Novel DCA based algorithms for a special class of nonconvex problems with application in machine learning

Hoai An Le Thi, Hoai Minh Le, Duy Nhat Phan, Bach Tran

Version of Record: https://www.sciencedirect.com/science/article/pii/S0096300320308572
Manuscript_3493a330690d36d9cf973ed694fe0f3e

Novel DCA Based Algorithms for a special class of nonconvex problems with Application in Machine learning

Hoai An Le Thi[a,*], Hoai Minh Le[a], Duy Nhat Phan[a], Bach Tran[a]

[a]*Department Computer Science and Application, LGIPM, University of Lorraine, France*

Abstract

We address the problem of minimizing the sum of a nonconvex, differentiable function and composite functions by DC (Difference of Convex functions) programming and DCA (DC Algorithm), powerful tools of nonconvex optimization. The main idea of DCA relies on DC decompositions of the objective function, it consists in approximating a DC (nonconvex) program by a sequence of convex ones. We first develop a standard DCA scheme especially dealing with the very specific structure of this problem. Furthermore, we extend DCA to give rise to the so-named DCA-Like, which is based on a new and efficient way to approximate the DC objective function without knowing a DC decomposition. We further improve DCA based algorithms by incorporating the Nesterov's acceleration technique into them. The convergence properties and the convergence rate under Kurdyka-Łojasiewicz assumption of extended DCAs are rigorously studied. We prove that DCA-Like and the accelerated versions subsequently converge from every initial point to a critical point of the considered problem. Finally, we investigate the proposed algorithms for an important problem in machine learning: the t-distributed stochastic neighbor embedding. Numerical experiments on several benchmark datasets illustrate the efficiency of our algorithms.

Keywords: DC Programming, DCA, DCA-Like, Accelerated DCA, Accelerated DCA-Like, Sum of Nonconvex Function and Composite

*Corresponding author

Email addresses: hoai-an.le-thi@univ-lorraine.fr (Hoai An Le Thi), minh.le@univ-lorraine.fr (Hoai Minh Le), duy-nhat.phan@univ-lorraine.fr (Duy Nhat Phan), bach.tran@univ-lorraine.fr (Bach Tran)

Functions, Dimensionality Reduction, t-distributed Stochastic Neighbor Embedding (t-SNE)

---

1. Introduction

We consider the sum of a nonconvex differentiable function and composite functions minimization problem which takes the form

$$\min_{\mathbf{x}\in X}\left\{F(\mathbf{x}) := f(\mathbf{x}) + \sum_{i=1}^{m} h_i(g_i(\mathbf{x}_i))\right\}, \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is nonconvex differentiable with L-Lipschitz continuous gradient, $g_i : \mathbb{R}^{n_i} \to \mathbb{R}$ are continuous convex (possibly nonsmooth) where $\mathbf{x}_i$ (for $i = 1, \ldots, m$) are sub-vectors of $\mathbf{x}$ with $\sum_{i=1}^{m} n_i = n$, $h_i : \mathbb{R} \to \mathbb{R}$ are concave increasing and $X \subset \mathbb{R}^n$ is closed convex.

The problem (1) is a common mathematical model of numerous classes of problems arising from several domains such as machine learning, computational biology, image processing, etc. For instance, several problems in machine learning are formulated as an optimization problem of the form

$$\min_{\mathbf{x}\in X} f(\mathbf{x}) + \lambda r(\mathbf{x}), \tag{2}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a loss function, r is a regularizer function, and the parameter $\lambda > 0$ makes the trade-off between the two terms f and r. As the function f is nonconvex, it covers several loss functions such as least-squares, hinge loss, ramp loss, logistic loss, negative log-likelihoods, etc. On the other hand, most existing regularization functions in the literature can be expressed as a sum of composite functions $h_i(g_i(\cdot))$. In particular, in learning with sparsity problems (e.g. sparse signal recovery [8], sparse inverse covariance (resp. covariance) estimation [20], sparse binary logistic regression [19]) involving the ℓ_0-norm ($\|\mathbf{x}\|_0$ stands for the number of nonzero components of $\mathbf{x}$), one often approximates the ℓ_0-norm by one of the sparse reducing regularizers ([11]) and then solves the resulting problems which have the form of the problem (1).

Paper's contribution. In this work, we investigate novel and efficient methods based on DC (Difference of Convex functions) programming and DCA (DC Algorithm) to solve (1). DCA was introduced in 1985 by T. Pham Dinh, and extensively developed by H.A. Le Thi and T. Pham Dinh since 1994 to

become now classic and increasingly popular ([9, 10, 17, 16, 18] and references therein). DCA has been successfully applied to various nonconvex/nonsmooth programs thanks to its versatility, flexibility, robustness, inexpensiveness and ability to adapt to the specific structure of considered problems. The readers are referred to the recent paper [10] for an extensive overview of thirty years of development of DCA. Recently, several variants of DCA have been developed in order to improve the standard DCA: Approximate DCA, DCA with successive DC decomposition, Stochastic DCA, Online DCA, Accelerated DCA, etc. This work follows the new trend in the development of DCA, i.e., develop novel DCA variants to improve standard DCA and to deal with very large-scale optimization problems. Our contributions are multiple.

Firstly, we propose a standard DCA scheme for solving (1). With the assumptions on f, g_i, and h_i, one can in most cases show that F is a DC function (see the Section 2.2 below) and then DCA can be applied. However, it is not evident to highlight a DC composition of the composite functions $h_i(g_i(\cdot))$ to get a DC decomposition of F. To tackle this difficulty, we equivalently reformulate (1) as an DC program whose the DC decomposition of the objective function is explicit, and then investigate DCA for solving the equivalent problem. It turns out that in this DCA scheme, one has to compute a parameter μ greater or equal to the L-Lipschitz constant of f. In practice, it is difficult (or even impossible) to determine the exact value of L, and one usually estimates L by a quite large number. However, a large value of L could lead to a bad convex approximation of F, then DCA may converge rapidly to a *biased* critical point.

To overcome this drawback and improve the standard DCA, we propose a new variant of DCA, named DCA-Like. This constitutes our second contribution. DCA-Like is "like" DCA in the sense that they iteratively approximate the DC program (1) by a sequence of convex ones. However, DCA-Like is "unlike" DCA as it relaxes two key requirements in DCA: i) in the manner to decompose $F = G_\rho - H_\rho$, with a parameter ρ, in which H_ρ is not necessarily convex (i.e. we possibly do not have a DC decomposition of F), and ρ is updated so that the value of the surrogate convex function of F at the current solution could be as close as possible to F; ii) the affine minorant of H_ρ may not be a lower bound of H_ρ on the whole space but rather only at the current solution. We prove that, fortunately, in spite of these modifications, the whole bounded sequence $\{\mathbf{x}^k\}$ generated by DCA-Like still converges to a critical point of (1) under some conditions. Furthermore, we demonstrate that, under Kurdyka-Łojasiewicz assumption, the convergence

rate of DCA-Like is at least sublinear $\mathcal{O}(1/k^\alpha)$ with $\alpha > 1$.

Thirdly, we incorporate a Nesterov's acceleration technique into DCA based algorithms and give rise to the so-named Accelerated DCA (ADCA) and Accelerated DCA-Like (ADCA-Like) algorithms. We prove that the accelerated versions enjoy similar convergence properties and convergence rate as DCA and DCA-Like respectively.

Fourthly, we develop the above proposed methods for solving an important problem in machine learning that is the t-distributed Stochastic Neighbor Embedding (t-SNE). t-SNE is an efficient dimensionality reduction technique which has many applications in various areas. We carefully conduct the numerical experiments and provide a comparison of proposed algorithms with existing methods on several benchmark datasets.

The remainder of the paper is organized as follows. In Section 2, we briefly provide an overview of DCA programming and DCA, then develop a standard DCA version for solving (1). DCA-Like and its convergence properties as well as its convergence rate are studied in Section 3 while Section 4 is devoted to the accelerated versions of DCA and DCA-Like. In Section 5 we show how to apply the proposed algorithms to t-SNE. Finally, Section 6 concludes the paper.

2. Standard DCA

Let us first recall some basis notations that will be used in the sequel. The modulus of strong convexity of ϕ on Ω, denoted by $\mu(\phi, \Omega)$ or $\mu(\phi)$ if $\Omega = \mathbb{R}^n$, is given by $\mu(\phi, \Omega) = \sup\{\mu \geq 0 : \phi - (\mu/2)\|.\|^2 \text{ is convex on } \Omega\}$. One says that ϕ is *strongly convex* on Ω if $\mu(\phi, \Omega) > 0$.

For a convex function ϕ, the subdifferential of ϕ at $\mathbf{x}_0 \in \text{dom}\phi := \{\mathbf{x} \in \mathbb{R}^n : \phi(x_0) < +\infty\}$, denoted by $\partial\phi(x_0)$, is defined by

$$\partial\phi(\mathbf{x}_0) := \{\mathbf{y} \in \mathbb{R}^n : \phi(\mathbf{x}) \geq \phi(\mathbf{x}_0) + \langle \mathbf{x} - \mathbf{x}_0, y\rangle, \forall \mathbf{x} \in \mathbb{R}^n\}.$$

The subdifferential $\partial\phi(\mathbf{x}_0)$ generalizes the derivative in the sense that ϕ is differentiable at $\mathbf{x}_0$ if and only if $\partial\phi(\mathbf{x}_0) \equiv \{\nabla_{\mathbf{x}}\phi(\mathbf{x}_0)\}$.

For a given $\mathbf{x}^* \in \text{dom } \sigma$, the Frechet subdifferential [14] of a lower semi-continous function σ at $\mathbf{x}^*$ is defined by

$$\partial^F\sigma(\mathbf{x}^*) = \left\{\mathbf{u} : \liminf_{\mathbf{x}\to\mathbf{x}^*} \frac{\sigma(\mathbf{x}) - \sigma(\mathbf{x}^*) - \langle \mathbf{u}, \mathbf{x} - \mathbf{x}^*\rangle}{\|\mathbf{x} - \mathbf{x}^*\|} \geq 0\right\}.$$

The limiting subdifferential [14] of σ at $\mathbf{x}^* \in \operatorname{dom}\sigma$ is defined by

$$\partial^L\sigma(\mathbf{x}^*) = \left\{\mathbf{u} : \exists \mathbf{x}^k \xrightarrow{\sigma} \mathbf{x}^*, \mathbf{u}^k \in \partial^F\sigma(\mathbf{x}^k), \mathbf{u}^k \to \mathbf{u}\right\},$$

where the notation $\mathbf{x}^k \xrightarrow{\sigma} \mathbf{x}^*$ means that $\mathbf{x}^k \to \mathbf{x}^*$ and $\sigma(\mathbf{x}^k) \to \sigma(\mathbf{x}^*)$. We note that if $\sigma = \phi + \psi$, where ϕ is lower semicontinuous, and ψ is continuously differentiable on a neighborhood of $\mathbf{x}^*$, then

$$\partial^L\sigma(\mathbf{x}^*) = \partial^L\phi(\mathbf{x}^*) + \nabla\psi(\mathbf{x}^*).$$

Let $\eta \in (0, +\infty]$. Denote by $\mathcal{M}_\eta$ the class of continuous concave functions $\psi : [0, \eta) \to [0, +\infty)$ satisfying

- $\psi(0) = 0$, and ψ is continuously differentiable on $(0, \eta)$.
- $\psi'(t) > 0$ for all $t \in (0, \eta)$.

Definition 1. *A lower semicontinuous function σ satisfies the KL property [2] at $\mathbf{u}^* \in$ dom $\partial^L\sigma$ if there exists $\eta > 0$, a neighborhood $\mathcal{V}$ of $\mathbf{u}^*$, and $\psi \in \mathcal{M}_\eta$ such that for all $u \in \mathcal{V} \cap \{u : \sigma(\mathbf{u}^*) < \sigma(\mathbf{u}) < \sigma(\mathbf{u}^*) + \eta\}$, one has*

$$\psi'(\sigma(\mathbf{u}) - \sigma(\mathbf{u}^*))\,dist(0, \partial^L\sigma(\mathbf{u})) \geq 1.$$

The class of functions σ satisfying the KL property at all points in dom $\partial^L\sigma$ is very ample, for example, semi-algebraic, subanalytic, and log-exp functions. In particular, these classes of functions satisfy the KL property with $\psi(s) = cs^{1-\theta}$, for some $\theta \in [0, 1)$ and $c > 0$.

2.1. Overview of DC programming and DCA

DC programming and DCA constitute the backbone of smooth/nonsmooth nonconvex programming and global optimization. A standard DC program takes the form

$$\alpha = \inf\{F(\mathbf{x}) := G(\mathbf{x}) - H(\mathbf{x}) \,|\, \mathbf{x} \in \mathbb{R}^n\} \quad (P_{dc}),$$

where G, H are lower semi-continuous proper convex functions on $\mathbb{R}^n$. Such a function F is called a DC function, and $G - H$ is a DC decomposition of F while G and H are the DC components of F. Note that any convex

constrained DC program can be rewritten in the standard form (P_{dc}) by using the indicator function on C, defined by $\chi_C(\mathbf{x}) = 0$ if $\mathbf{x} \in C$, $+\infty$ otherwise.

$$\inf\{F(\mathbf{x}) := G(\mathbf{x}) - H(\mathbf{x}) : \mathbf{x} \in C\} = \inf\{\chi_C(\mathbf{x}) + G(\mathbf{x}) - H(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n\}.$$

A point $\mathbf{x}^*$ is called a *critical point* of $G - H$, or a generalized Karush-Kuhn-Tucker point (KKT) of (P_{dc}) if $\partial H(\mathbf{x}^*) \cap \partial G(\mathbf{x}^*) \neq \emptyset$.

The main idea of DCA is simple: each iteration k of DCA approximates the concave part $-H$ by its affine majorization (that corresponds to taking $\mathbf{y}^k \in \partial H(\mathbf{x}^k)$) and computes $\mathbf{x}^{k+1}$ by solving the resulting convex problem,

$$\min\{G(\mathbf{x}) - \langle \mathbf{x}, \mathbf{y}^k \rangle : \mathbf{x} \in \mathbb{R}^n\} \quad (P_k).$$

The sequence $\{\mathbf{x}^k\}$ generated by DCA enjoys the following properties ([9, 16]):

(i) The sequence $\{F(\mathbf{x}^k)\}$ is decreasing.

(ii) If $F(\mathbf{x}^{k+1}) = F(\mathbf{x}^k)$, then $\mathbf{x}^k$ is a critical point of (P_{dc}) and DCA terminates at the k-th iteration.

(iii) If $\mu(G) + \mu(H) > 0$ then the series $\{\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2\}$ converges.

(iv) If the optimal value α of (P_{dc}) is finite and the infinite sequence $\{\mathbf{x}^k\}$ is bounded then every limit point of the sequence $\{\mathbf{x}^k\}$ is a critical point of $G - H$.

For a complete study of DC programming and DCA the reader is referred to [9, 10, 17, 16, 18] and the references therein.

2.2. A standard DCA scheme for solving (1)

As f is a differentiable function with L-Lipschitz continuous gradient, it is DC with, for example, the following DC decomposition, for any $\rho \geq L$:

$$f(\mathbf{x}) = \frac{\rho}{2}\|\mathbf{x}\|^2 - \left[\frac{\rho}{2}\|\mathbf{x}\|^2 - f(\mathbf{x})\right]. \tag{3}$$

The second term of F, e.g. $\sum_{i=1}^{m} h_i(g_i(\cdot))$, can be generally rewritten as a DC function. For instance, it has been proved that if $g_i : \mathbb{A} \to \mathbb{B}$ and $h_i : \mathbb{B} \to \mathbb{R}$ are DC (with $\mathbb{A} \subset \mathbb{R}^{n_i}$ being a convex set, $\mathbb{B}$ being a convex and open set) then $h_i(g_i(\cdot))$ is DC on $\mathbb{A}$ [3, 6]. Nevertheless, as mentioned above, it is not easy to highlight a DC decomposition of $h_i(g_i(\mathbf{x}_i))$. To apply standard DCA, we first reformulate the problem (1) as follows:

$$\min\left\{\varphi(\mathbf{x}, \mathbf{z}) := \chi_\Omega(\mathbf{x}, \mathbf{z}) + f(\mathbf{x}) + \sum_{i=1}^{m} h_i(z_i) : (\mathbf{x}, \mathbf{z}) \in \mathbb{R}^n \times \mathbb{R}^m\right\}, \tag{4}$$

where $\Omega := \{(\mathbf{x}, \mathbf{z}) : \mathbf{x} \in X, g_i(\mathbf{x}_i) \leq z_i, i = 1, ..., m\}$. Let $g(\mathbf{x})$ be the function defined by $g(\mathbf{x}) = (g_1(\mathbf{x}_1), ..., g_m(\mathbf{x}_m))$. The problems (1) and (4) are equivalent in the following sense.

Proposition 1. *A point $\mathbf{x}^* \in X$ is a global (resp. local) solution to the problem (1) if and only if $(\mathbf{x}^*, g(\mathbf{x}^*))$ is a global (resp. local) solution to the problem (4).*

The proof of Proposition 1 is given in Appendix A.

Since the problems (1) and (4) are equivalent, in the remaining of the paper, we consider the problem (4) instead of (1). The main advantage of (4) is that the objective function φ does not contain composite functions, and it is easy to highlight its DC decompositions. As h_i is concave, the function $\sum_{i=1}^m h_i$ is concave too. Then using the DC decomposition (3) of f, we get the following DC formulation of (4), for any $\mu \geq L$:

$$\min \{\varphi(\mathbf{x}, \mathbf{z}) = G_\mu(\mathbf{x}, \mathbf{z}) - H_\mu(\mathbf{x}, \mathbf{z}) : (\mathbf{x}, \mathbf{z}) \in \mathbb{R}^n \times \mathbb{R}^m\}, \tag{5}$$

where $G_\mu(\mathbf{x}, \mathbf{z}) := \frac{\mu}{2}\|\mathbf{x}\|^2 + \chi_\Omega(\mathbf{x}, \mathbf{z})$, $H_\mu(\mathbf{x}, \mathbf{z}) := \frac{\mu}{2}\|\mathbf{x}\|^2 - f(\mathbf{x}) - \sum_{i=1}^m h_i(z_i)$. In fact, since Ω is a convex set, the function χ_Ω is convex, and so is $G_\mu(\mathbf{x}, \mathbf{z})$. On the other hand, h_i is concave, and $\frac{\mu}{2}\|\mathbf{x}\|^2 - f(\mathbf{x})$ is convex if $\mu \geq L$. Consequently, $H_\mu(\mathbf{x}, \mathbf{z})$ is convex.

We apply DCA on the DC program (5) as follows. At each iteration k, DCA approximates the second DC component H_μ by its affine minorant

$$H_\mu^k(\mathbf{x}, \mathbf{z}) = H_\mu(\mathbf{x}^k, \mathbf{z}^k) + \langle (\mathbf{x}, \mathbf{z}) - (\mathbf{x}^k, \mathbf{z}^k), (\mathbf{y}^k, \xi^k) \rangle$$

with $(\mathbf{y}^k, \xi^k) \in \partial H_\mu(\mathbf{x}^k, \mathbf{z}^k)$, and computes $(\mathbf{x}^{k+1}, \mathbf{z}^{k+1})$ by solving the following convex problem

$$\min \left\{\varphi_\mu^k(\mathbf{x}, \mathbf{z}) := G_\mu(\mathbf{x}, \mathbf{z}) - H_\mu^k(\mathbf{x}, \mathbf{z}) : (\mathbf{x}, \mathbf{z}) \in \mathbb{R}^n \times \mathbb{R}^m\right\}. \tag{6}$$

This problem is equivalent to (we delete the constant terms in $H_\mu^k(\mathbf{x}, \mathbf{z})$):

$$\min \left\{ \frac{\mu}{2}\|\mathbf{x}\|^2 - \langle \mathbf{y}^k, \mathbf{x} \rangle + \sum_{i=1}^m (-\xi_i^k) z_i : (\mathbf{x}, \mathbf{z}) \in \Omega \right\}, \tag{7}$$

where $\xi_i^k \in \partial(-h_i)(z_i^k)$ and $\mathbf{y}^k = \mu \mathbf{x}^k - \nabla f(\mathbf{x}^k)$.

Proposition 2. *If* $\mathbf{x}^{k+1}$ *is an optimal solution to the following strongly convex problem*

$$\min\{\frac{\mu}{2}\|\mathbf{x}\|^2 - \langle \mathbf{y}^k, \mathbf{x}\rangle + \sum_{i=1}^{m}(-\xi_i^k)g_i(\mathbf{x}_i) : \mathbf{x} \in X\}, \tag{8}$$

then $(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1}))$ *is an optimal solution to* (7).

The proof of Proposition 2 is given in Appendix B. Finally, DCA for solving (4) is described in Algorithm 1.

Algorithm 1: DCA for solving (4)

Input : an initial point $\mathbf{x}_0 \in \mathbb{R}^n$, L - the Lipschits constant of f.
Initialization : Choose $\mu \geq L$ and set $k = 0$.
repeat
1: Compute $\xi_i^k \in \partial(-h_i)(g_i(\mathbf{x}_i^k))$, $\mathbf{y}^k = \mu\mathbf{x}^k - \nabla f(\mathbf{x}^k)$;
2: Compute $\mathbf{x}^{k+1}$ by solving the strongly convex problem (8);
3: $k \leftarrow k + 1$.
until *Stopping criterion*;
Output: $\mathbf{x}^k$ - a critical point of (4).

Remark 1. *Thanks to the use of the variable* $\mathbf{z}$ *in* (4) *and the DC decomposition* (5)*, the above DCA scheme is not affected by the fact that the function* g_i *in the composite function* $h_i(g_i(\cdot))$ *is smooth or not.*

The convergence properties of Algorithm 1 are provided in Theorem 1. We recall that, by the definition of the critical point mentioned above, a point $(\mathbf{x}^*, \mathbf{z}^*) \in \mathbb{R}^n \times \mathbb{R}^m$ is called a critical point of the problem (4) if and only if

$$[(\nabla f(\mathbf{x}^*), 0_m) + \mathcal{N}_\Omega(\mathbf{x}^*, \mathbf{z}^*)] \bigcap [0_n \times \partial(-h_1)(z_1^*) \times ... \times \partial(-h_m)(z_m^*)] \neq \emptyset,$$

where 0_d denotes the zero vector in $\mathbb{R}^d$, and $\mathcal{N}_\Omega(\mathbf{u}^*)$ is the normal cone of Ω at $\mathbf{u}^*$ which is defined by $\mathcal{N}_\Omega(\mathbf{u}^*) = \{\mathbf{v} : \langle \mathbf{v}, \mathbf{u} - \mathbf{u}^*\rangle \leq 0, \forall \mathbf{u} \in \Omega\}$.

Theorem 1. *Let* $\{\mathbf{x}^k\}$ *be the sequence generated by Algorithm 1. The following statements hold.*

(i) The sequence $\{\varphi(\mathbf{x}^k, g(\mathbf{x}^k))\}$ *is decreasing.*

(ii) If $\alpha = \inf \varphi(\mathbf{x}, \mathbf{z}) > -\infty$ *then* $\sum_{k=0}^{+\infty} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 < +\infty$*, and therefore* $\lim_{k\to+\infty} \|\mathbf{x}^{k+1} - \mathbf{x}^k\| = 0$.

(iii) If $\alpha = \inf \varphi(\mathbf{x}, \mathbf{z}) > -\infty$*, then any limit point of* $\{(\mathbf{x}^k, g(\mathbf{x}^k))\}$ *is a critical point of* (4).

These results are direct consequences of Theorem 3 about convergence properties of generic DCA in [16].

3. DCA-Like

In Algorithm 1, if μ is very large, then the convex approximation of the DC objective function F could be bad. Hence, we introduce DCA-Like, aiming to get better convex approximations than the one in DCA. The main idea of DCA-Like is to keep the parameter μ as small as possible while finding a convex approximation of F. DCA-Like relaxes the two key requirements of standard DCA which are i) the second component H_μ must be convex, and ii) the affine minorant of H_μ^k must be a lower bound of H_μ on the whole space (note that in DCA the condition i) and the way to define H_μ^k imply the condition ii)). In fact, at each iteration k, we only need to find μ_k such that

$$H_{\mu_k}(\mathbf{x}^{k+1}, \mathbf{z}^{k+1}) \geq H_{\mu_k}^k(\mathbf{x}^{k+1}, \mathbf{z}^{k+1}), \text{ with } (\mathbf{x}^{k+1}, \mathbf{z}^{k+1}) \in \arg\min \varphi_{\mu_k}^k(\mathbf{x}, \mathbf{z}). \tag{9}$$

The DCA-Like algorithm for solving (4) is described in Algorithm 2.

Algorithm 2: DCA-Like for solving (4)

Input : an initial point $\mathbf{x}_0 \in \mathbb{R}^n$, a small enough positive parameter μ_0, $\eta > 1$ and $0 < \delta < 1$.

Initialization : Set $k = 0$.

repeat

1: Compute $\xi_i^k \in \partial(-h_i)(g_i(\mathbf{x}_i^k))$ and $\nabla f(\mathbf{x}^k)$;

2: Set $\mu_k = \max\{\mu_0, \delta\mu_{k-1}\}$ if $k > 0$;

3: Compute $\mathbf{x}^{k+1}$ by solving (8) with $\mu = \mu_k$ and $\mathbf{y}^k = \mu_k\mathbf{x}^k - \nabla f(\mathbf{x}^k)$;

4: **while** $H_{\mu_k}(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1})) < H_{\mu_k}^k(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1}))$ **do**

  $\mu_k \leftarrow \eta\mu_k$;

  Update $\mathbf{x}^{k+1}$ by solving (8) with $\mu = \mu_k$ and $\mathbf{y}^k = \mu_k\mathbf{x}^k - \nabla f(\mathbf{x}^k)$;

end

5: $k \leftarrow k + 1$.

until *Stopping criterion*;

Output: $\mathbf{x}^k$ - a critical point of (4).

Remark 2. *a) It is easy to show that the while loop in STEP 4 stops after finite steps. Indeed, it follows from the convexity of $-h_i$ that for $i = 1, ..., m$*

$$-h_i(g_i(\mathbf{x}^{k+1})) \geq -h_i(g_i(\mathbf{x}^k)) + \langle \xi_i^k, g_i(\mathbf{x}^{k+1}) - g_i(\mathbf{x}^k) \rangle. \tag{10}$$

Since f has L-Lipschitz gradient, for $\mu_k \geq L$ we have

$$f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{\mu_k}{2} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \geq f(\mathbf{x}^{k+1}). \tag{11}$$

Summing inequalities (10) *and* (11) *we see that the condition* (9) *holds if $\mu_k \geq L$. Therefore, there also exists $\beta > 0$ such that $\mu_k \leq \beta L$ for all k.*

b) The condition (9) *does not imply that μ_k is large enough to ensure the convexity of H_{μ_k}. However, we can prove that the convergence properties of DCA-Like are still guaranteed.*

c) In fact, the condition (9) *can be expressed as*

$$F(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{\mu_k}{2} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 - \langle \xi^k, g(\mathbf{x}^{k+1}) - g(\mathbf{x}^k) \rangle \geq F(\mathbf{x}^{k+1}).$$

d) The hyper parameters δ and η are used to suitably update μ_k. From computational point of view, their values may affect the behavior of DCA-Like. A neutral and reasonable choice could be $\eta = 2$, $\delta = 1/2$. The choice of μ_0 is also important. As we want to keep μ_k as small as possible, we have an interest in choosing a small enough μ_0 (for instance 10^{-6} in our experiments).

We now study the convergence properties of DCA-Like in the following theorem.

Theorem 2. *Let $\{\mathbf{x}^k\}$ be the sequence generated by DCA-Like (Algorithm 2). The following statements hold.*

(i) The sequence $\{\varphi(\mathbf{x}^k, g(\mathbf{x}^k))\}$ is decreasing. Moreover, we have

$$\varphi(\mathbf{x}^k, g(\mathbf{x}^k)) - \varphi(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1})) \geq \frac{\mu_k}{2} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2.$$

(ii) If $\alpha = \inf \varphi(\mathbf{x}, \mathbf{z}) > -\infty$ then $\sum_{k=0}^{+\infty} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 < +\infty$, and therefore $\lim_{k \to +\infty} \|\mathbf{x}^{k+1} - \mathbf{x}^k\| = 0$.

(iii) If $\alpha = \inf \varphi(\mathbf{x}, \mathbf{z}) > -\infty$, then any limit point of $\{(\mathbf{x}^k, g(\mathbf{x}^k))\}$ is a critical point of (4).

Proof. (i) From the μ_k-convexity of G_{μ_k} on $\mathbf{x}$ and $(\mathbf{y}^k, \xi^k) \in \partial G_{\mu_k}(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1}))$, we have

$$\begin{aligned} G_{\mu_k}(\mathbf{x}^k, g(\mathbf{x}^k)) \geq G_{\mu_k}(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1})) + \langle \mathbf{y}^k, \mathbf{x}^k - \mathbf{x}^{k+1} \rangle \\ + \langle \xi^k, g(\mathbf{x}^k) - g(\mathbf{x}^{k+1}) \rangle + \frac{\mu_k}{2} \|\mathbf{x}^k - \mathbf{x}^{k+1}\|^2. \end{aligned}$$

This inequality and $H_{\mu_k}(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1})) \geq H^k_{\mu_k}(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1}))$ imply that

$$\varphi(\mathbf{x}^k, g(\mathbf{x}^k)) - \varphi(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1})) \geq \frac{\mu_k}{2} \|\mathbf{x}^k - \mathbf{x}^{k+1}\|^2. \tag{12}$$

(ii) For $\mu_0 \leq \mu_k$, we obtain from (12) that

$$\varphi(\mathbf{x}^k, g(\mathbf{x}^k)) - \varphi(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1})) \geq \frac{\mu_0}{2} \|\mathbf{x}^k - \mathbf{x}^{k+1}\|^2.$$

Summing the above inequality over $k = 0, ..., N$ we have

$$\varphi(\mathbf{x}^0, g(\mathbf{x}^0)) - \varphi(\mathbf{x}^{N+1}, g(\mathbf{x}^{N+1})) \geq \frac{\mu_0}{2} \sum_{k=0}^{N} \|\mathbf{x}^k - \mathbf{x}^{k+1}\|^2. \tag{13}$$

It follows from (13), $\mu_0 > 0$ and $\varphi(\mathbf{x}^{N+1}, g(\mathbf{x}^{N+1})) \geq \alpha$ that

$$\frac{2}{\mu_0}[\varphi(\mathbf{x}^0, g(\mathbf{x}^0)) - \alpha] \geq \sum_{k=0}^{N} \|\mathbf{x}^k - \mathbf{x}^{k+1}\|^2.$$

Finally, passing to the limit over the sequence $\{N\}_{N \in \mathbb{N}}$ we get $\sum_{k=0}^{+\infty} \|\mathbf{x}^k - \mathbf{x}^{k+1}\|^2 < +\infty$, and therefore $\lim_{k \to +\infty} \|\mathbf{x}^{k+1} - \mathbf{x}^k\| = 0$.
(iii) Let $(\mathbf{x}^*, \mathbf{z}^*)$ be a limit point of the sequence $\{(\mathbf{x}^k, g(\mathbf{x}^k))\}$. By the continuity of g, we can take a subsequence $\{\mathbf{x}^{k_j}\}$ of $\{\mathbf{x}^k\}$ that converges to $\mathbf{x}^*$ and $\mathbf{z}^* = g(\mathbf{x}^*) = \lim_{j \to +\infty} g(\mathbf{x}^{k_j})$. It follows from (ii) that $\lim_{j \to +\infty} \mathbf{x}^{k_j+1} = \mathbf{x}^*$. In addition, without loss of generality, we suppose that the subsequence $\{\xi^{k_j}\}$ converges to ξ^*. By the continuity of g_i and closed property of the subdifferential mapping $\partial(-h_i)$, we have $\xi_i^* \in \partial(-h_i)(g_i(\mathbf{x}_i^*))$.
On another hand, since $(\mathbf{x}^{k_j+1}, g(\mathbf{x}^{k_j+1}))$ is an optimal solution of the convex problem

$$\min \left\{ G_{\mu_{k_j}}(\mathbf{x}, \mathbf{z}) - \langle \mathbf{y}^{k_j}, \mathbf{x} \rangle - \langle \xi^{k_j}, \mathbf{z} \rangle \right\}, \tag{14}$$

we have

$$\begin{aligned} &\tfrac{\mu_{k_j}}{2} \|\mathbf{x}^{k_j+1}\|^2 + \chi_\Omega(\mathbf{x}^{k_j+1}, g(\mathbf{x}^{k_j+1}) - \langle \mu_{k_j} \mathbf{x}^{k_j} - \nabla f(\mathbf{x}^{k_j}), \mathbf{x}^{k_j+1} \rangle - \langle \xi^{k_j}, g(\mathbf{x}^{k_j+1}) \rangle \\ &\leq \tfrac{\mu_{k_j}}{2} \|\mathbf{x}^*\|^2 + \chi_\Omega(\mathbf{x}^*, g(\mathbf{x}^*)) - \langle \mu_{k_j} \mathbf{x}^{k_j} - \nabla f(\mathbf{x}^{k_j}), \mathbf{x}^* \rangle - \langle \xi^{k_j}, g(\mathbf{x}^*) \rangle. \end{aligned}$$

Therefore

$$\begin{array}{ll}\frac{\mu_{k_j}}{2}\|\mathbf{x}^{k_j+1}-\mathbf{x}^{k_j}\|^2 & +\chi_\Omega(\mathbf{x}^{k_j+1},g(\mathbf{x}^{k_j+1}))\leq\frac{\mu_{k_j}}{2}\|\mathbf{x}^*-\mathbf{x}^{k_j}\|^2+\chi_\Omega(\mathbf{x}^*,g(\mathbf{x}^*))\\ & +\langle\nabla f(\mathbf{x}^{k_j}),\mathbf{x}^*-\mathbf{x}^{k_j+1}\rangle-\langle\xi^{k_j},g(\mathbf{x}^*)-g(\mathbf{x}^{k_j+1})\rangle.\end{array}$$

From Remark 2 (a), the sequence $\{\mu_{k_j}\}$ is bounded, i.e., $\mu_0 \leq \mu_k \leq \beta L$. Hence, taking $j \to +\infty$ we get

$$\limsup_{j\to+\infty}\chi_\Omega(\mathbf{x}^{k_j+1},g(\mathbf{x}^{k_j+1}))\leq\chi_\Omega(\mathbf{x}^*,g(\mathbf{x}^*)). \tag{15}$$

Combining (15) with the lower semi-continuity of the function χ we get

$$\limsup_{j\to+\infty}\chi_\Omega(\mathbf{x}^{k_j+1},g(\mathbf{x}^{k_j+1}))=\chi_\Omega(\mathbf{x}^*,g(\mathbf{x}^*)). \tag{16}$$

Similarly, it follows from (14) and $\mu_0 \leq \mu_k \leq \beta L$ that

$$\begin{array}{ll}\frac{\mu_0}{2}\|\mathbf{x}^{k_j+1}-\mathbf{x}^{k_j}\|^2 & +\chi_\Omega(\mathbf{x}^{k_j+1},g(\mathbf{x}^{k_j+1}))\leq\frac{\beta L}{2}\|\mathbf{x}-\mathbf{x}^{k_j}\|^2+\chi_\Omega(\mathbf{x},\mathbf{z})\\ & +\langle\nabla f(\mathbf{x}^{k_j}),\mathbf{x}-\mathbf{x}^{k_j+1}\rangle-\langle\xi^{k_j},\mathbf{z}-g(\mathbf{x}^{k_j+1})\rangle.\end{array} \tag{17}$$

Passing to the limit and combining with (16) we get

$$\chi_\Omega(\mathbf{x}^*,g(\mathbf{x}^*))\leq\frac{\beta L}{2}\|\mathbf{x}-\mathbf{x}^*\|^2+\chi_\Omega(\mathbf{x},\mathbf{z})+\langle\nabla f(\mathbf{x}^*),\mathbf{x}-\mathbf{x}^*\rangle-\langle\xi^*,\mathbf{z}-g(\mathbf{x}^*)\rangle.$$

This implies

$$0\in(\nabla f(\mathbf{x}^*),-\xi^*)+\mathcal{N}_\Omega(\mathbf{x}^*,g(\mathbf{x}^*)). \tag{18}$$

It follows from (18) and $\xi_i^* \in \partial(-h_i)(g_i(\mathbf{x}_i^*))$ that

$$\begin{aligned}(0,\xi^*)\in&[(\nabla f(\mathbf{x}^*),0_m)+\mathcal{N}_\Omega(\mathbf{x}^*,g(\mathbf{x}^*))]\cap\\ &[0_n\times\partial(-h_1)(g_1(\mathbf{x}_1^*))\times\ldots\times\partial(-h_m)(g_m(\mathbf{x}_m^*))].\end{aligned}$$

Therefore, $(\mathbf{x}^*, g(\mathbf{x}^*))$ is a critical point of (4). ■

Remark 3. *Since the basic convergence properties of DCA were proved by using the conjugate of convex functions, we cannot use these techniques in the case that the second component may be nonconvex. Unlike DCA, we justified the convergence properties of DCA-Like based on three key facts:*

i) There exists $\beta > 0$ such that $\mu_0 \leq \mu_k \leq \beta L$ for all k.

ii) $\varphi(\mathbf{x}^k, g(\mathbf{x}^k)) - \varphi(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1})) \geq \frac{\mu_k}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2$ *for all* k.

iii) If h_i *is differentiable with locally Lipschitz derivative, there exists a non-negative integer* N *and* $\pi > 0$ *such that*

$$dist(0, \partial^L \varphi(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1}))) \leq (\mu_k + L + \pi)\|\mathbf{x}^{k+1} - \mathbf{x}^k\| \ \forall k > N \quad (19)$$

Clearly, these key properties were obtained without the convexity of the second component.

We now study the convergence properties of the sequence generated by DCA-Like under the Kurdyka-Łojasiewicz (KL) assumption. In Theorem 3, we provide the sufficient conditions that guarantee the convergence of the whole sequence $\{\mathbf{x}^k\}$ generated by DCA-Like. These conditions include the KL property of φ and the differentiability with locally Lipschitz derivative of h_i. Moreover, if the function ψ appearing in the KL inequality has the form $\psi(s) = cs^{1-\theta}$ with $\theta \in [0, 1)$ and $c > 0$, then we obtain the rates of convergence for both sequences $\{\mathbf{x}^k\}$ and $\{\varphi(\mathbf{x}^k, g(\mathbf{x}^k))\}$.

Theorem 3. *Suppose that* $\inf \varphi(\mathbf{x}, \mathbf{z}) > -\infty$, *and* h_i *is differentiable with locally Lipschitz derivative. Assume further that* φ *has the KL property at any point* $(\mathbf{x}, \mathbf{z}) \in dom\ \partial^L \varphi$. *If* $\{\mathbf{x}^k\}$ *generated by DCA-Like is bounded, then the whole sequence* $\{\mathbf{x}^k\}$ *converges to* $\mathbf{x}^*$ *and* $(\mathbf{x}^*, g(\mathbf{x}^*))$ *is a critical point of* (4). *Moreover, if the function* ψ *appearing in the KL inequality has the form* $\psi(s) = cs^{1-\theta}$ *with* $\theta \in [0, 1)$ *and* $c > 0$, *then the following statements hold:*

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

ii) If $\theta \in (0, 1/2]$, *then the sequences* $\{\mathbf{x}^k\}$ *and* $\{\varphi(\mathbf{x}^k, g(\mathbf{x}^k))\}$ *converge linearly to* $\mathbf{x}^*$ *and* φ^*, *respectively.*

iii) If $\theta \in (1/2, 1)$, *then there exist positive constants* δ_1, δ_2, *and* N_0 *such that* $\|\mathbf{x}^k - \mathbf{x}^*\| \leq \delta_1 k^{-\frac{1-\theta}{2\theta-1}}$ *and* $\varphi(\mathbf{x}^k, g(\mathbf{x}^k)) - \varphi^* \leq \delta_2 k^{-\frac{1}{2\theta-1}}$ *for all* $k \geq N_0$.

The proof of Theorem 3 is given in Appendix C.

4. Accelerated versions of DCA based algorithms

4.1. Accelerated DCA

The first work using the Nesterov's acceleration in DCA has been developed in the conference paper [19] for the standard DC program of the form (P_{dc})

given in Section 2. The idea is to find a point $\mathbf{w}^k$ which is better than $\mathbf{x}^k$, based on which $\mathbf{x}^{k+1}$ will be determined. In this work, we use the same idea for the problem (5), but the acceleration affects only on the variable x and not on the variable z. We consider $\mathbf{w}^k$ as an extrapolated point of the current iterate $\mathbf{x}^k$ and the previous iterate $\mathbf{x}^{k-1}$:

$$\mathbf{w}^k = \mathbf{x}^k + \frac{t_k - 1}{t_{k+1}}\left(\mathbf{x}^k - \mathbf{x}^{k-1}\right), \text{ where } t_{k+1} = (1 + \sqrt{1 + 4t_k^2})/2.$$

If $\mathbf{w}^k$ is better than the last iterate x^k, i.e., $\varphi(\mathbf{w}^k, g(\mathbf{w}^k)) \leq \varphi(\mathbf{x}^k, g(\mathbf{x}^k))$ then $\mathbf{w}^k$ will be used instead of $\mathbf{x}^k$ to compute $\mathbf{x}^{k+1}$. Note that, our technique does not require any particular property of the sequence $\{t_k\}$. Here we choose the above sequence as it results to interesting convergence rate [4].

The accelerated version of DCA, named ADCA, is described in Algorithm 3. In a similar way to the proof of convergence properties in [19], we can prove that any limit point of the sequence generated by ADCA is a critical point of (4).

Algorithm 3: ADCA for solving (4)

Input : an initial point $\mathbf{x}_0 \in \mathbb{R}^n$, L - the Lipschits constant of f.
Initialization : Set $\mathbf{w}^0 = \mathbf{x}^0$, $\mu \geq L$, $t_0 = (1 + \sqrt{5})/2$ and $k = 0$.
repeat
1: If $\varphi(\mathbf{w}^k, g(\mathbf{w}^k)) \leq \varphi(\mathbf{x}^k, g(\mathbf{x}^k))$ then set $\mathbf{v}^k = \mathbf{w}^k$; else set $\mathbf{v}^k = \mathbf{x}^k$.
2: Compute $\xi_i^k \in \partial(-h_i)(z_i^k)$ with $z_i^k = g_i(\mathbf{v}_i^k)$ and $\mathbf{y}^k = \mu\mathbf{v}^k - \nabla f(\mathbf{v}^k)$.
3: Compute $\mathbf{x}^{k+1}$ by solving the strongly convex problem (8).
4: Compute $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$ and $\mathbf{w}^{k+1} = \mathbf{x}^{k+1} + \frac{t_k - 1}{t_{k+1}}\left(\mathbf{x}^{k+1} - \mathbf{x}^k\right)$.
5: $k \leftarrow k + 1$.
until *Stopping criterion*;
Output: $\mathbf{x}^k$ - a critical point of (4).

4.2. Accelerated DCA-Like

Similarly to ADCA, we introduce an accelerated version of DCA-Like (ADCA-Like), which is described in Algorithm 4.

In Theorem 4, we provide the convergence properties of ADCA-Like whose proof is given in Appendix D.

Algorithm 4: ADCA-Like for solving (4)

Input : an initial point $\mathbf{x}_0 \in \mathbb{R}^n$, a small enough positive parameter μ_0, $\eta > 1$ and $0 < \delta < 1$.

Initialization : Set $\mathbf{w}^0 = \mathbf{x}^0$, $t_0 = (1+\sqrt{5})/2$ and $k = 0$.

repeat

1: If $\varphi(\mathbf{w}^k, g(\mathbf{w}^k)) \leq \varphi(\mathbf{x}^k, g(\mathbf{x}^k))$ then set $\mathbf{v}^k = \mathbf{w}^k$; else set $\mathbf{v}^k = \mathbf{x}^k$.

2: Compute $\xi_i^k \in \partial(-h_i)(z_i^k)$ with $z_i^k = g_i(\mathbf{v}_i^k)$ and $\nabla f(\mathbf{v}^k)$;

3: Set $\mu_k = \max\{\mu_0, \delta\mu_{k-1}\}$ if $k > 0$;

4: Compute $\mathbf{x}^{k+1}$ by solving (8) with $\mu = \mu_k$ and $\mathbf{y}^k = \mu_k\mathbf{v}^k - \nabla f(\mathbf{v}^k)$;

5: **while** $H_{\mu_k}(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1})) < H_{\mu_k}^{(\mathbf{v}^k, g(\mathbf{v}^k))}(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1}))$ **do**

$\mu_k \leftarrow \eta\mu_k$;

Update $\mathbf{x}^{k+1}$ by solving (8) with $\mu = \mu_k$ and $\mathbf{y}^k = \mu_k\mathbf{v}^k - \nabla f(\mathbf{v}^k)$;

end

6: Compute $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$ and $\mathbf{w}^{k+1} = \mathbf{x}^{k+1} + \frac{t_k - 1}{t_{k+1}}\left(\mathbf{x}^{k+1} - \mathbf{x}^k\right)$;

7: $k \leftarrow k + 1$.

until *Stopping criterion*;

Output: $\mathbf{x}^k$ - a critical point of (4).

Theorem 4. *Let $\{x^k\}$ be the sequence generated by Algorithm 4. The following statements hold.*

(i) The sequence $\{\varphi(\mathbf{x}^k, g(\mathbf{x}^k))\}$ is decreasing. More precisely, we have

$$\varphi(\mathbf{x}^k, g(\mathbf{x}^k)) - \varphi(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1})) \geq \frac{\mu_k}{2}\|\mathbf{x}^{k+1} - \mathbf{v}^k\|^2.$$

(ii) If $\alpha = \inf \varphi(\mathbf{x}, \mathbf{z}) > -\infty$ then $\sum_{k=0}^{+\infty} \|\mathbf{x}^{k+1} - \mathbf{v}^k\|^2 < +\infty$ and therefore $\lim_{k\to+\infty} \|\mathbf{x}^{k+1} - \mathbf{v}^k\| = 0$.

(iii) If $\alpha = \inf \varphi(\mathbf{x}, \mathbf{z}) > -\infty$, then any limit point of $\{(\mathbf{x}^k, g(\mathbf{x}^k))\}$ is a critical point of (4).

The sufficient descent property (i) of Theorem 4 is different from Theorem 2 due to the intermediate variable $\mathbf{v}^k$. Hence, neither the convergence of the whole sequence $\{\mathbf{x}^k\}$ nor convergence rate for $\{\|\mathbf{x}^k - \mathbf{x}^*\|\}$ can be achieved. However, we can still obtain some interesting results for the

sequence $\{\varphi(\mathbf{x}^k, g(\mathbf{x}^k))\}$ under the KL assumption. These properties are presented in Theorem 5.

Theorem 5. *Suppose that* $\inf \varphi(\mathbf{x}, \mathbf{z}) > -\infty$, *and* h_i *is differentiable with locally Lipschitz derivative. Assume further that* φ *has the KL property at any point* $(\mathbf{x}, \mathbf{z}) \in dom\ \partial^L \varphi$ *with* $\psi(s) = cs^{1-\theta}$ *for some* $\theta \in [0, 1)$ *and* $c > 0$. *If* $\{\mathbf{x}^k\}$ *generated by accelerated DCA-Like is bounded, then the following statements hold.*

(i) If $\theta = 0$, *then* $\{\varphi(\mathbf{x}^k, g(\mathbf{x}^k))\}$ *converges in a finite number of steps to* φ^*.

(ii) If $\theta \in (0, 1/2]$, *then* $\{\varphi(\mathbf{x}^k, g(\mathbf{x}^k))\}$ *converges linearly to* φ^*.

(iii) If $\theta \in (1/2, 1)$, *then there exist positive constants* δ *and* N_0 *such that* $\varphi(\mathbf{x}^k, g(\mathbf{x}^k)) - \varphi^* \leq \delta k^{-\frac{1}{2\theta-1}}$ *for all* $k \geq N_0$.

The proof of Theorem 5 is provided in Appendix E.

5. Application on the t-distributed Stochastic Neighbor Embedding problem

In this section, we develop the four proposed DCA based algorithms for solving the t-distributed Stochastic Neighbor Embedding (t-SNE) problem. t-SNE is a machine learning algorithm which models each high-dimensional object by a lower dimensional point (specifically two or three dimensions in case of visualization) in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability. t-SNE was first introduced in [13] as a visualization technique for high dimensional data, and later, it has been applied in many applications such as bioinformatic, cancer research, feature visualization in neural networks, etc.

More precisely, given a data set of N (high-dimensional) input objects $\mathcal{D} = \{\mathbf{a}_1, ..., \mathbf{a}_N\}$ with $\mathbf{a}_i \in \mathbb{R}^d$, the t-SNE aims to find, for each input object $\mathbf{a}_i$, a low-dimensional embedding point $\mathbf{x}_i \in \mathbb{R}^s$ in a way that respects similarities between points. To this end, t-SNE first defines joint probabilities p_{ij} (using Gaussian distribution) and q_{ij} (using a Student t-distribution with one degree of freedom ([13])) that measure, respectively, the pairwise similarity between objects $\mathbf{a}_i$ and $\mathbf{a}_j$ in the original space, and the one between two points $\mathbf{x}_i$ and $\mathbf{x}_j$ in the embedding space $\mathcal{E}$, which are given by

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N};\ p_{j|i} = \frac{\exp(-\|\mathbf{a}_i - \mathbf{a}_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{a}_i - \mathbf{a}_k\|^2/2\sigma_i^2)} \quad \text{if } i \neq j,\ 0 \text{ otherwise,}$$

where σ_i is the variance of the Gaussian that is centered at a_i, and

$$q_{ij} = \frac{(1 + \|\mathbf{x}_i - \mathbf{x}_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|\mathbf{x}_k - \mathbf{x}_l\|^2)^{-1}} \text{ if } i \neq j, \text{ and 0 otherwise.}$$

Then t-SNE learns a s-dimension map $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^s$ which minimizes the Kullback-Leibler divergence between the two joint distributions $P = (p_{ij})$ and $Q = (q_{ij})$:

$$\min_{\mathbf{x}} \left\{ F(\mathbf{x}) = KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \right\}. \tag{20}$$

The nonconvex optimization problem (20) has been studied in several works ([13, 23, 22]), and the most noticeable was presented in [24] where the authors presented and compared Majorization Minimization algorithm (MM) with five state-of-the-arts methods, such as gradient descent, gradient descent with momentum [13], spectral direction [22], FPHSSNE [23] and Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS [15]. The numerical results showed that MM [24] outperforms all five state-of-the-art optimization methods.

First, we show that (20) takes the form of (1). Indeed, the function F in (20) can be rewritten as

$$F(\mathbf{x}) = f(\mathbf{x}) + \sum_{i,j} h_{ij}(g_{ij}(\mathbf{x}_i, \mathbf{x}_j)), \tag{21}$$

where

$$f(\mathbf{x}) := \sum_{i \neq j} p_{ij} \log p_{ij} + \log(\sum_{i \neq j}(1 + \|\mathbf{x}_i - \mathbf{x}_j\|^2)^{-1}), \tag{22}$$

$$h_{ij}(t) = p_{ij} \log(1 + t),\ g_{ij}(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|^2. \tag{23}$$

It is obvious that g_{ij} are convex functions, and h_{ij} are concave increasing functions whose derivatives are non-negatives and Lipschitz continuous on $[0, +\infty)$. Moreover, the function f is differentiable with L-Lipschitz continuous gradient by the following proposition.

Proposition 3. *The function $f(\mathbf{x})$ in (22) is smooth with Lipschitz gradient, where we can choose a Lipschitz constant $L = 4$.*

The proof of Proposition 3 is provided in Appendix F.

As (20) takes the form of (1), we can investigate our proposed DCA based algorithms to solve it. Note that our algorithms are also applicable for other variants of SNE such as SNE [7], Symmetric SNE [13], etc.

For applying DCA based algorithms on (20), we have to specify the computations of t_{ij}^k, $\nabla f(\mathbf{x}^k)$, and $\mathbf{x}^{k+1}$ - the solution of convex subproblems in Algorithms 1–4.

From the definitions of f and h_{ij}, g_{ij} in (22) and (23), the computations of t_{ij}^k and $\nabla f(\mathbf{x}^k)$ are given by

$$t_{ij}^k = \frac{-p_{ij}}{1+\|\mathbf{x}_i^k - \mathbf{x}_j^k\|^2};\ \nabla f(\mathbf{x}^k) = \sum_{j=1}^{N} \frac{-4(\mathbf{x}_i^k - \mathbf{x}_j^k)(1+\|\mathbf{x}_i^k - \mathbf{x}_j^k\|^2)^{-2}}{\sum_{l\neq m}(1+\|\mathbf{x}_l^k - \mathbf{x}_m^k\|^2)^{-1}}, \quad (24)$$

while $\mathbf{x}^{k+1}$ is the optimal solution of the following convex problem

$$\min_{\mathbf{x}}\{\frac{\mu_k}{2}\|\mathbf{x}-\mathbf{x}^k\|^2 + \langle \nabla f(\mathbf{x}^k), \mathbf{x}\rangle + \sum_{i,j} -t_{ij}^k \|\mathbf{x}_i - \mathbf{x}_j\|^2\}. \quad (25)$$

The solution $\mathbf{x}^{k+1}$ of (25) can be explicitly computed as

$$\mathbf{x}^{k+1} = (2\mathcal{L}_{-\mathbf{T}^k-(\mathbf{T}^k)^T} + \mu_k I)^{-1}(-\nabla f(\mathbf{x}^k) + \mu_k \mathbf{x}^k), \quad (26)$$

where the matrix $\mathbf{T}^k$ is defined by the elements t_{ij}^k and $\mathcal{L}_A$ is the matrix given by $(\mathcal{L}_A)_{ij} = -A_{ij}$ if $i \neq j$ and $-A_{ii} + \sum_{l=1}^{N} A_{il}$ otherwise.

From the update rule (26) for $\mathbf{x}^{k+1}$ and this stopping criterion for searching μ_k, we see that algorithm MM in [24] for (20) is a special version of DCA-Like (see Appendix G for more details). On another hand, it is worth noting that all the above DCA based algorithms are in explicit form and very inexpensive.

We recall that all semi-algebraic functions and subanalysis functions satisfy the KL property [2], for examples, real polynomial functions, logarithm function, ℓ_p-norm with $p \geq 0$. In addition, finite sums, products, generalized inverse, compositions of semi-algebraic functions are also semi-algebraic. This implies that the objective function of (20) satisfies the KL property. Hence, DCA-Like and ADCA-Like for solving (20) enjoy all convergence properties provided in Theorems 2–5.

Experiment setting

Recall that Yang et al. (2015) [24] have developed a MM method for t-SNE and showed that their method outperformed many other methods for

t-SNE such as gradient descent, gradient descent with momentum, spectral direction, FPHSSNE, etc. And as MM for t-SNE ([24]) is a special case of DCA-Like, we do not need to compare our methods with other non DCA-based methods. Hence, we intend to compare our methods with other DCA-based methods. DC Proximal Newton (DC-PN) [21]) is chosen since it is the most recent paper and has been successfully applied to several non-convex optimization problems, especially in Machine Learning. DC-PN addresses the problem of minimizing $f(\mathbf{x}) + h(\mathbf{x})$, where $f(\mathbf{x}) = f_1(\mathbf{x}) - f_2(\mathbf{x})$ is a DC function, twice differentiable, with $f(\mathbf{x})$ verifying the L-Lipschitz gradient property, and $h(\mathbf{x}) = h_1(\mathbf{x}) - h_2(\mathbf{x})$ where both h_1 and h_2 are convex functions and (possibly) non-differentiable. Thus, we perform comparative numerical experiments on 5 algorithms - DCA, DCA-Like, ADCA, ADCA-Like and DC-PN on a PC Intel (R) Xeon (R) E5-2630 v4 @2.20 GHz of 32GB RAM.

For solving the convex sub-problem (25), we use Conjugate Gradient method, implemented in the package *cgLanczos* (`https://web.stanford.edu/group/SOL/software/cgLanczos/`). In DC-PN, L-BFGS is employed for approximating the Hessian matrix H_k; and the sub-problem is solved by `minFunc`, a unconstrained differentiable multivariate optimization solver in Matlab (`http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html`).

We consider six real datasets taken from UCI data repository (*gisette*, *usps*, *magic*, *letters*, *shuttle*, *sensorless*, *mnist*, *miniboone* and *covertype*).

As mentioned before, in DCA and ADCA schemes for solving (20) we have to estimate the L-Lipschitz constant of f. According to Proposition 3, we can choose $L = 4$. However, in practice, this value is still large and consequently DCA could converge rapidly to a bad solution. Hence, we incorporate a μ updating procedure into DCA and ADCA. We start with a small value of μ and increase μ if the objective value increases in DCA/ADCA scheme. For all algorithms, the initial value of μ is set to be $\mu_0 = 10^{-6}$. We set $\eta = \frac{1}{\delta} = 2$ as proposed in [24].

We adopt the same test procedure as described in [24]. The initial point $\mathbf{x}^0$ is drawn from normal distribution $\mathcal{N}(0, 10^{-8})$ for all methods. We use the early exaggeration technique as proposed in [13] which consists in running the algorithm in 2 stages. In the first stage, we run the algorithm from the initial point $\mathbf{x}^0$ with modified value of p_{ij} (for instance, p_{ij} is multiplied by 4). Then the solution of the first stage is used as a starting point for the second stage in which we use the true value of p_{ij}. By using large value of p_{ij} in the first stage, the algorithm is encouraged to focus on modelling p_{ij} by fairly large q_{ij} in low dimensional space and consequently it tends to form tight

widely separated clusters in the map. This creates a lot of empty space in the map, which makes it much easier for the clusters to move around in order to find a good global organization in the second stage ([13]).
Recent research on t-SNE suggests using Barnes-Hut tree approximation to reduce the running time by computing approximately the objective function and gradient in t-SNE [12]. This technique is well-known in Neighbor Embedding problems, which allows to reduce greatly the running time while having a small loss in gradients and the objective function. The parameter $\theta_{\text{Barnes-Hut}}$ is set to be 0.5. It has also been shown that Barnes-Hut tree approximation works very well with sparse matrix P. In our experiment, k-Nearest Neighbor (with $k = 10$) is employed to construct a sparse matrix P as follows: $p_{ij} = \frac{\bar{p}_{ij}}{\sum_{k,l} \bar{p}_{kl}}$ where $\bar{p}_{ij} = 1$ if data point j (reps. i) is one of k nearest neighbors of data point j (reps. i) and $\bar{p}_{ij} = 0$ otherwise. We use the same way to compute variances σ_i of Gaussian centered at a_i as in [13]. More precisely, one performs a binary search for the value of σ_i that produces a probability distribution P_i with a fixed perplexity given by the user.

The stopping conditions of all algorithms are the same, by either (a) number of iterations exceeds 10000 or (b) $\|\mathbf{x}^k - \mathbf{x}^{k-1}\|/\|\mathbf{x}^{k-1}\| \leq 10^{-8}$. Throughout our experiment, the number of embedding dimension is set to $s = 2$.

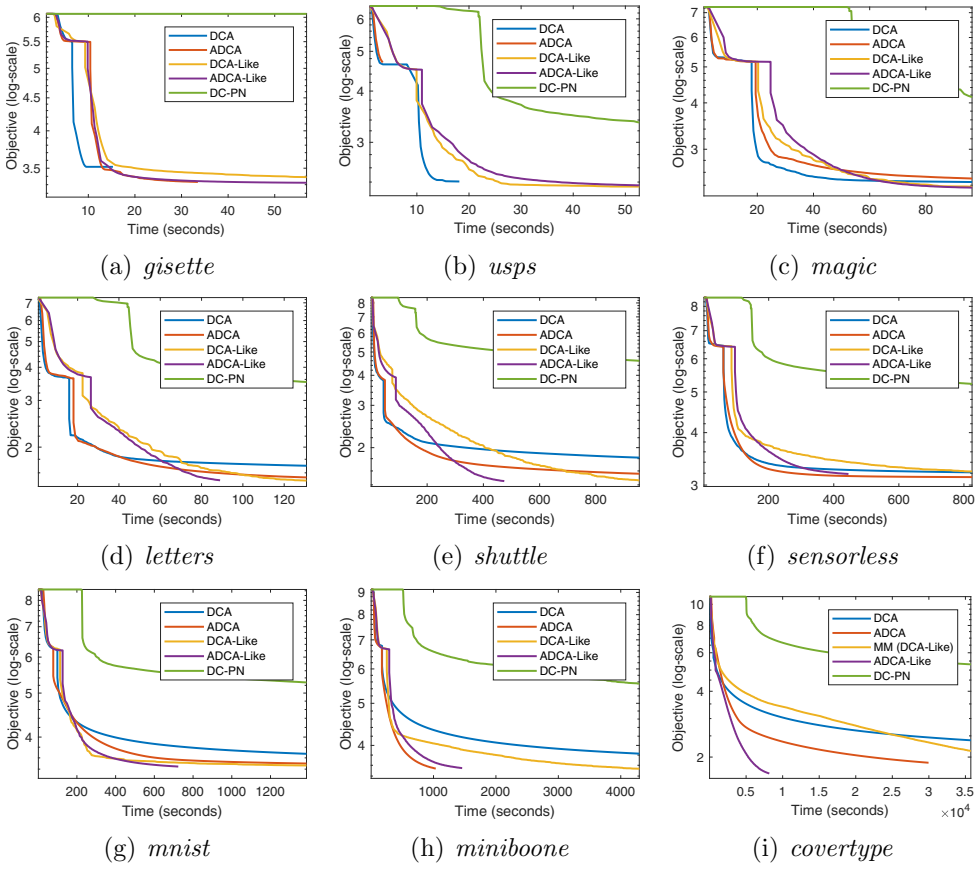
The performance is evaluated on three criteria: the objective value $F(\mathbf{x})$, the number of iterations and the running time (measured in seconds). We run each algorithm 10 times and then report the mean and standard deviation for each criterion in Table 1.

Table 1: Comparative results on t-SNE problem. Bold values indicate best results.

Dataset	Algorithm	Objective		Iteration		Time (sec.)	
		Mean	STD	Mean	STD	Mean	STD
gisette	DCA	3.52	0.04	**27**	0	**15.3**	0.2
$N = 7000$	ADCA	3.33	0.00	118	31	39.5	7.9
$d = 5000$	DCA-Like	3.34	0.01	283	81	209.0	57.9
	ADCA-Like	**3.32**	0.02	133	24	62.9	10.0
	DC-PN	3.53	0.02	2187	186	1218.3	194.3
usps	DCA	2.41	0.01	**29**	1	**23.1**	1.8
$N = 9298$	ADCA	3.92	1.36	70	107	27.9	39.0
$d = 256$	DCA-Like	**2.34**	0.00	106	18	62.5	11.0

	ADCA-Like	**2.34**	0.01	107	15	64.9	6.0
	DC-PN	2.57	0.07	2083	538	1315.2	171.7
magic	DCA	2.40	0.00	850	373	413.4	170.4
$N = 19020$	ADCA	2.46	0.01	215	9	150.5	9.8
$d = 10$	DCA-Like	**2.36**	0.00	168	39	135.2	31.1
	ADCA-Like	**2.36**	0.00	**106**	7	**101.7**	4.8
	DC-PN	2.80	0.07	3498	237	4321.0	557.8 s
letters	DCA	1.58	0.02	1186	159	652.8	98.2
$N = 20000$	ADCA	1.49	0.02	369	104	268.4	86.8
$d = 16$	DCA-Like	**1.48**	0.01	164	24	149.0	19.3
	ADCA-Like	**1.48**	0.01	**90**	7	**96.8**	7.4
	DC-PN	2.14	0.07	1326	253	1997.1	539.0
shuttle	DCA	1.60	0.02	3520	459	6056.5	782.9
$N = 58000$	ADCA	1.48	0.04	760	96	1781.7	215.1
$d = 9$	DCA-Like	1.45	0.02	333	9	981.7	33.1
	ADCA-Like	**1.42**	0.00	**152**	23	**553.7**	123.6
	DC-PN	2.54	0.03	4693	68	26784.4	3633.2
sensorless	DCA	3.18	0.02	1788	454	3232.1	837.3
$N = 58509$	ADCA	**3.13**	0.01	418	24	912.8	50.0
$d = 48$	DCA-Like	3.21	0.02	313	41	953.7	114.5
	ADCA-Like	3.18	0.02	**153**	28	**499.3**	91.2
	DC-PN	3.81	0.04	4255	144	20586.9	4488.6
mnist	DCA	3.44	0.00	3894	111	13752.7	644.9
$N = 70000$	ADCA	3.45	0.01	960	60	2659.1	63.3
$d = 784$	DCA-Like	3.46	0.01	371	67	1576.8	259.7
	ADCA-Like	**3.43**	0.01	**196**	27	**834.1**	127.5
	DC-PN	4.12	0.01	3703	308	21486.7	3643.1
miniboone	DCA	3.55	0.05	3401	1151	20473.3	6471.2
$N = 130064$	ADCA	**3.47**	0.06	454	338	2917.1	2050.2
$d = 50$	DCA-Like	3.53	0.05	469	61	4841.7	691.5
	ADCA-Like	3.53	0.02	**170**	12	**1560.6**	209.1
	DC-PN	5.36	0.15	471	29	7071.7	844.0
covertype	DCA	2.14	0.00	3998	35	71591.8	1407.2

$N = 581012$	ADCA	1.88	0.00	1670	0	29954.8	0.0
$d = 54$	DCA-Like	2.10	0.04	1217	66	37123.1	3546.5
	ADCA-Like	**1.68**	0.01	**330**	10	**8338.4**	191.8
	DC-PN	4.24	0.00	5741	45	140122.8	5126.4



(a) *gisette* (b) *usps* (c) *magic*

(d) *letters* (e) *shuttle* (f) *sensorless*

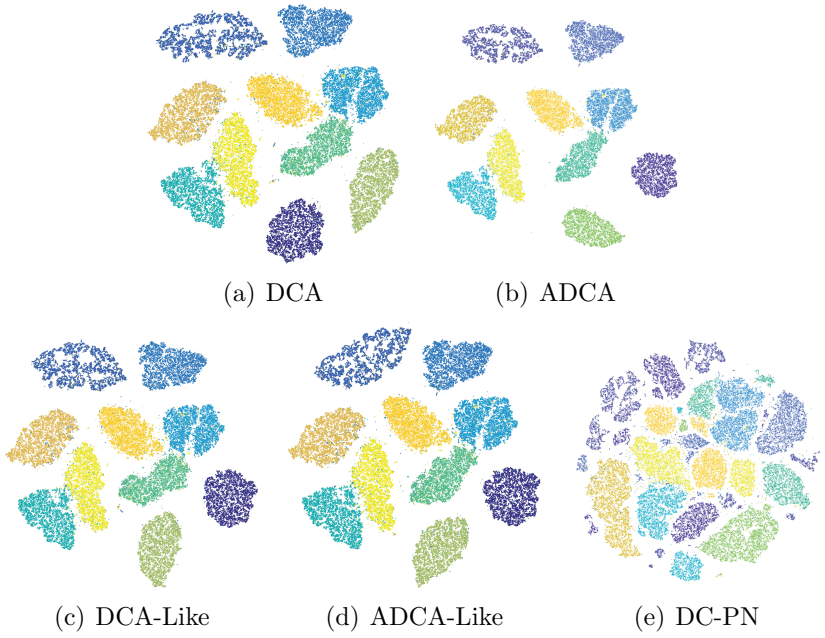
(g) *mnist* (h) *miniboone* (i) *covertype*

Figure 1: Objective value versus running time (average of ten runs)

Comments on numerical results

• First, we are interested in the performance gain of DCA-Like versus DCA.

For two smallest datasets (*gisette* and *ups*), DCA is faster than DCA-Like while the later is better than the former in terms of objective value. One possible reason is that the value of μ is quite large in DCA and then it converges quickly to a "bad" critical point.



(a) DCA (b) ADCA

(c) DCA-Like (d) ADCA-Like (e) DC-PN

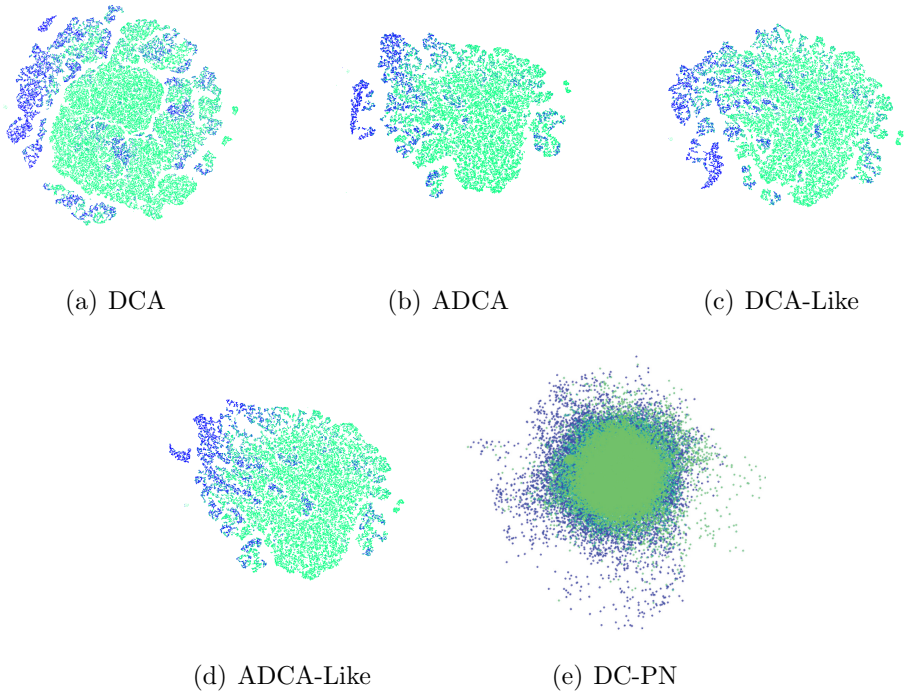
Figure 2: Visualization of embedding space on *mnist* dataset. Colors represent classes of data (0-9).

In datasets of over $10,000$ samples, DCA-Like is superior to DCA in all three criteria. In terms of running time, the number of iterations of DCA-Like is from 3.2 to 10 times less than DCA. Consequently, the computing time of DCA-Like is improved from 1.9 to 8.7 times comparing to DCA. Furthermore, DCA-Like gives lower objective value than DCA in six out of nine datasets (*gisette*, *usps*, *magic*, *letters*, *shuttle*, *miniboone*, and *covertype*), whereas the difference of two algorithms is neglectable for other three datasets.

In summary, DCA-Like improves the standard DCA on both quality of solution and rapidity.

• We study now the benefit of acceleration technique via the comparison between two pairs: ADCA versus DCA and ADCA-Like versus DCA-Like.

Not surprisingly, the number of iterations of ADCA-Like is always smaller (up to 3.6 times smaller) than that of DCA-Like. The gains in running time are also considerable, ADCA-Like is faster than DCA-Like from 1.5 to 5 times. On another hand, the two algorithms furnish the same objective value on three datasets (*ups*, *magic* and *letters*) while ADCA-Like gives better results



(a) DCA (b) ADCA (c) DCA-Like

(d) ADCA-Like (e) DC-PN

Figure 3: Visualization of embedding space on *miniboone* dataset (2 classes). Colors represent classes of data.

than DCA-Like in the other six datasets. Thus, we can say that ADCA-Like further improves the performance of DCA-Like.

As for the comparison of DCA and ADCA, in most of the cases, ADCA is better than DCA. In five out of nine datasets (*letters*, *shuttle*, *sensorless*, *miniboone* and *covertype*), ADCA is superior to DCA in all three aspects. The ratio of gains in terms of running time is from 2.4 to 7.0 times. In the two datasets (*magic* and *mnist*), ADCA is faster, but is slightly worse in terms of objective value.

• Among the five comparative algorithms, ADCA-Like gives the best results. In terms of running time, ADCA-Like, followed by DCA-Like, is the fastest in seven over nine datasets. As for the objective value, ADCA-Like gives the best result in seven datasets, followed by DCA-Like in three datasets. Note that, the gains of ADCA-Like versus the second-best algorithm, in each comparative criterion, are noticeable. This illustrates the superior of

ADCA-Like versus the other algorithms.

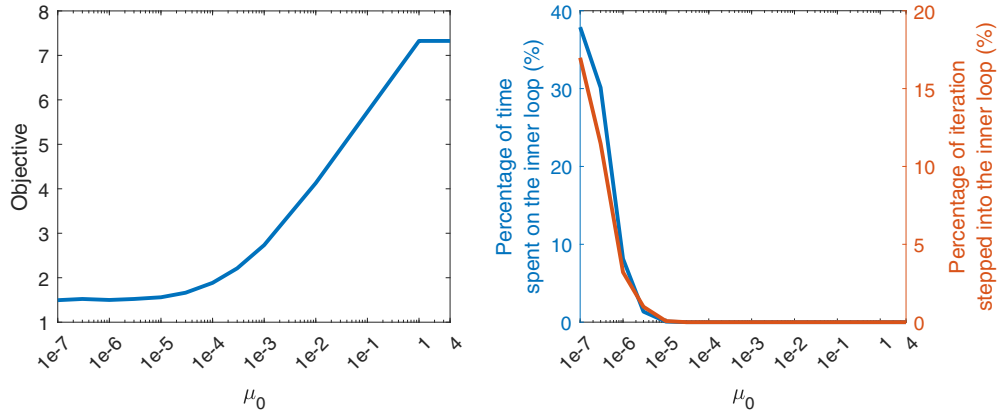
In Figure 1, we plot the objective value against the running time.

We observe that DCA performs thoroughly at the beginning but then it is left behind, while DCA-Like, ADCA and ADCA-Like improve swiftly over time. It is noticeable that, in the plot of *magic*, *letters*, *shuttle*, although ADCA-Like struggle at first iterations, it soon catches up and surpasses others algorithms to produce the best result while being the fastest. Also, for the biggest dataset (*covertype*), the effectiveness of accelerated algorithms is clearly demonstrated: ADCA-Like (reps. ADCA) surpass DCA-Like (reps. DCA) after a short amount of time.

Figure 2 visualizes the results on the well-known dataset *mnist*. This dataset consists of 70000 gray-scale 28×28 images over 10 classes of handwritten digits. *mnist* can be considered as the benchmark dataset for SNE-based algorithms, since they are able to capture both local and global structure of this dataset, especially in 2D embedding space. As we can see, in the embedding space, all four DCA-based algorithms managed to keep the structure of dataset of the original space, whereas DC-PN failed to maintain the structure. Four images of DCA-based algorithms in Figure 2 are quite similar since their objective values are fairly similar.

We now study the effect of μ_0 on the performance of DCA-Like. We run DCA-Like with different values of μ_0 taken from the interval $[10^{-7}, \ldots, 1, 4]$. This experiment is performed on *letters* - an arbitrary chosen dataset. In Figure 4 (a) we report the objective value given by DCA-Like as μ_0 varies. We observe that the best values of objective function are achieved when μ_0 is in the vicinity of 10^{-7} , and the objective value increases quickly as $\mu_0 \geq 10^{-6}$. Figure 4 (b) shows the percentage of number of iterations in inner while loop in DCA-Like over the total number of iterations (blue curve) and the percentage of time spent on the inner while loop over the total running time (red curve). We observe that with the smallest value $\mu_0 = 10^{-7}$, DCA-Like needs a large number of iterations in the inner loop for finding a suitable value of μ_k. In contrast, the number of inner iterations decreases drastically as μ_0 increases and for $\mu_0 \geq 10^{-5}$ it is equal to 0. It means that for a too large value of μ_0, the condition (9) is always verified and DCA-Like does not enjoy its advantages. It seems that $\mu_0 = 10^{-6}$ could be a good initial value of μ_0 to realize the trade-off between the objective value and the running time.

Overall, ADCA-Like gives the best results in all three comparison criteria while the results furnished by DC-PN are worst off. We can safely conclude that DCA-Like considerably improves DCA and ADCA-Like further enhances



(a) Objective value given by DCA-Like as μ_0 varies

(b) Percentage of number of iterations in DCA-Like where the inner while loop is activated and the time spent on inner while loop. Values of both lines are 0 for $\mu_0 \geq 10^{-5}$.

Figure 4: Result of DCA-Like with different value of μ_0 on dataset letters.

DCA-Like.

6. Conclusion

We have proposed the four new algorithms in DC programming and DCA framework for solving a special class of nonconvex programs which have many applications in various areas, in particular in machine learning. Due to the presence of the composite function, it is not easy to highlight a DC decomposition of the considered problem. Hence, we equivalently reformulate the original problem as a DC program without composite terms and then design a standard DCA for the resulting problem. Furthermore, to get better convex approximations of the objective functions, DCA-Like, a novel extension of DCA was developed. The idea of DCA-Like is original and its advantage is double. Firstly, while standard DCA works with a convex majorization of the objective function on the whole space via an available DC decomposition, DCA-Like seeks a better convex approximation at the current solution through a decomposition which is not necessarily DC. Secondly, and more importantly in several practical problems, DCA-Like works even when one can not highlight a DC decomposition. It turns out that, fortunately, despite these modifications we can prove that DCA-Like

still enjoys the convergence properties of DCA. Considering the Kurdyka-Łojasiewics assumption, we proved that each bounded sequence generated by DCA-Like globally converges to a critical point. Furthermore, under Kurdyka-Łojasiewicz assumption, the convergence rate of DCA-Like is proved to be at least sublinear $\mathcal{O}(1/k^{\alpha})$ with $\alpha > 1$. In addition, we propose accelerated versions of DCA and ADCA-Like by incorporating the Nesterov's acceleration technique into them. We showed that the convergence properties of these algorithms are still valid to their accelerated version.

The proposed algorithms successfully solved the t-distributed Stochastic Embedding (t-SNE). Our algorithms are in explicit form, say, the solution of convex subproblems are explicitly determined and therefore no convex optimization solver is needed. Numerical experiments carefully conducted on several benchmark datasets show that DCA-Like outperforms standard DCA and the corresponding accelerated version enhances further the rapidity as well as the quality of DCA and DCA-Like.

We are convinced that DCA-Like and the accelerated versions can be efficiently exploited for solving wider classes of nonconvex problems, and it is the purpose of our future works.

References

[1] Attouch, H., Bolte, J., 2009. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Math. Program. Ser. B 116, 5–16.

[2] Attouch, H., Bolte, J., Redont, P., Soubeyran, A., Apr. 2010. Proximal Alternating Minimization and Projection Methods for Nonconvex Problems: An Approach Based on the Kurdyka-Łojasiewicz Inequality. Mathematics of Operations Research 35 (2), 438–457.

[3] Bacak, M., Borwein, J. M., 2011. On difference convexity of locally lipschitz functions. Optimization 60 (8-9), 961–978.

[4] Beck, A., Teboulle, M., 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imag. Sci. 2, 183–202.

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

[6] Hartman, P., 1959. On functions representable as a difference of convex functions. Pacific J. Math 9, 707713.

[7] Hinton, G. E., Roweis, S. T., 2003. Stochastic neighbor embedding. In: Advances in Neural Information Processing Systems. pp. 857–864.

[8] Le Thi, H. A., Nguyen, B. T., Le, H. M., 2013. Sparse signal recovery by difference of convex functions algorithms. In: Intelligent Information and Database Systems, Lecture Notes in Computer Science. Vol. 7803. pp. 387–397.

[9] Le Thi, H. A., Pham Dinh, T., Jan. 2005. The DC (Difference of Convex Functions) Programming and DCA Revisited with DC Models of Real World Nonconvex Optimization Problems. Annals of Operations Research 133 (1), 23–46.

[10] Le Thi, H. A., Pham Dinh, T., May 2018. DC programming and DCA: Thirty years of developments. Mathematical Programming 169 (1), 5–68.

[11] Le Thi, H. A., Pham Dinh, T., Le, H. M., Vo, X. T., Jul. 2015. DC approximation approaches for sparse optimization. European Journal of Operational Research 244 (1), 26–46.

[12] Maaten, L. v. d., 2014. Accelerating t-sne using tree-based algorithms. Journal of machine learning research 15 (1), 3221–3245.

[13] Maaten, L. v. d., Hinton, G., 2008. Visualizing data using t-sne. Journal of Machine Learning Research 9 (Nov), 2579–2605.

[14] Mordukhovich, B. S., 2006. Variational Analysis and Generalized Differentiation I. Grundlehren Der Mathematischen Wissenschaften, Variational Analysis and Generalized Differentiation. Springer-Verlag, Berlin Heidelberg.

[15] Nocedal, J., 1980. Updating Quasi-Newton Matrices with Limited Storage. Mathematics of Computation 35 (151), 773–782.

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

[17] Pham Dinh, T., Le Thi, H. A., Feb. 1998. A D. C. Optimization Algorithm for Solving the Trust-Region Subproblem. SIAM Journal of Optimization 8 (2), 476–505.

[18] Pham Dinh, T., Le Thi, H. A., 2014. Recent advances in dc programming and dca. In: Transactions on Computational Intelligence XIII. pp. 1–37.

[19] Phan, D. N., Le, H. M., Le Thi, H. A., 2018. 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. pp. 1369–1375.

[20] Phan, D. N., Le Thi, H. A., Pham Dinh, T., 2017. Sparse covariance matrix estimation by DCA-based algorithms. Neural Computation 29 (11), 3040–3077.

[21] Rakotomamonjy, A., Flamary, R., Gasso, G., Mar. 2016. DC Proximal Newton for Nonconvex Optimization Problems. IEEE Transactions on Neural Networks and Learning Systems 27 (3), 636–647.

[22] Vladymyrov, M., Carreira-Perpinan, M., 2012. Partial-Hessian strategies for fast learning of nonlinear embeddings. arXiv preprint arXiv:1206.4646.

[23] Yang, Z., King, I., Xu, Z., Oja, E., 2009. Heavy-tailed symmetric stochastic neighbor embedding. In: Advances in Neural Information Processing Systems. pp. 2169–2177.

[24] Yang, Z., Peltonen, J., Kaski, S., 2015. Majorization-minimization for manifold embedding. In: Artificial Intelligence and Statistics. pp. 1088–1097.

Appendix A. Proof of Proposition 1

Let $\mathbf{x}^*$ be a local solution of the problem (1). Hence there exists $\epsilon > 0$ such that $F(\mathbf{x}) \geq F(\mathbf{x}^*)$, for all $\mathbf{x} \in \mathcal{B}(\mathbf{x}^*, \epsilon)$ where $\mathcal{B}(\mathbf{x}^*, \epsilon)$ is the ball of center $\mathbf{x}^*$ and radius ϵ in $\mathbb{R}^n$. We have $\|\mathbf{x} - \mathbf{x}^*\| \leq \|(\mathbf{x}, g(\mathbf{x})) - (\mathbf{x}^*, g(\mathbf{x}^*))\|$. This implies that if $(\mathbf{x}, g(\mathbf{x})) \in \mathcal{B}((\mathbf{x}^*, g(\mathbf{x}^*)), \epsilon)$ then $\mathbf{x} \in \mathcal{B}(\mathbf{x}^*, \epsilon)$. Hence, for all $(\mathbf{x}, \mathbf{z}) \in \mathcal{B}((\mathbf{x}^*, g(\mathbf{x}^*)), \epsilon)$, we have $\varphi(\mathbf{x}, \mathbf{z}) \geq F(\mathbf{x}) \geq F(\mathbf{x}^*) = \varphi(\mathbf{x}^*, g(\mathbf{x}^*))$. It follows that $(\mathbf{x}^*, g(\mathbf{x}^*))$ is a local solution to (4).

Inversely, let $(\mathbf{x}^*, g(\mathbf{x}^*))$ be a local solution to (4). There exists $\epsilon > 0$ such that $\varphi(\mathbf{x}, \mathbf{z}) \geq \varphi(\mathbf{x}^*, g(\mathbf{x}^*))$, for all $(\mathbf{x}, \mathbf{z}) \in \mathcal{B}((\mathbf{x}^*, g(\mathbf{x}^*)), \epsilon)$. Since g is convex, it is locally Lipschitz continuous around $\mathbf{x}^*$, and by shrinking ϵ if necessary, we can find $L_* > 0$ such that $\|g(\mathbf{x}) - g(\mathbf{x}^*)\| \leq L_* \|\mathbf{x} - \mathbf{x}^*\|$ for all $\mathbf{x} \in \mathcal{B}(\mathbf{x}^*, \epsilon)$. Therefore $\|(\mathbf{x}, g(\mathbf{x})) - (\mathbf{x}^*, g(\mathbf{x}^*))\| \leq \sqrt{L_*^2 + 1}\|\mathbf{x} - \mathbf{x}^*\|$. It follows that, for all $\mathbf{x} \in \mathcal{B}(\mathbf{x}^*, \epsilon/\sqrt{L_*^2 + 1})$, $F(\mathbf{x}) = \varphi(\mathbf{x}, g(\mathbf{x})) \geq \varphi(\mathbf{x}^*, g(\mathbf{x}^*)) = F(\mathbf{x}^*)$. This implies that $\mathbf{x}^*$ is a local solution of (1).

Suppose that $\mathbf{x}^* \in X$ is a global solution of (1). For any $(\mathbf{x}', \mathbf{z}')$, let $\epsilon > \|(\mathbf{x}', \mathbf{z}') - (\mathbf{x}^*, g(\mathbf{x}^*))\|$. Clearly, $F(\mathbf{x}) \geq F(\mathbf{x}^*)$, for all $\mathbf{x} \in \mathcal{B}(\mathbf{x}^*, \epsilon)$. As above, we have $\varphi(\mathbf{x}, \mathbf{z}) \geq \varphi(\mathbf{x}^*, g(\mathbf{x}^*))$ for all $(\mathbf{x}, \mathbf{z}) \in \mathcal{B}((\mathbf{x}^*, g(\mathbf{x}^*)), \epsilon)$. Thus, $\varphi(\mathbf{x}', \mathbf{z}') \geq \varphi(\mathbf{x}^*, g(\mathbf{x}^*))$. This is true for any $(\mathbf{x}', \mathbf{z}')$, hence $(\mathbf{x}^*, g(\mathbf{x}^*))$ is a global solution to (4).

Inversely, suppose that $(\mathbf{x}^*, g(\mathbf{x}^*))$ is a global solution to (4). For any $\mathbf{x}'$, let $\epsilon > \sqrt{L_*^2 + 1}\,\|\mathbf{x}' - \mathbf{x}^*\|$. Obviously, $\varphi(\mathbf{x}, \mathbf{z}) \geq \varphi(\mathbf{x}^*, g(\mathbf{x}^*))$ for all $(\mathbf{x}, \mathbf{z}) \in \mathcal{B}((\mathbf{x}^*, g(\mathbf{x}^*)), \epsilon)$. As above, for all $\mathbf{x} \in \mathcal{B}(\mathbf{x}^*, \epsilon/\sqrt{L_*^2 + 1})$, $F(\mathbf{x}) \geq F(\mathbf{x}^*)$. Thus, $F(\mathbf{x}') \geq F(\mathbf{x}^*)$. This is true for any $\mathbf{x}'$, so $\mathbf{x}^*$ is a global solution to (1). ■

Appendix B. Proof of Proposition 2

Assume that $\mathbf{x}^{k+1}$ is the optimal solution to (8). We have

$$\begin{aligned}
&\tfrac{\mu}{2}\|\mathbf{x}\|^2 - \langle \mathbf{y}^k, \mathbf{x}\rangle + \sum_{i=1}^{m}(-\xi_i^k) z_i \geq \tfrac{\mu}{2}\|\mathbf{x}\|^2 - \langle \mathbf{y}^k, \mathbf{x}\rangle + \sum_{i=1}^{m}(-\xi_i^k) g_i(\mathbf{x}_i) \\
&\geq \tfrac{\mu}{2}\|\mathbf{x}^{k+1}\|^2 - \langle \mathbf{y}^k, \mathbf{x}^{k+1}\rangle + \sum_{i=1}^{m}(-\xi_i^k) g_i(\mathbf{x}_i^{k+1}),\ \forall (\mathbf{x}, \mathbf{z}) \in \Omega,
\end{aligned}$$

where the first inequality follows from $-\xi_i^k \geq 0$ and $z_i \geq g(\mathbf{x}_i)$, and the second inequality comes from the fact that $\mathbf{x}^{k+1}$ is the optimal solution to (8). This implies that $(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1}))$ is an optimal solution to (7). ■

Appendix C. Proof of Theorem 3

To prove Theorem 3, we use the following Lemma 2 and Lemma 3. We also need the Lemma below given in [5].

Lemma 1 ([5]). *Let Λ be a compact set and let $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function. Assume that σ satisfies the KL property at each point of Λ on which σ is constant. Then, there exist*

$\eta > 0, \epsilon > 0$ *and* $\psi \in \mathcal{M}_\eta$ *such that for all* $\mathbf{u} \in \{\mathbf{u} : dist(\mathbf{u}, \Lambda) < \epsilon\} \cap \{\mathbf{u} : \sigma(\mathbf{u}^*) < \sigma(\mathbf{u}) < \sigma(\mathbf{u}^*) + \eta\}$, *one has*

$$\psi'(\sigma(\mathbf{u}) - \sigma(\mathbf{u}^*)) dist(0, \partial^L \sigma(\mathbf{u})) \geq 1.$$

Lemma 2. *Let* $\{\mathbf{x}^k\}$ *be the sequence generated by DCA-Like (Algorithm 2). If* $\inf \varphi(\mathbf{x}, \mathbf{z}) > -\infty$, *and the set of limit points* $\mathcal{C}$ *of* $\{\mathbf{x}^k\}$ *is not empty, then* $\lim_{k\to+\infty} \varphi(\mathbf{x}^k, g(\mathbf{x}^k)) = \varphi(\mathbf{x}^*, g(\mathbf{x}^*))$ *for some* $\mathbf{x}^* \in \mathcal{C}$. *Thus,* φ *has the same value on* $\Lambda = \{(\mathbf{x}^*, g(\mathbf{x}^*)) : \mathbf{x}^* \in \mathcal{C}\}$.

Proof of Lemma 2. Since $\inf \varphi(\mathbf{x}, \mathbf{z}) > -\infty$, it follows from (i) of Theorem 2 that $\{\varphi(\mathbf{x}^k, g(\mathbf{x}^k))\}$ is non-increasing and bounded below. Thus, there exists $\varphi^* = \lim_{k\to+\infty} \varphi(\mathbf{x}^k, g(\mathbf{x}^k))$. Let $\mathbf{x}^*$ be a limit point of $\{\mathbf{x}^k\}$. We therefore can find a subsequence $\{\mathbf{x}^{k_j}\}$ that converges to $\mathbf{x}^*$. We have

$$\begin{aligned}
\limsup_{j\to+\infty} \varphi(\mathbf{x}^{k_j}, g(\mathbf{x}^{k_j})) &= \limsup_{j\to+\infty} [\chi_\Omega(\mathbf{x}^{k_j}, g(\mathbf{x}^{k_j})) + f(\mathbf{x}^{k_j}) - \sum_{i=1}^m (-h_i)(g_i(\mathbf{x}_i^{k_j}))] \\
&\leq \limsup_{j\to+\infty} \chi_\Omega(\mathbf{x}^{k_j}, g(\mathbf{x}^{k_j})) + \limsup_{j\to+\infty} f(\mathbf{x}^{k_j}) - \liminf_{j\to+\infty} \sum_{i=1}^m (-h_i)(g_i(\mathbf{x}_i^{k_j})) \\
&\leq \chi_\Omega(\mathbf{x}^*, g(\mathbf{x}^*)) + f(\mathbf{x}^*) - \sum_{i=1}^m (-h_i)(g_i(\mathbf{x}_i^*)) = \varphi(\mathbf{x}^*, g(\mathbf{x}^*)),
\end{aligned}$$

where the last inequality holds by (16) and the lower semicontinuity of $-h_i$. From the above inequality and the lower semicontinuity of φ, we obtain $\lim_{j\to+\infty} \varphi(\mathbf{x}^{k_j}, g(\mathbf{x}^{k_j})) = \varphi(\mathbf{x}^*, g(\mathbf{x}^*))$. Hence, by the uniqueness of the limit, we have $\varphi^* = \varphi(\mathbf{x}^*, g(\mathbf{x}^*))$. ■

Lemma 3. *Let* $\{s_k\}$ *be a sequence in* $\mathbb{R}_+$ *and let* α, γ *be some non-negative constants. Assume that* $s_k \to 0$ *as* $k \to +\infty$ *and there exists* $N > 0$ *such that*

$$s_{k+1}^\alpha \leq \gamma(s_k - s_{k+1}), \forall k \geq N. \tag{C.1}$$

Then, the following statements hold.

i) If $\alpha = 0$, *then* $\{s_k\}$ *converges in a finite number of steps to* 0.

ii) If $\alpha \in (0, 1]$, *then* $\{s_k\}$ *converges linearly to* 0 *with rate* $\frac{\gamma}{1+\gamma}$.

iii) If $\alpha > 1$, *then there exist positive constants* δ *and* N_0 *such that*

$$s_k \leq \delta k^{-\frac{1}{\alpha-1}}, \forall k \geq N_0.$$

Proof of Lemma 3. The proof of Lemma 3 is similar to the one of Theorem 2 in [1].

(i) If $\alpha = 0$, then we have $s_k - s_{k+1} \geq \frac{1}{\gamma}$. This implies (i).

(ii) Assume that $\alpha \in (0, 1]$. Since $\{s_k\}$ converges to 0, there exists N_1 such that $s_k < 1$ whenever $k > N_1$. Hence, we have $s_{k+1} \leq s_{k+1}^{\alpha} \leq \gamma(s_k - s_{k+1})$. Therefore $s_{k+1} \leq \frac{\gamma}{1+\gamma} s_k, \forall k > N_1$, which implies (ii).

(iii) The condition (C.1) implies that the sequence $\{s_k\}_{k \geq N}$ is decreasing and non-negative. Thus, if there is $k_0 \geq N$ such that $s_{k_0} = 0$ then $s_k = 0$ for any $k \geq k_0$ and iii) trivially holds. Now we assume that $s_k > 0$ for all $k \geq N$. Due to convexity of the function $s \in (0, \infty) \mapsto s^{1-\alpha}$ and the condition (C.1), we have

$$s_{k+1}^{1-\alpha} - s_k^{1-\alpha} \geq (1-\alpha) s_k^{-\alpha}(s_{k+1} - s_k) \geq \frac{s_{k+1}^{\alpha}}{s_k^{\alpha}} \frac{\alpha - 1}{\gamma}, \quad \forall k \geq N.$$

Therefore, if $s_k^{\alpha} \leq 2 s_{k+1}^{\alpha}$, we have

$$s_{k+1}^{1-\alpha} - s_k^{1-\alpha} \geq \frac{\alpha - 1}{2\gamma}, \quad \forall k \geq N.$$

Otherwise, if $s_k^{\alpha} > 2 s_{k+1}^{\alpha}$, by taking both sides to the power of $\frac{1-\alpha}{\alpha} < 0$ we obtain

$$s_k^{1-\alpha} < 2^{\frac{1-\alpha}{\alpha}} s_{k+1}^{1-\alpha} \quad \Longrightarrow \quad s_{k+1}^{1-\alpha} - s_k^{1-\alpha} > (1 - 2^{\frac{1-\alpha}{\alpha}}) s_{k+1}^{1-\alpha}.$$

Since $1 - 2^{\frac{1-\alpha}{\alpha}} > 0$ and $0 < s_k \to 0$, there exists $\bar{\mu} > 0$ such that

$$(1 - 2^{\frac{1-\alpha}{\alpha}}) s_{k+1}^{1-\alpha} \geq \bar{\mu}, \quad \forall k \geq N.$$

Letting $\mu = \min\{\frac{\alpha - 1}{2\gamma}, \bar{\mu}\} > 0$, we have

$$s_{k+1}^{1-\alpha} - s_k^{1-\alpha} \geq \mu, \quad \forall k \geq N.$$

By summing these inequalities, we obtain that

$$s_k^{1-\alpha} = \sum_{i=N}^{k-1} (s_{i+1}^{1-\alpha} - s_i^{1-\alpha}) + s_N^{1-\alpha} \geq (k - N)\mu, \quad \forall k \geq N.$$

Thus, with $N_0 = 2N$ and $\delta = (\frac{\mu}{2})^{\frac{1}{1-\alpha}} > 0$, we have

$$s_k^{1-\alpha} \geq \frac{\mu}{2} k \quad \text{or} \quad s_k \leq \delta k^{\frac{1}{1-\alpha}}, \quad \forall k \geq N_0.$$

Proof of Theorem 3. Denote by $\mathcal{C}$ the set of limit points of $\{\mathbf{x}^k\}$. It follows from the continuity of g that the set of limit points Λ of $\{(\mathbf{x}^k, g(\mathbf{x}^k))\}$ takes the form $\Lambda = \{(\mathbf{x}, g(\mathbf{x})) : \mathbf{x} \in \mathcal{C}\}$. By the Lemma 2, φ has the same value on Λ, which is denoted by φ^*. Thus, $\lim_{k\to+\infty} \varphi(\mathbf{x}^k, g(\mathbf{x}^k)) = \varphi^*$. If there exists $k \geq 1$ such that $\varphi(\mathbf{x}^k, g(\mathbf{x}^k)) = \varphi^*$, then $\varphi(\mathbf{x}^k, g(\mathbf{x}^k)) = \varphi(\mathbf{x}^{k+p}, g(\mathbf{x}^{k+p}))$ for any $p \geq 0$ due to the decreasing property of $\{\varphi(\mathbf{x}^k, g(\mathbf{x}^k))\}$. Hence, $\mathbf{x}^k = \mathbf{x}^{k+p}$ for all $p \geq 0$ and DCA-Like terminates after a finite number of steps. Hence, without loss of generality, we can assume that $\varphi(\mathbf{x}^k, g(\mathbf{x}^k)) > \varphi^*$ for all k.

Since $\{\mathbf{x}^k\}$ is bounded, $\mathcal{C}$ is a compact set. Define $H(\mathbf{z}) = \sum_{i=1}^m (-h_i)(z_i)$. By the locally Lipschitz property of ∇H and g, for each $\mathbf{x} \in \mathcal{C}$, there exist L_x and ϵ_x such that $\|\nabla H(g(\mathbf{u})) - \nabla H(g(\mathbf{v}))\| \leq L_x\|\mathbf{u} - \mathbf{v}\|$ for all $\mathbf{u}, \mathbf{v} \in \mathcal{B}(\mathbf{x}, \epsilon_x)$. By the compactness of $\mathcal{C}$, there exist $\mathbf{w}_1, ..., \mathbf{w}_p \in \mathcal{C}$ such that $\mathcal{C} \subset \cup_{i=1}^p \mathcal{B}(\mathbf{w}_i, \epsilon_{w_i}/4)$. Set $\pi = \max\{L_{w_i} : i = 1, ..., p\}$ and $\epsilon = \min\{\epsilon_{w_i}/2 : i = 1, ..., p\}$. Since $\mathcal{C}$ is the set of limit points of $\{\mathbf{x}^k\}$, we have $\lim_{k\to+\infty} \text{dist}(\mathbf{x}^k, \mathcal{C}) = 0$. From this and (ii) of Theorem 2, there exists $N_1 > 0$ such that $\mathbf{x}^{k-1} \in \cup_{i=1}^p \mathcal{B}(\mathbf{w}_i, \epsilon_{w_i}/2)$ and $\|\mathbf{x}^k - \mathbf{x}^{k-1}\| < \epsilon$ whenever $k \geq N_1$. Hence, there are some $\mathbf{w}_i$ such that $\mathbf{x}^k, \mathbf{x}^{k-1} \in \mathcal{B}(\mathbf{w}_i, \epsilon_{w_i})$. Thus,

$$\|\nabla H(g(\mathbf{x}^k)) - \nabla H(g(\mathbf{x}^{k-1}))\| \leq L_{w_i}\|\mathbf{x}^k - \mathbf{x}^{k-1}\| \leq \pi\|\mathbf{x}^k - \mathbf{x}^{k-1}\|, \forall k \geq N_1. \tag{C.2}$$

Thank to the compactness of $\mathcal{C}$ and the continuity of g, Λ is also a compact set. Applying Lemma 1 to the function φ and by shrinking ϵ if necessary, we can find $\eta > 0$ and $\psi \in \mathcal{M}_\eta$ such that

$$\psi'(\varphi(\mathbf{x}, \mathbf{z}) - \varphi^*)\text{dist}(0, \partial^L\varphi(\mathbf{x}, \mathbf{z})) \geq 1 \tag{C.3}$$

$\forall (\mathbf{x}, \mathbf{z}) \in U := \{(\mathbf{x}, \mathbf{z}) : \text{dist}((\mathbf{x}, \mathbf{z}), \Lambda) < \epsilon\} \cap \{(\mathbf{x}, \mathbf{z}) : \varphi^* < \varphi(\mathbf{x}, \mathbf{z}) < \varphi^* + \eta\}$. From the definition of Λ: $\lim_{k\to+\infty} \text{dist}((\mathbf{x}^k, g(\mathbf{x}^k)), \Lambda) = 0$. Thus, there exists $N_2 > 0$ such that $\text{dist}((\mathbf{x}^k, g(\mathbf{x}^k)), \Lambda) < \epsilon$ whenever $k \geq N_2$. By (i) of Theorem 2 and Lemma 2, there exists $N_3 > 0$ such that $\varphi^* < \varphi(\mathbf{x}^k, g(\mathbf{x}^k)) < \varphi^* + \eta$ for all $k \geq N_3$. Taking $N = \max\{N_1, N_2, N_3\}$ we get $(\mathbf{x}^k, g(\mathbf{x}^k)) \in U$ for all $k \geq N$. Hence, it follows from (C.3) that for any $k \geq N$,

$$\psi'(\varphi(\mathbf{x}^k, g(\mathbf{x}^k)) - \varphi^*)\text{dist}(0, \partial^L\varphi(\mathbf{x}^k, g(\mathbf{x}^k))) \geq 1. \tag{C.4}$$

By the differentiability of h_i and the definition of $\{\mathbf{x}^k\}$, we have

$$(\mathbf{y}^{k-1}, \xi^{k-1}) \in \partial G_{\mu_k}(\mathbf{x}^k, g(\mathbf{x}^k)) = \mu_k\mathbf{x}^k + \partial\chi_\Omega(\mathbf{x}^k, g(\mathbf{x}^k)).$$

This implies that

$$(\mu_k(\mathbf{x}^{k-1} - \mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}) + \nabla f(\mathbf{x}^k), \xi^{k-1}) \in \partial\chi_\Omega(\mathbf{x}^k, g(\mathbf{x}^k)) + (\nabla f(\mathbf{x}^k), 0).$$

Therefore

$$\begin{aligned}&(\mu_k(\mathbf{x}^{k-1}-\mathbf{x}^k)-\nabla f(\mathbf{x}^{k-1})+\nabla f(\mathbf{x}^k),\xi^{k-1}-\xi^k)\\&\in\partial\chi_\Omega(\mathbf{x}^k,g(\mathbf{x}^k))+(\nabla f(\mathbf{x}^k),0)-(0,\nabla H(g(\mathbf{x}^k)))\partial^L\varphi(\mathbf{x}^k,g(\mathbf{x}^k)).\end{aligned}\tag{C.5}$$

By the Lipschitz continuity of ∇f and (C.2), for all $k\geq N$, we have

$$\begin{aligned}\|(\mu_k(\mathbf{x}^{k-1}-\mathbf{x}^k)-&\nabla f(\mathbf{x}^{k-1})+\nabla f(\mathbf{x}^k),\xi^{k-1}-\xi^k)\|\\&\leq(\mu_k+L+\pi)\|\mathbf{x}^{k-1}-\mathbf{x}^k\|\leq((\beta+1)L+\pi)\|\mathbf{x}^{k-1}-\mathbf{x}^k\|,\end{aligned}$$

where the last inequality follows from $\mu_k\leq\beta L$. Combining this with (C.5) we obtain, for any $k\geq N$,

$$\operatorname{dist}(0,\partial^L\varphi(\mathbf{x}^k,g(\mathbf{x}^k)))\leq M\|\mathbf{x}^{k-1}-\mathbf{x}^k\|,\ \text{where } M=(\beta+1)L+\pi.\tag{C.6}$$

It follows from this, (C.4), and the concavity of ψ that

$$\begin{aligned}\forall k\geq N:\ &M\|\mathbf{x}^{k-1}-\mathbf{x}^k\|[\psi(\varphi(\mathbf{x}^k,g(\mathbf{x}^k))-\varphi^*)-\psi(\varphi(\mathbf{x}^{k+1},g(\mathbf{x}^{k+1}))-\varphi^*)]\\\geq\operatorname{dist}(0,\partial^L\varphi(\mathbf{x}^k,g(\mathbf{x}^k)))&\psi'(\varphi(\mathbf{x}^k,g(\mathbf{x}^k))-\varphi^*)(\varphi(\mathbf{x}^k,g(\mathbf{x}^k))-\varphi(\mathbf{x}^{k+1},g(\mathbf{x}^{k+1})))\\&\geq\varphi(\mathbf{x}^k,g(\mathbf{x}^k))-\varphi(\mathbf{x}^{k+1},g(\mathbf{x}^{k+1}))\geq\frac{\mu_0}{2}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|^2,\end{aligned}$$

where the last inequality comes from (i) of Theorem 2. This implies that

$$\begin{aligned}\forall k\geq N:&\psi(\varphi(\mathbf{x}^k,g(\mathbf{x}^k))-\varphi^*)-\psi(\varphi(\mathbf{x}^{k+1},g(\mathbf{x}^{k+1}))-\varphi^*)\\&\geq\frac{\mu_0}{2M}\frac{\|\mathbf{x}^k-\mathbf{x}^{k+1}\|^2}{\|\mathbf{x}^{k-1}-\mathbf{x}^k\|}\geq\frac{\mu_0}{2M}[\|\mathbf{x}^k-\mathbf{x}^{k+1}\|-\frac{1}{4}\|\mathbf{x}^{k-1}-\mathbf{x}^k\|],\end{aligned}$$

where the last inequality follows from $a^2/b\geq a-b/4$ for all $a,b>0$. Thus,

$$\begin{aligned}\forall k\geq N:\ \|\mathbf{x}^k-\mathbf{x}^{k+1}\|\leq&\frac{1}{4}\|\mathbf{x}^{k-1}-\mathbf{x}^k\|+\frac{2M}{\mu_0}[\psi(\varphi(\mathbf{x}^k,g(\mathbf{x}^k))-\varphi^*)\\&-\psi(\varphi(\mathbf{x}^{k+1},g(\mathbf{x}^{k+1}))-\varphi^*)].\end{aligned}$$

Summing the above inequality from N to $N+k-1$ we get

$$\begin{aligned}\sum_{j=1}^{k}\|\mathbf{x}^{N+j}-\mathbf{x}^{N+j-1}\|\leq\ &\tfrac{8M}{3\mu_0}[\psi(\varphi(\mathbf{x}^N,g(\mathbf{x}^N))-\varphi^*)+\tfrac{1}{3}\|\mathbf{x}^{N-1}-\mathbf{x}^N\|\\&-\psi(\varphi(\mathbf{x}^{N+k},g(\mathbf{x}^{N+k}))-\varphi^*)].\end{aligned}$$

By the non-negativity of φ, we have

$$\sum_{j=1}^{k} \|\mathbf{x}^{N+j} - \mathbf{x}^{N+j-1}\| \leq \frac{1}{3}\|\mathbf{x}^{N-1} - \mathbf{x}^{N}\| + \frac{8M}{3\mu_0}\psi(\varphi(\mathbf{x}^N, g(\mathbf{x}^N)) - \varphi^*).$$

Taking the limit we obtain

$$\sum_{k=N}^{+\infty} \|\mathbf{x}^{k+1} - \mathbf{x}^{k}\| \leq \frac{1}{3}\|\mathbf{x}^{N-1} - \mathbf{x}^{N}\| + \frac{8M}{3\mu_0}\psi(\varphi(\mathbf{x}^N, g(\mathbf{x}^N)) - \varphi^*) < +\infty. \quad \text{(C.7)}$$

This implies that $\{\mathbf{x}^k\}$ is a Cauchy sequence, and hence it converges to a point $\mathbf{x}^*$. By Theorem 2, $(\mathbf{x}^*, g(\mathbf{x}^*))$ is a critical point of (4). From (C.4) and (C.6), for all $k \geq N$, we have

$$\psi'(\varphi(\mathbf{x}^k, g(\mathbf{x}^k)) - \varphi^*)M\|\mathbf{x}^{k-1} - \mathbf{x}^{k}\| \geq 1. \quad \text{(C.8)}$$

By the assumption $\psi(t) = ct^{1-\theta}$, (ii) of Theorem 2, and the above inequality, we obtain, for all $k \geq N$,

$$\begin{aligned} \|\varphi(\mathbf{x}^k, g(\mathbf{x}^k)) - \varphi^*\|^{2\theta} &\leq [c(1-\theta)M]^2\|\mathbf{x}^{k-1} - \mathbf{x}^{k}\|^2 \\ &\leq \frac{2[c(1-\theta)M]^2}{\mu_0}(\varphi(\mathbf{x}^{k-1}) - \varphi(\mathbf{x}^k)). \end{aligned} \quad \text{(C.9)}$$

Set $r_k = \varphi(\mathbf{x}^k, g(\mathbf{x}^k)) - \varphi^*$. It follows from (C.9) that

$$r_k^{2\theta} \leq \frac{2[c(1-\theta)M]^2}{\mu_0}(r_{k-1} - r_k), \forall k \geq N. \quad \text{(C.10)}$$

Therefore, taking $\alpha = 2\theta$, $\gamma = \frac{2[c(1-\theta)M]^2}{\mu_0}$ and applying Lemma 3 on the sequence $\{r_k\}$, we get the convergence rate of $\{\varphi(\mathbf{x}^k, g(\mathbf{x}^k))\} \to \varphi^*$ related to different values of θ as stated in this theorem.

Consider now the convergence rate of $\{\mathbf{x}^k\}$. By setting $s_i = \sum_{k=i}^{+\infty} \|\mathbf{x}^{k+1} - \mathbf{x}^{k}\|$, it follows from (C.7) that s_i is finite. Since $\|\mathbf{x}^i - \mathbf{x}^*\| \leq s_i$, the rate of convergence of $\{\mathbf{x}^i\}$ to $\mathbf{x}^*$ can be deduced from the rate of convergence of $\{s_i\}$ to 0. By (C.7) and $\psi(t) = ct^{1-\theta}$ we get

$$(\varphi(\mathbf{x}^i, g(\mathbf{x}^i)) - \varphi^*)^{1-\theta} \geq \frac{\mu_0}{8Mc}[3s_i - \|\mathbf{x}^{i-1} - \mathbf{x}^{i}\|], \ \forall i \geq N. \quad \text{(C.11)}$$

Combining $\psi(t) = ct^{1-\theta}$ and (C.8), we have

$$c(1-\theta)M\|\mathbf{x}^{i-1} - \mathbf{x}^i\| \geq (\varphi(\mathbf{x}^i, g(\mathbf{x}^i)) - \varphi^*)^\theta, \forall i \geq N. \tag{C.12}$$

This implies that if $\theta = 0$, then $\|\mathbf{x}^{i-1} - \mathbf{x}^i\| \geq \frac{1}{cM}$ for all $i \geq N$. Hence, by (i) of Theorem 2, we have

$$\varphi(\mathbf{x}^{i-1}, g(\mathbf{x}^{i-1})) - \varphi(\mathbf{x}^i, g(\mathbf{x}^i)) \geq \frac{\mu_k}{2}\|\mathbf{x}^{i-1} - \mathbf{x}^i\|^2 \geq \frac{\mu_0}{2[cM]^2}.$$

Hence, DCA-Like must terminate after a finite number of iterations. If $\theta \in (0, 1)$, according to (C.12) and (C.11) we have, for all $i \geq N$,

$$(c(1-\theta)M\|\mathbf{x}^{i-1} - \mathbf{x}^i\|)^{\frac{1-\theta}{\theta}} \geq \frac{\mu_0}{8Mc}[3s_i - \|\mathbf{x}^{i-1} - \mathbf{x}^i\|]. \tag{C.13}$$

Since $\|\mathbf{x}^{i-1} - \mathbf{x}^i\| \to 0$ as $i \to +\infty$, by increasing N if necessary, we have $\|\mathbf{x}^{i-1} - \mathbf{x}^i\| < 1$ for all $i \geq N$. Let $\phi(\theta) = \min(1, \frac{1-\theta}{\theta})$. Hence, it follows from (C.13) that

$$\begin{aligned} \frac{3\mu_0}{8Mc}s_i &\leq [(c(1-\theta)M)^{\frac{1-\theta}{\theta}} + \frac{\mu_0}{8Mc}]\|\mathbf{x}^{i-1} - \mathbf{x}^i\|^{\phi(\theta)} \\ &= [(c(1-\theta)M)^{\frac{1-\theta}{\theta}} + \frac{\mu_0}{8Mc}](s_{i-1} - s_i)^{\phi(\theta)}. \end{aligned} \tag{C.14}$$

This implies that $s_i^{\frac{1}{\phi(\theta)}} \leq (\frac{8Mc}{3\mu_0}(c(1-\theta)M)^{\frac{1-\theta}{\theta}} + \frac{1}{3})^{\frac{1}{\phi(\theta)}}(s_{i-1} - s_i), \forall i \geq N$. Taking $\alpha = 1/\phi(\theta)$, $\gamma = (\frac{8Mc}{3\mu_0}(c(1-\theta)M)^{\frac{1-\theta}{\theta}} + \frac{1}{3})^{1/\phi(\theta)}$ and applying (ii) and (iii) of Lemma 3 on the sequence $\{s_i\}$, we get the convergence rate of $\{s_i\} \to 0$, which is the convergence rate of $\{x^k\} \to x^*$, related to different values of θ as stated in (ii) and (iii) of this theorem. ■

Appendix D. Proof of Theorem 4

Similar to DCA-Like, the proof of convergence properties of ADCA-Like are mainly based on the key results analogue to (ii) and (iii) of Remark 3, with the use of the intermediate variable $\mathbf{v}^k$.

(i) By the definition of $\mathbf{x}^{k+1}$, we have $(\mathbf{y}^k, \xi^k) \in \partial G_{\mu_k}(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1}))$. Hence, it follows from the μ_k-convexity of G_{μ_k} on $\mathbf{x}$ that

$$\begin{aligned} G_{\mu_k}(\mathbf{v}^k, g(\mathbf{v}^k)) \geq\ & G_{\mu_k}(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1})) + \langle \mathbf{y}^k, \mathbf{v}^k - \mathbf{x}^{k+1}\rangle \\ & + \langle \xi^k, g(\mathbf{v}^k) - g(\mathbf{x}^{k+1})\rangle + \frac{\mu_k}{2}\|\mathbf{v}^k - \mathbf{x}^{k+1}\|^2. \end{aligned}$$

Combining the above inequality with $H_{\mu_k}(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1})) \geq H_{\mu_k}(\mathbf{v}^k, g(\mathbf{v}^k)) + \langle \mathbf{y}^k, \mathbf{x}^{k+1} - \mathbf{v}^k \rangle + \langle \xi^k, g(\mathbf{x}^{k+1}) - g(\mathbf{v}^k) \rangle$ we get

$$\varphi(\mathbf{v}^k, g(\mathbf{v}^k)) - \varphi(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1})) \geq \frac{\mu_k}{2} \|\mathbf{v}^k - \mathbf{x}^{k+1}\|^2.$$

From this and $\varphi(\mathbf{v}^k, g(\mathbf{v}^k)) \leq \varphi(\mathbf{x}^k, g(\mathbf{x}^k))$, we obtain

$$\varphi(\mathbf{x}^k, g(\mathbf{x}^k)) - \varphi(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1})) \geq \frac{\mu_k}{2} \|\mathbf{v}^k - \mathbf{x}^{k+1}\|^2. \tag{D.1}$$

(ii) From (D.1) and $\mu_k \geq \mu_0$, we have

$$\varphi(\mathbf{x}^k, g(\mathbf{x}^k)) - \varphi(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1})) \geq \frac{\mu_0}{2} \|\mathbf{v}^k - \mathbf{x}^{k+1}\|^2.$$

Summing the above inequality over $k = 0, ..., N$ we get

$$\varphi(\mathbf{x}^0, g(\mathbf{x}^0)) - \varphi(\mathbf{x}^{N+1}, g(\mathbf{x}^{N+1})) \geq \frac{\mu_0}{2} \sum_{k=0}^{N} \|\mathbf{v}^k - \mathbf{x}^{k+1}\|^2. \tag{D.2}$$

It follows from (D.2), $\mu_0 > 0$ and $\varphi(\mathbf{x}^{N+1}, g(\mathbf{x}^{N+1})) \geq \alpha$ that

$$\frac{2}{\mu_0}[\varphi(\mathbf{x}^0, g(\mathbf{x}^0)) - \alpha] \geq \sum_{k=0}^{N} \|\mathbf{v}^k - \mathbf{x}^{k+1}\|^2.$$

Passing to the limit over the sequence $\{N\}_{N \in \mathbb{N}}$, we obtain $\sum_{k=0}^{+\infty} \|\mathbf{v}^k - \mathbf{x}^{k+1}\|^2 < +\infty$, and therefore $\lim_{k \to +\infty} \|\mathbf{x}^{k+1} - \mathbf{v}^k\| = 0$.
(iii) Let $(\mathbf{x}^*, \mathbf{z}^*)$ be a limit point of the sequence $\{(\mathbf{x}^k, g(\mathbf{x}^k))\}$. By the continuity of g, we can take a subsequence $\{\mathbf{x}^{k_j+1}\}$ of $\{\mathbf{x}^k\}$ that converges to $\mathbf{x}^*$ and $\mathbf{z}^* = g(\mathbf{x}^*) = \lim_{j \to +\infty} g(\mathbf{x}^{k_j+1})$. It follows from (ii) that $\lim_{j \to +\infty} \mathbf{v}^{k_j} = \mathbf{x}^*$. In addition, without loss of generality, we can suppose that the subsequence $\{\xi^{k_j}\}$ converges to ξ^*. By the continuity of g_i and the property of the subdifferential mapping $\partial(-h)$, we have $\xi_i^* \in \partial(-h)(g_i(\mathbf{x}_i^*))$. We note that $(\mathbf{x}^{k_j+1}, g(\mathbf{x}^{k_j+1}))$ is a solution of the following convex problem

$$\min \left\{ G_{\mu_{k_j}}(\mathbf{x}, \mathbf{z}) - \langle \mathbf{y}^{k_j}, \mathbf{x} \rangle - \langle \xi^{k_j}, \mathbf{z} \rangle \right\}. \tag{D.3}$$

This implies that

$$\begin{aligned} &\tfrac{\mu_{k_j}}{2} \|\mathbf{x}^{k_j+1}\|^2 + \chi_\Omega(\mathbf{x}^{k_j+1}, g(\mathbf{x}^{k_j+1})) - \langle \mu_{k_j} \mathbf{v}^{k_j} - \nabla f(\mathbf{v}^{k_j}), \mathbf{x}^{k_j+1} \rangle - \langle \xi^{k_j}, g(\mathbf{x}^{k_j+1}) \rangle \\ &\leq \tfrac{\mu_{k_j}}{2} \|\mathbf{x}^*\|^2 + \chi_\Omega(\mathbf{x}^*, g(\mathbf{x}^*)) - \langle \mu_{k_j} \mathbf{v}^{k_j} - \nabla f(\mathbf{v}^{k_j}), \mathbf{x}^* \rangle - \langle \xi^{k_j}, g(\mathbf{x}^*) \rangle. \end{aligned}$$

Therefore

$$\begin{array}{ll} \frac{\mu_{k_j}}{2}\|\mathbf{x}^{k_j+1}-\mathbf{v}^{k_j}\|^2 & +\chi_\Omega(\mathbf{x}^{k_j+1},g(\mathbf{x}^{k_j+1}))\leq \frac{\mu_{k_j}}{2}\|\mathbf{x}^*-\mathbf{v}^{k_j}\|^2+\chi_\Omega(\mathbf{x}^*,g(\mathbf{x}^*)) \\ & +\langle\nabla f(\mathbf{v}^{k_j}),\mathbf{x}^*-\mathbf{x}^{k_j+1}\rangle-\langle\xi^{k_j},g(\mathbf{x}^*)-g(\mathbf{x}^{k_j+1})\rangle. \end{array}$$

Since there exists $\beta>0$ such that $\mu_0\leq\mu_k\leq\beta L$, we have

$$\begin{array}{ll} \frac{\mu_0}{2}\|\mathbf{x}^{k_j+1}-\mathbf{v}^{k_j}\|^2 & +\chi_\Omega(\mathbf{x}^{k_j+1},g(\mathbf{x}^{k_j+1}))\leq \frac{\beta L}{2}\|\mathbf{x}^*-\mathbf{v}^{k_j}\|^2+\chi_\Omega(\mathbf{x}^*,g(\mathbf{x}^*)) \\ & +\langle\nabla f(\mathbf{v}^{k_j}),\mathbf{x}^*-\mathbf{x}^{k_j+1}\rangle-\langle\xi^{k_j},g(\mathbf{x}^*)-g(\mathbf{x}^{k_j+1})\rangle. \end{array}$$

As $\lim_{j\to+\infty}\mathbf{x}^{k_j+1}=\lim_{j\to+\infty}\mathbf{v}^{k_j}=\mathbf{x}^*$, taking $j\to+\infty$ we obtain

$$\limsup_{j\to+\infty}\chi_\Omega(\mathbf{x}^{k_j+1},g(\mathbf{x}^{k_j+1}))\leq\chi_\Omega(\mathbf{x}^*,g(\mathbf{x}^*)). \tag{D.4}$$

Combining (D.4) with the lower semi-continuity of the function χ, we get

$$\limsup_{j\to+\infty}\chi_\Omega(\mathbf{x}^{k_j+1},g(\mathbf{x}^{k_j+1}))=\chi_\Omega(\mathbf{x}^*,g(\mathbf{x}^*)). \tag{D.5}$$

Similarly, it follows from (D.3) that

$$\begin{array}{ll} \frac{\mu_0}{2}\|\mathbf{x}^{k_j+1}-\mathbf{v}^{k_j}\|^2 & +\chi_\Omega(\mathbf{x}^{k_j+1},g(\mathbf{x}^{k_j+1}))\leq \frac{\beta L}{2}\|\mathbf{x}-\mathbf{v}^{k_j}\|^2+\chi_\Omega(\mathbf{x},\mathbf{z}) \\ & +\langle\nabla f(\mathbf{v}^{k_j}),\mathbf{x}-\mathbf{x}^{k_j+1}\rangle-\langle\xi^{k_j},\mathbf{z}-g(\mathbf{x}^{k_j+1})\rangle. \end{array} \tag{D.6}$$

Passing to the limit and combining with (D.5) we obtain

$$\chi_\Omega(\mathbf{x}^*,g(\mathbf{x}^*))\leq\frac{\beta L}{2}\|\mathbf{x}-\mathbf{x}^*\|^2+\chi_\Omega(\mathbf{x},\mathbf{z})+\langle\nabla f(\mathbf{x}^*),\mathbf{x}-\mathbf{x}^*\rangle-\langle\xi^*,\mathbf{z}-g(\mathbf{x}^*)\rangle.$$

This implies that $0\in(\nabla f(\mathbf{x}^*),-\xi^*)+\mathcal{N}_\Omega(\mathbf{x}^*,g(\mathbf{x}^*))$. Consequently

$$\begin{array}{ll} (0,\xi^*) & \in[(\nabla f(\mathbf{x}^*),0)+\mathcal{N}_\Omega(\mathbf{x}^*,g(\mathbf{x}^*))] \\ & \cap[0\times\partial(-h_i)(g_1(\mathbf{x}_1^*))\times\ldots\times\partial(-h_i)(g_m(\mathbf{x}_m^*))]. \end{array}$$

Therefore, $(\mathbf{x}^*,g(\mathbf{x}^*))$ is a critical point of (4). ■

Appendix E. Proof of Theorem 5

To prove Theorem 5, we use the following Lemma 4 whose the proof is similar to the proof of Lemma 2.

Lemma 4. *Let $\{\mathbf{x}^k\}$ be the sequence generated by ADCA-Like (Algorithm 4). If $\inf \varphi(\mathbf{x}, \mathbf{z}) > -\infty$, and the set of limit points $\mathcal{C}$ of $\{\mathbf{x}^k\}$ is not empty, then $\lim_{k\to+\infty} \varphi(\mathbf{x}^k, g(\mathbf{x}^k)) = \varphi(\mathbf{x}^*, g(\mathbf{x}^*))$ for some $\mathbf{x}^* \in \mathcal{C}$. Thus, φ has the same value on $\Lambda = \{(\mathbf{x}^*, g(\mathbf{x}^*)) : \mathbf{x}^* \in \mathcal{C}\}$.*

Proof of Theorem 5. (i) By Theorem 4 (ii), the sequences $\{\mathbf{x}^k\}$ and $\{\mathbf{v}^k\}$ share the same set of limit points $\mathcal{C}$. Since $\{\mathbf{x}^k\}$ is bounded, $\mathcal{C}$ is compact. As $H(\mathbf{z}) = \sum_{i=1}^m (-h_i)(z_i)$ is differentiable with locally Lipschitz derivative, by the continuity of g and by (ii) of Theorem 4 we see that there exist positive constants N_1 and π such that

$$\|\nabla H(g(\mathbf{x}^{k+1})) - \nabla H(g(\mathbf{v}^k))\| \leq \pi \|\mathbf{x}^{k+1} - \mathbf{v}^k\|, \forall k \geq N_1. \tag{E.1}$$

Since g is continuous, the set of limit points of $\{(\mathbf{x}^k, g(\mathbf{x}^k))\}$ is given by $\Lambda = \{(\mathbf{x}, g(\mathbf{x})) : \mathbf{x} \in \mathcal{C}\}$. Thus, Λ is also a compact set. By Lemma 4, φ has the same value φ^* on Λ. According to Lemma 1, there exist $\epsilon > 0$, $\eta > 0$ and $\psi \in \mathcal{M}_\eta$ such that for all $\forall (\mathbf{x}, \mathbf{z}) \in U := \{(\mathbf{x}, \mathbf{z}) : \text{dist}((\mathbf{x}, \mathbf{z}), \Lambda) < \epsilon\} \cap \{(\mathbf{x}, \mathbf{z}) : \varphi^* < \varphi(\mathbf{x}, \mathbf{z}) < \varphi^* + \eta\}$, we have

$$\psi'(\varphi(\mathbf{x}, \mathbf{z}) - \varphi^*) \text{dist}(0, \partial^L \varphi(\mathbf{x}, \mathbf{z})) \geq 1. \tag{E.2}$$

Similar to the proof of Theorem 3, without loss of generality, we can assume that $\varphi(\mathbf{x}^k, g(\mathbf{x}^k)) > \varphi^*$ for all k. By Lemma 4, there exists N_2 such that $\varphi^* < \varphi(\mathbf{x}^k, g(\mathbf{x}^k)) < \varphi^* + \eta$ whenever $k \geq N_2$. On the other hand, it follows from $\lim_{k\to+\infty} \text{dist}((\mathbf{x}^k, g(\mathbf{x}^k)), \Lambda) = 0$ that there is N_3 such that $\text{dist}((\mathbf{x}^k, g(\mathbf{x}^k)), \Lambda) < \epsilon$ for all $k \geq N_3$. Let $N = \max\{N_1, N_2, N_3\}$, we have $(\mathbf{x}^k, g(\mathbf{x}^k)) \in U$. It follows this and (E.2) that for all $k \geq N$,

$$\psi'(\mathbf{x}^k, g(\mathbf{x}^k)) - \varphi^*) \text{dist}(0, \partial^L \varphi(\mathbf{x}^k, g(\mathbf{x}^k))) \geq 1. \tag{E.3}$$

Since H is differentiable, by the definition of $\mathbf{x}^{k+1}$ we have

$$(\mathbf{y}^k, \nabla H(g(\mathbf{v}^k)) \in \partial G_{\mu_k}(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1})) = \mu_k \mathbf{x}^{k+1} + \partial \chi_\Omega(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1})).$$

This implies that

$$\begin{aligned}(\mu_k(\mathbf{v}^k - \mathbf{x}^{k+1}) - \nabla f(\mathbf{v}^k) + \nabla f(\mathbf{x}^{k+1}), \nabla H(g(\mathbf{v}^k))) \\ \in \partial \chi_\Omega(\mathbf{x}^{k+1}, g(\mathbf{x}^{k+1})) + (\nabla f(\mathbf{x}^{k+1}), 0).\end{aligned}$$

Therefore

$$\begin{aligned}&(\mu_k(\mathbf{v}^k-\mathbf{x}^{k+1})-\nabla f(\mathbf{v}^k)+\nabla f(\mathbf{x}^{k+1}),\nabla H(g(\mathbf{v}^k))-\nabla H(g(\mathbf{x}^{k+1})))\\ \in&\partial\chi_\Omega(\mathbf{x}^{k+1},g(\mathbf{x}^{k+1}))+(\nabla f(\mathbf{x}^{k+1}),0)-(0,\nabla H(g(\mathbf{x}^{k+1})))\\ =&\partial^L\varphi(\mathbf{x}^{k+1},g(\mathbf{x}^{k+1})).\end{aligned}\tag{E.4}$$

By the Lipschitz continuity of ∇f, $\mu_k \le \beta L$ and (E.1) for any $k \ge N$, we have $\|(\mu_k(\mathbf{v}^k-\mathbf{x}^{k+1})-\nabla f(\mathbf{v}^k)+\nabla f(\mathbf{x}^{k+1}),\nabla H(g(\mathbf{v}^k))-\nabla H(g(\mathbf{x}^{k+1})))\| \le M\|\mathbf{x}^{k+1}-\mathbf{v}^k\|$, where $M=\beta L+L+\pi$. Combining this and (E.4) we get $\mathrm{dist}(0,\partial^L\varphi(\mathbf{x}^{k+1},g(\mathbf{x}^{k+1}))) \le M\|\mathbf{x}^{k+1}-\mathbf{v}^k\|$. Denote $r_{k+1}=\varphi(\mathbf{x}^{k+1},g(\mathbf{x}^{k+1}))-\varphi^* > 0$. It follows from the above inequality and (E.3) that, for all $k \ge N$,

$$\begin{aligned}1 &\le [\psi'(\varphi(\mathbf{x}^{k+1},g(\mathbf{x}^{k+1}))-\varphi^*)M\|\mathbf{x}^{k+1}-\mathbf{v}^k\|]^2\\ &\le [\psi'(r_{k+1})]^2M^2\frac{2(\varphi(\mathbf{x}^k,g(\mathbf{x}^k))-\varphi(\mathbf{x}^{k+1},g(\mathbf{x}^{k+1})))}{u_k}\\ &\le [\psi'(r_{k+1})]^2M^2\frac{2(r_k-r_{k+1})}{u_0},\end{aligned}\tag{E.5}$$

where the second inequality follows from (i) of Theorem 4, and the last inequality holds by $\mu_k \ge \mu_0$. Since ψ takes the form $\psi(s)=cs^{1-\theta}$, we get $\psi'(s)=c(1-\theta)s^{-\theta}$. Therefore, the inequality (E.5) becomes

$$r_{k+1}^{2\theta} \le \frac{2c(1-\theta)M^2}{\mu_0}(r_k-r_{k+1}),\ \forall k \ge N.\tag{E.6}$$

Finally, applying Lemma 3 with $s_k=r_k$, $\alpha=2\theta$ and $\gamma=\frac{2c(1-\theta)M^2}{\mu_0}$ gives us that the sequence $\{r_k\}$ converges to 0 with the rates corresponding to different values of θ as started in this theorem. ■

Appendix F. Proof of Proposition 3

We recall the definition of function f

$$f(\mathbf{x}) := \sum_{i\neq j} p_{ij}\log p_{ij} + \log\left(\sum_{i\neq j}(1+\|\mathbf{x}_i-\mathbf{x}_j\|^2)^{-1}\right).$$

Note that p_{ij} is known in advance hence the first term of f is a constant. Let's consider the second term of f

$$\varphi(\mathbf{x}) = \log\left(\sum_{i\neq j}(1+\|\mathbf{x}_i-\mathbf{x}_j\|^2)^{-1}\right).$$

For $i \neq j$, let's define the function

$$z_{ij}(\mathbf{x}) = \|\mathbf{x}_i - \mathbf{x}_j\|^2.$$

Then, we have

$$\varphi(\mathbf{x}) = \log\left(\sum_{i\neq j} \frac{1}{1 + z_{ij}(\mathbf{x})}\right),$$

or equivalently

$$e^{\varphi(\mathbf{x})} = \sum_{i\neq j} \frac{1}{1 + z_{ij}(\mathbf{x})}.$$

Taking derivatives in $\mathbf{x}$ both sides of the above equation, we get

$$\nabla\varphi(\mathbf{x})\ e^{\varphi(\mathbf{x})} = \sum_{i\neq j} \frac{-1}{(1 + z_{ij}(\mathbf{x}))^2} \nabla z_{ij}(\mathbf{x}).$$

Taking derivatives in $\mathbf{x}$ both sides of the above equation, we get

$$\begin{aligned}
&\nabla^2\varphi(\mathbf{x})\ e^{\varphi(\mathbf{x})} + \nabla\varphi(\mathbf{x})(\nabla\varphi(\mathbf{x}))^T\ e^{\varphi(\mathbf{x})} \\
&\qquad = \sum_{i\neq j} \left(\frac{2}{(1 + z_{ij}(\mathbf{x}))^3} \nabla z_{ij}(\mathbf{x})(\nabla z_{ij}(\mathbf{x}))^T + \frac{-1}{(1 + z_{ij}(\mathbf{x}))^2} \nabla^2 z_{ij}(\mathbf{x})\right)
\end{aligned}$$

Note that z_{ij} are convex then $\nabla^2 z_{ij}(\mathbf{x}) \succeq 0$. The above equation implies that

$$\begin{aligned}
\nabla^2\varphi(\mathbf{x})\ e^{\varphi(\mathbf{x})} &\preceq \sum_{i\neq j} \frac{2}{(1 + z_{ij}(\mathbf{x}))^3} \nabla z_{ij}(\mathbf{x})(\nabla z_{ij}(\mathbf{x}))^T \\
&\preceq \sum_{i\neq j} \frac{2\|\nabla z_{ij}(\mathbf{x})\|^2}{(1 + z_{ij}(\mathbf{x}))^3}\ \mathbf{I}.
\end{aligned}$$

Moreover,

$$\|\nabla z_{ij}(\mathbf{x})\|^2 = 8\|\mathbf{x}_i - \mathbf{x}_j\|^2 = 8z_{ij}(\mathbf{x}) \leq 2(1 + z_{ij}(\mathbf{x}))^2.$$

Thus,

$$\nabla^2\varphi(\mathbf{x})\ e^{\varphi(\mathbf{x})} \preceq \sum_{i\neq j} \frac{4}{1 + z_{ij}(\mathbf{x})}\ \mathbf{I} = 4\ e^{\varphi(\mathbf{x})}\ \mathbf{I},$$

or equivalently,

$$\nabla^2\varphi(\mathbf{x}) \preceq 4\mathbf{I}.$$

Therefore, $\varphi(\mathbf{x})$ is Lipschitz gradient with a Lipschitz constant $L = 4$. ■

Appendix G. MM method for t-SNE

We will show that the MM method ([24]) for t-SNE is a special version of DCA-Like. Indeed, the MM method first contructs a majorization function $\bar{F}$ such that $F(x) \leq \bar{F}(\mathbf{x}, \mathbf{y}), \forall \mathbf{x}, \mathbf{y}$ and $F(\mathbf{x}) = \bar{F}(\mathbf{x}, \mathbf{x}), \forall \mathbf{x}$, then update $\mathbf{x}$ at each iteration k by $\mathbf{x}^{k+1} \in arg\min \bar{F}(\mathbf{x}, \mathbf{x}^k)$. In [24], the authors rewrote the objective function F of t-SNE by the same way as in (21) and then proposed the QL (Quadratification and Lipschitzation) majorization for finding the majorization function $\bar{F}$. More precisely, the step quadratification consists in majorizing the part $\sum\limits_{i,j} p_{ij} \log\left(1 + \|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$ by a pairwise-quadratic terms while in the step Lipschitzation the part $\log(\sum\limits_{i \neq j}(1+\|\mathbf{x}_i - \mathbf{x}_j\|^2)^{-1})$ is majorized by its Lipschitz surrogate. The majorization function $\bar{F}(\mathbf{x}, \mathbf{y})$ is given by

$$\begin{aligned}\bar{F}(\mathbf{x}, \mathbf{y}) :=& F(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2 \\ &+ \sum_{i,j} \frac{p_{ij}}{1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2}(\|\mathbf{x}_i - \mathbf{x}_j\|^2 - \|\mathbf{y}_i - \mathbf{y}_j\|^2)\end{aligned} \tag{G.1}$$

As mentioned above, at each iteration, MM compute the next iterate $\mathbf{x}^{k+1}$ by solving the optimization problem $\min \bar{F}(\mathbf{x}, \mathbf{x}^k)$. We see that this optimization problem is nothing else but the sub-problem (25) in our DCA-Like scheme. Furthermore, in [24], the authors used the backtracking method to compute the Lipschitz constant μ. Finally, the MM method for t-SNE is described in Algorithm 5.

We observe that the MM method for t-SNE (Algorithm 5) is a special case of DCA-Like for t-SNE where $\delta = \frac{1}{\eta}$ and we omit max operator in the Step 2 of DCA-Like while updating μ_k. In fact, even if the DC structure is hidden, with the choice of the surrogate function by using QL-majorization, the MM method results to a DCA version. Furthermore, the use of backtracking method for the computation of μ in MM is motivated by the same reason of DCA-Like, e.g. keeping μ as small as possible.

Algorithm 5: MM for t-SNE ([24])

Input : an initial point $\mathbf{x}_0$, an initial Lipschitz constant μ_0, an inflating factor $\eta > 1$.

Initialization : Set $k = 0$.

repeat

1: Compute t_{ij}^k and $\nabla f(\mathbf{x}^k)$ following (24);

2: Set $\mu_k = \frac{1}{\delta}\mu_{k-1}$ if $k > 0$;

3: Compute $\mathbf{x}^{k+1}$ by (26);

4: **while** $\bar{F}(\mathbf{x}^{k+1}, \mathbf{x}^k) < F(\mathbf{x}^{k+1})$ **do**

$\mu_k \leftarrow \eta\mu_k$;

Update $\mathbf{x}^{k+1}$ by (26);

end

5: $k \leftarrow k + 1$.

until *Stopping criterion*;

Output: $\mathbf{x}^k$ - a critical point of (20).

Highlights

- We address the problem of minimizing the sum of a nonconvex, differentiable function and composite functions by DC (Difference of Convex functions) programming and DCA (DC Algorithm).
- We first develop a standard DCA scheme especially dealing with the very specific structure of this problem.
- Furthermore, we extend DCA to give rise to the so-named DCA-Like, which is based on a new and efficient way to approximate the DC objective function without knowing a DC decomposition.
- We further improve DCA based algorithms by incorporating the Nesterov's acceleration technique into them.
- Finally, we investigate the proposed algorithms for an important problem in machine learning: the t-distributed stochastic neighbor embedding.