locally Lipschitz. Furthermore, the sum of two locally Lipschitz functions is locally Lipschitz. □

Since $Q(x)$ is locally Lipschitz, we can search for the generalized subgradient, denoted ∂.

Definition 3. *The generalized subgradient [16] of a function $f : \mathbb{R}^N \to \mathbb{R}$ (which is locally Lipschitz) is defined by*

$$\partial f(x) := \{\xi \in \mathbb{R}^N : f^0(x, v) \geq < v, \xi >, \forall v \in \mathbb{R}^N\}$$

where $f^0(x, v)$ is the generalized directional derivative in the direction v,

$$f^0(x, v) = \limsup_{\substack{y \to x \\ \eta \downarrow 0}} \frac{f(y + \eta v) - f(y)}{\eta}.$$

Theorem 3. *Let $x \in \mathbb{R}^N$, and let $T_k(x)$ be as defined in Proposition 2. The subgradient of $G_Q(x)$ is*

$$\partial G_Q(x) = A^*(Ax - d) - x + y(x) \tag{16}$$

where $y(x)$ is the argument where the supremum is reached in Lemma 1.

Proof. G_Q is sum of three functions, $\sup_w -\frac{1}{2}\sum_{i=1}^k w_i^{\downarrow 2} + < w, x >$, $\frac{1}{2}\|Ax - d\|^2$ and $-\frac{1}{2}\|x\|^2$. From [16, Proposition 2.3.3 and Corollary 1] and since the two last functions are differentiable, we can write the generalized subgradient of G_Q as the sum of the gradient of the two last functions and the generalized subgradient of the first, i.e.

$$\partial G_Q = \nabla[\frac{1}{2}\|A \cdot -d\|^2](x) - \nabla[\frac{1}{2}\|\cdot\|^2](x) + \partial[\sup_{w \in \mathbb{R}^N} -\frac{1}{2}\sum_{i=1}^k w_i^{\downarrow 2} + < w, \cdot >](x). \tag{17}$$

Thus, the difficulty is to calculate $\partial[\sup_w -\frac{1}{2}\sum_{i=1}^k w_i^{\downarrow 2} + < w, \cdot >](x)$.

From [24, Theorem 2.93], the subgradient of the supremum is the convex envelop of the subgradients where the supremum is reached. We define $g(w, x) = -\frac{1}{2}\sum_{i=1}^k w_i^{\downarrow 2} + < w, x >$. The subgradient of g with respect to x is $\partial(g(w, \cdot))(x) = w$. Now, we need to find the supremum in $\sup_w -\frac{1}{2}\sum_{i=1}^k w_i^{\downarrow 2} + < w, x >$. From Lemma 1, we know that the supremum is reached at $y(x)$, given in (11). We insert $y(x)$ into (17) and this concludes the proof. □

3.2 A numerical example of the relaxation in two dimensions

In order to obtain a clearer view of what is gained with the proposed relaxation, we study two numerical examples in two dimensions. We set $k = 1$ and the initial problem is

$$G_k(x) = \frac{1}{2}\|Ax - d\|^2 + \chi_{\|\cdot\|_0 \leq 1}(x).$$

In two dimensions the problem $G_{k=1}$ is a simple problem to minimise. The solution is either when the first component, $\hat{x}_1$ is 0, or when the second component $\hat{x}_2 = 0$, or both. For $k = 1$ we have that $T_k(x) = 1$, and the relaxed formulation is then

$$G_Q(x) = \frac{1}{2}\|Ax - d\|^2 + |x_1||x_2|.$$

We consider the case where $A \in \mathbb{R}^{2\times 2}$, and the two following examples:

$$A = \begin{pmatrix} 3 & 2 \\ 1 & 3 \end{pmatrix} \Lambda_{1/\|a_i\|} \quad \text{and } d = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \tag{18}$$

$$A = \begin{pmatrix} -3 & -2 \\ 1 & 3 \end{pmatrix} \Lambda_{1/\|a_i\|} \quad \text{and } d = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \tag{19}$$

where $\Lambda_{1/\|a_i\|}$ is a diagonal matrix with $\frac{1}{\|a_i\|}$ on its diagonal, and $\|a_i\|$ is the norm of the ith column of A. Figure 1 presents the contour lines of G_k and G_Q. The red semi-transparency layer over the contour line of the G_k represents the infinite value, and the blue semi-transparency layer over the relaxation marks the axes. The figures show the advantages of using G_Q as relaxation. The relaxation is continuous, and in Example (18), the relaxation is exact. This can be observed in the upper row in Fig. 1. Example (19) gives an example when the relaxation is not exact. In the lower row of Fig. 1 we observe the effect of the relaxation, as it is a product of the absolute value of x_1 and x_2. The global minima for the relaxation in this case is situated in $(-0.086, 1.0912)$ and the two minima for G_k are $(-0.3162, 0)$ and $(0, 1.094)$.

4 Algorithms to deal with G_Q

The analysis of the relaxation shows that it promotes sparsity. The function G_Q is non-convex and non-smooth, but G_Q is continuous, which is not the case for G_k. One could implement a subgradient method, either by using gradient bundle methods (see [10] for an overview), or classical subgradient methods. However, there are no convergence guarantees for the latter. Both methods are also known to be slow compared to the classical Forward-Backward splitting algorithm (FBS). The FBS algorithm is proven to converge when the objective function has the Kurdyka-Łojasiewicz (K-Ł) property. More recent algorithms propose accelerations of the FBS, such as the Nonmonotone Accelerated Proximal gradient algorithm (nmAPG) [21] which is used in the numerical experiences of this paper. The algorithm is

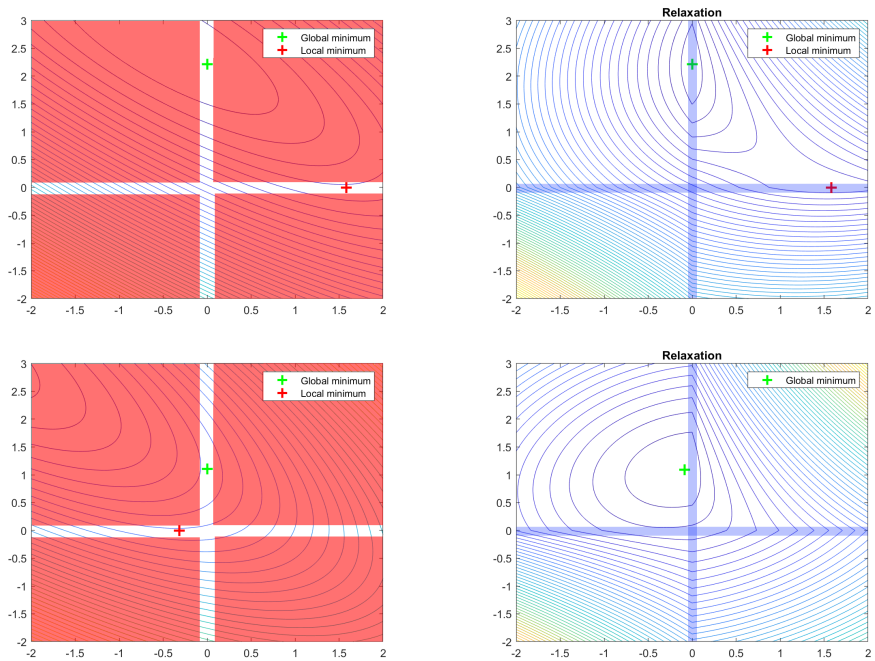


Figure 1: Top: Level lines of the function G_k and G_Q for the example (18). Bottom: Level lines of the function G_k and G_Q for the example (19).

presented in the Appendix A.4. It is designed to work on problems of the form

$$\hat{x} \in \arg\min_x F(x) := f(x) + g(x) \tag{20}$$

where f is a differentiable function, ∇f is L-Lipschitz, and the proximal operator of g can be calculated. It is possible to add a fail-safe to be sure that the algorithm always converges to a solution that satisfies the sparsity constraint. A simple projection to the constraint $\|x\|_0 \leq k$ using the proximal of the constraint and then the calculation of the optimal intensity for the given support would suffice. To use the FBS and its variants, we need to calculate the proximal operator of $Q(x)$. To do so, we present some preliminary results before presenting the proximal operator.

Lemma 3. *G_Q satisfies the K-Ł property.*

Proof. $\frac{1}{2}\|Ax-d\|^2$ is semi-algebraic. Using the definition of $Q(x)$ in (15) we can prove

that $Q(x)$ is semi-algebraic. First, note that $\|x\|_2^2$ is semi-algebraic. Furthermore:

$$\sum_{i=1}^{k} x_i^{\downarrow 2} = \sup_y g(x,y) := -\chi_{\|\cdot\|_0 \leq k}(y) - \frac{1}{2}\|x-y\|^2$$

and $g(x,y)$ is semi-algebraic [7], then $\sum_{i=1}^{k} x_i^{\downarrow 2}$ is semi-algebraic. Thus, $f(x,y) := -\sum_{i=1}^{k} y_i^{\downarrow 2} + < x, y >$ is semi-algebraic, and the supremum as well. We can conclude that $Q(x)$ is semi-algebraic, and thus G_Q satisfies the K-Ł property. □

The expression of $Q(x)$ in (6) is not on a closed-form expression because of the function $T_k(x)$ and calculating the proximal operator directly from this expression is difficult. The following proposition facilitates the calculation of prox_Q. The proposition is inspired by [13, Proposition 3.3], and the proof is omitted in this article as it follows the same steps and arguments as in the referenced article.

Proposition 3. *Let $\rho > 1$ and $z = prox_{-(\frac{\rho-1}{\rho})\sum_{i=k+1}^{N}(\cdot)^{\downarrow 2}}(y)$. We have*

$$prox_{\frac{Q}{\rho}}(y) = \frac{\rho y - z}{\rho - 1}. \tag{21}$$

Thus, it suffices to calculate the proximal operator of $\zeta(x) := -(\frac{\rho-1}{\rho})\sum_{i=k+1}^{N} x_i^{\downarrow 2}$. This is done in Lemma 8 in the Appendix A.3. The following theorem presents the proximal operator of Q

Theorem 4. *The proximal operator of Q for $\rho > 1$ is such that*

$$prox_{\frac{Q}{\rho}}(y)_i^{\downarrow y} = \begin{cases} \frac{\rho y_i^{\downarrow} - \text{sign}(y_i^{\downarrow})\max(|y_i^{\downarrow}|,\tau)}{\rho-1} & \text{if } i \leq k \\ \frac{\rho y_i^{\downarrow} - \text{sign}(y_i^{\downarrow})\min(\tau,\rho|y_i^{\downarrow}|)}{\rho-1} & \text{if } i > k \end{cases}$$

or, equivalently

$$prox_{\frac{Q}{\rho}}(y)_i^{\downarrow y} = \begin{cases} y_i^{\downarrow} & \text{if } i \leq k^* \\ \frac{\rho y_i^{\downarrow} - \text{sign}(y_i^{\downarrow})\tau}{\rho-1} & \text{if } k^* < i < k^{**} \\ 0 & \text{if } k^{**} \leq i. \end{cases}$$

where k^* is the first index such that $\tau > |y_i^{\downarrow}|$ and k^{**} is the first index such that $\rho|y_i^{\downarrow}| < \tau$. τ is a value in the interval $[|y_k^{\downarrow}|, \rho|y_{k+1}^{\downarrow}|]$, and is defined as

$$\tau = \frac{\rho\sum_{i\in n_1}|y_i^{\downarrow}| + \rho\sum_{i\in n_2}|y_i^{\downarrow}|}{\rho\#n_1 + \#n_2} \tag{22}$$

where n_1 and n_2 are two groups of indices such that $\forall\, i \in n_1, y_i^{\downarrow} < \tau$ and $\forall\, i \in n_2$, $\tau \leq \rho|y_i^{\downarrow}|$ for an $\#n_1$ and $\#n_2$ are the sizes of $n1$ and $n2$. To go from $prox_{\frac{Q}{\rho}}(y)^{\downarrow y}$ to $prox_{\frac{Q}{\rho}}(y)$ we apply the inverse permutation that sort y to $y^{\downarrow}$.

Proof. The result is direct by applying Proposition 3 and Lemma 8 which present the proximal operator of $\text{prox}_{-(\frac{\rho-1}{\rho})\sum_{i=k+1}^{N}(\cdot)^{\downarrow 2}}(y)$, the latter is presented in Appendix A.3. □

Note that the proximal operator of Q is only a relaxation of the proximal operator of $\|x\|_0 \leq k$, which keeps the k largest values of x. Further note that the search for τ can be done iteratively by sorting in descending order all values of $y_i^{\downarrow}$ $i \leq k$ and $\rho y_i^{\downarrow}$ $i > k$ that are (with respect to their absolute value) in the interval $[|y_k^{\downarrow}|, \rho|y_{k+1}^{\downarrow}|]$. The elements in the interval are sorted, and denoted p_i. n_1,n_2 must calculated for each interval $[p_{i+1}, p_i]$. The search is over if τ is $\in [p_{i+1}, p_i]$.

The codes to compute the proximal operator and the cost function are available online: `https://github.com/abechens/SMLM-Constraint-Relaxation`

5 Application to 2D single-molecule localization microscopy

In this section, we compare the minimization of the relaxation with other 2D grid-based sparse algorithms. The algorithms are applied to the problem of 2D Single-Molecule Localization Microscopy (SMLM).

SMLM is a microscopy method that is used to obtain images with a higher resolution than what is possible with traditional optical microscopes. The method was first introduced in [19, 5, 30]. Fluorescent microscopy uses photoactivatable fluorophores that can emit light when they are excited with a laser. The fluorophores are observed with an optical microscope, and, since the fluorophores are smaller than the diffraction limit, what is observed is not each fluorophore, but rather a diffraction pattern (or equivalently the Point Spread Function (PSF)) larger than the fluorophores. This limits the resolution of the image. SMLM exploits photoactivatable fluorophores, and, instead of activating all the fluorophores at once as done by other fluorescent microscopy methods, one activates a sparse set of fluorescent fluorophores. The probability that two fluorophores are in the same PSF is low when only a few fluorophores are activated (low-density images), and precise localization

of each is therefore possible. The localization becomes harder if the density of emitting fluorophores is higher because of the possibility of overlapping PSF's. Once each molecule has been precisely localized, they are switched off and the process is repeated until all the fluorophores have been activated. The total acquisition time may be long when activating few fluorophores at a time, which is unfortunate as SMLM may be used on living samples that can move during this time. We are, in this paper, interested in high-density acquisitions.

The localization problem of SMLM can be described as a $\ell_2 - \ell_0$ minimization problem such as (1) and (2) with an added positivity constraint since we reconstruct the intensity of the fluorophores. For G_Q, this is done by using the distance function to the non-negative space since the proximal operator of the sum of $Q(x)$ and the positivity constraint is not known. A is the matrix operator that performs a convolution with the Point Spread function and a reduction of dimensions. The fluorophores are reconstructed on a finer grid $\in \mathbb{R}^{ML \times ML}$ than the observed image $\in \mathbb{R}^{M \times M}$, with $L > 1$. A detailed description of the mathematical model can be found in [2]. Note that an estimation of the number of excited fluorophores is possible to do beforehand as this is dependent on the intensity of the excitation-laser. Thus the *constrained* sparse formulation (1) may be more suitable to use compared to the penalized sparse formulation (2) as the sparsity parameter k is the maximum number of non-zero pixels to reconstruct, and one pixel can be roughly equivalent to one *observed* excited fluorophore.

We compare first G_Q with Iterative Hard Thresholding (IHT) [17] which minimizes the constrained initial function (1). This gives a clear comparison between the initial function and the proposed relaxation. We construct an image artificially with 213 of fluorophores randomly scattered on a 256×256-grid, where each square measures 25×25 nm. The observed image is 64×64-pixel image, where each pixel measures 100×100 nm, with a simulated Gaussian PSF with an FWHM of 258.21nm. Note that we use these parameters as this is representative of the simulated 2D-ISBI data presented in the next section. We then construct 100 observations by applying different realizations of Poisson noise to the same image. The Signal-to-noise ratio is around 20dB for each observation.

We compare the ability of G_Q and Constrained IHT to minimize the ℓ_2 data fidelity term under the constraint that only 213 pixels are non-zero.

In Fig. 3, we compare the results of G_Q and Constrained IHT using the data fidelity term. The results of the 100 image reconstructions are presented with boxplots. The red mark in the box is the median of the reconstruction result of the 100 noisy, blurred, and downsampled images. The upper (respectively lower) part of the box indicates the 75th (25th) percentiles median. We can observe that G_Q always

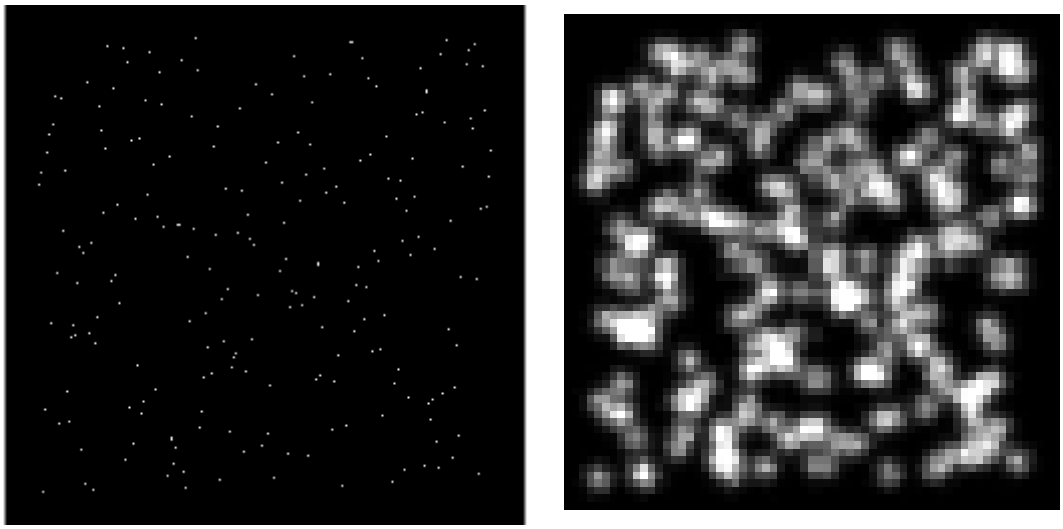


Figure 2: Example of the simulated dataset. The number of fluorophores is 213. To the left: Ground truth. To the right: One of the 100 observations.

minimize better than Constrained IHT in terms of the data fidelity term. Thus it manages more efficiently to solve the initial problem.

G_Q reconstructs the 100 images with a median data fidelity value of 1.55. To compare, Constrained IHT has 2.74 as a median data fidelity value. This small example shows clearly the advantage of using G_Q compared to Constrained IHT. In the next section we compare G_Q and Constrained IHT with other $\ell_2 - \ell_0$ based algorithms.

5.1 Comparison on 2013 ISBI data

We compare G_Q and Constrained IHT with CoBic [2], which is designed to minimize the constrained $\ell_2 - \ell_0$ problem. We further compare the algorithms with two algorithms: $CEL0$ [18] and the ℓ_1 relaxation, both relaxations of the penalized formulation (2). The ℓ_1 relaxation is minimized using FISTA [4], and G_Q is minimized with the Nonmonotone Accelerated Proximal gradient algorithm (nmAPG) [21]. The algorithms are applied to the problem of 2D Single-Molecule Localization Microscopy (SMLM).

The algorithms are tested on two datasets with high-density acquisitions, accessible from the ISBI 2013 challenge [31]. For a review of the SMLM and the different localization algorithms, see the ISBI-SMLM challenge [31]. A more recent challenge

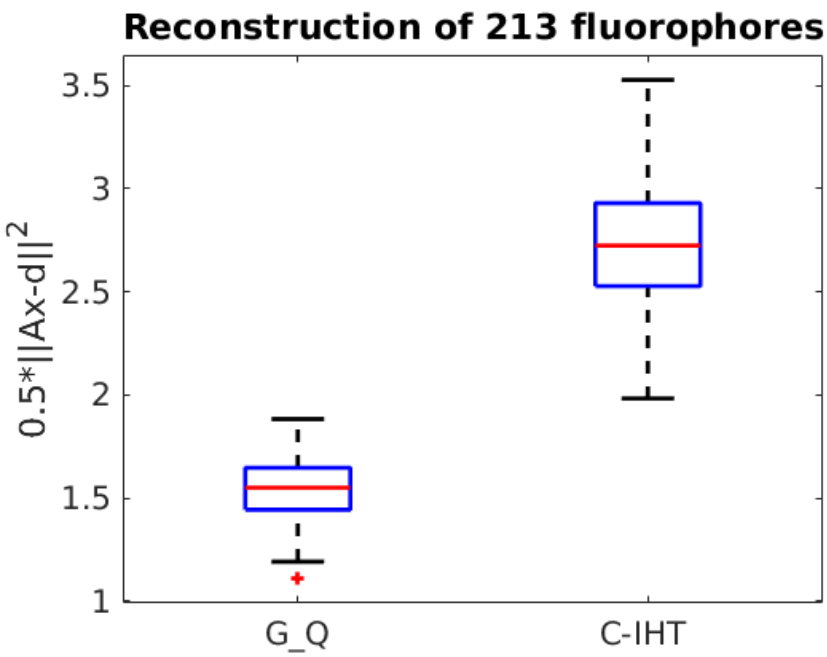


Figure 3: Comparison of the constrained-based algorithms: G_Q and Constrained IHT. The y-axis represent the value $\frac{1}{2}\|Ax - d\|^2$. The lower, the better.

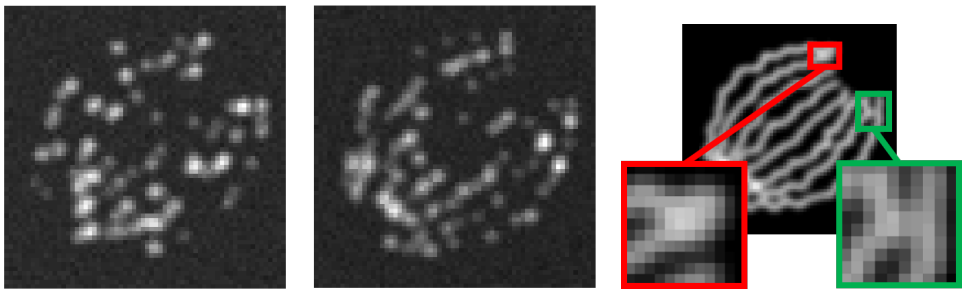


Figure 4: Simulated images, from left to right: 1st acquisition, 361th acquisition, and the sum of all the acquisitions.

was launched in 2016 [32]. We decided to use the 2013 challenge as the data is denser in the 2013 challenge. Furthermore, the 2D data in the 2016 challenge contains observations where some elements are not in the focal plane. Thus our image formation model is not optimized for this image acquisition method.

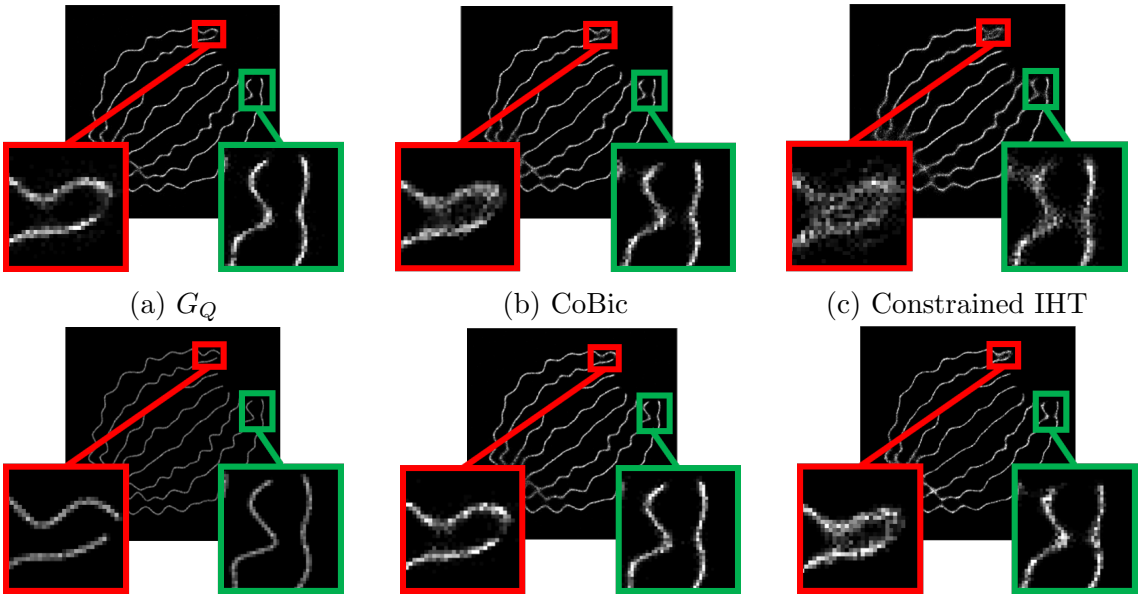
Figure 4 shows two of the 361 acquisitions of the simulated dataset as well as the sum of all the acquisitions. We apply the localization algorithm to each acquisition, and the sum of the results of the localization of the 361 acquisitions yield one super-resolution image. We use the Jaccard index to do a numerical evaluation of the reconstructions. The Jaccard index is known from probability and is used to evaluate similarities between sets. In this case it evaluates the localization of the reconstructed fluorophores (see [31]), and is defined as the ratio between the correctly reconstructed (CR) fluorophores and the sum of CR-, false negatives (FN)- and false positives (FP) fluorophores. The index is 1 for a perfect reconstruction, and the lower the index, the poorer the reconstruction. The Jaccard index includes a tolerance of error in its calculations when identifying the CR, FN and FP.

$$Jac = \frac{CR}{CR + FP + FN} \times 100\%.$$

5.2 Results of the ISBI simulated dataset

The simulated dataset represents 8 tubes of 30 nm diameter. The acquisition is captured on a 64×64 pixel grid with a pixel size of 100×100 nm^2. The acquisition used a simulated Point Spread Function (PSF) modeled by a Gaussian function with a Full Width at Half Maximum (FWHM) of 258.21 nm. Among the 361 images there are 81 049 fluorophores.

The algorithms localize the fluorophores with higher precision on a 256×256



(a) G_Q (b) CoBic (c) Constrained IHT

(d) Ground truth (e) $CEL0$ (f) ℓ_1-relaxation

Figure 5: Reconstructed images from the simulated ISBI dataset, 99 non-zero pixels on average. Top: From left to right: G_Q, CoBic and IHT. Bottom: From left to right: Ground Truth, $CEL0$, and ℓ_1-relaxation.

grid, where each pixel measures 25×25 nm^2. This can be written as a reconstruction of $x \in \mathbb{R}^{ML \times ML}$ with an acquisition $d \in \mathbb{R}^{M \times M}$, where $L = 4$ and $M = 64$. The position of the fluorophore is estimated using the center of the pixel.

We test the reconstruction ability of G_Q with the sparsity constraint k, set to three different values and the Jaccard index is presented in Table 1. The λ parameters for the penalized functional (2) is set such that the same number of non-zero pixels is reconstructed as for the constrained problem. The reconstructions for 99 non-zero pixels from the different algorithms are presented in Fig. 5. The proposed relaxation performs slightly better than CELO. The relaxation performs better than any of the constrained formulation algorithms (CoBic and Constrained IHT), moreover, CoBic does not reconstruct more than 99 non-zero pixels on average. The average reconstruction time for one acquisition is found in Table 2.

Table 1: The Jaccard index obtained for an reconstruction of around 90, 100 and 142 non zero pixels on average. In bold: Best reconstruction for the tolerance and the number of pixel reconstructed.

	Jaccard index (%) for 90 | 99 |142 non-zero pixels on average		
Method/Tolerance	50nm	100nm	150nm
Constrained IHT	20.2|21.3|22.0	35.0|37.8|42.2	38.9|42.9|51.0
$CEL0$	26.7|29.3|**32.7**	**37.7**|41.3|46.9	38.8|42.4|49.2
CoBic	23.9|25.2|-	36.3|40.0 |-	38.2|43.2|-
G_Q	**27.3**|**29.5**|32.5	37.4|**41.9**|42.5	**39.5**|**43.5**|44.0
ℓ_1-relaxation	20.1|22.4|27.5	33.5|37.7|**47.3**	37.5|42.4|**54.1**

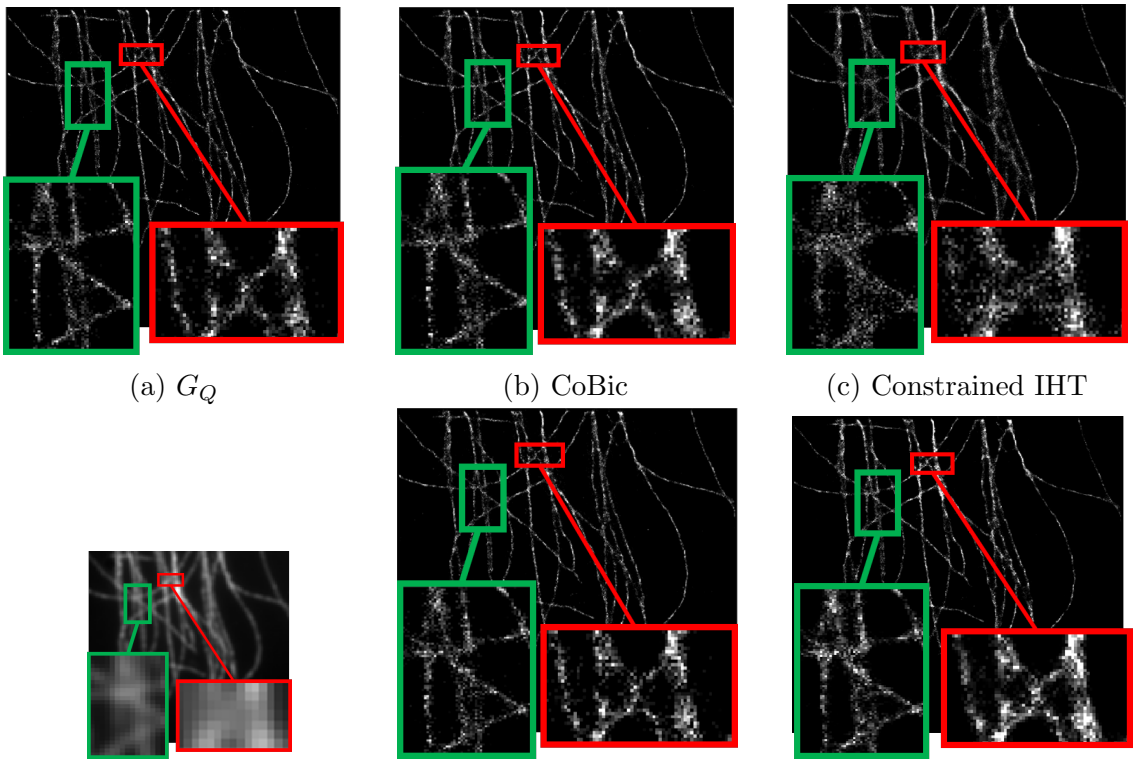
Table 2: Average reconstruction time for one image acquisition for the different methods.

	Average reconstruction time				
Method	G_Q	C. IHT	$CEL0$	CoBic	ℓ_1
Time (s)	84	67	105	87	49

5.3 Results of the real dataset

The algorithms are applied to the real high-density dataset, provided from the 2013 ISBI SMLM challenge [31]. In total there are 500 acquisitions and each acquisition is of size 128×128 pixels and each pixel measures 100×100 nm^2. The FWHM is evaluated to be 351.8 nm [15]. The localization is done on a fine 512×512 pixel grid, where each pixel measures 25×25 nm^2. Extensive testing of the sparsity parameters has been done to obtain the results, presented in Fig. 6, as we have no prior knowledge of the solution. The parameters were chosen such that the parts in red and green had distinctive tubes, as well as the overall tubulins, were reconstructed. The results of the real dataset confirm the results of the simulated data, where the Constrained IHT perform not good, and the ℓ_1 relaxation seems to tighten the holes which are observed in red.

An important note In these numerical experiences, the proposed relaxed formulation converges *always* to a critical point that satisfies the sparsity constraint, and thus the "fail-safe" strategy is never activated.



(a) G_Q (b) CoBic (c) Constrained IHT

(d) Data (e) $CEL0$ (f) ℓ_1-relaxation

Figure 6: Reconstructed images from the Real ISBI dataset. Top: From left to right: G_Q, CoBic and IHT. Bottom: From left to right: Sum of all acquisitions, $CEL0$, and ℓ_1-relaxation.

6 Conclusion

We have investigated in this paper a continuous relaxation of the constrained $\ell_2 - \ell_0$ problem. We compute the convex hull of G_k when A is orthogonal. We further propose to use the same relaxation for any A and name this relaxation G_Q. This is the same procedure as the authors used to obtain $CEL0$ [35]. The question that has driven us has been answered, the proposed relaxation, G_Q is not exact for every observation matrix A. However, it promotes sparsity and is continuous. We propose an algorithm to minimize the relaxed function. We further add a "fail-safe" strategy which ensures convergence to a critical point of the initial functional. In the case of SMLM, the relaxation performs as good as the other grid-based methods, and it converges towards a critical point of the initial problem *each time* without the "fail-safe" strategy activated. Furthermore, the constraint parameter of G_Q is usually easier to fix than the regularizing parameter λ in $CEL0$ in many sparse optimization problems.

A Appendix

A.1 Preliminary results for Lemma 1

Proposition 2 (Reminder) Let $x \in \mathbb{R}^N$. There exists $j \in \mathbb{N}$ such that $0 < j \leq k$ and

$$|x^{\downarrow}_{k-j+1}| \leq \frac{1}{j} \sum_{i=k-j+1}^{N} |x^{\downarrow}_i| \leq |x^{\downarrow}_{k-j}| \tag{23}$$

where the left inequality is strict if $j \neq 1$, and where $x_0 = +\infty$. Furthermore, $T_k(x)$ is defined as the smallest integer that verifies the double inequality.

Proof. First, we suppose that (23) is not true for $j \in \{1, 2, \ldots, k-1\}$, i.e., either

$$|x^{\downarrow}_{j+1}| > \frac{1}{j} \sum_{i=k-j+1}^{N} |x^{\downarrow}_i|, \tag{24}$$

or

$$\frac{1}{j} \sum_{i=k-j+1}^{N} |x^{\downarrow}_i| > |x^{\downarrow}_{k-j}|, \tag{25}$$

or both.

We investigate the case $j = 1$:

$$\sum_{i=k}^{N} |x_i^{\downarrow}| = |x_k^{\downarrow}| + \sum_{i=k+1}^{N} |x_i^{\downarrow}| \geq |x_k^{\downarrow}|. \tag{26}$$

The above inequality is obvious, and we can conclude that for $j = 1$, (24) is false, and thus (25) must be true, i.e.,

$$\sum_{i=k}^{N} |x_i^{\downarrow}| > |x_{k-1}^{\downarrow}|. \tag{27}$$

We investigate $j = 2$:

$$\frac{1}{2} \sum_{i=k-1}^{N} |x_i^{\downarrow}| = \frac{1}{2} \left(|x_{k-1}^{\downarrow}| + \sum_{i=k}^{N} |x_i^{\downarrow}| \right) > |x_{k-1}^{\downarrow}|. \tag{28}$$

We have used (27) to obtain (28). We can conclude that for $j = 2$, (24) is false, and thus (25) is true, i.e.,

$$\frac{1}{2} \sum_{i=k-1}^{N} |x_i^{\downarrow}| > |x_{k-2}^{\downarrow}|. \tag{29}$$

Now, we suppose that $\forall\, j \in \{1, 2, \ldots, k-2\}$ that (24) is false, and (25) is true. We investigate $j = k - 1$:

$$\frac{1}{k-1} \sum_{i=2}^{N} |x_i^{\downarrow}| = \frac{1}{k-1} \left(|x_2^{\downarrow}| + \frac{k-2}{k-2} \sum_{i=3}^{N} |x_i^{\downarrow}| \right) > \frac{1}{k-1} \left(|x_2^{\downarrow}| + (k-2)|x_2^{\downarrow}| \right) = |x_2^{\downarrow}|. \tag{30}$$

We get (30) since we have supposed for $j = k - 2$ that (24) is false, and (25) is true. Thus, we can conclude for $j = k - 1$, (24) is false, and (25) is true, i.e.,

$$\frac{1}{k-1} \sum_{i=2}^{N} |x_i^{\downarrow}| > |x_1^{\downarrow}|. \tag{31}$$

Now, we investigate $j = k$:

$$\frac{1}{k} \sum_{i=1}^{N} |x_i^{\downarrow}| = \frac{1}{k} \left(|x_1^{\downarrow}| + \frac{k-1}{k-1} \sum_{i=2}^{N} |x_i^{\downarrow}| \right) > \frac{1}{k} \left(|x_1^{\downarrow}| + (k-1)|x_1^{\downarrow}| \right) = |x_1^{\downarrow}|. \tag{32}$$

We use (31) to obtain the above inequality. Thus that (24) is false. By definition $x_0^{\downarrow} = +\infty$, and thus (25) is also false. Thus $T_k(x) = k$ verifies the double inequality in (23).

To conclude: Either $T_k(x) = k$, or there exists $j \in \{1, 2, \ldots, k-1\}$ such that $T_k(x) = j$. □

Definition 4. *Let $P^{(x)} \in \mathbb{R}^{N\times N}$ the permutation matrix such that $P^{(x)}x = x^{\downarrow}$. The space $\mathcal{D}(x)$ is defined as:*

$$\mathcal{D}(x) = \{b; P^{(x)}b = b^{\downarrow}\}.$$

$z \in \mathcal{D}(x)$ means $< z, x >=< z^{\downarrow}, x^{\downarrow} >$.

Remark 3. *$\mathcal{D}(x) = \mathcal{D}(|x|)$, since we have $|x^{\downarrow}| = |x|^{\downarrow}$.*

Proposition 4. *Let $(a, b) \in \mathbb{R}^N_{\geq 0} \times \mathbb{R}^N_{\geq 0}$. Then*

$$\sum_i a_i b_i \leq \sum_i a_i^{\downarrow} b_i^{\downarrow}$$

and the inequality is strict if $b \notin \mathcal{D}(a)$.

Proof. [34, Lemma 1.8] proves it without proving the strict inequality.

We assume that a is not on the form $a = t(1, 1 \ldots, 1)^T$, i.e., there exists $i \neq j$, $a_i \neq a_j$. Moreover, for simplicity, without loss of generality, we suppose $a = a^{\downarrow}$. We write

$$\sum_i^N a_i b_i = a_N \sum_{i=1}^N b_i + (a_{N-1} - a_N) \sum_{i=1}^{N-1} b_i + \cdots + (a_1 - a_2) b_1.$$

As it is obvious that $\forall\, j = 1, \ldots N$

$$\sum_{i=1}^j b_i \leq \sum_{i=1}^j b_i^{\downarrow}, \tag{33}$$

and since $a_{j-1} - a_j \geq 0 \,\forall\, j$, we get

$$\sum_{i=1}^N a_i b_i \leq \sum_{i=1}^N a_i b_i^{\downarrow} = \sum_{i=1}^N a_i^{\downarrow} b_i^{\downarrow} \tag{34}$$

The goal of Proposition 4 is to show that the inequality in (34) is strict if $b \notin \mathcal{D}(a)$.

First, we can remark if $b \notin \mathcal{D}(a)$, then there exists $j_0 \in \{2, 3, \ldots, N\}$ such

$$\sum_{i=1}^{j_0-1} b_i < \sum_{i=1}^{j_0-1} b_i^{\downarrow}. \tag{35}$$

By contradiction, if (35) is not true, we have $\forall\, j \in \{2, 3, \ldots, N\}$

$$\sum_{1=1}^{j-1} b_i^{\downarrow} \leq \sum_{1=1}^{j-1} b_i,$$

and with (33), we get

$$\sum_{1=1}^{j-1} b_i^{\downarrow} = \sum_{1=1}^{j-1} b_i. \tag{36}$$

From (36), we easily obtain $\forall\, j$,

$$b_j = b_j^{\downarrow},$$

which means $b^{\downarrow} = b$, i.e., $b \in \mathcal{D}(a)$, which contradicts the hypothesis $b \notin \mathcal{D}(a)$. So there exists j_0 such that (35) is true, and if $a_{j_0-1} \neq a_{j_0}$

$$(a_{j_0-1} - a_{j_0}) \sum_{i=1}^{j_0-1} b_i < (a_{j_0-1} - a_{j_0}) \sum_{i=1}^{j_0-1} b_i^{\downarrow},$$

which, with (33), implies

$$\sum_{i=1}^{N} a_i b_i < \sum_{i=1}^{N} a_i b_i^{\downarrow}.$$

It remains to examine the case where $a_{j_0-1} = a_{j_0}$. In this case we claim there exists $j_1 \in \{1, \ldots, j_{0-2}\}$ or $\{j_0, \ldots, N\}$ such that

$$\sum_{i=1}^{j_1} b_i < \sum_{i=1}^{j_1} b_i^{\downarrow} \text{ or } \sum_{i=j_0}^{j_1} b_i < \sum_{i=j_0}^{j_1} b_i^{\downarrow}. \tag{37}$$

If not, with the same proof as before we get

$$b_i^{\downarrow} = b_i \quad i \in \{1, \ldots, j_0 - 2\} \cup \{j_0 + 1, \ldots, N\},$$

i.e., we have

$$\begin{pmatrix} b_1^{\downarrow} \\ b_2^{\downarrow} \\ \vdots \\ b_{j_0-2}^{\downarrow} \\ x_1^{\downarrow} \\ x_2^{\downarrow} \\ b_{j_0+1}^{\downarrow} \\ \vdots \\ b_N^{\downarrow} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{j_0-2} \\ x_1 \\ x_2 \\ b_{j_0+1} \\ \vdots \\ b_N \end{pmatrix}$$

where $(x_1, x_2) = (b_{j_0-1}, b_{j_0})$ or (b_{j_0}, b_{j_0-1}). The order does not matter since $a_{j_0-1} = a_{j_0}$. This implies that $b \in \mathcal{D}(a)$, which contradicts the hypothesis. So (37) is true and we get for example

$$(a_{j_1-1} - a_{j_1}) \sum_{i=1}^{j_1-1} b_i < (a_{j_1-1} - a_{j_1}) \sum_{i=1}^{j_1-1} b_i^{\downarrow},$$

and if $a_{j_1-1} - a_{j_1} \neq 0$ we deduce

$$\sum_i a_i b_i < \sum_i a_i b_i^{\downarrow}. \tag{38}$$

If $a_{j_1-1} = a_{j_1}$ we repeat the same argument and proof as above, and we are sure to find an index j_w such that $a_{j_w-1} - a_{j_w} \neq 0$ since we have supposed that $a \neq t(1, 1, \ldots, 1)^T$. Therefore (38) is always true which concludes the proof. □

Proposition 5 ([38])**.** *$g(x) : \mathbb{R}^N \to \mathbb{R}$ defined as $g(x) = \frac{1}{2}\sum_{i=1}^k x_i^{\downarrow 2}$, is convex. Furthermore, note that $g(|x|) = g(x)$.*

Lemma 4. *Let $f_1(z, x) \in \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ be defined as*

$$f_1(z, x) := -\frac{1}{2} \sum_{i=1}^k z_i^{\downarrow 2} + < z^{\downarrow}, x^{\downarrow} > .$$

Let us consider the concave problem

$$\sup_{z \in \mathbb{R}^N_{\geq 0}} f_1(z, |x|). \tag{39}$$

Problem (39) has the following optimal arguments

$$\arg\sup_{z\in\mathbb{R}^N_{\geq 0}} f_1(z,|x|) = \{z; \exists\, P \in \mathbb{R}^{N\times N} \text{ a permutation matrix s.t. } Pz = \hat{z}\}, \tag{40}$$

where $\hat{z}$ is defined as

$$\hat{z}_j = \begin{cases} \frac{1}{T_k(x)} \sum_{i=k-T_k(x)+1}^{N} |x_i^{\downarrow}| & \text{if } k \geq j \geq k - T_k(x) + 1 \\ & \text{or if } j > k \text{ and } x_j^{\downarrow} \neq 0 \\ \left[0, \frac{1}{T_k(x)} \sum_{i=k-T_k(x)+1}^{N} |x_i^{\downarrow}|\right] & \text{if } j > k \text{ and } x_j^{\downarrow} = 0 \\ |x_j^{\downarrow}| & \text{if } j < k - T_k(x) + 1. \end{cases} \tag{41}$$

We can remark that $\hat{z} = \hat{z}^{\downarrow}$, and $T_k(x)$ is defined in Proposition 2. The value of the supremum problem is

$$\frac{1}{2} \sum_{i=1}^{k-T_k(x)} x_i^{\downarrow 2} + \frac{1}{2T_k(x)} \left(\sum_{i=k-T_k(x)+1}^{N} |x_i^{\downarrow}| \right)^2. \tag{42}$$

Proof. Problem (39) can be written as

$$\sup_{z\in\mathbb{R}^N_{\geq 0}} \sum_{i=1}^{k} |x_i^{\downarrow}| z_i^{\downarrow} - \frac{1}{2} \sum_{i=1}^{k} z_i^{\downarrow 2} + \sum_{i=k+1}^{N} |x_i^{\downarrow}| z_i^{\downarrow}. \tag{43}$$

We remark that finding the supremum for $z_i^{\downarrow}, i > k$ reduces to find the supremum of the following term, knowing that $z_i^{\downarrow}$ is upper bounded by $z_{i-1}^{\downarrow}$:

$$\sum_{i=k+1}^{N} |x_i^{\downarrow}| z_i^{\downarrow}. \tag{44}$$

Let $z_k^{\downarrow}$ be a constant. The sum in (44) is non-negative and increasing with respect to $z_j^{\downarrow}$ and the supremum is obtained when $z_j^{\downarrow}$ reaches its upper bound, i.e $z_j^{\downarrow} = z_{j-1}^{\downarrow} \forall j > k$ and $|x_j^{\downarrow}| \neq 0$. By recursion, $z_j^{\downarrow} = z_k^{\downarrow} \forall j > k$ and $|x_j^{\downarrow}| \neq 0$. When $\exists\, j > k, |x_j^{\downarrow}| = 0$, we observe that $z_j^{\downarrow}$ is multiplied with zero, and can take on every value between its lower bound and upper bound, which is between 0 and $z_k^{\downarrow}$. Then, obviously, the supremum arguments for (44) is

$$z_i^{\downarrow} \begin{cases} = z_k^{\downarrow} \text{ if } |x_i^{\downarrow}| \neq 0 \\ \in [0, z_k^{\downarrow}] \text{ if } |x_i^{\downarrow}| = 0 \end{cases} \tag{45}$$

Further, from (43), we observe that for $i < k$, the optimal argument is

$$z_i^{\downarrow} = \max(|x_i^{\downarrow}|, z_{i+1}^{\downarrow}). \tag{46}$$

By recursion, we can write this as

$$z_i^{\downarrow} = \max(|x_i^{\downarrow}|, z_k^{\downarrow}). \tag{47}$$

It remains to find the value of $z_k^{\downarrow}$.

Inserting (45) and (47) into (43), and we obtain:

$$\sup_{z_k^{\downarrow}} \sum_{i=1}^{k} |x_i^{\downarrow}| \max(|x_i^{\downarrow}|, z_k^{\downarrow}) - \frac{1}{2} \sum_{i=1}^{k} \max(|x_i^{\downarrow}|, z_k^{\downarrow})^2 + \sum_{i=k+1}^{N} |x_i^{\downarrow}| z_k^{\downarrow}. \tag{48}$$

To treat the term $\max(|x_i^{\downarrow}|, z_k^{\downarrow})$, we introduce $j^*(k) = \sup_j \{j : z_k^{\downarrow} \leq |x_j^{\downarrow}|\}$, i.e., $j^*(k)$ is the largest index such that $|x_{j^*(k)}^{\downarrow}| \geq z_k^{\downarrow}$, and we define $x_0^{\downarrow} = +\infty$. Therefore, (48) rewrites as

$$\sup_{z_k^{\downarrow}} \sum_{i=1}^{j^*(k)} |x_i^{\downarrow}|^2 - \frac{1}{2} \sum_{i=1}^{j^*(k)} |x_i^{\downarrow}|^2 + \sum_{i=j^*(k)+1}^{k} |x_i^{\downarrow}| z_k^{\downarrow} - \frac{1}{2} \sum_{i=j^*(k)+1}^{k} z_k^{\downarrow 2} + \sum_{i=k+1}^{N} |x_i^{\downarrow}| z_k^{\downarrow}. \tag{49}$$

(49) is a concave problem, and the optimiality conditions yields

$$- \sum_{i=j^*(k)+1}^{k} z_k^{\downarrow} + \sum_{j^*(k)+1}^{N} |x_i^{\downarrow}| = 0. \tag{50}$$

We define $\sum_{i=j^*(k)+1}^{k} 1 = S$. Then $j^*(k) = k - S$ and

$$z_k^{\downarrow} = \frac{1}{S} \sum_{k-S+1}^{N} |x_i^{\downarrow}|. \tag{51}$$

Furthermore, since $j^*(k) = k - S$ was the largest index such that $|x_{k-S}| \geq z_k^{\downarrow} > |x_{k-S+1}|$. This translates to

$$|x_{k-S}^{\downarrow}| \geq \frac{1}{S} \sum_{k-S+1}^{N} |x_i^{\downarrow}| > |x_{k-S+1}^{\downarrow}|,$$

which implies $S = T_k(x)$ (see Proposition 2). Note that if $j^*(k) = k$ (which is the same to say $T_k(x) = 1$), then the right part of the above inequality is not strict.

Now, assume $|x^{\downarrow}_{j^*(k)}| = z^{\downarrow}_k$. Then, the max function can both take $z^{\downarrow}_k$ or $|x^{\downarrow}_{j^*(k)}|$. If it is the latter, than the expression above is correct. In the former case $\max(|x^{\downarrow}_{j^*(k)}|, z^{\downarrow}_k) = z^{\downarrow}_k$. We obtain

$$z^{\downarrow}_k = \frac{1}{T_k(x)+1} \sum_{k-T_k(x)}^{N} |x^{\downarrow}_i|. \tag{52}$$

Furthermore, we use the fact that $|x^{\downarrow}_{j^*(k)}| = z^{\downarrow}_k$ and $j^*(k) = k - T_k(x)$, and develop (52) as,

$$z^{\downarrow}_k = \frac{1}{T_k(x)+1} \left(x_{k-T_k(x)} + \sum_{k-T_k(x)+1}^{N} |x^{\downarrow}_i| \right) \tag{53}$$

$$(T_k(x)+1) z^{\downarrow}_k = z^{\downarrow}_k + \sum_{k-T_k(x)+1}^{N} |x^{\downarrow}_i| \tag{54}$$

$$T_k(x) z^{\downarrow}_k = \sum_{k-T_k(x)+1}^{N} |x^{\downarrow}_i| \tag{55}$$

$$z^{\downarrow}_k = (51) \tag{56}$$

The unique value of $z^{\downarrow}_k$ is given by (51).

□

Lemma 5. *Let $x \in \mathbb{R}^N$ and $f_2(y, x) \in \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$, defined as*

$$f_2(y, x) = -\frac{1}{2} \sum_{i=1}^{k} y_i^{\downarrow 2} + < y, x >$$

The following concave supremum problem

$$\sup_{y \in \mathbb{R}^N} f_2(y, x) \tag{57}$$

is equivalent to

$$\sup_{z \in \mathbb{R}^N_{\geq 0}} f_2(z, |x|). \tag{58}$$

The arguments are such that $\hat{y}^{\downarrow}_i = \text{sign}^*(x_i^{\downarrow \hat{z}}) \hat{z}^{\downarrow}_i$.

Proof. Let $\hat{z} \in \mathbb{R}^N_{\geq 0}$ be the argument of the supremum in (58), $\hat{y}$ be such that $\hat{y}_i = \text{sign}(x_i)\hat{z}_i$, and note that $f_2(y,x) = -g(y) + < y, x >$ with g defined as in Proposition 5 in Appendix A.1. First, $f_2(y,x)$ is a concave function in y (see Proposition 5). Furthermore, $f_2(y,x)$ is such that $-f_2(y,x)$ is coercive in y. Thus a supremum exists. Further note that $g(\hat{y}) = g(|\hat{y}|) = g(\hat{z})$. Then the following sequence of equalities/inequalities completes the proof:

$$
\begin{aligned}
(58) &= \sup_{z \in \mathbb{R}^N_{\geq 0}} f_2(z, |x|) = -g(\hat{z}) + \sum_{i=1}^{N} \hat{z}_i |x_i| = -g(\hat{z}) + \sum_{i=1}^{N} \text{sign}(x_i)\hat{z}_i x_i \\
&= -g(\hat{y}) + \sum_{i=1}^{N} \hat{y}_i x_i \leq (57) = \sup_{y \in \mathbb{R}^N} f_2(y,x) \underset{<y,x> \leq <|y|,|x|>}{\leq} \sup_{y \in \mathbb{R}^N} f_2(|y|, |x|) \\
&= \sup_{z \in \mathbb{R}^N_{\geq 0}} f_2(z, |x|) = (58)
\end{aligned}
$$

□

A.2 Proof of Lemma 1

Proof. Note that a similar problem has been studied in [1]. They do however work with low-rank approximation, therefore they did not have the problem of how to permute x since they work with matrices. First, let $\mathcal{D}(x)$ be as defined in Definition 4.

We are interested in

$$\sup_{y \in \mathbb{R}^N} f_2(y,x),$$

and its arguments, with f_2 defined in Lemma 5. From this lemma, we know that we can rather study

$$\sup_{z \in \mathbb{R}^N_{\geq 0}} f_2(z, |x|).$$

Furthermore, from Lemma 4, we know the expression of $\sup_{z \in \mathbb{R}^N_{\geq 0}} f_1(z, |x|)$ and its arguments. We want to show that $\sup_{z \in \mathbb{R}^N_{\geq 0}} f_2(z, |x|) = \sup_{z \in \mathbb{R}^N_{\geq 0}} f_1(z, |x|)$, and to find a connection between the arguments of f_2 and f_1.

First, note that

$$\sup_{z \in \mathbb{R}^N_{\geq 0}} f_2(z, |x|) \geq \sup_{z \in \mathbb{R}^N_{\geq 0} \in \mathcal{D}(x)} f_2(z, |x|). \tag{59}$$

From [34, Lemma 1.8] and Proposition 4, we have that $\forall (y,x) \in \mathbb{R}^N_{\geq 0} \times \mathbb{R}^N_{\geq 0}$:

$$< y, x > \leq < y^{\downarrow}, x^{\downarrow} >,$$

and the inequality is strict if $y \notin \mathcal{D}(x)$, and thus

$$\sup_{z \in \mathbb{R}^N_{\geq 0}} f_2(z, |x|) \leq \sup_{z \in \mathbb{R}^N_{\geq 0}} f_1(z, |x|). \tag{60}$$

Note that we have $\mathcal{D}(|x|) = \mathcal{D}(x)$, then $\forall z \in \mathcal{D}(x)$, $f_2(z, |x|) = f_1(z, |x|)$ and:

$$\sup_{z \in \mathbb{R}^N_{\geq 0} \in \mathcal{D}(x)} f_2(z, |x|) = \sup_{z \in \mathbb{R}^N_{\geq 0}} \sum_{i=1}^{N} z_i^{\downarrow} |x_i^{\downarrow}| - \frac{1}{2} \sum_{i=1}^{k} z_i^{\downarrow 2} = \sup_{z \in \mathbb{R}^N_{\geq 0}} f_1(z, |x|). \tag{61}$$

Using inequalities (59) and (60) and connecting them to (61), we obtain

$$\sup_{z \in \mathbb{R}^N_{\geq 0}} f_1(z, |x|) = \sup_{z \in \mathbb{R}^N_{\geq 0} \in \mathcal{D}(x)} f_2(z, |x|) \leq \sup_{z \in \mathbb{R}^N_{\geq 0}} f_2(z, |x|) \leq \sup_{z \in \mathbb{R}^N_{\geq 0}} f_1(z, |x|).$$

$f_2(z, |x|)$ is upper and lower bounded by the same value, thus we have

$$\sup_{z \in \mathbb{R}^N_{\geq 0}} f_2(z, |x|) = \sup_{z \in \mathbb{R}^N_{\geq 0}} f_1(z, |x|) \tag{62}$$

The $\sup_{z \in \mathbb{R}^N_{\geq 0}} f_1(z, |x|)$ is known from Lemma 4:

$$\sup_{z \in \mathbb{R}^N_{\geq 0}} f_1(z, |x|) = \frac{1}{2} \sum_{i=1}^{k-T_k(x)} x_i^{\downarrow 2} + \frac{1}{2T_k(x)} \left(\sum_{i=k-T_k(x)+1}^{N} |x_i^{\downarrow}| \right)^2 \tag{63}$$

with the optimal arguments:

$$\arg\sup_{z \in \mathbb{R}^N_{\geq 0}} f_1(z, |x|) = \{z; \exists\, P \in \mathbb{R}^{N \times N} \text{ a permutation matrix s.t. } Pz = \hat{z}\}, \tag{64}$$

where $\hat{z}$ is such that:

$$\hat{z}_j = \begin{cases} \frac{1}{T_k(x)} \sum_{i=k-T_k(x)+1}^{N} |x_i^{\downarrow}| & \text{if } k \geq j \geq k - T_k(x) + 1 \\ & \text{or if } j > k \text{ and } x_j^{\downarrow} \neq 0 \\ \left[0, \frac{1}{T_k(x)} \sum_{i=k-T_k(x)+1}^{N} |x_i^{\downarrow}|\right] & \text{if } j > k \text{ and } |x_j^{\downarrow}| = 0 \\ |x_j^{\downarrow}| & \text{if } j < k - T_k(x) + 1. \end{cases} \tag{65}$$

Now we are interested in the optimal arguments of f_2. Let $P^{(x)}$ be such that $P^{(x)}x = x^{\downarrow}$. We define $z^* = P^{(x)^{-1}}\hat{z}$. Evidently, $P^{(x)}z^* = \hat{z}$, and since $\hat{z}$ is sorted by its absolute value, $P^{(x)}z^* = z^{*\downarrow}$, and thus $z^* \in \mathcal{D}(x)$. Furthermore, from Lemma 4, z^* is a optimal argument of f_1.

We have then $f_2(z^*, |x|) = f_1(z^*, |x|) = \sup_{z\in\mathbb{R}^N_{\geq 0}} f_1(z, |x|)$. z^* is therefore an optimal argument of f_2 since (62) shows the equality between the supremum value of f_1 and f_2.

We have shown that there exists $\hat{z} \in \arg\sup_{z\in\mathbb{R}^N_{\geq 0}} f_1(z, |x|)$, from which we can construct $z^* \in \mathcal{D}(x)$, an optimal argument of f_2. Now, by contradiction, we show that all optimal arguments of f_2 are in $\mathcal{D}(x)$. Assume $\hat{z} = \arg\sup_{z\in\mathbb{R}^N_{\geq 0}} f_2(z, |x|)$ and that $\hat{z} \notin \mathcal{D}(x)$. We can construct z^*, such that $z^{*\downarrow} = \hat{z}^{\downarrow}$, and $z^* \in \mathcal{D}(x)$. We have then

$$\begin{aligned} f_2(z^*, |x|) - f_2(\hat{z}, |x|) &= -\frac{1}{2}\sum_i^k z_i^{*\downarrow 2} + < z^*, |x| > + \frac{1}{2}\sum_i^k \hat{z}_i^{\downarrow 2} - < \hat{z}, |x| > \\ &= < z^*, |x| > - < \hat{z}, |x| > = < z^{*\downarrow}, |x^{\downarrow}| > - < \hat{z}, |x| > > 0. \end{aligned}$$

The last equality is due to $z^* \in \mathcal{D}(x)$, and the last inequality is from Proposition 4. Thus $\hat{z}$ is not an optimal argument for f_2, and all optimal arguments of f_2 must be in $\mathcal{D}(x)$.

Furthermore, thus it suffices to study $\sup_{z\in\mathbb{R}^N_{\geq 0}\in\mathcal{D}(z)} f_2(z, |x|)$, and from (61), we can rather study f_1, and construct all supremum arguments of f_2 from f_1.

$$\arg\sup_{z\in\mathbb{R}^N_{\geq 0}} f_2(z, |x|) = P^{(x)^{-1}}\hat{z} \tag{66}$$

where $\hat{z}$ is defined in (65). □

A.3 Calculation of Proximal operator of $\zeta(x)$

As preliminary results, we state and prove the two following lemmas 6 and 7.

Lemma 6. *Let $j : \mathbb{R} \to \mathbb{R}$ be a strictly convex and coercive function, let $w = \arg\min_t j(t)$, and let us suppose that j is symmetric with respect to its minimum, i.e. $j(w - t) = j(w + t)\,\forall t \in \mathbb{R}$. The problem*

$$z = \arg\min_{b\leq|t|\leq a} j(t)$$

with a and b positive, has the following solution

$$z = \begin{cases} w & \text{if } b \leq |w| \leq a \\ \text{sign}^*(w)a & \text{if } |w| \geq a \\ \text{sign}^*(w)b & \text{if } |w| \leq b. \end{cases}$$

Proof. Since j is symmetric with respect to its minimum $j(w+t_1) \leq j(w+t_2)\, \forall |t_1| \leq |t_2|$. Assume that $0 < w \leq b$. We can write $j(b) = j(w+\alpha)$, $\alpha > 0$ and $j(-b) = j(w+\beta), \beta < 0$. Since $w > 0$ then $|\alpha| < |\beta|$ and thus the minimum is reached with $z = b$ on the interval $[b, a]$. Similar reasoning can be used to prove the other cases. □

Lemma 7. *Let $g_i : \mathbb{R} \to \mathbb{R}, i \in [1..N]$ be strictly convex and coercive. Let $w = (w_1, w_2, \ldots w_N)^T = \arg\min_{t_i} \sum g_i(t_i)$, i.e., $w_i = \arg\min_{t_i} g_i(t_i)$. Assume that $|w_1| \geq |w_2| \geq \cdots \geq |w_k|$ and $|w_{k+1}| \geq |w_{k+2}| \geq \cdots \geq |w_N|$. Let g_i be symmetric with respect to its minimum. Consider the following problem*

$$\arg\min_{|t_1| \geq \cdots \geq |t_N|} \sum_i^N g_i(t_i). \tag{67}$$

The optimal solution is

$$t_i(\tau) = \begin{cases} \text{sign}^*(w_i) \max(|w_i|, \tau) & \text{if } 1 \leq i \leq k \\ \text{sign}^*(w_i) \min(|w_i|, \tau) & \text{if } i > k \end{cases} \tag{68}$$

where $\tau \in \mathbb{R}$ is in $[\min(|w_k|, |w_{k+1}|), \max(|w_k|, |w_{k+1}|)]$ and is the value that minimizes $\sum g_i(t_i(\tau))$.

Proof. Note that this proof is inspired by [20, Theorem 2], with some modifications. First, if $|w_k| \geq |w_{k+1}|$, then w satisfies the constraints in Problem (67), and thus w is the optimal solution. If $|w_k| < |w_{k+1}|$ we must search a little more. In both cases we can, since each g_i is convex and symmetric with respect to its minimum, apply Lemma 6 for t_i, and the choices can be limited to the following choices:

$$t_i = \begin{cases} w_i & \text{if } |t_{i-1}| \geq |w_i| \geq |t_{i+1}| \\ \text{sign}^*(w_i)|t_{i+1}| & \text{if } |w_i| < |t_{i+1}| \\ \text{sign}^*(w_i)|t_{i-1}| & \text{if } |w_i| > |t_{i-1}| \end{cases} \tag{69}$$

This can be rewritten in a shorter form, at first in the case where $i \leq k$.

$$t_i = \text{sign}(w_i)^* \max\left(|w_i|, |t_{i+1}|\right). \tag{70}$$

This can be proved by recursion. In the case of $i = 1$, w_1 is the optimal argument if $|w_1| \geq |t_2|$, otherwise $\text{sign}^*(w_1)|t_2|$ is optimal. Therefore $t_1 = \text{sign}^*(w_1)\max(|w_1|, |t_2|)$. Assume that this is true for the i-th index.

$$t_{i+1} = \begin{cases} w_{i+1} & \text{if } |t_i| \geq |w_{i+1}| \geq |t_{i+2}| \text{ and } i+1 \leq k \\ \text{sign}^*(w_{i+1})|t_{i+2}| & \text{if } |w_{i+1}| < |t_{i+2}| \text{ and } i+1 \leq k \\ \text{sign}^*(w_{i+1})|t_i| & \text{if } |w_{i+1}| > |t_i| \text{ and } i+1 \leq k. \end{cases} \tag{71}$$

But $t_i = \text{sign}^*(w_i)\max(|w_i|, |t_{i+1}|)$, which yields $|t_i| \geq |w_i| \geq |w_{i+1}|$ and thus the third case of (71) can be ignored.

Now assume for an $i \leq k$ that $t_i \neq w_i$. This implies that

$$|t_i| = |t_{i+1}| > |w_i|.$$

Since w_i is non increasing for $i \leq k$, the following inequality $|t_{i+1}| > |w_{i+1}|$ is true. Furthermore, $|t_{i+1}| = \max(|w_{i+1}|, |t_{i+2}|) = |t_{i+2}|$. By recursion, we have

$$|t_i| = |t_{i+1}| = |t_{i+2}| = \cdots = |t_k|.$$

To facilitate the notations, $|t_k| = \tau$. The lemma is proved by inserting τ instead of $|t_{i+1}|$ and $|t_k|$ into equation (70)

When $i > k$, a similar proof of recursion gives:

$$t_i = \text{sign}^*(w_i)\min(|t_k|, |w_i|). \tag{72}$$

and by adopting the notation τ, we finish the proof. □

Remark 4. *Note that if w, defined in Lemma 7 is such that $|w_k| \geq |w_{k+1}|$, then w is solution of (67).*

Lemma 8. *Let $y \in \mathbb{R}^N$. Define $\zeta : \mathbb{R}^N \to \mathbb{R}$ as $\zeta(x) := -(\frac{\rho-1}{\rho})\sum_{i=k+1}^{N}(x_i)^{\downarrow 2}$. The proximal operator of ζ is such that*

$$prox_{\zeta(\cdot)}(y)^{\downarrow y} = \begin{cases} \text{sign}(y_i^{\downarrow})\max\left(|y_i^{\downarrow}|, \tau\right) & \textit{if } i \leq k \\ \text{sign}(y_i^{\downarrow})\min(\tau, |\rho y_i^{\downarrow}|) & \textit{if } i > k. \end{cases} \tag{73}$$

If $|y_k^{\downarrow}| < \rho|y_{k+1}^{\downarrow}|$ *then* τ *is a value in the interval* $[|y_k^{\downarrow}|, \rho|y_{k+1}^{\downarrow}|]$*, and is defined as*

$$\tau = \frac{\rho \sum_{i\in n_1} |y_i^{\downarrow}| + \rho \sum_{i\in n_2} |y_i^{\downarrow}|}{\rho \# n_1 + \# n_2} \tag{74}$$

where n_1 *and* n_2 *are two groups of indices such that* $\forall\, i \in n_1, y_i^{\downarrow} < \tau$ *and* $\forall\, i \in n_2$, $\tau \leq \rho|y_i^{\downarrow}|$ *for an* $\# n_1$ *and* $\# n_2$ *are the sizes of* $n1$ *and* $n2$*. To go from* $prox_{\zeta(\cdot)}(y)^{\downarrow y}$ *to* $prox_{\zeta(\cdot)}(y)$ *we apply the inverse permutation that sort* y *to* $y^{\downarrow}$*.*

Note that we search

$$\text{prox}_{-(\frac{\rho-1}{\rho})\sum_{i=k+1}^{N}(\cdot)^{\downarrow 2}}(y) = \arg\min_x -\frac{1}{2}\sum_{i=k+1}^{N} x_i^{\downarrow 2} + \frac{\rho}{2(\rho-1)}\|x-y\|_2^2$$

We define two functions, $l_1 : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ and $l_2 : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$.

$$l_1(z, a) = \frac{\rho}{2(\rho-1)}\sum_{i}^{N}(z_i - |a_i|)^2 - \frac{1}{2}\sum_{i=k+1}^{N} z_i^{\downarrow 2} \tag{75}$$

$$l_2(z, |a|) = \frac{\rho}{2(\rho-1)}\sum_{i}^{N}(z_i^{\downarrow} - |a_i^{\downarrow}|)^2 - \frac{1}{2}\sum_{i=k+1}^{N} z_i^{\downarrow 2}. \tag{76}$$

As in Lemma 1, we can create relations between l_1 and l_2, where l_2 can be solved using Lemma 7.
We omit the proof as it is similar to the one of Lemma 1.

A.4 The algorithm

Algorithm 1: Nonmonotone APG

Initialization:

$z^{(1)} = x^{(1)} = x^{(0)}$,

$t^{(1)} = 1, t^{(0)} = 0, \eta \in [0, 1), \delta > 0, c^{(1)} = F(x^{(1)}), q^{(1)} = 1, \alpha_x < \frac{1}{L}, \alpha_y < \frac{1}{L}$

Repeat:

$$y^{(p)} = x^{(p)} + \frac{t^{(p-1)}}{t^{(p)}}(z^{(p)} - x^{(p)}) + \frac{t^{(p-1)} - 1}{t^{(p)}}(x^{(p)} - x^{(p-1)})$$

$$z^{(p+1)} = \text{prox}_{\alpha_x g}(y^{(p)} - \alpha_y \nabla f(y^{(p)}))$$

if $F(z^{(p+1)}) \leq c^{(p)} - \delta \|z^{(p+1)} - y^{(p)}\|^2$ **then:**

$$x^{(p+1)} = z^{(p+1)}$$

else:

$$v^{(p+1)} = \text{prox}_{\alpha_x g}(x^{(p)} - \alpha_y \nabla f(x^{(p)}))$$

$$x^{(p+1)} = \begin{cases} z^{(p+1)} \text{ if } F(z^{(p+1)}) \leq F(v^{(p+1)}) \\ v^{(p+1)} \text{ otherwise} \end{cases}$$

end if.

$$t^{(p+1)} = \frac{\sqrt{4(t^{(p)})^2 + 1} + 1}{2}$$

$$q^{(p+1)} = \eta q^{(p)} + 1$$

$$c^{(p+1)} = \frac{\eta q^{(p)} c^{(p)} + F(x^{(p+1)})}{q^{(p+1)}}$$

Until: Convergence

Acknowledgment The authors would like to thank the anonymous reviewers for their detailed comments and suggestions.
This work has been supported by the French government, through a financial Ph.D. allocation from MESRI and through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] Andersson, F., Carlsson, M., Olsson, C.: Convex envelopes for fixed rank approximation. Optimization Letters **11**(8), 1783–1795 (2017)

[2] Bechensteen, A., Blanc-Féraud, L., Aubert, G.: New $l_2 - l_0$ algorithm for single-molecule localization microscopy. Biomedical Optics Express **11**(2), 1153–1174 (2020)

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

[4] Beck, A., Teboulle, M.: A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. SIAM Journal on Imaging Sciences **2**(1), 183–202 (2009). DOI 10.1137/080716542

[5] Betzig, E., Patterson, G.H., Sougrat, R., Lindwasser, O.W., Olenych, S., Bonifacino, J.S., Davidson, M.W., Lippincott-Schwartz, J., Hess, H.F.: Imaging Intracellular Fluorescent Proteins at Nanometer Resolution. Science **313**(5793), 1642–1645 (2006). DOI 10.1126/science.1127344

[6] Bi, S., Liu, X., Pan, S.: Exact penalty decomposition method for zero-norm minimization based on mpec formulation. SIAM Journal on Scientific Computing **36**(4), A1451–A1477 (2014)

[7] Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Mathematical Programming **146**(1), 459–494 (2014). DOI 10.1007/s10107-013-0701-9

[8] Bourguignon, S., Ninin, J., Carfantan, H., Mongeau, M.: Exact sparse approximation problems via mixed-integer programming: Formulations and computational performance. IEEE Transactions on Signal Processing **64**(6), 1405–1419 (2016)

[9] Breiman, L.: Better Subset Regression Using the Nonnegative Garrote. Technometrics **37**(4), 373–384 (1995). DOI 10.2307/1269730

[10] Burke, J.V., Curtis, F.E., Lewis, A.S., Overton, M.L., Simões, L.E.: Gradient sampling methods for nonsmooth optimization. arXiv preprint arXiv:1804.11003 (2018)

[11] Candes, E.J., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Transactions on Information Theory **52**(2), 489–509 (2006). DOI 10.1109/TIT.2005.862083

[12] Candes, E.J., Wakin, M.B., Boyd, S.P.: Enhancing sparsity by reweighted ℓ_1 minimization. Journal of Fourier analysis and applications **14**(5-6), 877–905 (2008)

[13] Carlsson, M.: On convexification/optimization of functionals including an l2-misfit term. arXiv:1609.09378 [math] (2016). ArXiv: 1609.09378

[14] Carlsson, M.: On convex envelopes and regularization of non-convex functionals without moving global minima. Journal of Optimization Theory and Applications **183**(1), 66–84 (2019)

[15] Chahid, M.: Echantillonnage compressif appliqué à la microscopie de fluorescence et à la microscopie de super résolution. Ph.D. thesis, Bordeaux (2014)

[16] Clarke, F.H.: Optimization and nonsmooth analysis, vol. 5. Siam (1990)

[17] Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward-backward splitting. Multiscale Modeling & Simulation **4**(4), 1168–1200 (2005)

[18] Gazagnes, S., Soubies, E., Blanc-Féraud, L.: High density molecule localization for super-resolution microscopy using CEL0 based sparse approximation. In: Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on, pp. 28–31. IEEE (2017)

[19] Hess, S.T., Girirajan, T.P.K., Mason, M.D.: Ultra-High Resolution Imaging by Fluorescence Photoactivation Localization Microscopy. Biophysical Journal **91**(11), 4258–4272 (2006). DOI 10.1529/biophysj.106.091116

[20] Larsson, V., Olsson, C.: Convex Low Rank Approximation. International Journal of Computer Vision **120**(2), 194–214 (2016). DOI 10.1007/s11263-016-0904-7

[21] Li, H., Lin, Z.: Accelerated proximal gradient methods for nonconvex programming. In: Advances in neural information processing systems, pp. 379–387 (2015)

[22] Lu, Z., Zhang, Y.: Sparse approximation via penalty decomposition methods. SIAM Journal on Optimization **23**(4), 2448–2478 (2013)

[23] Mallat, S.G., Zhang, Z.: Matching pursuits with time-frequency dictionaries. IEEE Transactions on Signal Processing **41**(12), 3397–3415 (1993). DOI 10.1109/78.258082

[24] Mordukhovich, B.S., Nam, N.M.: An easy path to convex analysis and applications. Synthesis Lectures on Mathematics and Statistics **6**(2), 1–218 (2013)

[25] Nikolova, M.: Relationship between the optimal solutions of least squares regularized with ℓ_0-norm and constrained by k-sparsity. Applied and Computational Harmonic Analysis **41**(1), 237–265 (2016). DOI 10.1016/j.acha.2015.10.010

[26] Pati, Y.C., Rezaiifar, R., Krishnaprasad, P.S.: Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In: Proceedings of 27th Asilomar Conference on Signals, Systems and Computers, pp. 40–44 vol.1 (1993). DOI 10.1109/ACSSC.1993.342465

[27] Peleg, D., Meir, R.: A Bilinear Formulation for Vector Sparsity Optimization. Signal Process. **88**(2), 375–389 (2008). DOI 10.1016/j.sigpro.2007.08.015

[28] Pilanci, M., Wainwright, M.J., El Ghaoui, L.: Sparse learning via boolean relaxations. Mathematical Programming **151**(1), 63–87 (2015)

[29] Rockafellar, R.T., Wets, R.J.B.: Variational analysis, vol. 317. Springer Science & Business Media (2009)

[30] Rust, M.J., Bates, M., Zhuang, X.: Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). Nature Methods **3**(10), 793–796 (2006). DOI 10.1038/nmeth929

[31] Sage, D., Kirshner, H., Pengo, T., Stuurman, N., Min, J., Manley, S., Unser, M.: Quantitative evaluation of software packages for single-molecule localization microscopy. Nature methods **12**(8), 717 (2015)

[32] Sage, D., Pham, T.A., Babcock, H., Lukes, T., Pengo, T., Chao, J., Velmurugan, R., Herbert, A., Agrawal, A., Colabrese, S., et al.: Super-resolution fight club: assessment of 2d and 3d single-molecule localization microscopy software. Nature methods **16**(5), 387–395 (2019)

[33] Selesnick, I.: Sparse regularization via convex analysis. IEEE Transactions on Signal Processing **65**(17), 4481–4494 (2017)

[34] Simon, B.: Trace ideals and their applications. 120. American Mathematical Soc. (2005)

[35] Soubies, E., Blanc-Féraud, L., Aubert, G.: A continuous exact ℓ_0 penalty (CEL0) for least squares regularized problem. SIAM Journal on Imaging Sciences **8**(3), 1607–1639 (2015)

[36] Soubies, E., Blanc-Féraud, L., Aubert, G.: A unified view of exact continuous penalties for \ell_2-\ell_0 minimization. SIAM Journal on Optimization **27**(3), 2034–2060 (2017)

[37] Soussen, C., Idier, J., Brie, D., Duan, J.: From bernoulli–gaussian deconvolution to sparse signal restoration. IEEE Transactions on Signal Processing **59**(10), 4572–4584 (2011)

[38] Tono, K., Takeda, A., Gotoh, J.y.: Efficient dc algorithm for constrained sparse optimization. arXiv preprint arXiv:1701.08498 (2017)