A SECOND-ORDER METHOD WITH ENRICHED HESSIAN INFORMATION FOR IMAGING COMPOSITE SPARSE OPTIMIZATION PROBLEMS

J.C. DE LOS REYES[‡] AND P. MERINO[‡]

Abstract. In this paper we propose a second–order method for solving *linear composite sparse optimization problems* consisting of minimizing the sum of a differentiable (possibly nonconvex function) and a nondifferentiable convex term. The composite nondifferentiable convex penalizer is given by 1–norm of a matrix multiplied with the coefficient vector. The algorithm that we propose for the case of the linear composite 1–norm problem relies on the three main ingredients that power the OESOM algorithm [6]: the minimum norm subgradient, a projection step and, in particular, the second–order information associated to the nondifferentiable term. By extending these devices, we obtain a full second–order method for solving composite sparse optimization problems which includes a wide range of applications. For instance, problems involving the minimization of a general class of *differential graph operators* can be solved with the proposed algorithm. We present several computational experiments to show the efficiency of our approach for different application examples.

1. Introduction

In areas like statistics, machine learning, optimal control and image processing among others, there are many minimization problems involving a cost function that forces to get some sparsity pattern in the solution. In data analysis, for instance, the 1–norm penalization ability to produce sparse structures is well exploited for variable selection [12, 13].

Also, the composite problem of minimizing the cost $f(x) + \beta\|Cx\|_1$, for some matrix C, is relevant in practice when sparsity is affected by a pattern matrix C. For example, in the *fused lasso* problem, C corresponds to the successive difference operator; then, term $\|Cx\|_1$ becomes the anisotropic total variation of x, and it has several applications in signal and image processing [15]. In addition, high order differential operators (e.g. graph Laplacian) are covered by C, e.g. in trend filtering over graphs [16] as well as the novel nonlocal differential operators in the spirit of [9] can be also considered. Moreover, for a sufficiently large β, this problem can also be interpreted as the exact penalization of the minimization of f subject to the constraint $Cx = 0$.

First order algorithms have been proposed for minimizing special cases of the objective function $f(x) + \beta\|Cx\|_1$. See for instance [3], where a modified monotone version of FISTA algorithm is proposed for image restoration problems. One the other hand,

1991 *Mathematics Subject Classification.* 49M15, 65K05, 90C53, 90C90.

Key words and phrases. Nonsmooth optimization, linear composite optimization, ℓ^1–norm.

[*]This research has been supported by project PIMI-17-01 granted by Escuela Politécnica Nacional, Quito–Ecuador.

despite of the convergence properties of second–order methods, they have not been exhaustively investigated in the context of nonsmooth optimization. One of the reasons for its lack of popularity is related to the high storage requirements and computation cost at every iteration. However, second–order methods can be practical and advantageous if combined with cost–reduction and parallelization techniques. One of the first second-order algorithms developed for solving composite problems was introduced in [8] for general composition of a nonsmooth and a smooth functions. In their approach, the cost function is approximated by smooth functions, then solve the surrogate smoothed model using a trust-region algorithm. The surrogate model is itself a nonsmooth composite problem which is solved by expressing the nonsmooth penalization as a polyhedral function which leads to a constrained quadratic optimization problem. However, this procedure, in the case of the ℓ_1-norm, needs a dense matrix of size $m \times 2^m$ to express $\|Cx\|_1$ as a polyhedral function using the columns of H, which might be prohibitive for large values of m.

More recently, a primal-dual approach second-order method was proposed in [7]. There, a new variable y which represents the nonsmooth composite term is introduced in order to cope with the composite penalization quantity Cx. Next, a quadratic penalization is added to the Lagrangian formulating an Augmented Lagrange Method, which introduces a second variable to deal with the composite term as a constraint. This constraint is penalized by introducing an additional dual variable at the cost of increasing the size of the problem. The variant in this approach formulated in [7] uses the proximal operator in order to represent the Lagrange function by means of its Moreau's envelope function. Then, a generalized second-order Hessian of the Lagrangian is introduced which by computing the Clarke subgradient of the associated proximal operator. This generalized Hessian provides second-order updates for the primal and the dual variables. However, its second–order system requires the computation of proximal operator of the nonsmooth penalizer, which doest not have a close form in the case of the term $\|Cx\|_1$. Also, the generalized Hessian provided is not symmetric, which is a potential numerical disadvantage.

In this paper we intend to devise a new algorithm inspired by the orthant–wise second-order algorithm from [6], which utilizes second–order from the regular part f and also from the nondifferentiable composite term $\|C \cdot\|_1$. As expected, the transformation of the variable by matrix C entails new numerical and theoretical challenges in which the sparsity term is no longer separable. Therefore, by extending several devices from [6] we are able to describe a full second–order algorithm to the composite case and derive the corresponding convergence analysis associated to this method by applying the ideas from [2] using the Łojasiewikcz condition.

We organize this paper by setting the problem in Section 2. Then, In Section 3 we describe the associated elements of the algorithm which leads to the numerical method. Section 4 is devoted to the convergence analysis and the derivation of the corresponding rate. Finally we present the numerical tests that shows how second–order information ins relevant for the numerical performance.

2. Problem formulation

Let $f : \mathbb{R}^m \to \mathbb{R}$ be a differentiable function and let be $\beta > 0$. We are interested in the numerical solution of the unconstrained optimization problem

$$(\mathbf{P}) \qquad \min_{x\in\mathbb{R}^m} \varphi(x) := f(x) + \beta\|Cx\|_1,$$

where $\|\cdot\|_1$ corresponds to the standard 1–norm in $\mathbb{R}^m$ and C is a real $n \times m$ matrix with rows c_i, for $i = 1, \dots, n$. We shall notice that modifying the matrix C, problem $(\mathbf{P})$ also covers the so-called *fused problem*

$$\min_{x\in\mathbb{R}^m} \varphi(x) := f(x) + \alpha\|x\|_1 + \beta\|Cx\|_1. \tag{1}$$

In order to obtain existence of solutions for problem $(\mathbf{P})$ the following conditions are assumed hereafter. The existence of solutions then follows from Weierstrass' theorem.

Assumption 1.

(i) f *is bounded from below.*
(ii) $f : \mathbb{R}^m \to \mathbb{R}$ *is continuously differentiable, with locally Lipschitz continuous gradient* ∇f.
(iii) $\varphi = f + \beta\|C\cdot\|_1$ *is coercive.*

2.1. **First order optimality conditions.** Let us denote by $\bar{x}$ the solution of $(\mathbf{P})$ and by $\partial\phi(x)$ the subdifferential of the function ϕ at x. Moreover, let us denote by g the convex nondifferentiable part of φ, that is $g(x) = \beta\|Cx\|_1$. By using the standard optimality condition

$$0 \in \nabla f(\bar{x}) + \partial g(\bar{x}), \tag{2}$$

we can derive first-order necessary optimality conditions for $(\mathbf{P})$. Indeed, from the relation (2) we establish the optimality conditions characterizing a a solution for problem $(\mathbf{P})$. By using computation rules from subdifferential calculus, we argue that if $\bar{x}$ is a solution for $(\mathbf{P})$, then there exists $\xi(\bar{x}) \in \mathbb{R}^n$ such that:

$$0 = \nabla f(\bar{x}) + \beta C^\top \xi(\bar{x}), \tag{3}$$

where the corresponding entries of $\xi(x) \in \partial\|\cdot\|_1(Cx)$ are given by

$$\xi(x)_i = \begin{cases} \{\text{sign}(\langle c_i, x\rangle)\}, & \text{if } \langle c_i, x\rangle \neq 0, \\ [-1, 1], & \text{if } \langle c_i, x\rangle = 0. \end{cases} \tag{4}$$

For a given x, let us define the index sets

$$\mathcal{P} = \{i : \langle c_i, x\rangle > 0\}, \quad \mathcal{N} = \{i : \langle c_i, x\rangle < 0\}, \quad \text{and } \mathcal{A} = \{i : \langle c_i, x\rangle = 0\}.$$

Therefore, if $\bar{\xi} = \xi(\bar{x})$, condition (3) is equivalent to the existence of $\bar{\xi}_i$, for $i \in \bar{\mathcal{A}}$

$$-\sum_{i\in\bar{\mathcal{A}}} \bar{\xi}_i c_i^\top = \frac{1}{\beta}\nabla f(\bar{x}) + \sum_{i\in\bar{\mathcal{P}}} c_i^\top - \sum_{i\in\bar{\mathcal{N}}} c_i^\top, \tag{5}$$

where $\bar{\mathcal{P}}$, $\bar{\mathcal{N}}$ and $\bar{\mathcal{A}}$ are the corresponding index sets associated to $\bar{x}$.

Remark 1. *Notice that the linear system* (5) *is of size $m \times p$, with $p \leq n$ being the cardinality of $\bar{\mathcal{A}}$. By denoting $C_{\mathcal{A}} \in \mathbb{R}^{m \times p}$ the matrix whose columns are formed by the transposed rows indexed in $\mathcal{A}$ and $\tilde{C}_{\bar{\mathcal{A}}}$ denoting the corresponding augmented matrix, i.e. the matrix with the extra column given by the right–hand side of* (5)*.*

In the following, we will assume the Rouché–Capelli theorem holds. That is, the system (5) *has at least one solution provided that rank$\{\tilde{C}_{\bar{\mathcal{A}}}\}$ = rank$\{C_{\bar{\mathcal{A}}}\}$.*

3. The second-order method

We start with the construction of a *descent direction*, for which, following [6], we consider a vector of the form $\nabla f(x) + \beta C^\top \xi(x)$ according to (4).

3.1. Computation of a descent direction.

In standard 1–norm penalized problems [6], the natural choice for the subgradient element is the one with the minimum 2-norm or, equivalently in the convex case, the steepest descent direction. Because of the particular structure of the 1–norm, the minimum norm subgradient is also known as *orthant direction.* In fact, it characterizes the orthant in which a descent direction has to be found.

However, in the case of composite optimization, the term $\|Cx\|_1$ is no longer separable. Therefore, there is not orthant–wise interpretation for the minimum norm subgradient, which is defined in general as:

$$\xi^*(x) \in \operatorname{argmin}\{\|\nabla f(x) + \beta C^\top \xi\|_2 : \xi \in \partial \|\cdot\|_1(Cx)\} \tag{6}$$

One of the drawbacks of using the minimum norm subgradient is that its computation requires the solution of an auxiliary quadratic optimization problem with box constraints. However, although an additional optimization subproblem is needed, it is not as expensive as it may appear at first sight. Indeed, since we already know that $\xi_i = \operatorname{sign}(\langle c_i, x\rangle)$, if $\langle c_i, x\rangle \neq 0$, we can exclude these components in the optimization problem (6). Let $p := |\mathcal{A}|$ and let us denote

$$\tilde{\nabla}\varphi(x) := \nabla f(x) + \beta \sum_{i \in \mathcal{P}} c_i^\top - \beta \sum_{i \in \mathcal{N}} c_i^\top .$$

Further, let $C_{\mathcal{A}}$ denote the matrix obtained by removing all rows c_i, with $i \in \mathcal{N} \cup \mathcal{P}$, from C. Hence, we may reformulate problem (6) as the following box–constrained quadratic optimization problem:

$$\min_{\tilde{\xi} \in [-1,1]^p} \frac{1}{2} \left\| \tilde{\nabla}\varphi(x) + \beta C_{\mathcal{A}}^\top \tilde{\xi} \right\|_2^2 \tag{MinSub}$$

Notice that this problem is of the same size as the active set cardinality at x. In many cases $C_{\mathcal{A}} C_{\mathcal{A}}^\top$ is nonsingular, thus problem (MinSub) has a unique solution. Moreover, the solution of (MinSub) is given by

$$\tilde{\xi} = \mathbb{P}_{[-1,1]^p}\{\tilde{\xi} - \beta C_{\mathcal{A}} \tilde{\nabla}\varphi(x) - \beta^2 C_{\mathcal{A}} C_{\mathcal{A}}^\top \tilde{\xi}\}, \tag{7}$$

where $\mathbb{P}_I$ denotes the projection on a set I. Formula (7) can not be computed as a closed–form solution. Indeed, its dual fits in a classical LASSO problem formulation.

3.2. **Second order information.** Weak second order information associated to the 1–norm was algorithmically introduced in [6] in order to compute generalized hessian based descent directions that incorporate components coming from both the smooth and nonsmooth terms. There, the regularization of the ℓ_1–norm by Huber smoothing allowed to obtain the targeted second order information using the second derivative of its regularization. This procedure is analogous to consider generalized Hessians in the Bouligand subdifferential of the proximal operator: $\partial_B \operatorname{prox}_{\frac{1}{\gamma}\|\cdot\|_1}$.

Here, we generalize this procedure to the case of composite sparse optimization. In the present case, however, the weak second order derivative of the nondifferentiable term is not longer a diagonal matrix. Indeed, recalling that the Huber regularization of the 1–norm, for $\gamma > 0$ is defined by

$$h_\gamma(x_i) = \begin{cases} \gamma\frac{x_i^2}{2} & \text{if } |x_i| \leq \frac{1}{\gamma}, \\ |x_i| - \frac{1}{2\gamma} & \text{if } |x_i| > \frac{1}{\gamma}, \end{cases} \tag{8}$$

we now regularize $\|C\cdot\|_1$ as follows:

$$h_\gamma(Cx) = \begin{cases} \frac{\gamma}{2}\langle c_i, x\rangle^2 & \text{if } |\langle c_i, x\rangle| \leq \frac{1}{\gamma}, \\ |\langle c_i, x\rangle| - \frac{1}{2\gamma} & \text{if } |\langle c_i, x\rangle| > \frac{1}{\gamma}. \end{cases}$$

Then, $\nabla h_\gamma(Cu)$ is given by

$$\nabla h_\gamma(Cx) = C^\top \left[\frac{\langle c_i, x\rangle}{\max\{1/\gamma, \langle c_i, x\rangle\}}\right]_{i=1}^m, \tag{9}$$

and the "weak Hessian" of $\|C\cdot\|_{\ell_1}$ is given by the matrix

$$\Gamma = \gamma C^\top DC, \quad \text{with } D = \operatorname{diag}\left(\left[\begin{cases} 1 & \text{if } |\langle c_i, x\rangle| \leq \frac{1}{\gamma} \\ 0 & \text{otherwise} \end{cases}\right]_{i=1}^{i=n}\right) \tag{10}$$

One could realize that $D \in \partial_B(\operatorname{prox}_{\frac{1}{\gamma}\|\cdot\|_1})$.

We will write Γ^k to specify that (10) is computed for $x = x^k$. Now, the computation of the descent direction is carried on with the help of matrix (10), requiring the solution of the following linear system:

$$\left[B^k + \beta\Gamma^k\right] d^k = -[\nabla f(x^k) + \beta C^\top \xi(x^k)], \tag{11}$$

where B^k stands either for the Hessian of f at x^k or an approximation of it (e.g., the BFGS matrix).

Assumption 2. *The matrix B^k is symmetric positive definite and satisfies*

$$\hat{c}\|d\|_2^2 \leq d^\top B^k d \leq \hat{C}\|d\|_2^2, \qquad \textit{for all } d \in \mathbb{R}^m. \tag{12}$$

3.3. **Projection step.** In our algorithm, at each iteration, the approximated solution x may be close to fulfill sparsity in the range of C, i.e., $\langle c_i, x\rangle \approx 0$ for some of the indexes i. Thereafter, small perturbations on x may cause undesired sign changing in $\langle c_i, x\rangle$. When, under small perturbations on x, we detect a change in the sign of the quantity $\langle c_i, x\rangle$ we might keep the updated approximated solution satisfying sparsity condition. Therefore, we project the perturbation of x to the closest point $\tilde{x}$ satisfying $\langle c_i, \tilde{x}\rangle = 0$.

Thus, for a given approximated solution x and a descent direction y, we identify those $\langle c_i, x\rangle$ which change sign with respect to the subgradient $\xi(x)$ (recall that the subgradient $\xi(x)$ has the same sign of $\langle c_i, x\rangle$, when it is not 0).

For the sign identification process we introduce the set

$$\mathcal{S}(y) = \{i = 1, \ldots, n : \text{sign}(\langle c_i, y\rangle) \neq \text{sign}(\xi_i(x))\}, \tag{13}$$

and define $C_s := C(\mathcal{S}(y), :)$. Then, we consider the projection over the set $\mathcal{A}_S$, defined by

$$\mathcal{A}_S = \{y \in \mathbb{R}^m : \langle c_i, y\rangle = 0, \text{for } i \in \mathcal{S}(y)\} \tag{14}$$

Thus, the projection $\mathcal{P}$ on the set $\mathcal{A}_S$ is obtained as the solution of the following problem:

$$\min_{\tilde{x} \in \mathcal{A}_S} \frac{1}{2}\|\tilde{x} - x\|_2^2 \Leftrightarrow \min_{C_s\,\tilde{x}=0} \frac{1}{2}\|\tilde{x} - x\|_2^2 \tag{Prj}$$

It is known that (Prj) is a saddle point problem. A particular but important case is when C_s has full rank. Then, (Prj) is equivalent to the linear equation (see [4])

$$\begin{bmatrix} I & C_s^\top \\ C_s & O \end{bmatrix} \begin{bmatrix} \tilde{x} \\ y \end{bmatrix} = \begin{bmatrix} x \\ 0 \end{bmatrix}. \tag{15}$$

Furthermore, by introducing the projections $\Pi := C_s^\top (C_s C_s^\top)^{-1} C_s$ and $\mathcal{P} = I - \Pi$, we can solve (15) explicitly and the solution of (Prj) reads:

$$\tilde{x} = \mathcal{P}\,x = x - \Pi x \quad \in \text{span}\{c_i : i \in \mathcal{S}\}^\perp, \tag{16a}$$

$$y = (C_s C_s^\top)^{-1} C_s\, x. \tag{16b}$$

Note that, Πx is characterized as the solution of

$$\min_{z \in \text{range} C_s^\top} \|x - z\|^2. \tag{17}$$

Moreover, feasibility of $\tilde{x}$ implies that $C_{\mathcal{S}}\tilde{x} = 0$. From these relations, we realize that $x = \tilde{x} + \Pi x$, that is, $x \in \text{span}\{c_i : i \in \mathcal{S}\}^\perp \oplus \text{span}\{c_i : i \in \mathcal{S}\}$. In other words, the projection step removes the part belonging to range(C_s) from the current approximation.

In the case that C_S is not full rank, it cannot be guaranteed the existence of $(C_s C_s^\top)^{-1}$. Then, the common practice is to consider instead a regularization $C_s C_s^\top + \epsilon I$ for small $\epsilon > 0$.

3.4. **Linesearch step.** In analogous fashion to [1, 5], we consider the projected linesearch rule using $\mathcal{P}$ given by (16a), for choosing the step s_k fulfilling the decreasing condition:

$$\varphi[\mathcal{P}(x^k + s_k d^k)] \leq \varphi(x^k) + \widetilde{\nabla}\varphi(x^k)^T[\mathcal{P}(x^k + s_k d^k) - x^k]. \tag{18}$$

The calculation of the step s_k fulfilling the last condition is performed using a backtracking scheme.

Algorithm 1 Generalized Second–Order Method for Sparse Composite Optimization (GSOM)

```
Initialize $x^0$.
while stoping criteria is false do
    Compute $\xi^k$ given by solving (MinSub)
    Compute $d^k$ by solving system (11)
    Compute $s_k$ using a line–search procedure
    Update $x^{k+1} \leftarrow \mathcal{P}(x^k + s_k d^k)$
    $k \leftarrow k + 1$.
end while
```

4. Convergence Analysis

Let x^k be the approximated solution computed by Algorithm (1) in the k-th iteration. Moreover, let $C_k := C_{\mathcal{S}^k}$, for $k = 1, 2, \ldots$, and $\xi^k := \xi(x^k)$. Hence, at every step $\Pi = C_k(C_k C_k^\top)^{-1} C_k$. In addition, for a vector $y \in \mathbb{R}^m$, according to (13), we consider the index set

$$\mathcal{S}_k = \{i = 1, \ldots, n : \operatorname{sign}\langle c_i, x^k + sd^k\rangle \neq \operatorname{sign}(\xi_i^k)\}. \tag{19}$$

Remark 2. *It follows from the definition of $\mathcal{S}_k$ that for s sufficiently small x^k belongs to the null space of C_k and the index set $\mathcal{S}_k$ may be equivalently defined as*

$$\mathcal{S}_k = \{i = 1, \ldots, n : \operatorname{sign}(\xi_i) \operatorname{sign}\langle c_i, d^k\rangle \leq 0\}.$$

Indeed, this can be seen from the fact that if $i \in \mathcal{S}_k$ then we have that if $\langle c_i, x^k\rangle \neq 0$ then $\xi_i^k = \operatorname{sign}\langle c_i, x^k\rangle$ and, for sufficiently small s, we have $\operatorname{sign}\langle c_i, x^k + sd^k\rangle = \operatorname{sign}\langle c_i, x^k\rangle \neq \xi_i^k$, which is a contradiction. Therefore, the only possibility is that $\langle c_i, x^k\rangle = 0$. Thus, $\operatorname{sign}\langle c_i, d^k\rangle \operatorname{sign}(\xi_i^k) \leq 0$.

Theorem 1. *Let Assumptions 1 and 2 hold, and let x^k be the approximated solution for* (**P**) *at the kth iteration of Algorithm 1 and let d^k be the corresponding direction computed using* (11)*. Let us assume that C_k defined in projection step* (Prj) *is full rank. Moreover, let us assume that at every step $\langle c_i, d^k\rangle \neq 0$ for some i, and that the parameter $\gamma = \gamma_{k+1}$ is chosen in each iteration such that*

$$\gamma_{k+1} > \frac{1}{2\beta}\left(\frac{\| |\nu^k| + \beta(|\xi^k| + n|\eta^k|)\|^2}{\min \langle c_i, d^k\rangle^2} + 1\right), \tag{20}$$

where the minimum is taken from those $\langle c_i, d^k\rangle \neq 0$, where ν^k and η^k being the vectors of coefficients of $\Pi\nabla f(x^k)$ and Πc_{i^*} on span$\{c_i : i \in \mathcal{S}^k\}$, respectively. Here i^* is such that $|\langle c_{i^*}, \Pi d^k\rangle)| = \max_{i\in\mathcal{S}_k} |\langle c_i, \Pi d^k\rangle)|$. Then, d^k is a descent direction, i.e.:

$$\varphi(x^{k+1}) < \varphi(x^k). \tag{21}$$

Proof. Taking into account that $x^{k+1} = P(x^k + sd^k) = x^k + sd^k - \Pi(x^k + sd^k)$ and C_k is full rank then, by (15), it follows that $C_k x^{k+1} = 0$. That is, $\langle c_i, x^{k+1}\rangle = 0$ for all $i \in \mathcal{S}_k$. Moreover, if $i \in \mathcal{S}_k$ we have either $\langle c_i, x^k\rangle = 0$ or $\langle c_i, x^k\rangle \neq 0$. In the first case, it is clear that $0 = |\langle c_i, x^k\rangle| \leq s|\langle c_i, d^k\rangle|$. On the other hand, if $\langle c_i, x^k\rangle \neq 0$, we have that $\operatorname{sign}(\langle c_i, x^k + sd^k\rangle) \neq \operatorname{sign}(\xi^k) = \operatorname{sign}(\langle c_i, x^k\rangle)$. Then, we conclude that $|\langle c_i, x^k\rangle| < s|\langle c_i, d^k\rangle|$. Hence,

$$\|C_k x^k\| = \|[\langle c_i, x^k\rangle]_{i\in\mathcal{S}_k}\| \leq s\|C_k\|\|d^k\|,$$

which implies that $\|\Pi x^k\| \leq s\|C_k\|^2\|(C_kC_k^\top)^{-1}\|\|d^k\| \leq sc\|d^k\|$, for some constant c depending on the matrix C and independent of k. Therefore, we obtain the estimate

$$\begin{aligned}\|x^{k+1} - x^k\| &= \|s\mathcal{P}d^k - \Pi x^k\| \\ &\leq s\|\mathcal{P}d^k\| + \|\Pi x^k\| \\ &\leq s(1+c)\|d^k\|. \end{aligned} \tag{22}$$

Now, using (22) and the first order Taylor expansion of the regular part of φ, we get

$$\begin{aligned}\varphi(x^{k+1}) - \varphi(x^k) &= f(x^{k+1}) - f(x^k) + \beta\|Cx^{k+1}\|_1 - \beta\|Cx^k\|_1 \\ &= \nabla f(x^k)^\top \left(\mathcal{P}(x^k + sd^k) - x^k\right) + o(s\|d^k\|) \\ &\quad + \beta \sum_i \left(|\langle c_i, x^{k+1}\rangle| - |\langle c_i, x^k\rangle|\right). \end{aligned} \tag{23}$$

From the second–order system (11) and the positive semidefiniteness of Π, we see that $x^{k+1} - x^k = \mathcal{P}(x^k + sd^k) - x^k = sd^k - \Pi(x^k + sd^k)$, therefore

$$\begin{aligned}\nabla f(x^k)^\top(\mathcal{P}(x^k + sd^k) - x^k) =& s\nabla f(x^k)^\top d^k - s\nabla f(x^k)^\top\Pi d^k - \nabla f(x^k)^\top \Pi x^k \\ =& -s{d^k}^\top\left[B^k + \beta\Gamma^k\right]d^k - s\beta{\xi^k}^\top Cd^k - \nabla f(x^k)^\top\Pi(x^k + sd^k). \end{aligned} \tag{24}$$

Note that $\Pi = \Pi^2$; moreover, it is also a symmetric positive semi–definite matrix. In addition, we have that $\Gamma^k = \gamma C^\top D^k C$ is symmetric and positive semidefinite by its construction. Further, by Assumption 2 we have that exists a positive constant $\hat{c}$, independent of k, such that ${d^k}^\top B^k d^k \geq \hat{c}\|d^k\|^2$. Therefore, these matrix properties imply

$$\begin{aligned}\nabla f(x^k)^\top(P(x^k + sd^k) - x^k) \leq& - {d^k}^\top B^k d^k - s\gamma\beta(Cd^k)^\top D^k(Cd^k) - s\beta{\xi^k}^\top Cd^k \\ & - \nabla f(x^k)^\top\Pi(x^k + sd^k) \\ \leq& - s\hat{c}\|d^k\|^2 - \gamma s\beta \sum_{i:|\langle c_i, x^k\rangle|\leq 1/\gamma} \langle c_i, d^k\rangle^2 - s\beta\sum_{i\in\mathcal{S}_k}\xi_i^k\langle c_i, d^k\rangle \\ & - s\beta\sum_{i\notin\mathcal{S}_k}\xi_i^k\langle c_i, d^k\rangle - \nabla f(x^k)^\top\Pi(x^k + sd^k). \end{aligned} \tag{25}$$

Let us focus on the sum on the right–hand side of (23). Since for all $i \in \mathcal{S}_k = \{i \in \{1,\dots,n\} : \text{sign}\langle c_i, x^k + sd^k\rangle \neq \text{sign}(\xi_i^k)\}$, we have $\langle c_i, x^{k+1}\rangle = 0$; then:

$$\begin{aligned}
\sum_i (|\langle c_i, x^{k+1}\rangle| - |\langle c_i, x^k\rangle|) &= \sum_{i\notin\mathcal{S}_k} (|\langle c_i, x^{k+1}\rangle| - |\langle c_i, x^k\rangle|) - \sum_{i\in\mathcal{S}_k} |\langle c_i, x^k\rangle| \\
&= \sum_{i\notin\mathcal{S}_k} (|\langle c_i, \mathcal{P}(x^k + sd^k)\rangle| - |\langle c_i, x^k\rangle|) - \sum_{i\in\mathcal{S}_k} |\langle c_i, x^k\rangle| \\
&\leq \sum_{i\notin\mathcal{S}_k} |\langle c_i, x^k + sd^k\rangle| + |\langle c_i, \Pi(x^k + sd^k)\rangle| - |\langle c_i, x^k\rangle| \\
&\leq \sum_{i\notin\mathcal{S}_k} \xi_i^k \langle c_i, x^k + sd^k\rangle + |\langle c_i, \Pi(x^k + sd^k)\rangle| - |\langle c_i, x^k\rangle| \\
&\leq \sum_{i\notin\mathcal{S}_k} \xi_i^k \langle c_i, sd^k\rangle + |\langle c_i, \Pi(x^k + sd^k)\rangle|
\end{aligned}$$

Using Remark 2, it follows that $C_k x^k = 0$ if s is small enough, hence

$$\sum_i (|\langle c_i, x^{k+1}\rangle| - |\langle c_i, x^k\rangle|) \leq \sum_{i\notin\mathcal{S}_k} \xi_i^k \langle c_i, sd^k\rangle + s|\langle c_i, \Pi d^k)\rangle|. \tag{26}$$

Inserting (25) and (26) in (23) obtain the relation:

$$\begin{aligned}
\varphi(x^{k+1}) - \varphi(x^k) \leq &- s\hat{c}\|d^k\|^2 - \gamma s\beta \sum_{i:|\langle c_i, x^k\rangle|\leq 1/\gamma} \langle c_i, d^k\rangle^2 - s\beta \sum_{i\in\mathcal{S}_k} \xi_i^k \langle c_i, d^k\rangle \\
&+ s\beta \sum_{i\notin\mathcal{S}_k} |\langle c_i, \Pi d^k)\rangle| - \nabla f(x^k)^\top \Pi(x^k + sd^k) + o(s\|d^k\|) \\
\leq &- s\hat{c}\|d^k\|^2 - \gamma s\beta \sum_{i:|\langle c_i, x^k\rangle|\leq 1/\gamma} \langle c_i, d^k\rangle^2 + o(s\|d^k\|) \\
&- s\beta \sum_{i\in\mathcal{S}_k} \xi_i^k \langle c_i, d^k\rangle + s\beta|\mathcal{S}_k^C||\langle c_{i^*}, \Pi d^k)\rangle| - \nabla f(x^k)^\top \Pi(x^k + sd^k),
\end{aligned} \tag{27}$$

where $|\mathcal{S}_k^C|$ denotes the cardinality of the complement of the set $\mathcal{S}_k$ and i^* is the index where the term $|\langle c_i, \Pi d^k)\rangle|$ attains it maximum in $\mathcal{S}_k$.

By using again Remark 2, and taking into account that Π projects onto $\operatorname{span}\{c_i : i \in \mathcal{S}_k\}$, we can be estimate the last three terms as follows:

$$
\begin{aligned}
&\Big|\nabla f(x^k)^\top \Pi(x^k + sd^k) + s\beta \sum_{i\in\mathcal{S}^k} \xi_i^k \langle c_i, d^k\rangle - s\beta|\mathcal{S}_k^C||\langle c_{i^*}, \Pi d^k))\rangle|\Big| \\
&\qquad = \Big|[\Pi\nabla f(x^k)]^\top (x^k + sd^k) + s\beta \sum_{\substack{i\in\mathcal{S}_k \\ \langle c_i, x^k\rangle = 0}} \xi_i^k \langle c_i, d^k\rangle - s\beta|\mathcal{S}_k^C||\langle \Pi c_{i^*}, d^k))\rangle|\Big| \\
&\qquad = |\sum_{i\in\mathcal{S}_k} \nu_i^k \langle c_i, x^k + sd^k\rangle - s\beta \sum_{\substack{i\in\mathcal{S}_k \\ \langle c_i, x^k\rangle = 0}} |\xi_i^k|\,|\langle c_i, d^k\rangle| - s\beta|\mathcal{S}_k^C||\langle \sum_{i\in\mathcal{S}_k} \eta_i^k c_i, d^k))\rangle|| \\
&\qquad \le s \sum_{\substack{i\in\mathcal{S}_k \\ \langle c_i, x^k\rangle = 0}} (|\nu_i^k| + \beta(|\xi_i^k| + |\mathcal{S}_k^C||\eta_i^k|))\,|\langle c_i, d^k\rangle| \\
&\qquad \le s\Big(\sum_{\substack{i\in\mathcal{S}_k \\ \langle c_i, x^k\rangle = 0}} \frac{1}{2}(|\nu_i^k| + \beta(|\xi_i^k| + n|\eta_i^k|))^2 + \frac{1}{2}|\langle c_i, d^k\rangle|^2\Big). \qquad (28)
\end{aligned}
$$

Notice that we have assumed that the set $\{i : \langle c_i, x^k\rangle \le 1/\gamma\} \neq \emptyset$, otherwise the right–hand side of (28) vanishes. Using $\gamma = \gamma_k$ given in (32) in the last relation and inserting in (27), we arrive to

$$
\varphi(x^{k+1}) - \varphi(x^k) \le -\, s\hat{c}\|d^k\|^2 + o(s\|d^k\|), \tag{29}
$$

which allows us to conclude that d^k is a descent direction. ■

There are nonconvex problems for which Assumption 2 can not be fullfilled, e.g. when f is concave. In this case, the last proof can be modified to cope with this situation. We will need the following assumption.

Assumption 3. *The matrix B^k satisfies*

$$
|d^\top B^k d| \le \hat{C}\|d\|_2^2, \qquad \textit{for all } d \in \mathbb{R}^m, \tag{30}
$$

for some positive constant $\hat{C}$.

Theorem 2. *Let Assumptions 1 and 3 hold. Consider x^k, ν^k and η^k as in Theorem 1. Moreover, assume in addition that there exist a constant $\tilde{C} > 0$ such that*

$$
0 < \tilde{C}\|d^k\|_2^2 \le \sum_{i:|\langle c_i, x^k\rangle| \le 1/\gamma} \langle c_i, d^k\rangle^2, \tag{31}
$$

for every k, and that the parameter γ is chosen at each iteration as follows

$$
\gamma_{k+1} > \frac{1}{2\beta}\left(\frac{2\hat{C}\|d^k\|_2^2}{\sum_{i:|\langle c_i, x^k\rangle| \le 1/\gamma} \langle c_i, d^k\rangle^2} + \frac{\||\nu^k| + \beta(|\xi^k| + n|\eta^k|)\|_2^2}{\min \langle c_i, d^k\rangle^2} + 1 \right), \tag{32}
$$

then, d^k is a descent direction, i.e.:

$$
\varphi(x^{k+1}) < \varphi(x^k).
$$

Proof. Following the same arguments and notation of the proof of Theorem 1, we have that

$$
\begin{aligned}
\varphi(x^{k+1}) - \varphi(x^k) \leq & - s{d^k}^\top \left[B^k + \beta\Gamma^k\right] d^k - s\beta \sum_{i\in\mathcal{S}_k} \xi_i^k \langle c_i, d^k\rangle \\
& + s\beta \sum_{i\notin\mathcal{S}_k} |\langle c_i, \Pi d^k)\rangle| - \nabla f(x^k)^\top \Pi(x^k + sd^k) + o(s\|d^k\|) \\
\leq & s\hat{C}\|d^k\|^2 - \gamma s\beta \sum_{i:|\langle c_i, x^k\rangle|\leq 1/\gamma} \langle c_i, d^k\rangle^2 + o(s\|d^k\|) \\
& - s\beta \sum_{i\in\mathcal{S}_k} \xi_i^k \langle c_i, d^k\rangle + s\beta|\mathcal{S}_k^C||\langle c_{i^*}, \Pi d^k)\rangle| - \nabla f(x^k)^\top \Pi(x^k + sd^k),
\end{aligned}
\tag{33}
$$

By the estimate (28) and (31) we get

$$
\begin{aligned}
\varphi(x^{k+1}) - \varphi(x^k) \leq & s\hat{C}\|d^k\|^2 - \gamma s\beta \sum_{i:|\langle c_i, x^k\rangle|\leq 1/\gamma} \langle c_i, d^k\rangle^2 + o(s\|d^k\|) \\
& + s\Big(\sum_{\substack{i\in\mathcal{S}_k \\ \langle c_i, x^k\rangle = 0}} \frac{1}{2}(|\nu_i^k| + \beta(|\xi_i^k| + n|\eta_i^k|))^2 + \frac{1}{2}|\langle c_i, d^k\rangle|^2\Big) \\
\leq & -\frac{s}{2} \sum_{i:|\langle c_i, x^k\rangle|\leq 1/\gamma} \langle c_i, d^k\rangle^2 + o(s\|d^k\|).
\end{aligned}
\tag{34}
$$

Finally, the right–han side of the last relation is negative for sufficiently small s. ■

Definition 1. *We will say that a function f is a KL–function if f satisfies the Kurdyka–Łojasiewicz inequality, that is: for every $y \in \mathbb{R}$ and for every bounded subset $E \subset \mathbb{R}^m$, there exist three constants $\kappa > 0$, $\zeta > 0$ and $\theta \in [0, 1[$ such that for all $z \in \partial f(x)$ and every $x \in E$ such that $|f(x) - y| \leq \zeta$, it follows that*

$$\kappa|f(x) - y|^\theta \leq \|z\|_2, \tag{35}$$

with the convention $0^0 = 0$.

Theorem 3. *Suppose that Assumptions 1–2 are satisfied and that φ is a KL–function (i.e. satisfies the Kurdyka–Łojasiewicz condition). Then, the sequence $\{x^k\}_{k\in\mathbb{N}}$ generated by Algorithm 1 converges to a point $\bar{x}$ such that $0 \in \nabla f(\bar{x}) + \beta\, C^\top \partial\|\cdot\|_1(C\bar{x})$.*

Proof. The proof of this convergence result is analogous to the proof of Theorem 2 in [6]. Indeed, notice that the sequence $\{x^k\}_{k\in\mathbb{N}}$ lies in the level set $\{x : \varphi(x) \leq \varphi(x^0)\}$, which in view of Assumption (1) is compact. Moreover, by Theorem 1, for s^k sufficiently small, there exists $\mu > 0$ such that the sequence $\{\varphi(x^k)\}_{k\in\mathbb{N}}$ enjoys the property:

$$\mu\|d^k\|_2^2 \leq f(x^k) + \beta\|Cx^k\|_1 - f(x^{k+1}) - \beta\|Cx^{k+1}\|_1, \tag{36}$$

and $\varphi(x^k)$ converges to some value φ_∞ as $k \to \infty$. By using the Kurdyka–Łojasiewicz condition and Assumption 2, there exist $\kappa > 0$ and $\theta \in [0,1)$ such that

$$\kappa|\varphi(x^k) - \varphi_\infty|^\theta \leq \|\nabla f(x^k) + \beta C^\top \xi^k\|_2 \leq \frac{\hat{C}}{\kappa}\|d^k\|_2, \quad \forall \xi \in \partial\left(\beta\|\cdot\|_1\right)(Cx^k). \tag{37}$$

holds. Therefore, majoring (36) using (37) it can be concluded the summability of the sequence $\{\|d^k\|\}_{k\in\mathbb{N}}$. Which in turn, by (22), implies that $\{x^k\}_{k\in\mathbb{N}}$ is a Cauchy sequence and thus convergent. Let us denote its limit by $\bar{x}$.

Since $\nabla f(x^k) + \beta C^\top \xi^k \in \nabla f(x^k) + \beta C^\top \partial\|\cdot\|_1)(Cx^k)$ then we have

$$(x^k, \nabla f(x^k) + \beta C^\top \xi^k) \in \operatorname{Graph}(\nabla f + \beta C^\top \partial\|\cdot\|_1(C\cdot))$$

Finally, using (37) and taking the limit $k \to \infty$ we obtain

$$(x^k, \nabla f(x^k) + \beta C^\top \xi^k) \to (\bar{x}, 0) \quad \text{as} \quad k \to +\infty.$$

Hence $(\bar{x}, 0)$ belongs to $\operatorname{Graph}(\nabla f + \partial(\beta\|\cdot\|_1))$ due to its closedness which is equivalent to the relation $0 \in \nabla f(\bar{x}) + \partial(\beta\|\cdot\|_1)(\bar{x})$. ■

Theorem 4 (Rate of convergence)**.** *Let Assumptions 1–2 hold and assume also that φ is a KL–function with Łojasiewicz exponent $\theta \in (0,1)$. Let $\{x^k\}_{k\in\mathbb{N}}$ be a sequence generated by Algorithm 1, converging to a local solution $\bar{x}$. Then, the following rates hold:*

(i) If $\theta \in (0, \frac{1}{2})$, then there exist $c > 0$ and $\tau \in [0,1)$ such that

$$\|x^k - \bar{x}\| \leq c\tau^k$$

.

(ii) If $\theta \in (\frac{1}{2}, 1)$, then there exist $c > 0$ such that

$$\|x^k - \bar{x}\| \leq ck^{-\frac{1-\theta}{2\theta-1}}.$$

Proof. We follow the ideas from [2]. From (22) and the quadratic growth (29), for sufficiently small s, there is a positive constant c such that

$$\|x^{k+1} - x^k\|_2^2 \leq c\|d^k\|^2 \leq \varphi(x^k) - \varphi(x^{k+1}), \tag{38}$$

Without loss of generality, we assume that $\varphi(\bar{x}) = 0$ (we can always replace $\varphi(\cdot)$ by $\varphi(\cdot) - \varphi(\bar{x})$) and by multiplying relation (38) by $\varphi(x^k)^{-\theta}$and using the fact that the real function $\mathbb{R}_+ \ni t \mapsto t^{1-\theta}$ is a concave differentiable function

$$\begin{aligned}\|x^{k+1} - x^k\|_2^2 \varphi(x^k)^{-\theta} \leq& (\varphi(x^k) - \varphi(x^{k+1}))\varphi(x^k)^{-\theta} \\ \leq& \frac{1}{1-\theta}(\varphi(x^k)^{1-\theta} - \varphi(x^{k+1})^{1-\theta}).\end{aligned}$$

On the other hand, φ is a KL–function thus, from the last relation, we get

$$\begin{aligned}\|x^{k+1} - x^k\|_2^2 \leq& \frac{1}{1-\theta}(\varphi(x^k)^{1-\theta} - \varphi(x^{k+1})^{1-\theta})\varphi(x^k)^\theta \\ \leq& \frac{1}{1-\theta}(\varphi(x^k)^{1-\theta} - \varphi(x^{k+1})^{1-\theta})\|\nabla f(x^k) + \beta C^\top \xi^k\|_2. \qquad (39)\end{aligned}$$

Fhurther, ξ^k corresponds to the minimum norm subgradient solving (MinSub); therefore, by feasibility of ξ^{k-1} we have that $\|\nabla f(x^k) + \beta C^\top \xi^k\|_2 \leq \|\nabla f(x^k) + \beta C^\top \xi^{k-1}\|_2$ which can be inserted in (39) and combined with (11) and Assumption 1 to obtain that

$$\begin{aligned}\|x^{k+1} - x^k\|_2^2 \leq & \frac{1}{1-\theta}(\varphi(x^k)^{1-\theta} - \varphi(x^{k+1})^{1-\theta})\|\nabla f(x^k) + \beta C^\top \xi^{k-1}\|_2 \\ \leq & \frac{1}{1-\theta}(\varphi(x^k)^{1-\theta} - \varphi(x^{k+1})^{1-\theta})\Big(\|\nabla f(x^k) - \nabla f(x^{k-1})\|_2 \\ & + \|\nabla f(x^{k-1}) + \beta C^\top \xi^{k-1}\|_2\Big). \\ \leq & \frac{1}{1-\theta}(\varphi(x^k)^{1-\theta} - \varphi(x^{k+1})^{1-\theta})(L_f\|x^k - x^{k-1}\|_2 + \frac{\hat{C}}{\kappa}\|d^{k-1}\|_2).\end{aligned}$$

As before, we invoke Remark 2 to infer that for sufficiently small s it follows that $\Pi x^{k-1} = 0$ then $s\|d^{k-1}\|_2 \leq s\|\mathcal{P}d^{k-1}\|_2 = \|x^k - x^{k-1} + \Pi x^{k-1}\|_2 \leq \|x^k - x^{k-1}\|_2$ which together with the above inequality imply that there exist a constant $c > 0$ such that

$$\begin{aligned}2\|x^{k+1} - x^k\|_2 \leq & 2\left(\frac{c}{1-\theta}(\varphi(x^k)^{1-\theta} - \varphi(x^{k+1})^{1-\theta})\right)^{\frac{1}{2}}\|x^k - x^{k-1}\|_2^{\frac{1}{2}}. \\ \leq & \frac{c}{1-\theta}(\varphi(x^k)^{1-\theta} - \varphi(x^{k+1})^{1-\theta}) + \|x^k - x^{k-1}\|_2. \\ \leq & M_\theta(\varphi(x^k)^{1-\theta} - \varphi(x^{k+1})^{1-\theta}) + \|x^k - x^{k-1}\|_2, \qquad (40)\end{aligned}$$

where M_θ is a positive constant depending on θ. Let us sum (40) over k from $k = n$ up to $N > k$:

$$\sum_{k=n}^{N}\|x^{k+1} - x^k\|_2 + \|x^{N+1} - x^N\|_2 \leq M_\theta(\varphi(x^n)^{1-\theta} - \varphi(x^{N+1})^{1-\theta}) + \|x^n - x^{n-1}\|_2,$$

hence, recalling Theorem 3 that $\{\|x^{k+1} - x^k\|_2\}_{k\in\mathbb{N}}$ is summable in virtude of the sumability of the sequence $\{\|d^k\|_2\}_{k\in\mathbb{N}}$ and taking $N \to \infty$, we get

$$\sum_{k=n}^{\infty}\|x^{k+1} - x^k\|_2 \leq M_\theta\varphi(x^n)^{1-\theta} + \|x^n - x^{n-1}\|_2.$$

The last relation in terms of $\Delta^n := \sum_{k=n}^{\infty}\|x^{k+1} - x^k\|_2$ can be rewritten as follows:

$$\begin{aligned}\Delta^n \leq & M_\theta\varphi(x^n)^{1-\theta} + \Delta^{n-1} - \Delta^n. \\ \leq & M_\theta\varphi(x^{n-1})^{1-\theta} + \Delta^{n-1} - \Delta^n. \qquad (41)\end{aligned}$$

Using again that φ is a KL–function, we have from (35) and monotonicity that $\varphi(x^{n-1})^{1-\theta} \leq \frac{1}{\kappa}\|\nabla f(x^{n-1}) + \beta C^\top \xi^{n-1}\|_2^{\frac{1-\theta}{\theta}}$. Thus, observing that $\Delta^{n-1} - \Delta^n = \|x^n - x^{n-1}\|_2$, we obtain

$$\begin{aligned}\Delta^n \leq & \frac{M_\theta}{\kappa}\|\nabla f(x^{n-1}) + \beta C^\top \xi^{n-1}\|_2^{\frac{1-\theta}{\theta}} + \Delta^{n-1} - \Delta^n. \\ \leq & M(\Delta^{n-1} - \Delta^n)^{\frac{1-\theta}{\theta}} + \Delta^{n-1} - \Delta^n, \qquad (42)\end{aligned}$$

where M is a positive constant. Here, we rely on the analysis of a sequence satisfying relation (42) done in [2, pg. 13–15] henceforth (i) and (ii) hold. ■

5. Active–set identification strategy

Second-order methods are known to be expensive when it comes to the computation of the descent direction. Without any additional strategy regarding the numerical solution in the computation of the descent direction (system (11)), the method would hardly become practical for large problems. Therefore, it is important to look at the structure of the pattern matrix C and take it into account in order to improve the computation process.

In an effort to reduce the numerical cost, we extend the notion of *active sets* used in [6] in order to define an effective identification process of the components of the optimization variable which are known to fulfill optimality conditions and therefore, they can be excluded when seeking the descent direction. In this way, the optimization process takes place in a lower dimensional subspace, resulting in a reduction of the computation cost.

A common situation occurs when the matrix C possesses a known structure e.g., when C is the successive difference matrix or "discrete gradient"; in this case, C is a band matrix. We notice that in the multiplication Cx^k, not all the entries of x^k are taking part in the computation of a particular entry of the product Cx^k.

Observing the optimality condition (5), for each $i \in \mathcal{A}^k$ we consider the index set noted by $\mathcal{I}_i^k$, consisting of indexes $j \in \{1,\ldots,m\}$ such that $c_{ij} \neq 0$, satisfying:

$$|[\nabla f(x^k) + \beta C^\top \xi^k]_j| \approx 0. \tag{43}$$

Then, we define the set of active entries of x^k by

$$\mathcal{I}_0^k := \cup_{i\in\mathcal{A}^k} \mathcal{I}_i^k, \tag{44}$$

which corresponds to the set of indexes that are close to satisfy optimality conditions which are active. Thus, we would not move the current approximation x^k in the entries indexed by $\mathcal{I}_0^k$. By contrast, we define the set of indexes $\mathcal{I}_F^k := \{1,\ldots,m\} \setminus \mathcal{I}_0^k$, in which the variable is free to move. Thus, we consider the reduced system:

$$\left[\tilde{B}^k + \beta\tilde{\Gamma}^k\right] \tilde{d}^k = -[\nabla f(x^k) + \beta C^\top \xi(x^k)]_{j\in\mathcal{I}_F}, \tag{45}$$

where

$$\tilde{B}^k := [B_{ij}^k]_{i\in\mathcal{I}_F, j\in\mathcal{I}_F}, \quad \text{and} \quad \tilde{\Gamma}^k := [\Gamma_{ij}^k]_{i\in\mathcal{I}_F, j\in\mathcal{I}_F}.$$

Then, the step 4 of Algorithm 1 can be modified using (45) and by choosing the descend direction d computed according to the formula

$$d_j = \begin{cases} \tilde{d}_j & \text{if } j \in \mathcal{I}_F, \\ 0 & \text{if } j \in \mathcal{I}_0. \end{cases} \tag{46}$$

6. Numerical experiments

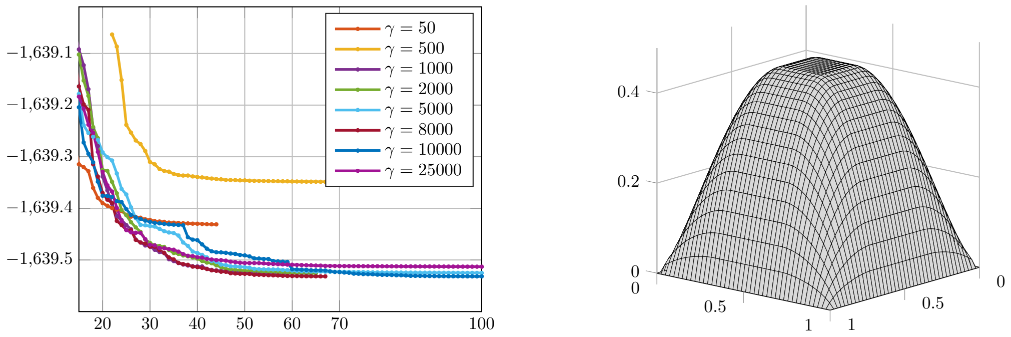
In this section, we conduct numerical experiments of the previous algorithm exhaustively. Several applications of the generalized 1–norm penalization are conducted in order to illustrate the general class of problems that can be handled by our algorithm. The GSOM algorithm was implemented in Matlab. The (MinSub) problem of step 3 was solved by using `quadprog` package from the optimization toolbox whereas the linear system (11) of step 4 was solved using direct methods from Matlab ("backslash solver").

In step 6 we implemented the line–search using a standard backtracking algorithm, by checking the condition (18). For the stopping criteria we use a certain tolerance for the difference of consecutive values for the approximated solution and its respective costs. The numerical experiments presented considered compare different problems with algorithms designed specifically for the problem structure. Therefore, we compare our algorithm with the corresponding algorithm for each problem.

6.1. **Anisotropic total variation in function spaces.** We consider the following problem resembling a simplified version of an anisotropic Bingham fluid flow model:

$$\min_{u\in H_0^1(\Omega)} \frac{1}{2}\int_\Omega |\nabla u|^2 dx - \int_\Omega zu\,dx + \beta\int_\Omega |\nabla(u)|_1\,dx. \tag{47}$$

After discretizing using finite differences the infinite-dimensional problem is reformulated as an energy minimization problem of the form (**P**). Hence, the regular part $f(u) = \frac{1}{2}u^\top Au - b^\top u$ of our minimization problem is written using the matrix A associated to the discrete laplacian and b is the vector corresponding to the discretization of the forcing term z.



(A) History of the cost function in every iteration for different values of γ.

(B) Solution for problem (47) for $\beta = 0.5$ and parameter $\gamma = 25000$

Figure 1. Anisotropic TV model

We observe in Table 1 the effect of using *generalized second–order information* introduced in Section 3.2. The cost values of the objective function were computed by varying the regularization parameter γ for different values of β which are shown in table regt:1 for the first 50 iterations of the algorithm. The first row (in red) shows the cost values achieved by the algorithm when no generalized second–order information is utilized for the computation of the descent direction ($\gamma = 0$). In this case, we notice that without generalized second-order information the cost is bigger in all tests. Moreover, this also can be observed in Figure 1 (A) where the history of the cost is shown for different values of γ for $\beta = 0.5$. There, $\gamma = 0$ is out of chart.

Next, we test with the active–set identification strategy. Here, we compare the execution time with respect to an implementation not hacking this strategy. We confirm the efficiency of using this strategy by measuring time execution for this particular problem, see Table 2.

	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.5$	$\beta = 0.7$	$\beta = 0.9$
$\gamma = 0$	-2640.5471	-2095.9578	-1638.1323	-1258.8208	-946.8398
$\gamma = 50$	-2640.5586	-2096.427	-1639.4316	-1261.8356	-955.7
$\gamma = 500$	-2640.5623	**-2096.4514**	-1639.3464	**-1261.9043**	**-956.3495**
$\gamma = 1000$	-2640.5623	-2096.4502	**-1639.5237**	-1261.8978	-955.7178
$\gamma = 2000$	-2640.5623	-2096.252	-1639.521	-1261.5852	-952.9638
$\gamma = 5000$	-2640.5623	-2096.3521	-1639.515	-1261.3488	-956.0101
$\gamma = 8000$	-2640.5623	-2096.3442	-1639.527	-1261.2423	-954.4819
$\gamma = 10000$	**-2640.5625**	-2096.3564	-1639.4932	-1261.5385	-953.5706
$\gamma = 25000$	**-2640.5625**	-2096.3142	-1639.506	-1261.5549	-955.643

Table 1. Cost function values varying parameters γ and β

	$\beta = 0.35$	$\beta = 0.4$	$\beta = 0.45$	$\beta = 0.5$	$\beta = 0.7$	$\beta = 1$
Active–set	0.0018	0.0017	0.0017	0.0016	0.0018	0.0018
none	0.0031	0.0030	0.0031	0.0033	0.0031	0.0030

Table 2. Average time (in seconds) of the numerical solution of system (11) with and without the Active–Set strategy within the execution of GSOM algorithm

6.2. **Graph trend filtering.** In [16] the authors introduced a technique of filtering data over graphs, and among other applications, this technique was applied in the denoising over graphs using the discrete laplacian as sparsity–inducting operator. There, it was showed that better results may be achieved than other denoising thechniques. In our setting, $C = \Delta^{(2)}$, where for a integer k the operator $\Delta^{(k)}$ is defined recursively by

$$\Delta^{(k+1)} := \begin{cases} (\Delta^{(1)})^\top \Delta^{(k)}, & \text{if } k \text{ is odd,} \\ \Delta^{(1)}\Delta^{(k)}, & \text{if } k \text{ is even,} \end{cases} \tag{48}$$

where $\Delta^{(1)}$ is the oriented incidence matrix of the graph. Notice that $\|\Delta^{(1)}x\|_1 = \sum_{(i,j)\in E} |x_i - x_j|$, where we denote the graph $G = \{N, E\}$. Therefore, $\Delta^{(2)} = {\Delta^{(1)}}^\top \Delta^{(1)}$.

In this example, we consider the denoising of COVID–19 data over a graph corresponding to the Pichincha province of Ecuador connecting adjacent areas or tracts. Hence, each node corresponds to a particular tract of the province territory. The signal data considered in each node consist of the reported number of cases of each tract, denoted by y. This data was provided by the team of the app SALVAVIDAS, developed for monitoring COVID–19 infection in Ecuador, which use the official data provided by the Ministry of Health of Ecuador. The noise source in this kind of data are due to imprecise assignments within tracts, counting errors, false positive or negative cases to name a few. In our example, we assume that the noise induced by these different sources is normally distributed $y \sim N(x_0, \sigma^2 I)$. The sparse graph filtering problem aims to minimize the following cost

$$f(x) = \frac{1}{2}\|x - y\|_2^2 + \beta_1 \|\Delta^{(2)}x\|_1 + \beta_2 \|x\|_1 \tag{49}$$

Figure 2 shows an expected behavior of a first order method (ADMM) compared with a second–order method (GSOM). We observed that GSOM is faster and more precise. However, its computational power requires the solution of a linear system, which may be costly. Nevertheless, the computational cost can be outstripped by utilizing parallelization and numerical techniques.

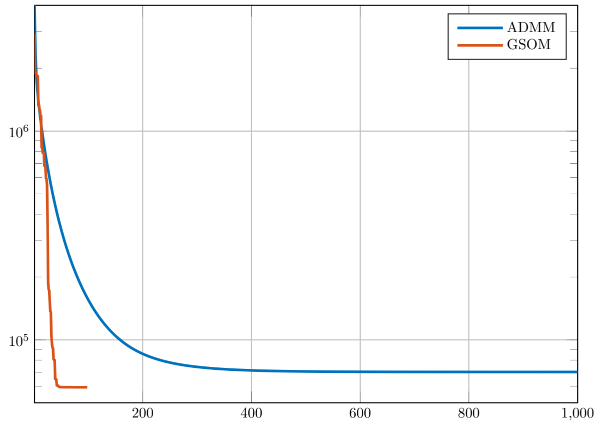
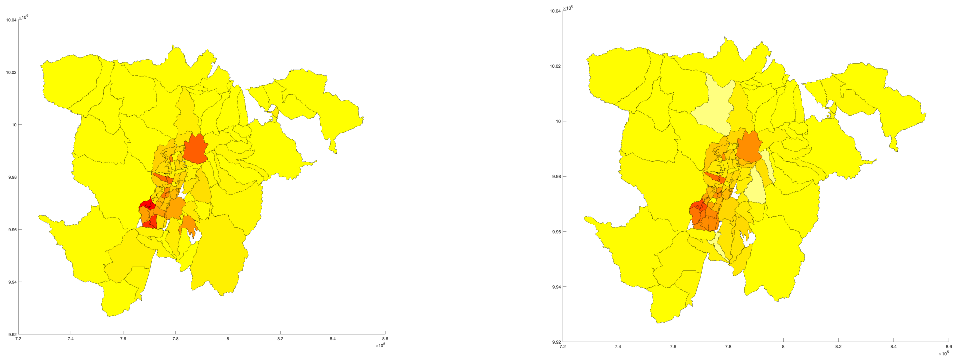


FIGURE 2. Comparison with Fast ADMM algorithm [16]



(A) Original graph data

(B) Filtered graph data

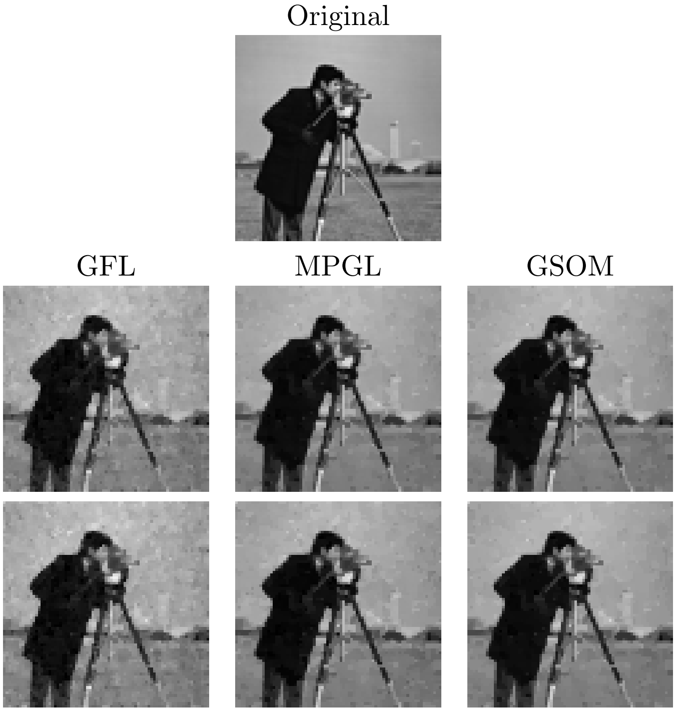
FIGURE 3. Graph trend filtering of COVID-19 data in a graph of Pichincha–Ecuador

6.3. **Image restoration.** Consider the image deconvolution example of [10]. The aim in this problem is to recover an image out of one convoluted with the random matrix A. For instance, this convolution occurs during the camera exposure, producing a blured image. If x is the original image, the contaminated one is modeled by $y = Ax + z$, where $A \in \mathbb{R}^{n\times n}$ and $z \in \mathbb{R}^n$. The recovering process consists in choosing the image x which best fits the observation and at the same time minimizes the term that computes the differences of each pixel with respect to its neighbors by means of a directed graph $G = \{N, E\}$. Thus, we look for the solution of the cost function defined by

$$f(x) = \frac{1}{2}\|Ax - y\|_2^2 + \alpha\|x\|_1 + \beta \sum_{(i,j)\in E} |x_i - x_j|$$

Notice that the last function fits in our settings using the incidence matrix C, associated to the graph G, in order to express the penalizing term as: $\sum_{(i,j)\in E} |x_i - x_j| = \|Cx\|_1$.

In the following example we consider the recovering of an image of size 77×77 from its corrupted observation $y = Ax + b$ with random noise z with standard deviation $\sigma = 0.05$. Here A is a random (uniformly distributed) convolution matrix of size 2000×5929.


Original

GFL

MPGL

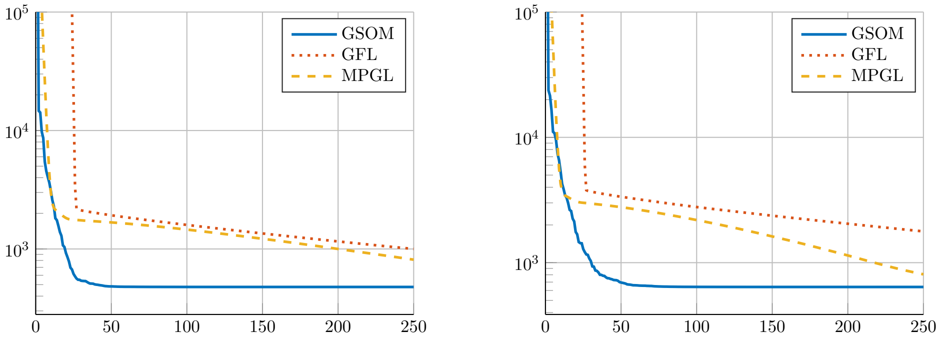
GSOM

Figure 4. Recovered images

Next, we test the Cauchy–denoising model characterized by its non-Gaussian and impulsive property that preserves edges and details of images (see [14] for the formulation of the Cauchy model in function spaces using the TV–norm). Here, the anisotropic version of the discrete Cauchy denoising problem, in a simplified setting, corresponds to the minimization of the nonconvex cost function:

$$\varphi(u) = \sum_i \log(a + (u_i - f_i)^2) + \beta\|Cu\|_1, \tag{50}$$

where C is the difference operator, f is the observed image perturbed with Cauchy noise and $a > 0$ is the scale parameter of the Cauchy distribution. Notice that the nonconvex structure of the optimization problem prevents the application of the standard convex methods.



(A) Noise level: $\sigma = 0.05$

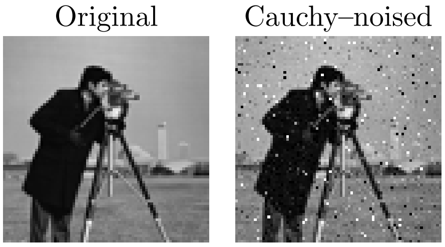
(B) Noise level: $\sigma = 0.2$

FIGURE 5. First 250 iterations of the cost function

Again, an image of size 77 × 77 pixels is considered and a Cauchy–noise is added to the original image according to the formula

$$f = u + v = u + \xi \frac{\eta_1}{\eta_2}, \tag{51}$$

suggested in [14], where $\xi > 0$ gives the noise level, and η_i, $i = 1, 2$, follow the Gaussian distribution with mean 0 and variance 1. In the next experiment we chose $\xi = 0.01$.x



The following set of pictures shows recovered images for different values of the scale parameter a and the composite sparsity penalizing parameter β. Both play an important role in the restoration process. Indeed, we observe that larger values of a result in a reduced level of Cauchy-noise. The same observation applies to higher values of b. As usual, in this type of problems, there is a compromise between the amount of removed noise and the preservation of the details inside the picture.

Because of the nonconvexity of the Cauchy problem, the previous methods can not be applied. There exist methods designed for nonconvex problems; for instance, the iPiano algorithm, see [11], which is based on a forward-backward splitting with inertial splitting techniques. In each step, iPiano requires the computation of the proximal mapping:

$$\hat{x} \mapsto \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \frac{1}{2} \|x - \hat{x}\|_2^2 + \alpha \|Cx\|_1 \right\} \tag{52}$$

which falls in the convex case of (**P**), consequently the previous methods used in the experiments (including GSOM) may be applied for evaluating (52). In our experiments,

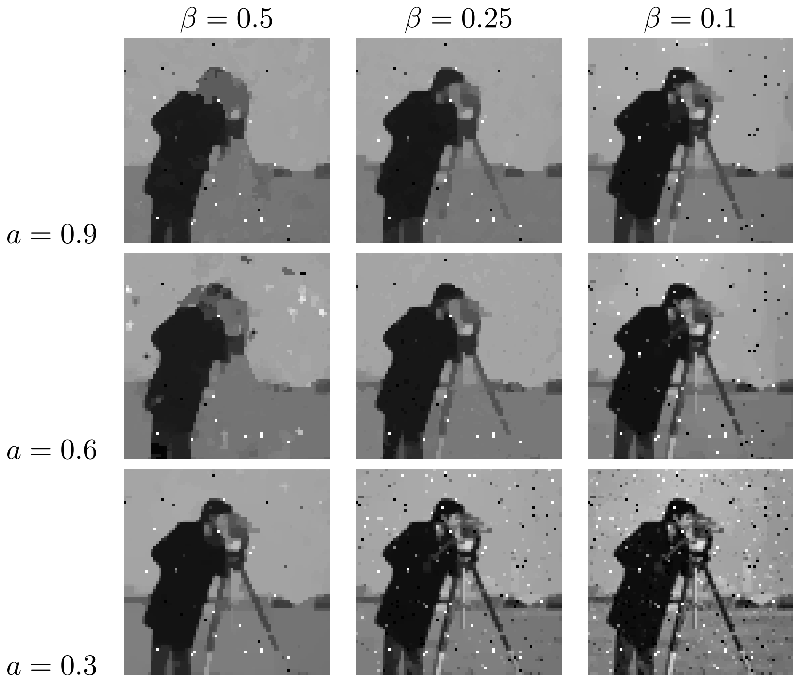


FIGURE 6. Recovered images by GSOM method

we have found that both methods are competitive, depending on the respective algorithms' chosen parameters. Many parameters are involved in iPiano setting, for which we obtained different performances. Nevertheless, there is not hard numerical evidence for the superiority of one of the methods in our tests. In our experiment, the parameters of Cauchy model are: $a = 0.3$ and $\beta = 0.1$. Figure (7) shows the firs 50 iterations of both algorithms with similar performance.

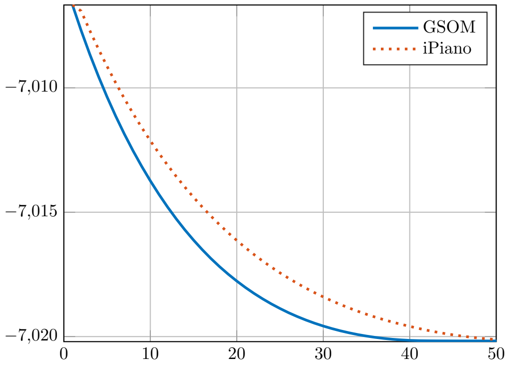


FIGURE 7. First 50 iterations of the cost function

References

[1] G. Andrew and J. Gao. Scalable training of ℓ_1—regularized log-linear models. *In Proceedings of the Twenty Fourth Conference on Machine Learning (ICML)*, 2007.

[2] Hedy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1-2):5–16, 2009.

[3] Amir Beck and Marc Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE transactions on image processing*, 18(11):2419–2434, 2009.

[4] Michele Benzi, Gene H Golub, and Jörg Liesen. Numerical solution of saddle point problems. *Acta numerica*, 14:1–137, 2005.

[5] R. Byrd, G. Chin, J. Nocedal, and Y. Wu. Sample size selection in optimization methods for machine learning. *Mathematical Programming*, 134(1), 2011.

[6] Juan Carlos De Los Reyes, Estefanía Loayza, and Pedro Merino. Second-order orthant-based methods with enriched hessian information for sparse ℓ_1-optimization. *Computational Optimization and Applications*, 67(2):225–258, 2017.

[7] Neil K Dhingra, Sei Zhen Khong, and Mihailo R Jovanović. A second order primal-dual algorithm for nonsmooth convex composite optimization. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 2868–2873. IEEE, 2017.

[8] R Fletcher. A model algorithm for composite nondifferentiable optimization problems. In *Nondifferential and Variational Techniques in Optimization*, pages 67–76. Springer, 1982.

[9] Guy Gilboa and Stanley Osher. Nonlocal operators with applications to image processing. *Multiscale Modeling & Simulation*, 7(3):1005–1028, 2009.

[10] Dong Gong, Mingkui Tan, Yanning Zhang, Anton van den Hengel, and Qinfeng Shi. Mpgl: An efficient matching pursuit method for generalized lasso. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[11] Peter Ochs, Yunjin Chen, Thomas Brox, and Thomas Pock. ipiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419, 2014.

[12] Irina Rish, Guillermo A Cecchi, Aurelie Lozano, and Alexandru Niculescu-Mizil. *Practical applications of sparse modeling.* MIT Press, 2014.

[13] Irina Rish and Genady Grabarnik. *Sparse modeling: theory, algorithms, and applications.* CRC press, 2014.

[14] Federica Sciacchitano, Yiqiu Dong, and Tieyong Zeng. Variational approach for restoring blurred images with cauchy noise. *SIAM Journal on Imaging Sciences*, 8(3):1894–1922, 2015.

[15] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, Feb 2005.

[16] Yu-Xiang Wang, James Sharpnack, Alexander J Smola, and Ryan J Tibshirani. Trend filtering on graphs. *The Journal of Machine Learning Research*, 17(1):3651–3691, 2016.

[‡] Research Center on Mathematical Modeling (MODEMAT) and Department of Mathematics, Escuela Politécnica Nacional, Quito, Ecuador