DCA BASED ALGORITHM WITH EXTRAPOLATION FOR NONCONVEX NONSMOOTH OPTIMIZATION

DUY NHAT PHAN* AND HOAI AN LE THI†

Abstract. In this paper, we focus on the problem of minimizing the sum of a nonconvex differentiable function and a DC (Difference of Convex functions) function, where the differentiable function is not restricted to the global Lipschitz gradient continuity assumption. This problem covers a broad range of applications in machine learning and statistics such as compressed sensing, signal recovery, sparse dictionary learning, and matrix factorization etc. We take inspiration from the Nesterov’s acceleration technique and the DC algorithm to develop a novel algorithm for the considered problem. Analyzing the convergence, we study the subsequential convergence of our algorithm to a critical point. Furthermore, we justify the global convergence of the whole sequence generated by our algorithm to a critical point, and establish its convergence rate under the Kurdyka-
Lojasiewicz condition. Numerical experiments on the nonnegative matrix completion problem are performed to demonstrate the efficiency of our algorithm and its superiority over well-known methods.

1. Introduction. In this paper, we specifically focus on a class of following nonconvex nonsmooth problems

\[ \min_{x \in \mathbb{E}} F(x) := f(x) + g(x) - h(x), \]

where \( \mathbb{E} \) is a finite dimensional real linear space and the functions \( f, g, \) and \( h \) satisfy the following mild assumption:

**Assumption 1.**

(a) \( f : \mathbb{E} \to \mathbb{R} \) is a nonconvex continuously differentiable function and there exists a strongly convex continuously differentiable function \( \phi \) such that \( L\phi - f \) and \( l\phi + f \) are convex for some \( L > 0 \) and \( l \geq 0 \).

(b) \( g : \mathbb{E} \to \mathbb{R} \cap \{+\infty\} \) is a proper and lower semicontinuous convex function.

(c) \( h : \mathbb{E} \to \mathbb{R} \) is a convex function.

(d) \( F \) is bounded from below.

Problem (1) covers a broad range of applications in machine learning and statistics, e.g., compressed sensing [14], signal recovery [5], sparse dictionary learning [1], sparse matrix factorization [10], quadratic inverse problem in phase retrieve [11], and Poisson linear inverse problems [7]. In particular, many optimization problems in these applications can be formulated as a regularized problem of the form

\[ \min_{x \in \mathbb{E}} \left\{ f(x) + r(x) \right\}, \]

where \( f \) is a loss function and \( r \) is a regularization term for promoting sparse solutions. We notice that if \( L \)-smooth and \( L \)-smooth adaptable [11] loss functions \( f \) satisfy Assumption 1 (a). Hence, the problem (2) can be recast into Problem (1) for some well-known penalty functions \( r \) which are DC functions (Difference of Convex functions) such as SCAD [15], MCP [35], and Exponential penalty function [12]. In this paper, we are interested in the nonnegative matrix completion problem corresponding to (1). Given a nonnegative sparse matrix \( A \in \mathbb{R}^{m \times n} \), in order to obtain factors \( U \in \mathbb{R}^{m \times t} \) and \( V \in \mathbb{R}^{t \times n} \) such that \( A \approx UV \), we need to solve the following

*Department of Mathematics and Informatics, HCMC University of Education, Vietnam (nhatpd@hcmue.edu.vn).

†Department Computer Science and Application, LGIPM, University of Lorraine, France (hoai-an.le-thi@univ-lorraine.fr).
nonconvex optimization problem

\[
\min_{U \in \mathbb{R}^{m \times t}, V \in \mathbb{R}^{t \times n}} \left\{ \frac{1}{2} \| P(A - UV) \|_F^2 + r(U, V) \right\},
\]

where \( f(U, V) := \frac{1}{2} \| P(A - UV) \|_F^2 \) is the data-fitting term, \( r \) is a regularization term. Here, \( P(Z)_{ij} = Z_{ij} \) if \( A_{ij} \) is observed and 0 otherwise. The matrix completion problem taking the form of (3) is one of the most successful approaches in recommendation system, see, e.g. [20]. The usage of matrix completion can also be found in a wide range of tasks from sensor networks [8], social network analysis [19] to image processing [23].

1.1. Related works. The considered problem (1) can be expressed as a DC program

\[
\min_{x \in E} \left\{ F(x) := G(x) - H(x) \right\},
\]

where \( G(x) = L\phi(x) + g(x) \) and \( H(x) = L\phi(x) - f(x) + h(x) \) are convex functions. Hence, the powerful classical DCA [21, 30] for handling (4) iteratively linearizes \( H \) by computing \( v^k \in \partial H(x^k) \) and computes the next iterate \( x^{k+1} \) by solving the convex sub-problem

\[
\min_{x \in E} \left\{ G(x) - \langle v^k, x \rangle \right\}.
\]

[13] proposed an inertial DCA (iDCA) by incorporating a weak inertial force of the heavy-ball method of [32] into DCA. More precisely, iDCA replaces \( v^k \) in the convex sub-problem (5) by \( v^k + \gamma (x^k - x^{k-1}) \) with \( \gamma \geq 0 \). This leads to the following convex sub-problem

\[
\min_{x \in E} \left\{ G(x) - \langle v^k + \gamma (x^k - x^{k-1}), x \rangle \right\}.
\]

In order to study the convergence of iDCA, [13] assumed that the second DC component \( H \) is strongly convex, but this assumption can be replaced by the strongly convexity of the first DC component \( G \). In [31], by taking inspiration from the Nesterov’s acceleration technique [29], the authors introduced an accelerated DCA (ADCA) that constructs an extrapolation point \( y^k \) of the current iterate \( x^k \) and the previous one \( x^{k-1} \)

\[
y^k = x^k + \frac{t_k - 1}{t_{k+1}} (x^k - x^{k-1}),
\]

where \( t_{k+1} = (1 + \sqrt{1 + 4t_k^2})/2 \). ADCA then verifies the following condition

\[
F(y^k) \leq \max_{t=[k-q]_+, \ldots, k} F(x^t),
\]

where \( [k-q]_+ = \max(0, k-q) \). If the condition (6) is satisfied, \( y^k \) will be used instead of \( x^k \) to linearize \( H \) by computing \( v^k \in H(y^k) \). Like DCA, ADCA computes \( x^{k+1} \) by solving the subproblem (5). However, the computing of the objective function values
For a special case of Problem (1) in which \( f \) admits a \( L \)-Lipschitz continuous gradient, we can choose the function \( \phi(x) = \frac{1}{2}\|x\|^2 \). Hence, the sub-problem (5) of DCA applied to this particular case becomes the proximal one

\[
\min_{x \in \mathbb{E}} \left\{ \frac{L}{2}\|x - x_k\|^2 + \langle \nabla f(x) - \xi^k, x \rangle + g(x) \right\},
\]

where \( \xi^k \in \partial h(x^k) \). This DCA scheme is noting else but the proximal DCA algorithm (pDCA) proposed in [17] for the particular case. In [33], the authors proposed a version of pDCA with extrapolation (pDCAe) for solving a special setting of (1) in which \( f \) is convex differentiable with \( L \)-Lipschitz continuous gradient. In particular, pDCAe first computes the extrapolation point \( y^k = x^k + \beta_k(x^k - x^{k-1}) \), where \( \beta_k \in [0, \beta] \) with \( \beta \in [0, 1) \), and then computes the next iterate \( x^{k+1} \) by solving the proximal sub-problem

\[
\min_{x \in \mathbb{E}} \left\{ \frac{L}{2}\|x - y^k\|^2 + \langle \nabla f(y^k) - \xi^k, x \rangle + g(x) \right\}.
\]

For this case, [25, 24] additionally assumed that \( h \) is the supremum of finitely many convex smooth functions and proposed several variants of pDCAe for solving the resulting problem.

Besides to DCA based algorithms, popular proximal gradient (PG) algorithms and their accelerated versions have developed for special cases of (1) see, e.g. [18, 4, 2, 22, 34]. However, these methods also require the global Lipschitz gradient continutity of \( f \). In fact, this requirement may not often be satisfied for many practical problems, such as the quadratic inverse problem in phase retrieve [11], Poisson linear inverse problems [7], and the matrix factorization (3), that arise in many real world applications. Recently, in [11], the authors proposed a Bregman PG (BPG) algorithm for minimizing the sum of two nonconvex functions \( f \) and \( r \). [28, 27] developed an extrapolated version of BPG for the same problem. Although they do not impose the global Lipschitz gradient continuity on \( f \), they require the weakly convexity of \( r \).

In addition, the aforementioned PG algorithms have to compute the proximal map of nonconvex functions \( r \) which do not has closed form in many cases. Usually, this computation can be very expensive or impossible.

### 1.2. Contribution

Motivated by the advantages of the success of Nesterov’s acceleration technique for convex programming to which it was accelerated with an \( O(1/k^2) \) convergence rate [29, 6], we thereby aim to investigate a novel algorithm with extrapolation for solving the DC program (1), named DCAe, by incorporating the Nesterov-type acceleration into DCA. In particular, the DCAe algorithm linearizes the first part \( L\phi - f \) of \( H \) at the extrapolated point \( y^k \) and the second one \( h \) at the current iterate \( x^k \). DCAe subsumes pDCAe [33] as a special version. In addition, it is important noting that our problem (1) is more general than the problems considered in [17, 33, 25, 24]. Since we do not require the global Lipschitz gradient continuity or convexity of \( f \), the analysis in [33, 25, 24] is not applicable to Problem (1). Theoretically, we show that every limit point of the generated by DCAe is a critical point. Importantly, we study the global convergence of the whole sequence generated by DCAe, and establish its convergence rate with the Kurdyka-Lojasiewicz (KL) assumption. As application, we take the nonnegative matrix completion prob-
lem into consideration. We perform numerical experiments on different datasets to demonstrate the efficiency of the proposed algorithm.

1.3. Organization of the paper. In section 2, we present some basic concepts in optimization and Kurdyka-Łojasiewicz property. We introduce the novel DCAe algorithm and investigate its convergence properties in section in section 3. Numerical experiments are showed in section 4. Finally, in section 5, we provide conclusions of the study.

2. Preliminaries. Before introducing the proposed algorithm, let us first recall some basic notations that will be used in the sequel.

Given a lower semicontinuous function $J : \mathbb{E} \to (-\infty, +\infty]$, its domain is defined by
$$
dom(J) := \{x \in \mathbb{E} : J(x) < +\infty\}.
$$

The modulus of strong convexity of $J$ on $\Omega$, denoted by $\rho(J, \Omega)$ or $\rho(J)$ if $\Omega = \mathbb{E}$, is given by
$$
\rho(J, \Omega) = \sup\{\rho \geq 0 : J - (\rho/2)\|\cdot\|^2 \text{ is convex on } \Omega\}.
$$

One says that $J$ is $\rho$-strongly convex on $\Omega$ if $\rho(J, \Omega) > 0$.

For a given $x^* \in \dom(J)$, the Fréchet subdifferential [26] of a lower semicontinuous function $J$ at $x^*$ is defined by
$$
\partial^F J(x^*) = \left\{u : \liminf_{x \to x^*} J(x) - J(x^*) - \langle u, x - x^* \rangle /\|x - x^*\| \geq 0\right\}.
$$

The limiting subdifferential [26] of $J$ at $x^* \in \dom(J)$ is defined by
$$
\partial^L J(x^*) = \left\{u : \exists x^k \xrightarrow{J} x^*, u^k \in \partial^F J(x^k), u^k \to u\right\},
$$

where the notation $x^k \xrightarrow{J} x^*$ means that $x^k \to x^*$ and $J(x^k) \to J(x^*)$. If $J$ is differentiable at $x^*$, then $\partial^F J(x^*) = \{\nabla J(x^*)\}$. Moreover, if $J$ is continuously differentiable on a neighborhood of $x^*$, then $\partial^L J(x^*) = \{\nabla J(x^*)\}$. We note that if $J = \phi + \psi$, where $\phi$ is lower semicontinuous, and $\psi$ is continuously differentiable on a neighborhood of $x^*$, then
$$
\partial^L J(x^*) = \partial^L \phi(x^*) + \nabla \psi(x^*).
$$

If $J$ is convex, then the Fréchet and the limiting subdifferential reduce to the subdifferential in the sense of convex analysis:
$$
\partial J(x^*) := \{y : J(x) \geq J(x^*) + \langle x - x^*, y \rangle, \forall x \in \mathbb{E}\}.
$$

A point $x^*$ is called a critical point of the problem (1) if and only if $[\nabla f(x^*) + \partial g(x^*) \cap \partial h(x^*)] \neq \emptyset$.

Given $\eta \in (0, \infty]$, we denote by $\mathcal{M}_\eta$ the class of continuous concave functions $\psi : [0, \eta) \to [0, \infty)$ verifying

(i) $\psi(0) = 0$ and $\psi$ is continuously differentiable on $(0, \eta)$,

(ii) $\psi'(t) > 0$ for all $t \in (0, \eta)$.
A lower semicontinuous function $J$ satisfies the KL property \cite{3} at $u^* \in \text{dom}(\partial^L J)$ if there exists $\eta > 0$, a neighborhood $V$ of $u^*$, and $\psi \in M_\eta$ such that for all $u \in V \cap \{ u : J(u) < J(u^*) < J(u^*) + \eta \}$, one has
$$
\psi'(J(u) - J(u^*)) \text{dist}(0, \partial^L J(u)) \geq 1.
$$
We notice that the class of functions verifying the KL property is very ample, for example, semi-algebraic, real analytic, and log-exp functions see, e.g., \cite{4, 3, 9}.

3. Difference of Convex function Algorithm with Extrapolation.

3.1. Algorithm description. Let us introduce a novel algorithm to solve the nonconvex nonsmooth optimization problem \eqref{1}. In order to accelerate DCA, we propose its version with extrapolation, named DCAe that, at each iteration, constructs an extrapolation point $y^k$ of the current iterate $x^k$ and the previous one $x^{k-1}$
$$
y^k = x^k + \beta_k (x^k - x^{k-1}),
$$
where $\beta_k$ is an extrapolation parameter satisfying
$$
(L + l)D\phi(x^k, y^k) \leq \delta LD\phi(x^{k-1}, x^k),
$$
with $\delta \in (0, 1)$ and $D\phi(x, y) := \phi(x) - [\phi(y) + \langle \nabla \phi(y), x - y \rangle]$. The DCAe employs both $y^k$ and $x^k$ in order to linearize $H$. In particular, the first part $L\phi - f$ of $H$ is linearized at $y^k$ by computing $\nabla L\phi(y^k) - \nabla f(y^k)$ while the second one $h$ is linearized at $x^k$ by computing $\xi^k \in \partial h(x^k)$. The next iterate $x^{k+1}$ is then computed by solving the following convex sub-problem
$$
\min_{x \in \mathbb{X}} \left\{ G(x) - \left( \nabla L\phi(y^k) - \nabla f(y^k) + \xi^k, x \right) \right\}.
$$
The DCAe algorithm is summarized in Algorithm 1.

\begin{algorithm}
\caption{DCAe for solving \eqref{1}}
\label{algorithm1}
\begin{algorithmic}
\algorithmicrequire Initialization: Choose $x^0 = x^{-1}$, $\delta \in (0, 1)$, $L > 0$, $l \geq 0$, and $k \leftarrow 0$.
\Repeat
1. Compute $y^k = x^k + \beta_k (x^k - x^{k-1})$, where $\beta_k$ satisfies
$$
(L + l)D\phi(x^k, y^k) \leq \delta LD\phi(x^{k-1}, x^k),
$$
2. Compute $x^{k+1}$ by solving the convex sub-problem
$$
\min_{x \in \mathbb{X}} \left\{ G(x) - \left( L\phi(y^k) - \nabla f(y^k) + \xi^k, x \right) \right\},
$$
where $\xi^k \in \partial h(x^k)$.
3. $k \leftarrow k + 1$.
\Until Stopping criterion.
\end{algorithmic}
\end{algorithm}

Now, let us provide some special choices of the extrapolation parameter. We first consider the special case in which $f$ admits the $L$-Lipschitz continuous gradient.
Hence, we can choose the function $\phi(x) = \frac{1}{2}\|x\|^2$. Condition (8) becomes

$$(L + 1)\beta_k^2\|x^k - x^{k-1}\|^2 \leq \delta L\|x^k - x^{k-1}\|^2.$$ 

Therefore, we choose $\beta_k$ such that $\beta_k \leq \sqrt{\frac{L+1}{L}}$. Moreover, if $f$ is convex, we can take $l = 0$ and $\beta_k \leq \sqrt{\delta}$. In general case, [28, Lemma 4.1] showed that there always exists $\gamma_k > 0$ such that the condition (8) is satisfied for all $\beta_k \in [0, \gamma_k]$. Therefore, $\beta_k$ can be determined by a line search as follows. At each iteration, we initialize $\beta_k = \frac{\mu_k - 1}{\mu_k}$, and while the inequality (8) does not hold, we decrease $\beta_k$ by $\beta_k = \beta_k \eta$ with $\eta \in (0, 1)$, where $\mu_k = \frac{1}{2} \left(1 + \sqrt{1 + 4\mu_{k-1}^2}\right)$ and $\mu_0 = 1$.

### 3.2. Convergence Analysis.

#### 3.2.1. Sub-sequential convergence.

In this subsection, we study the convergence of the DCAe. The first result presents the sub-sequential convergence of the sequence $\{x_k^k\}$ generated by DCAe.

**Theorem 3.1.** Let $\{x_k\}$ be the sequence generated by DCAe. Suppose Assumption 1 is satisfied. The following statements are hold.

a) For $k = 0, 1, \ldots$, we have

$$(10) \quad F(x^{k+1}) \leq F(x^k) + \delta LD_\phi(x^{k-1}, x^k) - LD_\phi(x^k, x^{k+1}).$$

b) $\lim_{k \to +\infty} \|x^k - x^{k-1}\|^2 < +\infty$, and thus $\lim_{k \to +\infty} \|x^k - x^{k-1}\| = 0$. 

c) If $x^*$ is a limit point of $\{x_k\}$, then $x^*$ is a critical point of (1).

**Proof.** a) By the definition of the gradient $\xi^k \in \partial h(x^k)$, we have

$$h(x^{k+1}) \geq h(x^k) + \langle \xi^k, x^{k+1} - x^k \rangle,$$

which is equivalent to

$$-h(x^{k+1}) \leq -h(x^k) - \langle \xi^k, x^{k+1} - x^k \rangle$$

Similarly, we have

$$L\phi(x^{k+1}) - f(x^{k+1}) \geq L\phi(y^k) - f(y^k) + \langle L\nabla \phi(y^k) - \nabla f(y^k), x^{k+1} - y^k \rangle,$$

which is equivalent to

$$(12) \quad f(x^{k+1}) \leq f(y^k) + L(\phi(x^{k+1}) - \phi(y^k)) - \langle L\nabla \phi(y^k) - \nabla f(y^k), x^{k+1} - y^k \rangle.$$ 

Since $x^{k+1}$ is a solution to the convex sub-problem (5), we have

$$L\nabla \phi(y^k) - \nabla f(y^k) + \xi^k - L\nabla \phi(x^{k+1}) \in \partial g(x^{k+1}).$$

This implies that

$$(13) \quad g(x^{k+1}) \leq g(x^k) - \langle L\nabla \phi(y^k) + \xi^k - \nabla f(y^k) - L\nabla \phi(x^{k+1}), x^k - x^{k+1} \rangle$$
Summing the three inequalities above, we get

\[ F(x^{k+1}) \leq g(x^k) - h(x^k) + f(y^k) + L\phi(x^{k+1}) - L\phi(y^k) \]

\[ - \left\langle L\nabla \phi(y^k) - \nabla f(y^k), x^k - y^k \right\rangle - \left\langle L\nabla \phi(x^{k+1}), x^{k+1} - x^k \right\rangle \]

\[ = g(x^k) - h(x^k) + f(y^k) + \left\langle \nabla f(y^k), x^k - y^k \right\rangle + LD_\phi(x^k, y^k) - LD_\phi(x^k, x^{k+1}). \]

On the other hand, by the convexity of \( l\phi + f \), we have

\[ l\phi(x^k) + f(x^k) \geq l\phi(y^k) + f(y^k) + \left\langle l\nabla \phi(y^k) + \nabla f(y^k), x^k - y^k \right\rangle \]

which is equivalent to

\[ f(y^k) + \left\langle \nabla f(y^k), x^k - y^k \right\rangle \leq f(x^k) + lD_\phi(x^k, y^k) \]

Summing the two inequalities (14) and (16) above gives us that

\[ F(x^{k+1}) \leq F(x^k) + (L + l)D_\phi(x^k, y^k) - LD_\phi(x^k, x^{k+1}) \]

\[ \leq F(x^k) + \delta LD_\phi(x^{k-1}, x^k) - LD_\phi(x^k, x^{k+1}), \]

where the second inequality follows from the definition of the parameter \( \beta_k \).

b) Summing the inequality (10) over \( k = 0, ..., N \) gives that

\[ \sum_{k=1}^{N} (1 - \delta)LD_\phi(x^{k-1}, x^k) \leq F(x^0) + LD_\phi(x^{-1}, x^0) - F(x^{N+1}) - LD_\phi(x^N, x^{N+1}) \leq F(x^0) - \alpha, \]

where \( \alpha = \inf_{x \in \mathbb{R}} F(x) \). This inequality and the \( \rho \)-strongly convexity of \( \phi \) imply that

\[ \sum_{k=0}^{N} (1 - \delta)\rho \| x^k - x^{k-1} \| \leq F(x^0) - \alpha \]

Therefore, we have

\[ \sum_{k=0}^{N} \| x^k - x^{k-1} \|^2 \leq \frac{F(x^0) - \alpha}{(1 - \delta)\rho}. \]

By taking to the limit, we have

\[ \sum_{k=0}^{+\infty} \| x^k - x^{k-1} \|^2 < +\infty, \]

and \( \lim_{k \to +\infty} \| x^k - x^{k-1} \| = 0. \)

c) Let \( \{x^{k_j}\} \) be a subsequence of \( \{x^k\} \) that converges to \( x^* \). It follows from (b) and \( 0 \leq \beta_k < 1 \) that \( \lim x^{k_j+1} = \lim x^{k_j+1} = \lim y^{k_j} = x^* \). Without loss generality, we can assume that \( \xi^{k_j} \) converges to \( \xi^* \). By the closedness of \( \partial h(x^{k_j}) \), we get that \( \xi^* \in \partial h(x^*) \). From the definition of \( x^{k_{j+1}} \), we have

\[ L\nabla \phi(y^{k_j}) - \nabla f(y^{k_j}) + \xi^{k_j} - L\nabla \phi(x^{k_j+1}) \in \partial g(x^{k_j+1}). \]
Thanks to the continuity of $\nabla \phi$ and $\nabla f$, and the closedness of $\partial g$, passing to the limit gives us that

\[(22) \quad -\nabla f(x^*) + \xi^* \in \partial g(x^*).\]

Therefore, we obtain that

\[(23) \quad \xi^* \in [\nabla f(x^*) + \partial g(x^*)] \cap \partial h(x^*).\]

This completes the proof. \(\square\)

### 3.2.2. Global convergence.

Denote $\Phi$ be an auxiliary function by

\[(24) \quad \Phi(x, y) = F(x) + (1 + \delta) L D^2 \phi(y, x),\]

The following theorem shows the global convergence of the whole bounded sequence generated by DCAe under an additional assumption below.

**Assumption 2.**

(a) $\Phi$ satisfies the KL property at any point $x \in \text{dom} \partial^L \Phi$.

(b) $\nabla f$ and $\nabla \phi$ are Lipschitz continuous on any bounded subset of $E$.

(c) $h$ is differentiable with $\nabla h$ being Lipschitz continuous on any bounded subset of $E$.

**Theorem 3.2.** Let $\{x^k\}$ be a sequence generated by DCAe. Suppose that Assumptions 1 and 2 are satisfied. If $\{x^k\}$ is bounded, it converges to a critical point of (1).

**Proof.** Denote $z^k = (x^k, x^{k-1})$. We first prove that the sequence $\{z^k\}$ satisfy three conditions $\text{H1}$, $\text{H2}$, and $\text{H3}$ in [4].

**$\text{H1}$ Sufficient decrease condition.** From the inequality (10) and the $\rho$-strongly convexity of $\phi$, we derive that for all $k$

$$\Phi(z^k) - \Phi(z^{k+1}) \geq \frac{(1 - \delta) L \rho}{4} \|z^k - z^{k+1}\|^2.$$

**$\text{H2}$ Relative error condition.** We need to justify that there exists $\kappa$ such that for all $k$

$$\text{dist}(0, \partial^L \Phi(z^{k+1})) \leq \kappa \|z^k - z^{k+1}\|.$$

By the boundedness of $\{x^k\}$ and the locally Lipschitz continuity of $\nabla h, \nabla \phi, \nabla f$, there exists $L_h, L_\phi, L_f > 0$ such that for all $k$

$$\|\nabla h(x^k) - \nabla h(x^{k+1})\| \leq L_h \|x^{k+1} - x^k\|$$

$$\|\nabla \phi(y^k) - \nabla \phi(x^{k+1})\| \leq L_\phi \|y^k - x^{k+1}\|$$

$$\|\nabla f(y^k) - \nabla f(x^{k+1})\| \leq L_f \|y^k - x^{k+1}\|.$$

It follows from the definition of $x^{k+1}$ that

$$\|L \phi(y^k) - \nabla f(y^k) + \nabla h(x^k) - L \phi(x^{k+1}) \in \partial g(x^{k+1}).$$

Therefore, we have

$$\|L \left( \phi(y^k) - \phi(x^{k+1}) \right) + \nabla f(x^{k+1}) - \nabla f(y^k) + \nabla h(x^k) - \nabla h(x^{k+1}) \in \partial^L F(x^{k+1})$$
On the other hand, we have
\[ \partial^2 \Phi(z^{k+1}) = \left( \partial^2 F(x^{k+1}) + (1 + \delta)L/2 \nabla^2 \phi(x^{k+1}), x^{k+1} - x^k \right). \] (28)

From this, (27), and (25), we have
\[
\text{dist} \left( 0, \partial^2 \Phi(z^{k+1}) \right) \leq \| \nabla^2 \phi(x^{k+1}) \| + \| \phi(x) - \phi(z^{k+1}) \| + \| \phi(z^{k+1}) - \phi(x^{k+1}) \|
\]
\[
\leq (LL\phi + L_f)\| x^{k+1} - y^k \| + (Lh + (1 + \delta)L/2(M + L_\phi))\| x^{k+1} - x^k \|
\]
\[
\leq (LL\phi + L_f)\| x^{k+1} - x^k \| + (Lh + (1 + \delta)L/2(M + L_\phi))\| x^{k+1} - x^k \|,
\]
where we have used the facts that \( \| \nabla^2 \phi(x^{k+1}) \| \leq M \) and \( \beta_k \leq 1 \). Therefore, we achieve
\[
\text{dist} \left( 0, \partial^2 \Phi(z^{k+1}) \right) \leq \kappa \| z^k - z^{k+1} \|,
\]
where we have used the Cauchy-Schwarz inequality and \( \kappa = (LL\phi + L_f + L_h + (1 + \delta)L/2(M + L_\phi))\sqrt{2} \).

**H3 Continuity condition.** By the boundedness of \( \{x^k\} \) and the part (b) of Theorem 3.1, there exists a sub-sequence \( \{z^{k_j}\} \subset \{x^k\} \) that converges to a point \( z^* = (x^*, x^*) \). We need to prove that \( \Phi(z^{k_j}) \rightarrow \Phi(z^*) \) as \( j \rightarrow +\infty \). It follows from (b) of Theorem 3.1 and \( 0 \leq \beta_k < 1 \) that \( \lim x^{k_j-1} = \lim x^{k_j-2} = \lim y^{k_j-1} = x^* \). It follows from the definition of \( x^{k_j} \) that
\[ G(x^{k_j}) \leq G(x^*) - (L\phi(y^{k_j-1}) - \nabla f(y^{k_j-1}) + \xi^{k_j-1}, x^* - x^{k_j}) \] (29)

Thanks to the fact that \( \lim x^{k_j-1} = \lim x^{k_j-2} = \lim y^{k_j-1} = x^* \), the continuity of \( \nabla f, \nabla \phi \), and the boundedness of \( \{\xi^{k_j-1}\} \), taking to the limit in (29) gives us that
\[ \limsup_{j \rightarrow +\infty} G(x^{k_j}) \leq G(x^*). \] (30)

Therefore, we have
\[
\limsup_{j \rightarrow +\infty} F(x^{k_j}) = \limsup_{j \rightarrow +\infty} [G(x^{k_j}) - H(x^{k_j})]
\]
\[
\leq \limsup_{j \rightarrow +\infty} G(x^{k_j}) - \liminf_{j \rightarrow +\infty} H(x^{k_j})
\]
\[
\leq G(x^*) - \liminf_{j \rightarrow +\infty} H(x^{k_j})
\]
\[
\leq G(x^*) - H(x^*) = F(x^*).
\] (31)

On the other hand, from the lower semicontinuity of \( F \), we obtain
\[ \liminf_{j \rightarrow +\infty} F(x^{k_j}) \geq F(x^*). \] (32)

Hence, by the uniqueness of limit and the definition of \( \Phi \), we have \( \lim_{j \rightarrow +\infty} \Phi(z^{k_j}) = \lim_{j \rightarrow +\infty} F(x^{k_j}) = F(x^*) = \Phi(x^*) \). The result now follows from the same arguments of the proof for [4, Theorem 2.9].
Moreover, if the function $\psi$ appearing in the KL inequality takes the form $\psi(s) = cs^{1-\theta}$ with $\theta \in [0, 1)$ and $c > 0$, we can derive the convergence rates for the both sequences $\{x^k\}$ and $\{F(x^k)\}$.

Theorem 3.3. Let $\{x^k\}$ be a sequence generated by DCAe. Suppose that the assumptions 1 and 2 are satisfied. If $\{x^k\}$ is bounded and the function $\psi$ appearing in the KL inequality takes the form $\psi(s) = cs^{1-\theta}$ with $\theta \in [0, 1)$ and $c > 0$, the following statements hold.

a) If $\theta = 0$, the sequences $\{x^k\}$ and $\{F(x^k)\}$ converge in a finite number of steps to $x^*$ and $F^*$, respectively.

b) If $\theta \in (0, 1/2]$, the sequences $\{x^k\}$ and $\{F(x^k)\}$ converge linearly to $x^*$ and $F^*$, respectively.

c) If $\theta \in (1/2, 1)$, there exist positive constants $\delta_1$, $\delta_2$, and $N$ such that $\|x^k - x^*\| \leq \delta_1 k^{-\frac{1}{2\theta - 1}}$ and $F(x^k) - F^* \leq \delta_2 k^{-\frac{1}{2\theta - 1}}$ for all $k \geq N$.

Since this theorem can be proved by using the same techniques as in the proofs of [9, Theorem 2] and [16, Theorem 3.4], we omit the detail of the proof.

4. Numerical Experiments. We evaluate DCAe with regard to the nonconvex negative matrix completion problem (3), namely

$$
\min_{U \in \mathbb{R}^{m \times t}, V \in \mathbb{R}^{t \times n}} \left\{ F(U, V) := f(U, V) + \chi_{\mathbb{R}^{m \times t} \times \mathbb{R}^{t \times n}}(U, V) + r(U, V) \right\},
$$

where $f(U, V) := \frac{1}{2}\|P(A - UV)\|_F^2$ is the data-fitting term, $r$ is a regularization term, and $\chi_{\mathbb{R}^{m \times t} \times \mathbb{R}^{t \times n}}$ is the indicator function. Here, $P(Z)_{ij} = Z_{ij}$ if $A_{ij}$ is observed and 0 otherwise. We consider the exponential concave regularization [12]:

$$
r(U, V) = \lambda \left( \sum_{ij} \left( 1 - \exp(-\theta|u_{ij}|) \right) + \sum_{ij} \left( 1 - \exp(-\theta|v_{ij}|) \right) \right),
$$

where $u_{ij}$ are elements of $U$. We choose a function $\phi$ given by

$$
\phi(U, V) = c_1 \left( \frac{\|U\|_F^2 + \|V\|_F^2}{2} \right)^2 + c_2 \left( \frac{\|U\|_F^2 + \|V\|_F^2}{2} \right),
$$

where $c_1 = 3$ and $c_2 = \|P_0(A)\|_F$. In [27], the authors showed that $L\phi - f$ and $L\phi + f$ are convex for all $L \geq 1$. The exponential concave regularization can be expressed as a DC function $r = \lambda \theta \left( \|U\|_1 + \|V\|_1 \right) - h$, where $h$ is defined by

$$
h(U, V) = \lambda \theta \left( \|U\|_1 + \|V\|_1 \right) - r(U, V).
$$

Therefore, the problem (33) takes the form of (1) with $g = \chi_{\mathbb{R}^{m \times t} \times \mathbb{R}^{t \times n}}(U, V) + \lambda \theta \left( \|U\|_1 + \|V\|_1 \right)$. The DCAe algorithm iteratively computes $\xi^k \in \partial h(U^k, V^k)$ by

$$
\xi^k_{u_{ij}} = \lambda \theta(1 - \exp(-\theta|u^k_{ij}|))\text{sign}(u^k_{ij})
$$

$$
\xi^k_{v_{ij}} = \lambda \theta(1 - \exp(-\theta|v^k_{ij}|))\text{sign}(v^k_{ij}),
$$

□
and solves the following convex nonsmooth problem:

\[
\min_{\mathbf{U} \in \mathbb{R}_+^{n \times 1}, \mathbf{V} \in \mathbb{R}_+^{m \times 1}} \left\{ \frac{\lambda}{L}(\|\mathbf{U}\|_1 + \|\mathbf{V}\|_1) + \langle \mathbf{P}^k, \mathbf{U} \rangle + \langle \mathbf{Q}^k, \mathbf{U} \rangle \right. \\
+ \left. c_1 \left( \frac{\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2}{2} \right) + c_2 \left( \frac{\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2}{2} \right) \right\},
\]

where \( \mathbf{P}^k \) and \( \mathbf{Q}^k \) are defined by

\[
\mathbf{P}^k = \nabla_U f(\mathbf{Y}^k, \mathbf{Z}^k)/L - \xi_U^k/L - \nabla_U \phi(\mathbf{Y}^k, \mathbf{Z}^k)
\]

\[
\mathbf{Q}^k = \nabla_V f(\mathbf{Y}^k, \mathbf{Z}^k)/L - \xi_V^k/L - \nabla_V \phi(\mathbf{Y}^k, \mathbf{Z}^k),
\]

with \( \mathbf{Y}^k = \mathbf{U}^k + \beta_k(\mathbf{U}^k - \mathbf{U}^{k-1}) \) and \( \mathbf{Z}^k = \mathbf{V}^k + \beta_k(\mathbf{V}^k - \mathbf{V}^{k-1}) \). The solution to the convex sub-problem (35) is given by \( \mathbf{U}^{k+1} = \tau^* \mathbf{P}^k - \lambda/L \mathbf{I} \) and \( \mathbf{V}^{k+1} = \tau^* \mathbf{Q}^k - \lambda/L \mathbf{I} \), where \( \left( [-\mathbf{P} - \lambda/L \mathbf{I}] \right)_{ij} = \max(0, -p_{ij} - \lambda/L) \) and \( \tau^* \) is the unique positive real root of

\[
c_1 \left( \|[-\mathbf{P} - \lambda/L \mathbf{I}]_+\|_F^2 + \|[-\mathbf{Q} - \lambda/L \mathbf{I}]_+\|_F^2 \right) \tau^3 + c_2 \tau - 1 = 0.
\]

In our experiments, we choose \( L = l = 1 \) and \( \delta = 0.9999 \). At each iteration, we initialize \( \beta_k = \frac{4k - 1}{\mu_k} \), and while the inequality (8) does not hold, we decrease \( \beta_k \) by \( \beta_k = 0.9\beta_k \), where \( \mu_k = \frac{1}{2}(1 + \sqrt{1 + 4\mu_{k-1}^2}) \) and \( \mu_0 = 1 \).

To verify the effect of the extrapolation term, we compare our DCAe algorithm with its non-extrapolated version, which is DCA, and the inertial DCA algorithm (iDCA) [13]. All the experiments were conducted on a PC 2.3 GHz Intel Core i5 of 8GB RAM. The codes were written in MATLAB and are available from https://github.com/nhatpd/DCAe.

We set \( \lambda = 0.1 \) and \( \theta = 5 \) as proposed in [12]. We note that we do not optimize numerical results by tweaking the parameters as this is beyond the scope of this work. It is important noting that we evaluate the algorithms on the same models. We carried out the experiments on the two most widely used datasets in the field of recommendation systems, MovieLens and Netflix, which contain ratings of different users. The characteristics of the datasets are given in Table 1. We respectively choose \( t = 5, 8, \) and 13 for MovieLens 1M, 10M, and Netflix data set. We randomly used 70% of the observed ratings for training and the rest for testing. The process was repeated twenty times. We run each algorithm 20, 200, and 3600 seconds for MovieLens 1M, 10M, and Netflix data set, respectively. We are interested in the root mean squared error on the test set: \( \text{RMSE} = \sqrt{\| \mathbf{P}_T (\mathbf{A} - \mathbf{UV}) \|^2 / N_T} \), where \( \mathbf{P}_T (\mathbf{Z})_{ij} = \mathbf{Z}_{ij} \) if \( \mathbf{A}_{ij} \) belongs to the test set and 0 otherwise, \( N_T \) is the number of ratings in the test set. We plotted the curves of the average value of RMSE and the objective function value.

### Dataset Descriptions

The number of users, items, and ratings used in each dataset

| Dataset   | #users | #items | #ratings       |
|-----------|--------|--------|----------------|
| MovieLens | 1M     | 6,040  | 3,449          |
|           | 10M    | 69,878 | 10,677         |
| Netflix   | 480,189| 17,770 | 100,480,507    |


We observe that DCAe converges the fastest on all the data sets, providing a significant acceleration of DCA and iDCA. DCAe obtains not only the best final objective function values but also the best RMSE on the test set. This illustrates the usefulness of the extrapolation term in DCA.

5. Conclusions. We developed the DCAe algorithm for minimizing the sum of a nonconvex differentiable function and a DC function, a novel variant of DCA with extrapolation that subsumes the pDCAe algorithm in [33] as a special case. We studied the sub-sequential and global convergence of DCAe to a critical point under mild and stronger conditions, respectively. To evaluate the performance of our
proposed algorithm, we applied DCAe to the nonnegative matrix completion problem. The numerical results strongly confirmed the advantage of DCAe to DCA and iDCA.

REFERENCES

[1] M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

[2] N. T. An and N. M. Nam. Convergence analysis of a proximal point algorithm for minimizing differences of functions. *Optimization*, 66(1):129–147, 2017.

[3] H. Attouch, J. Bolte, P. Redont, and 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):438–457, 2010.

[4] H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1):91–129, Feb 2013.

[5] A. Beck and Y. C. Eldar. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *SIAM Journal on Optimization*, 23(3):1480–1509, 2013.

[6] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2:183–202, 2009.

[7] M. Bertero, P. Boccacci, G. Desidera, and G. Vicidomini. Image deblurring with poisson data: from cells to galaxies. *Inverse Problems*, 25(12):123006, Nov 2009.

[8] P. Biswas, T.-C. Lian, T.-C. Wang, and Y. Ye. Semidefinite programming based algorithms for sensor network localization. *ACM Trans. Sen. Netw.*, 2(2):188–220, 2006.

[9] H. A. J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116, 01 2009.

[10] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for non-convex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, 2014.

[11] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd. First order methods beyond convexity and lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018.

[12] P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In *Proceeding of international conference on machine learning ICML’98*, 1998.

[13] W. de Oliveira and M. P. Tcheou. An inertial algorithm for dc programming. *Set-Valued and Variational Analysis*, pages 1–25, 2018.

[14] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[15] R. Fan, Jianqing Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 2001.

[16] P. Frankel, G. Garrigos, and P. Peypouquet. Splitting methods with variable metric for kurdyka–lojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, 165(3):874–900, 2015.
[17] A. T. K. Gotoh, Jun-ya; Takeda. Dc formulations and algorithms for sparse optimization problems. *Mathematical Programming*, 7 2017.

[18] A. Kaplan and R. Tichatschke. Proximal point methods and nonconvex optimization. *Journal of Global Optimization*, 13(4):389–406, Dec 1998.

[19] M. Kim and J. Leskovec. The network completion problem: Inferring missing nodes and edges in networks. In *Proceedings of the 11th International Conference on Data Mining*, 2011.

[20] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[21] H. A. Le Thi and T. Pham Dinh. The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Annals of Operations Research*, 133:23–46, 2005.

[22] H. Li and Z. Lin. Accelerated proximal gradient methods for nonconvex programming. In *Advances in Neural Information Processing Systems*, pages 377–387, 2015.

[23] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013.

[24] Z. Lu, Zhaosong; Zhou. Nonmonotone enhanced proximal dc algorithms for a class of structured nonsmooth dc programming. *SIAM Journal on Optimization*, 29, 01 2019.

[25] Z. S. Z. Lu, Zhaosong; Zhou. Enhanced proximal dc algorithms with extrapolation for a class of structured nonsmooth dc minimization. *Mathematical Programming*, 9 2018.

[26] B. S. Mordukhovich. *Variational Analysis and Generalized Differentiation I*. 330. Springer-Verlag Berlin Heidelberg, 2006.

[27] M. C. Mukkamala and P. Ochs. Beyond alternating updates for matrix factorization with inertial bregman proximal gradient algorithms. In *Advances in Neural Information Processing Systems* 32, pages 4266–4276. 2019.

[28] P. P. T. S. S. Mukkamala, Mahesh Chandra; Ochs. Convex-concave backtracking for inertial bregman proximal gradient algorithms in nonconvex optimization. *SIAM Journal on Mathematics of Data Science*, 2, 2020.

[29] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27:372–376, 1983.

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

[31] D. Phan, H. Le, and H. A. Le Thi. Accelerated difference of convex functions algorithm and its application to sparse binary logistic regression. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence*, pages 1369–1375, 2018.

[32] B. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1 – 17, 1964.

[33] X. P. T. K. Wen, Bo; Chen. A proximal difference-of-convex algorithm with extrapolation. *Computational Optimization and Applications*, 10 2017.

[34] Q. Yao, J. T. Kwok, F. Gao, W. Chen, and T.-Y. Liu. Efficient inexact proximal gradient algorithm for nonconvex problems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI’17, pages 3308–3314, 2017.

[35] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38, 04 2010.