Applied Analysis and Optimization
Volume 2, Number 2, 2018, 285–297

Yokohama Publishers
ISSN 2189-1664 Online Journal

A DC REGULARIZATION OF SPLIT MINIMIZATION PROBLEMS

ABDELLATIF MOUDAFI AND HONG-KUN XU

Abstract. Numerous problems in signal processing and imaging, statistical learning and data mining, or computer vision can be formulated as optimization problems which consist in minimizing a sum of convex functions, not necessarily differentiable, possibly composed with linear operators and that in turn can be transformed to split minimization problems (SMP), see for example [5]. Each function is typically either a data fidelity term or a regularization term enforcing some properties on the solution, see for example [8] and references therein. In the spirit of the idea developed for Split Feasibility Problems (SFP) in [1] and [26], we introduced a non convex regularization for split minimization problems, proposed three algorithms and proved their convergence properties. The first algorithm is based on the DCA introduced by Pham Dinh Tao, the second one is nothing else than the celebrate forward-backward algorithm. An algorithm based on Mine-Fukushima method is also stated. It is worth mentioning that as special case, we recovered the SFP which model a number of applied problems arising from signal/image processing and specially optimization problems for intensity-modulated radiation therapy (IMRT) treatment planning, see for example [4].

1. Introduction and preliminaries

Recent developments in science and technology have caused a revolution in data processing, as large datasets are becoming increasingly available and important. To meet the need in big data area, the field of compressive sensing (CS) [11] is rapidly blooming. The process of CS consists of encoding and decoding. 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 say the signal $x \in I\!R^n$ can be compressed. The process of decoding is to recover x from b with an additional assumption that x is sparse. It can be expressed as an optimization problem,

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

with $\|\cdot\|_0$ being the l_0 norm, which counts the number of nonzero entries of x; that is

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

where $|\cdot|$ denotes here the cardinality, i.e., the number of elements of a set. So minimizing the l_0 norm is equivalent to finding the sparsest solution. One of the biggest obstacles in CS is solving the decoding problem above, as l_0 minimization is NP-hard. A popular approach is to replace l_0 by a convex l_1-norm, which often gives

2010 *Mathematics Subject Classification.* Primary, 49J53, 65K10; Secondary, 49M37, 90C25.
Key words and phrases. Split minimization, split feasibility, soft-thresholding, regularization, DCA algorithm, forward-backward iterations, proximity mapping.

a satisfactory sparse solution. This l_1 heuristic has been applied in many different fields such as geology and geophysics, spectroscopy, and ultrasound imaging.
Recently, there has been an increase in applying nonconvex metrics as alternative approaches to l_1. In particular, the nonconvex metric l_p for $p \in (0,1)$ in [7] can be regarded as a continuation strategy to approximate l_0 as $p \to 0$. The optimization strategies include iterative reweighting [7] and half thresholding [27], the scale-invariant l_1, formulated as the ratio of l_1 and l_2, was discussed in [13]. Other nonconvex l_1 variants include transformed l_1, sorted l_1 and capped l_1. However, due to errors of measurements, the constraint $Ax = b$ is actually inexact; It turns out that the so called Lasso problem of Tibshirani [25] (which is well-known to be equivalent to the basic pursuit (BP) of Chen et al. [9]) is reformulated as

$$\min_{x \in I\!R^n} \|x\|_1 \quad \text{subject to } \|Ax - b\|_p \leq \varepsilon,$$

where $\varepsilon > 0$ is the tolerance level of errors and p is often $1, 2$ or ∞. It is noticed in [1] that if we let $Q := B_\varepsilon(b)$, the closed ball in $I\!R^n$ with center b and radius ε, then the later is rewritten as

$$\min_{x \in I\!R^n} \|x\|_1 \quad \text{subject to } Ax \in Q.$$

With Q a nonempty closed convex set of $I\!R^m$ and P_Q the projection from $I\!R^m$ onto Q and since that the constraint is equivalent to the condition $Ax - P_Q(Ax) = 0$, this leads to the following equivalent Lagrangian formulation

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

with $\gamma > 0$ a Lagrangian multiplier. A connection is also made in [1] with the so-called split feasibility problem [5] which is stated as finding x verifying

$$x \in C, \quad Ax \in Q, \tag{1.3}$$

where C and Q are closed convex subsets of $I\!R^n$ and $I\!R^m$, respectively. An equivalent minimization formulation of (1.3) is

$$\min_{x \in C} \frac{1}{2}\|(I - P_Q)Ax\|_2^2. \tag{1.4}$$

Its l_1 regularization reads as

$$\min_{x \in C} \frac{1}{2}\|(I - P_Q)Ax\|_2^2 + \gamma\|x\|_1, \tag{1.5}$$

where $\gamma > 0$ is a regularization parameter.
Note that it reduces to the Lagrangian formulation above when $C = I\!R^n$.

This convex relaxation attracts considerable attention, see for example [1] and references there in. In [18] we studied a non-convex but Lipschitz continuous metric l_{1-2} for SFP. As illustrated in [15], the level curves of l_{1-2} are closer to l_0 than those of l_1 and it is demonstrated in a series of papers [15, 27] that the difference of the l_1 and l_2 norms, denoted as l_{1-2}, outperforms l_1 and l_p in terms of promoting sparsity when the sensing matrix A is highly coherent, this motivated us to consider in [18] the following nonconvex l_{1-2} regularization for split feasibility problem,

$$\min_{x \in C} \big(\frac{1}{2}\|(I - P_Q)Ax\|_2^2 + \gamma(\|x\|_1 - \|x\|_2)\big), \tag{1.6}$$

and propose three algorithms with numerical experiments.

In what follows we are interested in the minimization problem:

$$\min_{x\in\mathbb{R}^n} \big(f(x) + g_\lambda(Ax) + \gamma(\|x\|_1 - \|x\|_2)\big), \tag{1.7}$$

which is more general and we will focus our attention to the algorithmic aspect by devising three methods. The first algorithm uses the DCA which is a descent method without line search introduced by Tao and An [24] for minimizing a function f which is the difference of two lower semicontinuous proper convex functions g and h on the space $\mathbb{R}^n$. The second one is based on the gradient proximal method to solve the problem [24] by full splitting, that is, at every iteration, the only operations involved are evaluations of the proximal mappings of g and r separately. An algorithm which relies on the Mine-Fukushima method for minimizing a sum of two functions is also stated.

Problem (1.7) is the nonconvex l_{1-2} regularization for the following split minimization problem

$$\min_{x\in\mathbb{R}^n} \{f(x) + g_\lambda(Ax)\}, \tag{1.8}$$

where $f : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}, g : \mathbb{R}^m \to \mathbb{R}\cup\{+\infty\}$ are two proper, convex, lower semicontinuous functions and $A : \mathbb{R}^n \to \mathbb{R}^m$ a bounded linear operator,

$$g_\lambda(y) = \inf_{u\in\mathbb{R}^m} \left\{ g(u) + \frac{1}{2\lambda}\|u - y\|^2 \right\}$$

stands for the Moreau-Yosida approximate of the function g of parameter λ.

Note also that (1.8) is a partial regularization of the so-called graph form type problem, see [21]. A wide of convex optimization problems can be expressed in this form, including cone programs and a wide variety of regularized loss minimization problems from statistics, like logistic regression, the support vector machine, the lasso and the intensity modulated radiation treatment planning. Observe also that by taking $\lambda = 1$, $f = i_C$ and $g = i_Q$ the indicator functions of two nonempty, closed and convex sets C, Q of $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, (1.8) reduces to (1.4) when $C \cap A^{-1}(Q) \neq \emptyset$.

Finally, the differentiability of the Yosida-approximate g_λ ensures the additivity of the subdifferentials and we can write

$$\partial(f(x) + g_\lambda(Ax)) = \partial f(x) + A^t\nabla g_\lambda(Ax).$$

This implies that the optimality condition of (1.8) can then be written as

$$0 \in \partial f(x) + A^t\nabla g_\lambda(Ax), \tag{1.9}$$

and that of (1.7), for any nonzero point x, is given by

$$0 \in \partial f(x) + A^t\nabla g_\lambda(Ax) + \gamma(\partial\|x\|_1 - \frac{x}{\|x\|_2}). \tag{1.10}$$

2. Computational approaches

We use $\|x\|_p$ to denote the p-norm of a vector x, where $1 \le p \le \infty$, and $\|\cdot\|$ is reserved exclusively for the Euclidean 2-norm $\|\cdot\|_2$.

2.1. **DCA.** First, recall that the partial differential of a convex function h is defined as

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

It is easily seen that

$$\partial \frac{1}{2}\|Ax - b\|^2 = \nabla \frac{1}{2}\|Ax - b\|^2 = A^t(Ax - b), \tag{2.2}$$

and

$$(\partial \|x\|_1)_i = \begin{cases} \text{sgn}(x_i) & \text{if } \ x_i \neq 0; \\ \text{any element of } [-1, 1] & \text{if } \ x_i = 0. \end{cases} \tag{2.3}$$

The *characteristic function* of a set $C \subseteq I\!R^n$ is defined as

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

Such function is convenient to enforce hard constraints on the solution. Moreover, the *normal cone* to C at $x \in C$, denoted by $N_C(x)$ is defined

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

A known relation between the above definitions is that $\partial i_C = N_C$.

Remember that given an initial point x_0, the DCA seeks critical points of $f := g - h$ by constructing two sequences (x_k) and (y_k) by the following rules

$$\begin{cases} y_k \in \partial h(x_k); \\ x_{k+1} = \arg\min_{x \in I\!R^n} \big(g(x) - (h(x_k) + \langle y_k, x - x_k\rangle)\big). \end{cases} \tag{2.6}$$

Note that by the definition of subdifferential, we can write

$$h(x_{k+1}) \geq h(x_k) + \langle y_k, x_{k+1} - x_k\rangle. \tag{2.7}$$

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

$$g(x_{k+1}) - (h(x_k) + \langle y_k, x_{k+1} - x_k\rangle) \leq g(x_k) - h(x_k). \tag{2.8}$$

Combining the last inequalities, we obtain

$$f(x_k) = g(x_k) - h(x_k) \geq g(x_{k+1}) - (h(x_k) + \langle y_k, x_{k+1} - x_k\rangle) \geq f(x_{k+1}). \tag{2.9}$$

Therefore, the DCA provides a monotonically decreasing sequence $\{f(x_k)\}$ which converges provided that the objective function f is bounded below.

The objective function in (1.7) has the following DC decomposition

$$\min_{x \in I\!R^n} \big((f(x) + g_\lambda(Ax) + \gamma\|x\|_1) - \gamma\|x\|_2\big). \tag{2.10}$$

Observe that $\|x\|_2$ is differentiable with gradient $x/\|x\|_2$ for any $x \neq 0$ and that $0 \in \partial\|\cdot\|_2(0)$ which leads to the following iterates

$$x_{k+1} = \begin{cases} \arg\min_{x \in I\!R^n} f(x) + g_\lambda(Ax) + \gamma\|x\|_1 & \text{if } x_k = 0 \\ \arg\min_{x \in I\!R^n} f(x) + g_\lambda(Ax) + \gamma\|x\|_1 - \langle x, \gamma\frac{x_k}{\|x_k\|_2}\rangle & \text{if } x_k \neq 0. \end{cases} \tag{2.11}$$

Now, we define for all $\gamma > 0$, Γ by

$$\Gamma(x) = f(x) + g_\lambda(Ax) + \gamma(\|x\|_1 - \|x\|_2). \tag{2.12}$$

We are in a position to prove the following convergence properties of Algorithm 2.11.

Theorem 2.1. *Let (x_k) be the sequence generated by Algorithm 2.11 and assume that for all $\gamma > 0$ we have that $\lim_{\|x\|_2 \to +\infty} \Gamma(x) = +\infty$. Γ is therefore coercive in the sense that its levels sets are bounded, namely $\{x \in I\!R^n; \Gamma(x) \leq \Gamma(x_0)\}$ is bounded for any $x_0 \in I\!R^n$. Then,*

(i) *(x_k) is bounded.*
(ii) *Every nonzero limit point x^* of the sequence (x_k) is a stationary point of (1.7), namely*

$$0 \in A^t \nabla g_\lambda(Ax^*) + \partial f(x^*) + \gamma \left(\partial \|x^*\|_1 - \frac{x^*}{\|x^*\|_2} \right). \tag{2.13}$$

Proof. A simple computation which uses the fact that $\|a\|^2 - \|b\|^2 = \|a - b\|^2 + 2\langle b, a - b\rangle$, gives

$$\begin{aligned} \Gamma(x_k) - \Gamma(x_{k+1}) &= f(x_k) - f(x_{k+1}) + g_\lambda(Ax_k) - g_\lambda(Ax_{k+1}) \\ &\quad + \gamma(\|x_k\|_1 - \|x_{k+1}\|_1 - \|x_k\|_2 + \|x_{k+1}\|_2). \end{aligned} \tag{2.14}$$

The first-order optimality condition at x_{k+1} as the solution of the problem (2.11) and the fact that $\partial(\|\cdot\|_1 + f)(x) = \partial\|x\|_1 + \partial f(x)$ lead to

$$A^t \nabla g_\lambda(Ax_{k+1}) + \gamma(w_{k+1} - y_k) + p_{k+1} = 0, \tag{2.15}$$

where $y_k \in \partial\|x_k\|_2, w_{k+1} \in \partial\|x_{k+1}\|_1, p_{k+1} \in \partial f(x_{k+1})$ which combined with the fact that $\langle w_{k+1}, x_{k+1}\rangle = \|x_{k+1}\|_1$ gives

$$\begin{aligned} \langle \nabla g_\lambda(Ax_{k+1}), Ax_k - Ax_{k+1}\rangle + \gamma(\langle w_{k+1}, x_k\rangle - \|x_{k+1}\|_1 + \langle y_k, x_{k+1} - x_k\rangle) \\ - \langle p_{k+1}, x_{k+1} - x_k\rangle = 0. \end{aligned} \tag{2.16}$$

Combining (2.14) and (2.16), we can write $\Gamma(x_k) - \Gamma(x_{k+1})$ as

$$\begin{aligned} \Gamma(x_k) - \Gamma(x_{k+1}) &= f(x_k) - f(x_{k+1}) - \langle p_{k+1}, x_{k+1} - x_k\rangle \\ &+ g_\lambda(Ax_k) - g_\lambda(Ax_{k+1}) - \langle \nabla g_\lambda(Ax_{k+1}), Ax_k - Ax_{k+1}\rangle \\ &+ \gamma\big(\|x_{k+1}\|_2 - \|x_k\|_2 - \langle y_k, x_{k+1} - x_k\rangle\big) \\ &+ \gamma(\|x_k\|_1 - \langle w_{k+1}, x_k\rangle). \end{aligned} \tag{2.17}$$

Since $p_{k+1} \in \partial f(x_{k+1})$, using the subdifferentiability of f, we get

$$f(x_k) \geq f(x_{k+1}) + \langle p_{k+1}, x_k - x_{k+1}\rangle. \tag{2.18}$$

The $\frac{1}{\lambda}$-Lipschitz continuity of the gradient of g_λ assures that

$$\begin{aligned} g_\lambda(Ax_k) \geq g_\lambda(Ax_{k+1}) + \langle \nabla g_\lambda(Ax_{k+1}), Ax_k - Ax_{k+1}\rangle \\ + \frac{\lambda}{2}\|\nabla g_\lambda(Ax_{k+1}) - \nabla g_\lambda(Ax_k)\|^2. \end{aligned} \tag{2.19}$$

Since $y_k \in \partial\|x_k\|_2$, we obtain

$$\|x_{k+1}\|_2 - \|x_k\|_2 - \langle y_k, x_{k+1} - x_k\rangle \geq 0 \tag{2.20}$$

and also since and $\|w_{k+1}\|_\infty \leq 1$,

$$\|x_k\|_1 - \langle w_{k+1}, x_k\rangle \geq 0. \tag{2.21}$$

Substituting (2.18)-(2.21) into (2.17) yields

$$\begin{aligned}\Gamma(x_k) - \Gamma(x_{k+1}) \geq \frac{\lambda}{2}\|\nabla g_\lambda(Ax_k) - \nabla g_\lambda(Ax_{k+1})\|^2 \\ + \gamma(\|x_{k+1}\|_2 - \|x_k\|_2 - \langle y_k, x_{k+1} - x_k\rangle) \geq 0.\end{aligned} \tag{2.22}$$

This ensures that the sequence $(\Gamma(x_k))$ is monotonically decreasing, which in turn ensures that the sequence $(x_k) \subset \{x \in I\!R^n, \Gamma(x) \leq \Gamma(x_0)\}$ which is bounded since Γ is coercive.

If $x_1 = x_0 = 0$, we then stop the algorithm producing the solution $x^* = 0$. Otherwise, it follows from (2.22) that

$$\Gamma(x_0) - \Gamma(x_1) \geq \gamma\|x_1\|_2 > 0, \tag{2.23}$$

so $x_k \neq 0$ for all $k \geq 1$. Since $(\Gamma(x_k))$ is convergent, substituting $y_k = \frac{x_k}{\|x_k\|_2}$ into (2.22), leads to

$$\lim_{k\to+\infty} \|\nabla g_\lambda(Ax_k) - \nabla g_\lambda(Ax_{k+1})\| = 0 \tag{2.24}$$

and

$$\lim_{k\to+\infty} (\|x_k\|_2 \cdot \|x_{k+1}\|_2 - \langle x_k, x_{k+1}\rangle) = 0. \tag{2.25}$$

Now, let (x_{k_ν}) be a subsequence of (x_k) converging to $x^* \neq 0$. With no loss of generality, we may also assume that $(x_{k_\nu - 1})$ converges to $\hat{x}$. It then turns out from (2.25) that

$$\|x^*\|_2 \cdot \|\hat{x}\|_2 = \langle x^*, \hat{x}\rangle.$$

This implies that $x^* = \mu\hat{x}$ for some $\mu > 0$.

On the other hand, the optimality condition at the k_ν-th step of Algorithm (2.11) reads

$$-\big(A^t\nabla g_\lambda(Ax_{k_\nu}) - \gamma\frac{x_{k_\nu-1}}{\|x_{k_\nu-1}\|_2}\big) \in \gamma\partial\|x_{k_\nu}\|_1 + \partial f(x_{k_\nu}). \tag{2.26}$$

Taking the limit as $\nu \to \infty$ in (2.26) and observing the fact that the operator $\gamma\partial\|\cdot\|_1 + \partial f$ is maximal monotone, we obtain

$$-\big(A^t\nabla g_\lambda(Ax^*) - \gamma\frac{\hat{x}}{\|\hat{x}\|_2}\big) \in \gamma\partial\|x^*\|_1 + \partial f(x^*).$$

Upon substituting $x^* = \mu\hat{x}$ into the last relation, we immediately get

$$-\big(A^t\nabla g_\lambda(Ax^*) - \gamma\frac{x^*}{\|x^*\|_2}\big) \in \gamma\partial\|x^*\|_1 + \partial f(x^*).$$

That is, (2.13) holds and x^* is a stationary point. □

Remark 2.2. It should be noticed that the coerciveness assumption of the objective function was also assumed in [15] and it is worth mentioning that this hypothesis is valid in numerous settings and most often holds true in real world applications. This is the case, in instance, for:

(i) **Subset selection in regression**. More precisely, let $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, by setting $f \equiv 0$ and $g(x) = i_{\{b\}}$ so that $g_\lambda(Ax) = \frac{1}{2\lambda}\|Ax - b\|_2^2$, the problem (1.8) is least square estimation of a linear model equipped with variable selection. Such data-fidelity terms are currently used in denoising, in deblurring, and in numerous inverse problems.

(ii) **Split feasibility problems** which modeled, among others, *Intensity modulated radiation treatment planning* and that can be recovered by setting $f = i_C$, $g = i_Q$ so that $g_\lambda(Ax) = \frac{1}{2\lambda}\|(I - P_Q)Ax\|_2^2$ with C, Q being two closed convex sets.

(iii) **Support vector machine (SVM)** with feature selection, see for example [20] and also in **Sparse portfolio selection**, see for instance, [30]. Indeed, let $V \in \mathbb{R}^{n\times n}$ be a covariance matrix, $r \in \mathbb{R}^n$ a mean return vector, $l, u \in \mathbb{R}^n$, and $\tau \in \mathbb{R}$. When $f(x) = x^tVx$ and $S = \{x \in \mathbb{R}^n : r^tx \geq \tau, \mathbf{1}^tx = 1, l \leq x \leq u\}$ the problem (1.8) is a sparse portfolio selection and the coerciveness assumption holds true.

(iv) **Huber M-estimator problem**. Now, having in mind that the support of x is defined by $\mathrm{supp}(x) = \{1 \leq x \leq n; x_i \neq 0\}$ and that $\|x\|_0 = |\mathrm{supp}(x)|$ is the cardinality of $\mathrm{supp}(x)$ and remembering that for all $x \neq 0$, we have $\|x\|_1 - \|x\|_2 \geq 0$ and that $\|x\|_1 - \|x\|_2 = 0 \Leftrightarrow \|x\|_0 = 1$, we can get other cases of coerciveness of Γ by particularizing f and g (since for λ small enough, g and its Yosida approximate, g_λ, have the same asymptotical behavior, since g_λ convergences to g in the epi-convergence sense, see fo instance [23]). We can easily verify that the coerciveness assumption is again valid, for example, when f is a non negative function and g is positively homogeneous of degree 1. An interesting case is the absolute value function $|\cdot|$. Its Yosida approximate is given by

$$|x|_\lambda = \begin{cases} \frac{1}{2\lambda}x^2 & \text{if } |x| \leq \lambda; \\ |x| - \frac{1}{2}\lambda & \text{otherwise.} \end{cases}$$

This is clearly equal, up to a scaling factor λ, to the so-called *Huber's M-cost function*

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \leq \lambda; \\ \lambda|x| - \frac{1}{2}\lambda^2 & \text{otherwise.} \end{cases}$$

in the context of robust linear estimation theory, see [14]. The Huber's M-cost function has been used in *M-estimator problems* which is known as a robust alternative to the *Least Squares estimator* that is unfortunately sensitive against occurrence of outliers in the ill-conditioned linear regression systems. The Huber's M-cost function has also been used in many inverse problems as an excellent robust convex penalty function that grows linearly for x far from zero, hence it achieves least sensitivity to large outliers of large residual, see [28] and references therein.

Remark 2.3. Each DCA iteration requires solving an l_1-regularized split feasibility subproblem of the form

$$\min_{x\in\mathbb{R}^n} (f(x) + g_\lambda(x) + \langle x, v\rangle + \gamma\|x\|_1),$$

where $v \in I\!R^n$ is a constant vector. This problem can be solved by the two split proximal algorithms (coupling forward-backward and the Douglas-Rachford algorithms) proposed in [8], [17] and also by the alternating direction method of multipliers (ADMM) following the analysis developed in [29] for the special case where $f \equiv 0$ and $Q = i_{\{b\}}$.

2.2. **Forward-backward splitting algorithm.** To begin with, recall that the proximal mapping of a proper, convex and lower semicontinuous function φ of parameter $\lambda > 0$ is defined by

$$\text{prox}_{\gamma\varphi}(x) := \arg\min_{v \in I\!R^n} \{\varphi(v) + \frac{1}{2\gamma}\|v - x\|^2\},\ x \in I\!R^n, \tag{2.27}$$

and that it has closed-form expression in some important cases. For example, if $\varphi = \|\cdot\|_1$, then for $x \in I\!R^n$

$$\text{prox}_{\gamma\|\cdot\|_1}(x) = (\text{prox}_{\gamma|\cdot|}(x_1), \cdots, \text{prox}_{\gamma|\cdot|}(x_n)), \tag{2.28}$$

where $\text{prox}_{\gamma|\cdot|}(x_k) = \text{sgn}(x_k)\max\{|x_k| - \gamma, 0\}$.

If $\varphi = i_C$, we have

$$\text{prox}_{\gamma\varphi}(x) = \text{proj}_C(x) := \arg\min_{z \in C} \|x - z\|. \tag{2.29}$$

For the sake of simplicity and clarity, we set in what follows $f \equiv 0$. Observe that the minimization problem (1.7) can be written as

$$\min_{x \in I\!R^n} g_\lambda(Ax) + \gamma(\|x\|_1 - \|x\|_2). \tag{2.30}$$

It is worth mentioning that when $f \not\equiv 0$, this requires to compute the proximal operator of a sum, namely $\text{prox}_{f+\gamma_k(\|\cdot\|_1 - \|\cdot\|_2)}$, which may be performed with Douglas-Rachford iterations in the spirit of the analysis developed in [8] and [17].

A closed-form solution of $\text{prox}_{\|x\|_1 - \|x\|_2}$ was proposed in [15], in particular, we have the following lemma.

Lemma 2.4. *Given $y \in I\!R^n$, $\gamma > 0$ and setting $r(x) = \|\cdot\|_1 - \|\cdot\|_2$, we have*

(i) *When $\gamma < \|y\|_\infty$, then*

$$\text{prox}_{\gamma r}(y) = \frac{\gamma + \|\text{prox}_{\gamma\|\cdot\|_1}(y)\|_2}{\|\text{prox}_{\gamma\|\cdot\|_1}(y)\|_2}\text{prox}_{\gamma\|\cdot\|_1}(y). \tag{2.31}$$

(ii) *When $\gamma = \|y\|_\infty$, then $x^* \in \text{prox}_{\gamma r}(y)$ if and only if it satisfies $x_i^* = 0$ if $|y_i| < \gamma, \|x^*\|_2 = \gamma$ and $x_i^* y_i \geq 0$ for all i.*

(iii) *When $\gamma > \|y\|_\infty$, then $x^* \in \text{prox}_{\gamma r}(y)$ if and only if it is a 1-sparse vector satisfying $x_i^* = 0$ if $|y_i| < \|y\|_\infty, \|x^*\|_2 = \|y\|_\infty$ and $x_i^* y_i \geq 0$ for all i.*

By setting $l(x) = g_\lambda(Ax)$, one has $\nabla g_\lambda(Ax) = \frac{1}{\lambda}A^t(I - \text{prox}_{\lambda g})(Ax)$, the forward-backward splitting algorithm can be expressed as follows

$$x_{k+1} \in \text{prox}_{\gamma r}\left(x_k - \frac{\gamma}{\lambda}A^t(I - \text{prox}_{\lambda g})Ax_k)\right). \tag{2.32}$$

Since the two assumptions of [15, Theorem 3] are satisfied, namely the coerciveness of the objective function and differentiability of the function l with Lipschitz-continuity of its gradient, a direct application of this Theorem leads to the following convergence result:

Proposition 2.5. *If $\gamma < \frac{\lambda}{\|A\|^2}$, then the objective values are decreasing and there exists a subsequence of (x_k) generated by (2.29) that converges to a stationary point of (2.27). Furthermore, any limit point of (x_k) is a stationary point of (2.27).*

2.3. **Mine-Fukushima Algorithm.** At this stage, we would like to mention that in the case where f is strictly convex and that we can generate from an initial point x_0 a sequence x_k such that $x_k \neq 0$ for all $k \in I\!N$, then the Algorithm introduced by Mine-Fukushima in [16] is applicable. Indeed, problem (1.6) can be written as

$$\min_{x \in I\!R^n} (h(x) := \phi(x) + \psi(x)), \tag{2.33}$$

with $\phi(x) = g_\lambda(Ax) - \gamma\|x\|_2$ and $\psi(x) = \gamma\|x\|_1 + f(x)$. Observe that in this case, we have for $x \neq 0$, that

$$\nabla\phi(x) = A^t(\nabla g_\lambda)Ax - \gamma\frac{x}{\|x\|_2} \text{ and } \partial\psi(x) = \partial\|x\|_1 + \partial f(x).$$

Algorithm 2.1 of [16] takes the following form:

Algorithm:

Step 1. Let x_0 be any initial point. Set $k = 0$, and go to step 2.

Step 2. If $-\nabla\phi(x_k) \in \partial\psi(x_k)$, then stop; otherwise, go to Step 3.

Step 3. Find a minimum $\tilde{x}_k$ of

$$\min_{x \in I\!R^n} \big(\langle x, A^t(\nabla g_\lambda)Ax_k - \gamma\frac{x_k}{\|x_k\|_2}\rangle + \gamma\|x\|_1 + f(x)\big), \tag{2.34}$$

and go to Step 4.

Step 4. Find

$$x_{k+1} = \alpha_k\tilde{x}_k + (1 - \alpha_k)x_k, \tag{2.35}$$

such that $\alpha_k \geq 0$ and

$$\phi(x_{k+1}) \leq \phi(\alpha\tilde{x}_k + (1 - \alpha)x_k) \text{ for all } \alpha \geq 0.$$

Set $k = k + 1$, and go to Step 2.

Observe that solving (2.31) in Step 3 is equivalent to finding $\tilde{x}_k$ such that $-\nabla f(x_k) \in \partial g(\tilde{x}_k)$.

Since h is coercive in many interesting cases, a direct application of [16, Theorem 3.1] yields the result below.

Proposition 2.6. *The sequence (x_k) generated by the Algorithm above contains a subsequence which converges to a critical point x^* of (1.7), namely*

$$-A^t\nabla g_\lambda(Ax^*) - \gamma\frac{x^*}{\|x^*\|_2} \in \partial\gamma\|x^*\|_1 + \partial f(x^*).$$

Remark 2.7. The assumption of strict convexity on f can be removed by applying the following process: for some $\mu > 0$ consider the following decomposition of the objective function h: $h = \tilde{\phi}(x) + \tilde{\psi}$ with $\tilde{\phi}(x) = \phi(x) - \mu\frac{\|x\|_2^2}{2}$ and ψ by $\tilde{\psi}(x) = \psi(x) + \mu\frac{\|x\|_2^2}{2}$. Relation (2.31) becomes

$$\min_{x\in C}\big(\langle x, A^t\nabla g_\lambda(Ax_k) + \mu x_k - \gamma\frac{x_k}{\|x_k\|_2}\rangle + \gamma\|x\|_1 + f(x) + \mu\frac{\|x\|_2^2}{2}\big). \tag{2.36}$$

3. Split Feasibility Problems

We focus on the split feasibility problem ([5, 2, 3]) obtained by taking $f = i_C$, $g = i_Q$ the indicator functions of two nonempty closed convex sets C, Q of $\mathbb{R}^n$, $\mathbb{R}^m$, respectively, Indeed the problem (1.8) reduces to

$$\min_{x\in\mathbb{R}^n}\{i_C(x) + (i_Q)_\lambda(Ax)\} \quad \Leftrightarrow \quad \min_{x\in C}\{\frac{1}{2\lambda}\|(I - P_Q)(Ax)\|^2\},$$

which, when $C \cap A^{-1}(Q) \neq \emptyset$, is equivalent to the split feasibility problem (see [6] for an implicit version), namely

$$x \in C \text{ such that } Ax \in Q.$$

This problem was used for solving an inverse problem in radiation therapy treatment planning [5] and has been well studied both theoretically and practically, see for example [1, 4, 22] and the references therein.

Having in mind that in this special case the proximal mapping of g is nothing by the orthogonal projection on the closed convex set Q, Algorithm (2.11) with $\lambda = 1$ and $f \equiv 0$ reduces to

$$x_{k+1} = \begin{cases} \arg\min_{x\in C}\frac{1}{2}\|(I - P_Q)Ax\|_2^2 + \gamma\|x\|_1 & \text{if } x_k = 0 \\ \arg\min_{x\in C}\frac{1}{2}\|(I - P_Q)Ax\|_2^2 + \gamma\|x\|_1 - \langle x, \gamma\frac{x_k}{\|x_k\|_2}\rangle & \text{if } x_k \neq 0 \end{cases} \tag{3.1}$$

which is exactly the algorithm studied in [18]. In this setting Γ being coercive, we obtain

Proposition 3.1. *Let (x_k) be the sequence generated by Algorithm (3.1), we have*

(i) *For all $\gamma > 0$ we have that $\lim_{\|x\|_2\to+\infty}\Gamma(x) = +\infty$. Γ is therefore coercive in the sense that its levels sets are bounded, namely $\{x \in \mathbb{R}^n; \Gamma(x) \leq \Gamma(x_0)\}$ is bounded for any $x_0 \in \mathbb{R}^n$.*

(ii) *(x_k) is bounded.*

(iii) *Any nonzero limit point x^* of the sequence (x_k) is a stationary point of (1.7), namely*

$$0 \in A^t(I - P_Q)Ax^* + \gamma\big(\partial\|x^*\|_1 - \frac{x^*}{\|x^*\|_2}\big). \tag{3.2}$$

The proof follows directly from applying Theorem 2.1. We hence recover the main convergence result, [18, Proposition 2.1], of DCA presented in [18] without the asymptotic regularity assumption. It is worth mentioning that we also recover [29, Proposition 3.1] by taking $Q = \{b\}$.

Concerning the proximal splitting algorithm, we have that $l(x) = \frac{1}{2\lambda}\|(I-P_Q)Ax\|_2^2$ and hence the forward-backward splitting algorithm reduces to

$$x_{k+1} \in \operatorname{prox}_{\lambda r}\left(x_k - \frac{\gamma}{\lambda}A^t(I-P_Q)A(x_k)\right), \tag{3.3}$$

and we recover its convergence properties obtained in [18, Proposition 2.4] directly from Proposition 2.3 above. This is also the case for the Mine-Fukushima Algorithm convergence property, namely [18, Proposition 2.5] which follows directly from Proposition 2.4. Step 3, in this case, reads as: Find a minimum $\tilde{x}_k$ of

$$\min_{x\in C}\left(\left\langle x, A^t(I-P_Q)Ax_k - \gamma\frac{x_k}{\|x_k\|_2}\right\rangle + \gamma\|x\|_1\right), \tag{3.4}$$

and go to Step 4.

4. Conclusion

The main purpose of this paper is to investigate split minimization problems under a nonconvex Lipschitz continuous metric instead of conventional methods such as l_1 or l_1/l_2 minimization developed for example in [1], to present an iterative minimization method based on DCA introduced by Tao et al. for DC optimization and also to analyze its convergence to a stationary point. Furthermore, relying on a proximal operator for l_{1-2} for minimizing the sum of a convex function and a differentiable one, an other algorithm is presented and its convergence property is stated. Moreover, an additional algorithm based on Mine-Fukushima method with its convergence result is also provided. It is worth mentioning that our results extend and unify in a more general setting the corresponding main results presented in [29] and [18]. Effectiveness of the algorithms was illustrated by numerical experiments for split feasibility problems in [18]. Moreover, it was shown that l_{1-2} is always better than l_1, and is better than l_p for highly coherent matrices in [29], Proximal operator can accelerate the minimization, but it tends to obtain a suboptimal solution, see [15] in which it is mentioned that in general, nonconvex methods have better empirical performance compared to convex ones, but lack of provable grounds.
Finally, we would like to emphasize that much attention has been paid very recently to DC approaches, for instance, for optimization problem having a nonconvex constraint called *the cardinality constrain*:

$$\|x\|_0 \leq k, \text{ where } x \in \mathbb{R}^n \text{ and } k \in \{1, \cdots, n\}.$$

Due to the nonconvexity and discontinuity of the l_0-norm, it is well-known that the resulting optimization problem is intractable (due to NP-hardness of the l_0-minimization over a linear system). An alternative approach was proposed in [12] by rewriting the cardinality constraint as $\|x\|_1 - \|x\|_k$, with $\|x\|_k$ being the largest-k norm. One advantage of the use of the largest-k norm representation is that its subgradient can be efficiently computed, which can make DCAs efficient. More interestingly, this fact motivates to develop a soft thresholding technique, which is popular in the context of proximal methods [10], and thus allows to use a closed-form solution of the DCA subproblem. We refer to the interesting paper [12], which deserves to be better known in the community of applied nonlinear analysis and

which is at the origin of new penalty methods for Q-Lasso, based on the difference of two norms, that will be proposed by the first author in a forthcoming paper.

References

[1] M. A. Alghamdi, M. A. Alghamdi, N. Shahzad and H. K. Xu, *Properties and iterative methods for the Q-lasso*, Abstract Applied Analysis (2013). Article ID 250943, 8 pages.

[2] C. Byrne, *Iterative oblique projection onto convex sets and the split feasibility problem*, Inverse Problems **18** (2002), 441–453.

[3] C. Byrne, *A unified treatment of some iterative algorithms in signal processing and image reconstruction*, Inverse Problems **20** (2004), 103–120.

[4] Y. Censor, T. Bortfeld, B. Martin, and A. Trofimov, *A unified approach for inversion problems in intensity-modulated radiation therapy*, Physics in Medicine and Biology **51** (2006), 2353–2365.

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

[6] Y. Censor, A. Gibali, F. Lenzen, and Ch. Schnorr, *The implicit convex feasibility problem and its application to adaptive image denoising,* J. Comput. Math. (to appear).

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

[8] C. Chaux, J.-C. Pesquet, and N. Pustelnik, *Nested iterative algorithms for convex constrained image recovery problems,* SIAM Journal on Imaging Sciences **2** (2009), 730–762.

[9] S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit,* SIAM Journal on Scientific Computing **20** (1998), 33–61.

[10] P. L. Combettes and J.-C. Pesquet, *A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery,* IEEE J. Selected Topics Signal Process **1** (2007), 564–574.

[11] D. Donoho, *Compressed sensing,* IEEE Trans. Inform. Theory **52** (2006), 1289–1306.

[12] J. Gotoh, A. Takeda and K. Tono, *DC formulations and algorithms for sparse optimization problems,* Math. Program. (2017). https://doi.org/10.1007/s10107-017-1181-0

[13] E. Esser, Y. Lou, and J. Xin, *A method for finding structured sparse solutions to non-negative least squares problems with applications,* SIAM J. Imaging Sci. **6** (2013), 2010–2046.

[14] P. J. Huber, *Robust estimation of a location parameter.* Ann. Math. Statist.**35** (1964) 73–101.

[15] Y. Lou and M. Yan, *Fast $l_1 - l_2$ Minimization via a proximal operator,* arXiv:1609.09530.

[16] H. Mine and M. Fukushima, *A Minimization method for the sum of a convex function and a continuously differentiable function,* J. Optim. Theory Appl. **33** (1981), 9–23.

[17] A. Moudafi, *About proximal algorithms for Q-lasso,* Thai Math. J. **15** (2017), 1–7.

[18] A. Moudafi and A. Gibali, *$l_1 - l_2$ regularization of split feasibility problems*, Numerical Algorithms (2017) DOI 10.1007/s11075-017-0398-6.

[19] J.-L. Ndoutoume and M. Théra, *Generalized second-order derivatives of convex functions in reflexive Banach spaces,* Bull. Austral. Math. Soc. **51** (1995), 55–72.

[20] H. A. Le Thi, H. M. Le, V. V. Nguyen and T. P. Dinh, *A dc programming approach for feature selection in support vector machines learning,* Advances in Data Analysis and Classification **2** (2008), 259–278.

[21] N. Parikh and S. Boyd, *Block splitting for distributed optimization,* Mathematical Programming Computation **6** (2014), 77–102.

[22] S. Penfold, R. Zalas, M. Casiraghi, M. Brooke, Y. Censor and R. Schulte, *Sparsity constrained split feasibility for dose-volume constraints in inverse planning of intensity-modulated photon or proton therapy,* accepted for publication, Physics in Medicine and Biology.

[23] R. T. Rockafellar and R. J-B Wets, *Variational Analysis*, Springer-Verlag, 1998.

[24] P. D. Tao and L. T. H. An, *Convex analysis approach to dc programming: Theory, algorithms and applications,* Acta Math. Vietnam. **22** (1997), 289–355.

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

[26] H. K. Xu, M. A. Alghamdi and N. Shahzad, *Regularization for the split feasibility problem,* J. Nonlinear Convex Anal. **17** (2016), 513–525.
[27] Z. Xu, X. Chang, F. Xu and H. Zhang, l_{1-2} *Regularization: A thresholding representation theory and a fast solver,* IEEE Trans. Neural Netw. Learn. Syst. **23** (2012), 1013–1027.
[28] I. Yamada, M. Yukawa and M, M. Yamagishi, *Minimizing the Moreau Envelope of Nonsmooth Convex Functions over the Fixed Point Set of Certain Quasi-Nonexpansive Mappings.* in: Bauschke H., Burachik R., Combettes P., Elser V., Luke D., Wolkowicz H. (eds) Fixed-Point Algorithms for Inverse Problems in Science and Engineering. Springer Optimization and Its Applications, vol. 49, Springer, New York, NY, 2011.
[29] P. Yin, Y. Lou, Q. He and J. Xin, *Minimization of* l_{1-2} *for compressed sensing,* SIAM J. Sci. Comput. **37** (2015), 536–563.
[30] X. Zheng, X. Sun, D. Li and J. Sun, *Successive convex approximations to cardinality-constrained convex programs: a piecewise-linear dc approach,* Comput. Optim. Appl. **59** (2014), 379–397.

Manuscript received December 4 2017
revised December 30 2017

A. Moudafi
Aix Marseille Université, CNRS-L.I.S UMR 7296, Domaine Universitaire de Saint-Jérôme, Avenue Escadrille Normandie-Niemen, 13397 Marseille Cedex 20, France
E-mail address: abdellatif.moudafi@univ-amu.fr

H. K. Xu
Department of Mathematics, School of Science, Hangzhou Dianzi University, Hangzhou 310018, China
E-mail address: xuhk@hdu.edu.cn