Difference of two norms-regularizations for Q-Lasso

Abdellatif Moudafi
Aix-Marseille Université, L.I.S UMR CNRS 7020, Domaine Universitaire de Saint-Jérome, Marseille, France

Received 8 June 2018
Revised 6 July 2018
Accepted 8 July 2018

Abstract

The focus of this paper is in Q-Lasso introduced in Alghamdi et al. (2013) which extended the Lasso by Tibshirani (1996). The closed convex subset Q belonging in a Euclidean m-space, for $m \in \mathrm{IN}$, is the set of errors when linear measurements are taken to recover a signal/image via the Lasso. Based on a recent work by Wang (2013), we are interested in two new penalty methods for Q-Lasso relying on two types of difference of convex functions (DC for short) programming where the DC objective functions are the difference of l_1 and l_{σ_q} norms and the difference of l_1 and l_r norms with $r > 1$. By means of a generalized q-term shrinkage operator upon the special structure of l_{σ_q} norm, we design a proximal gradient algorithm for handling the DC $l_1 - l_{\sigma_q}$ model. Then, based on the majorization scheme, we develop a majorized penalty algorithm for the DC $l_1 - l_r$ model. The convergence results of our new algorithms are presented as well. We would like to emphasize that extensive simulation results in the case $Q = \{b\}$ show that these two new algorithms offer improved signal recovery performance and require reduced computational effort relative to state-of-the-art l_1 and l_p ($p \in (0, 1)$) models, see Wang (2013). We also devise two DC Algorithms on the spirit of a paper where exact DC representation of the cardinality constraint is investigated and which also used the largest-q norm of l_{σ_q} and presented numerical results that show the efficiency of our DC Algorithm in comparison with other methods using other penalty terms in the context of quadratic programing, see Jun-ya et al. (2017).

Keywords Q-Lasso, Split feasibility, Soft-thresholding, DC-regularization, Proximal gradient algorithm, Majorized penalty algorithm, Shrinkage, DCA algorithm

Paper type Original Article

1. Introduction and preliminaries

The process of compressive sensing (CS) [8], which consists of encoding and decoding, is rapidly consolidated year after year due to the blooming of large datasets which become increasingly important and available. The process of encoding involves taking a set of (linear) measurements, $b = Ax$, where A is a matrix of size $m \times n$. If $m < n$, we can compress the signal $x \in \mathrm{IR}^n$, whereas the process of decoding is to recover x from b where x is assumed to be sparse. It can be formulated as an optimization problem, namely

$$\min \|x\|_0 \text{ subject to } Ax = b, \tag{1.1}$$

Publishers note: The publisher wishes to inform readers that the article "Difference of two norms-regularizations for Q-Lasso" was originally published by the previous publisher of *Applied Computing and Informatics* and the pagination of this article has been subsequently changed. There has been no change to the content of the article. This change was necessary for the journal to transition from the previous publisher to the new one. The publisher sincerely apologises for any inconvenience caused. To access and cite this article, please use Moudafi, A. (2021), "Difference of two norms-regularizations for Q-Lasso", *Applied Computing and Informatics*. Vol. 17 No. 1, pp. 79-89. The original publication date for this paper was 19/17/2018.

Applied Computing and Informatics
Vol. 17 No. 1, 2021
pp. 79-89
Emerald Publishing Limited
e-ISSN: 2210-8327
p-ISSN: 2634-1964
DOI 10.1016/j.aci.2018.07.002

where $\|\cdot\|_0$ is the l_0 norm, which counts the number of nonzero entries of x; namely

$$\|x\|_0 = |\{x_i; x_i \neq 0\}| \tag{1.2}$$

with $|\cdot|$ being here the cardinality, i.e., the number of elements of a set. Hence minimizing the l_0 norm amounts to finding the sparsest solution. One of the difficulties in CS is solving the decoding problem above, since l_0 optimization is NP-hard. An approach that has gained popularity is to replace l_0 by the convex norm l_1 since it often gives a satisfactory sparse solution and has been applied in many different fields such as geology and ultrasound imaging.

More recently, nonconvex metrics were used as alternative approaches to l_1, especially the nonconvex metric l_p for $p \in (0, 1)$ in [6] which can be interpreted as a continued approximation strategy of l_0 as $p \to 0$. A great deal of research has been conducted into l_p problems including all kinds of variants and related algorithms, as you can see in [4] and references therein. The convex l_1 relaxation compared to the nonconvex problem (l_p) is generally more difficult to handle. However, it was shown in [12] that the potential reduction method can solve this special nonconvex problem in polynomial time with arbitrarily given accuracy.

Most recently, the majority of such sparsity inducing functions are unified as the notion of DC programming in [9], including log-sum, smoothly clipped absolute deviation and capped-l_1 penalty. Generally, DC programming problem can be solved through a primal–dual convex relaxations algorithm which is famous in the literature of DC Programming [11]. Other algorithms appeared as for solving application problems of DC programming in the area of finance and insurance, data analysis, machine learning as well as signal processing. However, as noted in [18], among the above mentioned DC programming approaches for sparse reconstruction, most of them are mainly preserving the separability properties of both l_0 and l_1 norms.

To begin with, let us recall that the lasso of Tibshirani [16] is given by the following minimization problem

$$\min_{x \in IR^n} \frac{1}{2} \|Ax - b\|_2^2 + \gamma \|x\|_1, \tag{1.3}$$

A being an $m \times n$ real matrix, $b \in IR^m$ and $\gamma > 0$ is a tuning parameter. The latter is nothing else than the basic pursuit (BP) of Chen et al. [7], namely

$$\min_{x \in IR^n} \|x\|_1 \text{ such that } Ax = b. \tag{1.4}$$

However, the constraint $Ax = b$ being inexact due to errors of measurements, the problem (1.4) can be reformulated as

$$\min_{x \in IR^n} \|x\|_1 \text{ subject to } \|Ax - b\|_p \leq \varepsilon, \tag{1.5}$$

where $\varepsilon > 0$ is the tolerance level of errors and p is often $1, 2$ or ∞. It is noticed in [1] that (1.5) can be rewritten as

$$\min_{x \in IR^n} \|x\|_1 \text{ subject to } Ax \in Q, \tag{1.6}$$

in the case when $Q := B_\varepsilon(b)$, the closed ball in IR^m with center b and radius ε.

Now, when Q is a nonempty closed convex set of IR^m and P_Q the orthogonal projection from IR^m onto the set Q and by observing that the constraint is equivalent to the condition $Ax - P_Q(Ax) = 0$, this leads to the following Lagrangian formulation

$$\min_{x\in IR^n} \frac{1}{2}\,\|(I-P_Q)Ax\|_2^2+\gamma\|x\|_1, \tag{1.7}$$

$\gamma>0$ being a Lagrangian multiplier.

A link is also made in [1] with split feasibility problems [5] which consist in finding x satisfying

$$x\in C,\ Ax\in Q, \tag{1.8}$$

with C and Q two nonempty closed convex subsets of IR^n and IR^m, respectively. An equivalent formulation of (1.8) as a minimization problem is given by

$$\min_{x\in C} \frac{1}{2}\,\|(I-P_Q)Ax\|_2^2, \tag{1.9}$$

and its l_1-regularization is

$$\min_{x\in C} \frac{1}{2}\,\|(I-P_Q)Ax\|_2^2+\gamma\|x\|_1, \tag{1.10}$$

with $\gamma>0$ a regularization parameter.

This convex relaxation approach was frequently employed, see for example [1,20] and references there in. As the level curves of l_1-l_2 are closer to l_0 than those of l_1, this motivated us in [14] to propose a regularization of split feasibility problems by means of the nonconvex l_1-l_2, namely

$$\min_{x\in C} \frac{1}{2}\,\|(I-P_Q)Ax\|_2^2+\gamma(\|x\|_1-\|x\|_2), \tag{1.11}$$

and present three algorithms with their convergence properties [14]. Unlike the separable sparsity inducing functions involved in the aforementioned DC programming for problem (l_0), we are interested in the two first sections of this work to two specific types of DC programming with un-separable objective functions, which are in the form of difference functions between two norms, namely the new notion l_{σ_q} denoting the sum of q largest elements of a vector in magnitude (i.e., the l_1 norm of q-term best approximation of a vector) introduced in [18] and the classical l_r norm with $r>1$. Obviously l_{σ_q} and l_r $(r>1)$ are regular convex norms. The corresponding DC programs are as follows:

$$\min_{x\in IR^n}\big(\|x\|_1-\varepsilon\|x\|_{\sigma_q}:Ax=b\big), \tag{1.12}$$

and

$$\min_{x\in IR^n}(\|x\|_1-\varepsilon\|x\|_r:Ax=b), \tag{1.13}$$

where $\varepsilon\in(0,1]$, $\|x\|_{\sigma_q}$ is defined as the sum of the q largest elements of x in magnitude, $q\in\{1,2,\cdots n\}$ and $r>1$. We would like to emphasize that the following least-squares variant of (1.12) and (1.13), were studied in the recent work by Wang [18]:

$$\min_x\left(f(x):=\frac{1}{2}\,\|Ax-b\|^2+\mu\big(\|x\|_1-\varepsilon\|x\|_{\sigma_q}\big)\right), \tag{1.14}$$

where $\mu>0$ and $\varepsilon\in(0,1)$, and

$$\min_x \left(\tilde{f}(x) := \frac{1}{2} \|Ax - b\|^2 + \mu(\|x\|_1 - \varepsilon \|x\|_r) \right), \tag{1.15}$$

where $r > 0$ and $\varepsilon \in (0, 1)$.

This paper proposes generalizations to Q-Lasso, namely

$$\min_x \left(f(x) := \frac{1}{2} \|(I - P_Q)Ax\|^2 + \mu\left(\|x\|_1 - \varepsilon \|x\|_{\sigma_q}\right) \right),$$

where $\mu > 0$ and $\varepsilon \in (0, 1)$, as well as

$$\min_x \left(\tilde{f}(x) := \frac{1}{2} \|(I - P_Q)Ax\|^2 + \mu(\|x\|_1 - \varepsilon \|x\|_r) \right),$$

where $r > 0$ and $\varepsilon \in (0, 1)$, and our attention will be focused on the algorithmic aspect.

The rest of the paper is organized as follows. In Sections 2 and 3, two DC-penalty methods instead of conventional methods such as l_1 or $l_1 - l_2$ minimization are proposed. Their convergence to a stationary point are also analyzed. The first iterative minimization method is based on the gradient proximal algorithm and the second one is designed by means of the majored penalty strategy. Furthermore, relying on DCA (difference of convex algorithm) two other algorithms are proposed and their convergence results are established in Section 3 and 4.

2. Proximal gradient algorithm

First, we recall that the subdifferential of a convex function ϕ is given by

$$\partial\phi(x) := \{u \in IR^n; \phi(y) \geq \phi(x) + \langle u, y - x \rangle \ \forall y \in IR^n\}. \tag{2.1}$$

Each element of $\partial\phi(x)$ is called subgradient. If $\phi(x) = \frac{1}{2} \| (I - P_Q)Ax\|^2$, it is well-known that

$$\partial\phi(x) = \nabla\phi(x) = A^T(I - P_Q)Ax, \tag{2.2}$$

and when $\phi(x) = \|x\|_1$, we have

$$(\partial\phi(x))_i = \begin{cases} sgn(x_i) & \text{if } x_i \neq 0; \\ [-1, 1] & \text{if } x_i = 0. \end{cases} \tag{2.3}$$

The *indicator function* of a set $C \subseteq IR^n$ is defined by

$$i_C(x) = \begin{cases} 0 & \text{if } x \in C; \\ +\infty & \text{otherwise.} \end{cases} \tag{2.4}$$

Moreover, the *normal cone* of a set C at $x \in C$, denoted by $N_C(x)$ is defined as

$$N_C(x) := \{d \in IR^n | \langle d, y - x \rangle \leq 0, \forall y \in C\}. \tag{2.5}$$

Connection between the above definitions is given by the key relation $\partial i_C = N_C$.

In this section our interest is in solving the DC programming

$$\min_x \left(f(x) := \frac{1}{2} \|(I - P_Q)Ax\|^2 + \mu\left(\|x\|_1 - \varepsilon \|x\|_{\sigma_q}\right) \right), \tag{2.6}$$

where $\mu > 0$ and $\varepsilon \in (0, 1)$.

Similar to l_1 norm, l_2 norm, etc., we adopt the notation $\|x\|_{\sigma_q}$ to denote the norm of l_{σ_q} which is defined a line below (1.13) and we design an iterative algorithm based both on a generalized q-term shrinkage operator and on the proximal gradient algorithm framework.

At this stage, observe that the restriction on ε guarantees that $f(x) \geq 0$ for all x. To solve (2.6), we consider the following standard proximal gradient algorithm:

1. **Initialization:** Let x_0 be given and set $L > \lambda_{\max(A^TA)}$ with $\lambda_{\max(A^TA)}$ the maximal eigenvalue.
2. For $k = 0, 1, \cdots$ find

$$x_{k+1} \in \underset{x \in \mathbb{R}^n}{Argmin}\left(\left\langle A^T(I - P_Q)Ax_k, x - x_k \right\rangle + \frac{L}{2}\|x - x_k\|^2 + \mu\left(\|x\|_1 - \varepsilon\|x\|_{\sigma_q}\right)\right). \quad (2.7)$$

Observe that subproblem (2.7) can equivalently formulated as

$$\min_x \left(\frac{L}{2}\|x - \left(x_k - \frac{1}{L}A^T(I - P_Q)Ax_k\right)\|^2 + \mu\left(\|x\|_1 - \varepsilon\|x\|_{\sigma_q}\right)\right). \quad (2.8)$$

Thus, it suffices to consider the solutions to the following minimization problem

$$\min_x \left(\frac{1}{2}\|x - y\|^2 + \lambda_1\|x\|_1 - \lambda_2\|x\|_{\sigma_q}\right), \quad (2.9)$$

with a given vector y and positive numbers $\lambda_1 > \lambda_2 > 0$. An explicit solution of this problem is given by the following result, see [18].

Proposition 2.1 *Let* $\{i_1, \cdots, i_n\}$ *be the indices such that*

$$|y_{i_1}| \geq |y_{i_2}| \geq \cdots \geq |y_{i_n}|.$$

Then $x^* := prox_{\lambda_1\|x\|_1 - \lambda_2\|x\|_{\sigma_q}}(y)$ *with*

$$x_i^* = \begin{cases} sign(y_i)\max\{|y_i| - (\lambda_1 - \lambda_2), 0\} & \text{if } i = i_1, i_2, \cdots, i_q; \\ sign(y_i)\max\{|y_i| - \lambda_1, 0\} & \text{otherwise} \end{cases} \quad (2.10)$$

is a solution of (2.9).

The proximal operator above (called the generalized q-term shrinkage operator in [18]) amounts to write the algorithm as follows:

Proximal Gradient Algorithm:

1. **Start:** Let x_0 be given and set $L > \lambda_{\max(A^TA)}$ with $\lambda_{\max(A^TA)}$ the maximal eigenvalue.
2. For $k = 0, 1, \cdots$ find

$$y_{k+1} = \left(x_k - \frac{1}{L}A^T(I - P_Q)Ax_k\right),$$

$$\text{Sort } y_{k+1} \text{ as } |y_{i_1}| \geq |y_{i_2}| \geq \cdots \geq |y_{i_n}|,$$

$$(x_{k+1})_i = \begin{cases} sign(y_i)\max\{|y_i| - \mu(1 - \varepsilon), 0\} & \text{if } i = i_l; \\ sign(y_i)\max\{|y_i| - \mu, 0\} & \text{otherwise} \end{cases} \quad (2.11)$$

where $l = 1, \cdots, q$.

End.

Now, we are in a position to show the following convergence result of the scheme (2.7):

Proposition 2.2 *The sequence* (x_k) *generated by the Proximal Gradient Algorithm above converges to a stationary point of problem (2.6).*

Proof. Remember that $h(x) = \frac{1}{2}\|(I-P_Q)Ax\|^2$ is differentiable and its gradient $\nabla h(x) = A^T(I-P_Q)Ax$ is Lipschitz continuous with constant $\tilde{L} := \lambda_{\max(A^TA)}$. By [3]-Proposition A.24, we have

$$f(x_{k+1}) \leq \frac{1}{2}\|(I-P_Q)Ax_k\|^2 + \left\langle A^T(I-P_Q)Ax_k, x_{k+1}-x_k\right\rangle + \frac{\tilde{L}}{2}\|x_{k+1}-x_k\|^2 + \mu\left(\|x_{k+1}\|_1 - \varepsilon\|x_{k+1}\|_{\sigma_q}\right).$$

Combining this with definition of x_{k+1}, we obtain

$$f(x_{k+1}) \leq f(x_k) - \frac{L-\tilde{L}}{2}\|x_{k+1}-x_k\|^2. \tag{2.12}$$

Since $L > \tilde{L}$, we see immediately that $f(x_{k+1}) \leq f(x_k)$ and thus the sequence $(f(x_k))$ is convergent since f is a non-negative function. Furthermore, we obtain that $\sum_k \|x_{k+1}-x_k\|^2 < +\infty$ which follows by summing (2.12) from $k=0$ to ∞. As a further consequence, we note that

$$\mu(1-\varepsilon)\|x_k\|_1 \leq \mu\left(\|x_k\|_1 - \varepsilon\|x_k\|_{\sigma_q}\right) \leq f(x_k) \leq f(x_0).$$

Since $\mu(1-\varepsilon) > 0$, we have that (x_k) is bounded. Moreover, the objective function f is square term plus a piecewise linear function which ensures that f is semi-algebric and hence satisfies Kurdyka-Lojasiewicz inequality. [2]-Theorem 5.1 is then applicable and obtain that (x_k) is convergent to a stationary point of (2.6). □

3. Majorized penalty algorithm

Consider the following minimization problem

$$\min_x \left(\tilde{f}(x) := \frac{1}{2}\|(I-P_Q)Ax\|^2 + \mu(\|x\|_1 - \varepsilon\|x\|_r\right), \tag{3.1}$$

where $A \in IR^{m\times n}$, Q a nonempty closed convex set of IR^m, $r \geq 0$ and $\varepsilon \in (0,1)$.

First, observe again that conditions on ε guarantees that $\tilde{f}(x) \geq 0$ for all x. We will now describe an algorithm for solving (3.1), based on the majorized penalty approach see, for example, [18] and references therein. Following the same lines as in [18], we start by constructing a majorization of $\tilde{f}$. To that end let $L > \lambda_{\max}(A^TA)$, then for any $x, y \in IR^n$, we have

$$\frac{1}{2}\|(I-P_Q)Ax\|_2^2 \leq \frac{1}{2}\|(I-P_Q)Ay\|_2^2 + \left\langle A^T(I-P_Q)Ay, x-y\right\rangle + \frac{L}{2}\|x-y\|_2^2.$$

Moreover, by invoking the convexity of the norm $\|x\|_r$ and definition of its subdifferential, we also have

$$\|x\|_r \geq \|y\|_r + \langle g(y), x-y\rangle \text{ with } g(y) \in \partial\|y\|_r,$$

where

$$[g(y)]_i = \begin{cases} \dfrac{sign(y_i)|y_i|^{r-1}}{\|y\|_r^{r-1}} & \text{if } y \neq 0; \\ 0 & \text{otherwise.} \end{cases} \tag{3.2}$$

Hence, if we define

$$F(x,y) = \frac{1}{2}\|(I-P_Q)\mathrm{Ay}\|_2^2 + \left\langle A^T(I-P_Q)Ay, x-y\right\rangle + \frac{L}{2}\|x-y\|^2 + \mu(\|x\|_1 - \varepsilon\|y\|_r - \varepsilon\langle g(y), x-y\rangle),$$

hence, for every $x, y \in IR^n$, we get

$$F(x,y) \geq \tilde{f}(x) \text{ and } F(y,y) = \tilde{f}(y).$$

Starting with an initial iterate x_0, the majorized penalty approach above updates x_k by solving

$$x_{k+1} = arg\min_x F(x, x_k). \tag{3.3}$$

This leads to the following explicit formulation of x_{k+1} by means of the proximity (shrinkage) operator of $\|x\|_1$:

$$\begin{aligned} x_{k+1} &= arg\min_x \left(\left\langle A^T(I-P_Q)Ax_k, x-x_k\right\rangle + \frac{L}{2}\|x-x_k\|^2 + \mu(\|x\|_1 - \varepsilon\langle g(x_k), x-x_k\rangle)\right) \\ &= arg\min_x \left(\frac{L}{2}\|x-x_k + \frac{1}{L}\left(A^T(I-P_Q)Ax_k - \mu\varepsilon g(x_k)\right)\|^2 + \mu\|x\|_1\right) \\ &= prox_{\frac{\mu}{L}\|x\|_1}\left(x_k - \frac{1}{L}\left(A^T(I-p_Q)Ax_k - \mu\varepsilon g(x_k)\right)\right) \\ &= sgn(v_k)\circ\max\left\{|v_k| - \frac{\mu}{L}, 0\right\}, \end{aligned}$$

where

$$v_k = x_k - \frac{1}{L}\left(A^T(I-P_Q)Ax_k - \mu\varepsilon g(x_k)\right) \text{with } g(x_k) \in \partial\|x_k\|_r.$$

We summarize the algorithm as follows:

Majorized Penalty Algorithm:

1. **Initialization:** Let x_0 be given and set $L > \lambda_{\max(A^TA)}$.
2. For $k = 0, 1, \cdots$ find

$$x_{k+1} = sgn\left(x_k - \frac{1}{L}\left(A^T(I-P_Q)Ax_k - \mu\varepsilon g(x_k)\right)\right)\circ\max\left\{|v_k| - \frac{\mu}{L}, 0\right\} \tag{3.4}$$

End.

The following proposition contains the convergence result of this Penalty Algorithm.

Proposition 3.1 *Let (x_k) be the sequence generated by the Majorized Penalty Algorithm above. Then*

$$\frac{L}{2}\|x_k - x_{k+1}\|^2 \leq \tilde{f}(x_k) - \tilde{f}(x_{k+1}). \tag{3.5}$$

Furthermore, the sequence (x_k) is bounded and any cluster point is a stationary point of problem (3.1).

Proof. Since x_{k+1} minimizes $F(x, x_k)$, thanks to the first-order optimality condition we can write

$$0 \in A^T(I-P_Q)Ax_k + L(x_{k+1} - x_k) + \mu\|x_{k+1}\|_1 - \mu\varepsilon g(x_k), \tag{3.6}$$

$g(x_k)$ being a subgradient of $\|x\|_r$ at x_{k+1}. This combined with the definition of the subdifferential of $\|x\|_1$ at x_{k+1} gives

$$\begin{aligned}\mu\ \|x_k\|_1 - \mu\|x_{k+1}\|_1 &\geq \left\langle -A^T(I-P_Q)Ax_k - L(x_{k+1}-x_k) + \mu\varepsilon g(x_k), x_k - x_{k+1}\right\rangle \\ &= \left\langle -A^T(I-P_Q)Ax_k + \mu\varepsilon g(x_k), x_k - x_{k+1}\right\rangle + L\|x_{k+1}-x_k\|_2^2 \\ &= \left\langle A^T(I-P_Q)Ax_k + \mu\varepsilon g(x_k), x_{k+1} - x_k\right\rangle + L\|x_{k+1}-x_k\|_2^2.\end{aligned}$$

Hence

$$\mu\|x_{k+1}\|_1 - \mu\|x_k\|_1 + \left\langle A^T(I-P_Q)Ax_k - \mu\varepsilon g(x_k), x_k - x_{k+1}\right\rangle \leq -L\|x_{k+1}-x_k\|_2^2.$$

This together with the definition of F, for any $k \geq 1$, leads to

$$\begin{aligned}\tilde{f}(x_{k+1}) - \tilde{f}(x_k) \leq F(x_{k+1}, x_k) - \tilde{f}(x_k) &= \left\langle A^T(I-P_Q)Ax_k, x_{k+1}-x_k\right\rangle + \frac{L}{2}\|x_{k+1}-x_k\|_2^2 \\ &\quad + \mu(\|x_{k+1}\|_1 - \|x_k\|_1 - \varepsilon g(x_k), x_k - x_{k+1}\rangle) \\ &= \frac{L}{2}\|x_{k+1}-x_k\|_2^2 + \mu\|x_{k+1}\|_1 - \mu\|x_k\|_1 \\ &\quad + \left\langle A^T(I-P_Q)Ax_k - \mu\varepsilon g(x_k), x_{k+1}-x_k\right\rangle \\ &\leq \frac{L}{2}\|x_{k+1}-x_k\|_2^2 - L\|x_{k+1}-x_k\|_2^2.\end{aligned}$$

Consequently,

$$\tilde{f}(x_{k+1}) - \tilde{f}(x_k) \leq -\frac{L}{2}\|x_{k+1}-x_k\|_2^2. \tag{3.7}$$

Hence $\tilde{f}(x_{k+1}) \leq \tilde{f}(x_k)$ and thus the sequence $(\tilde{f}(x_k))$ is convergent since $\tilde{f}$ is a non-negative function. Furthermore, the sequence (x_k) is such that

$$\sum_{k=0}^{\infty}\|x_{k+1}-x_k\|^2 < +\infty.$$

Indeed, by summing (3.7) from $k=0$ to ∞, we obtain that

$$\frac{L}{2}\sum_{k=0}^{\infty}\|x_{k+1}-x_k\|_2^2 \leq \tilde{f}(x_0) - \lim_{k\to+\infty}\tilde{f}(x_k) \leq \tilde{f}(x_0) < +\infty.$$

Consequently, the sequence (x_k) is asymptotically regular, i.e., $\lim_{k\to+\infty}\|x_k - x_{k+1}\| = 0$. On the other hand, observe that the definition of $\tilde{f}$ for any $k \geq 1$, leads to

$$\mu(\|x_k\|_1 - \varepsilon\|x_k\|_r) \leq \frac{1}{2}\|(I-P_Q)Ax_k\|_2^2 + \mu(\|x_k\|_1 - \varepsilon\|x_k\|_r) = \tilde{f}(x_k) \leq \tilde{f}(x_0).$$

Since $\|x_k\|_1 \geq \|x_k\|_r$, we obtain that $\mu(1-\varepsilon)\|x_k\|_r \leq \tilde{f}(x_0)$. This implies that (x_k) is bounded since $0 < \varepsilon < 1$. To conclude, we prove that every cluster point of (x_k) is a stationary point of (3.1). Let x^* be a cluster point of (x_k), then $x^* = \lim_V x_{k_v}$, (x_{k_v}) being subsequence of (x_k). By passing to the limit in (3.6) along the subsequence (x_{k_v}) and in the light of the upper semicontinuity of (Clarke) subdifferentials, we obtain the desired result, namely

$$0 \in A^T(I - P_Q)Ax^* + \mu\partial\|x^*\|_1 - \mu\varepsilon g(x^*),$$

which is nothing else than the first-order optimality condition of (3.1). □

4. DCA algorithm

Now we turn our attention to a DC Algorithm (DCA), where the dual step at each iteration can be efficiently carried out due to the accessible subgradients of the largest- q-norm $\|\cdot\|_{\sigma_q}$ and $\|\cdot\|_r$ norm. Remember that to find critical points of $f := \phi - \psi$, the DCA consists in designing of sequences (x_k) and (y_k) by the following rules

$$\begin{cases} y_k \in \partial\psi(x_k); \\ x_{k+1} = argmin_{x\in\mathbb{R}^n}(\phi(x) - (\psi(x_k) + \langle y_k, x - x_k\rangle)). \end{cases} \tag{4.1}$$

Note that by the definition of subdifferential, we can write

$$\psi(x_{k+1}) \geq \psi(x_k) + \langle y_k, x_{k+1} - x_k\rangle.$$

Since x_{k+1} minimizes $\phi(x) - (\psi(x_k) + \langle y_k, x - x_k\rangle)$, we also have

$$\phi(x_{k+1}) - (\psi(x_k) + \langle y_k, x_{k+1} - x_k\rangle) \leq \phi(x_k) - \psi(x_k).$$

Combining the last inequalities, we obtain

$$f(x_k) = \phi(x_k) - \psi(x_k) \geq \phi(x_{k+1}) - (\psi(x_k) + \langle y_k, x_{k+1} - x_k\rangle) \geq f(x_{k+1}).$$

Therefore, the DCA leads to a monotonically decreasing sequence $(f(x_k))$ that converges as long as the objective function f is bounded below.

Now, we can decompose the objective function in (2.6) as follows

$$\min_x \left(f(x) := \left(\frac{1}{2}\| (I - P_Q)Ax\|^2 + \mu\|x\|_1 \right) - \left(\mu\varepsilon\|x\|_{\sigma_q}\right) \right), \tag{4.2}$$

where $\mu > 0$, $\varepsilon \in (0,1)$, here $\phi(x) = \frac{1}{2}\|(I - P_Q)Ax\|^2 + \mu\|x\|_1$ and $\psi(x) = \mu\varepsilon\|x\|_{\sigma_q}$.

At each iteration, DCA solves The convex subproblem defined by linearizing the concave term $-\varepsilon\|x\|_{\sigma_q}$ is solved by DCA at each iteration until a convergence condition is satisfied. More precisely, we have

$$\begin{cases} y_k \in \mu\varepsilon\,\partial\|x_k\|_{\sigma_q}; \\ x_{k+1} = argmin_{x\in\mathbb{R}^n}\left(\frac{1}{2}\|(I - P_Q)Ax\|^2 + \mu\|x\|_1 - \left(\mu\varepsilon\|x_k\|_{\sigma_q} + \langle y_k, x - x_k\rangle\right) \right). \end{cases} \tag{4.3}$$

Especially, if either the function ϕ or ψ is polyhedral, the DCA is said to be polyhedral and terminates in finite iterations [15]. Note that the our proposed DCA is polyhedral since the largest-q norm term $-\varepsilon\|x\|_{\sigma_q}$ can be expressed as a pointwise maximum of $2^q C_n^q$ linear functions, see [10]. On the other hand, the subdifferential of $\|x\|_{\sigma_q}$ at a point x_k is given in, see for example [19],

$$\partial\|x_k\|_{\sigma_q} arg\max_{y}\left(\sum_{i=1}^{n}|[x_k]_i|y_i : \sum_{i=1}^{n} y_i = q, 0 \le y_i \le 1, i = 1,\cdots,n\right), \tag{4.4}$$

that is

$$\partial\|x_k\|_{\sigma_q} = \left\{(y_1,\cdots,y_n) : y_{i_1} = \cdots = y_{i_q} = 1, y_{i_{q+1}} = 0 = \cdots = y_{i_n} = 0\right\},$$

where y_{i_j} denotes the element of y corresponding to x_{i_j} in the linear program (4.4). Observe that a subgradient $y \in \partial\|x_k\|_{\sigma_q}$ can be computed efficiently by first sorting the elements $|x_i|$ in decreasing order, namely $|x_{i_1}| \ge |x_{i_2}| \ge \cdots \ge |x_{i_q}|$. Then, assign 1 to y_i which corresponds to $x_{i_1}\cdots,x_{i_q}$.

To conclude, let us consider the following DC formulation of (3.1):

$$\min_{x}\left(f(x) := \left(\frac{1}{2}\|(I-P_Q)Ax\|^2 + \mu\|x\|_1\right) - (\mu\varepsilon\|x\|_r)\right), \tag{4.5}$$

where $r > 0$, $\varepsilon \in (0,1)$, here $\phi(x) = \frac{1}{2}\|(I-P_Q)Ax\|^2 + \mu\|x\|_1$ and $\psi(x) = \mu\varepsilon\|x\|_r$.

The subgradient $y \in \partial\|x_k\|_r$ is also available via the formula (3.2) and the DCA in this context take the following form

$$\begin{cases} y_k \in \mu\varepsilon\partial\|x_k\|_r; \\ x_{k+1} = argmin_{x\in\mathbb{R}^n}\left(\frac{1}{2}\|(I-P_Q)Ax\|^2 + \mu\|x\|_1 - (\mu\varepsilon\|x_k\|_r + \langle y_k, x - x_k\rangle)\right). \end{cases} \tag{4.6}$$

where

$$[y_k]_i = \begin{cases} \dfrac{sign([x_k]_i)|[x_k]_i|^{r-1}}{\|x_k\|_r^{r-1}} & \text{if } x_k \ne 0; \\ 0 & \text{otherwise.} \end{cases}$$

For the details of DCA convergence properties, see [15].

5. Concluding remarks

The focus of this paper is on Q-Lasso relying on two new DC-penalty methods instead of conventional methods such as l_1 or $l_1 - l_2$ minimization developed in [13,17] and [21]. Two iterative minimization methods based on the gradient proximal algorithm as well as the majored penalty algorithm are designed and their convergence to a stationary point is proved. Furthermore, by means of DC (difference of convex) Algorithm, two other algorithms are devised and their convergence results are also stated.

References

[1] M.A. Alghamdi, M. Ali Alghamdi, Naseer Shahzad, H.-K. Xu, Properties and iterative methods for the Q-Lasso, Abstr. Appl. Anal. (2013), Article ID 250943, 8 pages.

[2] H. Attouch, J. Bolte, B.F. Svaiter, Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods, Math. Program., Ser. A 137 (2013) 91–129.

[3] D.-P. Bertsekas, Nonlinear Programming, Athena Scientific, 1999.

[4] A.M. Bruckstein, D.L. Donoho, M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images, SIAM Rev. 51 (2009) 34–81.

[5] Y. Censor, T. Elfving, A multiprojection algorithm using Bregman projections in a product space, Numer. Algorithms 8 (1994) 221–239.

[6] R. Chartrand, Exact reconstruction of sparse signals via nonconvex minimization, IEEE Signal Process. Lett. 14 (2007) 707–710.

[7] S.S. Chen, D.L. Donoho, M.A. Saunders, Atomic decomposition by basis pursuit, SIAM J. Sci. Comput. 20 (1998) 33–61.

[8] D. Donoho, Compressed sensing, IEEE Trans. Inf. Theory 52 (2006) 1289–1306.

[9] G. Gasso, A. Rakotomamonjy, S. Canu, Recovering sparse signals with a certain family of nonconvex penalties and dc programming, IEEE Trans. Signal Process. 57 (12) (2009) 4686–4698.

[10] Jun-ya Gotoh, Akiko Takeda, Katsuya Tono, DC formulations and algorithms for sparse optimization problems, Math. Program. (2017) 1–36.

[11] R. Horst, N.V. Thoai, Dc programming: overview, J. Optim. Theory Appl. 103 (1999) 1–41.

[12] S. Ji, K.-F. Sze, Z. Zhou, A.M.-C. So, Y. Ye, Beyond convex relaxation: A polynomial-time non-convex optimization approach to network localization, in: Proceedings of the 32nd IEEE International Conference on Computer Communications (INFOCOM 2013), Torino, 2013.

[13] Y. Lou, M. Yan, Fast l_1-l_2 Minimization via a proximal operator, J. Sci. Comput. (2017) 1–19.

[14] A. Moudafi, A. Gibali, l_1-l_2 regularization of split feasibility problems, Numer. Algorithms (2017) 1–19, http://dx.doi.org/10.1007/s11075-017-0398-6.

[15] T. Pham Dinh, H.A. Le Thi, Convex analysis approach to D.C. programming: Theory, algorithms and applications, Acta Math. Vietnamica 22 (1) (1997) 289–355.

[16] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc., Ser. B 58 (1996) 267–288.

[17] P. Yin, Y. Lou, Q. He, J. Xin, Minimization of l_{1-2} for compressed sensing, SIAM J. Sci. Comput. 37 (2015) 536–563.

[18] Y. Wang, New improved penalty methods for sparse reconstruction based on difference of two norms, Technical Report (2013) 1–11.

[19] B. Wu, C. Ding, D.F. Sun, K.C. Toh, On the Moreau-Yoshida regularization of the vector k-norm related functions, SIAM J. Optim. 24 (2014) 766–794.

[20] Xu. Hong-Kun, Maryam A. Alghamdi, Naseer Shahzad, Regularization for the split feasibility problem, J. Nonlinear Convex Anal. 17 (3) (2015) 513–525.

[21] Z. Xu, X. Chang, F. Xu, H. Zhang, l_{1-2} regularization: a thresholding representation theory and a fast solver, IEEE Trans. Neural Networks Learn. Syst. 23 (2012) 1013–1027.

Corresponding author
Abdellatif Moudafi can be contacted at: abdellatif.moudafi@univ-amu.fr