Algorithms for Difference-of-Convex (DC) Programs Based on Difference-of-Moreau-Envelopes Smoothing

Kaizhao Sun
School of Industrial and Systems Engineering, Georgia Institute of Technology, ksun46@gatech.edu
Xu Andy Sun
Operations Research and Statistics, Sloan School of Management, Massachusetts Institute of Technology, sunx@mit.edu

In this paper we consider minimization of a difference-of-convex (DC) function with and without linear equality constraints. We first study a smooth approximation of a generic DC function, termed difference-of-Moreau-envelopes (DME) smoothing, where *both* components of the DC function are replaced by their respective Moreau envelopes. The resulting smooth approximation is shown to be Lipschitz differentiable, capture stationary points, local, and global minima of the original DC function, and enjoy some growth conditions, such as level-boundedness and coercivity, for broad classes of DC functions. For a smoothed DC program without linear constraints, it is shown that the classic gradient descent method as well as an inexact variant converge to a stationary solution of the original DC function in the limit with a rate of $\mathcal{O}(K^{-1/2})$, where K is the number of proximal evaluations of both components. Furthermore, when the DC program is explicitly constrained in an affine subspace, we combine the smoothing technique with the augmented Lagrangian function and derive two variants of the augmented Lagrangian method (ALM), named LCDC-ALM and composite LCDC-ALM, targeting on different structures of the DC objective function. We show that both algorithms find an ϵ-approximate stationary solution of the original DC program in $\mathcal{O}(\epsilon^{-2})$ iterations. Comparing to existing methods designed for linearly constrained weakly convex minimization, the proposed ALM-based algorithms can be applied to a broader class of problems, where the objective contains a nonsmooth concave component. Finally, numerical experiments are presented to demonstrate the performance of the proposed algorithms.

Key words: difference-of-convex optimization; Moreau envelope; augmented Lagrangian method; proximal point method

1. Introduction

In this paper, we consider the following unconstrained difference-of-convex (DC) program

$$\min_{x\in\mathbb{R}^n} F(x) := \phi(x) - g(x), \tag{1}$$

and the linearly constrained DC (LCDC) program

$$\min_{x\in\mathbb{R}^n} F(x) \quad \text{s.t.} \quad Ax = b, \tag{2}$$

where $\phi:\mathbb{R}^n \to \overline{\mathbb{R}} := \mathbb{R}\cup\{+\infty\}$ is a weakly-convex function, $g:\mathbb{R}^n\to\mathbb{R}$ is a proper and closed convex function, $A\in\mathbb{R}^{m\times n}$, and $b\in\mathbb{R}^m$. We also work with a more structured setting, where $\phi = f + h$ in a composite form, $h:\mathbb{R}^n\to\overline{\mathbb{R}}$ is a proper, closed, and convex function, and $f:\mathbb{R}^n\to\mathbb{R}$ has a

Lipschitz gradient over the effective domain of h. DC functions include some important classes of nonconvex functions, such as twice continuously differentiable functions on any compact convex subset of $\mathbb{R}^n$ (Tuy 2010), continuous piecewise-linear functions (Melzer 1986), and multivariate polynomial functions (Ahmadi and Hall 2018). DC programs (1)-(2) appear in various applications such as compressed sensing (Yin et al. 2015), high-dimensional statistical regression (Cao et al. 2018), and power allocation in digital communication systems (Alvarado et al. 2014), just to name a few.

In (1) and (2), we only require the first objective component ϕ to be weakly convex, i.e., there exists $m_\phi \geq 0$ such that $\phi + \frac{m_\phi}{2}\|\cdot\|^2$ is a convex function, where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^n$. Notice that the objective F is still a DC function since $F = \left(\phi + \frac{m_\phi}{2}\|\cdot\|^2\right) - \left(g + \frac{m_\phi}{2}\|\cdot\|^2\right)$. sNevertheless, we allow ϕ to be merely weakly convex, as a convex decomposition of F may not be readily available or necessary. Moreover, we distinguish the DC program (1) from the LCDC program (2) for algorithmic purposes. In particular, directly turning (2) to (1) by incorporating linear constraints into the objective using an indicator function may complicate computation, and in situations such as distributed optimization, the data in the linear constraints may not be available to all agents in the optimization problem.

In this paper, we first study a smoothing technique to obtain Lipschitz differentiable DC approximations of a DC function. Then, we combine the smoothing technique with first-order algorithms to solve the DC programs (1) and (2). It is well-known that the Moreau envelope of a proper closed convex function is a Lipschitz differentiable convex function. Similarly, a nonsmooth weakly convex function can be smoothed by its Moreau envelope. However, directly applying Moreau envelope to the DC function $F = \phi - g$ as a whole might be problematic. On the one hand, the proximal mapping of F can be difficult to compute or even not well-defined; on the other hand, due to the concave component $-g$, the Moreau envelope of F might not be smooth (see Fig. 1 for an illustration). With these observations, we propose to study a simple smoothing technique, termed difference-of-Moreau-envelope (DME) smoothing, which replaces (weakly) convex components ϕ and g separately by their respective Moreau envelopes, i.e., we form

$$F_\mu := M_{\mu\phi} - M_{\mu g}, \tag{3}$$

where $M_{\mu f}(z) := \min_{x\in\mathbb{R}^n}\{f(x) + \frac{1}{2\mu}\|x - z\|^2\}$ is a Moreau envelope of f with parameter $\mu > 0$. It is easy to see that F_μ is a Lipschitz differentiable DC function for all $\mu > 0$. Moreover, it has been shown that F_μ preserves the global minimum of F when it exists, i.e. min $F_\mu =$ min F. Consequently, it is natural to use F_μ as a surrogate function in search of a minimizers of F.

As we finished the manuscript, we discovered that this smoothing technique has been considered as a realization of the Toland duality by Ellaia (1984) and applied to smooth DC functions by

Hiriart-Urruty (1985, 1991) as a result. In addition, a very recent paper by Themelis et al. (2020) proposed to run gradient descent on F_μ, i.e., through the update

$$z^{k+1} = z^k - \alpha \nabla F_\mu(z^k) \tag{4}$$

for some proper stepsize $\alpha > 0$, which coincides with one of the four algorithms studied in an early version of the present paper. Nevertheless, (i) some important properties of the smoothed DC function F_μ, such as the correspondence between the local minima of F and F_μ and the growth properties of F_μ, (ii) an inexact gradient descent method on F_μ, and (iii) the combination of DME technique with the augmented Lagrangian framework for LCDC (2) are new to our best knowledge. We summarize our contributions in the next subsection.

1.1. Contributions

1. We carry out a new study on the DME smoothing of DC functions using two separate Moreau envelopes of the convex components as in F_μ defined in (3). We show that local minimal solutions of F_μ can be mapped to those of F, complementing existing results on the correspondence between stationary points and global minima of F_μ and F. Moreover, we identify four general conditions on F such that, under each of them, the smoothed DC function F_μ enjoys coercivity or level-boundedness, which is important for the convergence analysis of the proposed algorithms that use F_μ as a surrogate for solving the DC programs.

2. We propose gradient-based algorithms on F_μ for the unconstrained DC program (1). As shown in Themelis et al. (2020) and also independently developed in this paper, the classic gradient descent (GD) algorithm applied to F_μ converges with rate $\mathcal{O}(K^{-1/2})$ where K is the number of proximal evaluations of ϕ and g, and finds a stationary point of F in the limit. When $\phi = f + h$, where f has Lispchitz gradient and h is proper, closed, and convex, we propose an inexact GD method for minimizing F_μ that admits a simpler subproblem: f can be replaced by its linearization during the proximal evaluation of ϕ. The inexact GD achieves the same $\mathcal{O}(K^{-1/2})$ convergence rate, and can be realized as a variant of the proximal DC Algorithm (pDCA) with a novel choice of the proximal center. A key feature in contrast to DCA-type algorithms is that, we perform a proximal instead of subgradient evaluation on g, and proximal mappings of h (or ϕ in GD) and g can be implemented in parallel.

3. For the linearly constrained DC program (2), we apply the smoothing property of Moreau envelopes to the classic augmented Lagrangian function in augmented Lagrangian method (ALM) and propose two ALM-base algorithms, LCDC-ALM and composite LCDC-ALM, for different realizations of ϕ (see Table 1 below). In each iteration of LCDC-ALM, the augmented Lagrangian function is smoothed by the DME technique, and then a primal descent step and a dual ascent step

are performed. In composite LCDC-ALM, we replace the negative component $-g$ by its linearization, and then smooth a linearized augmented Lagrangian function with its Moreau envelope. For both algorithms, we prove that an ϵ-stationary solution to (2) can be found in at most $\mathcal{O}(\epsilon^{-2})$ subproblem oracle calls. Moreover, we establish an $\tilde{\mathcal{O}}(\epsilon^{-3})$ first-order complexity estimate for composite LCDC-ALM when each subproblem is solved by a proper first-order algorithm invoking gradient oracles of f and proximal oracles of h. LCDC-ALM and composite LCDC-ALM extend ALM-based algorithms considered in Hajinezhad and Hong (2019), Hong (2016), Li and Xu (2021), Melo and Monteiro (2020), Melo et al. (2020), Zhang and Luo (2020b) by allowing a concave component $-g$ in the objective, though we believe that the aforementioned algorithms have the potential to handle DC objective functions.

In addition to the gradient descent scheme (4), whose theoretical properties have already been established by Themelis et al. (2020), we present three new algorithms in this paper for solving structured DC programs (1) and (2), summarized as follows.

Table 1 Summary of the Proposed Algorithms

Problem	f	$h = \phi - f$	g	Algorithm	$\mathcal{O}(\epsilon^{-2})$ Iter. Complexity
(1)	Lip. gradient	proper closed convex	$\mathcal{P}$	Inexact GD on F_μ (Algorithm 1)	Theorem 2
(2)	Lip. gradient	0	$\mathcal{P}$	LCDC-ALM (Algorithm 2)	Theorem 3
	Lip. gradient	convex and Lipschitz over a compact domain	$\mathcal{G}$	Composite LCDC-ALM (Algorithm 3)	Theorem 4

The function g is assumed to be a finite-valued convex function in all three algorithms, where either a proximal ($\mathcal{P}$) or subgradient ($\mathcal{G}$) oracle is used in the proposed algorithms.

1.2. Notations and Organization

We denote the set of positive integers up to integer p by $[p]$, the set of nonnegative integers by $\mathbb{N}$, the set of real numbers and nonnegative real numbers by $\mathbb{R}$ and $\mathbb{R}_+$, respectively, the extended real line by $\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$, and the n-dimensional real Euclidean space by $\mathbb{R}^n$. For $x, y \in \mathbb{R}^n$, the inner product of x and y is denoted by $\langle x, y \rangle$, and the Euclidean norm of x is denoted by $\|x\|$. For $A \in \mathbb{R}^{m\times n}$, $\|A\|$, $\sigma_{\min}(A)$, $\sigma^+_{\min}(A)$, $\lambda_{\max}(A)$, and $\mathrm{Im}(A)$ denote the matrix norm induced by the Euclidean norm, the smallest singular value, the smallest positive singular value, the largest eigenvalue (when $m = n$), and the column space of A, respectively. For $X \subseteq \mathbb{R}^n$, δ_X denotes its $0/\infty$-indicator function, i.e., $\delta_X(x) = 0$ for $x \in X$, and $\delta_X(x) = +\infty$ for $x \notin X$. For a proper function $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$, we denote its effective domain by $\mathrm{dom}\ \phi := \{x \in \mathbb{R}^n \mid \phi(x) < +\infty\}$.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 introduces the DME technique and studies some new properties of the smooth approximation F_μ. Section

4 shows that applying GD or inexact GD method to the smooth approximation can deliver a stationary solution of (1). Section 5 presents two ALM-based algorithms with convergence analysis for the LCDC problem (2) under different assumptions of ϕ. We present numerical experiments in Section 6, and finally give some concluding remarks in Section 7.

2. Related Work

In this section, we review some related work in smoothing, DC algorithms, and ALM.

2.1. Smooth Approximation of Nonsmooth Functions

There is a rich literature on approximation of nonsmooth nonconvex functions. Chen and Mangasarian (1996) proposed a class of smooth approximations for the max function $(t)_+ = \max\{0, t\}$ using integral convolution with some symmetric and piecewise-continuous density function, which is also closely related to mollification in functional analysis. Chen (2012) exploited this scheme in approximating composition of C^1 functions with $(t)_+$. Lu (2014) proposed a Lipschitz continuous ϵ-approximation of the function $\|\cdot\|_p^p$ for $p \in (0,1)$, where the tolerance $\epsilon > 0$ controls the approximation error. The smooth approximation (3) is a different smoothing technique, relying on inf-convolution, rather than on integral convolution. In Ellaia (1984), it is shown that F_μ is a special way to realize Toland duality of DC programs by adding $\frac{1}{2\mu}\|\cdot\|^2$ to the convex components. In Hiriart-Urruty (1985) and Hiriart-Urruty (1991), it is shown that, as a result of Toland duality, the stationary points and global optima of F_μ and F are closely related.

2.2. DC Algorithms

Algorithms for minimizing a weakly convex function have been studied in the literature (Davis and Drusvyatskiy 2018, Drusvyatskiy 2017). When $-g$ cannot be convexified through the addition of a quadratic form, a classic iterative approach to finding stationary solutions of (1) is the so-called DC algorithm (DCA) proposed by Tao and An (1997), (see also Tao et al. (2005)), where $-g$ is replaced by a linear over-approximation in each iteration, using subgradient information; see Artacho et al. (2018), Le Thi et al. (2012), Le Thi Hoai and Tao (1997), Tao and An (1998) for specific applications and convergence results of DCA.

An important variant of DCA is the proximal DCA (pDCA), which, in addition to replacing the concave function with a linear majorization, includes a proximal term in the minimization problem at iteration $k+1$:

$$x^{k+1} = \underset{x\in\mathbb{R}^n}{\arg\min} \left\{ \phi(x) - \langle \xi_g^k, x - x^k \rangle + \frac{1}{2c_k}\|x - x^k\|^2 \right\}, \tag{5}$$

where $c_k > 0$ and $\xi_g^k \in \partial g(x^k)$, e.g. see Souza et al. (2016), Sun et al. (2003). When $\phi = f + h$, where f has Lipschitz gradient and h is proper closed and convex, one can further linearize f (An and Nam 2017), and replace the proximal center x^k by a proper extrapolation (Wen et al. 2018). More recently, de Oliveira (2019, 2020) replaced the positive component ϕ with some minorant function such as a bundle of cutting planes.

Pang et al. (2017) characterized different types of stationary points and their relations for generic DC programs. They considered the case $g(x) = \max_{i \in I} g_i(x)$, where each g_i is continuously differentiable and convex and I is a finite index set, and proposed an enhanced DCA that subsequentially converges to a d(irectional)-stationary point. Furthermore, they showed that the algorithm can be extended to compute a B(ouligand)-stationary point for a class of DC constrained DC program under a suitable constraint qualification. This work also motivates several follow-up works. Lu et al. (2019) extended the problem setting to allow an infinite supremum in the definition of g; by incorporating the extrapolation technique into the enhanced DCA in Pang et al. (2017), they proposed an algorithm named EPDCA and established an iteration complexity of $\mathcal{O}(\epsilon^{-2})$ for computing an approximate stationary solution with tolerance $\epsilon > 0$. Lu and Zhou (2019) later applied nonmonotone line-search schemes and randomized update to accelerate the convergence of EPDCA in practice. Tao and Dong (2018) established linear convergence of the enhanced DCA by utilizing locally linear regularity and error bound conditions.

Another related work by Banert and Boț (2019) applied the proximal alternating linearized minimization (PALM) algorithm (Bolte et al. 2014) to a primal-dual formulation of a DC program. Convergence rate results are established under the Kurdyka–Łojasiewicz (KŁ) property (Kurdyka 1998, Lojasiewicz 1963).

2.3. ALM for Linearly-constrained Weakly Convex Minimization

The augmented Lagrangian method (ALM), which was proposed in the late 1960s by Hestenes (1969) and Powell (1969), provides a powerful algorithmic framework for constrained optimization. The convergence, local, and global convergence rate of ALM have been extensively studied for convex programs (Aybat and Iyengar 2013, Lan and Monteiro 2016, Li and Qu 2019, Liu et al. 2019, Lu and Zhou 2018, Rockafellar 1973, 1976, Xu 2021) and smooth nonlinear programs (Bertsekas 2014, Sahin et al. 2019). In the following, we review some recent developments in ALM-based algorithms applied to linearly-constrained nonconvex problems of the form

$$\min_{x \in \mathbb{R}^n} \phi(x) = f(x) + h(x) \quad \text{s.t.} \quad Ax = b, \tag{6}$$

where f has Lipschitz gradient and h is a possibly nonsmooth convex function.

Hong (2016) considered the case where $h = 0$, and proposed a proximal ALM: in iteration $k+1$, an additional proximal term $\frac{\rho}{2}\|x - x^k\|^2_{B^\top B}$ is added to the augmented Lagrangian (AL) function. Assuming h is a compactly supported convex function, Hajinezhad and Hong (2019) proposed a perturbed proximal ALM that will converge to a solution with bounded infeasibility; when the initial point is feasible, they established an iteration complexity of $\mathcal{O}(\epsilon^{-4})$, where $\epsilon > 0$ measures both stationarity and feasibility. Under the same perturbed AL framework, Melo et al. (2020) applied an accelerated composite gradient method (ACG) (Beck and Teboulle 2009) to solve each proximal ALM subproblem and obtained an improved iteration complexity of $\mathcal{O}(\log(\epsilon^{-1})\epsilon^{-3})$ total ACG iterations, which can be further reduced to $\mathcal{O}(\log(\epsilon^{-1})\epsilon^{-2.5})$ with mildly stronger assumptions. In Melo and Monteiro (2020), this inner acceleration scheme is embedded in the proximal ALM with full dual multiplier update. Li and Xu (2021) combined an inexact ALM and a quadratic penalty method to solve a convex-constrained program with a weakly-convex objective function, and they showed that an ϵ-KKT solution can be found in $\mathcal{O}(\log(\epsilon^{-1})\epsilon^{-2.5})$ adaptive accelerated proximal gradient steps.

Finally we highlight two recent works by Zhang and Luo (2020a,b) and a further generalization by Zeng et al. (2021). Zhang and Luo studied the case where h is an indicator function of a hypercube or a polyhedron. In practice, their specially chosen proximal term prevents iterates from oscillation, therefore stabilizing the algorithm; in the convergence proof, the authors constructed a novel potential function, which combines properties from primal descent, dual ascent, and proximal descent. Utilizing an error bound that exploits the polyhedral structure of h, the authors showed the proposed algorithm will find an ϵ-approximate KKT point in $\mathcal{O}(\epsilon^{-2})$ iterations. Moreover, their methodology has been successfully applied to min-max problems (Zhang et al. 2020) and distributed optimization (Chen et al. 2021). More recently, Zeng et al. (2021) generalized the setting (6) to any weakly-convex objective, and proposed a Moreau Envelope ALM (MEAL) that achieves the same iteration complexity under either the implicit Lipschitz subgradient condition or the implicit bounded subgradient condition. The novel conditions relax the standard smoothness and bounded subgradient conditions commonly adopted in weakly-convex minimization, and provide a new way to control dual iterates using primal iterates in the analysis of ALM.

3. Smooth Approximation of DC Functions

3.1. Assumptions and Stationarity

We make the following assumptions in this paper regarding the DC function $F = \phi - g$.

Assumption 1. *The function $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper, lower semi-continuous, and m_ϕ-weakly convex, i.e., there exists a finite $m_\phi \geq 0$ such that $\phi + \frac{m_\phi}{2}\|\cdot\|^2$ is a convex function. The function $g : \mathbb{R}^n \to \mathbb{R}$ is convex. Moreover, when $\phi = f + h$ admits a composite form, the function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper, closed, and convex, and $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz differentiable with modulus $L_f > 0$ over $\mathcal{H} := \mathrm{dom}\ h = \{x \in \mathbb{R}^n \mid h(x) < +\infty\}$.*

Assumption 2. *The set of global minimizers of F, $\arg\min F$, is nonempty.*

By Assumptions 1 and 2, it holds that

$$\mathbb{R}^n = \text{dom } g = \text{dom } \partial g := \{x \in \mathbb{R}^n : \ \partial g(x) \neq \emptyset\}, \text{ and } F^* := \min_{x \in \mathbb{R}^n} F(x) < +\infty. \tag{7}$$

In addition, we adopt the following definition for stationary points of F.

Definition 1. We say $x \in \mathbb{R}^n$ is a stationary point of F if

$$0 \in \partial\phi(x) - \partial g(x); \tag{8}$$

or equivalently, $\partial\phi(x) \cap \partial g(x) \neq \emptyset$. Furthermore, given $\epsilon > 0$, we say $x \in \mathbb{R}^n$ is an ϵ-stationary point of F if there exists $(\xi, y) \in \mathbb{R}^n \times \mathbb{R}^n$ such that

$$\xi \in \partial\phi(x) - \partial g(y), \text{ and } \max\{\|\xi\|, \|x - y\|\} \leq \epsilon. \tag{9}$$

When $\phi = f + h$, we simply replace $\partial\phi(x)$ by $\nabla f(x) + \partial h(x)$ in (8) and (9).

Remark 1. We give some comments regarding Definition 1.

1. We use $\partial\phi(x)$ to denote the general subdifferential of ϕ at x (Rockafellar and Wets 2009, Definition 8.3). If ϕ is continuously differentiable in a neighborhood of x, then $\partial\phi(x) = \{\nabla\phi(x)\}$; if $\phi = f + h$ with f continuously differentiable and h finite at x, then $\partial\phi(x) = \nabla f(x) + \partial h(x)$ (Rockafellar and Wets 2009, Exercise 8.8).

2. Compared to (8), a more natural stationarity condition of F would be $0 \in \partial(\phi - g)(x)$. In view of the previous comment, if g is continuously differentiable over dom ϕ, then $\partial(\phi - g)(x) = \partial\phi(x) - \nabla g(x)$ for all $x \in \text{dom } \phi$. However, the equality does not hold in general. The condition (8) is known as criticity in the DC literature. Some other stationary conditions include d(irectionary)-stationarity and c(larke)-stationarity, which are defined as the directional derivative or Clarke directional derivative being nonnegative in every feasible direction. Under suitable settings, e.g., see Pang et al. (2017), local minimum of F implies

$$\partial g(x) \subseteq \partial\phi(x) \equiv \text{d-stationary} \Rightarrow \text{c-stationary} \Rightarrow (8).$$

Clearly the above three types of stationarity are equivalent when $\partial g(x) = \{\nabla g(x)\}$. However, since we do not impose any smoothness structure on g, computing (or even verifying) a d-stationary point can be intractable, so we adopt the more general notion (8) in this paper. More details on different stationarity conditions can be found in Pang et al. (2017) and de Oliveira (2019).

3. Usually $\partial\phi$ and ∂g are not evaluated at the same point during DCA or pDCA. For example in (5), a subgradient $\xi_g^k \in \partial g(x^k)$ is used to set x^{k+1}. So (9) is a natural relaxation of (8), and can serve as the stopping criteria for some iterative approach in practice.

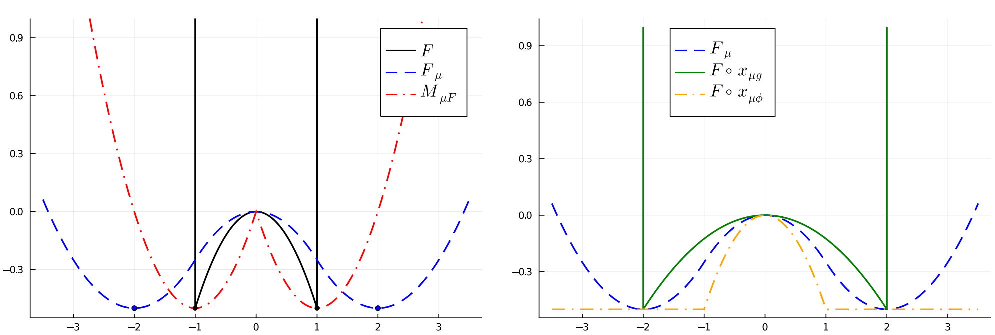
3.2. New Properties of the Difference-of-Moreau-Envelopes Smoothing

Given $0 < \mu < 1/m_\phi$, the Moreau envelope $M_{\mu\phi} : \mathbb{R}^n \to \mathbb{R}$ and the proximal mapping $x_{\mu\phi} : \mathbb{R}^n \to \mathbb{R}^n$ of ϕ are defined as

$$M_{\mu\phi}(z) := \min_{x\in\mathbb{R}^n} \left\{ \phi(x) + \frac{1}{2\mu}\|x - z\|^2 \right\}, \quad x_{\mu\phi}(z) := \arg\min_{x\in\mathbb{R}^n} \left\{ \phi(x) + \frac{1}{2\mu}\|x - z\|^2 \right\}. \tag{10}$$

It is known that $M_{\mu\phi}$ forms a smooth approximation of the possibly nonconvex nonsmooth function ϕ. Similarly, the Moreau envelope and proximal mapping of g are given by $M_{\mu g}$ and $x_{\mu g}$, respectively. Consequently, the function $F_\mu(z) = M_{\mu\phi}(z) - M_{\mu g}(z)$ constitutes a smooth approximation of F. Our motivation to form the Moreau envelopes of ϕ and g separately comes from the observation that the Moreau envelope of a concave function may not be smooth, therefore, the Moreau envelope $M_{\mu F}$ of the DC function $F = \phi - g$ as a whole may not be smooth. But smoothing each component of F separately will surely give a smooth DC function. This is shown in the next example. Consider $F = \phi - g$, where $\phi(x) = \delta_{[-1,1]}(x)$, and $g(x) = \frac{1}{2}x^2$. In Fig. 1(a) , we plot F, the smooth approximation F_μ, and the Moreau envelope $M_{\mu F}$ of F; we see that $M_{\mu F}$ is not smooth at the origin. In Fig. 1(b), we see the smooth approximation F_μ is further bounded by $F \circ x_{\mu\phi}$ from below and by $F \circ x_{\mu g}$ from above; this is formally stated in Lemma 1 and the proof follows directly from the definitions of F_μ, $x_{\mu g}(z)$, and $x_{\mu\phi}(z)$.

Figure 1 **An Example:** $F = \phi - g$, **where** $\phi(x) = \delta_{[-1,1]}(x)$, **and** $g(x) = \frac{1}{2}x^2$



(a) F, F_μ, and $M_{\mu F}$ $(\mu = 1)$ (b) $F \circ x_{\mu\phi} \le F_\mu \le F \circ x_{\mu g}$ $(\mu = 1)$

Lemma 1 **(Bounds on F_μ).** *Suppose Assumption 1 holds and $0 < \mu < 1/m_\phi$ in (10). For all $z \in \mathbb{R}^n$, it holds $(F \circ x_{\mu\phi})(z) \le F_\mu(z) \le (F \circ x_{\mu g})(z)$.*

In EC.1, we provide some preliminary properties on DME smoothing. In particular, it has been established that (i) F_μ has a Lipschitz gradient with

$$\nabla F_\mu(z) = \mu^{-1}(x_{\mu g}(z) - x_{\mu\phi}(z)) \tag{11}$$

for all $z \in \mathbb{R}^n$, (ii) if $z \in \mathbb{R}^n$ is a stationary point or a global minimizer of F_μ, then $x_{\mu\phi}(z) = x_{\mu g}(z)$ is a stationary point or a global minimizer of F, respectively, and (iii) the converse of (ii) holds as well. Moreover, when g is smooth or has bounded subgradients, we show below that a local minimum of F_μ can be used to construct a local minimum of F.

Proposition 1 **(Correspondence of Local Minima)**. *Suppose Assumptions 1 and 2 hold, $0 < \mu < 1/m_\phi$ in (10), and $\bar{z}$ is a local minimizer of F_μ, i.e., there exists $r > 0$ such that $F_\mu(\bar{z}) \leq F_\mu(z)$ for all $z \in \mathbb{R}^n$ satisfying $\|z - \bar{z}\| \leq r$. The following two claims hold.*

1. *Suppose g is Lipschitz differentiable with modulus L_g, then $F(x_{\mu\phi}(\bar{z})) \leq F(x)$ for all $x \in \mathbb{R}^n$ such that $\|x - x_{\mu\phi}(\bar{z})\| \leq \frac{r}{\mu L_g + 1}$.*
2. *Suppose $\sup_{\xi_g \in \partial g(\mathbb{R}^n)} \|\xi_g\| \leq M_{\partial g}$ for some $M_{\partial g} \geq 0$ and $r > 2\mu M_{\partial g}$, then $F(x_{\mu\phi}(\bar{z})) \leq F(x)$ for all $x \in \mathbb{R}^n$ such that $\|x - x_{\mu\phi}(\bar{z})\| \leq r - 2\mu M_{\partial g}$.*

Proof. See EC.2.1

As a consequence of Proposition EC.3 and Proposition 1, the smooth approximation F_μ can serve as a surrogate for the nonsmooth DC function F when solving problem (1). Before making this idea more concrete in Section 4, we consider an important class of properties defined as follows.

Definition 2 (Theorem 3.26 of Rockafellar and Wets (2009)). A function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ is said to be (i) *level-bounded* if its lower level set $\text{lev}_\alpha F := \{x : F(x) \leq \alpha\}$ is bounded (possibly empty) for any $\alpha \in \mathbb{R}$, (ii) *level-coercive* if there exist some $\alpha \in (0, +\infty)$ and $r \in \mathbb{R}$ such that $F \geq \alpha\|\cdot\| + r$, and (iii) *coercive* if for any $\alpha \in (0, +\infty)$, there exists $r \in \mathbb{R}$ such that $F \geq \alpha\|\cdot\| + r$.

Clearly, coercivity implies level-coercivity, which in turn implies level-boundedness. Level-boundedness is important for a descent algorithm to produce bounded iterates, so that at least one limit point exists and we can study the asymptotic behavior of the algorithm. We note that for a general DC function F satisfying Assumption 1, F_μ is not necessarily level-bounded. In the next proposition, we give some sufficient conditions that ensure F_μ is coercive or level-bounded.

Proposition 2. *Suppose Assumptions 1 and 2 hold, and $0 < \mu < 1/m_\phi$.*

1. *If F is level-bounded, and g is Lipschitz up to some constant, i.e., there exist $L > 0$ and $M \geq 0$ such that $\|g(x) - g(z)\| \leq L\|x - z\| + M$ for all $x, z \in \mathbb{R}^n$, then F_μ is level-bounded.*
2. *If F is level-coercive, then F_μ is level-bounded.*
3. *If $\text{dom}\, \phi := \{x : \phi(x) < +\infty\}$ is compact, then F_μ is coercive, and therefore level-bounded.*
4. *If F is level-bounded, and $\text{dom}\, \phi = \mathbb{R}^n$, then F_μ is level-bounded.*

Proof. See EC.2.2. $\square$

4. Unconstrained DC Optimization

In this section, we present two algorithms for solving the unconstrained DC program (1). Although we call problem (1) "unconstrained", under Assumption 1, implicit convex functional constraints can be incorporated into the definition of ϕ using an indicator function, as long as proximal evaluation of the resulting ϕ is available.

Minimizing F is in general challenging due to its nonconvexity and nonsmoothness. In contrast, the DME smooth approximation F_μ provides an attractive surrogate: the Lipschitzness of ∇f is a desirable property for a wide range of first-order methods, among which the GD (4) is probably the most classic one. By (11), each gradient update (4) requires the proximal evaluations of ϕ and g, which can be performed in parallel and is hence a major difference from DCA and pDCA. The convergence of GD (4) is stated in the next theorem. A similar result has been proved in (Themelis et al. 2020, Theorem 7). We provide a detailed proof for self-consistency.

THEOREM 1. *Suppose Assumptions 1 and 2 hold and let* $\epsilon > 0$, $0 < \mu < 1/m_\phi$, *and* $0 < \alpha \le 1/L_{F_\mu}$ *where* $L_{F_\mu} = \frac{2-\mu m_\phi}{\mu - \mu^2 m_\phi}$ *if* $m_\phi > 0$ *and* $L_{F_\mu} = 2\mu^{-1}$ *if* $m_\phi = 0$. *Let* $\{(z^k,\ x_{\mu\phi}(z^k),\ x_{\mu g}(z^k))\}_{k\in\mathbb{N}}$ *be the sequence generated by* (4) *Denote* $\xi^k = \mu^{-1}(x_{\mu g}(z^k)\ - x_{\mu\phi}(z^k))$. *Then for any positive integer* K, *there exists* $0 \le \bar{k} \le K-1$ *such that*

$$\xi^{\bar{k}} \in \partial\phi(x_{\mu\phi}(z^{\bar{k}})) - \partial g(x_{\mu g}(z^{\bar{k}})), \text{ and} \tag{12a}$$

$$\max\{\|\xi^{\bar{k}}\|, \|x_{\mu g}(z^{\bar{k}}) - x_{\mu\phi}(z^{\bar{k}})\|\} \le \max\{1,\mu^{-1}\}\left(\frac{2\mu^2(F_\mu(z^0)-F^*)}{\alpha K}\right)^{1/2}. \tag{12b}$$

Therefore, the GD (4) *finds an* ϵ*-stationary solution in the sense of* (9) *in no more than*

$$\left\lceil \frac{2\mu^2 \max\{1,\mu^{-1}\}^2(F_\mu(z^0)-F^*)}{\alpha\epsilon^2} \right\rceil = \mathcal{O}(\epsilon^{-2}) \tag{13}$$

iterations. Moreover, if F_μ *is level-bounded, then the sequence* $\{z^k\}_{k\in\mathbb{Z}_+}$ *stays bounded; for every limit point* z^*, *it holds* $x_{\mu\phi}(z^*) = x_{\mu g}(z^*)$, *and* $x_{\mu\phi}(z^*)$ *satisfies the exact stationary condition* (8).

Proof. See EC.3.1. □

Next we consider a more structured setting where $\phi = f + h$ admits a composite form. Instead of directly evaluating the proximal mapping of ϕ, which might still be difficult, we can replace the smooth component f with its linearization at the previous iterate. See Algorithm 1.

REMARK 2.

1. The update of x^{k+1} in line 4 takes the form $x^{k+1} = \arg\min_x h(x) + \frac{1}{2\mu}\|x - (z^k - \mu\nabla f(x^k))\|^2$. We first take a gradient (forward) step with respect to the smooth component f, and then take a proximal (backward) step with respect to the nonsmooth component h. A notable difference is that, we move z^k, instead of x^k, along the direction $-\nabla f(x^k)$.

Algorithm 1 : Inexact GD on F_μ

1: **Let** $0 < \mu < 1/L_f$ and $0 < \beta < 2$;

2: **Initialize** $x^0, z^0 \in \mathbb{R}^n$;

3: **for** $k = 0, 1, \cdots$ **do**

4: $\quad x^{k+1} = \underset{x\in\mathbb{R}^n}{\arg\min} \left\{ \langle \nabla f(x^k), x\rangle + h(x) + \frac{1}{2\mu}\|x - z^k\|^2 \right\}$;

5: $\quad x_{\mu g}(z^k) = \underset{x\in\mathbb{R}^n}{\arg\min} \left\{ g(x) + \frac{1}{2\mu}\|x - z^k\|^2 \right\}$;

6: $\quad z^{k+1} = z^k + \beta(x^{k+1} - x_{\mu g}(z^k))$;

7: **end for**

2. Our choice of the proximal center z^k is different from the extrapolation $x^k + \beta(x^k - x^{k-1})$ considered in Wen et al. (2018).

3. We view $\mu^{-1}(x^{k+1} - x_{\mu g}(z^k))$ as an inexact gradient of F_μ at z^k, and take a similar gradient step to update z^{k+1} with some properly chosen step size $\beta\mu$ as in line 6.

We propose the following potential function

$$\mathcal{F}(x, z) := \phi(x) + \frac{1}{2\mu}\|x - z\|^2 - M_{\mu g}(z) = f(x) + h(x) + \frac{1}{2\mu}\|x - z\|^2 - M_{\mu g}(z). \tag{14}$$

We show the sequence $\{\mathcal{F}(x^k, z^k)\}_{k\in\mathbb{N}}$ generated by Algorithm 1 is bounded and non-increasing.

LEMMA 2. *Suppose Assumptions 1 and 2 hold. For all $k \in \mathbb{N}$, $\mathcal{F}(x^k, z^k) \geq F_\mu(z^k) \geq F^*$, and*

$$\mathcal{F}(x^k, z^k) - \mathcal{F}(x^{k+1}, z^{k+1}) \geq c_1\|x^{k+1} - x^k\|^2 + c_2\|z^{k+1} - z^k\|^2, \tag{15}$$

where

$$c_1 = \frac{\mu^{-1} - L_f}{2} > 0, \quad c_2 = \frac{1}{\mu}\left(\frac{1}{\beta} - \frac{1}{2}\right) > 0. \tag{16}$$

Proof. See EC.3.2. □

THEOREM 2. *Suppose Assumptions 1 and 2 hold, and let $\epsilon > 0$. Let $\{(x^{k+1}, x_{\mu g}(z^k), z^{k+1})\}_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 1. Denote $\xi^{k+1} := \nabla f(x^{k+1}) - \nabla f(x^k) - \mu^{-1}(x^{k+1} - x_{\mu g}(z^k))$. Then for any positive integer K, there exists $0 \leq \bar{k} \leq K - 1$ such that*

$$\xi^{\bar{k}+1} \in \partial\phi(x^{\bar{k}+1}) - \partial g(x_{\mu g}(z^{\bar{k}})), \text{ and} \tag{17a}$$

$$\max\{\|\xi^{\bar{k}+1}\|, \|x_{\mu g}(z^{\bar{k}}) - x^{k+1}\|\} \leq \left(L_f + \frac{\mu^{-1}+1}{\beta}\right)\left(\frac{\mathcal{F}(x^0, z^0) - F^*}{\min\{c_1, c_2\}K}\right)^{1/2}. \tag{17b}$$

Therefore, Algorithm 1 finds an ϵ-stationary solution in the sense of (9) *in no more than*

$$\left\lceil \frac{(L_f + (\mu^{-1}+1)/\beta)^2(\mathcal{F}(x^0, z^0) - F^*)}{\min\{c_1, c_2\}\epsilon^2} \right\rceil = \mathcal{O}(\epsilon^{-2}) \tag{18}$$

iterations. Moreover, if F_μ is level-bounded, then the sequence $\{(x^k, z^k)\}_{k\in\mathbb{Z}_+}$ stays bounded; for every limit point (x^*, z^*), it holds $x^* = x_{\mu g}(z^*)$, and x^* satisfies the exact stationary condition (8).

Proof. See EC.3.3. □

We end this section with some additional remarks regarding comparisons with (proximal) DCA-type algorithms. Admittedly, (proximal) DCA is also able to achieve a similar iteration complexity, and usually a subgradient evaluation of g is easier than its proximal evaluation. We note that the DME technique provides a globally smooth approximation for any generic DC function with desirable properties established in Section 3, which a local linearization of g used in DCA does not possess. From a computation point of view, as ∇F_μ can be efficiently evaluated through proximal mappings of both (weakly) convex components, it is possible to further incorporate quasi-Netwon updates (Byrd et al. 1994) or Nesterov-type accelerations (Nesterov 1983) for solving problem (1). In addition, the proximal mappings of h (or ϕ in GD) and g can be computed simultaneously. Moreover, as we will see in the next section, the DME smoothing can be combined with the classical augmented Lagrangian to solve constrained programs (2).

5. Linearly Constrained DC (LCDC) Optimization

In this section we consider the LCDC program (2):

$$\min_{x\in\mathbb{R}^n} F(x)=\phi(x)-g(x) \quad \text{s.t.} \quad Ax=b.$$

It is known that under some mild regularity condition at a local minimizer x (Rockafellar and Wets 2009, Theorem 8.15), there exists $\lambda\in\mathbb{R}^m$ such that

$$0\in\partial F(x)+A^\top\lambda, \text{ and } Ax-b=0. \tag{19}$$

Based on different structures of ϕ, we propose two proximal ALM algorithms that will find an approximate solution to (19). In particular, we present LCDC-ALM in Section 5.1 when ϕ is a smooth (possibly nonconvex) function, and composite LCDC-ALM in Section 5.2 when ϕ is the sum of a smooth (possibly nonconvex) function and a nonsmooth convex function.

5.1. LCDC-ALM

In this subsection, we make the following additional assumptions on the problem data.

Assumption 3.

1. *$\phi=f:\mathbb{R}^n\to\mathbb{R}$ is Lipschitz differentiable with modulus $L_f>0$, i.e., h is the zero function.*
2. *The matrix $A\in\mathbb{R}^{m\times n}$ is nonzero and $b\in\mathbb{R}^m$ belongs to the column space of A.*
3. *There exist finite $\overline{\mu},\overline{\rho}>0$ such that for all $0<\mu<\overline{\mu}$ and $\rho>\overline{\rho}$,*

$$v(\mu,\rho):=\inf_{x,y\in\mathbb{R}^n}\left\{f(x)-g(y)+\frac{1}{2\mu}\|x-y\|^2+\frac{\rho}{2}\|Ax-b\|^2\right\}>-\infty. \tag{20}$$

We justify condition (20) in Lemma 3, which allows g to be any Lipschitz convex function, or problem (2) to be any quadratic program that is regular in some sense.

Lemma 3. *Condition (20) can be satisfied under each of the following cases.*

1. *The function $f-g$ achieves a finite minimal value, and g is L_g-Lipschitz continuous.*

2. *The function f is quadratic (possibly nonconvex), g is L_g-Lipschitz continuous, and $x^\top \nabla^2 f x > 0$ for all $x \in \mathbb{R}^n$ such that $Ax = 0$.*

3. *The function f is quadratic (possibly nonconvex), g is a convex quadratic function, and*

$$\min_{x \in \mathbb{R}^n} \{x^\top \nabla^2 f x \mid Ax = 0, \|x\| = 1\} > \lambda_{\max}(\nabla^2 g).$$

(Note that 2 and 3 coincide when g is affine.)

Proof. See EC.4.1. □

Next we define approximate stationarity for problem (2) as follows.

Definition 3. Given $\epsilon > 0$, we say $x \in \mathbb{R}^n$ is an ϵ-stationary point of (2) under Assumption 3 if there exists $(\xi, y, \lambda) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m$ such that

$$\xi \in \nabla f(x) - \partial g(y) + A^\top \lambda, \text{ and } \max\{\|\xi\|, \|x-y\|, \|Ax-b\|\} \leq \epsilon. \tag{21}$$

Before presenting the algorithm, we first introduce some notation. The classic augmented Lagrangian function is defined as

$$L_\rho(x, \lambda) := f(x) - g(x) + \langle \lambda, Ax + b \rangle + \frac{\rho}{2}\|Ax - b\|^2, \tag{22}$$

for some $\rho > 0$. We also introduce a smoothed proximal augmented Lagrangian function,

$$\psi(x, z, \lambda) := \left(\tilde{f}_\rho(x, \lambda) + \frac{1}{2\mu}\|x - z\|^2 \right) - M_{\mu g}(z), \tag{23}$$

for some $\mu > 0$, where $\tilde{f}_\rho(x, \lambda) := f(x) + \langle \lambda, Ax + b \rangle + \frac{\rho}{2}\|Ax - b\|^2$. Furthermore, the Moreau envelope $M_{\mu \tilde{f}_\rho}$ and proximal mapping $x_{\mu \tilde{f}_\rho}$ of $\tilde{f}_\rho$ are given below,

$$M_{\mu \tilde{f}_\rho}(z, \lambda) = \min_{x \in \mathbb{R}^n} \left\{ \tilde{f}_\rho(x, \lambda) + \frac{1}{2\mu}\|x - z\|^2 \right\}, \text{ and } x_{\mu \tilde{f}_\rho}(z, \lambda) = \arg\min_{x \in \mathbb{R}^n} \left\{ \tilde{f}_\rho(x, \lambda) + \frac{1}{2\mu}\|x - z\|^2 \right\}.$$

The LCDC-ALM is presented in Algorithm 2. To provide some intuition, notice that in analog to the unconstrained case, given some dual variable $\lambda \in \mathbb{R}^m$, the function

$$\mathcal{L}_{\mu\rho}(z, \lambda) := M_{\mu \tilde{f}_\rho}(z, \lambda) - M_{\mu g}(z)$$

is a smooth approximation of the augmented Lagrangian $L_\rho(x, \lambda)$ with gradient given by

$$\nabla_z \mathcal{L}_{\mu\rho}(z, \lambda) = \frac{1}{\mu}(z - x_{\mu \tilde{f}_\rho}(z, \lambda)) - \frac{1}{\mu}(z - x_{\mu g}(z)) = \frac{1}{\mu}(x_{\mu g}(z) - x_{\mu \tilde{f}_\rho}(z, \lambda)), \text{ and}$$
$$\nabla_\lambda \mathcal{L}_{\mu\rho}(z, \lambda) = A x_{\mu \tilde{f}_\rho}(z, \lambda) - b.$$

Algorithm 2 : LCDC-ALM

1: **Let** $0 < \mu < \min\{\overline{\mu}, (L_f)^{-1}\}$, $0 < \beta < 2$, $\rho > \overline{\rho}$;

2: **Initialize** $x^0 \in \mathbb{R}^n$, $z^0 \in \mathbb{R}^n$, $\lambda^0 \in \mathbb{R}^m$;

3: **for** $k = 0, 1, \cdots$ **do**

4:
$$x^{k+1} = \underset{x \in \mathbb{R}^n}{\arg\min} \left\{ \langle \nabla f(x^k), x - x^k \rangle + \langle \lambda^k, Ax - b \rangle + \frac{\rho}{2}\|Ax - b\|^2 + \frac{1}{2\mu}\|x - z^k\|^2 \right\}; \tag{24}$$

5: $x_{\mu g}(z^k) = \arg\min_{x \in \mathbb{R}^n} \left\{ g(x) + \frac{1}{2\mu}\|x - z^k\|^2 \right\}$;

6: $z^{k+1} = z^k + \beta(x^{k+1} - x_{\mu g}(z^k))$;

7: $\lambda^{k+1} = \lambda^k + \rho(Ax^{k+1} - b)$;

8: **end for**

Since the Moreau envelope $M_{\mu \tilde{f}_\rho}$ may still be difficult to evaluate, (24) replaces f with its linearization and thus computes an inexact gradient of $M_{\mu \tilde{f}_\rho}$, which is equivalent to solving the following positive definite linear system

$$\left(\rho A^\top A + \frac{1}{\mu} I \right) x = \frac{1}{\mu} z^k + \rho A^\top b - A^\top \lambda^k - \nabla f(x^k).$$

Consequently, line 6 and line 7 can be regarded as an inexact gradient descent of $\mathcal{L}_{\mu\rho}(z, \lambda)$ in z and gradient ascent of $\mathcal{L}_{\mu\rho}(z, \lambda)$ in λ, respectively. As we will show later, this inexact gradient descent on z is sufficient to ensure convergence of Algorithm 2. Meanwhile, when the Moreau envelope $M_{\mu \tilde{f}_\rho}$ is easy to compute, we can always replace (24) by its exact version, i.e., $x^{k+1} = x_{\mu \tilde{f}_\rho}(z^k, \lambda^k)$, whose analysis is similar and omitted in this paper. We firstly establish some descent properties of ψ in the next lemma.

Lemma 4. *Suppose Assumptions 1 and 3 hold. For all $k \in \mathbb{N}$, we have*

$$\psi(x^k, z^k, \lambda^k) - \psi(x^{k+1}, z^{k+1}, \lambda^{k+1}) \geq \left(\frac{\mu^{-1} - L_f}{2} \right) \|x^{k+1} - x^k\|^2 + \frac{1}{\mu} \left(\frac{1}{\beta} - \frac{1}{2} \right) \|z^{k+1} - z^k\|^2 - \frac{1}{\rho}\|\lambda^{k+1} - \lambda^k\|^2.$$

Proof. See EC.4.2. □

Next we show that, due to the smoothness of f, the difference of dual variables can be effectively bounded by the difference of primal variables. For notation consistency, denote $x^{-1} = x^0$ and $z^{-1} = x^0 + \mu(\nabla f(x^0) + A^\top \lambda^0)$.

Lemma 5. *Suppose Assumptions 1 and 3 hold. For all $k \in \mathbb{N}$,*

$$\|\lambda^{k+1} - \lambda^k\| \leq \frac{1}{\sigma^+_{\min}(A)} \left(\mu^{-1}\|x^{k+1} - x^k\| + L_f\|x^k - x^{k-1}\| + \mu^{-1}\|z^k - z^{k-1}\| \right).$$

Proof. See EC.4.3. □

Lemmas 4 and 5 suggest that the sequence defined by

$$\Psi_k := \psi(x^k, z^k, \lambda^k) + \frac{\nu}{2}\|x^k - x^{k-1}\|^2 + \frac{\nu}{2}\|z^k - z^{k-1}\|^2 \tag{25}$$

is a decreasing sequence for some properly chosen $\nu > 0$. To this end, recall c_1 and c_2 defined in (16), and we further let

$$c_3 = \frac{3\mu^{-2}}{\sigma^+_{\min}(AA^\top)}, \quad c_4 = \frac{3L_f^2}{\sigma^+_{\min}(AA^\top)}. \tag{26}$$

We assume that $\nu > 0$ and $\rho > 0$ are chosen such that

$$\kappa_1 = c_1 - \frac{c_3}{\rho} - \frac{\nu}{2} > 0,\ \kappa_2 = c_2 - \frac{\nu}{2} > 0,\ \kappa_3 = \frac{\nu}{2} - \frac{c_4}{\rho} > 0,\ \kappa_4 = \frac{\nu}{2} - \frac{c_3}{\rho} > 0. \tag{27}$$

This is always possible, for example, by first letting $\nu < 2\min\{c_1, c_2\}$ and then choosing $\rho > \max\left\{\frac{c_3}{c_1-\nu/2}, \frac{2c_3}{\nu}, \frac{2c_4}{\nu}\right\}$.

Lemma 6. *Suppose Assumptions 1 and 3 hold. The following claims hold for LCDC-ALM.*

1. *For all $k \in \mathbb{N}$, we have*

$$\Psi_k - \Psi_{k+1} \geq \kappa_1\|x^{k+1} - x^k\|^2 + \kappa_2\|z^{k+1} - z^k\|^2 + \kappa_3\|x^k - x^{k-1}\|^2 + \kappa_4\|z^k - z^{k-1}\|^2. \tag{28}$$

2. *Ψ_k is bounded from below by $v(\mu,\rho)$ defined in (20) for all $k \in \mathbb{N}$.*
3. *For any positive integer K, there exists an index $0 \leq \bar{k} \leq K-1$ such that*

$$\max\left\{\|x^{\bar{k}+1} - x^{\bar{k}}\|^2, \|z^{\bar{k}+1} - z^{\bar{k}}\|^2, \|x^{\bar{k}} - x^{\bar{k}-1}\|^2, \|z^{\bar{k}} - z^{\bar{k}-1}\|^2\right\} \leq \frac{\Psi_0 - v(\mu,\rho)}{\kappa_{\min}K},$$

where $\kappa_{\min} = \min\{\kappa_1, \kappa_2, \kappa_3, \kappa_4\} > 0$.

Proof. See EC.4.4 . □

Utilizing Lemma 4-Lemma 6, we are ready to present the convergence of LCDC-ALM.

Theorem 3. *Suppose Assumptions 1 and 3 hold. Let $\{(x^{k+1}, z^{k+1}, \lambda^{k+1})\}_{k\in\mathbb{N}}$ be the sequence generated by LCDC-ALM, and denote*

$$\xi^{k+1} := \nabla f(x^{k+1}) - \nabla f(x^k) + \mu^{-1}(x_{\mu g}(z^k) - x^{k+1}).$$

For any positive index K, there exists $0 \leq \bar{k} \leq K-1$ such that $x^{\bar{k}+1}$, together with

$$(\xi^{\bar{k}+1}, x_{\mu g}(z^{\bar{k}}), \lambda^{\bar{k}+1}) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m,$$

is an approximate stationary solution of (2) satisfying

$$\xi^{\bar{k}+1} \in \nabla f(x^{\bar{k}+1}) - \partial g(x_{\mu g}(z^{\bar{k}})) + A^\top \lambda^{k+1},\ \text{and} \tag{29a}$$

$$\max\left\{\|\xi^{\bar{k}+1}\|, \|x_{\mu g}(z^{\bar{k}}) - x^{\bar{k}+1}\|, \|Ax^{\bar{k}+1} - b\|\right\} \leq C\left(\frac{\Psi_0 - v(\mu,\rho)}{\kappa_{\min}K}\right)^{1/2}, \tag{29b}$$

where $v(\mu,\rho)$ is defined in (20)*,* $\kappa_{\min}=\min\{\kappa_1,\kappa_2,\kappa_3,\kappa_4\}>0$*,*

$$\Psi_0 = \psi(x^0,z^0,\lambda^0)+\frac{\nu}{2}\left\|x^0-z^0+\mu(\nabla f(x^0)+A^\top\lambda^0)\right\|^2, \text{ and}$$
$$C = L_f+\frac{\mu^{-1}+1}{\beta}+\frac{2\sqrt{c_3/3}}{\rho}+\frac{\sqrt{c_4/3}}{\rho}.$$

That is, LCDC-ALM finds an ϵ-approximate solution of (2) *in the sense of* (21) *in no more than*

$$\left\lceil\frac{C^2(\Psi_0-v(\mu,\rho))}{\kappa_{\min}\epsilon^2}\right\rceil=\mathcal{O}(\epsilon^{-2}) \tag{30}$$

iterations.

Proof. The optimality conditions of updates of x^{k+1} and $x_{\mu g}(z^k)$ are given as

$$0=\nabla f(x^k)+A^\top\lambda^{k+1}+\mu^{-1}(x^{k+1}-z^k), \text{ and } 0\in\partial g(x_{\mu g}(z^k))+\mu^{-1}(x_{\mu g}(z^k)-z^k).$$

As a result, we have

$$\xi^{k+1} = \nabla f(x^{k+1})-\nabla f(x^k)+\mu^{-1}(x_{\mu g}(z^k)-x^{k+1})\in\nabla f(x^{k+1})-\partial g(x_{\mu g}(z^k))+A^\top\lambda^{k+1}. \tag{31}$$

Since $x^{k+1}-x_{\mu g}(z^k)=\frac{1}{\beta}(z^{k+1}-z^k)$ and $Ax^{k+1}-b=\frac{1}{\rho}(\lambda^{k+1}-\lambda^k)$, we have

$$\begin{aligned}
&\max\left\{\|\xi^{k+1}\|,\ \|x_{\mu g}(z^k)-x^{k+1}\|,\ \|Ax^{k+1}-b\|\right\}\\
\le&\max\Big\{L_f\|x^{k+1}-x^k\|+\frac{1}{\mu\beta}\|z^{k+1}-z^k\|,\ \frac{1}{\beta}\|z^{k+1}-z^k\|,\\
&\qquad\frac{\sqrt{c_3/3}}{\rho}\|x^{k+1}-x^k\|+\frac{\sqrt{c_4/3}}{\rho}\|x^k-x^{k-1}\|+\frac{\sqrt{c_3/3}}{\rho}\|z^k-z^{k-1}\|\Big\}\\
\le& C\max\left\{\|x^{k+1}-x^k\|,\|x^k-x^{k-1}\|,\|z^{k+1}-z^k\|,\|z^k-z^{k-1}\|\right\},
\end{aligned}$$

where the first inequality is due to Lemma 5. Invoking Lemma 6, there exists $0\le\bar{k}\le K-1$ such that

$$\max\left\{\|\xi^{\bar{k}+1}\|,\|x_{\mu g}(z^{\bar{k}})-x^{\bar{k}+1}\|,\|Ax^{\bar{k}+1}-b\|\right\}\le C\left(\frac{\Psi_0-v(\mu,\rho)}{\kappa_{\min}K}\right)^{1/2}.$$

This proves (29b) and (30) and substituting $\bar{k}$ into (31) proves (29a). □

5.2. Composite LCDC-ALM

In this subsection, we consider a more challenging setup of problem (2), where we allow ϕ to have a nonsmooth component h. In particular, we make some additional assumptions on h.

ASSUMPTION 4. *In addition to being proper, closed, and convex, the function $h:\mathbb{R}^n\to\overline{\mathbb{R}}$ is L_h-Lipschitz continuous over its effective domain $\mathcal{H}=\{x\in\mathbb{R}^n \mid h(x)<+\infty\}$, which is nonempty, convex, and compact, i.e., there exists $0<D_{\mathcal{H}}<+\infty$ such that $\|x^1-x^2\|\le D_{\mathcal{H}}$ for all $x^1,x^2\in\mathcal{H}$. Moreover, there exists $\bar{x}\in\operatorname{int}\mathcal{H}$ (interior of $\mathcal{H}$) such that $A\bar{x}=b$;*

By Assumption 4, we can further define positive constants

$$M_{\nabla f} := \max_{x \in \mathcal{H}} \|\nabla f(x)\|, \ \bar{d} := \operatorname{dist}(\bar{x}, \partial\mathcal{H}), \text{ and } M_{\partial g} := \sup_{\xi_g \in \partial g(\mathcal{H})} \|\xi_g\|. \tag{32}$$

The constant $M_{\nabla f}$ is well-defined due to the continuity of ∇f over compact $\mathcal{H}$; $\partial\mathcal{H}$ is the boundary of $\mathcal{H}$, and hence $\bar{d}$ is strictly positive. The constant $M_{\partial g}$ is finite due to (Rockafellar 1970, Theorem 24.7). Next we define an approximate stationary solution under Assumptions 1 and 4.

Definition 4. Given $\epsilon > 0$, we say that $x \in \mathbb{R}^n$ is an ϵ-stationary point of (2) under Assumptions 1 and 4 if there exists $(\xi, y, \lambda) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m$ such that

$$\xi \in \nabla f(x) + \partial h(x) - \partial g(y) + A^\top \lambda, \text{ and } \max\{\|\xi\|, \|x - y\|, \|Ax - b\|\} \le \epsilon. \tag{33}$$

We present the composite LCDC-ALM in Algorithm 3.

Algorithm 3 : Composite LCDC-ALM

1: **Input** $0 < \mu < L_f^{-1}$, $0 < \beta \le 1$, $\rho > 0$, and $\{\epsilon_{k+1}\}_{k \in \mathbb{N}} \subset [0,1]$ with $E := \sum_{k=1}^{\infty} \epsilon_k^2 < +\infty$;
2: **Initialize** $x^0 \in \mathcal{H}$, $z^0 \in \mathcal{H}$, $\lambda^0 \in \operatorname{Im}(A)$;
3: **for** $k = 0, 1, \cdots$ **do**
4: evaluate $\xi_g^k \in \partial g(x^k)$ and find $(x^{k+1}, \zeta^{k+1}) \in \mathcal{H} \times \mathbb{R}^n$ with $\|\zeta^{k+1}\| \le \epsilon_{k+1}$ such that

$$\zeta^{k+1} \in \nabla f(x^k) - \xi_g^k + \partial h(x^{k+1}) + A^\top \lambda^k + \rho A^\top (Ax^{k+1} - b) + \frac{1}{\mu}(x^{k+1} - z^k); \tag{34}$$

5: $z^{k+1} = z^k + \beta(x^{k+1} - z^k)$;
6: $\lambda^{k+1} = \lambda^k + \rho(Ax^{k+1} - b)$;
7: **end for**

Remark 3. Due to the additional nonsmooth function h, we cannot effectively control the dual difference using primal iterates as in Lemma 5. We simply replace $-g$ by its linearization at x^k, and then smooth the linearized AL function: condition (34) says that x^{k+1} is an approximate solution of the following problem:

$$\min_{x \in \mathbb{R}^n} \langle \nabla f(x^k) - \xi_g^k, x - x^k \rangle + h(x) + \langle \lambda^k, Ax - b \rangle + \frac{\rho}{2}\|Ax - b\|^2 + \frac{1}{2\mu}\|x - z^k\|^2. \tag{35}$$

Though not directly using the DME smooth approximation (3), Algorithm 3 still relies on the smoothing property of the Moreau envelope. It can be regarded as a primal-dual version of the linearized pDCA, and is closely related to the proximal ALM proposed in Zhang and Luo (2020a,b) and LiMEAL in Zeng et al. (2021).

Throughout the analysis, we assume that $\{\epsilon_{k+1}\}_{k\in\mathbb{N}} \subset [0,1]$ with $E := \sum_{k=1}^{\infty} \epsilon_k^2 < +\infty$. We utilize the following proximal augmented Lagrangian function:

$$P(x,z,\lambda) := f(x) + h(x) - g(x) + \langle \lambda, Ax-b\rangle + \frac{\rho}{2}\|Ax-b\|^2 + \frac{1}{2\mu}\|x-z\|^2, \tag{36}$$

which will serve as a potential function as shown in the following lemma.

LEMMA 7. *Suppose Assumptions 1 and 4 hold. For all $k \in \mathbb{N}$, we have*

$$\begin{aligned} &P(x^k,z^k,\lambda^k) - P(x^{k+1},z^{k+1},\lambda^{k+1}) \\ &\geq \left(\frac{\mu^{-1}-2L_f}{4}\right)\|x^{k+1}-x^k\|^2 + \frac{1}{2\beta\mu}\|z^{k+1}-z^k\|^2 - \frac{1}{\rho}\|\lambda^{k+1}-\lambda^k\|^2 - \mu\epsilon_{k+1}^2. \end{aligned} \tag{37}$$

Proof. See EC.4.5. □

Next we show that $P(x^1,z^1,\lambda^1)$ is bounded from above by a constant independent of ρ.

LEMMA 8. *Suppose Assumptions 1 and 4 hold. Then $P(x^1,z^1,\lambda^1) \leq \overline{P}$, where*

$$\begin{aligned} \overline{P} :=& \max_{x\in\mathcal{H}}\{f(x)+h(x)-g(x)\} + 3(L_h + M_{\nabla f} + M_{\partial g})D_{\mathcal{H}} + \\ & \frac{L_f + 6\mu^{-1}}{2}D_{\mathcal{H}}^2 + \frac{3L_f^{-1}}{2} + 2\|\lambda^0\|\max_{x\in\mathcal{H}}\|Ax-b\|. \end{aligned} \tag{38}$$

Proof. See EC.4.6. □

Now we show that the sequence $\{\lambda^k\}_{k\in\mathbb{N}}$ stays bounded.

LEMMA 9. *Suppose Assumptions 1 and 4 hold. For all $k \in \mathbb{N}$, we have*

$$\|\lambda^k\| \leq \Lambda := \max\left\{\|\lambda^0\|, \frac{2D_{\mathcal{H}}}{\bar{d}\sigma_{\min}^+(A)}\left(M_{\nabla f} + M_{\partial g} + \frac{D_{\mathcal{H}}}{\mu} + L_h + 1\right)\right\}. \tag{39}$$

Proof. See EC.4.7. □

In the next lemma, we show that $P(x^k,z^k,\lambda^k)$ is bounded from below due to the boundedness of λ^k established in Lemma 9, which further allows us to bound the difference of consecutive primal iterates.

LEMMA 10. *Suppose Assumptions 1 and 4 hold. The following statements hold.*

1. *Recall Λ from Lemma 9. For all $k \in \mathbb{N}$, we have*

$$P(x^k,z^k,\lambda^k) \geq \underline{P} := \min_{x\in\mathbb{R}^n}\{f(x)+h(x)-g(x)\} - \Lambda\max_{x\in\mathcal{H}}\|Ax-b\| > -\infty. \tag{40}$$

2. *Recall $\overline{P}$ in* (38)*, the constant $E = \sum_{k=1}^{\infty}\epsilon_k^2 < +\infty$, and define*

$$\eta := \min\{\frac{1}{4}(\mu^{-1}-2L_f), (2\mu\beta)^{-1}\}. \tag{41}$$

For any positive $K \in \mathbb{N}$, there exists $1 \leq \bar{k} \leq K$ such that

$$\max\left\{\|x^{\bar{k}+1}-x^{\bar{k}}\|^2, \|z^{\bar{k}+1}-z^{\bar{k}}\|^2\right\} \leq \frac{\overline{P}-\underline{P}+\mu E}{\eta K} + \frac{8\Lambda^2}{\eta\rho}. \tag{42}$$

Proof. See EC.4.8. □

Now we are ready to present the convergence of composite LCDC-ALM.

THEOREM 4. *Suppose Assumptions 1 and 4 hold. Let* $\{(x^{k+1}, z^{k+1}, \lambda^{k+1})\}_{k\in\mathbb{N}}$ *be the sequence generated by composite LCDC-ALM, and define* $\xi^{k+1} := \zeta^{k+1} + \nabla f(x^{k+1}) - \nabla f(x^k) + \mu^{-1}(z^k - x^{k+1})$. *For any positive* $K \in \mathbb{N}$, *there exists* $1 \le \bar{k} \le K$ *such that* $x^{\bar{k}+1}$, *together with* $(\xi^{\bar{k}+1}, x^{\bar{k}}, \lambda^{\bar{k}+1}) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m$, *is an approximate stationary solution of problem* (2) *satisfying*

$$\xi^{\bar{k}+1} \in \nabla f(x^{\bar{k}+1}) + \partial h(x^{\bar{k}+1}) - \partial g(x^{\bar{k}}) + A^\top \lambda^{\bar{k}+1}, \tag{43a}$$

$$\max\left\{\|\xi^{\bar{k}+1}\|, \|x^{\bar{k}+1} - x^{\bar{k}}\|\right\} \le \gamma \left(\frac{\overline{P} - \underline{P} + \mu E}{\eta K} + \frac{8\Lambda^2}{\eta\rho}\right)^{1/2} + \epsilon_{\bar{k}+1}, \tag{43b}$$

$$\|Ax^{\bar{k}+1} - b\| \le \frac{2\Lambda}{\rho}, \tag{43c}$$

where $\overline{P}$ *is given in* (38), $\underline{P}$ *is defined in* (40), Λ *is defined in* (39), $E = \sum_{k=1}^{\infty} \epsilon_k^2$, $\gamma = L_f + 1/(\mu\beta) + 1$, *and* $\eta = \min\{(\mu^{-1} - 2L_f)/4, 1/(2\mu\beta)\}$. *In other words, further suppose* $\epsilon > 0$ *such that* $0 \le \epsilon_k \le \epsilon/2$ *for all positive integer* k. *If we choose*

$$\rho \ge \max\left\{\frac{2\Lambda}{\epsilon}, \frac{64\Lambda^2\gamma^2}{\eta\epsilon^2}\right\}, \tag{44}$$

then composite LCDC-ALM finds an ϵ*-approximate solution of* (2) *in the sense of* (33) *in no more than* $K + 1$ *iterations, where*

$$K \le \left\lceil \frac{8\gamma^2(\overline{P} - \underline{P} + \mu E)}{\eta\epsilon^2} \right\rceil = \mathcal{O}(\epsilon^{-2}) \tag{45}$$

iterations.

Proof. Let $\bar{k}$ be the index given in Lemma 10. The optimality condition of $x^{\bar{k}+1}$ gives (43a) immediately. It holds that $\max\left\{\|\xi^{\bar{k}+1}\|, \|x^{\bar{k}+1} - x^{\bar{k}}\|\right\}$ is bounded from above by

$$\begin{aligned}
&\max\left\{\|\zeta^{\bar{k}+1}\| + \|\nabla f(x^{\bar{k}+1}) - \nabla f(x^{\bar{k}})\| + \|\mu^{-1}(z^{\bar{k}} - x^{\bar{k}+1})\|, \|x^{\bar{k}+1} - x^{\bar{k}}\|\right\} \\
&\le \left(L_f + \frac{1}{\mu\beta} + 1\right) \max\left\{\|x^{\bar{k}+1} - x^{\bar{k}}\|, \|z^{\bar{k}+1} - z^{\bar{k}}\|\right\} + \epsilon_{\bar{k}+1} \le \gamma \left(\frac{\overline{P} - \underline{P} + \mu E}{\eta K} + \frac{8\Lambda^2}{\eta\rho}\right)^{1/2} + \epsilon_{\bar{k}+1};
\end{aligned}$$

in addition,

$$\|Ax^{\bar{k}+1} - b\| = \frac{\|\lambda^{\bar{k}+1} - \lambda^{\bar{k}}\|}{\rho} \le \frac{\|\lambda^{\bar{k}+1}\| + \|\lambda^{\bar{k}}\|}{\rho} \le \frac{2\Lambda}{\rho}.$$

This proves (43). Finally recall that $\epsilon_{\bar{k}+1} \le \epsilon/2$. It is straightforward to verify that the claimed lower bound of ρ in (44) and upper bound of K in (45) together ensure $\|Ax^{\bar{k}+1} - b\| \le \epsilon$ and

$$\max\left\{\|\xi^{\bar{k}+1}\|, \|x^{\bar{k}+1} - x^{\bar{k}}\|\right\} \le \gamma \left(\frac{\epsilon^2}{8\gamma^2} + \frac{\epsilon^2}{8\gamma^2}\right)^{1/2} + \frac{\epsilon}{2} = \epsilon.$$

This completes the proof. □

Before ending this section, we briefly discuss the *first-order* iteration complexity of composite LCDC-ALM. In view of (35), the major computational task in each iteration of composite LCDC-ALM is to solve a strongly convex optimization problem of the form $\min_{x\in\mathbb{R}^n}\psi(x)+h(x)$, where, for a fixed index k, $\psi(x)=\langle f(x^k)-\xi_g^k, x-x^k\rangle+\langle\lambda^k, Ax-b\rangle+\frac{\varrho}{2}\|Ax-b\|^2+\frac{1}{2\mu}\|x-x^k\|^2$ for $x\in\mathbb{R}^n$. It is easy to verify that ψ is m_ψ-strongly convex and has a L_ψ-Lipschitz gradient, where $m_\psi=\mu^{-1}$ and $L_\psi=\rho\|A\|^2+\mu^{-1}$. We take the accelerated proximal gradient (APG) method adopted by Li et al. (2021) as an example of an optimal first-order method for solving this strongly convex program. It finds $(x^{k+1},\zeta^{k+1})\in\mathcal{H}\times\mathbb{R}^n$ that satisfies $\|\zeta^{k+1}\|\le\epsilon_{k+1}$ and (34) in no more than

$$T_{k+1}:=\left\lceil\sqrt{\frac{L_\psi}{m_\psi}}\log\frac{64L_\psi^3D_{\mathcal{H}}^2}{\epsilon_{k+1}^2\mu}+1\right\rceil$$

iterations (Li et al. 2021, Lemma 1). We aim to bound the summation $\sum_{k=0}^K T_{k+1}$, where $K=\mathcal{O}(\epsilon^{-2})$ by (45). Now let us suppose $\rho\ge\mu^{-1}/\|A\|^2$ so that $L_\phi\le 2\rho\|A\|^2$, and set $\epsilon_k=\epsilon/(2k)$ for all positive integer k and some $\epsilon\in(0,2]$ (we consider μ as a constant independent of ϵ). We then have

$$\sqrt{L_\psi/m_\psi}=\sqrt{L_\psi\mu}\le\sqrt{2\rho\|A\|^2L_f^{-1}},\quad \frac{64L_\psi^3D_{\mathcal{H}}^2}{\epsilon_{k+1}^2\mu}=\frac{256L_\psi^3D_{\mathcal{H}}^2(k+1)^2}{\epsilon^2\mu}\le\frac{2048D_{\mathcal{H}}^2\|A\|^8\rho^4(k+1)^2}{\epsilon^2}.$$

By (44), it suffices to choose $\rho=\mathcal{O}(\epsilon^{-2})$; by the above, for $k\le K=\mathcal{O}(\epsilon^{-2})$, we can bound T_{k+1} by

$$T_{k+1}\le\left\lceil\sqrt{2\mathcal{O}(\epsilon^{-2})\|A\|^2L_f^{-1}}\log\left(2048D_{\mathcal{H}}^2\|A\|^8\mathcal{O}(\epsilon^{-14})\right)+1\right\rceil=\tilde{\mathcal{O}}(\epsilon^{-1}),$$

where the $\tilde{\mathcal{O}}$ notation hides logarithmic dependency on ϵ. Consequently, the overall first-order iteration complexity of composite LCDC-ALM is given by

$$\sum_{k=0}^K T_{k+1}\le(K+1)\tilde{\mathcal{O}}(\epsilon^{-1})=\tilde{\mathcal{O}}(\epsilon^{-3}).$$

We note that this complexity upper bound matches existing results of ALM-based algorithms (Melo et al. 2020, Melo and Monteiro 2020) for affine-constrained weakly convex minimization. An interesting future direction is to explore the possibility of reducing the iteration complexity to $\tilde{\mathcal{O}}(\epsilon^{-2.5})$, which has been achieved when the concave component $-g$ is absent (Li and Xu 2021).

6. Numerical Experiments

In this section, we present experiments to demonstrate the performance of proposed algorithms.

6.1. Unconstrained DC Regularized Least Square Problem

In this subsection, we consider the ℓ_{1-2} regularized least squares problem:

$$\min_{x\in\mathbb{R}^n}F(x)=\frac{1}{2}\|Cx-d\|^2+\varrho\|x\|_1-\varrho\|x\|, \tag{46}$$

and compare the Inexact GD method (Algorithm 1) with the pDCAe proposed in Wen et al. (2018). We generate data as described in (Wen et al. 2018, Section 5): first create $C \in \mathbb{R}^{m\times n}$ with standard Gaussian entries, and then normalize its columns to unit length; generate $\hat{x} \in \mathbb{R}^n$ such that $\|\hat{x}\|_0 = s$ with nonzero Gaussian entries, and finally set $d = C\hat{x} + 0.01\xi$, where $\xi \in \mathbb{R}^m$ has standard Gaussian entries. In our experiments, we choose $\beta = 1$ and $\mu = 1/\|A\|^2$. For pDCAe, the extrapolation parameters are chosen according to Section 3 of their paper, which are also popular choices used in FISTA (Beck and Teboulle 2009). Both algorithms are provided with the same initial point and will terminate if $\|x^{k+1} - y^k\|/\max\{1, \|x^{k+1}\|\} \le 10^{-5}$, where $y^k = x_{\mu g}(z^k)$ in Inexact GD and $y^k = x^k$ in pDCAe.

Table 2 Comparison of Inexact GD method (Algorithm 1) and pDCAe Wen et al. (2018)

		Avg. Iteration		Avg. Time (s)		Avg. Time/Iter (10^{-2}s)	
i	ϱ	Inexact GD	pDCAe	Inexact GD	pDCAe	Inexact GD	pDCAe
1	1	**124**	1920	**2.00**	4.29	1.60	**0.22**
	0.1	**174**	2988	**2.69**	6.34	1.54	**0.21**
	0.01	**1079**	1541	16.81	**3.27**	1.56	**0.21**
2	1	**107**	2273	**7.90**	22.28	7.35	**0.98**
	0.1	**194**	3005	**14.06**	29.35	7.26	**0.98**
	0.01	**1077**	1572	75.56	**15.38**	7.02	**0.98**
3	1	**104**	2429	**15.44**	60.19	14.79	**2.48**
	0.1	**196**	3005	**28.64**	72.32	14.61	**2.41**
	0.01	**1052**	1467	159.98	**37.11**	15.21	**2.53**

For each $(m,n,s) = (720i, 2560i, 80i)$ where $i \in [3]$ and $\varrho \in \{1.0, 0.1, 0.01\}$, we generate five instances and report the average iteration number and wall clock time in Table 2. Since both algorithms terminate with the same objective value in most cases, the average objective is the same up to four decimal places and therefore omitted from comparison. Inexact GD requires fewer iterations than pDCAe, and terminates faster in wall clock time when $\varrho \in \{1.0, 0.1\}$. We give two possible explanations for the slowness of Inexact GD when $\varrho = 0.01$: 1) since in general a proximal evaluation is more expensive than a subgradient evaluation, the per-iteration time of Inexact GD can be higher than pDCAe, while a proper parallel implement might be the remedy, and 2) when a smaller ϱ is used, the concave component $-\|x\|$ in problem (46) tends to vanish, and pDCAe behaves more like FISTA applied to convex composite optimization problems, whose efficiency is well-recognized.

6.2. Constrained DC Regularized Least Square Problem

In this subsection, we consider the DC regularized least squares problem with affine constraints:

$$\min_{x\in\mathbb{R}^n} F(x) = \frac{1}{2}\|Cx - d\|^2 + \delta_{\{x:\|x\|_1 \le M\}}(x) - \varrho\|x\| \quad \text{s.t.} \quad Ax = b, \tag{47}$$

where we explicitly require $\|x\|_1 \le M$ instead of penalizing $\|x\|_1$ as in (46). We generate $C \in \mathbb{R}^{m\times n}$ and $d \in \mathbb{R}^m$ the same way as in the previous subsection, $A \in \mathbb{R}^{m\times n}$ with standard Gaussian entries, and $b = A\tilde{x}$, where each component of $\tilde{x}$ is uniformly sampled from $[-M/(2n), M/(2n)]$. We compare the composite LCDC-ALM (Algorithm 3) with GD (4), DCA, and pDCA (5) according to the following problem decomposition: in GD, DCA, and pDCA, let $\phi(x) = \frac{1}{2}\|Cx - d\|^2 + \delta_{\{x: Ax=b, \|x\|_1 \le M\}}(x)$; in composite LCDC-ALM, let $f(x) = \frac{1}{2}\|Cx - d\|^2$ and $h(x) = \delta_{\{x:\|x\|_1 \le M\}}(x)$; in both algorithms, we set $g(x) = \varrho\|x\|$.

The infeasibility of problem (47) is measured by $\|Ax^{k+1} - b\|$ in composite LCDC-ALM, while the constraints $Ax = b$ are always satisfied in the GD, DCA, and pDCA as part of ϕ. Therefore we compare the four algorithms as follows: we first run composite LCDC-ALM until $\|Ax^k - b\| \le 10^{-5}$ and $|F(x^k) - F(x^{k-1})|/|F(x^k)| \le 10^{-3}$; then we execute GD, DCA, and pDCA for the same number of iterations. All nonlinear subproblems are solved by IPOPT with linear solver MA57. In our experiments, we use $\varrho = 1.0$, $M = 2.0$ for problem (47), and set $\beta = 0.1$, $\mu = 1/L_f$ (this is also the proximal coefficient used in pDCA). For each $(m, n, s) = (50i, 200i, 10i)$ where $i \in [5]$, we generate five instances and report the average iteration, objective value, and running time in Table 3. The objective function is evaluated at the last iterate for all algorithms. We observed that all algorithms converge to solutions with similar quality, while DCA and pDCA seem to achieve slightly better objective values. Meanwhile, the composite LCDC-ALM has a clear advantage in solution time.

Table 3 Comparison of Composite LCDC-ALM (ALM), GD, DCA, and pDCA

		Avg. Objective				Avg. Time (s)			
i	Iteration	ALM	GD	DCA	pDCA	ALM	GD	DCA	pDCA
1	104	2.7564	2.7574	2.7517	**2.7490**	**3.30**	4.55	4.89	4.75
2	113	8.8506	8.8514	**8.8494**	8.8500	**16.85**	22.25	24.28	22.61
3	115	10.4611	10.4612	10.4608	**10.4606**	**42.25**	70.28	80.10	71.95
4	94	17.4131	17.4134	17.4131	**17.4127**	**73.71**	114.13	127.00	111.19
5	89	24.1486	24.1489	**24.1483**	24.1483	**120.49**	170.40	193.03	176.45

6.3. Linearly Constrained Nonconvex Quadratic Program

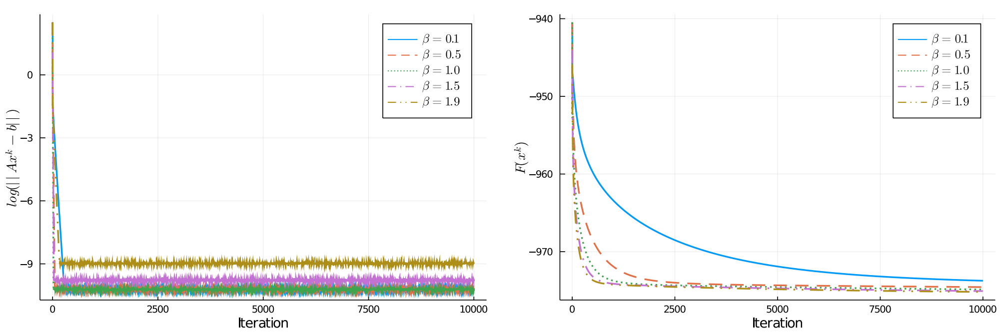
In this subsection, we consider the linearly constrained nonconvex quadratic program:

$$\min_{x\in\mathbb{R}^n} F(x) = f(x) - g(x) \quad \text{s.t.} \quad Ax = b, \tag{48}$$

where $f(x) = \frac{1}{2}x^\top Qx + q^\top x$ and $g(x) = \frac{1}{2}x^\top Gx$ for some $Q, G \in \mathbb{S}^{n\times n}_+$ and $q \in \mathbb{R}^n$. In particular, we would like to investigate how the choice of $\beta \in (0, 2)$ affects the convergence of LCDC-ALM (Algorithm 2), and then compare its performance with the proximal ALM proposed in Zhang and Luo (2020a). We first generate $A \in \mathbb{R}^{m\times n}$, $q \in \mathbb{R}^n$, and $\hat{x} \in \mathbb{R}^n$ with standard Gaussian entries, and set $b = A\hat{x}$. Suppose for simplicity $\text{rank}(A) = m$. Let $\{v_1, \cdots, v_{n-m}\}$ and $\{u_1, \cdots, u_m\}$ be an orthonormal basis of the null space of A and the column space of $A^\top$, respectively; further denote $V_+ = \{v_1, \cdots, v_m, u_1, \cdots, u_{\lfloor m/2 \rfloor}\}$ and $V_- = \{u_{\lfloor m/2 \rfloor + 1}, \cdots, u_m\}$. Then we let $Q = \sum_{v \in V_+} a(v) v v^\top$ and $G = \sum_{u \in V_-} b(u) u u^\top$, where each $a(v)$ is uniformly sampled from $[0, 10]$, and each $b(u)$ is uniformly sampled from $[0, 50]$. The construction ensures $\inf_x \{F(x) : Ax = b\} > -\infty$ and Assumption 3 is satisfied. In our experiments, we use $m = 200$ and $n = 500$; the spectrum of the generated matrix $Q - G$ lies between $[-49.74, 9.97]$.

Recall the definition of c_1, c_2 in (16) and c_3, c_4 in (26). For LCDC-ALM, we set $\nu = \min\{c_1, c_2\}$ and $\rho = 10 \max\{c_3/(c_1 - \nu/2), 2c_3/\nu, 2c_4/\nu\}$. We perform LCDC-ALM with different values of β, and plot the infeasibility and objective value as functions of iteration in Fig. 2. The infeasibility converges with similar speed, while a mild slowdown is observed as β moves towards its upper or lower bound. In contrast, the objective converges faster when a larger β is used.

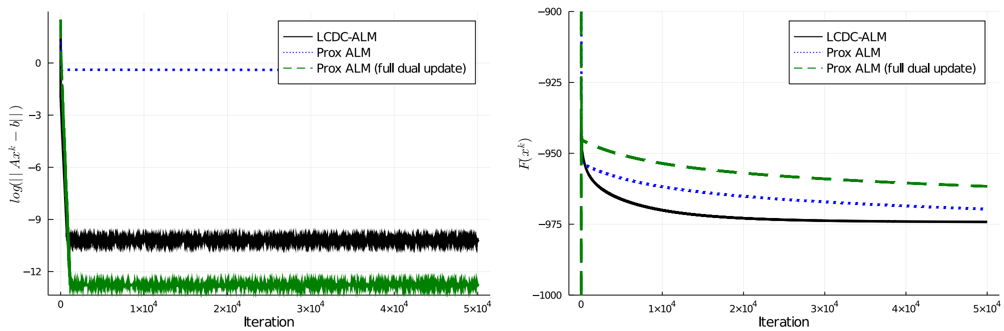
Figure 2 **Infeasibility and Objective Trajectory of LCDC-ALM (Algorithm 2)**



(a) Infeasibility (b) Objective Value

Next we compare LCDC-ALM with the proximal ALM proposed in Zhang and Luo (2020a) in Figure 3. For proximal ALM, all parameters are chosen according to (Zhang and Luo 2020a, Lemma 3.1): in particular, we choose $\beta = 1/30$ (also for LCDC-ALM), and $\alpha = L_f / [(\|A^\top A\| + 4)\|A^\top A\|]$, where α is the dual step size, i.e., $\lambda^{k+1} = \lambda^k + \alpha(Ax^{k+1} - b)$. For the generated instance, we realize the dual step size α is extremely small, and this causes proximal ALM to converge very slowly in both infeasibility and objective. Therefore we also apply a full dual update as $\lambda^{k+1} = \lambda^k + \rho(Ax^{k+1} - b)$, where $\rho = \|Q - G\| = 49.74$. The infeasibility level of the proximal ALM with full dual update then decreases even slightly faster than LCDC-ALM; however, such behavior of proximal ALM is not explained by the analysis in Zhang and Luo (2020a) and deserves further investigation. Meanwhile, LCDC-ALM achieves a better objective value than proximal ALM.

Figure 3 Comparison of LCDC-ALM (Algorithm 2) and Proximal ALM (Zhang and Luo 2020a)



(a) Infeasibility (b) Objective Value

7. Concluding Remarks

In this paper, we study the minimization of a DC function F in the form of (1) or (2). Our algorithmic developments are based on the the difference-of-Moreau-envelopes smooth approximation F_μ introduced in (3). We first study some important properties of F_μ, such as Lipschitz smoothness, the correspondence of stationary points, local, and global minima with F, and the coercivity and level-boundedness of F_μ. Then, we propose algorithms based on the smoothed objective function F_μ. In particular, we show that applying gradient-based updates on F_μ converges to a stationary solution of F with rate $\mathcal{O}(K^{-1/2})$. Since a local or global solution of F_μ can be used to construct a counterpart of F, future directions include exploiting high-order information or sharpness/error bound conditions of F_μ.

When the minimization of F is explicitly constrained in an affine subspace, we apply the smoothing technique to the classic augmented Lagrangian function and propose two ALM-based algorithms, LCDC-ALM and composite LCDC-ALM, that will find an ϵ-stationary solution in $\mathcal{O}(\epsilon^{-2})$ iterations. We note that due to the more challenging DC setting, the subproblem oracle in composite LCDC-ALM in general cannot be replaced by a single projected gradient step as in Zhang and Luo (2020a,b). We are interested in simplifying ALM subproblems and extending the smoothing idea to handle linear inequality or even DC constraints in future works.

References

Ahmadi AA, Hall G (2018) DC decomposition of nonconvex polynomials with algebraic techniques. *Mathematical Programming* 169(1):69–94.

Alvarado A, Scutari G, Pang JS (2014) A new decomposition method for multiuser DC-programming and its applications. *IEEE Transactions on Signal Processing* 62(11):2984–2998.

An NT, Nam NM (2017) Convergence analysis of a proximal point algorithm for minimizing differences of functions. *Optimization* 66(1):129–147.

Artacho FJA, Fleming RM, Vuong PT (2018) Accelerating the DC algorithm for smooth functions. *Mathematical Programming* 169(1):95–118.

Aybat NS, Iyengar G (2013) An augmented Lagrangian method for conic convex programming. arXiv:1302.6322.

Banert S, Boț RI (2019) A general double-proximal gradient algorithm for DC programming. *Mathematical Programming* 178(1-2):301–326.

Beck A (2017) *First-order Methods in Optimization* (SIAM).

Beck A, Teboulle M (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences* 2(1):183–202.

Bertsekas DP (2014) *Constrained Optimization and Lagrange Multiplier Methods* (Academic press).

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

Byrd RH, Nocedal J, Schnabel RB (1994) Representations of quasi-newton matrices and their use in limited memory methods. *Mathematical Programming* 63(1):129–156.

Cao S, Huo X, Pang JS (2018) A unifying framework of high-dimensional sparse estimation with difference-of-convex (DC) regularizations. arXiv:1812.07130.

Chen C, Mangasarian OL (1996) A class of smoothing functions for nonlinear and mixed complementarity problems. *Computational Optimization and Applications* 5(2):97–138.

Chen C, Zhang J, Shen L, Zhao P, Luo Z (2021) Communication efficient primal-dual algorithm for nonconvex nonsmooth distributed optimization. *International Conference on Artificial Intelligence and Statistics*, 1594–1602 (PMLR).

Chen X (2012) Smoothing methods for nonsmooth, nonconvex minimization. *Mathematical programming* 134(1):71–99.

Davis D, Drusvyatskiy D (2018) Stochastic subgradient method converges at the rate $\mathcal{O}(k^{-1/4})$ on weakly convex functions. arXiv:1802.02988.

de Oliveira W (2019) Proximal bundle methods for nonsmooth DC programming. *Journal of Global Optimization* 75(2):523–563.

de Oliveira W (2020) Sequential difference-of-convex programming. *Journal of Optimization Theory and Applications* 186(3):936–959.

Drusvyatskiy D (2017) The proximal point method revisited. arXiv:1712.06038.

Ellaia R (1984) *Contribution à l'analyse et l'optimisation de différence de fonctions convexes*. Ph.D. thesis, Université Paul Sabatier.

Hajinezhad D, Hong M (2019) Perturbed proximal primal–dual algorithm for nonconvex nonsmooth optimization. *Mathematical Programming* 176(1-2):207–245.

Hestenes MR (1969) Multiplier and gradient methods. *Journal of Optimization Theory and Applications* 4(5):303–320.

Hiriart-Urruty JB (1985) Generalized differentiability/duality and optimization for problems dealing with differences of convex functions. *Convexity and duality in optimization*, 37–70 (Springer).

Hiriart-Urruty JB (1991) How to regularize a difference of convex functions. *Journal of Mathematical Analysis and Applications* 162(1):196–209.

Hong M (2016) Decomposing linearly constrained nonconvex problems by a proximal primal dual approach: algorithms, convergence, and applications. arXiv:1604.00543.

Kurdyka K (1998) On gradients of functions definable in o-minimal structures. *Annales de l'institut Fourier*, volume 48, 769–783.

Lan G, Monteiro R (2016) Iteration-complexity of first-order augmented Lagrangian methods for convex programming. *Mathematical Programming* 155(1):511–547.

Le Thi HA, Dinh TP, Van Ngai H (2012) Exact penalty and error bounds in DC programming. *Journal of Global Optimization* 52(3):509–535.

Le Thi Hoai A, Tao PD (1997) Solving a class of linearly constrained indefinite quadratic problems by DC algorithms. *Journal of Global Optimization* 11(3):253–285.

Li F, Qu Z (2019) An inexact proximal augmented Lagrangian framework with arbitrary linearly convergent inner solver for composite convex optimization. arXiv:1909.09582.

Li Z, Chen PY, Liu S, Lu S, Xu Y (2021) Rate-improved inexact augmented lagrangian method for constrained nonconvex optimization. *International Conference on Artificial Intelligence and Statistics*, 2170–2178 (PMLR).

Li Z, Xu Y (2021) Augmented Lagrangian based first-order methods for convex-constrained programs with weakly-convex objective. arXiv:2003.08880v2.

Liu Yf, Liu X, Ma S (2019) On the nonergodic convergence rate of an inexact augmented Lagrangian framework for composite convex programming. *Mathematics of Operations Research* 44(2):632–650.

Lojasiewicz S (1963) Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles* 117:87–89.

Lu Z (2014) Iterative reweighted minimization methods for ℓ_p regularized unconstrained nonlinear programming. *Mathematical Programming* 147(1):277–307.

Lu Z, Zhou Z (2018) Iteration-complexity of first-order augmented Lagrangian methods for convex conic programming. arXiv:1803.09941.

Lu Z, Zhou Z (2019) Nonmonotone enhanced proximal DC algorithms for a class of structured nonsmooth DC programming. *SIAM Journal on Optimization* 29(4):2725–2752.

Lu Z, Zhou Z, Sun Z (2019) Enhanced proximal DC algorithms with extrapolation for a class of structured nonsmooth DC minimization. *Mathematical Programming* 176(1-2):369–401.

Melo JG, Monteiro RDC (2020) Iteration-complexity of an inner accelerated inexact proximal augmented Lagrangian method based on the classical Lagrangian function and a full lagrange multiplier update. arXiv:2008.00562.

Melo JG, Monteiro RDC, Wang H (2020) Iteration-complexity of an inexact proximal accelerated augmented Lagrangian method for solving linearly constrained smooth nonconvex composite optimization problems. arXiv:2006.08048.

Melzer D (1986) On the expressibility of piecewise-linear continuous functions as the difference of two piecewise-linear convex functions. *Quasidifferential Calculus*, 118–134 (Springer).

Nesterov YE (1983) A method for solving the convex programming problem with convergence rate $\mathcal{O}(1/k^2)$). *Dokl. akad. nauk Sssr* 269:543–547.

Pang JS, Razaviyayn M, Alvarado A (2017) Computing b-stationary points of nonsmooth DC programs. *Mathematics of Operations Research* 42(1):95–118.

Powell MJ (1969) A method for nonlinear constraints in minimization problems. *Optimization, R. Fletcher, ed.* 283–298.

Rockafellar RT (1970) *Convex Analysis*. Number 28 (Princeton university press).

Rockafellar RT (1973) The multiplier method of Hestenes and Powell applied to convex programming. *Journal of Optimization Theory and Applications* 12(6):555–562.

Rockafellar RT (1976) Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of Operations Research* 1(2):97–116.

Rockafellar RT, Wets RJB (2009) *Variational Analysis*, volume 317 (Springer Science & Business Media).

Sahin MF, Eftekhari A, Alacaoglu A, Latorre F, Cevher V (2019) An inexact augmented Lagrangian framework for nonconvex optimization with nonlinear constraints. *Advances in Neural Information Processing Systems*, 13943–13955.

Souza JCO, Oliveira PR, Soubeyran A (2016) Global convergence of a proximal linearized algorithm for difference of convex functions. *Optimization Letters* 10(7):1529–1539.

Sun Wy, Sampaio RJ, Candido M (2003) Proximal point algorithm for minimization of DC function. *Journal of Computational Mathematics* 451–462.

Tao M, Dong H (2018) On the linear convergence of difference-of-convex algorithms for nonsmooth DC programming. Optimization Online.

Tao PD, An LTH (1997) Convex analysis approach to DC programming: theory, algorithms and applications. *Acta Mathematica Vietnamica* 22(1):289–355.

Tao PD, An LTH (1998) A DC optimization algorithm for solving the trust-region subproblem. *SIAM Journal on Optimization* 8(2):476–505.

Tao PD, et al. (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(1-4):23–46.

Themelis A, Hermans B, Patrinos P (2020) A new envelope function for nonsmooth dc optimization. *2020 59th IEEE Conference on Decision and Control (CDC)*, 4697–4702 (IEEE).

Tuy H (2010) *Convex Analysis and Global Optimization* (Springer), second edition.

Wen B, Chen X, Pong TK (2018) A proximal difference-of-convex algorithm with extrapolation. *Computational optimization and applications* 69(2):297–324.

Xu Y (2021) Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming. *Mathematical Programming* 185:199–244.

Yin P, Lou Y, He Q, Xin J (2015) Minimization of 1-2 for compressed sensing. *SIAM Journal on Scientific Computing* 37(1):A536–A563.

Zeng J, Yin W, Zhou DX (2021) Moreau envelope augmented Lagrangian method for nonconvex optimization with linear constraints. *arXiv preprint arXiv:2101.08519* .

Zhang J, Luo Z (2020a) A global dual error bound and its application to the analysis of linearly constrained nonconvex optimization. arXiv:2006.16440.

Zhang J, Luo ZQ (2020b) A proximal alternating direction method of multiplier for linearly constrained nonconvex minimization. *SIAM Journal on Optimization* 30(3):2272–2302.

Zhang J, Xiao P, Sun R, Luo Z (2020) A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. *Advances in Neural Information Processing Systems* 33:7377–7389.

Electronic Companion

EC.1. Preliminaries of Difference-of-Moreau-Envelopes Smoothing

EC.1.1. Moreau Envelope

We summarize some known properties of Moreau envelope in the next proposition.

PROPOSITION EC.1. *Suppose Assumption 1 holds and* $0 < \mu < 1/m_\phi$ *in* (10). *Then the following claims hold.*

1. $x_{\mu\phi}$ *is Lipschitz continuous with modulus* $\frac{1}{1-\mu m_\phi}$.
2. $M_{\mu\phi}$ *is differentiable with gradient* $\nabla M_{\mu\phi}(z) = \mu^{-1}(z - x_{\mu\phi}(z))$.
3. $\nabla M_{\mu\phi}$ *is Lipschitz continuous with modulus* $\frac{2-\mu m_\phi}{\mu-\mu^2 m_\phi}$.

Proof. The first two claims are well-known, see, e.g., (Zhang and Luo 2020b, Lemma 3.5) and (Rockafellar and Wets 2009, Proposition 13.37), combining which proves the last one. □

Proposition EC.1 suggests that $M_{\mu\phi}$ forms a smooth approximation of the possibly nonconvex nonsmooth function ϕ. Similarly, the Moreau envelope and proximal mapping of g are given by $M_{\mu g}$ and $x_{\mu g}$, respectively. Since g is convex, it is known that $x_{\mu g}$ is 1-Lipschitz and $M_{\mu g}$ is differentiable, whose gradient $\nabla M_{\mu g}(z) = (z - x_{\mu g}(z))/\mu$ is $1/\mu$-Lipschitz (Beck 2017, Theorem 6.60).

EC.1.2. Lipschitz Differentiability of F_μ

PROPOSITION EC.2. *Suppose Assumption 1 holds and* $0 < \mu < 1/m_\phi$ *in* (10). F_μ *is differentiable, and* $\nabla F_\mu(z) = \mu^{-1}(x_{\mu g}(z) - x_{\mu\phi}(z))$ *is Lipschitz continuous with modulus* $L_{F_\mu} = \frac{2-\mu m_\phi}{\mu-\mu^2 m_\phi}$.

Proof. By the Lipschitz differentiability of the Moreau envelope shown in Proposition EC.1 and the definition of F_μ, we know that F_μ is differentiable, and $\nabla F_\mu(z) = \nabla M_{\mu\phi}(z) - \nabla M_{\mu g}(z) = \mu^{-1}(x_{\mu g}(z) - x_{\mu\phi}(z))$. Since $x_{\mu g}$ and $x_{\mu\phi}$ are Lipschitz with modulus 1 and $\frac{1}{1-\mu m_\phi}$, respectively, we obtain the claimed $L_{F_\mu} = \frac{1}{\mu}(1 + \frac{1}{1-\mu m_\phi})$. □

If ϕ is convex, then the Lipschitz constant of ∇F_μ can be improved to $2/\mu$ (Hiriart-Urruty 1991).

EC.1.3. Correspondence of Stationary Points and Global Minima of F and F_μ

In addition to being smooth, the approximation F_μ captures both the local and global structure of the original function F. In particular, some properties of F_μ established in Hiriart-Urruty (1991) are summarized in the next proposition.

PROPOSITION EC.3 **(Hiriart-Urruty (1991))**. *Suppose Assumptions 1 and 2 hold and* $0 < \mu < 1/m_\phi$ *in* (10). *Then the following claims hold.*

1. *The set of global minimizers of* F_μ, $\arg\min F_\mu$, *is nonempty, and* $F^* = \min_{z\in\mathbb{R}^n} F_\mu(z)$.

2. *(Correspondence of Stationary Point) If z is a stationary point of F_μ, i.e., $\nabla F_\mu(z)=0$, then $x_{\mu\phi}(z)=x_{\mu g}(z)$, and $x_{\mu\phi}(z)$ is a stationary point of F in the sense of (8) with $F(x_{\mu\phi}(z))=F_\mu(z)$; conversely, if $x\in\mathbb{R}^n$ is a stationary point of F, then there exists $z\in\mathbb{R}^n$ such that $\nabla F_\mu(z)=0$, $z=x_\phi(x)=x_g(x)$, and $F_\mu(z)=F(x)$.*

3. *(Correspondence of Global Minima) If $z\in\arg\min F_\mu$, then $x_{\mu\phi}(z)\in\arg\min F$; conversely, if $x\in\arg\min F$, then there exists $z\in\arg\min F_\mu$ such that $x=x_{\mu\phi}(z)=x_{\mu g}(z)$.*

EC.2. Proofs in Section 3

EC.2.1. Proof of Proposition 1

Proof. By Proposition EC.3 and Lemma 1, we know $x_{\mu\phi}(\bar z)=x_{\mu g}(\bar z)$, and for all z such that $\|z-\bar z\|\le r$, we have $F(x_{\mu g}(\bar z))=F(x_{\mu\phi}(\bar z))=F_\mu(\bar z)\le F_\mu(z)\le F(x_{\mu g}(z))$. It suffices to show that, for all x sufficiently close to $x_{\mu g}(\bar z)$, there exists some $z\in\mathbb{R}^n$ such that $\|z-\bar z\|\le r$ and $x=x_{\mu g}(z)$. In particular, take $\tilde\nabla g(x)\in\partial g(x)$ and let $z=x+\mu\tilde\nabla g(x)$. By construction, we have $0=\tilde\nabla g(x)+\mu^{-1}(x-z)\in\partial g(x)+\mu^{-1}(x-z)$, and therefore $x=x_{\mu g}(z)$. It follows that

$$\begin{aligned}\|z-\bar z\|&=\|x+\mu\tilde\nabla g(x)-x_{\mu g}(\bar z)-\mu\tilde\nabla g(x_{\mu g}(\bar z))\|\\&\le\|x-x_{\mu g}(\bar z)\|+\mu\|\tilde\nabla g(x)-\tilde\nabla g(x_{\mu g}(\bar z))\|\le\begin{cases}(1+\mu L_g)\|x-x_{\mu g}(\bar z)\|\le r\\ r-2\mu M_{\partial g}+2\mu M_{\partial g}=r,\end{cases}\end{aligned}$$

where the two cases above correspond to the two claims respectively. □

EC.2.2. Proof of Proposition 2

Proof. 1. We prove that $\mathrm{lev}_\alpha F_\mu$ is bounded. Without loss of generality, we consider $\alpha\ge F^*$, otherwise $\mathrm{lev}_\alpha F_\mu=\emptyset$ by Proposition EC.3. Firstly notice that

$$\mathrm{lev}_\alpha F_\mu\subseteq\{z:M_{\mu\phi}(z)-g(z)\le\alpha\}\subseteq\left\{z:\exists x\text{ s.t. }F(x)+\frac{1}{2\mu}\|x-z\|^2-L\|x-z\|\le\alpha+M\right\},\quad\text{(EC.1)}$$

where the first inclusion is due to $M_{\mu g}\le g$, and the second inclusion is due to our assumption on g. Using the fact that $L\|x-z\|\le\frac{tL^2}{2}+\frac{\|x-z\|^2}{2t}$ for any $t>0$ and taking $t=2\mu$, we have

$$F(x)+\frac{1}{2\mu}\|x-z\|^2-L\|x-z\|\ge F(x)+\frac{1}{4\mu}\|x-z\|^2-\mu L^2.\quad\text{(EC.2)}$$

Now (EC.1), (EC.2), and the fact that $F(x)+\frac{1}{4\mu}\|x-z\|^2\ge\max\{F(x),F^*+\frac{1}{4\mu}\|x-z\|^2\}$ imply that

$$\mathrm{lev}_\alpha F_\mu\subseteq\left\{z:\|z\|\le\sqrt{(4\mu)(\alpha+M+\mu L^2-F^*)}+\max_{x:F(x)\le\alpha+M+\mu L^2}\|x\|\right\}.\quad\text{(EC.3)}$$

Since $\mathrm{lev}_{\alpha+M+\mu L^2}F$ is compact, $\max_{x:F(x)\le\alpha+M+\mu L^2}\|x\|$ is finite, and hence $\mathrm{lev}_\alpha F_\mu$ is bounded.

2. Suppose $F=\phi-g\geq\alpha\|\cdot\|+r$ for some $a\in(0,+\infty)$ and $r\in R$. For a convex function f, we use f^* to denote its convex conjugate, i.e., $f^*(z)=\sup_x\langle z,x\rangle-f(x)$. By definition, we have

$$\mu F_\mu=\mu(M_{\mu\phi}-M_{\mu g})=\left(\mu g+\frac{1}{2}\|\cdot\|^2\right)^*-\left(\mu\phi+\frac{1}{2}\|\cdot\|^2\right)^*$$
$$\geq\left(\mu g+\frac{1}{2}\|\cdot\|^2\right)^*-\left(\mu g+\mu\alpha\|\cdot\|+\mu r+\frac{1}{2}\|\cdot\|^2\right)^*, \tag{EC.4}$$

where the inequality uses the fact that $f_1\geq f_2$ implies $f_1^*\leq f_2^*$ for any functions f_1 and f_2.

We first consider some properties of the first term in (EC.4). For simplicity, denote $p=\left(\mu g+\frac{1}{2}\|\cdot\|^2\right)^*$. Since $\mu g+\frac{1}{2}\|\cdot\|^2$ is strongly convex with modulus 1, its conjugate p is convex and has Lipschitz gradient with modulus 1 (Beck 2017, Theorem 5.26). Moreover, we claim that p is coercive: notice that for any $\bar\alpha\in(0,+\infty)$, we have

$$p(z)=\max_x\left\{\langle z,x\rangle-\mu g(x)-\frac{1}{2}\|x\|^2\right\}\geq\max_{x:\|x\|\leq\bar\alpha}\langle z,x\rangle-\max_{x:\|x\|\leq\alpha}\left\{\mu g(x)+\frac{1}{2}\|x\|^2\right\}=\bar\alpha\|z\|+\bar r,$$

where $\bar r=-\max_{x:\|x\|\leq\alpha}\left\{\mu g(x)+\frac{1}{2}\|x\|^2\right\}$ is finite, since $\mu g+\frac{1}{2}\|\cdot\|^2$ achieves a finite maximum over the compact set $\{x:\|x\|\leq\alpha\}$.

Next we rewrite the second term in (EC.4): for any $z\in\mathbb{R}^n$,

$$\left(\mu g+\mu\alpha\|\cdot\|+\mu r+\frac{1}{2}\|\cdot\|^2\right)^*(z)=\min_x\left\{p(x)+(\alpha\mu\|\cdot\|+\mu r)^*(z-x)\right\}$$
$$=\min_{w:\|w\|\leq\alpha\mu}p(z-w)-\mu r, \tag{EC.5}$$

where the first equality is due to (Beck 2017, Theorem 4.17), and the second equality uses the following facts: $\|\cdot\|^*=\delta_{\{x:\|x\|\leq1\}}$ (Beck 2017, Section 4.4.2), and $(\alpha\mu\|\cdot\|+\mu r)^*(w)=(\alpha\mu)\|\cdot\|^*(\frac{w}{\alpha\mu})-\mu r$ (Beck 2017, Theorem 4.13, 4.14). Combining (EC.4) and (EC.5), we have

$$\mu F_\mu(z)\geq p(z)-\min_{w:\|w\|\leq\alpha\mu}p(z-w)+\mu r=\max_{w:\|w\|\leq\alpha\mu}p(z)-p(z-w)+\mu r$$
$$\geq\max_{w:\|w\|\leq\alpha\mu}\langle\nabla p(z),w\rangle-\frac{1}{2}\|w\|^2+\mu r\geq\alpha\mu\|\nabla p(z)\|-\frac{1}{2}\alpha^2\mu^2+\mu r, \tag{EC.6}$$

where the second inequality is due to the Lipschitz differentiability of p, and the last inequality holds with $w=\alpha\mu\frac{\nabla p(z)}{\|\nabla p(z)\|}$ when $\|\nabla p(z)\|>0$, or any w with $\|w\|=\alpha\mu$ when $\|\nabla p(z)\|=0$. Notice that (EC.6) further suggests that

$$\liminf_{\|z\|\to\infty}F_\mu(z)\geq\liminf_{\|z\|\to\infty}\alpha\|\nabla p(z)\|-\frac{1}{2}\alpha^2\mu+r\geq\alpha\liminf_{\|z\|\to\infty}\frac{p(z)-p(0)}{\|z\|}-\frac{1}{2}\alpha^2\mu+r=+\infty, \tag{EC.7}$$

where the second inequality is due to the convexity of p: $\|\nabla p(z)\|\|z\|\geq\nabla p(z)^\top z\geq p(z)-p(0)$, and the last equality is due to p being coercive (see an equivalent characterization in (Rockafellar and Wets 2009, Definition 3.25)). Therefore, (EC.7) implies that F_μ is level-bounded.

3. Since dom ϕ is compact, there exists $R > 0$ such that dom $\phi \subseteq \{x : \|x\| \le R\}$. Notice that

$$\begin{aligned} F_\mu(z) \ge & \min_{x:\|x\|\le R} \left\{ \phi(x) + \frac{1}{2\mu}\|x\|^2 - \frac{1}{\mu}\langle x, z\rangle \right\} + \max_x \left\{ \frac{1}{\mu}\langle x, z\rangle - g(x) - \frac{1}{2\mu}\|x\|^2 \right\} \\ \ge & \hat{\phi}^* - \frac{R}{\mu}\|z\| + \max_x \left\{ \frac{1}{\mu}\langle x, z\rangle - g(x) - \frac{1}{2\mu}\|x\|^2 \right\}, \end{aligned}$$

where in the last inequality, $\hat{\phi}^* = \min_x\{\phi(x) + \frac{1}{2\mu}\|x\|^2\}$ is well-defined by the strong convexity of $\phi + \frac{1}{2\mu}\|\cdot\|^2$. Pick any $\alpha \in (0, +\infty)$. By a similar argument used in the previous part, we have

$$\max_x \left\{ \frac{1}{\mu}\langle x, z\rangle - g(x) - \frac{1}{2\mu}\|x\|^2 \right\} \ge \left(\frac{R}{\mu} + \alpha \right) \|z\| - \max_{x:\|x\|\le R+\mu\alpha} \left\{ g(x) + \frac{1}{2\mu}\|x\|^2 \right\}.$$

Combining the above two inequalities, we have $F_\mu(z) \ge \alpha\|z\| + r$, where $r = \hat{\phi}^* - \max_{x:\|x\|\le R+\mu\alpha} \left\{ g(x) + \frac{1}{2\mu}\|x\|^2 \right\}$ is finite. Therefore, we conclude that F_μ is coercive, and hence also level-bounded.

4. We show that $\mathrm{lev}_\alpha F_\mu$ is bounded. Let $z \in \mathrm{lev}_\alpha F_\mu$. By Lemma 1 and the assumption that F is level-bounded, we know that $x_{\mu\phi}(z) \in \mathrm{lev}_\alpha F$ and hence is bounded. The definition of $x_{\mu\phi}$ gives $z \in \partial(\mu\phi + \frac{1}{2}\|\cdot\|^2)(x_{\mu\phi}(z))$. Since $x_{\mu\phi}(z)$ is bounded, and $\mu\phi + \frac{1}{2}\|\cdot\|^2$ is a (strongly) convex function whose domain is $\mathbb{R}^n$, we conclude that z is bounded (Rockafellar 1970, Theorem 24.7). □

EC.3. Proofs in Section 4

EC.3.1. Proof of Theorem 1

Proof. Notice that since ∇F_μ is L_{F_μ}-Lipschitz and $\alpha \le 1/L_{F_\mu}$, we have

$$F_\mu(z^k) - F_\mu(z^{k+1}) \ge \left(\frac{1}{\alpha} - \frac{L_{F_\mu}}{2} \right) \|z^{k+1} - z^k\|^2 \ge \frac{1}{2\alpha}\|z^{k+1} - z^k\|^2 = \frac{\alpha}{2\mu^2}\|x_{\mu g}(z^k) - x_{\mu\phi}(z^k)\|^2.$$

Summing the above inequality over $k = 0, \cdots, K-1$ for some positive integer $K-1$, we have

$$\sum_{k=0}^{K-1} \|x_{\mu g}(z^k) - x_{\mu\phi}(z^k)\|^2 \le \frac{2\mu^2}{\alpha}(F_\mu(z^0) - F_\mu(z^K)) \le \frac{2\mu^2}{\alpha}(F_\mu(z^0) - F^*). \tag{EC.8}$$

Let $\bar{k} = \arg\min_{k=0,\cdots,K-1} \|x_{\mu g}(z^k) - x_{\mu\phi}(z^k)\|^2$, then from (EC.8) it holds

$$\|x_{\mu g}(z^{\bar{k}}) - x_{\mu\phi}(z^{\bar{k}})\| \le \left(\frac{2\mu^2(F_\mu(z^0) - F^*)}{\alpha K} \right)^{1/2}. \tag{EC.9}$$

For any $k \in \mathbb{Z}_+$, due to the optimality of the proximal mapping $x_{\mu\phi}(z^k)$ and $x_{\mu g}(z^k)$, we have

$$\xi^k = \mu^{-1}(z^k - x_{\mu\phi}(z^k)) - \mu^{-1}(z^k - x_{\mu g}(z^k)) \in \partial\phi(x_{\mu\phi}(z^k)) - \partial g(x_{\mu g}(z^k)). \tag{EC.10}$$

In view of (EC.9), we have (12) proved due to the claimed upper bound K in (13). Since F_μ is level-bounded and $\{F_\mu(z^k)\}_{k\in\mathbb{N}}$ is monotonically non-increasing, we know the sequence $\{z^k\}_{k\in\mathbb{N}}$ is bounded and therefore has at least one limit point z^*. Let $\{z^{k_j}\}_{j\in\mathbb{N}}$ denote the subsequence convergent to z^*. Since $x_{\mu\phi}$ and $x_{\mu g}$ are continuous, (EC.8) implies $x_{\mu\phi}(z^*) = x_{\mu g}(z^*)$. Since g is continuous, we have $\lim_{j\to\infty} g(x_{\mu g}(z^{k_j})) = g(x_{\mu g}(z^*))$; in addition,

$$\begin{aligned}\phi(x_{\mu\phi}(z^*)) &\le \liminf_{j\to\infty} \phi(x_{\mu\phi}(z^{k_j})) \le \limsup_{j\to\infty} \phi(x_{\mu\phi}(z^{k_j}))\\ &\le \lim_{j\to\infty}\left[\phi(x_{\mu\phi}(z^*)) + \frac{1}{2\mu}\|x_{\mu\phi}(z^*) - z^{k_j}\|^2 - \frac{1}{2\mu}\|x_{\mu\phi}(z^{k_j}) - z^{k_j}\|^2\right] = \phi(x_{\mu\phi}(z^*)),\end{aligned}$$

where the first inequality is due to the lower-semicontinuity of ϕ and the last inequality is due to the optimality of $x_{\mu\phi}(z^{k_j})$ in each Moreau envelope evaluation, and therefore we also have $\lim_{j\to\infty}\phi(x_{\mu\phi}(z^{k_j})) = \phi(x_{\mu\phi}(z^*))$. Taking limit on (EC.10) along the subsequence gives (8). □

EC.3.2. Proof of Lemma 2

Proof We first show that the sequence is bounded from below:

$$\mathcal{F}(x^k, z^k) \ge \min_x f(x) + h(x) + \frac{1}{2\mu}\|x - z^k\|^2 - M_{\mu g}(z^k) = F_\mu(z^k) \ge F^*,$$

where the last inequality is due to Proposition EC.2. Next we show the descent in x:

$$\begin{aligned}\mathcal{F}(x^{k+1}, z^k) \le& f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k\rangle + \frac{L_f}{2}\|x^{k+1} - x^k\|^2 + h(x^{k+1}) + \frac{1}{2\mu}\|x^{k+1} - z^k\|^2 - M_{\mu g}(z^k)\\ \le& f(x^k) + h(x^k) + \frac{1}{2\mu}\|x^k - z^k\|^2 - M_{\mu g}(z^k) + \left(\frac{L_f}{2} - \frac{1}{2\mu}\right)\|x^{k+1} - x^k\|^2\\ =& \mathcal{F}(x^k, z^k) - \left(\frac{\mu^{-1} - L_f}{2}\right)\|x^{k+1} - x^k\|^2, \qquad \text{(EC.11)}\end{aligned}$$

where the first inequality is due to the Lipschitz differentiability of f and the second inequality is due to x^{k+1} being the minimizer of some μ^{-1}-strongly convex function. The descent with respect to z is given as:

$$\begin{aligned}&\mathcal{F}(x^{k+1}, z^k) - \mathcal{F}(x^{k+1}, z^{k+1})\\ =&\frac{1}{\mu}\left(\frac{1}{\beta} - \frac{1}{2}\right)\|z^{k+1} - z^k\|^2 + M_{\mu g}(z^{k+1}) - M_{\mu g}(z^k) - \left\langle \frac{1}{\mu}(z^k - x_{\mu g}(z^k)), z^{k+1} - z^k\right\rangle\\ \ge&\frac{1}{\mu}\left(\frac{1}{\beta} - \frac{1}{2}\right)\|z^{k+1} - z^k\|^2, \qquad \text{(EC.12)}\end{aligned}$$

where we replace x^{k+1} by $x_{\mu g}(z^k) + \frac{1}{\beta}(z^{k+1} - z^k)$ to get the equality, and the inequality is due to $M_{\mu g}$ being convex and $\nabla M_{\mu g}(z^k) = \mu^{-1}(z^k - x_{\mu g}(z^k))$. Combining (EC.11) and (EC.12) gives (15). □

EC.3.3. Proof of Theorem 2

Proof. By Lemma 2, we have

$$\mathcal{F}(x^k, z^k) - \mathcal{F}(x^{k+1}, z^{k+1}) \ge \min\{c_1, c_2\}\left(\|x^{k+1} - x^k\|^2 + (\|z^{k+1} - z^k\|^2\right),$$

summing which from $k=0$ to some positive integer $K-1$ gives

$$\sum_{k=0}^{K-1}\left(\|x^{k+1}-x^k\|^2+(\|z^{k+1}-z^k\|^2\right)\le\frac{\mathcal{F}(x^0,z^0)-\mathcal{F}(x^K,z^K)}{\min\{c_1,c_2\}}\le\frac{\mathcal{F}(x^0,z^0)-F^*}{\min\{c_1,c_2\}}. \tag{EC.13}$$

Let $\bar{k}=\underset{k=0,\cdots,K-1}{\arg\min}\ \|x^{k+1}-x^k\|^2+\|z^{k+1}-z^k\|^2$, then (EC.13) implies

$$\max\{\|x^{\bar{k}+1}-x^{\bar{k}}\|,\|z^{\bar{k}+1}-z^{\bar{k}}\|\}\le\left(\frac{\mathcal{F}(x^0,z^0)-F^*}{\min\{c_1,c_2\}K}\right)^{1/2}. \tag{EC.14}$$

Due to the optimality condition of x^{k+1} and $x_{\mu g}(z^k)$, we have

$$\xi^{k+1}\in\nabla f(x^{k+1})+\partial h(x^{k+1})-\partial g(x_{\mu g}(z^k))=\partial\phi(x^{k+1})-\partial g(x_{\mu g}(z^k)),$$

which proves (17a). Now in view of (EC.14), we have

$$\max\{\|\xi^{\bar{k}+1}\|,\|x_{\mu g}(z^{\bar{k}})-x^{\bar{k}+1}\|\}\le\left(L_f+\frac{\mu^{-1}+1}{\beta}\right)\left(\frac{\mathcal{F}(x^0,z^0)-F^*}{\min\{c_1,c_2\}K}\right)^{1/2},$$

which proves (17b) and (18).

Next we show that if F_μ is level bounded, then $\{(x^k,z^k)\}_{k\in\mathbb{N}}$ stays bounded. Since F_μ is continuous and level-bounded, and $\mathcal{F}(x^0,z^0)\ge\mathcal{F}(x^k,z^k)\ge F_\mu(z^k)$, we know that z^k stays in some compact level set of F_μ. Since the mapping $x_{\mu g}$ is continuous, $x_{\mu g}(z^k)$ is also bounded. Consequently, $x^{k+1}=x_{\mu g}(z^k)+\frac{1}{\beta}(z^{k+1}-z^k)$ stays bounded for all $k\in\mathbb{N}$. Therefore, the sequence $\{(x^k,z^k)\}_{k\in\mathbb{N}}$ has a limit point, denoted as (x^*,z^*). Let $\{(x^{k_j},z^{k_j})\}_{j\in\mathbb{N}}$ be a subsequence converging to (x^*,z^*). Since $\|x^{k_j}-x^{k_j-1}\|\to0$ and $\|z^{k_j}-z^{k_j-1}\|\to0$, taking limit on $x^{k+1}=x_{\mu g}(z^k)+\frac{1}{\beta}(z^{k+1}-z^k)$ along the subsequence gives $x^*=x_{\mu g}(z^*)$. Finally the asymptotical convergence follows a similar argument as in the proof of Theorem 1. □

EC.4. Proofs in Section 5

EC.4.1. Proof of Lemma 3

Proof. We first verify condition (20) under the first two conditions. Since g is Lipschitz, we have $-g(y)\ge-g(x)-L_g\|x-y\|\ge-g(x)-\frac{1}{2}\|x-y\|^2-\frac{L_g^2}{2}$. For $0<\mu\le1$, it follows that

$$v(\mu,\rho)\ge\inf_{x\in\mathbb{R}^n}\left\{f(x)-g(x)+\frac{\rho}{2}\|Ax-b\|^2\right\}-\frac{L_g^2}{2}. \tag{EC.15}$$

The first case implies that (EC.15) is finite for any $\rho\ge0$. For the second case, notice that we can choose $\rho>0$ big enough so that $\nabla^2 f+\rho A^\top A\succ0$. Since $-g$ dominates an affine function, the objective in the right-hand side of (EC.15) is level-bounded, and hence $v(\mu,\rho)>-\infty$.

Next we verify condition (20) for the third case. Denote $\nabla^2 f=F$ and $\nabla^2 g=G$. The Hessian of the objective in (x,y) in the right-hand side of (20) is positive-definite if $\mu < \lambda_{\max}(G)^{-1}$ and its Schur complement

$$\begin{aligned} S(\mu,\rho) :=& F+\rho A^\top A+\frac{1}{\mu}I_n-\left(-\frac{1}{\mu}I_n\right)\left(\frac{1}{\mu}I_n-G\right)^{-1}\left(-\frac{1}{\mu}I_n\right)\\ =& F+\rho A^\top A+\frac{1}{\mu}I_n-\frac{1}{\mu^2}\left[\mu I_n+\mu^2 G(I_n-\mu G)^{-1}\right]=F+\rho A^\top A-G(I_n-\mu G)^{-1} \end{aligned}$$

is positive-definite, where we use the Woodbury matrix identity in the second equality, and $I_n \in \mathbb{R}^{n\times n}$ denotes the identity matrix. Since $F\succ 0$ over the null space of A by assumption, we can choose $\rho>0$ large enough so that $F+\rho A^\top A\succ 0$. Since $\lambda_{\max}(G(I-\mu G)^{-1})\to\lambda_{\max}(G)$ as $\mu\to 0$, we know $S(\mu,\rho)\succ 0$ if the smallest eigenvalue of F over the null space of A is strictly greater than $\lambda_{\max}(G)$. This completes the proof. □

EC.4.2. Proof of Lemma 4

Proof. Similar to the derivation in (EC.11)-(EC.12), the descent of ψ in x and z are given as

$$\begin{aligned} \psi(x^k,z^k,\lambda^k)-\psi(x^{k+1},z^k,\lambda^k) &\ge \left(\frac{\mu^{-1}-L_f}{2}\right)\|x^{k+1}-x^k\|^2,\\ \psi(x^{k+1},z^k,\lambda^k)-\psi(x^{k+1},z^{k+1},\lambda^k) &\ge \frac{1}{\mu}\left(\frac{1}{\beta}-\frac{1}{2}\right)\|z^{k+1}-z^k\|^2. \end{aligned}$$

In addition,

$$\psi(x^{k+1},z^{k+1},\lambda^k)-\psi(x^{k+1},z^{k+1},\lambda^{k+1})=\langle\lambda^k-\lambda^{k+1},Ax^{k+1}-b\rangle=-\frac{1}{\rho}\|\lambda^{k+1}-\lambda^k\|^2.$$

Adding the above three expressions completes the proof. □

EC.4.3. Proof of Lemma 5

Proof. For $k\in\mathbb{N}$, the update of x^{k+1} gives $\nabla f(x^k)+A^\top\lambda^{k+1}+\mu^{-1}(x^{k+1}-z^k)=0$, which implies that for $k\in\mathbb{N}$, $A^\top(\lambda^{k+1}-\lambda^k)=\mu^{-1}(x^k-x^{k+1})+(\nabla f(x^{k-1})-\nabla f(x^k))+\mu^{-1}(z^k-z^{k-1})$. Since $\lambda^{k+1}-\lambda^k=\rho(Ax^{k+1}-b)$ belongs to the column space of A, we have

$$\sigma_{\min}^+(A)\|\lambda^{k+1}-\lambda^k\|\le\|A^\top(\lambda^{k+1}-\lambda^k)\|\le\mu^{-1}\|x^{k+1}-x^k\|+L_f\|x^k-x^{k-1}\|+\mu^{-1}\|z^k-z^{k-1}\|,$$

where the first inequality is due to the min-max theorem of eigenvalues of a real symmetric matrix. Dividing both sides by $\sigma_{\min}^+(A)$ gives the desired inequality. □

EC.4.4. Proof of Lemma 6

Proof. 1. By Lemma 5, squaring both sides gives

$$\|\lambda^{k+1}-\lambda^k\|^2\le c_3\|x^{k+1}-x^k\|^2+c_4\|x^k-x^{k-1}\|^2+c_3\|z^k-z^{k-1}\|^2.$$

Then (28) follows from Lemma 4 and constants defined in (27).

2. Recall that

$$\Psi_k \geq \psi(x^k, z^k, \lambda^k) \geq f(x^k) - g(z^k) + \frac{\rho}{2}\|Ax^k - b\|^2 + \frac{1}{2\mu}\|x^k - z^k\|^2 + \frac{1}{\rho}\langle \lambda^k, \lambda^k - \lambda^{k-1}\rangle$$
$$\geq v(\mu, \rho) + \frac{1}{2\rho}(\|\lambda^k\|^2 - \|\lambda^{k-1}\|^2),$$

where the second inequality is due to $M_{\mu g}(z) \leq g(z)$ and $\lambda^k = \lambda^{k-1} + \rho(Ax^k - b)$, and the third inequality is due to (20). This further implies that

$$\sum_{k=1}^{K}(\Psi_k - v(\mu,\rho)) \geq \frac{1}{2\rho}(\|\lambda^K\|^2 - \|\lambda^0\|^2) \geq -\frac{1}{2\rho}\|\lambda^0\|^2 > -\infty,$$

for all positive integer K. Since Ψ_k is non-increasing, we must have $\Psi_k \geq v(\mu,\rho)$ for all $k \in \mathbb{Z}_+$; otherwise, there exists some $\delta > 0$ such that $\Psi_k - v(\mu,\rho) < -\delta$ for all large enough k, then the above summation would converge to $-\infty$ as $K \to \infty$.

3. For any positive integer K, summing (28) from 0 to $K-1$ gives

$$\kappa_{\min} K \min_{k=0,\cdots,K-1}\left\{\|x^{k+1} - x^k\|^2 + \|z^{k+1} - z^k\|^2 + \|x^k - x^{k-1}\|^2 + \|z^k - z^{k-1}\|^2\right\} \tag{EC.16}$$
$$\leq \sum_{k=0}^{K-1}\left(\kappa_1\|x^{k+1} - x^k\|^2 + \kappa_2\|z^{k+1} - z^k\|^2 + \kappa_3\|x^k - x^{k-1}\|^2 + \kappa_4\|z^k - z^{k-1}\|^2\right)$$
$$\leq \sum_{k=0}^{K-1}(\Psi_k - \Psi_{k+1}) = \Psi_0 - \Psi_K \leq \Psi_0 - v(\mu,\rho),$$

where the last inequality is due to $\Psi_K \geq v(\mu,\rho)$ for all $K \geq 1$. Now let $\bar{k}$ be the minimizer in (EC.16); it follows that

$$\max\left\{\|x^{\bar{k}+1} - x^{\bar{k}}\|^2, \|z^{\bar{k}+1} - z^{\bar{k}}\|^2, \|x^{\bar{k}} - x^{\bar{k}-1}\|^2, \|z^{\bar{k}} - z^{\bar{k}-1}\|^2\right\}$$
$$\leq \left(\|x^{\bar{k}+1} - x^{\bar{k}}\|^2 + \|z^{\bar{k}+1} - z^{\bar{k}}\|^2 + \|x^{\bar{k}} - x^{\bar{k}-1}\|^2 + \|z^{\bar{k}} - z^{\bar{k}-1}\|^2\right) \leq \frac{\Psi_0 - v(\mu,\rho)}{\kappa_{\min} K}.$$

This completes the proof. □

EC.4.5. Proof of Lemma 7

Proof. By (34) and the μ-strong convexity of the function in the following line, it holds for all $x \in \mathcal{H}$ that

$$\langle \nabla f(x^k) - \xi_g^k, x - x^k\rangle + h(x) + \langle \lambda^k, Ax - b\rangle + \frac{\rho}{2}\|Ax - b\|^2 + \frac{1}{2\mu}\|x - z^k\|^2$$
$$\geq \langle \nabla f(x^k) - \xi_g^k, x^{k+1} - x^k\rangle + h(x^{k+1}) + \langle \lambda^k, Ax^{k+1} - b\rangle + \frac{\rho}{2}\|Ax^{k+1} - b\|^2$$
$$+ \frac{1}{2\mu}\|x^{k+1} - z^k\|^2 + \langle \zeta^{k+1}, x - x^{k+1}\rangle + \frac{1}{2\mu}\|x^{k+1} - x\|^2. \tag{EC.17}$$

Using the Lipschitz condition of ∇f, the convexity of g, inequality (EC.17) with $x = x^k$, the fact that $\langle \zeta^{k+1}, x^{k+1} - x^k \rangle \le \frac{1}{4\mu}\|x^{k+1} - x^k\|^2 + \mu\|\zeta^{k+1}\|^2$, and the choice that $\|\zeta^{k+1}\| \le \epsilon_{k+1}$, the descent in x can be derived as follows:

$$P(x^{k+1}, z^k, \lambda^k) \le P(x^k, z^k, \lambda^k) + \mu\epsilon_{k+1}^2 - \frac{\mu^{-1} - 2L_f}{4}\|x^{k+1} - x^k\|^2.$$

The descent with respect to variable z can be derived as follows:

$$P(x^{k+1}, z^k, \lambda^k) - P(x^{k+1}, z^{k+1}, \lambda^k) = \frac{1}{2\mu}\left(\frac{2}{\beta} - 1\right)\left\|z^{k+1} - z^k\right\|^2 \ge \frac{1}{2\beta\mu}\left\|z^{k+1} - z^k\right\|^2,$$

where we replace x^{k+1} by $z^k + \frac{1}{\beta}(z^{k+1} - z^k)$ to get the equality, and the inequality is due to $\beta \le 1$. Finally, similar to Lemma 4, the change in λ is

$$P(x^{k+1}, z^{k+1}, \lambda^k) - P(x^{k+1}, z^{k+1}, \lambda^{k+1}) = -\frac{1}{\rho}\|\lambda^{k+1} - \lambda^k\|^2.$$

Combining the above three expressions completes the proof. □

EC.4.6. Proof of Lemma 8

Proof. The optimality condition (EC.17) with $k = 0$ and $x = \bar{x}$ (by Assumption 4, $A\bar{x} = b$) gives

$$\begin{aligned}&\langle \nabla f(x^0) - \xi_g^0, x^1 - x^0 \rangle + h(x^1) + \langle \lambda^0, Ax^1 - b \rangle + \frac{\rho}{2}\|Ax^1 - b\|^2 + \frac{1}{2\mu}\|x^1 - z^0\|^2 \\ \le& \langle \nabla f(x^0) - \xi_g^0, \bar{x} - x^0 \rangle + h(\bar{x}) + \frac{1}{2\mu}\|\bar{x} - z^0\|^2 + \langle \zeta^1, x^1 - \bar{x} \rangle. \end{aligned} \tag{EC.18}$$

Since $\|\zeta^1\| \le \epsilon_1 \le 1$, $\mu < L_f^{-1}$, and $x^1, \bar{x} \in \mathcal{H}$, we have

$$\langle \zeta^1, x^1 - \bar{x} \rangle \le \frac{\mu}{2}\|\zeta^1\|^2 + \frac{1}{2\mu}\|x^1 - \bar{x}\|^2 \le \frac{L_f^{-1}}{2} + \frac{1}{2\mu}D_{\mathcal{H}}^2. \tag{EC.19}$$

The above two inequalities together with the L_h-Lipschitz continuity of h and $\bar{x}, z^0 \in \mathcal{H}$ give that

$$\rho\|Ax^1 - b\|^2 \le 2(M_{\nabla f} + M_{\partial g} + L_h)D_{\mathcal{H}} + 2\mu^{-1}D_{\mathcal{H}}^2 + L_f^{-1} + 2\|\lambda^0\| \max_{x \in \mathcal{H}} \|Ax - b\|. \tag{EC.20}$$

Notice that due to ∇f being Lipschitz and g being convex, $P(x^1, z^0, \lambda^0)$ is bounded from above by

$$\begin{aligned}&f(x^0) - g(x^0) + \frac{L_f}{2}\|x^1 - x^0\|^2 + \langle \nabla f(x^0) - \xi_g^0, x^1 - x^0 \rangle + h(x^1) + \langle \lambda^0, Ax^1 - b \rangle + \frac{\rho}{2}\|Ax^1 - b\|^2 + \frac{1}{2\mu}\|x^1 - z^0\|^2 \\ \le& f(x^0) - g(x^0) + \frac{L_f}{2}\|x^1 - x^0\|^2 + \langle \nabla f(x^0) - \xi_g^0, \bar{x} - x^0 \rangle + h(\bar{x}) + \frac{1}{2\mu}\|\bar{x} - z^0\|^2 + \langle \zeta^1, x^1 - \bar{x} \rangle \\ \le& \max_{x \in \mathcal{H}}\{f(x) + h(x) - g(x)\} + (L_h + M_{\nabla f} + M_{\partial g})D_{\mathcal{H}} + \frac{L_f + 2\mu^{-1}}{2}D_{\mathcal{H}}^2 + \frac{L_f^{-1}}{2}, \end{aligned} \tag{EC.21}$$

where the first inequality is due to (EC.18), and the second inequality is due to the compactness of $\mathcal{H}$, the L_h-Lipschitz continuity of h, and (EC.19). By Lemma 7, (EC.20), and (EC.21), we have

$$P(x^1, z^1, \lambda^1) = P(x^1, z^1, \lambda^0) + \rho\|Ax^1 - b\|^2 \le P(x^1, z^0, \lambda^0) + \rho\|Ax^1 - b\|^2 \le \overline{P}.$$

This completes the proof. □

EC.4.7. Proof of Lemma 9

Proof. By step 4 in Algorithm 3, there exists $\xi_h^{k+1} \in \partial h(x^{k+1})$ such that

$$\zeta^{k+1} = \nabla f(x^k) - \xi_g^k + \xi_h^{k+1} + A^\top \lambda^{k+1} + \frac{1}{\mu}(x^{k+1} - z^k). \tag{EC.22}$$

Since $\lambda^0, b \in \mathrm{Im}(A)$ and $\lambda^{k+1} = \lambda^k + \rho(Ax^{k+1} - b)$, $\lambda^{k+1} \in \mathrm{Im}(A)$ for all $k \in \mathbb{N}$. Since $z^0 \in \mathcal{H}$, $x^k \in \mathcal{H}$, and $z^{k+1} = (1-\beta)z^k + \beta x^{k+1}$, $z^k \in \mathcal{H}$ for all $k \in \mathbb{N}$ as well. Consequently,

$$\|\lambda^{k+1}\| \le \frac{1}{\sigma_{\min}^+(A)} \left(M_{\nabla f} + M_{\partial g} + \frac{D_{\mathcal{H}}}{\mu} + 1 \right) + \frac{\|\xi_h^{k+1}\|}{\sigma_{\min}^+(A)}. \tag{EC.23}$$

By (Melo et al. 2020, Lemma 4.7), we can bound $\|\xi_h^{k+1}\|$ as follows:

$$\bar{d}\|\xi_h^{k+1}\| \le (\bar{d} + \|x^{k+1} - \bar{x}\|)L_h + \left\langle \xi_h^{k+1}, x^{k+1} - \bar{x} \right\rangle \le 2D_{\mathcal{H}}L_h + \left\langle \xi_h^{k+1}, x^{k+1} - \bar{x} \right\rangle. \tag{EC.24}$$

Using (EC.22), we can further bound the inner product term in (EC.24) by

$$\begin{aligned}
\langle \xi_h^{k+1}, x^{k+1} - \bar{x} \rangle &= \left\langle \zeta^{k+1} - \nabla f(x^k) + \xi_g^k - A^\top \lambda^{k+1} - \frac{1}{\mu}(x^{k+1} - z^k), x^{k+1} - \bar{x} \right\rangle \\
&\le \left(M_{\nabla f} + M_{\partial g} + \frac{D_{\mathcal{H}}}{\mu} + 1 \right) D_{\mathcal{H}} - \left\langle \lambda^{k+1}, Ax^{k+1} - b \right\rangle \\
&\le \left(M_{\nabla f} + M_{\partial g} + \frac{D_{\mathcal{H}}}{\mu} + 1 \right) D_{\mathcal{H}} + \frac{1}{\rho}\|\lambda^{k+1}\|\|\lambda^k\| - \frac{1}{\rho}\|\lambda^{k+1}\|^2,
\end{aligned} \tag{EC.25}$$

where we use the facts that $A\bar{x} = b$ and $\|\zeta^{k+1}\| \le \epsilon_{k+1} \le 1$ in the first inequality, and $Ax^{k+1} - b = \frac{1}{\rho}(\lambda^{k+1} - \lambda^k)$ to get the second inequality. Combining (EC.23), (EC.24) and (EC.25), we have

$$\frac{\|\lambda^{k+1}\|^2}{\rho\sigma_{\min}^+(A)} + \bar{d}\|\lambda^{k+1}\| \le \frac{\|\lambda^{k+1}\|\|\lambda^k\|}{\rho\sigma_{\min}^+(A)} + \frac{2D_{\mathcal{H}}}{\sigma_{\min}^+(A)} \left(M_{\nabla f} + M_{\partial g} + \frac{D_{\mathcal{H}}}{\mu} + L_h + 1 \right),$$

which further implies that, for all $k \in \mathbb{N}$,

$$\left(\frac{\|\lambda^{k+1}\|}{\rho\sigma_{\min}^+(A)} + \bar{d} \right) \|\lambda^{k+1}\| \le \frac{\|\lambda^{k+1}\|}{\rho\sigma_{\min}^+(A)} \|\lambda^k\| + \bar{d}\Lambda,$$

The claim then follows from an inductive argument: if $\|\lambda^k\| \le \Lambda$, then the above inequality implies that $\|\lambda^{k+1}\| \le \Lambda$ as well, and $\|\lambda^0\| \le \Lambda$ holds by the definition of Λ. □

EC.4.8. Proof of Lemma 10

Proof. 1. For all $k \in \mathbb{N}$,

$$\begin{aligned}
P(x^k, z^k, \lambda^k) &= f(x^k) + h(x^k) - g(x^k) + \langle \lambda, Ax^k - b \rangle + \frac{\rho}{2}\|Ax^k - b\|^2 + \frac{1}{2\mu}\|x^k - z^k\|^2 \\
&\ge \min_{x \in \mathbb{R}^n} \{f(x) + h(x) - g(x)\} - \Lambda \max_{x \in \mathcal{H}} \|Ax - b\| > -\infty,
\end{aligned}$$

where the inequality is due to the continuity of f, h, g, and $\|\cdot\|$ over compact domain $\mathcal{H}$.

2. Lemma 7 implies that, for all $k \in \mathbb{N}$,

$$\eta\left(\|x^{k+1}-x^k\|^2+\|z^{k+1}-z^k\|^2\right) \le P(x^k,z^k,\lambda^k)-P(x^{k+1},z^{k+1},\lambda^{k+1})+\frac{1}{\rho}\|\lambda^{k+1}-\lambda^k\|^2+\mu\epsilon_{k+1}^2.$$

Summing the above inequality over $k=1,\dots,K$,

$$\begin{aligned}
&\eta K\min_{k\in[K]}\left\{\|x^{k+1}-x^k\|^2+\|z^{k+1}-z^k\|^2\right\} \le \eta\sum_{k=1}^{K}\left\{\|x^{k+1}-x^k\|^2+\|z^{k+1}-z^k\|^2\right\}\\
\le& P(x^1,z^1,\lambda^1)-P(x^{K+1},z^{K+1},\lambda^{K+1})+\frac{1}{\rho}\sum_{k=1}^{K}\|\lambda^{k+1}-\lambda^k\|^2+\mu\sum_{k=1}^{K}\epsilon_{k+1}^2\\
\le& \overline{P}-\underline{P}+\mu E+\frac{4}{\rho}\sum_{k=1}^{K+1}\|\lambda^k\|^2 \le \overline{P}-\underline{P}+\mu E+\frac{4(K+1)\Lambda^2}{\rho},
\end{aligned}$$

where we use the the fact that $P(x^{K+1},z^{K+1},\lambda^{K+1}) \ge \underline{P}$ for all integer $K\in\mathbb{Z}_+$ and $\|\lambda^{k+1}-\lambda^k\|^2 \le 2\|\lambda^{k+1}\|^2+2\|\lambda^k\|^2$ in the third inequality, and Lemma 9 in the last inequality. Let $\bar{k}$ be the minimizer in the first line of the above chain of inequalities, then dividing both sides by any positive integer K gives

$$\max\left\{\|x^{\bar{k}+1}-x^{\bar{k}}\|^2,\|z^{\bar{k}+1}-z^{\bar{k}}\|^2\right\} \le \frac{\overline{P}-\underline{P}+\mu E}{\eta K}+\frac{8\Lambda^2}{\eta\rho}.$$

This completes the proof. □