Precompact convergence of the nonconvex Primal-Dual Hybrid Gradient algorithm

Tao Sun[*,a], Roberto Barrio[b], Lizhi Cheng[a,d], Hao Jiang[c]

[a]College of Science, National University of Defense Technology, Changsha, Hunan, China, 410073.
[b]IUMA and CoDy group, Universidad de Zaragoza. Departamento de Matemática Aplicada. Pedro Cerbuna, 12. E-50009. Zaragoza, Spain.
[c]College of Computer, National University of Defense Technology, Hunan, Changsha, China, 410073.
[d] The State Key Laboratory for High Performance Computation, National University of Defense Technology, Hunan, Changsha, China, 410073.

Abstract

The Primal-Dual Hybrid Gradient (PDHG) algorithm is a powerful algorithm used quite frequently in recent years for solving saddle-point optimization problems. The classical application considers convex functions, and it is well studied in literature. In this paper, we consider the convergence of an alternative formulation of the PDHG algorithm in the nonconvex case under the precompact assumption. The proofs are based on the Kurdyka-Łojasiewic functions, that covers a wide range of problems. A simple numerical experiment illustrates the convergence properties.

Key words: Primal-Dual Hybrid Gradient algorithms, Kurdyka-Łojasiewic functions, nonconvex optimization, convergence analysis

1. Introduction

This paper is devoted to solving the following well-studied primal problem

$$\min_{x} \Phi(x) := f(x) + g(Kx), \tag{1}$$

where $K \in \mathbb{R}^{m\times n}$ and $x \in \mathbb{R}^n$. If f and g are convex functions, one can represent model (1) as the following saddle-point problem

$$\min_{x} \max_{y} \Psi(x, y) := f(x) - y^{\top} Kx - g^{*}(y), \tag{2}$$

where g^* is the convex conjugate function of g. The saddle-point problem (2) is ubiquitous in different disciplines and applications, especially in the total variation regularization problem arising in imaging science [1, 2, 3].

*Corresponding author
Email addresses: nudtsuntao@163.com,tsun@math.ucla.edu (Tao Sun), rbarrio@unizar.es (Roberto Barrio), clzcheng@nudt.edu.cn (Lizhi Cheng), haojiang@nudt.edu.cn (Hao Jiang)

A classical and frequently-used method for solving problem (2) is the Primal-Dual Hybrid Gradient (PDHG) algorithm [4, 5]. Mathematically, PDHG algorithm can be described as the iterative process

$$\begin{cases} x^{k+1} = \arg\min_x \left\{ \Psi(x, y^k) + \frac{r}{2}\|x - x^k\|_2^2 \right\}, \\ y^{k+1} = \arg\max_y \left\{ \Psi(x^{k+1}, y) - \frac{s}{2}\|y - y^k\|_2^2 \right\}, \end{cases} \tag{3}$$

where r and s are the step sizes of the method. If f and g are convex functions, another description (obtained by choosing $\theta = 0$ in [6]) of the PDHG algorithm is given by

$$\begin{cases} y^{k+1} \in \arg\min_y \left\{ \frac{s}{2}\|y - Kx^k\|_2^2 - \langle y, q^k \rangle + g(y) \right\}, \\ q^{k+1} = q^k + s(Kx^k - y^{k+1}), \\ x^{k+1} \in \arg\min_x \left\{ \frac{1}{2t}\|x - x^k\|_2^2 + \langle Kx, q^{k+1} \rangle + f(x) \right\}, \end{cases} \tag{4}$$

where now K, r and t are the parameters and step sizes of the method. We can easily see that in (4) the main steps in each iteration just lay on calculating the proximal maps of f and g. Compared with (3), the scheme (4) can be directly used to the nonconvex case. This is because in (3) the convex conjugate function g^* is used in the definition (2) of the function $\Psi(x, y)$, however, the conjugate function of a nonconvex function has not been well defined yet in literature, but the proximal maps used in (4) exist for closed nonconvex functions.

Although scheme (4) is apparently similar to the Alternating Direction Method of Multipliers (ADMM) algorithm [7, 8, 9, 10], they are in fact quite different (the authors in [6] also point out this fact). Actually, PDHG has a deep relationship with the inexact Uzawa method [11]. In [12], the authors prove the convergence of PDHG under asymptotical assumptions on the step sizes. The sublinear convergence rate was established in [13] in an ergodic sense via variational inequalities, providing a very concise way to understand the convergence influenced by the step size. Meanwhile, paper [1] also proves the sublinear convergence rate of the PDHG, and this convergence problem is an active research subject in literature [14, 15, 16, 17].

In all the previous papers, the convergence problem was studied for the convex case. In this paper, we study the convergence of the PDHG algorithm for the nonconvex case. More precisely, we consider the scheme (4) where f and g are both nonconvex functions. With the help of Kurdyka-Łojasiewic function properties, we prove that the points generated by scheme (4) converge to a critical point of Φ under the precompact assumption (that is, assuming the sequence is bounded). The proofs of our results are motivated by previous recent works [18, 19, 20]. Finally, we present a simple numerical experiment to show the convergence properties. We remark that the convergence results presented in this paper are based on the precompact assumption. Then, we do not prove the generic convergence problem for the nonconvex PDHG algorithm.

The paper is organized as follows: Section 2 presents some basic results and definitions; Section 3 contains the convergence results for the nonconvex PDHG algorithm; Section 4 reports a simple numerical example; and finally, Section 5 presents some conclusions.

2. Preliminaries

In this section we present the definitions and basic properties in variational and convex analysis and in Kurdyka-Łojasiewicz functions used later in the convergence analysis.

2.1. Subdifferentials

We collect several definitions as well as some useful properties in variational and convex analysis (see for more details the excellent monographes [21, 22, 23]).

Given a lower semicontinuous function $J : \mathbb{R}^N \to (-\infty, +\infty]$, its *domain* is defined by

$$\mathrm{dom}(J) := \{x \in \mathbb{R}^N : J(x) < +\infty\}.$$

The *graph* of a real extended valued function $J : \mathbb{R}^N \to (-\infty, +\infty]$ is defined by

$$\mathrm{graph}(J) := \{(x, v) \in \mathbb{R}^N \times \mathbb{R} : v = J(x)\}.$$

The notation of subdifferential plays a central role in (non)convex optimization.

Definition 2.1 (Subdifferentials [21, 22])**.** Let $J : \mathbb{R}^N \to (-\infty, +\infty]$ be a proper and lower semicontinuous function.

1. For a given $x \in \mathrm{dom}(J)$, the *Fréchet subdifferential* of J at x, written as $\hat{\partial} J(x)$, is the set of all vectors $u \in \mathbb{R}^N$ which satisfy

$$\liminf_{\substack{y \neq x \\ y \to x}} \frac{J(y) - J(x) - \langle u, y - x\rangle}{\|y - x\|_2} \geq 0.$$

   When $x \notin \mathrm{dom}(J)$, we set $\hat{\partial} J(x) = \emptyset$.
2. The (limiting) subdifferential, or simply the *subdifferential*, of J at $x \in \mathbb{R}^N$, written as $\partial J(x)$, is defined through the following closure process

$$\partial J(x) := \{u \in \mathbb{R}^N : \exists x^k \to x, J(x^k) \to J(x) \text{ and } u^k \in \hat{\partial} J(x^k) \to u \text{ as } k \to \infty\}.$$

It is easy to verify that the Fréchet subdifferential is convex and closed while the subdifferential is closed. When J is convex, the definition agrees with the one used in convex analysis [23] which can be described as

$$\partial J(x) := \{v : J(y) \geq J(x) + \langle v, y - x\rangle \text{ for any } y \in \mathbb{R}^N\}.$$

Let $\{(x^k, v^k)\}_{k\in\mathbb{N}}$ be a sequence in $\mathbb{R}^N \times \mathbb{R}$ such that $(x^k, v^k) \in \mathrm{graph}\ (\partial J)$. If (x^k, v^k) converges to (x, v) as $k \to +\infty$ and $J(x^k)$ converges to v as $k \to +\infty$, then $(x, v) \in \mathrm{graph}(\partial J)$. This fact gives us the following simple proposition.

Proposition 2.2. If $v^k \in \partial J(x^k)$, $\lim_k v^k = v$ and $\lim_k x^k = x$, then we have that

$$v \in \partial J(x). \tag{5}$$

A necessary condition for $x \in \mathbb{R}^N$ to be a minimizer of a lower semicontinuous function $J(x)$ is

$$\mathbf{0} \in \partial J(x). \tag{6}$$

When J is convex, (6) is also sufficient. A point that satisfies (6) is called *(limiting) critical point*. The set of critical points of $J(x)$ is denoted by $\mathrm{crit}(J)$.

Proposition 2.3. If x^* satisfies that

$$\mathbf{0} \in \partial f(x^*) + K^\top \partial g(Kx^*). \tag{7}$$

Then, x^* is a critical point of Φ.

Proof. It is easy to obtain that

$$\mathbf{0} \in \partial f(x^*) + K^\top \partial g(Kx^*) \subseteq \partial \Phi(x^*). \tag{8}$$

So, x^* is a critical point of Φ. □

2.2. Kurdyka-Łojasiewicz functions

In this paper the convergence analysis is based on the Kurdyka-Łojasiewicz functions, originated in the seminal works of Łojasiewicz [24] and Kurdyka [25]. These kind of functions has played a key role in several recent convergence results on nonconvex minimization problems and they are ubiquitous in applications. For example, semi-algebraic, subanalytic and log-exp functions are Kurdyka-Łojasiewicz functions (see [26, 27, 28]).

Definition 2.4. [26, 27] (a) The lower semicontinuous function $J : \mathbb{R}^N \to (-\infty, +\infty]$ is said to have the *Kurdyka-Łojasiewicz (KL) property* at $\overline{x} \in \mathrm{dom}(\partial J)$ if there exist $\eta \in (0, +\infty]$, a neighborhood U of $\overline{x}$ and a continuous function (desingularizing function): $\varphi : [0, \eta) \to \mathbb{R}^+$ such that

1. $\varphi(0) = 0$.
2. φ is C^1 on $(0, \eta)$.
3. (Concavity) For all $s \in (0, \eta)$, $\varphi'(s) > 0$.
4. For all x in $U \bigcap \{x \,|\, J(\overline{x}) < J(x) < J(\overline{x}) + \eta\}$, the Kurdyka-Łojasiewicz inequality holds

$$\varphi'(J(x) - J(\overline{x}))\,\mathrm{dist}(\mathbf{0}, \partial J(x)) \geq 1. \tag{9}$$

(b) Proper lower semicontinuous functions which satisfy the Kurdyka-Łojasiewicz property at each point of $\mathrm{dom}(\partial J)$ are called *Kurdyka-Łojasiewicz (KL) functions*.

KL functions have several interesting properties. We present here a result used later in the main theorem (Theorem 3.7).

Lemma 2.5. *(Lemma 3.6, [26]) Let $J : \mathbb{R}^N \to \mathbb{R}$ be a proper lower semi-continuous function and Ω be a compact set. If J is constant on Ω and satisfies the KL property at each point on Ω, then there exist a function φ and constants $\eta, \varepsilon > 0$ such that for any $\overline{x} \in \Omega$ and any x satisfying that* $\mathrm{dist}(x, \Omega) < \varepsilon$ *and $f(\overline{x}) < f(x) < f(\overline{x}) + \eta$, it holds that*

$$\varphi'(J(x) - J(\overline{x}))\,\mathrm{dist}(\boldsymbol{0}, \partial J(x)) \geq 1. \tag{10}$$

The concept of semi-algebraic sets and functions can help to find and check a very rich class of Kurdyka-Łojasiewicz functions.

Definition 2.6 (Semi-algebraic sets and functions [27])**.** (i) A subset S of $\mathbb{R}^N$ is a real *semi-algebraic set* if there exists a finite number of real polynomial functions $g_{ij}, h_{ij} : \mathbb{R}^N \to \mathbb{R}$ such that

$$S = \bigcup_{j=1}^{p} \bigcap_{i=1}^{q} \{u \in \mathbb{R}^N : g_{ij}(u) = 0 \text{ and } h_{ij}(u) < 0\}.$$

(ii) A function $h : \mathbb{R}^N \to (-\infty, +\infty]$ is a *semi-algebraic function* if its graph

$$\{(u, t) \in \mathbb{R}^{N+1} : h(u) = t\}$$

is a semi-algebraic set of $\mathbb{R}^{N+1}$.

If J is a semi-algebraic function, for any fixed $x \in \mathrm{dom}(\partial J)$, the auxiliary desingularizing function φ in Definition 2.4 takes the form

$$\varphi(s) = cs^{1-\theta}, \tag{11}$$

where $c > 0$ and $0 \leq \theta < 1$.

Lemma 2.7. *[27] Let $J : \mathbb{R}^N \to \mathbb{R}$ be a proper and lower semicontinuous function. If J is semi-algebraic then it satisfies the KL property at any point of* $\mathrm{dom}(J)$*. In particular, if J is semi-algebraic and* $\mathrm{dom}(J) = \mathrm{dom}(\partial J)$*, then J is a KL function.*

We would like to point out that many common functions are semi-algebraic functions: real polynomial functions; indicator functions of semi-algebraic sets; addition and composition of semi-algebraic functions; and in matrix theory cone of PSD matrices, Stiefel manifolds and constant rank matrices. We refer the reader to the reference [27] for these and more examples and properties.

3. Convergence analysis

In this section we provide convergence results in the case f and g functions are semi-algebraic (and so they are KL functions). First we detail some technical lemmas, and we will use the following notation

$$w^k := (x^k,\, y^k,\, q^k,\, x^{k-1}), \qquad d^k := (x^k,\, q^k), \tag{12}$$

where we use the convention $w^0 = w^1$.

Lemma 3.1. *Let $\{(x^k, y^k, q^k)\}_{k=0,1,2,...}$ be the sequence generated by the nonconvex PDHG algorithm (4). If $\frac{1}{2t} - s\|K\|_2^2 > 0$, then we have that*

$$\mathcal{L}(w^k) - \mathcal{L}(w^{k+1}) \geq \nu \|d^{k+1} - d^k\|_2^2, \tag{13}$$

where $\nu := \min\{\frac{1}{2t} - s\|K\|_2^2, \frac{1}{2s}\}$ and

$$\mathcal{L}(x, y, q, \widehat{x}) := f(x) + g(y) + \langle Kx - y, q\rangle + s\|y - K\widehat{x}\|_2^2 + s\|K\|_2^2\|x - \widehat{x}\|_2^2. \tag{14}$$

Proof. Taking into account the iteration formula (4) of updating y^{k+1} we have that

$$\frac{s}{2}\|y^{k+1} - Kx^k\|_2^2 - \langle y^{k+1}, q^k\rangle + g(y^{k+1}) \leq \frac{s}{2}\|y^k - Kx^k\|_2^2 - \langle y^k, q^k\rangle + g(y^k). \tag{15}$$

And updating x^{k+1}

$$\frac{1}{2t}\|x^{k+1} - x^k\|_2^2 + \langle Kx^{k+1}, q^{k+1}\rangle + f(x^{k+1}) \leq \langle Kx^k, q^{k+1}\rangle + f(x^k). \tag{16}$$

Adding the above two inequalities we obtain

$$\begin{aligned}\frac{1}{2t}\|x^{k+1}-x^k\|_2^2+\frac{s}{2}\|y^{k+1}-Kx^k\|_2^2+\langle Kx^{k+1},q^{k+1}\rangle-\langle y^{k+1},q^k\rangle+f(x^{k+1})+g(y^{k+1})\\ \leq\frac{s}{2}\|y^k-Kx^k\|_2^2+\langle Kx^k,q^{k+1}\rangle-\langle y^k,q^k\rangle+f(x^k)+g(y^k).\end{aligned}\tag{17}$$

Inequality (17) can be represented as

$$\begin{aligned}&\langle Kx^{k+1}-y^{k+1},q^{k+1}\rangle+f(x^{k+1})+g(y^{k+1})\\ &\qquad+\frac{1}{2t}\|x^{k+1}-x^k\|_2^2+\frac{s}{2}\|y^{k+1}-Kx^k\|_2^2+\langle Kx^k-y^{k+1},q^k-q^{k+1}\rangle\\ &\leq\frac{s}{2}\|y^k-Kx^k\|_2^2+\langle Kx^k-y^k,q^k\rangle+f(x^k)+g(y^k).\end{aligned}\tag{18}$$

From the iteration of q^{k+1} in (4), we have that

$$\langle q^k-q^{k+1},Kx^k-y^{k+1}\rangle=s\|Kx^k-y^{k+1}\|_2^2.\tag{19}$$

Substituting (19) into (18), we obtain

$$\begin{aligned}&\langle Kx^{k+1}-y^{k+1},q^{k+1}\rangle+f(x^{k+1})+g(y^{k+1})\\ &\qquad+\frac{1}{2t}\|x^{k+1}-x^k\|_2^2+\frac{3s}{2}\|y^{k+1}-Kx^k\|_2^2\\ &\leq\frac{s}{2}\|y^k-Kx^k\|_2^2+\langle Kx^k-y^k,q^k\rangle+f(x^k)+g(y^k).\end{aligned}\tag{20}$$

Now, by using the Schwartz's inequality we infer

$$\begin{aligned}\frac{s}{2}\|y^k-Kx^k\|_2^2&=\frac{s}{2}\|y^k-Kx^{k-1}+Kx^{k-1}-Kx^k\|_2^2\\ &\leq s\|y^k-Kx^{k-1}\|_2^2+s\|Kx^{k-1}-Kx^k\|_2^2\\ &\leq s\|y^k-Kx^{k-1}\|_2^2+s\|K\|_2^2\cdot\|x^{k-1}-x^k\|_2^2.\end{aligned}\tag{21}$$

And by substituting (21) into (20)

$$\begin{aligned}&\langle Kx^{k+1}-y^{k+1},q^{k+1}\rangle+f(x^{k+1})+g(y^{k+1})\\ &\qquad+\frac{1}{2t}\|x^{k+1}-x^k\|_2^2+\frac{3s}{2}\|y^{k+1}-Kx^k\|_2^2\\ &\leq s\|y^k-Kx^{k-1}\|_2^2+s\|K\|_2^2\cdot\|x^{k-1}-x^k\|_2^2\\ &\qquad+\langle Kx^k-y^k,q^k\rangle+f(x^k)+g(y^k).\end{aligned}\tag{22}$$

Defining $\nu:=\min\{\frac{1}{2t}-s\|K\|_2^2,\frac{1}{2s}\}>0$, and using the operator $\mathcal{L}$ (Eq. 14) we will have that

$$\mathcal{L}(x^{k+1},y^{k+1},q^{k+1},x^k)+\nu\|x^{k+1}-x^k\|_2^2+\frac{s}{2}\|y^{k+1}-Kx^k\|_2^2\leq\mathcal{L}(x^k,y^k,q^k,x^{k-1}).\tag{23}$$

Taking into account that $y^{k+1}-Kx^k=(q^k-q^{k+1})/s$ and $\nu\leq\frac{1}{2s}$, we obtain the result. □

The next lemma relates the critical points of the functions $\mathcal{L}$ and Φ.

Lemma 3.2. *If $w^* = (x^*, y^*, q^*, \widehat{x}^*)$ is a critical point of the function $\mathcal{L}$ (14), then, x^* is a critical point of Φ.*

Proof. A direct differentiation calculus will give that

$$\partial\mathcal{L}(w) = \begin{pmatrix} \partial f(x) + K^\top q + 2s\|K\|_2^2(x - \widehat{x}) \\ \partial g(y) - q + 2s(y - K\widehat{x}) \\ Kx - y \\ 2sK^\top(K\widehat{x} - y) \\ 2s\|K\|_2^2(\widehat{x} - x) \end{pmatrix}. \tag{24}$$

As w^* is a critical point of $\mathcal{L}$ then $\mathbf{0} \in \partial\mathcal{L}(w^*)$. That is, we have

$$\begin{cases} \mathbf{0} \in \partial f(x^*) + K^\top q^* + 2s\|K\|_2^2(x^* - \widehat{x}^*), \\ \mathbf{0} \in \partial g(y^*) - q^* + 2s(y^* - K\widehat{x}^*), \\ \mathbf{0} = Kx^* - y^*, \\ \mathbf{0} = 2sK^\top(K\widehat{x}^* - y^*), \\ \mathbf{0} = 2s\|K\|_2^2(\widehat{x}^* - x^*). \end{cases} \tag{25}$$

So, we obtain that

$$\mathbf{0} \in \partial f(x^*) + K^\top \partial g(Kx^*) \subseteq \partial\Phi(x^*), \tag{26}$$

and from Proposition 2.3 we have that x^* is a critical point of Φ. □

Lemma 3.3. *Let $\{(x^k, y^k, q^k)\}_{k=0,1,2,\ldots}$ be the sequence generated by the nonconvex PDHG (4). Then, there exists $L > 0$ such that*

$$\text{dist}(\mathbf{0}, \partial\mathcal{L}(w^{k+1})) \leq L\|d^{k+1} - d^k\|_2. \tag{27}$$

Proof. Note that y^{k+1} minimizes $\frac{s}{2}\|y - Kx^k\|_2^2 - \langle y, q^k\rangle + g(y)$, then

$$\mathbf{0} \in \partial\left[\frac{s}{2}\|y - Kx^k\|_2^2 - \langle y, q^k\rangle + g(y)\right]\Bigg|_{y=y^{k+1}}.$$

That is, the optimization condition of the algorithm at each iteration gives that

$$q^k + s(Kx^k - y^{k+1}) \in \partial g(y^{k+1}). \tag{28}$$

Similarly, we can derive that

$$-K^\top q^{k+1} - \frac{1}{t}(x^{k+1} - x^k) \in \partial f(x^{k+1}). \tag{29}$$

Substituting (28) and (29) into (24), we have

$$z^{k+1} := \begin{pmatrix} 2s\|K\|_2^2(x^{k+1} - x^k) - \frac{1}{t}(x^{k+1} - x^k) \\ s(y^{k+1} - Kx^k) = q^k - q^{k+1} \\ Kx^{k+1} - y^{k+1} \\ 2sK^\top(Kx^k - y^{k+1}) = 2K^\top(q^{k+1} - q^k) \\ 2s\|K\|_2^2(x^k - x^{k+1}) \end{pmatrix} \in \partial\mathcal{L}(w^{k+1}), \tag{30}$$

where for convenience, we have defined the term z^{k+1}. Note that

$$Kx^{k+1} - y^{k+1} = Kx^k - y^{k+1} + Kx^{k+1} - Kx^k = q^{k+1} - q^k + Kx^{k+1} - Kx^k. \tag{31}$$

Thus, we have

$$\|Kx^{k+1} - y^{k+1}\|_2 \leq \|q^{k+1} - q^k\|_2 + \|K\|_2 \, \|x^{k+1} - x^k\|_2. \tag{32}$$

Therefore, we can further obtain that

$$\begin{aligned} \|z^{k+1}\|_2 &\leq \left[\left(\frac{1}{t} - 2s\|K\|_2^2\right) + \|K\|_2 + 2s\|K\|_2^2 \right] \|x^{k+1} - x^k\|_2 \\ &\quad + (2 + 2\|K\|_2)\|q^{k+1} - q^k\|_2 \\ &\leq L\|d^{k+1} - d^k\|_2, \end{aligned} \tag{33}$$

where $L = \sqrt{2} \max\{\frac{1}{t} + \|K\|_2, 2 + 2\|K\|_2\}$. Then, we have

$$\operatorname{dist}(\mathbf{0}, \partial\mathcal{L}(w^{k+1})) \leq \operatorname{dist}(\mathbf{0}, z^{k+1}) = \|z^{k+1}\|_2 \leq L\|d^{k+1} - d^k\|_2. \tag{34}$$

□

Lemma 3.4. *If $\frac{1}{2t} - s\|K\|_2^2 > 0$ and the sequence $\{(x^k, y^k, q^k)\}_{k=0,1,2,\ldots}$, generated by the nonconvex PDHG (4), is bounded. Then, we have*

$$\lim_k \|d^{k+1} - d^k\|_2 = 0, \tag{35}$$

being d^k defined in (12). Moreover, for any cluster point (x^*, y^*, q^*), x^* is also a critical point of Φ.

Proof. We note that the boundness of $\{(x^k, y^k, q^k)\}_{k=0,1,2,\ldots}$ indicates the boundness of $\{w^k\}_{k=0,1,2,\ldots}$, and the continuity of $\mathcal{L}$ indicates that $\{\mathcal{L}(w^k)\}_{k=0,1,2,\ldots}$ is also bounded. Moreover, from Lemma 3.1, $\mathcal{L}(w^k)$ is decreasing. Thus, the sequence $\{\mathcal{L}(w^k)\}_{k=0,1,2,\ldots}$ is convergent, i.e., $\lim_k[\mathcal{L}(w^k) - \mathcal{L}(w^{k+1})] = 0$. Now, using Lemma 3.1, we infer

$$\lim_k \|d^{k+1} - d^k\|_2 \leq \lim_k \sqrt{\frac{\mathcal{L}(w^k) - \mathcal{L}(w^{k+1})}{\nu}} = 0. \tag{36}$$

For any cluster point (x^*, y^*, q^*), there exists $\{k_j\}_{j=0,1,2,\ldots}$ such that $\lim_j(x^{k_j}, y^{k_j}, q^{k_j}) = (x^*, y^*, q^*)$. Then, $\lim_j(x^{k_j+1}, q^{k_j+1}) = (x^*, q^*)$. Note that

$$\lim_j y^{k_j+1} = \lim_j \left(Kx^{k_j} - \frac{q^{k_j} - q^{k_j+1}}{s}\right) = Kx^*.$$

From the definition of the PDHG scheme (4), we have the following conditions

$$\begin{aligned} q^{k_j} - s(y^{k_j+1} - Kx^{k_j}) &\in \partial g(y^{k_j+1}), \\ -K^\top q^{k_j+1} - \frac{x^{k_j+1} - x^{k_j}}{t} &\in \partial f(x^{k_j+1}). \end{aligned} \tag{37}$$

Letting $j \to +\infty$, and using Proposition 2.2, we have

$$\begin{aligned} q^* &\in \partial f(x^*), \\ -K^\top q^* &\in \partial g(Kx^*). \end{aligned} \tag{38}$$

Thus, $\mathbf{0} \in \partial f(x^*) + K^\top \partial g(Kx^*) \subseteq \partial\Phi(x^*)$, and therefore, x^* is a critical point of Φ. □

Now, we recall the definition of the limit point set introduced in [26].

Definition 3.5. Let d^{k_j} be a sequence generated by PDHG scheme from a starting point $d^0 \in \mathbb{R}^N$, then we define the set $\mathcal{M}(d^0)$ of all limit points

$$\mathcal{M}(d^0) := \{u \in \mathbb{R}^N : \exists \text{ an increasing sequence of integers } \{k_j\}_{j\in\mathbb{N}} \text{ such that } d^{k_j} \to u \text{ as } j \to \infty\}.$$

Before proving our main theorem the following result provides some properties of the limit point set.

Lemma 3.6. *Let $\{(x^k, y^k, q^k)\}_{k=0,1,2,...}$ be a sequence generated by the PDHG scheme (4) which is assumed to be bounded, and $\frac{1}{2t} - s\|K\|_2^2 > 0$. Then, the following assertions hold.*

(1) The limit point set $\mathcal{M}(w^0)$ is nonempty and $\mathcal{M}(w^0) \subseteq \mathrm{cri}(\mathcal{L})$ (the set of critical points of $\mathcal{L}$).
(2) $\lim_k \mathrm{dist}(w^k, \mathcal{M}(w^0)) = 0$.
(3) The objective function $\mathcal{L}$ is finite and constant on $\mathcal{M}(w^0)$.

Proof. (1) As the sequence $\{(x^k, y^k, q^k)\}_{k=0,1,2,...}$ is bounded, also the sequence $\{w^k\}_{k=0,1,2,...}$ is bounded, and then $\mathcal{M}(w^0)$ is nonempty. Assume that $w^* \in \mathcal{M}(w^0)$, from the definition, there exists a subsequence $w^{k_i} \to w^*$. From Lemma 3.1, we have $d^{k_i-1} \to d^*$, and from Lemma 3.3, we have that $z^{k_i} \in \partial\mathcal{L}(w^{k_i})$ and $z^{k_i} \to \mathbf{0}$ (using z^k defined in (30)). Therefore, Proposition 2.2 indicates that $\mathbf{0} \in \partial\mathcal{L}(w^*)$, i.e. $w^* \in \mathrm{cri}(\mathcal{L})$.

(2) We obtain the proof by "reductio ad absurdum". Assume this assertion does not hold. Then, there exist $\{k_j\}_{j=0,1,2,...}$ such that

$$\mathrm{dist}(w^{k_j}, \mathcal{M}(w^0)) \geq \varepsilon_0 > 0. \tag{39}$$

Note that $\{k_j\}_{j=0,1,2,...}$ is bounded, and without loss of generality we assume that $\{w^{k_j}\}_{j=0,1,2,...} \to w^*$. Obviously, $w^* \in \mathcal{M}(w^0)$ from the definition. That means $\mathrm{dist}(w^*, \mathcal{M}(w^0)) = 0$. However, from (39), we obtain that

$$\mathrm{dist}(w^*, \mathcal{M}(w^0)) > 0, \tag{40}$$

which leads to a contradiction.

(3) Let the subsequence $w^{k_i} \to w^* \in \mathrm{cri}(\mathcal{L})$. It is easy to see that $\mathcal{L}(w^{k_i})$ is decreasing and $\mathcal{L}(w^{k_i}) > -\infty$. Then, $\mathcal{L}(w^{k_i}) \to \mathcal{L}(w^*)$ which is a finite constant. Let $\overline{w} \in \mathrm{cri}(\mathcal{L})$ be another critical point. It is easy to see that there exists $w^{j_i} \to \overline{w}$ satisfying that $j_i \leq k_i$ for any i. The descend property of $\mathcal{L}$ gives that

$$\mathcal{L}(w^{k_i}) \leq \mathcal{L}(w^{j_i}).$$

Letting $i \to +\infty$, we have $\mathcal{L}(w^*) \leq \mathcal{L}(\overline{w})$. Similarly, we can obtain $\mathcal{L}(\overline{w}) \leq \mathcal{L}(w^*)$. Then, the function value of $\mathcal{L}$ is constant on $\mathcal{M}(w^0)$. □

Now, we present our main theorem about the convergence of the nonconvex PDHG algorithm approaching the set of critical points. We consider the precompact assumption in the convergence study of the nonconvex PDHG (in a similar way as Theorem 1 for the PALM algorithm of [26]), i.e., we assume the sequence is bounded.

Theorem 3.7 (Convergence result)**.** *Suppose that f, g are both semi-algebraic functions and $\frac{1}{2t} - s\|K\|_2^2 > 0$. Let $\{(x^k, y^k, q^k)\}_{k=0,1,2,\dots}$ be a sequence generated by the PDHG scheme (4) which is assumed to be bounded. Then, the sequence $\{d^k\}_{k=0,1,2,3,\dots}$ defined in (12) has finite length, i.e.*

$$\sum_{k=0}^{+\infty} \|d^{k+1} - d^k\|_2 < +\infty. \tag{41}$$

Moreover, $\{d^k\}_{k=0,1,2,3,\dots}$ converges to (x^*, q^*), where x^* is a critical point of Φ.

Proof. As f and g are both semi-algebraic functions, obviously $\mathcal{L}$ is also semi-algebraic. Besides, from Lemma 3.6, $\mathcal{L}$ is constant on $\mathcal{M}(w^0)$. Let w^* be a stationary point of $\{w^k\}_{k=0,1,2,\dots}$. Also from Lemmas 2.5 and 3.6, we have $\mathrm{dist}(w^k, \mathcal{M}(w^0)) < \varepsilon$ and there is a constant η such that $\mathcal{L}(w^k) < \mathcal{L}(w^*) + \eta$ for any $k > \tilde{k}$ for some $\tilde{k} \in \mathbb{N}$. Hence, also, from Lemma 2.5, we infer

$$\varphi'(\mathcal{L}(w^k) - \mathcal{L}(w^*))\,\mathrm{dist}(\mathbf{0}, \partial\mathcal{L}(w^k)) \geq 1, \tag{42}$$

which together with Lemma 3.3 gives that

$$\frac{1}{\varphi'(\mathcal{L}(w^k) - \mathcal{L}(w^*))} \leq \mathrm{dist}(\mathbf{0}, \partial\mathcal{L}(w^k)) \leq L\|d^k - d^{k-1}\|_2. \tag{43}$$

Then, the concavity of φ yields that

$$\begin{aligned}
\mathcal{L}(w^k) - \mathcal{L}(w^{k+1}) &= \mathcal{L}(w^k) - \mathcal{L}(w^*) - [\mathcal{L}(w^{k+1}) - \mathcal{L}(w^*)] \\
&\leq \frac{\varphi[\mathcal{L}(w^k) - \mathcal{L}(w^*)] - \varphi[\mathcal{L}(w^{k+1}) - \mathcal{L}(w^*)]}{\varphi'[\mathcal{L}(w^k) - \mathcal{L}(w^*)]} \\
&\leq \{\varphi[\mathcal{L}(w^k) - \mathcal{L}(w^*)] - \varphi[\mathcal{L}(w^{k+1}) - \mathcal{L}(w^*)]\} \times L\|d^k - d^{k-1}\|_2.
\end{aligned}$$

Taking into account Lemma 3.1, we have that

$$\nu\|d^{k+1} - d^k\|_2^2 \leq \{\varphi[\mathcal{L}(w^k) - \mathcal{L}(w^*)] - \varphi[\mathcal{L}(w^{k+1}) - \mathcal{L}(w^*)]\} \times L\|d^k - d^{k-1}\|_2,$$

which is equivalent to

$$\begin{aligned}
2\frac{\nu}{L}\|d^{k+1} - d^k\|_2 &\leq 2\sqrt{\varphi[\mathcal{L}(w^k) - \mathcal{L}(w^*)] - \varphi[\mathcal{L}(w^{k+1}) - \mathcal{L}(w^*)]} \\
&\quad \times \sqrt{\frac{\nu}{L}}\sqrt{\|d^k - d^{k-1}\|_2}.
\end{aligned} \tag{44}$$

Using the Schwartz's inequality, we infer

$$\begin{aligned}
2\frac{\nu}{L}\|d^{k+1} - d^k\|_2 &\leq \{\varphi[\mathcal{L}(w^k) - \mathcal{L}(w^*)] - \varphi[\mathcal{L}(w^{k+1}) - \mathcal{L}(w^*)]\} \\
&\quad + \frac{\nu}{L}(\|d^k - d^{k-1}\|_2).
\end{aligned} \tag{45}$$

Adding (45) from $\tilde{k}$ to $\tilde{k} + j$ yields that

$$\frac{\nu}{L}\sum_{k=\tilde{k}}^{\tilde{k}+j-1} \|d^{k+1} - d^k\|_2 + \frac{2\nu}{L}\|d^{\tilde{k}+j+1} - d^{\tilde{k}+j}\|_2 \leq \varphi[\mathcal{L}(w^{\tilde{k}}) - \mathcal{L}(w^*)] - \varphi[\mathcal{L}(w^{\tilde{k}+j+1}) - \mathcal{L}(w^*)]. \tag{46}$$

Taking the limit $j \to +\infty$, combined with Lemma 3.3, we have

$$\frac{\nu}{L}\sum_{k=\tilde{k}}^{+\infty}\|d^{k+1}-d^k\|_2 \le \varphi[\mathcal{L}(w^K)-\mathcal{L}(w^*)] < +\infty. \tag{47}$$

Note that sequence $\{(x^k, y^k, q^k)\}_{k=0,1,2,...}$ is assumed to be bounded. From the inequality above, we can see that $\{d^k\}_{k=0,1,2,...}$ is convergent. From Lemma 3.4, there exists a stationary point (x^*, y^*, q^*) of $\{(x^k, y^k, q^k)\}_{k=0,1,2,...}$ such that x^* is a critical point of Φ. Then, $\{d^k\}_{k=0,1,2,...}$ is convergent and (x^*, q^*) is a stationary point of $\{d^k\}_{k=0,1,2,...}$. That is to say, $\{d^k\}_{k=0,1,2,...}$ converges to (x^*, q^*). This completes the proof. □

Remark 3.8. From the second step in each iteration of the nonconvex PDHG, we also have that $y^k \to Kx^*$. Then, $y^* = Kx^*$ and $w^* = (x^*, Kx^*, q^*, x^*)$.

Theorem 3.9 (Convergence rate). *In the conditions of Theorem 3.7, and assuming that the concave desingularizing function of $\mathcal{L}$ at the point w^* is given by $\varphi(s) = c\, s^{1-\theta}$ (Eq. (11)). Then, the following items hold.*

(1) If $\theta = 0$, the sequence $\{w^k\}_{k=0,1,2,...}$ converges in a finite numbers of steps (and so, the sequence $\{x^k\}_{k=0,1,2,...}$ converges).

(2) If $0 < \theta \le \frac{1}{2}$, then there exist $\alpha > 0$ and $0 \le \tau < 1$ such that $\|x^k - x^*\|_2 \le \|w^k - w^*\|_2 \le \alpha\,\tau^k$.

(3) If $\frac{1}{2} < \theta < 1$, then there exist $\alpha > 0$ such that $\|x^k - x^*\|_2 \le \|w^k - w^*\|_2 \le \alpha\, k^{-\frac{1-\theta}{2\theta-1}}$.

Proof. This is a classical result that follows convergence theorems in KL theory. The proof is similar to Theorem 5 in [29] and it will not be presented here. □

We remark that the previous results establish the convergence of the nonconvex PDHG method, and Theorem 3.9 gives some idea of the convergence rates, but without giving a complete analysis of it. In the following, we present sufficient conditions to guarantee boundedness of the sequence generated by the nonconvex PDHG.

Lemma 3.10. *Assume that f is differentiable, K^* is invertible (i.e., $\|K^*q\|_2 \ge \sigma\|q\|_2$ for any $q \in \mathbb{R}^m$), and there exists $\tau > 0$ such that*

$$\inf\{f(x) - \frac{1}{\tau}\|\nabla f(x)\|_2^2\} > -\infty. \tag{48}$$

If the parameters satisfy the following condition

$$\frac{256\tau\|K\|_2^2}{\sigma^2} < 1, \tag{49}$$

and the step sizes t and s are chosen such that

$$\frac{2}{1+\sqrt{1-\frac{256\tau\|K\|_2^2}{\sigma^2}}} < t < \frac{2}{1-\sqrt{1-\frac{256\tau\|K\|_2^2}{\sigma^2}}}, \qquad \frac{32\tau}{\sigma^2} + \frac{1}{2t^2\|K\|_2^2} < s < \frac{1}{2t\|K\|_2^2}, \tag{50}$$

then the sequence generated by the nonconvex PDHG algorithm is bounded if one of the following conditions is satisfied:

1. *function g is coercive and K is invertible,*

2. *function* $f(x) - \frac{1}{\tau}\|\nabla f(x)\|_2^2$ *is coercive and* $\inf g > -\infty$.

Proof. The definition of the function $\mathcal{L}$ (14) gives that

$$\mathcal{L}(w^k) = f(x^k) + g(y^k) + \langle Kx^k - y^k, q^k\rangle + s\|y^k - Kx^{k-1}\|_2^2 + s\|K\|_2^2\|x^k - x^{k-1}\|_2^2. \quad (51)$$

The second step in the iteration process of the nonconvex PDHG algorithm gives

$$K^*q^k = \frac{x^{k-1} - x^k}{t} - \nabla f(x^k). \quad (52)$$

Note that K^* is invertible, so, there must exist $\sigma > 0$ such that

$$\|q^k\|_2^2 \leq \frac{\|K^*q^k\|_2^2}{\sigma^2} \leq \frac{2\|\nabla f(x^k)\|_2^2}{\sigma^2} + \frac{2\|x^k - x^{k-1}\|_2^2}{\sigma^2 t^2}. \quad (53)$$

On the other hand, we have that

$$\begin{aligned} \langle Kx^k - y^k, q^k\rangle &= \langle Kx^k - Kx^{k-1}, q^k\rangle + \langle Kx^{k-1} - y^k, q^k\rangle \\ &\leq \frac{32\tau}{\sigma^2}\|K\|_2^2 \cdot \|x^k - x^{k-1}\|_2^2 + \frac{\sigma^2}{8\tau}\|q^k\|_2^2 \\ &\quad + \frac{32\tau}{\sigma^2}\|Kx^{k-1} - y^k\|_2^2 + \frac{\sigma^2}{8\tau}\|q^k\|_2^2. \end{aligned} \quad (54)$$

Substituting (53) and (54) into (51), we can obtain that

$$\begin{aligned} \mathcal{L}(w^0) \geq \;& \mathcal{L}(w^k) \geq g(y^k) + f(x^k) - \frac{1}{\tau}\|\nabla f(x^k)\|_2^2 + \frac{1}{2\tau}\|\nabla f(x^k)\|_2^2 \\ & + [(s - \frac{32\tau}{\sigma^2})\|K\|_2^2 - \frac{1}{\tau t^2}] \cdot \|x^k - x^{k-1}\|_2^2 + (s - \frac{32\tau}{\sigma^2})\|y^k - Kx^{k-1}\|_2^2. \end{aligned} \quad (55)$$

Under the parameters assumptions, we can see that $(s - \frac{32\tau}{\sigma^2})\|K\|_2^2 - \frac{1}{\tau t^2} \geq 0$ and $s - \frac{32\tau}{\sigma^2} \geq 0$. Now, in the case of condition –1– (function g is coercive and K is invertible), $\{y^k\}_{k=0,1,2,...}$, $\{x^k - x^{k-1}\}_{k=1,2,...}$, $\{\nabla f(x^k)\}_{k=0,1,2,...}$ and $\{y^k - Kx^{k-1}\}_{k=1,2,...}$ are bounded. And so, by the scheme of PDHG, $\{w^k\}_{k=0,1,2,...}$ is bounded. And in the case of condition –2– (function $f(x) - \frac{1}{\tau}\|\nabla f(x)\|_2^2$ is coercive and $\inf g > -\infty$), $\{x^k\}_{k=0,1,2,...}$, $\{x^k - x^{k-1}\}_{k=1,2,...}$, $\{\nabla f(x^k)\}_{k=0,1,2,...}$ and $\{y^k - Kx^{k-1}\}_{k=1,2,...}$ are bounded. And therefore, $\{w^k\}_{k=0,1,2,...}$ is bounded. □

Remark 3.11. Just as illustrative examples, and to show that the conditions imposed in the previous theorem are easily satisfied for some problems, we present two examples for the two cases.

Example 1 (condition –1–): Minimization of

$$\min_x \frac{1}{2}\|b - Ax\|_2^2 + \lambda\|x\|_q^q, \quad (56)$$

where $\lambda > 0$, $0 < q < 1$ and $\|\cdot\|_q$ denotes the ℓ_q norm. In paper [18], $\|b - Ax\|_2^2$ satisfies (48) with $\tau = \frac{1}{2\sqrt{2}\|A\|_2^2}$ and $\lambda\|x\|_q^q$ is coercive.

Example 2 (condition –2–): Minimization of

$$\min_x \frac{1}{2}\|b - Ax\|_2^2 + g(Kx), \quad (57)$$

where $\inf g > -\infty$, K is invertible and $\|A\|_2$ is small.

4. Numerical results

In this section, we present some numerical tests. The first part illustrates the convergence of the nonconvex PDHG algorithm under a general matrix K. While the second part is devoted to compare the performances of different first-order methods that have been extensively studied and widely used for solving a subset of the problem analyzed in this paper (PDHG, Peaceman-Rachford splitting (PRS) [30, 31], Douglas-Rachford Splitting algorithm (DRS) [32], Forward-Backward splitting algorithm (FB) [33] and Alternating Direction Method of Multipliers algorithm (ADMM) [34]).

4.1. Performance of the nonconvex PDHG

This section contains a simple experimental result on the performance of the nonconvex PDHG method in order to show the convergence of the scheme. We consider the following problem

$$\min_x \left\{ \|Kx\|_0 + \frac{\lambda}{2}\|b - Ax\|_2^2 \right\}, \tag{58}$$

where $x \in \mathbb{R}^n$, $K, A \in \mathbb{R}^{m\times n}$ and $\|x\|_0 = \sharp(\mathrm{supp}(x))$. Letting $g(\cdot) := \|\cdot\|_0$ and $f(x) = \frac{\lambda}{2}\|b - Ax\|_2^2$, then problem (58) can be rewritten in the same form as (1). In our tests we consider $m = n = 100$. The PDHG algorithm applied to (58) takes the following form

$$\begin{cases} y^{k+1} \in \arg\min_y \left\{ \frac{s}{2}\|y - Kx^k\|_2^2 - \langle y, q^k\rangle + \|y\|_0 \right\}, \\ q^{k+1} = q^k + s(Kx^k - y^{k+1}), \\ x^{k+1} = \left(\lambda A^\top A + \frac{\mathbb{I}_n}{t}\right)^{-1}\left(\frac{x^k}{t} - K^\top q^{k+1} + \lambda A^\top b\right), \end{cases} \tag{59}$$

with $\mathbb{I}_n$ the identity matrix. In fact, the first step is the proximal map of $\|\cdot\|_0$, which is easy to compute [27]. More precisely, y^{k+1} can be calculated as

$$y^{k+1} = \begin{cases} 0, & \text{if } Kx^k + \frac{q^k}{s} < \sqrt{\frac{2}{s}}, \\ Kx^k + \frac{q^k}{s}, & \text{in other case.} \end{cases} \tag{60}$$

In order to study the error of the convergence process we will show the relative error of a sequence $\{x^k\}_{k=0,1,2,\ldots}$, that is given by

$$\mathrm{RelErr}_x^k = \frac{\|x^k - x^*\|_2}{\|x^*\|_2}, \tag{61}$$

where x^* is the convergent point of sequence $\{x^k\}_{k=0,1,2,\ldots}$. In a similar manner we can obtain RelErr_y^k.

In any numerical iterative method a quite useful problem is the automatic error control, that is, how to detect if the method has reached (within a required tolerance error) or not an approximate solution of the problem. In our case, in the numerical tests we use the following error estimators to study the convergence process: $\mathrm{Est}_f^k := |f^k - f^{k+1}|$, $\mathrm{Est}_x^k := \|x^{k+1} - x^k\|_2$ and D^k, where in each iteration f^k and D^k are defined as:

1. The function values
$$f^k := \|Kx^k\|_0 + \frac{\lambda}{2}\|b - Ax^k\|_2^2. \tag{62}$$
2. The distance of the subdifferential to the origin
$$D^k := \text{dist}\Big[\mathbf{0}, \partial\Big(\|Kx^k\|_0 + \frac{\lambda}{2}\|b - Ax^k\|_2^2\Big)\Big]. \tag{63}$$

Note that $\frac{\lambda}{2}\|b - Ax\|_2^2$ is differentiable, thus we have that
$$\partial\Big(\|Kx^k\|_0 + \frac{\lambda}{2}\|b - Ax^k\|_2^2\Big) = K^\top\partial\|Kx^k\|_0 + \lambda A^\top(Ax^k - b). \tag{64}$$

Besides, as
$$(\partial\|x\|_0)_i = \partial|x_i|_0, \tag{65}$$
and $\partial|x_i|_0 = \begin{cases} 0, & x_i \neq 0, \\ (-\infty, +\infty), & x_i = 0, \end{cases}$ we obtain a simple formula for the distance
$$D^k = \arg\min_{y\subseteq \text{supp}(Kx^k)} \|\lambda A^\top(b - Ax^k) - K^\top y\|_2. \tag{66}$$

In order to stop the iterative process we have to choose three different error tolerances (one per error estimator) and to impose $\text{Est}_f^k < \texttt{Tol}_f$, $\text{Est}_x^k < \texttt{Tol}_x$ and $D^k < \texttt{Tol}_D$.

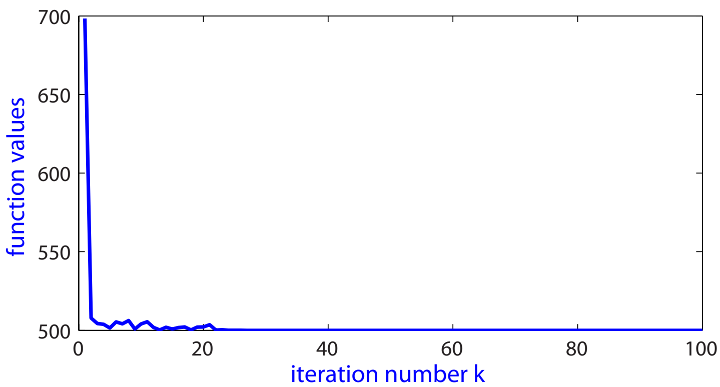

Figure 1: Function values versus the iteration number using the nonconvex PDHG in the test problem (58).

In our test problem (58), we have chosen a matrix $A \in \mathbb{R}^{200\times1000}$ which entries are generated by a normal (Gaussian) random variable. Figure 1 shows the function values f^k versus the iteration number k. We observe that the function value is approaching to a constant value, showing the convergence process. In our tests we have considered that the algorithm has converged to x^* if $\text{Est}_f^k < 10^{-6}$, $\text{Est}_x^k < 10^{-4}$ and $D^k < 10^{-4}$, and the process it would stop at 30 iterations (although we present the results for more iterations just to show the behaviour of the method).

The evolution of the relative error of variables $\{x^k\}_{k=0,1,...}$ and $\{y^k\}_{k=0,1,...}$ (that is, RelErr_x^k and RelErr_y^k) versus the iteration number using the nonconvex PDHG is presented in Figure 2. We have used two different scales (linear and logarithmic) in the pictures in order to see the convergence process. From the pictures we observe clearly the convergence of the algorithm, but

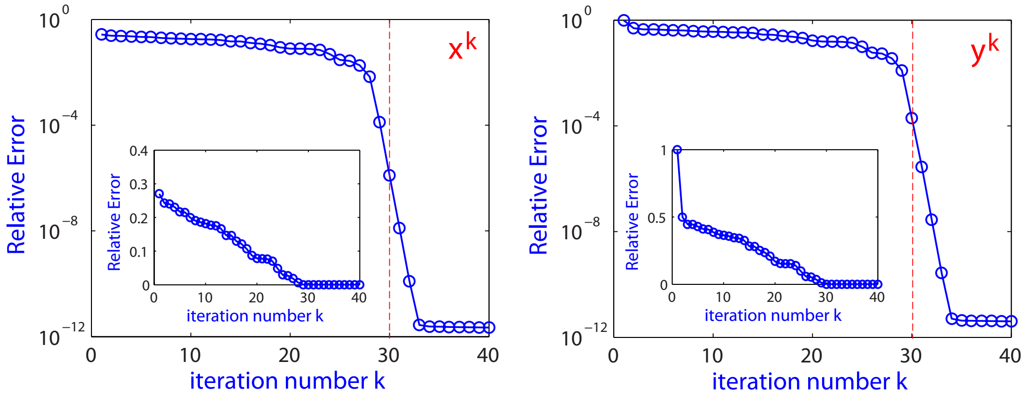

Figure 2: Relative errors RelErr_x^k and RelErr_y^k of variables x^k and y^k, respectively, versus the iteration number using the nonconvex PDHG in the test problem (58).

also two different slopes, on one hand we have at the beginning a sublinear/linear convergence process approaching the solution, and when the numerical solution is close enough the convergence is much faster (lineal/superlinear). Obviously this numerical test is just an illustration of the theory developed in this paper showing the convergence of the nonconvex PDHG method, and a much more detailed numerical analysis is out of the scope of this paper and it is part of the future research of our group.

4.2. Comparisons of different algorithms

This section considers the comparisons of the PDHG, FB, PRS, DRS and ADMM algorithms on the problem

$$\min \frac{\lambda}{2}\left\{\|Ax-b\|_2^2+\|x\|_0\right\}, \tag{67}$$

where $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$ and $\|x\|_0 = \sharp(\mathrm{supp}(x))$. The parameters and stopping criterion are the same as the ones used in the previous numerical test. Compared with the model used above (Subsection 4.1), now the matrix K is given by $K = \mathbb{I}_n$. We set this matrix K because the algorithm FB, PRS and DRS need to calculate the proximal of $\|Kx\|_0$. But the proximal of $\|Kx\|_0$ can be numerically computed only if K is invertible, and therefore a general case is not solvable with those algorithms. Note that these algorithms do not have the dual iterative points and ADMM has two iterative points. Therefore, from that point of view, the nonconvex PDHG method provides an interesting algorithm that can scan a much wider range of problems. Thus, in this section we just compare the function values of the algorithms in a simple problem that is approachable by all of them.

In Figure 3 we observe that all the methods work well on this problem, as expected, but from the tests the best performance, as the convergence seems to be the fastest (see in the magnification how the PDHG curve is the smallest), is obtained with the nonconvex PDHG method. In Table 1, we present, for all the algorithms and problems of different sizes n, the error estimate using the function values Est_f^k and the number of iterations needed to reach $\mathrm{Est}_f^k < 10^{-3}$. In all the simulations the nonconvex PDHG method has required the lowest number of iterations to reach the tolerance error. Moreover, we remark that the nonconvex PDHG method can be used for a wider range of problems.

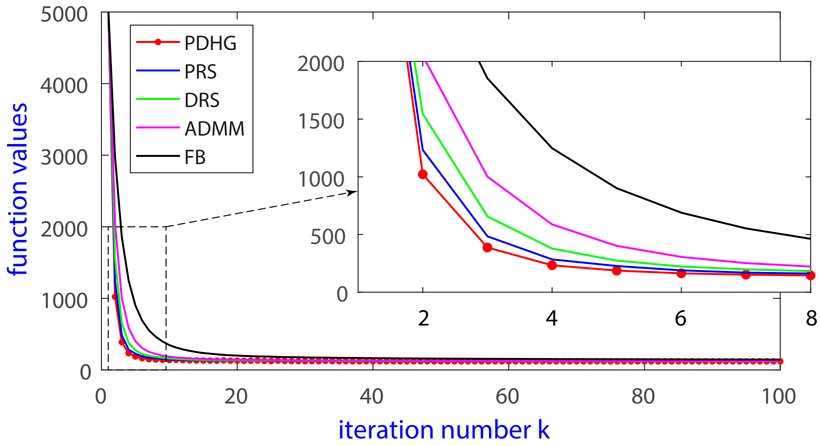

Figure 3: Comparisons of function values versus the iteration number using different algorithms in the test problem (67).

Table 1: Error estimate using the function values Est_f^k and the number of iterations to reach $\mathrm{Est}_f^k < 10^{-3}$ for the PDHG, FB, PRS, DRS and ADMM algorithms.

scenario	PDHG		ADMM		DRS		PRS		FB	
n	Iter	Est_f^k	Iter	Est_f^k	Iter	Est_f^k	Iter	Est_f^k	Iter	Est_f^k
100	78	9.87e-4	97	7.42e-4	93	8.62e-4	85	9.65e-4	112	8.42e-4
200	68	8.34e-4	82	9.55e-4	75	9.61e-4	71	8.44e-4	95	8.22e-4
300	74	9.66e-4	96	8.41e-4	91	9.37e-4	81	9.12e-4	106	9.17e-4
400	76	9.47e-4	94	9.27e-4	92	9.06e-4	80	9.22e-4	108	8.84e-4
500	72	9.07e-4	95	9.58e-4	92	9.48e-4	82	9.47e-4	111	9.24e-4
600	73	8.94e-4	93	8.95e-4	90	8.92e-4	83	9.52e-4	115	9.43e-4
700	77	9.74e-4	96	9.67e-4	91	9.57e-4	83	9.38e-4	120	8.74e-4
800	70	9.01e-4	93	9.84e-4	93	9.68e-4	79	9.49e-4	119	9.16e-4
900	78	8.89e-4	95	9.76e-4	89	9.97e-4	84	9.57e-4	104	9.40e-4
1000	79	9.36e-4	98	9.28e-4	93	9.69e-4	85	9.62e-4	118	9.03e-4

5. Conclusions

The Primal-Dual Hybrid Gradient (PDHG) algorithm is a powerful algorithm used quite frequently for solving saddle-point optimization problems. In this paper, we study the precompact convergence of the nonconvex PDHG, i.e., we prove the convergence if the sequence is bounded (but the global convergence is not proved). The proofs are motivated by existing results in Kurdyka-Łojasiewic function theory, giving rise results when the optimization functions are semi-algebraic. A simple numerical test illustrates the convergence process of the nonconvex PDHG algorithm.

Acknowledgments

T.S., L.C. and H.J. have been supported during this research by the National Natural Science Foundation of Hunan Province, China (13JJ2001), the Science Project of National University of

Defense Technology (JC120201) and the National Science Foundation of China (No.61402495 and No.61571008). R.B. has been supported during this research by the Spanish Research project MTM2015-64095-P (MINECO/FEDER) and the European Social Fund and Diputación General de Aragón (Grant E48).

References

[1] A. Chambolle, T. Pock, A first-order primal-dual algorithm for convex problems with applications to imaging, Journal of Mathematical Imaging and Vision 40 (1) (2011) 120–145.

[2] P. Weiss, L. Blanc-Féraud, G. Aubert, Efficient schemes for total variation minimization under constraints in image processing, SIAM journal on Scientific Computing 31 (3) (2009) 2047–2080.

[3] X. Zhang, M. Burger, S. Osher, A unified primal-dual algorithm framework based on Bregman iteration, Journal of Scientific Computing 46 (1) (2011) 20–46.

[4] M. Zhu, T. Chan, An efficient primal-dual hybrid gradient algorithm for total variation image restoration, UCLA CAM Report (2008) 08–34.

[5] T. Goldstein, M. Li, X. Yuan, E. Esser, R. Baraniuk, Adaptive Primal-Dual Hybrid Gradient Methods for Saddle-Point Problems, arXiv:1305.0546.

[6] T. Möllenhoff, E. Strekalovskiy, M. Moeller, D. Cremers, The primal-dual hybrid gradient method for semiconvex splittings, SIAM Journal on Imaging Sciences 8 (2) (2015) 827–857.

[7] D. Gabay, B. Mercier, A dual algorithm for the solution of nonlinear variational problems via finite element approximation, Computers & Mathematics with Applications 2 (1) (1976) 17–40.

[8] J. Eckstein, M. Fukushima, Large Scale Optimization: State of the Art, Springer US, Boston, MA, 1994, Ch. Some Reformulations and Applications of the Alternating Direction Method of Multipliers, pp. 115–134.

[9] R. Glowinski, P. Le Tallec, Augmented Lagrangian and operator-splitting methods in nonlinear mechanics, Vol. 9, SIAM, 1989.

[10] R. Kellogg, A nonlinear alternating direction method, Mathematics of Computation 23 (105) (1969) 23–27.

[11] K.-J. Arrow, L. Hurwicz, H. Uzawa, H.-B. Chenery, S.-M. Johnson, S. Karlin, T. Marschak, Studies in linear and non-linear programming, Stanford University Press, John Wiley & Sons, 1959.

[12] E. Esser, X. Zhang, T. F. Chan, A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science, SIAM Journal on Imaging Sciences 3 (4) (2010) 1015–1046.

[13] B. He, Y. You, X. Yuan, On the convergence of primal-dual hybrid gradient algorithm, SIAM Journal on Imaging Sciences 7 (4) (2014) 2526–2537.

[14] R. I. Boţ, E. R. Csetnek, A. Heinrich, C. Hendrich, On the convergence rate improvement of a primal-dual splitting algorithm for solving monotone inclusion problems, Mathematical Programming 150 (2) (2015) 251–279.

[15] R. I. Boţ, E. R. Csetnek, An inertial forward-backward-forward primal-dual splitting algorithm for solving monotone inclusion problems, Numerical Algorithms (2015) 1–22.

[16] R. I. Boţ, E. R. Csetnek, On the convergence rate of a forward-backward type primal-dual splitting algorithm for convex optimization problems, Optimization 64 (1) (2015) 5–23.

[17] R. I. Boţ, E. R. Csetnek, A. Heinrich, A primal-dual splitting algorithm for finding zeros of sums of maximal monotone operators, SIAM Journal on Optimization 23 (4) (2013) 2011–2036.

[18] G. Li, T. K. Pong, Global convergence of splitting methods for nonconvex composite optimization, SIAM Journal on Optimization 25 (4) (2015) 2434–2460.

[19] G. Li, T. K. Pong, Peaceman-Rachford splitting for a class of nonconvex optimization problems, arXiv:1507.00887.

[20] G. Li, T. K. Pong, Douglas–rachford splitting for nonconvex optimization with application to nonconvex feasibility problems, Mathematical Programming 159 (1-2) (2016) 371–401.

[21] B. S. Mordukhovich, Variational analysis and generalized differentiation I: Basic theory, Vol. 330, Springer Science & Business Media, 2006.

[22] R. T. Rockafellar, R. J.-B. Wets, Variational analysis, Vol. 317, Springer Science & Business Media, 2009.

[23] R. T. Rockafellar, Convex analysis, Princeton university press, 2015.

[24] S. Łojasiewicz, Les Équations aux Dérivées Partielles, Éditions du centre National de la Recherche Scientifique, Paris, 1963, Ch. Une propriété topologique des sous-ensembles analytiques réels, pp. 8–89.

[25] K. Kurdyka, On gradients of functions definable in o-minimal structures, Annales de l'Institut Fourier 48 (1998) 769–783.

[26] J. Bolte, S. Sabach, M. Teboulle, Proximal alternating linearized minimization for nonconvex and nonsmooth problems, Mathematical Programming 146 (1-2) (2014) 459–494.

[27] H. Attouch, J. Bolte, B. F. Svaiter, Convergence of descent methods for semi-algebraic and tame problems: prox-

imal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods, Mathematical Programming 137 (1-2) (2013) 91–129.

[28] H. Attouch, J. Bolte, P. Redont, A. Soubeyran, Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Lojasiewicz inequality, Mathematics of Operations Research 35 (2) (2010) 438–457.

[29] H. Attouch, J. Bolte, On the convergence of the proximal algorithm for nonsmooth functions involving analytic features, Mathematical Programming 116 (1-2) (2009) 5–16.

[30] R. Glowinski, T. Kärkkäinen, K. Majava, On the convergence of operator-splitting methods, in: Numerical Methods for Scientific Computing, Variational Problems and Applications, CIMNE, Barcelona, 2003, pp. 67–79.

[31] D. W. Peaceman, H. H. Rachford, The numerical solution of parabolic and elliptic differential equations, Journal of the Society for Industrial and Applied Mathematics 3 (1) (1955) 28–41.

[32] J. Eckstein, D. P. Bertsekas, On the douglas-rachford splitting method and the proximal point algorithm for maximal monotone operators, Mathematical Programming 55 (1) (1992) 293–318.

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

[34] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Foundations and Trends in Machine Learning 3 (1) (2011) 1–122.