A continuous relaxation of the constrained $L\_2 - -L\_0$ problem

Arne Bechensteen, Laure Blanc-Féraud, Gilles Aubert

Journal of Mathematical Imaging and Vision manuscript No.
(will be inserted by the editor)

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

Arne Bechensteen · Laure Blanc-Féraud · Gilles Aubert

Received: date / Accepted: date

Abstract We focus on the minimization of the least square loss function under a k-sparse constraint with 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 from 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 relaxation, 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.

This work has been supported by the French government, through a financial PhD 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

Arne Bechensteen
Université Côte d'Azur, CNRS, Inria, Laboratoire I3S UMR 7271, 06903 Sophia Antipolis, France
E-mail: arne-henrik.bechensteen@inria.fr

Laure Blanc-Féraud
Université Côte d'Azur, CNRS, Inria, Laboratoire I3S UMR 7271, 06903 Sophia Antipolis, France,
E-mail: blancf@i3s.unice.fr

Gilles Aubert
Université Côte d'Azur, UNS, Laboratoire J. A. Dieudonné UMR 7351, 06100 Nice, France,
E-mail: 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-zeros entries. This type of problem appears in many applications, such as source separation, machine learning, and single-molecule localization microscopy, 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 [24] for 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 fix 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 [22], and the Orthogonal Matching Pursuit (OMP) [25]. Newer algorithms add and *subtract* components at each iteration, among them are the algorithm Greedy sparse simplex [3] or Single Best Replacement (SBR) [35].

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 [7] and Boolean relaxation [27] are two among the many algorithms based on this method. Algorithms of this method have been proposed to solve sparse problems (see [5,21], 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 [10]. 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 [8], the Log-Sum penalty [11] or Capped-ℓ_1 [26] to mention some. The continuous Exact ℓ_0 penalty introduced in [33] proposes an exact relaxation for problem (2) and a unified view of these functions is given in [34]. A recent convex relaxation has been proposed in [31], 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) [33], 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 to 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 [13], choosing the right parameters. However, in [13] the author uses an additional parameter that depends on the observation matrix A. In section 3, 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" 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.

Notations and Assumption

- The vector $x^{\downarrow} \in \mathbb{R}^N$ is the vector x sorted by its magnitude until the index k, i.e. $|x_1^{\downarrow}| \geq |x_2^{\downarrow}| \geq \dots |x_k^{\downarrow}|$, $|x_k^{\downarrow}| \geq |x_i^{\downarrow}|, \forall i > k$.
- The vector $x^{\downarrow y}$ is the vector that sorted such that $< y, x >=< y^{\downarrow}, x^{\downarrow y} >$.
- a_i is the i-th column of A. We suppose $\|a_i\| \neq 0 \, \forall i$.
- The indicator function $\iota_{x \in X}$ is defined for $X \subset \mathbb{R}^N$ as

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

- $\text{sign}^*(x)$ 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

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

where A_n is a matrix deduced from A where the norm of the columns are 1. □

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

2 Convex envelope

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, 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 [33]). The problem (1) can be written as

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

where ι is the indicator function defined in Notations. Before calculating the convex envelope, some preliminary results are needed.

2.1 Preliminary results

Proposition 2 ([36]) *$g(x) : \mathbb{R}^N \to \mathbb{R}$ defined as $g(x) = \frac{1}{2}\sum_{i=1}^k x_i^{\downarrow 2}$, where $x^\downarrow$ is defined as in Notations, is convex.*

Lemma 1 *Let $x \in \mathbb{R}^N$ a non-increasing and non-negative vector, with $x_i \geq 0$ and $x_1 \geq x_2 \geq \cdots \geq x_N$. Let us consider the concave problem*

$$\sup_{\substack{y_1\geq\cdots\geq y_k\geq 0 \\ y_k\geq y_i\ ,\forall i=k+1..N}} -\frac{1}{2}\sum_{i=1}^k y_i^2 + < y, x > . \tag{4}$$

Then there exists a function $T : \mathbb{R}^N \to \mathbb{N}$, $0 < T(x) \leq k$ defined as the smallest integer satisfying the double inequality

$$x_{k-T(x)+1} \leq \frac{1}{T(x)} \sum_{i=k-T(x)+1}^{N} x_i \leq x_{k-T(x)} \tag{5}$$

where the left inequality is strict if $T(x) \neq 1$, and where $x_0 = +\infty$ and such that problem (4) has the following optimal solution

$$y_j = \begin{cases} \frac{1}{T(x)}\sum_{i=k-T(x)+1}^{N} x_i & \text{if } k \geq j \geq k-T(x)+1 \\ & \text{or if } j > k \text{ and } x_j \neq 0 \\ \left[0, \frac{1}{T(x)}\sum_{i=k-T(x)+1}^{N} x_i\right] & \text{if } j > k \text{ and } x_j = 0 \\ x_j & \text{if } j < k-T(x)+1. \end{cases} \tag{6}$$

The value of the supremum problem is

$$\frac{1}{2}\sum_{i=1}^{k-T(x)} x_i^2 + \frac{1}{2T(x)}\left(\sum_{i=k-T(x)+1}^{N} |x_i|\right)^2. \tag{7}$$

Proof Let $x \in \mathbb{R}^N$ be a non negative non increasing vector. Problem (4) can be written as

$$\sup_{\substack{y_1 \geq \cdots \geq y_k \geq 0 \\ y_k \geq y_i \geq 0 \, \forall i = k+1..N}} \sum_{i=1}^{k} x_i y_i - \frac{1}{2}\sum_{i=1}^{k} y_i^2 + \sum_{i=k+1}^{N} x_i y_i. \tag{8}$$

We remark that finding the supremum for $y_i\,, i > k$ reduces to

$$\sup_{\substack{y_k \geq y_i \geq 0 \\ \forall i = k+1..N}} \sum_{i=k+1}^{N} x_i y_i. \tag{9}$$

We divide the proof into different scenarios.

1 a) $\exists\, j > k, x_j = 0$**.** When $x_j = 0$, for $j > k$, we observe that y_j is multiplied with zero, and can take on every value between 0 and its upper bound y_k. Thus $y_j \in [0, y_k]$ if $x_j = 0\, j > k$.

1 b) $x_k = 0$. Since x is non-increasing, this means that $x_i = 0$, $i > k$. So problem (8) can be written as

$$\sup_{y_1 \geq \cdots \geq y_{k-1} \geq 0} \sum_{i=1}^{k-1} x_i y_i - \frac{1}{2}\sum_{i=1}^{k-1} y_i^2 + \sup_{y_{k-1} \geq y_k \geq 0} -\frac{1}{2} y_k^2. \tag{10}$$

The supremum of y_k in (10) is $y_k = 0$. From **1 a)**, $y_j = [0, y_k]$, and thus $y_j = 0\, \forall j > k$. Furthermore, for y_j, $j < k$ the problems are strictly concave and thus the derivative can be used to find its supremum, which yields $y_j = x_j, \forall j < k$. Moreover, the solution y satisfies the inequality constraints since x is non-increasing.

1 c) $\exists\, j < k, x_j = 0$. Since x is sorted, this means also $x_i = 0$, $i \geq j$. Then the results form **1 a)** and **1 b)** are valid. In fact this does not change the results and we will have the same results as in **1 b)** with $y_i = 0\, \forall i \geq j$, and $y_i = x_i\, \forall i < j$.

2 a) $\forall j > k, x_j \neq 0$ **and** $\sum_{i=k}^{N} x_i \leq x_{k-1}$. The sum in (9) is non-negative and increasing with respect to y_j and the supremum is obtained when y_j reaches its upper bound, i.e $y_j = y_k$, $\forall j > k$ and $x_j \neq 0$ and we have

$$\sup_{\substack{y_k \geq y_i \geq 0 \\ \forall i = k+1..N}} \sum_{i=k+1}^{N} x_i y_i = y_k \sum_{i=k+1}^{N} x_i.$$

Thus in this case problem (8) is equivalent to

$$\sup_{y_1 \geq \cdots \geq y_{k-1} \geq 0} \sum_{i=1}^{k-1} x_i y_i - \frac{1}{2}\sum_{i=1}^{k-1} y_i^2 + \sup_{y_{k-1} \geq y_k \geq 0} y_k \sum_{i=k+1}^{N} x_i + x_k y_k - \frac{1}{2} y_k^2. \tag{11}$$

Since problem (11) is concave with respect to y_k the solution is given by the zeros of the gradient:

$$x_k - y_k + \sum_{i=k+1}^{N} x_i = 0 \Leftrightarrow y_k = \sum_{i=k}^{N} x_i,$$

and the value of the supremum in y_k is reached at $\frac{1}{2}\left(\sum_{i=k}^{N} x_i\right)^2$. Furthermore, for each y_j, $j < k$ we have the following expression

$$y_j = x_j.$$

We further assumed that $\sum_{i=k}^{N} x_i \leq x_{k-1}$ and thus the inequality constraints in (4) are verified, and the solution is of the form

$$y_j = \begin{cases} \sum_{i=k}^{N} x_i \text{ if } j \geq k \text{ and } x_j \neq 0 \\ [0, \sum_{i=k}^{N} x_i] \text{ if } j > k \text{ and } x_j = 0 \\ x_j \text{ if } j < k. \end{cases} \tag{12}$$

The second line of the solution comes from **1 a)**.

2 b) $\forall j > k, x_j \neq 0$ **and** $\sum_{i=k}^{N} x_i > x_{k-1}$. If $\sum_{i=k}^{N} x_i > x_{k-1}$, then y obtained in (12) does not verify $y_k < y_{k-1}$ and thus cannot be a solution. The optimal argument of y_k is reached at its constraint border, this yields $y_k = y_{k-1}$. Furthermore, $y_i = y_k, i > k$. Problem (11) can therefore be written as

$$\sup_{y_1 \geq \cdots \geq y_{k-2} \geq 0} \sum_{i=1}^{k-2} x_i y_i - \frac{1}{2}\sum_{i=1}^{k-2} y_i^2 + \sup_{y_{k-2} \geq y_{k-1} \geq 0} y_{k-1} \sum_{i=k-1}^{N} x_i - y_{k-1}^2. \tag{13}$$

We solve the above equation by differentiating with respect to $y_j, j = \{1, \ldots, k-1\}$ as the problem is concave in y_j. The optimal solution is given by

$$y_j = \begin{cases} \frac{1}{2}\sum_{i=k-1}^{N} x_i \text{ if } j \geq k-1 \\ [0, \frac{1}{2}\sum_{i=k-1}^{N} x_i] \text{ if } j > k \text{ and } x_j = 0 \\ x_j \text{ if } j < k-1. \end{cases} \tag{14}$$

The second line of (14) comes from **1 a)**. The constraints are satisfied if $x_{k-2} \geq \frac{1}{2}\sum_{i=k-1}^{N} x_i$. Furthermore, such a solution can be found if $x_{k-1} < \sum_{i=k}^{N} x_i$, otherwise (12) is the solution. What lower bound has $\frac{1}{2}\sum_{i=k-1}^{N} x_i$? Comparing it with x_{k-1} yields

$$\frac{1}{2}\sum_{i=k-1}^{N} x_i - x_{k-1} = \frac{1}{2}\left(\sum_{i=k}^{N} x_i - x_{k-1}\right) > 0.$$

This shows that the optimal arguments (14) satisfies the constraints if $x_{k-1} < \frac{1}{2}\sum_{i=k-1}^{N} x_i \leq x_{k-2}$. However, as in the last case, it is not certain that $x_{k-2} \geq \frac{1}{2}\sum_{i=k-1}^{N} x_i$. If the inequality is false, it yields that $y_k = y_{k-1} = y_{k-2}$, in addition to $y_k = y_i, i > k$. In this case problem (13) can be rewritten. By recursion, if there

exist $T(x) : \mathbb{R}^N \to \mathbb{N}$, $0 < T(x) \leq k$ such that $y_k = y_{k-1} = \cdots = y_{k-T(x)+1}$ then problem (4) can be written as :

$$\sup_{y_1 \geq \cdots \geq y_{k-(T(x)+1)+1} \geq 0} \sum_{i=1}^{k-(T(x)+1)+1} x_i y_i - \frac{1}{2} \sum_{i=1}^{k-(T(x)+1)+1} y_i^2$$
$$+ \sup_{y_{k-(T(x)+1)+1} \geq y_{k-T(x)+1} \geq 0} y_{k-T(x)+1} \sum_{i=k-T(x)+1}^{N} x_i - \frac{T(x)}{2} y_{k-T(x)+1}^2.$$

Then the above problem yields an optimal argument of the form

$$y_j = \begin{cases} \frac{1}{T(x)} \sum_{i=k-T(x)+1}^{N} x_i \text{ if } j \geq k - T(x) + 1 \text{ and } x_j \neq 0 \\ [0, \frac{1}{T(x)} \sum_{i=k-T(x)+1}^{N} x_i] \text{ if } j > k \text{ and } x_j = 0 \\ x_j \text{ if } j < k - T(x) + 1. \end{cases} \tag{15}$$

This is a solution only if

$$\sum_{i=k}^{N} x_i > x_{k-1}, \frac{1}{2} \sum_{i=k-1}^{N} x_i > x_{k-2}, \ldots, \frac{1}{T(x)-1} \sum_{i=k-(T(x)-1)-1}^{N} x_i > x_{k-T(x)+1} \tag{16}$$

and $\frac{1}{T(x)} \sum_{i=k-T(x)+1}^{N} x_i \leq x_{k-T(x)}$, since, for example, if the first inequality in (16) was not true, then the solution found in **2 a)** would verify the constraints and would be a solution. If the second inequality of (16) was not true, then (14) would be a solution. Thus, $T(x)$ is the smallest integer to satisfy $\frac{1}{T(x)} \sum_{i=k-T(x)+1}^{N} x_i \leq x_{k-T(x)}$. A lower bound can be found of $\frac{1}{T(x)} \sum_{i=k-T(x)}^{N} x_i$ by comparing it with $x_{k-T(x)-1}$.

$$\frac{1}{T(x)} \sum_{i=k-T(x)+1}^{N} x_i - x_{k-T(x)+1} =$$
$$\frac{1}{T(x)} \left(\sum_{i=k-(T(x)-1)+1}^{N} x_i - (T(x)-1) x_{k-T(x)+1} \right) > 0.$$

The inequality comes from the last inequality in (16). Thus we can write the optimal argument on a general form:

$$y_j = \begin{cases} \frac{1}{T(x)} \sum_{i=k-T(x)+1}^{N} x_i \text{ if } j \geq k - T(x) + 1 \text{ and } x_j \neq 0 \\ [0, \frac{1}{T(x)} \sum_{i=k-T(x)+1}^{N} x_i] \text{ if } j > k \text{ and } x_j = 0 \\ x_j \text{ if } j < k - T(x) + 1 \end{cases} \tag{17}$$

where $0 < T(x) \leq k$ so that

$$x_{k-T(x)+1} \leq \frac{1}{T(x)} \sum_{i=k-T(x)+1}^{N} x_i \leq x_{k-T(x)}$$

and the left inequality is strict if $T(x) \neq 1$. We denote $x_0 = +\infty$. The function $T(x)$ ensures that the solution verifies the constraints.

Note that if $x_k = 0$, then $T(x) = 1$ as $\frac{1}{1}\sum_{i=k}^{N} x_i = 0 \leq x_{k-1}$, and thus the solution in (17), with a slight modification to, can serve as a general solution of problem (4). □

Lemma 2 *Let $x \in \mathbb{R}^N$. The following concave supremum problem*

$$\sup_{\substack{|y_1^{\downarrow}| \geq \cdots \geq |y_k^{\downarrow}| \geq 0 \\ |y_k^{\downarrow}| \geq |y_i^{\downarrow}|\,, \forall i = k+1..N}} -\frac{1}{2}\sum_{i=1}^{k} y_i^{\downarrow 2} + < y^{\downarrow}, x^{\downarrow y} > \tag{18}$$

is equivalent to

$$\sup_{\substack{z_1^{\downarrow} \geq \cdots \geq z_k^{\downarrow} \geq 0 \\ z_k^{\downarrow} \geq z_i^{\downarrow}\,, \forall i = k+1..N}} -\frac{1}{2}\sum_{i=1}^{k} z_i^{\downarrow 2} + < z^{\downarrow}, |x^{\downarrow z}| > . \tag{19}$$

The arguments are such that $y_i^{\downarrow} = \operatorname{sign}^*(x_i^{\downarrow z}) z_i^{\downarrow}$.

Proof

$$\text{Problem (18)} = \sup_{\substack{|y_1^{\downarrow}| \geq \cdots \geq |y_k^{\downarrow}| \geq 0 \\ |y_k^{\downarrow}| \geq |y_i^{\downarrow}|\,, \forall i = k+1..N}} -\frac{1}{2}\sum_{i=1}^{k} y_i^{\downarrow 2} + \sum_{i=1}^{N} \operatorname{sign}(x_i^{\downarrow y}) y_i^{\downarrow} |x_i^{\downarrow y}|. \tag{20}$$

Note that $\sum_{i=1}^{k} (\operatorname{sign}(x_i^{\downarrow y}) y_i^{\downarrow})^2 = \sum_{i=1}^{k} y_i^{\downarrow 2}$. Let us define $z_i = \operatorname{sign}(x_i) y_i$, then problem (18) writes as

$$\sup_{\substack{|z_1^{\downarrow}| \geq \cdots \geq |z_k^{\downarrow}| \geq 0 \\ |z_k^{\downarrow}| \geq |z_i^{\downarrow}|\,, \forall i = k+1..N}} -\frac{1}{2}\sum_{i=1}^{k} z_i^{\downarrow 2} + \sum_{i=1}^{N} z_i^{\downarrow} |x_i^{\downarrow z}|. \tag{21}$$

Let's assume that there is a $\hat{z}_i^{\downarrow} < 0$, optimal argument for the ith component of (21).

If $i > k$, then $\hat{z}_i^{\downarrow}$ realises the supremum of

$$\sup_{|z_k^{\downarrow}| \geq |z_i^{\downarrow}|} z_i^{\downarrow} |x_i^{\downarrow z}|.$$

We have thus $\hat{z}_i^{\downarrow} |x_i^{\downarrow z}| \geq z_i^{\downarrow} |x_i^{\downarrow z}|, \forall |z_i| \leq |z_k|$. Let chose a $\tilde{z}_i = -\hat{z}_i^{\downarrow}$, then we have $\hat{z}_i^{\downarrow} |x_i^{\downarrow z}| \geq -\hat{z}_i^{\downarrow} |x_i^{\downarrow z}| \Leftrightarrow 2\hat{z}_i^{\downarrow} |x_i^{\downarrow z}| \geq 0$, which is impossible if $x_i^{\downarrow z} \neq 0$. If $x_i = 0$, then $|\hat{z}_i^{\downarrow}| \leq |\hat{z}_k^{\downarrow}|$, and could take negative values, however the value of supremum does not change as $z_i^{\downarrow}$ is only multiplied with 0, and thus we can suppose $z_i \geq 0, i > k$.

If $i \leq k$, then $\hat{z}_i^{\downarrow}$ is the supremum of

$$\sup_{|z_{i-1}^{\downarrow}| \geq |z_i^{\downarrow}| \geq |z_{i+1}^{\downarrow}|} -\frac{1}{2} z_i^{\downarrow 2} + |x_i^{\downarrow z}| z_i^{\downarrow}.$$

Assume that $\hat{z}_i^{\downarrow} < 0$ then

$$\hat{z}_i^{\downarrow 2} + |x_i^{\downarrow z}| \hat{z}_i^{\downarrow} \geq z^2 + |x_i^{\downarrow z}| z, \forall\, z, |z_{i-1}| \geq |z| \geq |z_{i+1}|.$$

In particular, we chose $z = -\hat{z}_i^{\downarrow}$, and we have

$$\hat{z}_i^{\downarrow 2} + |x_i^{\downarrow z}|\hat{z}_i^{\downarrow} \geq (-\hat{z}_i^{\downarrow})^2 - |x_i^{\downarrow z}|\hat{z}_i^{\downarrow} \Leftrightarrow |x_i^{\downarrow z}|\hat{z}_i^{\downarrow} \geq -|x_i^{\downarrow z}|\hat{z}_i^{\downarrow} \Leftrightarrow$$
$$2|x_i^{\downarrow z}|\hat{z}_i^{\downarrow} \geq 0$$

which is absurd if $|x_i^{\downarrow z}|$ is positive and $\hat{z}_i^{\downarrow}$ is negative. If $x_i^{\downarrow z} = 0$, then we have $\hat{z}_i^{\downarrow} = 0$ or $|\hat{z}_i^{\downarrow}| = |z_{i+1}|$, and thus we can assume $\hat{z}_i^{\downarrow}$ non-negative $\forall i \leq k$. In order to recover y_i we define $y_i^{\downarrow} = \text{sign}^*(x_i^{\downarrow z})z_i^{\downarrow}, \text{sign}^*(0) = \{-1, 1\}$. □

Corollary 1 *Let $x \in \mathbb{R}^N$. The problem (19) can be written as a sum of strictly concave problems and linear problems*

$$\sum_{i=1}^{k} \left[\sup_{z_{i-1}^{\downarrow} \geq z_i^{\downarrow} \geq z_{i+1}^{\downarrow}} -\frac{1}{2} z_i^{\downarrow 2} + |x_i^{\downarrow z}| z_i^{\downarrow} \right] + \sum_{i=k+1}^{N} \sup_{z_k^{\downarrow} \geq z_i^{\downarrow}} \left[|x_i^{\downarrow z}| z_i^{\downarrow} \right]. \tag{22}$$

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

$$\sup_{\substack{|y_1^{\downarrow}| \geq \cdots \geq |y_k^{\downarrow}| \geq 0 \\ |y_k^{\downarrow}| \geq |y_i^{\downarrow}|, \forall i = k+1..N}} -\frac{1}{2} \sum_{i=1}^{k} y_i^{\downarrow 2} + < y^{\downarrow}, x^{\downarrow y} > . \tag{23}$$

Let $T(x)$ as in Lemma 1. Then the value of the supremum problem (23) is

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

The supremum argument is given by

$$y_j^{\downarrow}(x) = \begin{cases} \text{sign}(x_j^{\downarrow}) \frac{1}{T(x)} \sum_{i=k-T(x)+1}^{N} |x_i^{\downarrow}| & \text{if } k \geq j \geq k - T(x) + 1 \\ & \text{or if } j > k \text{ and } x_j^{\downarrow} \neq 0 \\ [-1, 1] \frac{1}{T(x)} \sum_{i=k-T(x)+1}^{N} |x_i^{\downarrow}| & \text{if } j > k \text{ and } x_j^{\downarrow} = 0 \\ x_j^{\downarrow} & \text{if } j < k - T(x) + 1. \end{cases} \tag{24}$$

$y(x)$ can be reconstructed using the same permutation as $x^{\downarrow}$ uses to go to x.

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 the $x^{\downarrow y}$ since they work with matrices. This inspired us to this proof. From Lemma 2, we can rather study (19).
We construct two functions and we prove a relation between them. First, defining the space

$$\mathcal{D}(a) = \{b; b \text{ is sorted the same way as } a\}.$$

$z \in \mathcal{D}(a)$ means $a^{\downarrow z} = a^{\downarrow}$.

We define two functions $f_1 : \mathbb{R}^N_{\geq 0} \times \mathbb{R}^N \to \mathbb{R}$ and $f_2 : \mathbb{R}^N_{\geq 0} \times \mathbb{R}^N \to \mathbb{R}$.

$$f_1(z, a) = < z, |a| > -\frac{1}{2} \sum_{i=1}^{k} z_i^{\downarrow 2} \tag{25}$$

$$f_2(z, a) = < z^{\downarrow}, |a^{\downarrow}| > -\frac{1}{2} \sum_{i=1}^{k} z_i^{\downarrow 2}. \tag{26}$$

First, we can observe that

$$\sup_z f_1(z,a) \geq \sup_{z\in\mathcal{D}(a)} f_1(z,a).$$

We want to establish a relation between the supremum of f_1 and f_2. From [32, Lemma 1.8], we have that $\forall (z,a) \in \mathbb{R}^N_{\geq 0} \times \mathbb{R}^N_{\geq 0}$ we have the following inequality:

$$< z, a > \leq < z^{\downarrow}, a^{\downarrow} >$$

and thus

$$\sup_z f_1(z,a) \leq \sup_z f_2(z,a).$$

The $\sup_z f_2(z,a)$ is known from Lemma 1:

$$\sup_z f_2(z,a) = \frac{1}{2}\sum_{i=1}^{k-T(a)} a_i^{\downarrow 2} + \frac{1}{2T(a)}\left(\sum_{i=k-T(a)+1}^{N} |a_i^{\downarrow}|\right)^2. \tag{27}$$

We have further:

$$\sup_{z\in\mathcal{D}(a)} f_1(z,a) = \sup_z \sum_{i=1}^{N} z_i^{\downarrow}|a_i^{\downarrow}| - \frac{1}{2}\sum_{i=1}^{k} z_i^{\downarrow 2} = \sup_z f_2(z,a).$$

Since $\sup_z f_1(z,a)$ is upper and lower bounded by the same value, we have

$$\sup_z f_1(z,a) = \frac{1}{2}\sum_{i=1}^{k-T(a)} a_i^{\downarrow 2} + \frac{1}{2T(a)}\left(\sum_{i=k-T(a)+1}^{N} |a_i^{\downarrow}|\right)^2. \tag{28}$$

It remains to find all z for which f_1 reaches its supremum value. Let $|a_i| = |x_i|$. Problem (23) can be decomposed into a sum of strict concave problems and linear bounded problems as seen in Corollary 1 and we have

$$\sum_{i=1}^{k}\left[\sup_{z_{i-1}^{\downarrow}\geq z_i^{\downarrow}\geq z_{i+1}^{\downarrow}} -\frac{1}{2}z_i^{\downarrow 2} + |x_i^{\downarrow z}|z_i^{\downarrow}\right] + \sum_{i=k+1}^{N}\sup_{z_k^{\downarrow}\geq z_i^{\downarrow}}\left[|x_i^{\downarrow z}|z_i^{\downarrow}\right] =$$

$$\sum_{i=1}^{k-T(x)}\left[\sup_{z_{i-1}^{\downarrow}\geq z_i^{\downarrow}\geq z_{i+1}^{\downarrow}} -\frac{1}{2}z_i^{\downarrow 2} + |x_i^{\downarrow z}|z_i^{\downarrow}\right] + \tag{29}$$

$$\sum_{i=k-T(x)+1}^{k}\left[\sup_{z_{i-1}^{\downarrow}\geq z_i^{\downarrow}\geq z_{i+1}^{\downarrow}} -\frac{1}{2}z_i^{\downarrow 2} + |x_i^{\downarrow z}|z_i^{\downarrow}\right] + \sum_{i=k+1}^{N}\sup_{z_k^{\downarrow}\geq z_i^{\downarrow}}\left[|x_i^{\downarrow z}|z_i^{\downarrow}\right]. \tag{30}$$

We observe from the supremum value (28) the following equality

$$\frac{1}{2}\sum_{i=1}^{k-T(X)} x_i^{\downarrow 2} = \sum_{i=1}^{k-T(X)}\left[-\frac{1}{2}x_i^{\downarrow 2} + |x_i^{\downarrow}||x_i^{\downarrow}|\right]. \tag{31}$$

Letting z be such that $x^{\downarrow z} = x^{\downarrow}$ and replacing $z_j^{\downarrow}$ in (29) by $|x_j^{\downarrow}|$ for $j \leq k - T(x)$ yields (31), and is thus a solution of (29) for $z_i^{\downarrow}, i \leq k - T(x)$. Further, by

identification, we remark that we can replace $z_j^{\downarrow}$ in (30) by $\frac{1}{T(x)} \sum_{i=1}^{k-T(X)} |x_i^{\downarrow}|$ for $j > k - T(x)$ and obtain the second part of (28). We have thus identified optimal arguments of problem (22). Furthermore, the supremum problem (22) is strictly concave for $z_j^{\downarrow}$, $j < k$, as observed in Corollary 1, and thus the supremum argument is unique. For $z_j^{\downarrow}$, $j > k$, we observe the non-uniqueness in the case where $x_j^{\downarrow} = 0$, as $z_j^{\downarrow}$ can take the values $[0, \frac{1}{T(x)} \sum_{i=1}^{k-T(X)} |x_i^{\downarrow}|]$. Let $P \in \mathbb{R}^{N \times N}$ be such that $Px = x^{\downarrow}$, then z can be constructed from $z^{\downarrow}$ by letting $z = P^{-1} z^{\downarrow}$ and y_j is reconstructed from z_j by multiplying with the $\text{sign}^*(x_j)$. □

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

Following [33], we suppose that $A^T A$ is a diagonal matrix D^2 (which is equivalent to say that the columns of A is orthogonal). Note that $D = I$, the identity matrix since A has normalized columns. In this case the function G_k (3) can be rewritten as

$$G_k(x) = \iota_{\|\cdot\|_0 \leq k}(x) + \frac{1}{2} \|d - b\|_2^2 + \frac{1}{2} \|x - z\|_2^2 \tag{32}$$

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.
To calculate the convex envelope of a function $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ we use the Legendre-Fenchel transformation 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. We apply the Legendre transformation on the functional (32)

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

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

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

Writing differently the expression inside the supremum we get

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

To simplify the notations, we denote $w = z + y$.

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

The supremum is reached when $x_i = w_i^{\downarrow}$, $i \leq k$, 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 get 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 of

$$G_k^{**}(x) = \frac{1}{2}\|Ax-d\|_2^2 - \frac{1}{2}\|x\|^2 + \sup_{\substack{|w_1^{\downarrow}|\geq\cdots\geq|w_k^{\downarrow}| \\ |w_k^{\downarrow}|\geq|w_i^{\downarrow}|\,\forall i=k+1..N}} < x^{\downarrow w}, w^{\downarrow} > - \frac{1}{2}\sum_{i=1}^{k} w_i^{\downarrow 2}. \tag{33}$$

The above supremum problem can be solved by using Theorem 1 with $x^{\downarrow w} = x^{\downarrow y}$ and $w^{\downarrow} = y^{\downarrow}$. This yields

$$G_k^{**}(x) = \frac{1}{2}\|Ax-d\|_2^2 - \frac{1}{2}\sum_{i=k-T(x)+1}^{N} x_i^{\downarrow 2} + \frac{1}{2T(x)}\left(\sum_{i=k-T(x)+1}^{N}|x_i^{\downarrow}|\right)^2 \tag{34}$$

or

$$G_k^{**}(x) = \frac{1}{2}\|Ax-d\|_2^2 + Q(x) \tag{35}$$

where

$$Q(x) = -\frac{1}{2}\sum_{i=k-T(x)+1}^{N} x_i^{\downarrow 2} + \frac{1}{2T(x)}\left(\sum_{i=k-T(x)+1}^{N}|x_i^{\downarrow}|\right)^2 \tag{36}$$

and where $T(x)$ is defined as in Lemma 1. This expression of the convex envelope may be hard to grasp since the expression is on a non-closed form. To understand better $Q(x)$ and its properties we have the following lemmas.

Lemma 3 $Q(x) : \mathbb{R}^n \to [0, \infty[$.

Proof Let us show that $Q(x) \geq 0$, $\forall x$. We use equation (36) as starting point.

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

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

Lemma 4 *The function $Q(x)$ is continuous.*

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 [28, 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 $\mathbb{R}^N$. $dom(G_k)$ 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. □

Lemma 5 *Let $\|x\|_0 \leq k$. Then $T(x)$ in Lemma 1 is such that $T(x) = 1$. The inverse it not necessarily true.*

Proof From Lemma 1 we know that T(x) satisfies

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

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

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

$$\sum_{i=k}^{N} |x_i^{\downarrow}| = 6 \leq |x_{k-1}^{\downarrow}| = 6.$$

$T(x) = 1$ but the constraint is clearly not satisfied. □

Lemma 6 $Q(x) = 0$ ***if and only if*** $\|x\|_0 \leq k$.

Proof From Lemma 3, $Q(x) \geq 0$ and the inequality is strict if $T(x) > 1$. Thus, it suffices to investigate $T(x) = 1$. 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 $\forall j,\, j > k,\, x_j^{\downarrow} = 0$.

□

Corollary 2 *$\forall x,\ \|x\|_0 \leq k$ we have that $G_Q(x) = G_k(x)$. This comes from Lemma 6 since $Q(x) = 0$ and in the case of G_k, $\iota_{\|\cdot\|_0 \leq k}(x) = 0$, and thus the values are the same.*

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

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

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

$$G_Q(x) = \frac{1}{2}\|Ax - d\|^2 + Q(x) \tag{37}$$

where

$$Q(x) = -\frac{1}{2} \sum_{i=k-T(x)+1}^{N} x_i^{\downarrow 2} + \frac{1}{2T(x)} \left(\sum_{i=k-T(x)+1}^{N} |x_i^{\downarrow}| \right)^2 \tag{38}$$

with $T(x)$ is the function defined in Lemma 1, and is the smallest positive integer satisfying

$$|x_{k-T(x)+1}^{\downarrow}| \leq \frac{1}{T(x)} \sum_{i=k-T(x)+1}^{N} |x_i^{\downarrow}| \leq |x_{k-T(x)}^{\downarrow}| \tag{39}$$

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

Remark that $Q(x)$ can be written as

$$Q(x) = -\frac{1}{2} \sum_{i=1}^{N} x_i^2 + \sup_{w} -\frac{1}{2} \sum_{i=1}^{k} w_i^{\downarrow 2} + < w, x > . \tag{40}$$

Note that the properties of $Q(x)$ proved in Section 2.2 are valid for any A. The exactness of a relaxation signifies 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 is 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.

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 Corollary 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.

3.1 The subgradient, preliminaries

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 1 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 7 *$Q(x)$ 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's use the definition of $Q(x)$ from (40). 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 2 The generalized subgradient [15] 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 $x \in \mathbb{R}^N$, let $T(x)$ be as defined in Lemma 1, and let $y(x)$ be defined as in Theorem 1 on page 9. The subgradient of $G_Q(x)$ is*

$$\partial G_Q(x) = A^*(Ax - d) - x + co(y(x)) \tag{41}$$

where $co(y(x))$ is the convex envelope of the set of subgradients where the supremum is reached in Theorem 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 $-\frac{1}{2}\sum_{i=1}^N x_i^2$. From [15, 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.

$$\partial G_Q = \nabla[\frac{1}{2}\|A\cdot - d\|^2](x) - \nabla[\frac{1}{2}\sum_{i=1}^N \cdot_i^2](x) + \partial_c[\sup_w -\frac{1}{2}\sum_{i=1}^k {w_i^\downarrow}^2 + < w, \cdot >](x). \tag{42}$$

From [23, Theorem 2.93], the subgradient of the supremum is the convex envelop of the subgradient where the supremum is reached. Our problem is

$$\partial_c[\sup_y -\frac{1}{2}\sum_{i=1}^k {y_i^\downarrow}^2 + < y, \cdot >](x).$$

We define $g(y, x) = -\frac{1}{2}\sum_{i=1}^k {y_i^\downarrow}^2 + < y, x >$. The subgradient of g with respect to x is $\partial_c(g(y, \cdot))(x) = y$. It suffices therefore to find y that realises the supremum. Note that we can write the supremum problem as follows

$$\sup_{\substack{|y_1^\downarrow| \geq \cdots \geq |y_k^\downarrow| \geq 0 \\ |y_k^\downarrow| \geq |y_i^\downarrow|, \forall i = k+1..N}} -\frac{1}{2}\sum_{i=1}^k y_i^{\downarrow 2} + < y^\downarrow, x^{\downarrow y} > . \tag{43}$$

Theorem 1 identifies the optimal arguments $y(x)$. Inserting the result into the expression of subgradient in (42) we obtain

$$\partial G_Q(x) = A^*(Ax - d) - x + co(y(x))$$

where $co(y(x))$ denotes the convex envelop of the supremum set. □

A natural question that arises is what is $co(y(x))$? We have seen that $y(x)$ is not unique when $x_i^\downarrow = 0$ for $i > k$. Moreover the permutation to recover $y(x)$ from $y^\downarrow(x)$ is not unique thus $y(x)$ is not unique. Let us take an artificial example: Let $x \in \mathbb{R}^4$ be such that $|x_1| = |x_2|$, with $|x_4|$ the largest coordinate in absolute value and $|x_3|$ the smallest, and $x^\downarrow$ can be either written as $\tilde{x}^\downarrow = (x_4, x_1, x_2, x_3)^T$ or $\dot{x}^\downarrow = (x_4, x_2, x_1, x_3)^T$. Let $k = 3$, and suppose $T(x) = 1$. Then, with $\tilde{x}^\downarrow$ we will to obtain $\tilde{y}^\downarrow(x) = (x_4, x_1, \text{sign}(x_2)(|x_2| + |x_3|), \text{sign}(x_3)(|x_2| + |x_3|))^T$ and $\dot{x}^\downarrow$ we will to obtain $\dot{y}^\downarrow(x) = (x_4, x_2, \text{sign}(x_1)(|x_1| + |x_3|), \text{sign}(x_3)(|x_1| + |x_3|))^T$. We reconstruct $\tilde{y}$ and $\dot{y}$ by applying the inverse permutation that was applied to x in order to obtain $\tilde{x}^\downarrow$ and $\dot{x}^\downarrow$. $\tilde{y}(x) = (x_1,\ \text{sign}(x_2)(|x_2| + |x_3|), \text{sign}(x_3)(|x_2| + |x_3|), x_4)^T$ is not equal to $\dot{y}(x) = (\text{sign}(x_1)(|x_1| + |x_3|), x_2,\ \text{sign}(x_3)(|x_1| + |x_3|), x_4)^T$. This is an artificial example and in the following corollary we will show that the function $T(x)$ is such that $y(x)$ is always unique up to a permutation.

Corollary 3 *The subgradient of G_Q is non-unique only if $\exists x_i = 0$, i.e. $y(x)$ is always unique with respect to permutations, and $co(y(x)) = y(x)$. Thus*

$$\partial G_Q(x) = A^*(Ax - d) - x + y(x). \tag{44}$$

Proof In the artificial example above, we have seen that $y(x)$ is not unique if there exist multiple elements of x which are equal in their absolute value, and, depending on the permutations that yield $x^{\downarrow}$, would yield different $y(x)$. This would happen if x is such that for $|x_j| = |x_i|$, and for one permutation we would have $x_j = x^{\downarrow}_{k-T(x)+1}$ and $x_i = x^{\downarrow}_{k-T(x)}$, and for another permutation, we would have the opposite. This would lead to $y(x)$ being not unique in the sense of permutation. Let us investigate if the function $T(x)$ would allow this to happen. Lets call $x_i = z$ and $x_j = v$. To simplify the notation, let us set $T(x) = T > 1$ which then satisfies

$$\frac{1}{T}\sum_{i=k-T+1}^{N} |x_i^{\downarrow}| \leq |x^{\downarrow}_{k-T}|$$

where $x^{\downarrow}_{k-T} = z$ and $x^{\downarrow}_{k-T+1} = v$. We can rewrite the above inequality

$$\frac{1}{T}|v| + \frac{1}{T}\sum_{i=k-(T-1)+1}^{N} |x_i^{\downarrow}| \leq |z| \Leftrightarrow \frac{1}{T}\sum_{i=k-(T-1)+1}^{N} |x_i^{\downarrow}| \leq |z| - \frac{1}{T}|v|.$$

Since $|z| = |v|$ we have

$$\frac{1}{T-1}\sum_{i=k-(T-1)+1}^{N} |x_i^{\downarrow}| \leq |v|$$

and more precisely,

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

This is not possible since T is the smallest value that satisfies the inequality in (5), and here $T-1$ verifies the inequality.

If $T(x) = T = 1$, then

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

can only be true if $x^{\downarrow}_{k+1} = 0$ as $|x^{\downarrow}_{k-1}| = |x^{\downarrow}_{k}|$. And in that case $y(x)$ is unique as $y^{\downarrow}_k = \text{sign}(x^{\downarrow}_k)(\sum_{i=k}^{N} |x_i^{\downarrow}|) = x^{\downarrow}_k$, and thus $y(x) = x$.

Thus $T(x)$ will ensure that $y(x)$ is unique with respect to permutation, and the $co(y(x)) = y(x)$. □

3.1.1 Examples of the uniqueness of $y(x)$

Let $x \in \mathbb{R}^3$ be such that $|x_2| > |x_1| = |x_3|$. We choose $k = 2$, let x be such that $T(x)$ in (5) is $T(x) = 1$, note that $T(x)$ is not necessarily 1, but is assumed here. Then the permutation that obtains $x^{\downarrow}$ is not unique, and we note $\tilde{x}^{\downarrow}$ and $\dot{x}^{\downarrow}$ for the two different permuted vectors. $y^{\downarrow}(x)$, solution from Theorem 1, is thus either

$$\tilde{y}^{\downarrow} = \begin{pmatrix} x_2 \\ \text{sign}(x_1)(|x_1| + |x_3|) \\ \text{sign}(x_3)(|x_1| + |x_3|) \end{pmatrix} \quad \text{or} \quad \dot{y}^{\downarrow} = \begin{pmatrix} x_2 \\ \text{sign}(x_3)(|x_1| + |x_3|) \\ \text{sign}(x_1)(|x_1| + |x_3|). \end{pmatrix}.$$

In order to reconstruct $y(x)$ from $y^{\downarrow}(x)$, we perform the inverse permutation that defined $x^{\downarrow}$. In both cases we have that

$$\tilde{y} = \dot{y} = \begin{pmatrix} \text{sign}(x_1)(|x_1| + |x_3|) \\ x_2 \\ \text{sign}(x_3)(|x_1| + |x_3|) \end{pmatrix}.$$

We observe that $y(x)$ is unique.

3.1.2 A second example involving a zero coordinate

Let $x \in \mathbb{R}^3$ and let $|x_1| = |x_2|$ and $|x_3| = 0$. $k = 2$, $T(x)$ is defined in (5) and we have $T(x) = 1$ (since x verify the constraint). And $y^{\downarrow}$ can be these 2 vectors

$$\tilde{y}^{\downarrow} = \begin{pmatrix} x_1 \\ \text{sign}(x_2)(|x_2|) \\ [-|x_2|, |x_2|] \end{pmatrix} \quad \text{or} \quad \dot{y}^{\downarrow} = \begin{pmatrix} x_2 \\ \text{sign}(x_1)(|x_1|) \\ [-|x_1|, |x_1|]) \end{pmatrix}.$$

In order to reconstruct $y(x)$ from $y^{\downarrow}(x)$, we perform the inverse permutation that defined $x^{\downarrow}$, and we obtain

$$\tilde{y} = \begin{pmatrix} x_1 \\ x_2 \\ [-|x_1|, |x_1|] \end{pmatrix} \quad \text{or} \quad \dot{y} = \begin{pmatrix} x_1 \\ x_2 \\ [-|x_2|, |x_2|] \end{pmatrix}.$$

However, since $|x_1| = |x_2|$, $\tilde{y} = \dot{y}$, and thus y is unique.

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

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

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 $T(x) = 1$, 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:

$$A = \begin{pmatrix} 3 & 2 \\ 1 & 3 \end{pmatrix} \Lambda_{1/\|a_i\|} \quad \text{and } d = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \tag{45}$$

$$A = \begin{pmatrix} -3 & -2 \\ 1 & 3 \end{pmatrix} \Lambda_{1/\|a_i\|} \quad \text{and } d = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \tag{46}$$

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 i-th 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 (45), the relaxation is exact. This can be observed in the upper row in Fig. 1. Example (46) 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 $(0, 1.094)$.

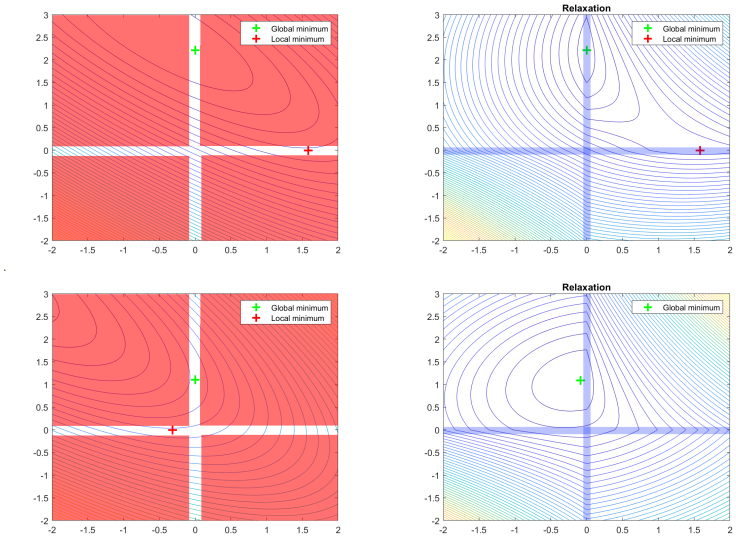


Fig. 1: Top: Level lines of the function G_k and G_Q for the example (45). Bottom: Level lines of the function G_k and G_Q for the example (46).

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 in comparison to the initial function G_k, G_Q is continuous. One could implement a subgradient method, either by using gradient

bundle methods (see [9] 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) [20] which is used in the numerical experiences of this paper. They are designed to work on problems on the form

$$\hat{x} \in \arg\min_x J(x) := f(x) + g(x) \tag{47}$$

where f is a differential 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 8 *G_Q satisfies the K-Ł property.*

Proof $\frac{1}{2}\|Ax - d\|^2$ is semi-algebraic. Using the definition of $Q(x)$ in (40) we can prove that $Q(x)$ is semi-algebraic. $\|x\|_2^2$ is semi-algebraic. Since

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

and $g(x,y)$ is semi-algebraic [6], 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-Łproperty. □

Lemma 9 *Let $g : \mathbb{R} \to \mathbb{R}$ be a strictly convex function, let $w = \arg\min_t g(t)$, and let us suppose that g is symmetric with respect to its minimum, i.e. $g(w-t) = g(w+t) \forall t \in \mathbb{R}$. The problem*

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

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 g is symmetric over its minimum $g(w+t_1) \leq g(w+t_2) \forall |t_1| \leq |t_2|$. Assume that $0 < w \leq b$. We can write $g(b) = g(w+\alpha)$, $\alpha > 0$ and $g(-b) = g(w+\beta), \beta < 0$. 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 10 *Let $g_i, i \in [1..N]$ be strictly convex and let $w = (w_1, w_2, \dots w_N)^T = \arg\min_{t_i} \sum g_i(t_i)$ and let us 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*

$$\underset{\substack{|t_1| \geq \cdots \geq |t_k| \\ |t_k| \geq |t_i| \, \forall i = k+1..N}}{\arg\min} \sum_i^N g_i(t_i). \tag{48}$$

The optimal solution is

$$t_i(\tau) = \begin{cases} \operatorname{sign}^*(w_i) \max(|w_i|, \tau) & \text{if } 1 \leq i \leq k \\ \operatorname{sign}^*(w_i) \min(|w_i|, \tau) & \text{if } i > k \end{cases} \tag{49}$$

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 [19, Theorem 2], with some modifications. First, if $|w_k| \geq |w_{k+1}|$, then w satisfies the constraints in Problem (48), 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 9 for t_i, and the choices can be limited to the following choices:

$$t_i = \begin{cases} w_i & \text{if } |t_{i-1}| \geq |w_i| \geq |t_{i+1}| \text{ and } i \leq k \\ \operatorname{sign}^*(w_i)|t_{i+1}| & \text{if } |w_i| < |t_{i+1}| \text{ and } i \leq k \\ \operatorname{sign}^*(w_i)|t_{i-1}| & \text{if } |w_i| > |t_{i-1}| \text{ and } i \leq k \\ w_i & \text{if } |t_k| > |w_i| \text{ and } i > k \\ \operatorname{sign}^*(w_i)|t_k| & \text{if } |t_k| < |w_i| \text{ and } i > k. \end{cases} \tag{50}$$

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

$$t_i = \operatorname{sign}(w_i)^* \max(|w_i|, |t_{i+1}|). \tag{51}$$

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 $\operatorname{sign}^*(w_1)|t_2|$ is optimal. Therefore $t_1 = \operatorname{sign}^*(w_1) \max(|w_1|, |t_2|)$. Assume that this is true for the i-th index.

$$t_{i+1} = \begin{cases} w_{i+1} & \text{if } |t_i| \geq |w_{i+1}| \geq |t_{i+2}| \text{ and } i+1 \leq k \\ \operatorname{sign}^*(w_{i+1})|t_{i+2}| & \text{if } |w_{i+1}| < |t_{i+2}| \text{ and } i+1 \leq k \\ \operatorname{sign}^*(w_{i+1})|t_i| & \text{if } |w_{i+1}| > |t_i| \text{ and } i+1 \leq k. \end{cases} \tag{52}$$

But $t_i = \operatorname{sign}^*(w_i) \max(|w_i|, |t_{i+1}|)$, which yields $|t_i| \geq |w_i| \geq |w_{i+1}|$ and thus the third case of (52) can be ignored.
When $i > k$ the following solution is evident

$$t_i = \operatorname{sign}^*(w_i) \min(|t_k|, |w_i|). \tag{53}$$

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, $|t_{i+1}| = \max(|w_{i+1}|, |t_{i+2}|) = |t_{i+2}|$. This argument can be repeated just till k so that

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

To facilitate the notations, $|t_k| = \tau$. The theorem is proved by inserting τ instead of $|t_{i+1}|$ and $|t_k|$ into equation (51) and (53). □

Remark 1 Note that if w, defined in Lemma (10) is such that $|w_k| \geq |w_{k+1}|$, then $t(\tau) = w$.

Theorem 4 *Let $y \in \mathbb{R}^N$. The proximal operator of $-(\frac{\rho-1}{\rho})\sum_{i=k+1}^{N}(x)^{\downarrow 2}$, denoted $prox_{-(\frac{\rho-1}{\rho})\sum_{i=k+1}^{N}(\cdot)^{\downarrow 2}}(y)$ is*

$$prox_{-(\frac{\rho-1}{\rho})\sum_{i=k+1}^{N}(\cdot)^{\downarrow 2}}(y)^{\downarrow} = \begin{cases} \operatorname{sign}(y_i^{\downarrow})\max\left(|y_i^{\downarrow}|, \tau\right) & \text{if } i \leq k \\ \operatorname{sign}(y_i^{\downarrow})\min(\tau, |\rho y_i^{\downarrow}|) & \text{if } i > k. \end{cases} \tag{54}$$

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

$$\tau = \frac{\rho\sum_{i\in n_1}|y_i^{\downarrow}| + \rho\sum_{i\in n_2}|y_i^{\downarrow}|}{\rho\#n_1 + \#n_2} \tag{55}$$

where n_1 and n_2 are two groups of indices that indicates when $y_i^{\downarrow} < \tau$ for an $i \leq k$ and $\tau \leq \rho|y_i^{\downarrow}|$ for an $i > k$ respectively. $\#n_1$ and $\#n_2$ are the sizes of $n1$ and $n2$.

The proof is similar to the proof of Theorem 1 and thus is omitted.

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 sorted elements defined in Theorem 4 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 expression of $Q(x)$ in (36) is not on a closed-form expression because of the function $T(x)$ and calculating the proximal operator directly from the expression is difficult. The following proposition facilitates the calculation of prox_Q. The proposition is inspired by [12, 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*

$$prox_{\frac{Q}{\rho}}(y) = \frac{\rho y - z}{\rho - 1}. \tag{56}$$

Theorem 5 *The proximal operator of Q for $\rho > 1$ is*

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

or, equivalently

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

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 as defined in Theorem 4.

Proof The result is direct by applying Proposition 3 and Theorem 4. □

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.

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 state-of-the-art sparse algorithms. Namely CoBic [2] and the Iterative Hard Thresholding (IHT) [16] which minimize functions on the constrained form (1), and the $CEL0$ [17] and the ℓ_1 relaxation, both relaxations of the penalized formulation (2). G_Q is minimized with the Nonmonotone Accelerated Proximal gradient algorithm (nmAPG) [20]. 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 normal optical microscopes. The method was first introduced in [18,4,29]. Fluorescent microscopy uses photoactivable 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 disk (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

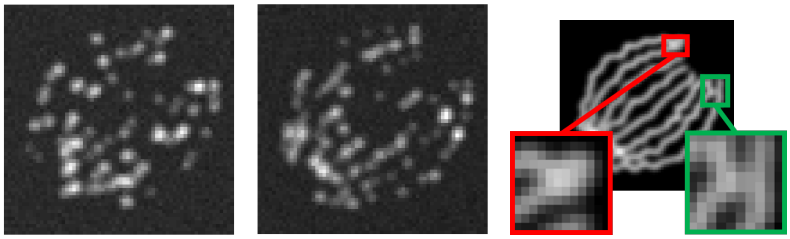


Fig. 2: Simulated images, from left to right: 1^{st} acquisition, 361^{th} acquisition, and the sum of all the acquisitions.

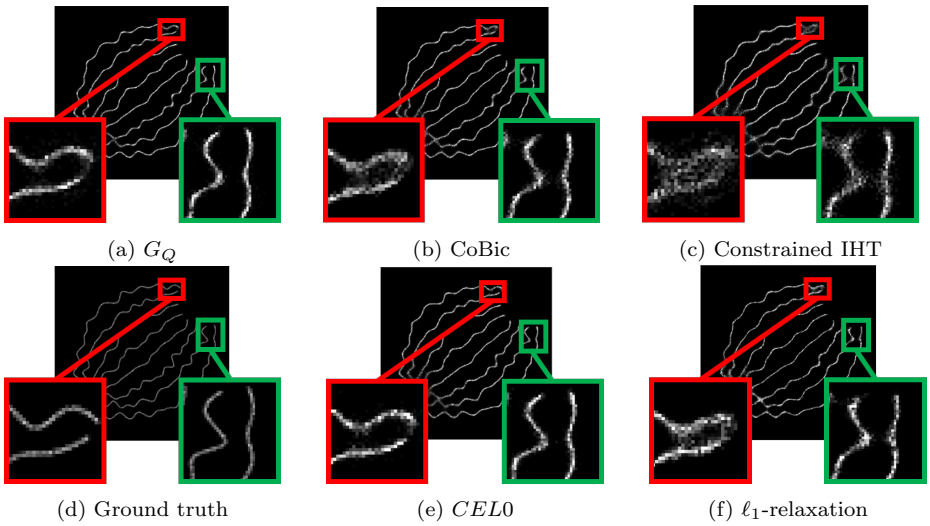
function to the non-negative space since the proximal operator of the sum of $Q(x)$ and the positivity constraint is not known. A is the matrix operator that performs a convolution with the Point Spread function and a reduction of dimensions. The fluorophores are reconstructed on a finer grid $\in \mathbb{R}^{ML\times ML}$ than the observed image $\in \mathbb{R}^{M\times M}$, with $L > 1$. A complete lecture of the mathematical model can be found in [2]. Note that an estimation 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.

The algorithms are tested on two datasets with high-density acquisitions, accessible from the ISBI 2013 challenge [30]. For a review of the SMLM and the different localization algorithms, see the ISBI-SMLM challenge [30]. Figure 2 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 [30]), 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.

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

5.1 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) is 258.21 nm. Among the 361 images are 81 049 fluorophores.



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

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

Fig. 3: 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.

The algorithms localize the fluorophores with higher precision on a 256×256 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. 3. The proposed relaxation performs better or equivalent to the state-of-the-art 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

5.2 Results of the real dataset

The algorithms are applied to the real high-density dataset, provided from the 2013 ISBI SMLM challenge [30]. 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 [14]. 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. 4, 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. An example of this trade-off is in the top right corner of Fig. 4 (f), where tubulins visible in the original data are barely visible in the reconstruction. Furthermore, increasing the penalty parameter λ deteriorates the resolution in the red and green parts. The results of the real dataset confirm the results of the simulated data.

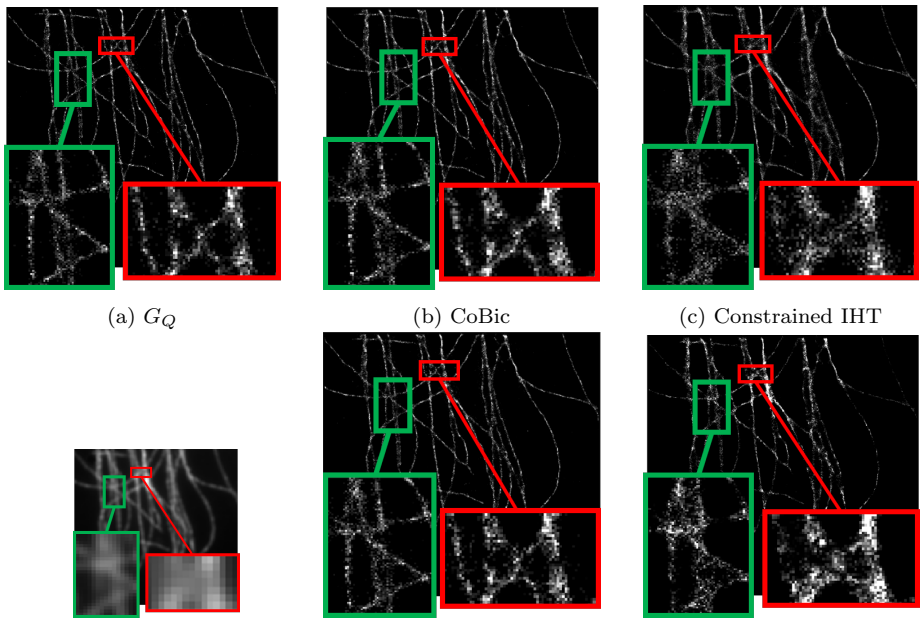
An important note In these numerical experiences, the proposed relaxed formulation converges *always* to a critical point that satisfies the sparsity constraint, and thus the "fail-safe" is never activated.

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 $CEL0$ [33]. 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" which ensures convergence to a critical point of the initial functional. In the case of SMLM, the relaxation performs as good as state-of-the-art, and it converges towards a critical point of the initial problem *each time* without the "fail-safe" 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.

Conflict of interest

The authors declare that they have no conflict of interest.



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

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

Fig. 4: 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.

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. 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
5. 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)
6. 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
7. 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)
8. Breiman, L.: Better Subset Regression Using the Nonnegative Garrote. Technometrics **37**(4), 373–384 (1995). DOI 10.2307/1269730
9. 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)
10. 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

11. 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)
12. Carlsson, M.: On convexification/optimization of functionals including an l2-misfit term. arXiv:1609.09378 [math] (2016). ArXiv: 1609.09378
13. 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)
14. Chahid, M.: Echantillonnage compressif appliqué à la microscopie de fluorescence et à la microscopie de super résolution. Ph.D. thesis, Bordeaux (2014)
15. Clarke, F.H.: Optimization and nonsmooth analysis, vol. 5. Siam (1990)
16. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward-backward splitting. Multiscale Modeling & Simulation **4**(4), 1168–1200 (2005)
17. 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)
18. 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
19. 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
20. Li, H., Lin, Z.: Accelerated proximal gradient methods for nonconvex programming. In: Advances in neural information processing systems, pp. 379–387 (2015)
21. Lu, Z., Zhang, Y.: Sparse approximation via penalty decomposition methods. SIAM Journal on Optimization **23**(4), 2448–2478 (2013)
22. 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
23. 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)
24. 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
25. 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
26. 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
27. Pilanci, M., Wainwright, M.J., El Ghaoui, L.: Sparse learning via boolean relaxations. Mathematical Programming **151**(1), 63–87 (2015)
28. Rockafellar, R.T., Wets, R.J.B.: Variational analysis, vol. 317. Springer Science & Business Media (2009)
29. 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
30. 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)
31. Selesnick, I.: Sparse regularization via convex analysis. IEEE Transactions on Signal Processing **65**(17), 4481–4494 (2017)
32. Simon, B.: Trace ideals and their applications. 120. American Mathematical Soc. (2005)
33. 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)
34. 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)
35. 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)
36. Tono, K., Takeda, A., Gotoh, J.y.: Efficient dc algorithm for constrained sparse optimization. arXiv preprint arXiv:1701.08498 (2017)