SIAM J. OPTIM.
Vol. 34, No. 2, pp. 1679–1707

DUAL DESCENT AUGMENTED LAGRANGIAN METHOD AND ALTERNATING DIRECTION METHOD OF MULTIPLIERS*

KAIZHAO SUN† AND XU ANDY SUN‡

Abstract. Classical primal-dual algorithms attempt to solve $\max_\mu \min_x \mathcal{L}(x,\mu)$ by alternately minimizing over the primal variable x through primal descent and maximizing the dual variable μ through dual ascent. However, when $\mathcal{L}(x,\mu)$ is highly nonconvex with complex constraints in x, the minimization over x may not achieve global optimality and, hence, the dual ascent step loses its valid intuition. This observation motivates us to propose a new class of primal-dual algorithms for nonconvex constrained optimization with the key feature to reverse dual ascent to a conceptually new *dual descent,* in a sense, elevating the dual variable to the same status as the primal variable. Surprisingly, this new dual scheme achieves some best iteration complexities for solving nonconvex optimization problems. In particular, when the dual descent step is scaled by a fractional constant, we name it scaled dual descent (SDD), otherwise, unscaled dual descent (UDD). For nonconvex multiblock optimization with nonlinear equality constraints, we propose SDD-alternating direction method of multipliers (SDD-ADMM) and show that it finds an ϵ-stationary solution in $\mathcal{O}(\epsilon^{-4})$ iterations. The complexity is further improved to $\mathcal{O}(\epsilon^{-3})$ and $\mathcal{O}(\epsilon^{-2})$ under proper conditions. We also propose UDD-augmented Lagrangian method (UDD-ALM), combining UDD with ALM, for weakly convex minimization over affine constraints. We show that UDD-ALM finds an ϵ-stationary solution in $\mathcal{O}(\epsilon^{-2})$ iterations. These complexity bounds for both algorithms either achieve or improve the best-known results in the ADMM and ALM literature. Moreover, SDD-ADMM addresses a long-standing limitation of existing ADMM frameworks.

Key words. augmented Lagrangian method, alternating direction method of multipliers

MSC codes. 65K05, 90C26, 90C30, 90C46

DOI. 10.1137/21M1449099

1. Introduction.

In this paper, we consider the following problem:

$$\min_{x\in\mathbb{R}^n} \left\{ f(x) + \sum_{i=1}^p g_i(x_i) \,\Big|\, \sum_{i=1}^p h_i(x_i) = 0 \right\}, \tag{1.1}$$

where the variable $x \in \mathbb{R}^n$ has the block-coordinate form $x = [x_1^\top, \ldots, x_p^\top]^\top$ with each $x_i \in \mathbb{R}^{n_i}$ for $i \in [p] := \{1, 2, \ldots, p\}$ and $\sum_{i=1}^p n_i = n$. We assume $f : \mathbb{R}^n \to \mathbb{R}$ has a Lipschitz gradient, and each $g_i : \mathbb{R}^{n_i} \to \overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ is proper, lower-semicontinuous, and possibly nonconvex; in addition, for each $i \in [p]$, constraints $h_i : \mathbb{R}^{n_i} \to \mathbb{R}^m$ are continuously differentiable over the domain of g_i. Denote $g(x) := \sum_{i=1}^p g_i(x_i)$ and $h(x) := \sum_{i=1}^p h_i(x_i)$.

The augmented Lagrangian method (ALM), which was proposed in the late 1960s [21, 42], provides a powerful algorithmic framework for constrained optimization problems including (1.1). Define the augmented Lagrangian function as

$$\mathcal{L}_\rho(x,\mu) := f(x) + g(x) + \langle \mu, h(x)\rangle + \frac{\rho}{2}\|h(x)\|^2, \tag{1.2}$$

*Received by the editors September 27, 2021; accepted for publication (in revised form) November 16, 2023; published electronically May 7, 2024.
https://doi.org/10.1137/21M1449099
Funding: The research is partially supported by the NSF CAREER Award 1751747.
†Decision Intelligence Lab, DAMO Academy, Alibaba Group (U.S.) Inc., Bellevue, WA 98004 USA (kaizhao.s@alibaba-inc.com).
‡Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02142 USA (sunx@mit.edu).

where $\mu \in \mathbb{R}^m$ and $\rho > 0$. In the $(k+1)$th iteration, the ALM first obtains the primal iterate x^{k+1} by minimizing the augmented Lagrangian function with dual variable μ^k fixed, possibly in an inexact way,

$$x^{k+1} \approx \operatorname*{argmin}_{x \in \mathbb{R}^n} \mathcal{L}_\rho(x, \mu^k), \tag{1.3}$$

and then updates the dual variable using primal residuals,

$$\mu^{k+1} = \mu^k + \varrho_k h(x^{k+1}), \tag{1.4}$$

where $\varrho_k > 0$ is a positive dual step size.

The ALM framework is flexible: x^{k+1} in (1.3) is allowed to be (some approximate counterpart of) a global minimum [45], a local minimum [4], or just a stationary point [1]. Another possibility is to update blocks of variables $(x_1, \dots, x_p)$ in a coordinate fashion, i.e., through a Gauss–Seidel or Jacobi sweep; when h is affine, algorithms of this type are commonly known as the alternating direction method of multipliers (ADMM). The dual update (1.4) is motivated by the fact that the augmented Lagrangian dual function

$$d(\mu) := \min_{x \in \mathbb{R}^n} \mathcal{L}_\rho(x, \mu) \tag{1.5}$$

is concave, and $-h(x^{k+1}) \in \partial_\epsilon(-d)(\mu^k)$ with any x^{k+1} such that $\mathcal{L}_\rho(x^{k+1}, \mu^k) \le d(\mu^k) + \epsilon$. In this case, the update (1.4) is essentially maximizing the concave function d using an inexact subgradient of $-d$. We refer to (1.4) as a *dual ascent* step. A motivation for this paper is that the classic interpretation of dual ascent of (1.4) is not valid anymore if the gap between $d(\mu^k)$ and $\mathcal{L}_\rho(x^{k+1}, \mu^k)$ is large or cannot be uniformly bounded over iterations, especially when x^{k+1} is a local minimum, a stationary point, or a coordinatewise solution of the nonconvex function $\mathcal{L}_\rho(\cdot, \mu^k)$.

This observation opens up new possibilities for algorithmic design within the augmented Lagrangian framework. Given $\mu^k \in \mathbb{R}^m$, let x^{k+1} represent a coordinatewise solution of $\mathcal{L}(\cdot, \mu^k)$. Notably, since x^{k+1} does not provide valid zero-/first-order information of d at μ^k, the intuition of the dual ascent (1.4) is lost. Instead of maximizing d, the fact that $\nabla_\mu \mathcal{L}\rho(x^{k+1}, \cdot) = h(x^{k+1})$ suggests an alternative approach. By "minimizing" $\mathcal{L}\rho$ with respect to μ and assuming an approximate stationary point can be attained, it is expected that the primal residual $\|h(x^{k+1})\|$ will be small. To pursue this idea, one might be inclined to employ block-coordinate descent algorithms [57, 58] for $\mathcal{L}\rho(x_1, \dots, x_p, \mu)$; however, the linearity of the function $\mathcal{L}\rho(x, \cdot)$ renders $\mathcal{L}_\rho$ potentially unbounded in the dual variable μ. To address this, we introduce a regularized augmented Lagrangian function,

$$\mathcal{P}(x, \mu) := \mathcal{L}_\rho(x, \mu) + \frac{\omega}{2\rho}\|\mu\|^2, \tag{1.6}$$

where we include a quadratic term in μ with $\omega > 0$. Once x^{k+1} is obtained, for example, through a Gauss–Seidel sweep of proximal gradient updates, we can update μ^{k+1} using the following formulation,

$$\mu^{k+1} = \operatorname*{argmin}_{\mu \in \mathbb{R}^m} \mathcal{P}(x^{k+1}, \mu) + \frac{\tau\omega}{2\rho}\|\mu - \mu^k\|^2 = \underbrace{\frac{\tau}{1+\tau}}_{\text{scaled}} \underbrace{\left(\mu^k - \tau^{-1}\omega^{-1}\rho h(x^{k+1})\right)}_{\text{dual descent}}, \tag{1.7}$$

where $\tau > 0$ and this update is referred to as the *scaled dual descent* (SDD).

The above update ensures (1) sufficient descent and lower-boundedness of $\mathcal{P}$ and (2) boundedness of the sequence $\{\mu^{k+1}\}_{k\in\mathbb{N}}$, which are critical for the convergence rate analysis of ALM-based algorithms. In particular, we show that, with a near-feasible initialization, the SDD update gives an ϵ-stationary solution of (1.1) in $\mathcal{O}(\epsilon^{-4})$ iterations, which can be further improved to $\mathcal{O}(\epsilon^{-3})$ and $\mathcal{O}(\epsilon^{-2})$ under additional verifiable assumptions. Inspired by a comment from a referee, we have made an intriguing observation. By using a different proximal center $\hat{\mu}^k$ instead of μ^k in (1.7), defined as

$$\hat{\mu}^k := \mu^k + \frac{\rho}{\omega\tau} h(x^{k+1}), \tag{1.8}$$

we find that the dual variable vanishes in all iterations when initialized with zeros. This realization highlights the versatility of the SDD framework, as it not only introduces a novel class of dual updates but also encompasses the classic penalty method when combined with a traditional dual ascent step (1.8). Consequently, we provide a unified convergence analysis for both SDD and the penalty method, with the complexity results for the latter being novel contributions to the literature.

A natural question then arises: what will happen if we simply perform an *unscaled dual descent* (UDD) update, i.e.,

$$\mu^{k+1} = \mu^k - \varrho \nabla_\mu \mathcal{L}_\rho(x^{k+1}, \mu^k) = \mu^k - \varrho h(x^{k+1}), \tag{1.9}$$

where $\varrho > 0$ is a fixed dual step size. The analysis of UDD presents a main technical challenge in establishing the boundedness of the dual variable to prevent the augmented Lagrangian function from becoming unbounded from below. In this paper, we provide some positive theoretical results and preliminary empirical observations for the UDD update. From a theoretical perspective, we demonstrate that when some regularity condition holds at the primal limit point, regardless of the choice of $\varrho > 0$, the dual sequence has a bounded subsequence and hence the augmented Lagrangian function is lower bounded; as a result, the UDD update finds an ϵ-stationary solution in $\mathcal{O}(\epsilon^{-2})$ iterations. On the empirical side, we observe that UDD converges on a simple consensus problem when the step size ϱ is close to zero. In this scenario, the UDD update (1.9) describes the limiting behavior of the SDD update (1.7) with τ converging to $+\infty$ and ω remaining constant. Essentially, the empirical convergence of UDD could be attributed to the penalty method, where the update of the dual variable is relatively negligible.

To our knowledge, references [25, 26] are the only two works that use a related idea of dual descent. The authors consider a special case of (1.1) with two blocks of variables with continuously differentiable coupling constraints. In each iteration, the proposed algorithm first solves two quadratic programs (QPs), and then uses the solutions to determine the primal and dual descent directions of the augmented Lagrangian function. Assuming boundedness of dual variables, convergence to a stationary point is proved with an iteration complexity of $\mathcal{O}(\epsilon^{-2})$ QP oracles. The proposed dual descent framework in this paper is different from [25, 26] in several nontrivial perspectives. We summarize our contributions in the next subsection.

1.1. Contributions. We summarize our contributions as follows. We introduce SDD within the augmented Lagrangian framework to solve nonlinear constrained nonconvex problem (1.1). In iteration $k+1$, we obtain x^{k+1} through a Gauss–Seidel sweep of proximal gradient updates, and then update the dual variable μ^{k+1} via an SDD step. We call the resulting algorithm SDD-ADMM when $p > 1$, and SDD-ALM

when $p = 1$. In contrast to most existing ADMM and ALM works considering only affine constraints (see section 2 for a detailed review), the proposed SDD-ADMM and SDD-ALM are able to handle nonlinear smooth coupling constraints of the form $\sum_{i=1}^{p} h_i(x_i) = 0$, and therefore are applicable to a broader class of problems.

In addition to being able to handle nonlinear constraints, SDD-ADMM ($p > 1$) achieves better iteration complexities under a more general setting. Compared to existing multiblock nonconvex ADMM works [17, 24, 27, 38, 37, 55], we do not impose restrictive assumptions on problem data (see section 2.2), and we show that SDD-ADMM obtains an ϵ-stationary solution in $\mathcal{O}(\epsilon^{-4})$ iterations, which can be further improved to $\mathcal{O}(\epsilon^{-3})$ or $\mathcal{O}(\epsilon^{-2})$ under suitable conditions. Our $\mathcal{O}(\epsilon^{-4})$ and $\mathcal{O}(\epsilon^{-3})$ estimates significantly improve the existing $\mathcal{O}(\epsilon^{-6})$ [27] and $\mathcal{O}(\epsilon^{-4})$ [54] complexities, respectively, and our $\mathcal{O}(\epsilon^{-2})$ estimate complements the above-mentioned references. Moreover, our iteration complexities are measured by first-order oracles, i.e., gradient oracles of f and h and proximal oracles of g_i's, which are in general more tractable than the subproblem oracles considered in [27, 54].

For SDD-ALM ($p = 1$), our iteration complexities slightly improve the best-known results in [33] (without a technical assumption) and [31] (with a technical assumption) by getting rid of the logarithmic dependency on ϵ^{-1}. Another feature of SDD-ALM is that the algorithm is single-looped, which might be preferable over double- or triple-looped ALM and penalty methods [31, 33, 48, 56] from an implementation point of view, i.e., the technicality of choosing the inner-loop stopping criteria is avoided. In addition, our convergence analysis and complexity estimates also apply to an interesting single-looped first-order penalty method.

To further understand the behavior of dual descent, we introduce UDD within the augmented Lagrangian framework and name the resulting algorithm UDD-ALM. We first investigate UDD-ALM for weakly convex minimization with affine constraints and show that when a certain regularity condition holds at the primal limit point, UDD-ALM finds an ϵ-stationary solution in $\mathcal{O}(\epsilon^{-2})$ iterations. UDD-ALM is single-looped and our iteration complexity is again measured by first-order oracles of f and g. We do not restrict g to be an indicator function of a box or polyhedron and, hence, our result complements those in [60, 61]. Finally, we extend the analysis of UDD-ALM to handle nonconvex g and nonlinear constraints h by assuming a novel descent oracle of each proximal augmented Lagrangian relaxation over the domain of g. We would like to acknowledge that there is still a need for a deeper understanding of the behavior of UDD, particularly concerning the implicit impact of the dual step size on the regularity assumption we imposed on the primal limit point. We do not claim or advocate the superiority of UDD over existing algorithms, but simply share our current theoretical understanding and empirical observations on this counterintuitive approach. Our hope is that both SDD and UDD can serve as catalysts for inspiring further algorithmic developments that go beyond traditional approaches.

1.2. Notations. We denote the set of positive integers up to p by $[p]$, the set of nonnegative integers by $\mathbb{N}$, the set of real numbers by $\mathbb{R}$, the set of an 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 $A \in \mathbb{R}^{m\times n}$, $\|A\|$ and $\sigma_{\min}(A)$ denote the largest and smallest singular value of A, respectively. For $X \subseteq \mathbb{R}^n$, we use δ_X to denote the $0/\infty$-indicator function of X. Denote $\operatorname{dist}(x, x) := \inf_{y\in X} \|y - x\|$. When $x \in \mathbb{R}^n$ has the block-coordinate form $[x_1^\top, \ldots, x_p^\top]^\top$ with $x_i \in \mathbb{R}^{n_i}$ for $i \in [p]$ and $\sum_{i=1}^{p} n_i = n$, we denote $x_{\le i} = [x_1^\top, \ldots, x_i^\top]^\top$ and $x_{\ge i} = [x_i^\top, \ldots, x_p^\top]^\top$ for $i \in [p]$, and similarly for $x_{<i}$ and $x_{>i}$. We use the following notations for x when it is necessary to distinguish variable blocks (specifically when x is an argument of a function): $(x_{\le i}, x_{>i})$, $(x_{<i}, x_i, x_{>i})$, $(x_{<i}, x_{\ge i})$,

or $(x_{\neq i}, x_i)$. Finally, we adopt the following notations for complexity analysis. Let $\varepsilon > 0$ and $K > 0$. We write $K = \mathcal{O}(\varepsilon)$ if $K \leq B\varepsilon$ for some $0 < B < +\infty$, and $K = \Theta(\varepsilon)$ if $b\varepsilon \leq K \leq B\varepsilon$ for some $0 < b < B < +\infty$.

1.3. Organization. The rest of the paper is organized as follows. Section 2 reviews related works in ALM and ADMM. In section 3, we introduce the SDD-ADMM algorithm, present its convergence analysis as well as an adaptive version, and discuss its connection with existing algorithms. In section 4, we establish the convergence of UDD-ALM under the setting where $p = 1$, $h(x)$ is affine, and g is convex, and further extend the analysis to handle nonconvex g and nonlinear h. We present some numerical experiments in section 5 and finally conclude this paper in section 6.

2. Related works.

This section reviews the literature of ALM and ADMM.

2.1. ALM.

The asymptotic convergence and convergence rate of ALM have been extensively studied for convex programs by [30, 44, 45] and smooth nonlinear programs [1, 4]. In this subsection, we review some recent developments on ALM-based algorithms applied to nonconvex problems of the form

$$\min_{x \in \mathbb{R}^n} \{F(x) \mid h_E(x) = 0, h_I(x) \leq 0\}. \tag{2.1}$$

Often the objective F is assumed to admit a composite form $f + g$, where f has Lipschitz gradient and g is a nonsmooth convex function.

2.1.1. Convex constraints. Works [18, 22, 29, 39, 53, 59, 60, 61] consider affine constraints, i.e., $h_E(x) = Ax - b = 0$, while inequality constraints $h_I(x) \leq 0$ are not present. For a special case with $g = 0$, Hong [22] proposed a proximal primal-dual algorithm (prox-PDA) that finds an approximate stationary point in $\mathcal{O}(\epsilon^{-2})$ iterations, where $\epsilon > 0$ measures both first-order stationarity and feasibility ("ϵ-stationary point" hereafter). When g is a compactly supported convex function, Hajinezhad and Hong [18] proposed a perturbed prox-PDA that achieves an iteration complexity of $\mathcal{O}(\epsilon^{-4})$. Zeng, Yin, and Zhou [59] proposed a Moreau envelope ALM for handling a general weakly convex objective function F, which achieves the $\mathcal{O}(\epsilon^{-2})$ iteration complexity. Authors of [53] proposed two variants of ALM with $\mathcal{O}(\epsilon^{-2})$ iteration estimates when F is a difference-of-convex function.

In contrast to previously mentioned works where iteration complexities are measured by the number of times a (proximal) augmented Lagrangian relaxation is solved, the following works study iteration complexities in terms of *first-order oracles,* i.e., the number of proximal gradient steps. Melo, Monteiro, and Wang [39] applied an accelerated composite gradient (ACG) method [2] to solve each proximal ALM subproblem and showed that an ϵ-stationary point can be found in $\tilde{\mathcal{O}}(\epsilon^{-3})$[1] ACG iterations, which can be further reduced to $\tilde{\mathcal{O}}(\epsilon^{-5/2})$ with mildly stronger assumptions. Later, Kong, Melo, and Monteiro [29] embedded this inner acceleration scheme into a proximal ALM with full dual multiplier update and derived an iteration complexity of $\tilde{\mathcal{O}}(\epsilon^{-3})$ ACG iterations. Zhang and Luo proposed a single-looped proximal ALM and established an $\mathcal{O}(\epsilon^{-2})$ iteration estimate when g is an indicator function of a hypercube [60] or a polyhedron [61].

Works [28, 32] study the setting where convex nonlinear constraints $h_I(x) \leq 0$ are explicitly present. When F is weakly convex and $h_E(x)$ is affine, Li and Xu [32] combined an inexact ALM and a quadratic penalty method, and they showed that an ϵ-stationary point can be found in $\tilde{\mathcal{O}}(\epsilon^{-5/2})$ adaptive accelerated proximal gradient (APG) steps. Kong, Melo, and Monteiro [28] studied a more general setting: convex

[1] The notation $\tilde{\mathcal{O}}$ hides logarithmic dependence on ϵ^{-1}.

nonlinear constraints take the form $-h_I(x) \in \mathcal{K}$, where $\mathcal{K}$ is a closed convex cone. Under the same inner acceleration scheme as in [29], they showed that the proposed proximal ALM finds an ϵ-stationary solution in $\tilde{\mathcal{O}}(\epsilon^{-3})$ ACG iterations.

2.1.2. Nonconvex constraints. Works [31, 33, 48, 56] consider nonlinear and nonconvex constraints. Sahin et al. [48] studied problem (2.1) with only the equality constraints $h_E(x) = 0$, and proposed a double-looped inexact ALM (iALM), where the augmented Lagrangian relaxation is solved by the accelerated gradient method in [14], and then the dual variable is updated with a small step size, which ensures that the sequence of dual variables is uniformly bounded. Assuming a technical *regularity condition* that provides a convenient workaround to control primal infeasibility using dual infeasibility, the proposed iALM achieves an $\tilde{\mathcal{O}}(\epsilon^{-4})$ iteration complexity.[2] Assuming the same regularity condition, Li et al. [31] later improved the iteration complexity to $\tilde{\mathcal{O}}(\epsilon^{-3})$, which is achieved through a triple-looped iALM.

For problems where the aforementioned regularity condition is not satisfied, the inexact proximal point penalty method proposed by Lin, Ma, and Xu [33] finds an ϵ-stationary solution in $\tilde{\mathcal{O}}(\epsilon^{-4})$ adaptive APG steps under the requirement that the initial point is feasible. Xie and Wright [56] proposed a proximal ALM that finds an ϵ-stationary solution in $\mathcal{O}(\epsilon^{-5.5})$ Newton-CG iterations.

2.2. ADMM. The ADMM was proposed in the mid-1970s [13, 15], while the underlying idea has deep roots in maximal monotone operator theory [11] and numerical methods for partial differential equations [10, 41]. Commonly regarded as a variant of ALM, ADMM solves the augmented Lagrangian relaxation by alternately optimizing through blocks of variables and in this way subproblems become decoupled. Such a feature has gained ADMM considerable attention in distributed optimization [6, 36, 51]. The convergence of ADMM with two block variables is proved for convex optimization problems [11, 12, 13, 15] and a convergence rate of $\mathcal{O}(1/K)$ is established [19, 20, 40], where K is the iteration index. See also [7, 9, 16, 23, 34] on convex multiblock ADMM.

In recent years, researchers have extended the ADMM framework to solve nonconvex multiblock problem (1.1), where $h_i(x_i) = A_i x_i$ is affine for all $i \in [p]$ [17, 24, 27, 38, 37, 55]. The asymptotic convergence and an iteration complexity of $\mathcal{O}(\epsilon^{-2})$ are established based on two crucial conditions on the problem data: (a) $g_p = 0$ and (b) the column space of A_p contains the column space of the concatenated matrix $[A_1, \ldots, A_{p-1}]$. Condition (a) provides a way to control dual iterates by primal iterates, while condition (b) is required for ADMM to locate a feasible solution in the limit. See also [54, Table 1] for a summary of other assumptions. These two assumptions are almost necessary for the convergence of nonconvex ADMM. Namely, when either one of the two assumptions fails to hold, divergent examples have been found.

There are also several works investigating the convergence of nonconvex ADMM without conditions (a) and (b). In particular, Jiang et al. [27] proposed to run ADMM on a penalty relaxation of (1.1). Sun and Sun [54, 52] proposed a two-level framework that embeds a structured three-block ADMM inside an ALM. For weakly convex minimization over affine constraints, works [59, 60] demonstrate two ADMM variants that do not require assumptions (a) or (b).

Zhu, Zhao, and Zhang [62] considered nonlinear coupling constraints of the form $h(x_1) + Bx_2 = 0$. Assuming condition (a) and a straightforward extension of condition (b), i.e., the range of the nonlinear mapping h belongs to the column space of the matrix B, the authors derived the $\mathcal{O}(\epsilon^{-2})$ iteration complexity.

[2]The complexity presented in [48] is claimed to be wrong and corrected to $\tilde{\mathcal{O}}(\epsilon^{-4})$ by [31].

3. SDD ADMM.

3.1. Assumptions and stationarity. In this subsection, we formally state our assumptions on the problem data and define stationarity for problem (1.1).

Assumption 3.1.

1. For $i \in [p]$, the function $g_i : \mathbb{R}^{n_i} \to \overline{\mathbb{R}}$ is proper and lower-semicontinuous with an effective domain denoted by $X_i = \operatorname{dom} g_i := \{x \in \mathbb{R}^n \mid g_i(x) < +\infty\}$. Moreover, the proximal oracle of g_i is available, i.e., given $z_i \in \mathbb{R}^{n_i}$ and a sufficiently large constant $\eta > 0$, we can solve the following problem:
$$\min_{x_i \in \mathbb{R}^{n_i}} g_i(x_i) + \frac{\eta}{2}\|x_i - z_i\|^2. \tag{3.1}$$
Denote $g(x) = \sum_{i=1}^p g_i(x_i)$ and $X = \prod_{i=1}^p X_i$.
2. The function $f : \mathbb{R}^n \to \mathbb{R}$ has Lipschitz gradient over X, i.e., there exists a positive constant L_f such that $\|\nabla f(x) - \nabla f(z)\| \le L_f \|x - z\|$ for any $x, z \in X$.
3. The mapping $h : \mathbb{R}^n \to \mathbb{R}^m$ is given by $h(x) = \sum_{i=1}^p h_i(x_i)$, where $h_i : \mathbb{R}^{n_i} \to \mathbb{R}^m$ is continuously differentiable over X_i, and there exist positive constants M_{h_i}, K_{h_i}, J_{h_i}, and L_{h_i} such that for all $i \in [p]$ and $x_i, z_i \in X_i$, we have
$$\max_{x_i \in X_i} \|h_i(x_i)\| \le M_{h_i}, \quad \|h_i(x_i) - h_i(z_i)\| \le K_{h_i}\|x_i - z_i\|, \tag{3.2a}$$
$$\max_{x_i \in X_i} \|\nabla h_i(x_i)\| \le J_{h_i}, \|\nabla h_i(x_i) - \nabla h_i(z_i)\| \le L_{h_i}\|x_i - z_i\|, \tag{3.2b}$$
where $\nabla h_i(x_i) = [\nabla h_{i1}(x_i), \ldots, \nabla h_{im}(x_i)] \in \mathbb{R}^{n_i \times m}$, and $\|\cdot\|$ denotes the Euclidean norm for vectors or the induced norm for matrices.
4. The following constants are finite:
$$\overline{\mathcal{P}} := \sup_{x \in X} f(x) + \sum_{i=1}^p g_i(x_i) < +\infty \text{ and} \tag{3.3a}$$
$$\underline{\mathcal{P}} := \inf_{x \in X} f(x) + \sum_{i=1}^p g_i(x_i) > -\infty. \tag{3.3b}$$

For ease of later presentation, let
$$M_h := \sum_{i=1}^p M_{h_i},\ K_h := \max_{i \in [p]}\{K_{h_i}\},\ J_h := \max_{i \in [p]}\{J_{h_i}\}, \text{ and } L_h := \max_{i \in [p]}\{L_{h_i}\}. \tag{3.4}$$

Remark 3.2. We make some remarks regarding the above assumptions.

1. We allow g_i to be nonconvex as long as its proximal oracle is available. Without loss of generality, we may assume any $\eta \ge L_f$ suffices to carry out the minimization in (3.1) exactly.
2. Any g_i of the form $\delta_{X_i} + \tilde{g}_i$, where X_i is a compact set and $\tilde{g}_i$ is continuous over X_i ensures that constants in (3.2) and (3.3) are well-defined. However, we do not explicitly require the compactness of X_i's because many nonconvex g_i's such as smoothly clipped absolute deviation (SCAD), minimax concave penalty (MCP), and capped-ℓ_1 are defined over $\mathbb{R}^{n_i}$ and uniformly bounded from above, and nonlinear mappings including the sine, cosine, arctangent, and sigmoid functions can ensure (3.2) without X_i's being compact. In addition, it is possible to further relax condition (3.3a); see Remark 3.9.

3. Suppose each $h_{ij}:\mathbb{R}^n\to\mathbb{R}$ satisfies that $\|\nabla h_{ij}(x_i)-\nabla h_{ij}(z_i)\|\le L_{h_{ij}}\|x_i-z_i\|$ and $\|\nabla h_{ij}(x_i)\|\le J_{h_{ij}}$ for all $x_i,z_i\in X_i$, then we can obtain the following estimates: $L_{h_i}=\sqrt{m}\max_{j\in[m]}L_{h_{ij}}$ and $J_{h_i}=K_{h_i}=\sqrt{m}\max_{j\in[m]}J_{h_{ij}}$.

We define approximate stationarity as follows.

DEFINITION 3.3 (approximate stationary point). *Given $\epsilon>0$, we say $x\in X$ is an ϵ-stationary point of problem* (1.1) *if there exists $\lambda\in\mathbb{R}^m$ such that*

$$\max_{i\in[p]}\Big\{\operatorname{dist}\Big(-\nabla_i f(x)-\nabla h_i(x_i)\lambda,\partial g_i(x_i)\Big)\Big\}\le\epsilon \text{ and } \|h(x)\|\le\epsilon,$$

where $\partial g_i(x_i)$ denotes the general subdifferential of g_i at $x_i\in X_i$ [46, *Definition* 8.3].

3.2. The proposed algorithm. Recall the augmented Lagrangian function in (1.2). We also separate out the smooth components in $\mathcal{L}_\rho(x,\mu)$ as

$$\mathcal{K}_\rho(x,\mu):=\mathcal{L}_\rho(x,\mu)-\sum_{i=1}^p g_i(x_i)=f(x)+\langle\mu,h(x)\rangle+\frac{\rho}{2}\|h(x)\|^2. \tag{3.5}$$

Notice that for $i\in[p]$, $\nabla_{x_i}\mathcal{K}_\rho(x,\mu)=\nabla_i f(x)+\nabla h_i(x_i)(\mu+\rho h(x))$. It can be verified that $\nabla_{x_i}\mathcal{K}_\rho(x,\mu)$ is Lipschitz. The calculation is straightforward and hence omitted.

LEMMA 3.4. *For any $i\in[p]$, $x_i,z_i\in X_i$, and fixed $x_j\in X_j$ with $j\ne i$, we have*

$$\|\nabla_{x_i}\mathcal{K}_\rho(x_{<i},x_i,x_{>i},\mu)-\nabla_{x_i}\mathcal{K}_\rho(x_{<i},z_i,x_{>i},\mu)\|\le\operatorname{Lip}(\mu,\rho)\|x_i-z_i\|,$$

where

$$\operatorname{Lip}(\mu,\rho):=L_f+\|\mu\|L_h+\rho(J_hK_h+M_hL_h). \tag{3.6}$$

Lemma 3.4 allows us to update each x_i via a single proximal gradient step. The SDD ADMM (SDD-ADMM) is presented in Algorithm 3.1.

Remark 3.5. We make some remarks on Algorithm 3.1.

1. We update x_i^{k+1}'s through p proximal gradient updates, and then obtain μ^{k+1} by minimizing $\mathcal{P}(x^{k+1},\mu)$ plus a proximal term as in (3.8). Due to the strong convexity of $\mathcal{P}$ in μ, τ is allowed to be 0.

Algorithm 3.1. `SDD-ADMM`.

1: **Input** $x^0\in X$, $\rho>0$, $\omega\ge4$, $\theta>1$, $\tau\ge0$;
2: **initialize** $\mu^0=0\in\mathbb{R}^m$;
3: **for** $k=0,1,2,\cdots$ **do**
4: **for** $i=1,2,\cdots,p$ **do**
5: update x_i^{k+1} through a proximal gradient step:

$$x_i^{k+1}=\operatorname*{argmin}_{x_i\in\mathbb{R}^{n_i}} g_i(x_i)+\langle\nabla_{x_i}\mathcal{K}_\rho(x_{<i}^{k+1},x_{\ge i}^k,\mu^k),x_i-x_i^k\rangle+\frac{\theta\operatorname{Lip}(\mu^k,\rho)}{2}\|x_i-x_i^k\|^2; \tag{3.7}$$

6: **end for**
7: update μ^{k+1} by

$$\mu^{k+1}=\operatorname*{argmin}_{\mu\in\mathbb{R}^m}\mathcal{P}(x^{k+1},\mu)+\frac{\tau\omega}{2\rho}\|\mu-\mu^k\|^2=\frac{1}{1+\tau}\left(\tau\mu^k-\omega^{-1}\rho h(x^{k+1})\right); \tag{3.8}$$

8: **end for**

2. If we perturb the proximal term by a specific *linear* term to cancel out the inner product in $\mathcal{P}(x^{k+1},\mu)$, and update μ^{k+1} as

$$\mu^{k+1} = \operatorname*{argmin}_{\mu\in\mathbb{R}^m} \mathcal{P}(x^{k+1},\mu) + \frac{\tau\omega}{2\rho}\|\mu-\mu^k\|^2 + \langle -h(x^{k+1}),\mu\rangle, \tag{3.9}$$

then, since we initialize with $\mu^0 = 0$, (3.9) recovers the penalty method, i.e., $\mu^{k+1} = 0$ for all $k \in \mathbb{N}$. When $\tau > 0$, (3.9) is also equivalent to the SDD update with a different proximal center:

$$\mu^{k+1} = \operatorname*{argmin}_{\mu\in\mathbb{R}^m} \mathcal{P}(x^{k+1},\mu) + \frac{\tau\omega}{2\rho}\left\|\mu - \left(\mu^k + \frac{\rho}{\omega\tau}h(x^{k+1})\right)\right\|^2. \tag{3.10}$$

It is interesting that the new proximal center is a dual ascent iterate, which recovers the penalty method in combination with the SDD framework.

3. The lower bound of the parameter ω is chosen to be 4 mainly for the ease of analysis, e.g., in Lemma 3.10. Other values are possible as well.

When $p = 1$, we call the algorithm SDD-ALM. One key observation is that, although x_i's are updated in a Gauss–Seidel fashion in Algorithm 3.1, we can also apply a Jacobi-type update, i.e., replace $\nabla_{x_i}\mathcal{K}_\rho(x^{k+1}_{<i}, x^k_{\geq i}, \mu^k)$ by $\nabla_{x_i}\mathcal{K}_\rho(x^k,\mu^k)$ in (3.7) and then solve the p subproblems in parallel. This is a special case when we treat $x = [x_1^\top,\dots,x_p^\top]^\top$ as a single block variable and apply SDD-ALM, where decomposition is achieved within this single block update and assembles a Jacobi sweep. Such a feature might be favored in a distributed optimization setting. We present the convergence of SDD-ADMM in the form of Algorithm 3.1 in the next subsection.

3.3. Convergence analysis. In this subsection, we analyze the convergence of Algorithm 3.1 under dual updates (3.8) and (3.9) in a unified framework. We first show that when the initial point is almost feasible, SDD-ADMM finds an ϵ-stationary solution in $\mathcal{O}(\epsilon^{-4})$ iterations. Then under an additional assumption regarding h and g, we can further improve the complexity by one or even two orders of magnitude to $\mathcal{O}(\epsilon^{-3})$ and $\mathcal{O}(\epsilon^{-2})$, respectively.

Although our analysis encompasses both the SDD update (3.8) and the penalty method update (3.9), the technical claims presented below are stated within the context of the SDD update (3.8), which explicitly involves the dual variable μ. This choice is motivated by our initial goal of minimizing the augmented Lagrangian function in both the primal and dual variables. However, throughout each proof, we accommodate the analysis to incorporate the penalty method update (3.9) whenever necessary. By adopting this approach, we ensure that our results are applicable to both the SDD and the penalty method, providing a comprehensive understanding of their convergence properties.

3.3.1. An $\mathcal{O}(\epsilon^{-4})$ iteration complexity. Due to that fact that ∇f is Lipschitz, we have $\|\nabla_i f(x_{<i},x_i,x_{>i}) - \nabla_i f(x_{<i},z_i,x_{>i})\| \leq L_f\|x_i - z_i\|$ for any $i \in [p]$, $x_i, z_i \in X_i$, and fixed x_j for $j \neq i$; this will be invoked several times in the analysis. We first show that the sequence $\{\mathcal{P}(x^k,\mu^k)\}_{k\in\mathbb{N}}$ is nonincreasing.

LEMMA 3.6 (one-step progress of SDD-ADMM). *Suppose Assumption* 3.1 *holds. For all* $k \in \mathbb{N}$,

$$\begin{aligned}&\mathcal{P}(x^k,\mu^k) - \mathcal{P}(x^{k+1},\mu^{k+1})\\ &\geq \left(\frac{\theta-1}{2}\right)\mathrm{Lip}(\mu^k,\rho)\sum_{i=1}^p \|x_i^{k+1}-x_i^k\|^2 + \left(\tau+\frac{1}{2}\right)\frac{\omega}{\rho}\|\mu^{k+1}-\mu^k\|^2.\end{aligned}$$

Proof. We first establish the descent in x. Let $i \in [p]$, then it holds that

$$\begin{aligned}
&\mathcal{P}(x_{\le i}^{k+1}, x_{>i}^{k}, \mu^k) = \sum_{j\le i} g_j(x_j^{k+1}) + \sum_{j>i} g_j(x_j^k) + \mathcal{K}_\rho(x_{\le i}^{k+1}, x_{>i}^k, \mu^k) + \frac{\omega}{2\rho}\|\mu^k\|^2 \\
&\le \sum_{j\le i} g_j(x_j^{k+1}) + \sum_{j>i} g_j(x_j^k) + \mathcal{K}_\rho(x_{<i}^{k+1}, x_{\ge i}^k, \mu^k) + \langle \nabla_{x_i}\mathcal{K}_\rho(x_{<i}^{k+1}, x_{\ge i}^k, \mu^k), x_i^{k+1} - x_i^k\rangle \\
&\quad + \frac{\mathrm{Lip}(\mu^k,\rho)}{2}\|x_i^{k+1} - x_i^k\|^2 + \frac{\omega}{2\rho}\|\mu^k\|^2 \\
&\le \sum_{j<i} g_j(x_j^{k+1}) + \sum_{j\ge i} g_j(x_j^k) + \mathcal{K}_\rho(x_{<i}^{k+1}, x_{\ge i}^k, \mu^k) + \frac{\omega}{2\rho}\|\mu^k\|^2 \\
&\quad - \Big(\frac{\theta-1}{2}\Big)\mathrm{Lip}(\mu^k,\rho)\|x_i^{k+1} - x_i^k\|^2 \\
&= \mathcal{P}(x_{<i}^{k+1}, x_{\ge i}^k, \mu^k) - \left(\frac{\theta-1}{2}\right)\mathrm{Lip}(\mu^k,\rho)\|x_i^{k+1} - x_i^k\|^2,
\end{aligned}$$

where the first inequality is due to $\nabla_{x_i}\mathcal{K}_\rho(x,\mu^k)$ being Lipschitz and the second inequality is due to the optimality of x_i^{k+1} in (3.7). Summing the above inequality from $i=1$ to p, we have

$$\mathcal{P}(x^k,\mu^k) - \mathcal{P}(x^{k+1},\mu^k) \ge \left(\frac{\theta-1}{2}\right)\mathrm{Lip}(\mu^k,\rho)\sum_{i=1}^{p}\|x_i^{k+1} - x_i^k\|^2. \tag{3.11}$$

Next we derive the descent in μ. The strong convexity of the objective in (3.8) implies

$$\mathcal{P}(x^{k+1},\mu^k) - \mathcal{P}(x^{k+1},\mu^{k+1}) \ge \left(\tau + \frac{1}{2}\right)\frac{\omega}{\rho}\|\mu^{k+1} - \mu^k\|^2. \tag{3.12}$$

In view of (3.9), the above inequality holds as well since $\mu^k = \mu^{k+1} = 0$. Combining (3.11) and (3.12) proves the lemma. ∎

Remark 3.7. We assume that the proximal mapping (3.1) of g_i can be carried out exactly, which can be satisfied for many nonconvex functions such as SCAD, MCP, capped-ℓ_1, and indicator functions of sphere or annulus constraints. This assumption is mainly used to establish the descent property of $\mathcal{P}(x_{<i}^{k+1}, \cdot, x_{>i}^k, \mu^k)$, whereas the global optimality of x_i^{k+1} in (3.7) is not necessary. As an alternative, we may directly assume a descent oracle on g_i: we can find a stationary point x_i^{k+1} of (3.7) such that $\mathcal{P}(x_{<i}^{k+1}, x_i^{k+1}, x_{>i}^k, \mu^k) \le \mathcal{P}(x_{<i}^{k+1}, x_i^k, x_{>i}^k, \mu^k) - \nu\|x_i^{k+1} - x_i^k\|^2$ for some $\nu > 0$. See also Assumption B.3. This is in general more realistic when g_i is highly complicated and reasonable if some nonconvex solver can be warm-started.

Next we show that the sequence $\{\mathcal{P}(x^k,\mu^k)\}_{k\in\mathbb{N}}$ is bounded from below; consequently, we can further control $\|h(x^k)\|$ and $\|\mu^k\|$. To this end, we require the infeasibility of the initial point x^0 to be controlled in the following sense.

Assumption 3.8. There exists a constant $C \ge 0$ such that for any finite $\alpha > 0$, we can find an initial point $x^0 \in X$ such that $\|h(x^0)\|^2 \le \frac{C}{\alpha}$.

Remark 3.9. Assumption 3.8 is a slight relaxation of the requirement that $h(x^0) = 0$. As we will see in Theorem 3.12, in order to find an ϵ-stationary solution, we will need to satisfy Assumption 3.8 with $\alpha = \rho = \Theta(\epsilon^{-2})$ and hence the initial point is required to be almost feasible, i.e., $\|h(x^0)\| = \mathcal{O}(\epsilon)$. It can be satisfied by solving a convex program when X is convex and h is affine, or when there exists

$i \in [p]$ such that $h_i(x_i) = A_i x_i - b$, where A_i has full row rank. We also note that this (near-)feasibility assumption on the initial point is commonly adopted in the literature to establish iteration complexity estimates for nonlinear programs [5, 33, 56]. In addition, it suffices to replace $\overline{\mathcal{P}}$ defined in (3.3a) by any finite upper bound on $f(x^0) + g(x^0)$ in case f is not bounded from above over X.

LEMMA 3.10 (bounds on dual variable and primal residual). *Suppose Assumptions* 3.1 *and* 3.8 *hold. Recall* $\overline{\mathcal{P}}$ *and* $\underline{\mathcal{P}}$ *from* (3.3)*, constant* C *from Assumption* 3.8*, and further define*

$$\Delta := \overline{\mathcal{P}} - \underline{\mathcal{P}} + \frac{C}{2}, \tag{3.13}$$

which is a constant independent of the penalty ρ*. Then* $\mathcal{P}(x^k, \mu^k) \geq \underline{\mathcal{P}}$ *for all* $k \in \mathbb{N}$*. Moreover, it holds that*

$$\|h(x^k)\| \leq \left(\frac{4\Delta}{\rho}\right)^{1/2} \quad \text{and} \quad \|\mu^k\| \leq (\rho\Delta)^{1/2}. \tag{3.14}$$

Proof. Let x^0 be an initial point supplied to SDD-ADMM satisfying Assumption 3.8 with $\alpha = \rho$. Moreover, since $\mu^0 = 0$, we have

$$\mathcal{P}(x^0, \mu^0) = f(x^0) + \sum_{i=1}^{p} g_i(x_i^0) + \frac{\rho}{2}\|h(x^0)\|^2 \leq \overline{\mathcal{P}} + \frac{C}{2}.$$

By Lemma 3.6, for all $k \in \mathbb{N}$, we have $\mathcal{P}(x^0, \mu^0)$ is greater than

$$\begin{aligned}
&\mathcal{P}(x^k, \mu^k) = f(x^k) + \sum_{i=1}^{p} g_i(x_i^k) + \langle \mu^k, h(x^k)\rangle + \frac{\rho}{2}\|h(x^k)\|^2 + \frac{\omega}{2\rho}\|\mu^k\|^2 \\
&\geq \inf_{x\in X}\left\{ f(x) + \sum_{i=1}^{p} g_i(x_i)\right\} + \frac{\rho}{4}\|h(x^k)\|^2 + \frac{1}{\rho}\|\mu^k\|^2 = \underline{\mathcal{P}} + \frac{\rho}{4}\|h(x^k)\|^2 + \frac{1}{\rho}\|\mu^k\|^2 \geq \underline{\mathcal{P}},
\end{aligned}$$

where the second inequality is due to $\langle \mu^k, h(x^k)\rangle \geq -\frac{\rho}{4}\|h(x^k)\|^2 - \frac{1}{\rho}\|\mu^k\|^2$ and $\omega \geq 4$. The above inequality further gives the bounds in (3.14). ∎

Lemma 3.10 holds under both dual updates (3.8) and (3.9). If (3.9) is performed, then (3.14) can be improved to $\|h(x^k)\| \leq (2\Delta/\rho)^{1/2}$ and $\|\mu^k\| = 0$. Though x_i^{k+1} is obtained by a single proximal gradient step, it still is an approximate stationary solution in the following sense.

LEMMA 3.11 (bound on dual residual). *Suppose Assumption* 3.1 *holds. For all* $k \in \mathbb{N}$ *and* $i \in [p]$*,*

$$\operatorname{dist}\Big(-\nabla_i f(x^{k+1}) - \nabla h_i(x_i^{k+1})\tilde{\mu}^{k+1}, \partial g_i(x_i^{k+1})\Big) \leq (\theta+1)\operatorname{Lip}(\mu^k, \rho)\sum_{j\geq i}\|x_j^{k+1} - x_j^k\|,$$

where $\tilde{\mu}^{k+1} := \mu^k + \rho h(x^{k+1})$.

Proof. The update of x_i^{k+1} gives $\xi_i^{k+1} \in \nabla_i f(x^{k+1}) + \partial g_i(x_i^{k+1}) + \nabla h_i(x_i^{k+1})\tilde{\mu}^{k+1}$, where

$$\begin{aligned}
\xi_i^{k+1} :=\ & \nabla_i f(x^{k+1}) - \nabla_i f(x_{<i}^{k+1}, x_{\geq i}^k) - \theta\operatorname{Lip}(\mu^k, \rho)(x_i^{k+1} - x_i^k) \\
& + \nabla h_i(x_i^{k+1})(\mu^k + \rho h(x^{k+1})) - \nabla h_i(x_i^k)\left(\mu^k + \rho\sum_{j<i} h_i(x_j^{k+1}) + \rho\sum_{j\geq i} h_j(x_j^k)\right).
\end{aligned}$$

The last two terms in the definition of ξ_i^{k+1} can be bounded by

$$\begin{aligned}&\|(\nabla h_i(x_i^{k+1}) - \nabla h_i(x_i^k))(\mu^k + \rho h(x^{k+1})\| + \rho\|\nabla h_i(x_i^k)\| \sum_{j\geq i} \|h_j(x_j^{k+1}) - h_j(x_j^k)\| \\ &\leq (L_h\|\mu^k\| + \rho L_h M_h)\|x_i^{k+1} - x_i^k\| + \rho J_h K_h \sum_{j\geq i} \|x_j^{k+1} - x_j^k\|;\end{aligned}$$

by the smoothness of f and the definition of $\mathrm{Lip}(\mu^k, \rho)$ in (3.6), $\|\xi_i^k\|$ is bounded by

$$\begin{aligned}&(L_f + \rho J_h K_h) \sum_{j\geq i} \|x_j^{k+1} - x_j^k\| + (L_h\|\mu^k\| + \rho L_h M_h + \theta \mathrm{Lip}(\mu^k, \rho))\|x_i^{k+1} - x_i^k\| \\ &\leq (\theta + 1)\mathrm{Lip}(\mu^k, \rho) \sum_{j\geq i} \|x_j^{k+1} - x_j^k\|.\end{aligned}$$

This completes the proof. ∎

With the help of the previous lemmas, we are now ready to present an iteration complexity upper bound for SDD-ADMM.

THEOREM 3.12. *Suppose Assumptions* 3.1 *and* 3.8 *hold, and let* $\epsilon > 0$. *Recall parameters* (M_h, K_h, J_h, L_h) *from* (3.4) *and* Δ *in* (3.13), *and define constants*

$$\kappa_1 := J_h K_h + M_h L_h,\ \kappa_2 := L_h\sqrt{\Delta} + \kappa_1 + 1. \tag{3.15}$$

Further choose $\rho \geq \max\{1, L_f, 4\Delta\epsilon^{-2}\}$, *and let* $x^0 \in X$ *be an initial point satisfying Assumption* 3.8 *with* $\alpha = \rho$. *Then SDD-ADMM with input* $(x^0, \rho, \omega, \theta, \tau)$ *finds an* ϵ*-stationary solution of* (1.1) *in at most* $K(\rho)$ *iterations, where*

$$K(\rho) := \left\lceil \frac{2p\Delta(\theta+1)^2\kappa_2^2\rho}{(\theta-1)\kappa_1\epsilon^2} \right\rceil = \mathcal{O}(\rho\epsilon^{-2}). \tag{3.16}$$

In particular, if we choose $\rho = \Theta(\epsilon^{-2})$, *then* $K(\rho) = \mathcal{O}(\epsilon^{-4})$.

Proof. We first show that $\mathrm{Lip}(\mu^k, \rho) = \Theta(\rho)$. Since $\rho \geq \max\{1, L_f\}$, by the second inequality in (3.14) of Lemma 3.10, we have $\|\mu^k\| \leq \sqrt{\rho\Delta} \leq \rho\sqrt{\Delta}$ and

$$\rho\kappa_1 \leq \mathrm{Lip}(\mu^k, \rho) = L_f + \|\mu^k\|L_h + \rho(J_h K_h + M_h L_h) \leq \rho\kappa_2. \tag{3.17}$$

The above lower bound of $\mathrm{Lip}(\mu^k, \rho)$ and Lemma 3.6 together give

$$\frac{(\theta-1)\kappa_1}{2}\rho\sum_{i=1}^p \|x_i^{k+1} - x_i^k\|^2 \leq \mathcal{P}(x^k, \mu^k) - \mathcal{P}(x^{k+1}, \mu^{k+1}).$$

Summing the above inequality from $k = 0$ to some positive index $K-1$, we have

$$\frac{(\theta-1)\kappa_1}{2}\rho\sum_{k=0}^{K-1}\sum_{i=1}^p \|x_i^{k+1} - x_i^k\|^2 \leq \Delta. \tag{3.18}$$

As a result, there exists an index $0 \leq \bar{k} \leq K-1$ such that

$$\sum_{i=1}^p \|x_i^{\bar{k}+1} - x_i^{\bar{k}}\| \leq \sqrt{p}\left(\sum_{i=1}^p \|x_i^{\bar{k}+1} - x_i^{\bar{k}}\|^2\right)^{1/2} \leq \left(\frac{2p\Delta}{\rho(\theta-1)\kappa_1 K}\right)^{1/2}. \tag{3.19}$$

By (3.14) in Lemma 3.10 and the choice that $\rho \geq 4\Delta\epsilon^{-2}$, we have $\|h(x^{\bar{k}+1})\| \leq \epsilon$. Moreover, recall $\tilde{\mu}^{\bar{k}+1} = \mu^{\bar{k}} + \rho h(x^{\bar{k}+1})$; Lemma 3.11, the upper bound in (3.17), and (3.19) imply that

$$\begin{aligned}&\max_{i\in[p]}\Big\{\mathrm{dist}\Big(-\nabla_i f(x^{\bar{k}+1}) - \nabla h_i(x^{\bar{k}+1})\tilde{\mu}^{\bar{k}+1}, \partial g_i(x_i^{\bar{k}+1})\Big)\Big\}\\ &\leq (\theta+1)\mathrm{Lip}(\mu^{\bar{k}},\rho)\sum_{i=1}^{p}\|x_i^{\bar{k}+1} - x_i^{\bar{k}}\| \leq (\theta+1)\kappa_2\rho\left(\frac{2p\Delta}{\rho(\theta-1)\kappa_1 K}\right)^{1/2} \leq \epsilon,\end{aligned}$$

where the last inequality holds by the upper bound $K = K(\rho)$ in (3.16). This completes the proof. ∎

In view of Lemma 3.10 and Theorem 3.12, the primal infeasibility is bounded by $\sqrt{4\Delta/\rho}$ while the dual infeasibility can be reduced to ϵ in $\mathcal{O}(\rho\epsilon^{-2})$ iterations. Such measures can be informative if different primal and dual tolerances are preferred.

3.3.2. Improve iteration complexity to $\mathcal{O}(\epsilon^{-3})$ and $\mathcal{O}(\epsilon^{-2})$. Next we show that under an additional technical assumption, we can further improve the iteration complexity of SDD-ADMM. Given $r > 0$ and $i \in [p]$, define

$$X(r) := \{x \in X \mid \|h(x)\| \leq r\}, \tag{3.20}$$

$$X_i(r) := \{x_i \in X_i \mid (x_{\neq i}, x_i) \in X(r) \text{ for some } x_j \in X_j,\ j \in [p]\setminus\{i\}\}. \tag{3.21}$$

By Assumption 3.8, we know that $X(r)$ is nonempty for any $r > 0$ and thus its projection $X_i(r)$ is also nonempty. Now we further make the following assumption.

Assumption 3.13. There exist $i \in [p]$, $(r,\sigma) \in \mathbb{R}^2_{++}$, and $(M_g, \nabla_f) \in \mathbb{R}^2_+$ such that

$$\sigma\|\mu\| \leq \mathrm{dist}(-\nabla h_i(x_i)\mu, \partial g_i(x_i)) + M_g \quad \forall \mu \in \mathbb{R}^m, x_i \in X_i(r), \text{ and} \tag{3.22}$$

$$\sup_{x\in X(r)} \|\nabla_i f(x)\| \leq \nabla_f. \tag{3.23}$$

Remark 3.14. We make some comments regarding Assumption 3.13.

1. Suppose that $\nabla h_i(x_i)$ has full rank over $X_i(r)$, and their smallest singular values are bounded away from zero, i.e.,

$$\inf_{x_i\in X_i(r)} \sigma_{\min}(\nabla h_i(x_i)) > 0. \tag{3.24}$$

   In Appendix A, we show that broad classes of g_i functions can ensure condition (3.22) with the help of (3.24) or a similar constraint qualification. In particular, g_i can be (Example A.1) a possibly nonconvex Lipschitz function, (Example A.2) a function of the form $\delta_{X_i} + \tilde{g}_i$, where X_i is a sufficiently large full-dimensional closed convex set and $\tilde{g}_i$ is continuous and convex over X_i, or (Example A.3) an indicator function of a set defined by continuously differentiable constraints satisfying a constraint qualification.
2. Clearly, $h_i(x_i) = Ax - b$ with full row rank always implies (3.24) as $\sigma_{\min}(A) > 0$. Even in this case, (3.22) is still weaker than conditions (a) and (b) commonly adopted in existing ADMM works (reviewed in section 2.2) as we allow the presence of some nonsmooth g_i.
3. Condition (3.23) is rather mild and can be satisfied under the boundedness of either $\nabla_i f$ or $X(r)$.
4. A reasonable direction to further weaken Assumption 3.13 is to restrict the regions of x_i and x on which (3.22) and (3.23) hold. For example, one can directly assume (3.22) and (3.23) hold on all algorithmic iterates $\{x^{k+1}\}_{k\in\mathbb{N}}$.

We further comment on condition (3.24). As a concrete example, consider $i = p = 1$ and $h(x) = x^\top x - R$ for some $R > 0$. Then given any $0 < r < R$, it holds that $X(r) \subset \{x \in \mathbb{R}^n \mid R - r \le x^\top x \le R + r\}$ and, hence,

$$\sigma_{\min}(\nabla h(x)) = 2\|x\| \ge 2(R-r)^{1/2} > 0$$

for all $x \in X(r)$. In nonlinear programs, this condition is closely related to the well-known linearly independence constraint qualification (LICQ) commonly assumed on KKT points. It is worth noting that our condition is primarily imposed on $X_i(0)$, the feasible region of x_i, which is justified by Sard's theorem. Consequently, we extend this condition to $X_i(r)$ through the continuity of the rank of ∇h_i. In the context of nonlinear and nonconvex constraints, existing algorithms such as those based on ALM/penalty methods [31, 48, 56, 33], sequential quadratic programs (SQPs) [3, 8], and proximal point methods (PPMs) [5, 35] all rely on specific regularity conditions to control the behavior of dual variables. In contrast, our condition (3.24) differs from those used in the literature. It not only generalizes the classic rank condition employed for the convergence of affine-constrained ADMM but also enables us to derive a novel first-order iteration complexity estimate of $\mathcal{O}(\epsilon^{-2})$.

With Assumption 3.13, we can derive new bounds on dual variables.

LEMMA 3.15. *Suppose Assumptions* 3.1, 3.8, *and* 3.13 *hold. Further define constants*

$$\kappa_3 := \frac{(\theta+1)\kappa_2}{\sigma}\sqrt{\frac{2p\Delta}{(\theta-1)\kappa_1}},\ \kappa_4 := \frac{\kappa_3 + (1+\tau)\sqrt{2\Delta\omega}}{3}. \tag{3.25}$$

Suppose that $\rho \ge \max\{1, L_f, 4\Delta/r^2\}$. *For any positive integer* $K > 0$, *there exists an index* $0 \le \bar{k} \le K-1$ *such that*

$$\|\tilde{\mu}^{\bar{k}+1}\| \le \kappa_3\sqrt{\frac{\rho}{K}} + \frac{\nabla_f + M_g}{\sigma} \ \text{ and } \ \|\mu^{\bar{k}}\| \le \kappa_4\sqrt{\frac{\rho}{K}} + \frac{\nabla_f + M_g}{3\sigma}, \tag{3.26}$$

where $\tilde{\mu}^{\bar{k}+1} = \mu^{\bar{k}} + \rho h(x^{\bar{k}+1})$.

Proof. By Lemma 3.10, the choice $\rho \ge 4\Delta/r^2$ ensures that $\{x^{k+1}\}_{k\in\mathbb{N}} \subset X(r)$, and hence Assumption 3.13 can be applied. Let $i \in [p]$ be the index specified in Assumption 3.13. By Lemma 3.11, there exists $\xi_i^{k+1} \in \mathbb{R}^{n_i}$ such that

$$\begin{aligned}
&\xi_i^{k+1} \in \nabla_i f(x^{k+1}) + \nabla h_i(x_i^{k+1})\tilde{\mu}^{k+1} + \partial g_i(x_i^{k+1}),\\
&\|\xi_i^{k+1}\| \le (\theta+1)\mathrm{Lip}(\mu^k,\rho)\sum_{j=1}^{p}\|x_j^{k+1} - x_j^k\|.
\end{aligned}$$

Hence, by Assumption 3.13 and the fact that $\mathrm{Lip}(\mu^k,\rho) \le \kappa_2\rho$ from (3.17), we have

$$\begin{aligned}
\sigma\|\tilde{\mu}^{k+1}\| &\le \mathrm{dist}(-\nabla h_i(x_i)\tilde{\mu}^{k+1}, \partial g_i(x_i^{k+1})) + M_g \le \|\xi_i^{k+1}\| + \|\nabla f_i(x^{k+1})\| + M_g\\
&\le (\theta+1)\kappa_2\rho\sum_{j=1}^{p}\|x_j^{k+1} - x_j^k\| + \nabla_f + M_g.
\end{aligned}$$

By (3.19) and the above inequality, we have

$$\|\tilde{\mu}^{\bar{k}+1}\| \le \frac{(\theta+1)\kappa_2\rho}{\sigma}\left(\frac{2p\Delta}{\rho(\theta-1)\kappa_1 K}\right)^{1/2} + \frac{\nabla_f + M_g}{\sigma}. \tag{3.27}$$

Hence the first inequality in (3.26) is proved for both dual updates (3.8) and (3.9). Since in the penalty method we have $\mu^k = 0$, it remains to prove the second inequality in (3.26) under the SDD update (3.8). By the SDD update and the definition of $\tilde{\mu}^{k+1}$, we have

$$0 = \rho h(x^{k+1}) + (1+\tau)\omega\mu^{k+1} - \tau\omega\mu^k = \tilde{\mu}^{k+1} + (1+\tau)\omega(\mu^{k+1} - \mu^k) + (\omega - 1)\mu^k,$$

which implies that

$$\|\mu^k\| \le \frac{1}{\omega - 1}\|\tilde{\mu}^{k+1}\| + \frac{(1+\tau)\omega}{\omega - 1}\|\mu^{k+1} - \mu^k\|. \tag{3.28}$$

Next we bound the two terms on the right-hand side of (3.28) at a specific index $\bar{k}$. By Lemma 3.6 and a similar argument as in the proof of Theorem 3.12, we have for any positive integer K, it holds that

$$\Delta \ge K\left(\frac{(\theta-1)\kappa_1}{2}\rho\sum_{i=1}^{p}\|x_i^{\bar{k}+1} - x_i^{\bar{k}}\|^2 + \frac{\omega}{2\rho}\|\mu^{\bar{k}+1} - \mu^{\bar{k}}\|^2\right),$$

where

$$\bar{k} := \operatorname*{argmin}_{k\in\{0,\cdots,K-1\}}\left\{\frac{(\theta-1)\kappa_1}{2}\rho\sum_{i=1}^{p}\|x_i^{k+1} - x_i^{k}\|^2 + \frac{\omega}{2\rho}\|\mu^{k+1} - \mu^{k}\|^2\right\},$$

then we know that (3.19) holds and

$$\|\mu^{\bar{k}+1} - \mu^{\bar{k}}\| \le \left(\frac{2\rho\Delta}{\omega K}\right)^{1/2}. \tag{3.29}$$

Combining (3.27), (3.28), and (3.29), we have

$$\|\mu^{\bar{k}}\| \le \frac{(\theta+1)\kappa_2\rho}{(\omega-1)\sigma}\left(\frac{2p\Delta}{\rho(\theta-1)\kappa_1 K}\right)^{1/2} + \frac{(1+\tau)\omega}{\omega-1}\left(\frac{2\rho\Delta}{\omega K}\right)^{1/2} + \frac{\nabla_f + M_g}{(\omega-1)\sigma}.$$

This completes the proof in view of (κ_3, κ_4) defined in (3.25) and the fact that $\omega \ge 4$. ∎

THEOREM 3.16. *Suppose Assumptions* 3.1, 3.8, *and* 3.13 *hold, and let* $\epsilon > 0$. *Recall* $(r, \sigma, \nabla_f, M_g)$ *is required in Assumption* 3.13, (κ_3, κ_4) *is defined in* (3.25), *and* $K(\rho)$ *is defined in* (3.16). *Choose* $\rho \ge \max\{1, L_f, 4\Delta/r^2\}$.

- *If* $\nabla_f + M_g > 0$, *then further let*

$$\rho \ge \left(\kappa_3 + \kappa_4 + \frac{4(\nabla_f + M_g)}{3\sigma}\right)\epsilon^{-1}, \tag{3.30}$$

  and let $x^0 \in X$ *be an initial point satisfying Assumption* 3.8 *with* $\alpha = \rho$. *Then SDD-ADMM with input* $(x^0, \rho, \omega, \theta, \tau)$ *finds an* ϵ*-stationary solution of* (1.1) *in at most*

$$K'(\rho) := \max\{\lceil\rho\rceil, K(\rho)\} \tag{3.31}$$

  iterations. In particular, if we choose $\rho = \Theta(\epsilon^{-1})$, *then* $K'(\rho) = \mathcal{O}(\epsilon^{-3})$.
- *If* $\nabla_f = M_g = 0$, *let* $x^0 \in X$ *be an initial point satisfying Assumption* 3.8 *with* $\alpha = \rho$. *Then SDD-ADMM with input* $(x^0, \rho, \omega, \theta, \tau)$ *finds an* ϵ*-stationary solution of* (1.1) *in at most*

$$K''(\rho) := \max\{\lceil(\kappa_3 + \kappa_4)^2\epsilon^{-2}\rceil, K(\rho)\} \tag{3.32}$$

  iterations. In particular, if we choose $\rho = \Theta(1)$, *then* $K''(\rho) = \mathcal{O}(\epsilon^{-2})$.

Proof. By a similar argument as in the proof of Theorem 3.12, at the index $\bar{k}$ specified in Lemma 3.15 with $K = K(\rho)$, the dual residual is bounded by ϵ, i.e.,

$$\max_{i\in[p]} \Big\{ \operatorname{dist}\Big(-\nabla_i f(x^{\bar{k}+1}) - \nabla h_i(x^{\bar{k}+1})\tilde{\mu}^{\bar{k}+1}, \partial g_i(x_i^{\bar{k}+1})\Big)\Big\} \le \epsilon.$$

It remains to show $\|h(x^{\bar{k}+1})\| \le \epsilon$. By the definition of $\tilde{\mu}^{\bar{k}+1}$, we have

$$\|h(x^{\bar{k}+1})\| \le \frac{1}{\rho}(\|\tilde{\mu}^{\bar{k}+1}\| + \|\mu^{\bar{k}}\|) \le \frac{1}{\rho}\left((\kappa_3+\kappa_4)\sqrt{\frac{\rho}{K}} + \frac{4(\nabla_f + M_g)}{3\sigma}\right), \tag{3.33}$$

where the second inequality is due to Lemma 3.15. Next we consider the two cases separately.

- If $\nabla_f + M_g > 0$, then (3.33) gives

  $$\|h(x^{\bar{k}+1})\| \le \frac{1}{\rho}\left(\kappa_3+\kappa_4 + \frac{4(\nabla_f + M_g)}{3\sigma}\right) \le \epsilon,$$

  where the first inequality holds with any $K \ge \rho$, and the second inequality is due to the choice of ρ in (3.30).
- If $\nabla_f = M_g = 0$, then (3.33) gives

  $$\|h(x^{\bar{k}+1})\| \le (\kappa_3+\kappa_4)\sqrt{\frac{1}{\rho K}} \le (\kappa_3+\kappa_4)\sqrt{\frac{1}{K}} \le \epsilon,$$

  which the second inequality is due to $\rho \ge 1$ and the last inequality holds with any $K \ge (\kappa_3+\kappa_4)^2\epsilon^{-2}$.

This completes the proof. ∎

Remark 3.17. In view of Examples A.2 and A.3, it is possible to have $M_g = 0$ with g_i being the indicator function of some proper sets. While the condition $\nabla_f = 0$ is a little restrictive as it means that f is constant with respect to x_i, this is not impossible as we work with multiblock problems. As a result, our $\mathcal{O}(\epsilon^{-2})$ complexity estimate in Theorem 3.16, if not stronger than, complements the previous ADMM works in the sense that we do not rely on both conditions (a) and (b) discussed in section 2.2.

3.4. Adaptive SDD-ADMM. In Algorithm 3.1, we use a fixed penalty ρ, which is on the order of $\Theta(\epsilon^{-2})$ in view of Theorem 3.12, or $\Theta(\epsilon^{-1})$ and $\Theta(1)$ in view of Theorem 3.16. The exact value of ρ depends on the problem data, i.e., parameters $(L_f, M_h, K_h, J_h, L_h)$ and $(r, \sigma, \nabla_f, M_g)$ required in Assumption 3.13, and may not be straightforward to estimate for some applications. In this subsection, we show that it is possible to find an ϵ-stationary point of problem (1.1) through multiple calls of SDD-ADMM with increasing ρ's. Moreover, this adaptive version does not deteriorate the iteration estimates established in Theorems 3.12 and 3.16.

The proposed adaptive version of SDD-ADMM is presented in Algorithm 3.2. Essentially we start SDD-ADMM with a relatively small penalty ρ_t for some iterations, and rerun SDD-ADMM with $\rho_{t+1} = 2\rho_t$ until an ϵ-stationary solution is located. One technical issue is that, we need to initialize the tth SDD-ADMM with a proper $x^0 \in X$ satisfying Assumption 3.8 with $\alpha = \rho_t$. Of course, if some $x^0 \in X \cap \{x | h(x) = 0\}$ is available, then we can set $x^{t,0} = x^0$ for all index $t \ge 1$. Otherwise, any primal iterate in the tth SDD-ADMM satisfying Assumption 3.8 with $\alpha = \rho_{t+1}$ can serve as $x^{t+1,0}$.

Though invoking a sequence of calls to SDD-ADMM, this adaptive version preserves the same iteration complexities established in Theorems 3.12 and 3.16. To see

Algorithm 3.2. `Adaptive SDD-ADMM`.

Input $(\rho_0, \omega, \theta, \tau, \epsilon) \in (0, +\infty) \times [4, +\infty) \times (1, \infty) \times [0, +\infty) \times (0, +\infty)$;
initialize index $t \leftarrow 0$;
while an ϵ-stationary solution of (1.1) is not found **do**
  $t \leftarrow t + 1$;
  find $x^{t,0} \in X$ satisfying Assumption 3.8 with $\alpha = \rho_t := 2^t \rho_0$;
  run `SDD-ADMM`$(x^{t,0}, \rho_t, \omega, \theta, \tau)$ for at most $K(\rho_t)$ iterations;
end while

this, denote the total number of SDD-ADMM calls by T. Notice that in view of Theorem 3.12, there exists a constant $B > 0$ such that $K(\rho) \le B\rho\epsilon^{-2}$. The total number of SDD-ADMM iterations can be then bounded by

$$\mathcal{T} := \sum_{t=1}^{T} K(\rho_t) \le B\epsilon^{-2} \times \sum_{t=1}^{T} \rho_0 2^t = B\epsilon^{-2} \times 2\rho_0(2^T - 1) \le 2B\epsilon^{-2}\rho_0 2^T. \tag{3.34}$$

By Theorem 3.12, it suffices to find T such that $\rho_T = \rho_0 2^T = \Theta(\epsilon^{-2})$, plugging which into (3.34) gives the same $\mathcal{T} = \mathcal{O}(\epsilon^{-4})$ iteration complexity estimate. Similarly under assumptions of Theorem 3.16, the orders of $K'(\rho) = \mathcal{O}(\epsilon^{-3})$ and $K''(\rho) = \mathcal{O}(\epsilon^{-2})$ can be preserved as well.

4. UDD ALM. The success of SDD motivates us to ask a natural question: what if we skip the scaling step? In this section, we investigate the UDD update for solving the following special case of problem (1.1), where $p = 1$, g is convex, and h is affine:

$$\min_{x \in \mathbb{R}^n} \{f(x) + g(x) \mid h(x) := Ax - b = 0\}. \tag{4.1}$$

We note that the analysis in this section can be applied to a more general multiblock setting, while focusing on $p = 1$ suffices to demonstrate the behavior of UDD. Differently from SDD-ADMM, the convergence of UDD-ALM requires certain regularity or constraint qualification to hold at the primal limit point, so that the sequence of dual variables has a bounded subsequence and the augmented Lagrangian function may then serve as a potential function. We focus on the structured setup (4.1) in this section. In Appendix B, we generalize the analysis to handle a more challenging setting with nonconvex g and nonlinear h by assuming a stronger subproblem oracle.

Formally, we adopt the following assumptions.

Assumption 4.1. We make the following assumptions regarding problem (4.1).

1. The function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ can be decomposed as $g_0 + \delta_X$, where $X \subseteq \mathbb{R}^n$ is convex and compact, and $g_0 : \mathbb{R}^n \to \mathbb{R}$ is continuous and convex over X.
2. The function $f : \mathbb{R}^n \to \mathbb{R}$ has an L_f-Lipschitz gradient over X.
3. The constraints $h(x) = Ax - b$ are affine with $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, and $X \cap \{x | Ax = b\} \neq \emptyset$.

Recall the definition of $\mathcal{K}_\rho$ in (3.5). We see $\nabla_x \mathcal{K}_\rho(x, \mu) = \nabla f(x) + A^\top \mu + \rho A^\top (Ax - b)$ is Lipschitz with modulus $L_f + \rho \|A^\top A\|$, which is independent of μ due to the linearity of constraints. This fact allows us to use a single proximal gradient step to update x. See Algorithm 4.1.

Algorithm 4.1. UDD-ALM for problem (4.1).

1: **Initialize** $x^0 \in X$, $\mu^0 \in \mathbb{R}^m$, $\rho \geq 0$, $\varrho > 0$, and $\theta > 1$; set $L_{\mathcal{K}} = L_f + \rho\|A^\top A\|$;
2: **for** $k = 0, 1, 2 \cdots$ **do**
3: perform a proximal gradient step:

$$x^{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^n} g(x) + \langle \nabla f(x^k) + A^\top(\mu^k + \rho(Ax^k - b)), x - x^k \rangle + \frac{\theta L_{\mathcal{K}}}{2}\|x - x^k\|^2; \tag{4.2}$$

4: update μ^{k+1} through a UDD update:

$$\mu^{k+1} = \mu^k - \varrho(Ax^{k+1} - b); \tag{4.3}$$

5: **end for**

First we establish the descent of the augmented Lagrangian function.

LEMMA 4.2 (one-step progress of UDD-ALM). *Suppose Assumption* 4.1 *holds. For all* $k \in \mathbb{Z}_{++}$, *we have*

$$L_\rho(x^k, \mu^k) - L_\rho(x^{k+1}, \mu^{k+1}) \geq \left(\frac{2\theta - 1}{2}\right) L_{\mathcal{K}}\|x^{k+1} - x^k\|^2 + \varrho\|Ax^{k+1} - b\|^2. \tag{4.4}$$

Proof. Similarly to (3.11) and using the fact that g is convex, the descent in x is given as $L_\rho(x^k, \mu^k) - L_\rho(x^{k+1}, \mu^k) \geq \left(\frac{2\theta-1}{2}\right) L_{\mathcal{K}}\|x^{k+1} - x^k\|^2$. The change with respect to μ is given as $L_\rho(x^{k+1}, \mu^{k+1}) - L_\rho(x^{k+1}, \mu^k) = \langle \mu^{k+1} - \mu^k, Ax^{k+1} - b)\rangle = -\varrho\|Ax^{k+1} - b\|^2$, where the last equality is due to the UDD update. Combining the inequality and the equality proves the claim. ∎

We then bound the dual residuals of iterates produced by UDD-ALM.

LEMMA 4.3 (bound on dual residual in UDD-ALM). *Suppose Assumption* 4.1 *holds. For all* $k \in \mathbb{N}$, *it holds that*

$$\begin{aligned} &\operatorname{dist}\left(\partial g(x^{k+1}), -\nabla f(x^{k+1}) - A^\top \mu^{k+1}\right) \\ &\leq (\theta + 1) L_{\mathcal{K}}\|x^{k+1} - x^k\| + (\rho + \varrho)\|A\|\|Ax^{k+1} - b\|. \end{aligned}$$

Proof. The claim follows from the optimality of x^{k+1} in (4.2), the fact that $\nabla_x \mathcal{K}_\rho$ is Lipschitz, and straightforward derivations. ∎

Lemma 4.2 suggests that values of the augmented Lagrangian function form a nonincreasing sequence. We aim to show that this sequence is actually bounded from below if a certain regularity condition is satisfied.

DEFINITION 4.4 (modified Robinson's condition). *We say* $x \in X = \operatorname{dom} g$ *satisfies the modified Robinson's condition if* $\{Ad \mid d \in T_X(x)\} = \mathbb{R}^m$, *where* $T_X(x)$ *denotes the tangent cone of* X *at* x:

$$T_X(x) = \left\{ d \in \mathbb{R}^n \mid d = \lim_{k \to \infty} \frac{x^k - x}{\tau_k}, x^k \to x, \tau_k \downarrow 0, \{x_k\}_{k \in \mathbb{N}} \subseteq X \right\}.$$

The above definition is slightly different from the standard Robinson's condition, e.g., in [47, section 3.3.2]: we do not require x to satisfy $Ax = b$ in Definition 4.4.

Despite this difference, [47, Lemma 3.16] still gives a sufficient condition: A has full row rank and $x + \mathrm{Null}(A) \cap \mathrm{int}\, X \neq \emptyset$, where $\mathrm{Null}(A)$ denotes the null space of A and $\mathrm{int}\, X$ denotes the interior of X. This modified Robinson's condition has been adopted to ensure a certain boundedness condition on the dual sequence and verified in specific applications [18, 49, 50]. Since the techniques are not new, we present the following lemma in the context of problem (4.1) while skipping the proof.

LEMMA 4.5 (existence of dual limit point). *Suppose Assumption* 4.1 *holds. Let* $x^* \in X$ *be a limit point of the sequence* $\{x_k\}_{k\in\mathbb{N}}$ *generated by UDD-ALM, and* $\{x^{k_r}\}_{r\in\mathbb{N}}$ *be the subsequence convergent to* x^*. *If* x^* *satisfies the modified Robinson's condition, then* $\{\mu^{k_r}\}_{r\in\mathbb{R}^n}$ *has a bounded subsequence and hence a limit point* μ^*.

THEOREM 4.6. *Suppose Assumption* 4.1 *holds. Let* x^* *be a limit point of the sequence* $\{x^k\}_{k\in\mathbb{N}}$ *generated by UDD-ALM that satisfies the modified Robinson's condition. Then the following statements hold.*

1. *(asymptotic convergence) The point* x^* *is a stationary point of problem* (4.1).
2. *(iteration complexity) Let* $\epsilon > 0$. *Define constants*

$$\delta_1 := \min\left\{\frac{(2\theta-1)L_{\mathcal{K}}}{2}, \varrho\right\}, \quad \delta_2 := (\theta+1)L_{\mathcal{K}} + (\rho+\varrho)\|A\|.$$

UDD-ALM finds an ϵ*-stationary solution in at most* K *iterations, where*

$$K \le \left\lceil \frac{\max\{1,\delta_2\}^2 (L_\rho(x^0,\mu^0) - f(x^*) - g(x^*))}{\delta_1 \epsilon^2} \right\rceil = \mathcal{O}(\epsilon^{-2}). \tag{4.5}$$

Proof. Let $\{x^{k_r}\}_{r\in\mathbb{N}}$ be the subsequence convergent to x^*. By Lemma 4.5, we may assume $\mu^{k_r} \to \mu^* \in \mathbb{R}^m$ as $r \to \infty$ without loss of generality. Consequently, $L_\rho(x^{k_r}, \mu^{k_r}) \to L_\rho(x^*, \mu^*)$ due to the continuity of the augmented Lagrangian function over X. By Lemma 4.2, the sequence $\{L_\rho(x^k, \mu^k)\}_{r\in\mathbb{N}}$ is nonincreasing, so the whole sequence is bounded from below by $L_\rho(x^*, \mu^*)$. Summing the inequality claimed in Lemma 4.2 from $k = 0$ to some positive integer $K-1$, we have

$$\begin{aligned} &\min\left\{\frac{(2\theta-1)L_{\mathcal{K}}}{2}, \varrho\right\} \sum_{k=0}^{K-1} \|x^{k+1} - x^k\|^2 + \|Ax^{k+1} - b\|^2 \\ &\le \sum_{k=0}^{K-1} \frac{(2\theta-1)L_{\mathcal{K}}}{2}\|x^{k+1} - x^k\|^2 + \varrho\|Ax^{k+1} - b\|^2 \le L_\rho(x^0,\mu^0) - L_\rho(x^*,\mu^*). \end{aligned} \tag{4.6}$$

Now we are ready to prove the two claims.

1. Let $K \to \infty$ in (4.6) and focusing on the subsequence $\{x^{k_r}, \mu^{k_r}\}_{r\in\mathbb{N}}$, we see $\lim_{r\to\infty} \max\{\|x^{k_r} - x^{k_r-1}\|, \|Ax^{k_r} - b\|\} = 0$, from which it immediately follows that $\|Ax^* - b\| = \lim_{r\to\infty} \|Ax^{k_r} - b\| = 0$. Moreover, by Lemma 4.3,

$$\begin{aligned} &\mathrm{dist}\left(-\nabla f(x^*) - A^\top \mu^*, \partial g(x^*)\right) \le \lim_{r\to\infty} \mathrm{dist}\left(-\nabla f(x^{k_r}) - A^\top \mu^{k_r}, \partial g(x^{k_r})\right) \\ &\le \lim_{r\to\infty} (\theta+1)L_{\mathcal{K}}\|x^{k_r} - x^{k_r-1}\| + (\rho+\varrho)\|A\|\|Ax^{k_r} - b\| = 0. \end{aligned}$$

   This suggests that x^* is a stationary point of problem (4.1).
2. Since $Ax^* - b = 0$, we have $L_\rho(x^*, \mu^*) = f(x^*) + g(x^*)$. By (4.6), there exists an index $0 \le \bar{k} \le K-1$ such that

$$\max\{\|x^{\bar{k}+1} - x^{\bar{k}}\|, \|Ax^{\bar{k}+1} - b\|\} \le \left(\frac{L_\rho(x^0,\mu^0) - f(x^*) - g(x^*)}{\delta_1 K}\right)^{1/2}.$$

By Lemma 4.3 and the above inequality,

$$\max\left\{\operatorname{dist}\left(\partial g(x^{\bar{k}+1}), -\nabla f(x^{\bar{k}+1}) - A^\top \mu^{\bar{k}+1}\right), \|Ax^{\bar{k}+1} - b\|\right\}$$
$$\leq \max\{1, \delta_2\}\left(\frac{L_\rho(x^0, \mu^0) - f(x^*) - g(x^*)}{\delta_1 K}\right)^{1/2} \leq \epsilon,$$

where the last inequality holds by the claimed upper bound of K in (4.5). ∎

Remark 4.7. We make some remarks on UDD-ALM.

1. Different from the $\mathcal{O}(\epsilon^{-2})$ established in [61, 60], Theorem 4.6 relies on the modified Robinson's condition at the limit point of iterates produced by UDD-ALM. Though assuming a certain constraint qualification at the limit point is common in nonlinear programs, this specific condition may not be satisfied by general instances of (4.1).
2. In Appendix B, we extend the iteration complexity result to handle nonconvex g and nonlinear h by assuming a stronger subproblem oracle.

The value of ϱ deserves more attention in deriving the $\mathcal{O}(\epsilon^{-2})$ complexity in Theorem 4.6. We treat ϱ as a constant in our analysis and do not impose explicit requirements. However, it is observed that a larger ϱ usually leads to iterates staying on the boundary of X and, hence, the limit point is more likely to violate the modified Robinson's condition. As we illustrate in section 5.3, the behavior of UDD-ALM is very sensitive to the choice of ϱ. For certain instances, the numerical value of ϱ needs to be even smaller than ϵ in order for UDD-ALM to exhibit convergence. In this case, the $\mathcal{O}(\epsilon^{-2})$ complexity may not be practically informative. We share more empirical observations in section 5.3.

5. Numerical experiments.

5.1. SDD-ALM for nonconvex QCQP. In this section, we consider the following nonconvex quadratically constrained QP (QCQP)

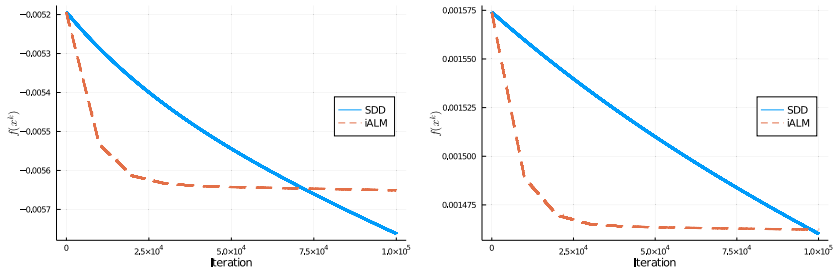
$$\min_{x\in\mathbb{R}^n} \{f(x) := x^\top Q x + q^\top x \mid h(x) := x^\top B x - 1 = 0, \|x\| \leq r\}, \tag{5.1}$$

and compare it with the iALM proposed in [31]. Let $g(x) = \delta_{\{x \mid \|x\| \leq r\}}(x)$, whose projection operator can be computed explicitly. We generate data as follows: first create $\tilde{Q} \in \mathbb{R}^{n\times n}$ with standard Gaussian entries, and set $Q = 0.5(\tilde{Q} + \tilde{Q}^\top)$; generate $\bar{B}$ in the same way as Q, and set $B = \bar{B} + (\|\bar{B}\| + 1)I_n$, where I_n denotes the identity matrix; finally we set q to be the zero vector and $r = n/10$. For SDD-ADMM, we choose $(\omega, \theta, \tau) = (4, 2, 1)$ and simply set $\rho = 10n$. For iALM, we limit the number of outer-level updates by 10, and the penalty used in each outer level is $\beta_k = \beta_0\sigma^k$, where $\beta_0 = 1$ and $\sigma = 2(\rho/\beta_0)^{1/10}$, so that β_k in iALM should be able to quickly catch up the SDD-ALM penalty ρ; we set the input tolerance to the inner-level APG and middle-level iPPM to be 1e-3. For both algorithms, we first generate a vector with standard Gaussian entries, then scale it to get x^0 so that $\|h(x^0)\| = 0.5/\sqrt{\rho}$.

For each run of an algorithm, we record the primal residual "`pres`" (measured by $\|h(x^{k+1})\|$), dual residual "`dres`" (measured by $\|x^{k+1} - x^k\|$), the iteration index "`iter`" when both `pres` and `dres` drop below 1e-3 for the first time, as well as the wall clock time "`time`" over 100,000 proximal gradient iterations; if either `pres` or `dres` does not drop below 1e-3, we record their values where the sum of `pres` and `dres` is the minimum, and set `iter` = 100,000. For $n \in \{100, 200, 300\}$, we generate 5 instances and report the average metrics in Table 1. For $n \in \{100, 200\}$, SDD-ALM

TABLE 1
Averaged metrics of SDD-ALM and iALM [31].

	SDD-ALM				iALM			
n	`pres`	`dres`	`iter`	`time`	`pres`	`dres`	`iter`	`time`
100	1.00e-3	2.19e-8	16,158	13.97	1.03e-2	2.68e-10	100,000	25.73
200	1.00e-3	3.02e-8	81,729	37.08	1.03e-2	6.51e-13	100,000	73.37
300	3.11e-3	2.97e-9	100,000	69.43	8.86e-3	6.78e-13	100,000	147.29



(a) An instance with $n = 200$ (b) An instance with $n = 300$

FIG. 1. *Objective trajectories of SDD-ALM and iALM* [31]. *Note: color appears only in the online article.*

reduces `pres` below 1e-3, while for $n = 300$, SDD-ALM achieves a slightly better `pres`. In constrast, iALM maintains a smaller `dres` in all runs. On average, SDD-ALM takes less time to perform 100,000 proximal gradient updates.

For the generated instances, we also observe that iALM usually reduces the objective value faster than SDD-ALM, while SDD-ALM seems to converge to solutions with better qualities in the long run. The objective trajectories of two instances are plotted in Figure 1.

5.2. SDD-ADMM for robust tensor PCA. In this section, we test SDD-ADMM on the robust tensor PCA problem, and compare it with the ADMM-g algorithm proposed in [27]. Given a tensor $\mathcal{T} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, the goal is to decompose $\mathcal{T}$ as $\mathcal{Z} + \mathcal{E} + \mathcal{B}$, where $\mathcal{E}$ has a low rank, $\mathcal{E}$ is sparse, and $\mathcal{B}$ represents a small noise. The problem can be cast as the following multiblock problem:

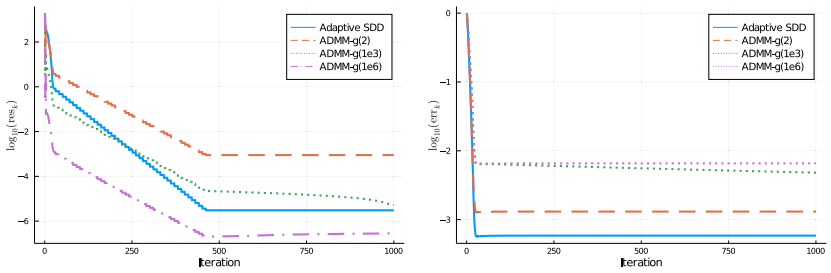
$$\min_{A,B,C,\mathcal{Z},\mathcal{E},\mathcal{B}} \{\|\mathcal{Z} - [\![A, B, C]\!]\|^2 + \alpha\|\mathcal{E}\|_1 + \alpha_N\|\mathcal{B}\|_F^2 \mid \mathcal{Z} + \mathcal{E} + \mathcal{B} = \mathcal{T}\}, \tag{5.2}$$

where $A \in \mathbb{R}^{I_1 \times R}, B \in \mathbb{R}^{I_2 \times R}, C \in \mathbb{R}^{I_3 \times R}$, R is an estimate of the true CP-rank, $[\![A, B, C]\!]$ denotes the summation of columnwise outer products of A, B, and C, and $\|\cdot\|_F$ denotes the Frobenius norm. See [27] and references therein for a detailed background of the problem. Following standard notations, the Khatri–Rao product, the Hadamard product, and the soft shrinkage operator are denoted by $\odot$, $\circ$, and $\mathbf{S}$, respectively; we use $\mathcal{Z}_{(i)}$ to denote the mode-i unfolding of a tensor $\mathcal{Z}$, and use $\mathbf{0}^{I_1 \times I_2 \times I_3}$ to denote the zero tensor. An adaptive SDD-ADMM tailored to problem (5.2) performs the following updates.

Lines 4-9 are standard proximal ADMM updates, while we update the dual variable in line 10 via an SDD step. Due to the linearity of constraints, each partial proximal augmented Lagrangian subproblem in SDD-ADMM admits closed-form solutions and, hence, we do not necessarily need to perform a proximal gradient step

Algorithm 5.1. `Adaptive SDD-ADMM` for problem (5.2).

Input $\rho, p, \omega, \tau, \gamma, \bar{\rho}, k_{\text{interval}} > 0$;
Initialize $A^0, B^0, C^0, \mathcal{Z}^0$ $\mathcal{E}^0$, $\mathcal{B}^0$; $\mu^0 = \mathbf{0}^{I_1 \times I_2 \times I_3}$;
for $k = 0, 1, 2, \ldots$ **do**
  $A^{k+1} = [(Z)^k_{(1)}(C^k \odot B^k) + 0.5pA^k][((C^k)^\top C^k) \circ ((B^k)^\top B^k) + 0.5pI_R]^{-1}$;
  $B^{k+1} = [(Z)^k_{(2)}(C^k \odot A^k) + 0.5pB^k][((C^k)^\top C^k) \circ ((A^k)^\top A^k) + 0.5pI_R]^{-1}$;s
  $C^{k+1} = [(Z)^k_{(3)}(B^k \odot A^k) + 0.5pC^k][((B^k)^\top B^k) \circ ((C^k)^\top C^k) + 0.5pI_R]^{-1}$;
  $\mathcal{E}^{k+1}_{(1)} = \mathbf{S}\left(\frac{\rho}{\rho+p}(\mathcal{T}_{(1)} - \mu^k_{(1)}/\rho - \mathcal{B}^k_{(1)} - \mathcal{Z}^k_{(1)}) + \frac{p}{\rho+p}\mathcal{E}^k_{(1)}, \frac{\alpha}{\rho+p}\right)$;
  $\mathcal{Z}^{k+1}_{(1)} = \frac{1}{2+2p+\rho}(2A^{k+1}\left(C^{k+1} \odot B^{k+1}\right)^\top + 2p\mathcal{Z}^k_{(1)} - \mu^k_{(1)} - \rho\left(\mathcal{E}^{k+1}_{(1)} + \mathcal{B}^k_{(1)} - \mathcal{T}_{(1)}\right))$;
  $\mathcal{B}^{k+1}_{(1)} = \frac{1}{\rho+2\alpha_N+p}\left(p\mathcal{B}^k_{(1)} - \mu^k_{(1)} - \rho(\mathcal{Z}^{k+1}_{(1)} + \mathcal{E}^{k+1}_{(1)} - \mathcal{T}_{(1)})\right)$;
  $\mu^{k+1}_{(1)} = \frac{1}{1+\tau}(\tau\mu^{k+1}_{(1)} - \omega\rho^{-1}(\mathcal{Z}^{k+1}_{(1)} + \mathcal{E}^{k+1}_{(1)} + \mathcal{B}^{k+1}_{(1)} - \mathcal{T}_{(1)}))$;
  if $(k+1)\%k_{\text{interval}} = 0$ **then**
    $\rho \leftarrow \min(\bar{\rho}, (1+\gamma)\rho)$;
  end if
end for



(a) Geo. Mean of Primal Residual (b) Geo. Mean of Relative Error

FIG. 2. $(I_1, I_2, I_3, R_{cp}) = (30, 50, 70, 40)$. *Note: color appears only in the online article.*

for each block variable. Moreover, motivated by the adaptive SDD-ADMM in section 3.4, we simply increase the penalty ρ by some factor $(1+\gamma)$ after every fixed number of iterations, until some upper bound $\bar{\rho}$ is reached. We also acknowledge that slow convergence of SDD-ADMM is observed with a fixed penalty.

Given a parameter $R_{cp} > 0$ and dimensions (I_1, I_2, I_3), we guess the initial CP rank by $R = R_{cp} + \lceil 0.2R_{cp} \rceil$, and generate all data in exactly the same way as in [27]. For adaptive SDD-ADMM, we choose $\gamma = 1/3$, $\tau = 1/(1+\gamma)$, $\omega = (1+\gamma)/\gamma$, $p = 1$, and the initial penalty $\rho = 2$; moreover, we set $k_{\text{interval}} = 10$ and $\bar{\rho} = 1\text{e}6$. For each instance, we run ADMM-g with three different values of ρ: $\rho = 2$ (the initial penalty passed to adaptive SDD-ADMM), $\rho = 1\text{e}6$ (the maximum penalty used in adaptive SDD-ADMM), and $\rho = 1\text{e}3$ (an intermediate value). For all algorithms, we initialize (A, B, C) with standard Gaussian entries, $\mathcal{Z}$ with the zero tensor, and $(\mathcal{B}, \mathcal{E})$ with the tensors used to generate $\mathcal{T}$.

For $(I_1, I_2, I_3, R_{cp}) = (30, 50, 70, 40)$, we plot $\texttt{res}_k$ and $\texttt{err}_k$ as functions of iteration index k in Figure 2: here $\texttt{res}_k$ is the geometric mean of the primal residual $\|\mathcal{Z}^k + \mathcal{E}^k + \mathcal{B}^k - \mathcal{T}\|_F$ over 3 instances, and $\texttt{err}_k$ is the geometric mean of the relative error $\|\mathcal{Z}^k - \mathcal{Z}^*\|_F / \|\mathcal{Z}^*\|$ over 3 instances, where $\mathcal{Z}^*$ is the low-rank ground truth. The

adaptive SDD-ADMM is able to reduce the primal residual close to zero, and recover a $\mathcal{Z}^k$ whose relative error is less than 1% or even close to 0.1%. In contrast, the performance of ADMM-g is sensitive to the choice of ρ: a smaller ρ usually results in a large primal residual, while a larger ρ leads to $\mathcal{Z}^k$ with poor quality. Tests on other problem scales exhibit similar behaviors and hence are omitted from presentation.

5.3. Observations for UDD-ALM. In this last subsection, we present some experiments on UDD-ALM applied to weakly convex minimization over affine constraints. A key observation in our experiments is that, although the dual step size ϱ is chosen as a constant in our analysis, UDD-ALM is very sensitive to its numerical value. In particular, UDD-ALM may indeed fail to converge when a relatively large ϱ is used, and the order of constraint violation $\|Ax-b\|$ and the order of ϱ are closely related. We consider a simple consensus problem

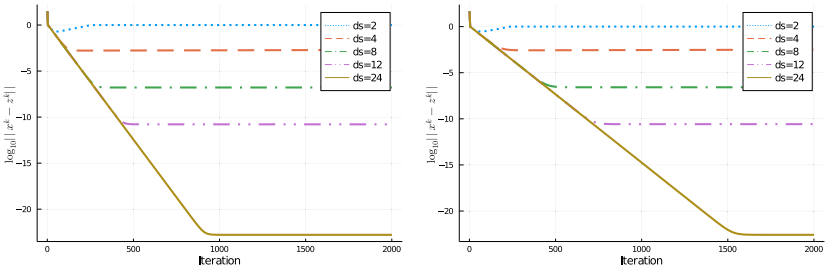
$$\min_{x,z\in\mathbb{R}^n} \{f(x)+\alpha\|z\|_1 \mid x-z=0, \|x\|\le r\}, \tag{5.3}$$

where $f(x)=-x^\top(U^\top U)x$ and $U\in\mathbb{R}^{n\times n}$ has standard Gaussian entries. We fix $\alpha=r=1$, $\rho=1000$, and test UDD-ALM with different dual scaling factors $\texttt{ds}>0$: the dual stepsize is chosen as $\varrho=\rho\times 0.1^{\texttt{ds}}$. In all runs of UDD-ALM, x^0 and z^0 are initialized with standard Gaussian entries. The objective values at the end of 2000 iterations are recorded in Table 2, and we plot the trajectories of the primal residuals in Figure 3. We observe that for both instances, UDD-ALM converges to the zero vector as ds gets smaller, which is indeed a stationary point of (5.3); moreover, the order of constraint violation drops significantly as well.

We note that when $\texttt{ds}=24$, the value of ϱ can be orders-of-magnitude smaller than ϵ and hence the theoretical $\mathcal{O}(\epsilon^{-2})$ complexity in Theorem 4.6 is invalidated. In particular, when we choose ϱ to be close to zero, UDD becomes the limiting behavior of SDD with $\tau\to+\infty$, and the resulting algorithm resembles the penalty method, where the dual variables stay close to zero. In other words, the empirical convergence of UDD, to some extent, can be attributed to the penalty method. It is important to note

TABLE 2
Objective values obtained by UDD-ALM.

n	ds = 2	ds = 4	ds = 8	ds = 12	ds = 24
500	17.59	3.51e-2	2.89e-6	2.89e-10	2.89e-22
1000	25.43	6.68e-6	6.68e-6	6.68e-10	6.68e-22



(a) An instance with $n=500$ (b) An instance with $n=1000$

FIG. 3. *UDD-ALM with different* ds. *Note: color appears only in the online article.*

that despite the convergence results presented in Theorem 4.6, we acknowledge that the dual step size ϱ might implicitly affect the fulfillment of the proposed regularity condition at the primal limit point. In practical terms, a large value of ϱ often leads to primal iterates approaching the boundary of X, which in turn increases the likelihood of the regularity condition being violated. As a result, we do not claim that UDD-ALM outperforms existing algorithms. Instead, our objective is to share our initial observations on this seemingly counterintuitive scheme in order to stimulate further exploration and understanding of its potential advantages and limitations.

6. Conclusions. This paper proposes two new algorithms based on the concept of dual descent: SDD-ADMM and UDD-ALM. We apply SDD-ADMM to solve nonlinear equality-constrained multiblock problems, and establish an $\mathcal{O}(\epsilon^{-4})$ iteration complexity upper bound, or $\mathcal{O}(\epsilon^{-3})$ and $\mathcal{O}(\epsilon^{-2})$ under additional technical assumptions. When UDD-ALM is applied for weakly convex minimization over affine constraints, we show that under a regularity condition, the algorithm asymptotically converges to a stationary point and finds an approximate solution in $\mathcal{O}(\epsilon^{-2})$ iterations. Our iteration complexities for both algorithms either achieve or improve the best-known results in the ADMM and ALM literature. Moreover, SDD-ADMM addresses a long-standing limitation of existing ADMM frameworks.

Nevertheless, the behavior of UDD-ALM is somehow not fully understood. Theoretically the dual stepsize ϱ is treated as a constant, while, as we illustrate numerically, the convergence of UDD-ALM can be very sensitive to its numerical value. We conjecture that the modified Robinson's condition required on the limit point can be implicitly affected by the dual step size. This issue seems to be highly problem dependent, and we leave it as our future work.

Appendix A. Examples satisfying Assumption 3.13. We give some examples where Assumption 3.13 can be satisfied. Suppose $i = p = 1$ for simplicity. We assume that for all $x \in X(r)$, $\nabla h(x)$ has full column rank, and their smallest singular values are bounded away from zero, i.e., $\sigma(r) := \inf_{x\in X(r)} \sigma_{\min}(\nabla h(x)) > 0$.

Example A.1. Let g be a possibly nonconvex function with

$$M_g := \sup\{\|\xi_g\| : \xi_g \in \partial g(x), x \in X(r)\} < +\infty,$$

i.e., g is Lipschitz over $X(r)$. Then we have

$$\begin{aligned}\operatorname{dist}(-\nabla h(x)\mu, \partial g(x)) &= \inf_{\xi_g\in\partial g(x)} \|\nabla h(x)\mu + \xi_g\| \\ &\geq \|\nabla h(x)\mu\| + \inf_{\xi_g\in\partial g(x)} -\|\xi_g\| \geq \sigma\|\mu\| - M_g.\end{aligned}$$

Hence Assumption 3.13 is satisfied with $\sigma = \sigma(r)$.

Example A.2. Let $g = \delta_X + \tilde{g}$, where X is a full-dimensional compact convex set and $\tilde{g}$ is a convex function over X. By [43, Theorems 24.7 and 23.8],

$$M_g := \sup\{\|\xi_{\tilde{g}}\| :\ \xi_{\tilde{g}} \in \partial\tilde{g}(x), x \in X\} < +\infty \ \text{ and } \ \partial g(x) = \partial\tilde{g}(x) + N_X(x) \ \ \forall x \in X,$$

where $N_X(x)$ denotes the normal cone of X at $x \in X$. Further assume that $X(r)$ belongs to $\operatorname{int} X$, the interior of X, so that $N_X(x) = \{0\}$ for all $x \in X(r)$. As a result,

$$\begin{aligned}\operatorname{dist}(-\nabla h(x)\mu, \partial g(x)) &= \inf\{\|\nabla h(x)\mu + \xi_{\tilde{g}} + d_g\| :\ \xi_{\tilde{g}} \in \partial\tilde{g}(x), d_g \in N_X(x)\} \\ &\geq \|\nabla h(x)\mu\| - M_g \geq \sigma(r)\|\mu\| - M_g.\end{aligned}$$

So again Assumption 3.13 is satisfied with $\sigma = \sigma(r)$.

Example A.3. Suppose $g = \delta_X$ and $X := \{x \in \mathbb{R}^n \mid F(x) \in D\}$, where $F : \mathbb{R}^n \to \mathbb{R}^p$ are continuously differentiable and $D \subset \mathbb{R}^p$ ($p \le n - m$). Further suppose that for any $x \in X(r)$, the Jacobian matrix $J(x) := [\nabla h(x), \nabla F(x)] \in \mathbb{R}^{n\times(m+p)}$ has full column rank, and $\sigma := \min_{x\in X(r)} \sigma_{\min}(J(x)) > 0$. Then by [46, Theorem 6.14], for all $x \in X(r)$, it holds that

$$\partial g(x) = N_X(x) \subset \{\nabla F(x)y \mid y \in N_D(F(x))\}.$$

Denote $u = [\mu^\top, y^\top]^\top$; since $\|J(x)u\| \ge \sigma\|u\| \ge \sigma\|\mu\|$, we have

$$\begin{aligned}\operatorname{dist}(-\nabla h(x)\mu, \partial g(x)) &\ge \inf\{\|\nabla h(x)\mu + \nabla F(x)y\| : y \in N_D(F(x))\}\\ &= \inf\{\|J(x)u\| : y \in N_D(F(x))\} \ge \sigma\|\mu\|.\end{aligned}$$

In particular, consider $h(x) = Ax - b$ and $X = \{x \in \mathbb{R}^n \mid l \le Cx \le u\}$, where $C \in \mathbb{R}^{p\times n}$ and $l, u \in \mathbb{R}^p$. Then Assumption 3.13 holds as long as rows of A and C are linearly independent.

Appendix B. UDD-ALM with nonlinear constraints. In this section we apply UDD-ALM to deal with nonlinear constraints and establish its convergence by assuming a descent solution oracle of each augmented Lagrangian relaxation.

Assumption B.1. We make the following assumptions regarding problem (4.1).

1. The function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ can be decomposed as $\tilde{g} + \delta_X$, where $X \subseteq \mathbb{R}^n$ is compact and described by a finite number of inequality constraints, i.e., $X = \{x \in \mathbb{R}^n | q_l(x) \le 0, \forall l \in [L]\}$, where $q_l : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable for $l \in [L]$, and $\tilde{g} : \mathbb{R}^n \to \mathbb{R}$ is continuous and convex over X.
2. The function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable over X.
3. The constraints $h : \mathbb{R}^n \to \mathbb{R}^m$ are given by $h(x) = [h_1(x), \dots, h_m(x)]^\top$, where $h_j : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable over X for $j \in [m]$.

Denote $\nabla h(x) = [\nabla h_1(x), \dots, \nabla h_m(x)] \in \mathbb{R}^{n\times m}$ in this subsection. We also define an approximate KKT point for problem (4.1) under Assumption B.1 as follows.

Definition B.2. *Let $\epsilon > 0$. We say x is an ϵ-KKT point for problem* (4.1) *if*

$$\operatorname{dist}\left(-\nabla f(x) - \nabla h(x)\mu - \sum_{l=1}^{L} \nabla q_l(x) y_l, \partial\tilde{g}(x)\right) \le \epsilon,\ \|h(x)\| \le \epsilon, \tag{B.1a}$$

$$q_l(x) \le 0,\ y_l g_l(x) = 0 \quad \forall l \in [L] \tag{B.1b}$$

for some $\mu \in \mathbb{R}^m$ and $y \in \mathbb{R}^L_+$. We simply say x is a KKT point when $\epsilon = 0$.

The UDD-ALM with nonlinear constraints is almost the same as Algorithm 4.1, except that we replace the primal update (4.2) by the following nonlinear program:

$$\min_{x\in X} L_\rho(x, \mu^k) + \frac{c}{2}\|x - x^k\|^2, \tag{B.2}$$

where $L_\rho(x, \mu) = f(x) + \tilde{g}(x) + \langle\mu, h(x)\rangle + \frac{\rho}{2}\|h(x)\|^2$. Next we define a descent solution oracle for problem (B.2).

Assumption B.3. Given $x^k \in X$ and $\mu^k \in \mathbb{R}^m$, we can find x^{k+1} such that

$$L_\rho(x^{k+1}, \mu^k) \le L_\rho(x^k, \mu^k) - \nu\|x^{k+1} - x^k\|^2 \tag{B.3}$$

for some $\nu > 0$, and there exists $y^{k+1} \in \mathbb{R}^L_+$ such that

$$0 \in \partial_x L_\rho(x^{k+1}, \mu^k) + c(x^{k+1} - x^k) + \sum_{l=1}^{L} \nabla q_l(x^{k+1}) y_l^{k+1}, \tag{B.4a}$$

$$q_l(x^{k+1}) \le 0,\ y_l^{k+1} q_l(x^{k+1}) = 0 \ \ \forall l \in [L]. \tag{B.4b}$$

Remark B.4. Assumption B.3 requires x^{k+1} to be a KKT point of problem (B.2) with an improved objective value compared to the previous iterate x^k. Notice that the sufficient descent condition (B.3) can be satisfied with $\nu = c/2$ if some global solver for problem (B.2) is available. To this end, we also note that given $x^k \in X$ and $\mu^k \in \mathbb{R}^m$, problem (B.2) is convex if X is convex, f and $h_1, \dots, h_m$ have continuous Hessians over X, and c is sufficiently large. We adopt (B.4) to avoid unnecessary technicality, while it is possible to allow x^{k+1} to be an inexact KKT solution of (B.2).

Since X is assumed to be compact, the sequence $\{x^k\}_{k\in\mathbb{N}}$ has at least one limit point $x^* \in X$. The next lemma shows that if x^* satisfies the LICQ, then $\{\mu^k\}_{k\in\mathbb{N}}$ has a bounded subsequence.

LEMMA B.5. *Suppose Assumptions* B.1 *and* B.3 *hold. Let $x^* \in X$ be a limit point of $\{x^k\}_{k\in\mathbb{N}}$ generated by UDD-ALM, and $\{x^{k_r}\}_{r\in\mathbb{N}}$ be the corresponding convergent subsequence. Denote $I(x^*) = \{l \in [L] \mid q_l(x^*) = 0\}$. Suppose that the matrix*

$$H^* := [\nabla h_1(x^*), \dots, \nabla h_m(x^*), \{\nabla q_l(x^*)\}_{l\in I(x^*)}] \in \mathbb{R}^{n\times(m+|I(x^*)|)} \tag{B.5}$$

has full column rank, then the sequence $\{\mu^{k_r}\}_{r\in\mathbb{N}}$ is bounded.

Proof. For $l \notin I(x^*)$, we have $g_l(x^{k_r}) < 0$ and thus $y_l^{k_r} = 0$ by (B.4b) for all sufficiently large $r \in \mathbb{N}$. Hence, (B.4a) becomes $H_{k_r}[(\mu^{k_r})^\top, (y^{k_r}_{I(x^*)})^\top]^\top = e_{k_r}$, where

$$\begin{aligned} H_{k_r} &:= [\nabla h_1(x^{k_r}), \dots, \nabla h_m(x^{k_r}), \{\nabla q_l(x^{k_r})\}_{l\in I(x^*)}], \\ e_{k_r} &:= -\nabla f(x^{k_r}) - \xi_{\tilde g}^{k_r} - (\rho + \varrho)\nabla h(x^{k_r}) h(x^{k_r}) - c(x^{k_r} - x^{k_r - 1}), \end{aligned}$$

$y^{k_r}_{I(x^*)} \in \mathbb{R}^{|I(x^*)|}_+$ is the subvector of y^{k_r} specified by indices in $I(x^*)$, and $\xi_{\tilde g}^{k_r} \in \partial \tilde g(x^{k_r})$. Since H^* has full column rank, so does H^{k_r} for sufficiently large $r \in \mathbb{N}$, which suggests that $\|\mu^{k_r}\| \le \|\mu^{k_r}\| + \|y^{k_r}_{I(x^*)}\| \le \|(H_{k_r}^\top H_{k_r})^{-1} H_{k_r}^\top\| \|e_{k_r}\|$. Due to the compactness of X, the continuity of $\tilde g$, and the continuous differentiability of f, g_l's, and h_j's, we know that $\|e_{k_r}\|$ is bounded by some finite constant depending on the problem data $(f, \tilde g, X, h)$ as well as parameters (ρ, ϱ, c). As a result of the previous inequality, the sequence $\{\mu^{k_r}\}_{r\in\mathbb{N}}$ is bounded. ∎

THEOREM B.6. *Suppose Assumptions* B.1–B.3 *hold. Let $x^* \in X$ be a limit point of $\{x^k\}_{k\in\mathbb{N}}$ generated by UDD-ALM that satisfies the LICQ condition, i.e., H^* defined in* (B.5) *has full column rank. Then the following statements hold.*

1. *(asymptotic convergence) The point x^* is a KKT point of problem* (4.1).
2. *(iteration complexity) Let $\epsilon > 0$. Define constants $\sigma_1 := \min\{\nu, \varrho\}$ and $\sigma_2 := c + (\rho + \varrho)\max_{x\in X} \|\nabla h(x)\|$. UDD-ALM finds an ϵ-KKT point in at most K iterations, where*

$$K \le \left\lceil \frac{\max\{1, \sigma_2\}^2 (L_\rho(x^0, \mu^0) - f(x^*) - \tilde g(x^*))}{\sigma_1 \epsilon^2} \right\rceil = \mathcal{O}(\epsilon^{-2}).$$

Proof. First note that $L_\rho(x^k,\mu^k)-L_\rho(x^{k+1},\mu^{k+1}) \geq \nu\|x^{k+1}-x^k\|^2+\varrho\|h(x^{k+1})\|^2$ for all $k\in\mathbb{N}$ by (B.3) and the UDD update. Then with the help of Lemma B.5, the claims can be proved via straightforward modification of the proof of Theorem 4.6. ∎

Remark B.7. We note that the $\mathcal{O}(\epsilon^{-2})$ complexity bound in Theorem 4.6 is measured by first-order oracles of the problem data, whereas the iteration complexity in Theorem B.6 is measured by the subproblem oracle defined in Assumption B.3. Information including $\|\mu^k\|$ and ρ may affect the computation effort to evaluate the subproblem oracle, which we do not consider explicitly in Theorem B.6.

REFERENCES

[1] R. Andreani, E. G. Birgin, J. M. Martínez, and M. L. Schuverdt, *On augmented Lagrangian methods with general lower-level constraints*, SIAM J. Optim., 18 (2007), pp. 1286–1309.

[2] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.

[3] A. S. Berahas, F. E. Curtis, D. Robinson, and B. Zhou, *Sequential quadratic optimization for nonlinear equality constrained stochastic optimization*, SIAM J. Optim., 31 (2021), pp. 1352–1379.

[4] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 2014.

[5] D. Boob, Q. Deng, and G. Lan, *Stochastic first-order methods for convex and nonconvex functional constrained optimization*, Math. Program., 197 (2023), pp. 215–279.

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Found. Trends Mach. Learn., 3 (2011), pp. 1–122.

[7] C. Chen, B. He, Y. Ye, and X. Yuan, *The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent*, Math. Program., 155 (2016), pp. 57–79.

[8] F. E. Curtis, D. P. Robinson, and B. Zhou, *Inexact Sequential Quadratic Optimization for Minimizing a Stochastic Objective Function Subject to Deterministic Nonlinear Equality Constraints*, preprint, arXiv:2107.03512, 2021.

[9] D. Davis and W. Yin, *A three-operator splitting scheme and its optimization applications*, Set-valued Var. Anal., 25 (2017), pp. 829–858.

[10] J. Douglas and H. H. Rachford, *On the numerical solution of heat conduction problems in two and three space variables*, Trans. Amer. Math. Soc., 82 (1956), pp. 421–439.

[11] J. Eckstein and D. P. Bertsekas, *On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators*, Math. Program., 55 (1992), pp. 293–318.

[12] D. Gabay, *Applications of the method of multipliers to variational inequalities*, in Augmented Lagrangian Methods: Applications to the Solution of Boundary Value Problems, M. Fortin and R. Glowinski, eds., North-Holland, Amsterdam, 1983, pp. 299–330.

[13] D. Gabay and B. Mercier, *A dual algorithm for the solution of nonlinear variational problems via finite element approximation*, Comput. Math. Appl., 2 (1976), pp. 17–40.

[14] S. Ghadimi and G. Lan, *Accelerated gradient methods for nonconvex nonlinear and stochastic programming*, Math. Program., 156 (2016), pp. 59–99.

[15] R. Glowinski and A. Marroco, *Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires,* French J. Autom. Inform. Oper. Res. Numer. Anal., 9 (1975), pp. 41–76.

[16] M. L. Gonçalves, J. G. Melo, and R. D. Monteiro, *Extending the Ergodic Convergence Rate of the Proximal ADMM*, preprint, arXiv:1611.02903, 2016.

[17] M. L. Gonçalves, J. G. Melo, and R. D. Monteiro, *Convergence rate bounds for a proximal ADMM with over-relaxation stepsize parameter for solving nonconvex linearly constrained problems*, Pac. J. Optim., 15 (2019), pp. 379–398.

[18] D. Hajinezhad and M. Hong, *Perturbed proximal primal–dual algorithm for nonconvex nonsmooth optimization*, Math. Program., 176 (2019), pp. 207–245.

[19] B. He and X. Yuan, *On the o(1/n) convergence rate of the Douglas–Rachford alternating direction method*, SIAM J. Numer. Anal., 50 (2012), pp. 700–709.

[20] B. He and X. Yuan, *On non-ergodic convergence rate of Douglas–Rachford alternating direction method of multipliers*, Numer. Math., 130 (2015), pp. 567–577.

[21] M. R. Hestenes, *Multiplier and gradient methods*, J. Optim. Theory Appl., 4 (1969), pp. 303–320.
[22] M. Hong, *Decomposing Linearly Constrained Nonconvex Problems by a Proximal Primal Dual Approach: Algorithms, Convergence, and Applications*, preprint, arXiv:1604.00543, 2016.
[23] M. Hong and Z.-Q. Luo, *On the linear convergence of the alternating direction method of multipliers*, Math. Program., 162 (2017), pp. 165–199.
[24] M. Hong, Z.-Q. Luo, and M. Razaviyayn, *Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems*, SIAM J. Optim., 26 (2016), pp. 337–364.
[25] J. Jian, P. Liu, J. Yin, C. Zhang, and M. Chao, *A QCQP-based splitting SQP algorithm for two-block nonconvex constrained optimization problems with application*, J. Comput. Appl. Math., 390 (2021), 113368.
[26] J. Jian, C. Zhang, J. Yin, L. Yang, and G. Ma, *Monotone splitting sequential quadratic optimization algorithm with applications in electric power systems*, J. Optim. Theory Appl., 186 (2020), pp. 226–247.
[27] B. Jiang, T. Lin, S. Ma, and S. Zhang, *Structured nonconvex and nonsmooth optimization: Algorithms and iteration complexity analysis*, Comput. Optim. Appl., 72 (2019), pp. 115–157.
[28] W. Kong, J. G. Melo, and R. D. Monteiro, *Iteration complexity of a proximal augmented Lagrangian method for solving nonconvex composite optimization problems with nonlinear convex constraints*, Math. Oper. Res., 48 (2023), pp. 1066–1094.
[29] W. Kong, J. G. Melo, and R. D. Monteiro, *Iteration complexity of an inner accelerated inexact proximal augmented Lagrangian method based on the classical Lagrangian function*, SIAM J. Optim., 33 (2023), pp. 181–210.
[30] G. Lan and R. Monteiro, *Iteration-complexity of first-order augmented Lagrangian methods for convex programming*, Math. Program., 155 (2016), pp. 511–547.
[31] Z. Li, P.-Y. Chen, S. Liu, S. Lu, and Y. Xu, *Rate-improved inexact augmented Lagrangian method for constrained nonconvex optimization*, Proc. Mach. Learn. Res. (PMLR), 130 (2021), pp. 2170–2178.
[32] Z. Li and Y. Xu, *Augmented Lagrangian–based first-order methods for convex-constrained programs with weakly convex objective*, INFORMS J. Optim., 3 (2021), pp. 373–397.
[33] Q. Lin, R. Ma, and Y. Xu, *Complexity of an inexact proximal-point penalty method for constrained smooth non-convex optimization,* Comput. Optim. Appl., 82 (2022), pp. 175–224, https://doi.org/10.1007/s10589-022-00358-y.
[34] T. Lin, S. Ma, and S. Zhang, *Global convergence of unmodified 3-block ADMM for a class of convex minimization problems*, J. Sci. Comput., 76 (2018), pp. 69–88.
[35] R. Ma, Q. Lin, and T. Yang, *Proximally Constrained Methods for Weakly Convex Optimization with Weakly Convex Constraints*, preprint, arXiv:1908.01871, 2019.
[36] A. Makhdoumi and A. Ozdaglar, *Broadcast-based distributed alternating direction method of multipliers*, in 2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton), IEEE, Monticello, IL, IEEE, Piscataway, NJ, 2014, pp. 270–277, https://doi.org/10.1109/ALLERTON.2014.7028466.
[37] J. G. Melo and R. D. Monteiro, *Iteration-complexity of a Jacobi-type Non-Euclidean ADMM for Multi-Block Linearly Constrained Nonconvex Programs*, preprint, arXiv:1705.07229, 2017.
[38] J. G. Melo and R. D. Monteiro, *Iteration-complexity of a Linearized Proximal Multi-block ADMM Class for Linearly Constrained Nonconvex Optimization Problems*, 2017, https://optimization-online.org/wp-content/uploads/2017/04/5964.pdf.
[39] J. G. Melo, R. D. C. Monteiro, and H. Wang, *Iteration-complexity of an Inexact Proximal Accelerated Augmented Lagrangian Method for Solving Linearly Constrained Smooth Nonconvex Composite Optimization Problems*, arXiv:2006.08048, 2020.
[40] R. D. C. Monteiro and B. F. Svaiter, *Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers*, SIAM J. Optim., 23 (2013), pp. 475–507.
[41] D. W. Peaceman and H. H. Rachford, Jr., *The numerical solution of parabolic and elliptic differential equations*, J. Soc. Ind. Appl. Math., 3 (1955), pp. 28–41.
[42] M. J. Powell, *A method for nonlinear constraints in minimization problems*, in Optimization, R. Fletcher, ed., Academic, London, 1969, pp. 283–298.
[43] R. T. Rockafellar, *Convex Analysis*, Princeton Math. Ser. 28, Princeton University Press, Princeton, NJ, 1970.
[44] R. T. Rockafellar, *The multiplier method of Hestenes and Powell applied to convex programming*, J. Optim. Theory Appl., 12 (1973), pp. 555–562.

[45] R. T. Rockafellar, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res., 1 (1976), pp. 97–116.
[46] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer, Berlin, 2009.
[47] A. Ruszczynski, *Nonlinear Optimization*, Princeton University Press, Princeton, NJ, 2011.
[48] M. F. Sahin, A. Eftekhari, A. Alacaoglu, F. Latorre, and V. Cevher, *An inexact augmented Lagrangian framework for nonconvex optimization with nonlinear constraints*, in Advances in Neural Information Processing Systems, Curran Associates, Red Hook, NY, 2019, pp. 13966–13978.
[49] Q. Shi and M. Hong, *Penalty dual decomposition method for nonsmooth nonconvex optimization-part* I*: Algorithms and convergence analysis*, IEEE Trans. Signal Process., 68 (2020), pp. 4108–4122.
[50] Q. Shi, M. Hong, X. Fu, and T.-H. Chang, *Penalty dual decomposition method for nonsmooth nonconvex optimization-part* II*: Applications*, IEEE Trans. Signal Process., 68 (2020), pp. 4242–4257.
[51] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, *On the linear convergence of the ADMM in decentralized consensus optimization*, IEEE Trans. Signal Process., 62 (2014), pp. 1750–1761, https://doi.org/10.1109/TSP.2014.2304432.
[52] K. Sun and X. A. Sun, *A two-level ADMM algorithm for AC OPF with global convergence guarantees*, IEEE Trans. Power Syst., 36 (2021), pp. 5271–5281.
[53] K. Sun and X. A. Sun, *Algorithms for difference-of-convex programs based on difference-of-Moreau-envelopes smoothing*, INFORMS J. Optim., 5 (2023), pp. 321–339.
[54] K. Sun and X. A. Sun, *A two-level distributed algorithm for nonconvex constrained optimization*, Comput. Optim. Appl., 84 (2023), pp. 609–649.
[55] Y. Wang, W. Yin, and J. Zeng, *Global convergence of ADMM in nonconvex nonsmooth optimization*, J. Sci. Comput., 78 (2019), pp. 29–63.
[56] Y. Xie and S. J. Wright, *Complexity of proximal augmented Lagrangian for nonconvex optimization with nonlinear equality constraints*, J. Sci. Comput., 86 (2021), pp. 1–30.
[57] Y. Xu and W. Yin, *A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion*, SIAM J. Imaging Sci., 6 (2013), pp. 1758–1789.
[58] Y. Xu and W. Yin, *A globally convergent algorithm for nonconvex optimization based on block coordinate update*, J. Sci. Comput., 72 (2017), pp. 700–734.
[59] J. Zeng, W. Yin, and D.-X. Zhou, *Moreau envelope augmented Lagrangian method for nonconvex optimization with linear constraints*, J. Sci. Comput., 91 (2022), 61.
[60] J. Zhang and Z.-Q. Luo, *A proximal alternating direction method of multiplier for linearly constrained nonconvex minimization*, SIAM J. Optim., 30 (2020), pp. 2272–2302.
[61] J. Zhang and Z.-Q. Luo, *A global dual error bound and its application to the analysis of linearly constrained nonconvex optimization*, SIAM J. Optim., 32 (2022), pp. 2319–2346.
[62] D. Zhu, L. Zhao, and S. Zhang, *A first-order primal-dual method for nonconvex constrained optimization based on the augmented Lagrangian*, Math. Oper. Res., to appear.