Rate-improved Inexact Augmented Lagrangian Method for Constrained Nonconvex Optimization

Zichong Li[1] **Pin-Yu Chen**[*,2] **Sijia Liu**[*,3] **Songtao Lu**[*,2] **Yangyang Xu**[*,1]
[1]Rensselaer Polytechnic Institute [2]IBM Research [3]Michigan State University

Abstract

First-order methods have been studied for nonlinear constrained optimization within the framework of the augmented Lagrangian method (ALM) or penalty method. We propose an improved inexact ALM (iALM) and conduct a unified analysis for nonconvex problems with either affine equality or nonconvex constraints. Under certain regularity conditions (that are also assumed by existing works), we show an $\tilde{O}(\varepsilon^{-\frac{5}{2}})$ complexity result for a problem with a nonconvex objective and affine equality constraints and an $\tilde{O}(\varepsilon^{-3})$ complexity result for a problem with a nonconvex objective and nonconvex constraints, where the complexity is measured by the number of first-order oracles to yield an ε-KKT solution. Both results are the best known. The same-order complexity results have been achieved by penalty methods. However, two different analysis techniques are used to obtain the results, and more importantly, the penalty methods generally perform significantly worse than iALM in practice. Our improved iALM and analysis close the gap between theory and practice. Numerical experiments on nonconvex problems with affine equality or nonconvex constraints are provided to demonstrate the effectiveness of our proposed method.

Proceedings of the 24$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

1 INTRODUCTION

First-order methods (FOMs) have been extensively used for solving large-scale optimization problems, partly due to its nice scalability. Compared to second-order or higher-order methods, FOMs generally have much lower per-iteration complexity and much lower requirement on machine memory. A majority of existing works on FOMs focus on problems without constraints or with simple constraints, e.g., (Nesterov, 2013; Beck and Teboulle, 2009; Ghadimi and Lan, 2016; Carmon et al., 2018; Lu et al., 2020a). Several recent works have made efforts on analyzing FOMs for problems with complicated functional constraints, e.g., (Yu and Neely, 2017; Lin et al., 2018, 2019; Xu, 2019, 2020; Lu and Zhou, 2018; Li and Qu, 2019).

In this paper, we consider *nonconvex* problems with (*possibly nonlinear*) equality constraints, formulated as

$$f_0^* := \min_{\mathbf{x}\in\mathbb{R}^n} \left\{ f_0(\mathbf{x}) := g(\mathbf{x}) + h(\mathbf{x}), \text{ s.t. } \mathbf{c}(\mathbf{x}) = \mathbf{0} \right\}, \quad (1)$$

where g is continuously differentiable but possibly nonconvex, h is closed convex but possibly nonsmooth, and $\mathbf{c} = (c_1, \ldots, c_l) : \mathbb{R}^n \to \mathbb{R}^l$ is a vector function with continuously differentiable components. Note that an inequality constraint $d(\mathbf{x}) \le 0$ can be equivalently formulated as an equality constraint $d(\mathbf{x}) + s = 0$ by enforcing the nonnegativity of s. In addition, the stationary conditions of an inequality-constrained problem and its reformulation can be equivalent, as we will see later at the end of Section 3. Hence, we do not lose generality by focusing on equality-constrained problems in the form of (1). A large class of nonlinear constraint problems can be covered by our formulation. Examples include Neyman-Pearson classification with type-I error (Neyman and Pearson, 1933), resource allocation with nonlinear budgets (Katoh and Ibaraki, 1998), and quadratically constrained quadratic program.

1.1 Related works

The augmented Lagrangian method (ALM) is one of the most popular approaches for solving nonlinear constrained problems. It first appeared in (Powell, 1969;

Hestenes, 1969). Based on the augmented Lagrangian (AL) function, ALM alternatingly updates the primal variable by minimizing the AL function and the Lagrangian multiplier by dual gradient ascent. If the multiplier is fixed to zero, then ALM reduces to a standard penalty method. Early works often used second-order methods, such as the Newton's method, to solve primal subproblems of ALM. With the rapid increase of problem size in modern applications and/or existence of non-differentiable terms, second-order methods become extremely expensive or even inapplicable. Recently, more efforts have been made on integrating first-order solvers into the ALM framework and analyzing the AL-based FOMs.

For convex affinely-constrained problems, (Lan and Monteiro, 2016) presents an AL-based FOM that can produce an ε-KKT point with $O(\varepsilon^{-1}|\log\varepsilon|)$ gradient evaluations and matrix-vector multiplications. This result was extended to convex conic programming (Lu and Zhou, 2018; Aybat and Iyengar, 2013) and to convex nonlinear constrained problems (Li and Xu, 2020; Li and Qu, 2019). When an ε-optimal solution is desired, $O(\varepsilon^{-1})$ complexity results have been established for AL-based FOMs in several works, e.g., (Xu, 2019, 2017; Ouyang et al., 2015; Li and Qu, 2019; Nedelcu et al., 2014). For strongly-convex problems, the complexity results can be respectively improved to $O(\varepsilon^{-\frac{1}{2}}|\log\varepsilon|)$ for an ε-KKT point and $O(\varepsilon^{-\frac{1}{2}})$ for an ε-optimal solution; see (Li and Xu, 2020; Li and Qu, 2019; Xu, 2019; Nedelcu et al., 2014; Necoara and Nedelcu, 2014) for example.

For nonconvex problems with affine equality constraints, (Jiang et al., 2019) can find an ε-KKT solution to a similar variant of our problem (1) with $\tilde{O}(\varepsilon^{-2})$ complexity. Also, (Zhang and Luo, 2020) achieved $O(\varepsilon^{-2})$ complexity for nonconvex smooth problems with polyhedral constraints. However, both of their analysis heavily exploited the affinity of $c(\cdot)$ and didn't include the case of nonconvex $c(\cdot)$.

For problems with nonconvex constraints, early works designed and analyzed FOMs in the framework of a penalty method. (Cartis et al., 2011) first presents an FOM for minimizing composite functions and then applies it to nonlinear constrained nonconvex optimization within the framework of an exact-penalty method. To obtain an ε-KKT point, the FOM by Cartis et al. (2011) needs $O(\varepsilon^{-5})$ gradient evaluations. A follow-up paper by Cartis et al. (2014) gives a trust-region based FOM and shows an $O(\varepsilon^{-2})$ complexity result to produce an ε-Fritz-John point, which is weaker than an ε-KKT point. On solving affinely-constrained nonconvex problems, (Kong et al., 2019) gives a quadratic-penalty-based FOM and establishes an $O(\varepsilon^{-3})$ complexity result to obtain an ε-KKT point. When Slater's condition holds, $\tilde{O}(\varepsilon^{-\frac{5}{2}})$ complexity results have been shown in (Li and Xu, 2020; Lin et al., 2019), which consider nonconvex problems with nonlinear convex constraints. While the FOMs in (Li and Xu, 2020; Lin et al., 2019) are penalty-based, the recent work (Melo et al., 2020) proposes a first-order proximal ALM for affinely-constrained nonconvex problems and obtains an $\tilde{O}(\varepsilon^{-\frac{5}{2}})$ result.

Besides AL and penalty-based FOMs, several other FOMs have been designed to solve nonlinear-constrained problems, such as the level-set FOM by Lin et al. (2018) and the primal-dual method by Yu and Neely (2016) for convex problems. FOMs have also been proposed for minimax problems. For example, (Hien et al., 2017; Hamedani and Aybat, 2018) study FOMs for convex-concave minimax problems, and (Lu et al., 2019; Lu et al., 2020b; Lin et al., 2020) analyzes FOMs for nonconvex-concave minimax problems. While a nonlinear-constrained optimization problem can be formulated as a minimax problem, its KKT conditions are stronger than the stationarity conditions of a nonconvex-concave minimax problem, because the latter with a compact dual domain cannot guarantee primal feasibility. Therefore, stationarity of a minimax problem in (Lin et al., 2020) does not imply primal feasibility of our problem.

1.2 Contributions

Our contributions are three-fold. First, we propose a novel FOM in the framework of *inexact ALM* (iALM) for nonconvex optimization problems with nonlinear (possibly nonconvex) constraints. Due to nonlinearity and large-scale, it is impossible to exactly solve primal subproblems of ALM, and the iALM instead solves each subproblem approximately to a certain desired accuracy. Different from existing works on iALMs, we use an inexact proximal point method (iPPM) to solve each ALM subproblem. The use of iPPM leads to more stable numerical performance and also better theoretical results. Second, we conduct complexity analysis to the proposed iALM. Under a regularity condition, we obtain an $\tilde{O}(\varepsilon^{-\frac{5}{2}})$ result if the constraints are convex and an $\tilde{O}(\varepsilon^{-3})$ result if the constraints are nonconvex. This yields a substantial improvement over the best known complexity results of AL-based FOMs, $\tilde{O}(\varepsilon^{-3})$ (Li and Xu, 2020) and $\tilde{O}(\varepsilon^{-4})$ (Sahin et al., 2019) (see Remark 1) respectively for the aforementioned convex and nonconvex constrained cases. While quadratic-penalty-based FOMs (under the same regularity condition as what we assume for nonconvex-constraint problems) (Lin et al., 2019) have achieved the same-order results as ours, but their empirical performance is generally (much) worse. Hence, our results close the gap between theory and practice. Thirdly,

Table 1: Comparison of the complexity results of several methods in the literature to our method to produce an ε-KKT solution to (1).

Method	type	objective	constraint	regularity	complexity
iALM (Li and Xu, 2020)	AL	strongly convex	convex	none	$\tilde{O}(\varepsilon^{-\frac{1}{2}})$
		convex	convex	none	$\tilde{O}(\varepsilon^{-1})$
QP-AIPP (Kong et al., 2019)	penalty	nonconvex	convex	none	$\tilde{O}(\varepsilon^{-3})$
HiAPeM (Li and Xu, 2020)	hybrid	nonconvex	convex	Slater's condition	$\tilde{O}(\varepsilon^{-\frac{5}{2}})$
iPPP (Lin et al., 2019)	penalty	nonconvex	convex	Slater's condition	$\tilde{O}(\varepsilon^{-\frac{5}{2}})$
			nonconvex	none	$\tilde{O}(\varepsilon^{-4})$
			nonconvex	Assumption 3	$\tilde{O}(\varepsilon^{-3})$
iALM (Sahin et al., 2019)	AL	nonconvex	nonconvex	Assumption 3	$\tilde{O}(\varepsilon^{-4})$
this paper	AL	nonconvex	convex	Assumption 3	$\tilde{O}(\varepsilon^{-\frac{5}{2}})$
			nonconvex	Assumption 3	$\tilde{O}(\varepsilon^{-3})$

our algorithm and analysis are unified for the convex-constrained and nonconvex-constrained cases. Existing works on penalty-based FOMs such as (Lin et al., 2019) need different algorithmic designs and also different analysis techniques to obtain the $\tilde{O}(\varepsilon^{-\frac{5}{2}})$ and $\tilde{O}(\varepsilon^{-3})$ results, separately for the convex-constrained and nonconvex-constrained cases.

Remark 1. *An $\tilde{O}(\varepsilon^{-3})$ complexity is claimed in Corollary 4.2 in (Sahin et al., 2019). However, this complexity is based on an existing result that was not correctly referred to. The authors claimed that the complexity of solving each nonconvex composite subproblem is $O\left(\frac{\lambda_{\beta_k}^2\rho^2}{\varepsilon_{k+1}}\right)$, which should be $O\left(\frac{\lambda_{\beta_k}^2\rho^2}{\varepsilon_{k+1}^2}\right)$; see (Sahin et al., 2019) for the definitions of $\lambda_{\beta_k}, \rho, \varepsilon_{k+1}$. Using the correctly referred result and following the same proof in (Sahin et al., 2019), we get a total complexity of $\tilde{O}(\varepsilon^{-4})$.*

1.3 Complexity comparison on different methods

In Table 1, we summarize our complexity results and several existing ones of first order methods to produce an ε-KKT solution to (1). We consider several cases based on whether the objective and the constraints are convex. Here, constraints being convex means that the feasible set is convex, or in other words, equality constraint functions must be affine and inequality constraint functions must be convex. Our result matches the best-known existing results, which are achieved by penalty-type methods such as the iPPP by Lin et al. (2019). In practice, AL-type methods usually significantly outperform penalty-type methods. Hence, our method is competitive in theory and can be significantly better in practice, as we demonstrated in the numerical experiments.

1.4 Notations, definitions, and assumptions

We use $\|\cdot\|$ for the Euclidean norm of a vector and the spectral norm of a matrix. For a positive integer, $[n]$ denotes the set $\{1,\ldots,n\}$. The big-O notation is used with standard meaning, while $\tilde{O}$ suppresses all logarithmic terms of ε. Given $\mathbf{x}\in\mathrm{dom}(h)$, we denote $J_c(\mathbf{x})$ as the Jacobi matrix of $\mathbf{c}$ at $\mathbf{x}$. We denote the distance function between a vector $\mathbf{x}$ and a set $\mathcal{X}$ as $\mathrm{dist}(\mathbf{x},\mathcal{X})=\min_{\mathbf{y}\in\mathcal{X}}\|\mathbf{x}-\mathbf{y}\|$. The augmented Lagrangian (AL) function of (1) is

$$\mathcal{L}_\beta(\mathbf{x},\mathbf{y})=f_0(\mathbf{x})+\mathbf{y}^\top\mathbf{c}(\mathbf{x})+\frac{\beta}{2}\|\mathbf{c}(\mathbf{x})\|^2, \qquad (2)$$

where $\beta>0$ is a penalty parameter, and $\mathbf{y}\in\mathbb{R}^l$ is the multiplier vector.

Definition 1 (ε-KKT point). *Given $\varepsilon\geq 0$, a point $\mathbf{x}\in\mathbb{R}^n$ is called an ε-KKT point to (1) if there is a vector $\mathbf{y}\in\mathbb{R}^l$ such that*

$$\|\mathbf{c}(\mathbf{x})\|\leq\varepsilon,\quad \mathrm{dist}\left(\mathbf{0},\partial f_0(\mathbf{x})+J_c^\top(\mathbf{x})\ \mathbf{y}\right)\leq\varepsilon. \qquad (3)$$

Definition 2 (L-smoothness). *A differentiable function f on $\mathbb{R}^n$ is L-smooth if $\|\nabla f(\mathbf{x}_1)-\nabla f(\mathbf{x}_2)\|\leq L\|\mathbf{x}_1-\mathbf{x}_2\|$ for all $\mathbf{x}_1,\mathbf{x}_2\in\mathbb{R}^n$.*

Definition 3 (ρ-weakly convex). *A function g is ρ-weakly convex if $g+\frac{\rho}{2}\|\cdot\|^2$ is convex.*

Remark 2. *If f is L-smooth, then it is also L-weakly convex. However, the weak-convexity constant of a differentiable function can be much smaller than its smoothness constant. For example, if $f(\mathbf{x})=\frac{1}{2}\mathbf{x}^\top\mathbf{Q}\mathbf{x}+\mathbf{c}^\top\mathbf{x}$ where $\mathbf{Q}$ is a symmetric but indefinite matrix, then the smoothness constant of f is $\|\mathbf{Q}\|$, and its weak-convexity constant is the negative of the smallest eigenvalue of $\mathbf{Q}$.*

Throughout the paper, we make the following assumptions about (1). Examples that satisfy these assumptions will be given in the experimental section.

Assumption 1 (smoothness and weak convexity)**.** *The function g in the objective of* (1) *is L_0-smooth and ρ_0-weakly convex. For each $j \in [l]$, c_j is L_j-smooth and ρ_j-weakly convex.*

Assumption 2 (bounded domain)**.** *h is a simple closed convex function with a compact domain, i.e.,*

$$D =: \max_{\mathbf{x},\mathbf{x}' \in \mathrm{dom}(h)} \|\mathbf{x} - \mathbf{x}'\| < \infty. \tag{4}$$

2 A NOVEL AL-BASED FOM WITH IMPROVED CONVERGENCE RATE

In this section, we present a novel FOM (see Algorithm 3 below) for solving (1). It follows the standard ALM framework, similar to AL-based FOMs (Sahin et al., 2019; Xu, 2019). Notably, different from existing works, we use an inexact proximal point method (iPPM) to approximately solve each ALM subproblem. The complexity result of iPPM has the best dependence on the smoothness constant. This enables us to obtain order-reduced complexity results by geometrically increasing the penalty parameter in ALM, as compared to the AL-based FOMs (Li and Xu, 2020; Sahin et al., 2019) for nonconvex constrained optimization. Our whole algorithm has three layers. We analyze the inner algorithm 1 in Section 2.1, the middle algorithm 2 in Section 2.2, and the outer algorithm 3 in Section 2.3.

Algorithm 1: Accelerated proximal gradient method: APG$(G, H, \mu, L_G, \varepsilon)$

Initialization: choose $\bar{\mathbf{x}}^{-1} \in \mathrm{dom}(H)$ and set $\alpha = \sqrt{\frac{\mu}{L_G}}$; let

$$\bar{\mathbf{x}}^0 = \mathbf{x}^0 = \arg\min_{\mathbf{x}} \langle \nabla G(\bar{\mathbf{x}}^{-1}), \mathbf{x} \rangle + \frac{L_G}{2} \left\| \mathbf{x} - \bar{\mathbf{x}}^{-1} \right\|^2 + H(\mathbf{x}).$$

for $t = 0, 1, \ldots$ **do**

  Update the iterate by

$$\mathbf{x}^{t+1} = \arg\min_{\mathbf{x}} \langle \nabla G(\bar{\mathbf{x}}^t), \mathbf{x} \rangle + \frac{L_G}{2} \left\| \mathbf{x} - \bar{\mathbf{x}}^t \right\|^2 + H(\mathbf{x}), \tag{5}$$

$$\bar{\mathbf{x}}^{t+1} = \mathbf{x}^{t+1} + \tfrac{1-\alpha}{1+\alpha}(\mathbf{x}^{t+1} - \mathbf{x}^t). \tag{6}$$

  if $\mathrm{dist}\big(-\nabla G(\mathbf{x}^{t+1}), \partial H(\mathbf{x}^{t+1})\big) \le \varepsilon$, **then** output $\mathbf{x}^{t+1}$ and stop.

2.1 Accelerated proximal gradient (APG) method for convex composite problems

The kernel problems that we solve are a sequence of convex composite problems in the form of

$$\operatorname*{minimize}_{\mathbf{x} \in \mathbb{R}^n} \ F(\mathbf{x}) := G(\mathbf{x}) + H(\mathbf{x}), \tag{7}$$

where G is μ-strongly convex and L_G-smooth, and H is a closed convex function. Various optimal FOMs (e.g., Nesterov (2013, 2004); Beck and Teboulle (2009)) have been designed to solve (7). We choose the FOM used by Li and Xu (2020) for the purpose of obtaining near-stationary points. Its pseudocode is given in Algorithm 1.

The next lemma is from (Li and Xu, 2020, Lemma 3). It gives the complexity result of Algorithm 1.

Lemma 1. *Given $\varepsilon > 0$, within at most T iterations, Algorithm 1 will output a solution $\mathbf{x}^T$ that satisfies $\mathrm{dist}\big(\mathbf{0}, \partial F(\mathbf{x}^T)\big) \le \varepsilon$, where*

$$T = \left\lceil \sqrt{\tfrac{L_G}{\mu}} \log \frac{64 L_G^2 \big(L_G \|\mathbf{x}^{-1} - \mathbf{x}^*\|^2 + \mu \|\mathbf{x}^* - \mathbf{x}^0\|^2\big)}{\varepsilon^2 \mu} + 1 \right\rceil.$$

2.2 Inexact proximal point method (iPPM) for nonconvex composite problems

Each primal subproblem of the ALM for (1) is a nonconvex composite problem in the form of

$$\Phi^* = \operatorname*{minimize}_{\mathbf{x} \in \mathbb{R}^n} \ \big\{\Phi(\mathbf{x}) := \phi(\mathbf{x}) + \psi(\mathbf{x})\big\}, \tag{8}$$

where ϕ is L_ϕ-smooth and ρ-weakly convex, and ψ is closed convex. We propose to use the iPPM to approximately solve the ALM subproblems. The iPPM framework has appeared in (Kong et al., 2019). Different from (Kong et al., 2019), we propose to use APG in Algorithm 1 to solve each iPPM subproblem. The pseudocode of our iPPM is shown in Algorithm 2. It appears that our iPPM has more stable numerical performance.

Algorithm 2: Inexact proximal point method (iPPM) for (8): iPPM$(\phi, \psi, \mathbf{x}^0, \rho, L_\phi, \varepsilon)$

Input: $\mathbf{x}^0 \in \mathrm{dom}(\psi)$, smoothness L_ϕ, weak convexity ρ, stationarity tolerance ε

for $k = 0, 1, \ldots,$ **do**

  Let $G(\cdot) = \phi(\cdot) + \rho \| \cdot - \mathbf{x}^k \|^2$

  Call Algorithm 1 to obtain $\mathbf{x}^{k+1} \leftarrow \mathrm{APG}(G, \psi, \rho, L_\phi + 2\rho, \frac{\varepsilon}{4})$

  if $2\rho \|\mathbf{x}^{k+1} - \mathbf{x}^k\| \le \frac{\varepsilon}{2}$, **then** return $\mathbf{x}^{k+1}$.

The next theorem gives the complexity result. Its proof is given in the supplementary materials.

Theorem 1. *Suppose Φ^* is finite. Algorithm 2 must stop within T iterations, where*

$$T = \left\lceil \frac{32\rho}{\varepsilon^2}(\Phi(\mathbf{x}^0) - \Phi^*) \right\rceil. \tag{9}$$

The output $\mathbf{x}^S$ must be an ε-stationary point of (8), i.e., $\mathrm{dist}(\mathbf{0}, \partial\Phi(\mathbf{x}^S)) \le \varepsilon$. In addition, if $\mathrm{dom}(\psi)$ is compact and has diameter $D_\psi < \infty$, then the total complexity is $O\left(\frac{\sqrt{\rho L_\phi}}{\varepsilon^2}[\Phi(\mathbf{x}^0) - \Phi^*] \log \frac{D_\psi}{\varepsilon}\right)$.

Remark 3. *A similar result has been shown by Kong et al. (2019). It has better dependence on L_ϕ than that by Ghadimi and Lan (2016). In addition, in the worst case, $\Phi(\mathbf{x}^0) - \Phi^*$ is in the same order of L_ϕ. However, we will see that for our case, $\Phi(\mathbf{x}^0) - \Phi^*$ can be uniformly bounded when Algorithm 2 is applied to solve subproblems of ALM even if the penalty parameter (that is proportional to the smooth constant) in the AL function geometrically increases. As a result, when the smoothness and weak convexity parameters are both $O(\varepsilon^{-1})$, we can obtain a total complexity of $\tilde{O}\left(\frac{\sqrt{\rho L_\Phi}}{\varepsilon^2}\right) = \tilde{O}(\varepsilon^{-3})$, which is better than $\tilde{O}(\frac{L_\Phi^2}{\varepsilon^2}) = \tilde{O}(\varepsilon^{-4})$ obtained if the method in (Ghadimi and Lan, 2016) is applied. This is the key for us to have order-reduced complexity results, as compared to (Sahin et al., 2019).*

2.3 Inexact augmented Lagrangian method (iALM) for nonlinear constrained problems

Now we are ready to present an improved AL-based FOM for solving (1). Different from existing AL-based FOMs, our method uses iPPM, given in Algorithm 2, to approximately solve each subproblem, and also its dual step size is adaptive to the primal residual. The pseudocode is shown in Algorithm 3.

In the algorithm and the later analysis, we denote

$$B_0 = \max_{\mathbf{x} \in \mathrm{dom}(h)} \max\left\{|f_0(\mathbf{x})|, \|\nabla g(\mathbf{x})\|\right\},$$

$$B_c = \max_{\mathbf{x} \in \mathrm{dom}(h)} \|J_c(\mathbf{x})\|, \tag{10a}$$

$$B_i = \max_{\mathbf{x} \in \mathrm{dom}(h)} \max\left\{|c_i(\mathbf{x})|, \|\nabla c_i(\mathbf{x})\|\right\}, \forall\, i \in [l], \tag{10b}$$

$$\bar{B}_c = \sqrt{\textstyle\sum_{i=1}^l B_i^2}, \quad \bar{L} = \sqrt{\textstyle\sum_{i=1}^l L_i^2},$$

$$\rho_c = \sum_{i=1}^l B_i \rho_i, \quad L_c = \sum_{i=1}^l B_i L_i + B_i^2, \tag{10c}$$

where $\{\rho_i\}$ and $\{L_i\}$ are given in Assumption 1. Note that the above constants are all finite under Assumptions 1 and 2, and we do not need to evaluate them exactly but only need upper bounds.

Algorithm 3 follows the standard framework of the ALM. The existing method that is the closest to ours is

Algorithm 3: Inexact augmented Lagrangian method (iALM) for (1)

Initialization: choose $\mathbf{x}^0 \in \mathrm{dom}(f_0), \mathbf{y}^0 = \mathbf{0}$, $\beta_0 > 0$ and $\sigma > 1$

for $k = 0, 1, \ldots,$ **do**

Let $\beta_k = \beta_0\sigma^k$, $\phi(\cdot) = \mathcal{L}_{\beta_k}(\cdot, y^k) - h(\cdot)$, and

$$\hat{\rho}_k = \rho_0 + \bar{L}\|\mathbf{y}^k\| + \beta_k\rho_c, \qquad \hat{L}_k = L_0 + \bar{L}\|\mathbf{y}^k\| + \beta_k L_c. \tag{11}$$

Call Algorithm 2 to obtain $\mathbf{x}^{k+1} \leftarrow \mathrm{iPPM}(\phi, h, \mathbf{x}^k, \hat{\rho}_k, \hat{L}_k, \varepsilon)$

Update $\mathbf{y}$ by

$$\mathbf{y}^{k+1} = \mathbf{y}^k + w_k \mathbf{c}(\mathbf{x}^{k+1}), \tag{12}$$

where

$$w_k = w_0 \min\left\{1, \frac{\gamma_k}{\|\mathbf{c}(\mathbf{x}^{k+1})\|}\right\}. \tag{13}$$

the iALM by Sahin et al. (2019). The main difference is that we use the iPPM to solve ALM subproblems, while (Sahin et al., 2019) applies the FOM by Ghadimi and Lan (2016). This change of subroutine, together with our new analysis, leads to order-reduced complexity results under the same assumptions. In addition, we observed from the experiments that our iPPM is more stable and more efficient on solving nonconvex subproblems than the subsolver by Sahin et al. (2019).

3 COMPLEXITY RESULTS

In this section, we analyze the complexity result of Algorithm 3. In general, it is difficult to show convergence rates of AL-based FOMs on nonconvex constrained problems mainly due to two reasons. First, a stationary point of the AL function may not be (near) feasible, even a large penalty parameter is used. This is essentially different from penalty-based FOMs. Second, the Lagrangian multiplier cannot be bounded if the dual step size is not carefully set. We show that, with a regularity condition and a well-controlled dual step size, our AL-based FOM can circumvent both issues and achieve best-known convergence rates.

For simplicity, we let

$$\gamma_k = \frac{(\log 2)^2 \|\mathbf{c}(\mathbf{x}^1)\|}{(k+1)[\log(k+2)]^2}, \tag{14}$$

which has been adopted by Sahin et al. (2019). This choice of γ_k will lead to a uniform bound on $\{\mathbf{y}^k\}$ and simplify our analysis. More complicated analysis with general $\{\gamma_k\}$ is given in the supplementary materials.

It is impossible to find a (near) feasible solution of a general nonlinear system in polynomial time. Hence, a certain regularity condition is necessary in order to guarantee near-feasibility. Following (Sahin et al., 2019; Lin et al., 2019), we assume the regularity condition below on (1).

Assumption 3 (regularity)**.** *There is some $v > 0$ such that for any $k \geq 1$,*

$$v\|\mathbf{c}(\mathbf{x}^k)\| \leq \operatorname{dist}\left(-J_c(\mathbf{x}^k)^\top \mathbf{c}(\mathbf{x}^k), \frac{\partial h(\mathbf{x}^k)}{\beta_{k-1}}\right). \quad (15)$$

Remark 4. *The intuition of using Assumption 3 is to ensure near feasibility of a near-stationary point to the AL function. Without any regularity conditions on the nonconvex constraints, one cannot even achieve feasibility. It is unclear whether Assumption 3 is stronger or weaker than other common regularity conditions such as the Slater's condition and the MFCQ condition.*

Several nonconvex examples that satisfy the regularity condition are given in (Sahin et al., 2019) and (Lin et al., 2019), such as the EV and clustering problems in our experiments. In Section 3.1, we prove that Assumption 3 holds for all affine equality constrained problems possibly with an additional polyhedral constraint set or a ball constraint set. Hence, the LCQP problem in our experiments also has this property. Notice that we only require the existence of v but do not need to know its value in our algorithm.

3.1 Convex constraint examples with regularity condition

In this subsection, we show that the regularity condition in Assumption 3 can hold for the LCQP problem (22) that we will test. We prove this for a broader class of problems, namely, affine-equality constraints problems with an additional polyhedral constraint set or $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} = \mathbf{b}, \|\mathbf{x}\| \leq 1\}$. The proofs are given in the supplementary materials.

3.1.1 polyhedral constraint

Let $X \subseteq \mathbb{R}^n$ be a compact polyhedral set and $h(\cdot) = \iota_X(\cdot)$ be the indicator function on X. Then for any $\beta > 0$ and $\mathbf{x} \in X$, $\frac{\partial h(\mathbf{x})}{\beta} = \mathcal{N}_X(\mathbf{x})$, where $\mathcal{N}_X$ denotes the normal cone. We have the result in the claim below.

Claim 1. *If $X \cap \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} = \mathbf{b}\} \neq \emptyset$, then there is a constant $v > 0$ such that $\forall\, \mathbf{x} \in X$,*

$$v\|\mathbf{A}\mathbf{x} - \mathbf{b}\| \leq \operatorname{dist}\left(\mathbf{0}, \mathbf{A}^\top(\mathbf{A}\mathbf{x} - \mathbf{b}) + \mathcal{N}_X(\mathbf{x})\right), \quad (16)$$

which implies (15) *with* $\mathbf{c}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$ *and* $h(\mathbf{x}) = \iota_X(\mathbf{x})$.

By this claim, we let $X = \{\mathbf{x} \in \mathbb{R}^n : l_i \leq x_i \leq u_i, \forall\, i \in [n]\}$ and immediately have that the LCQP problem (22) satisfies the regularity condition in Assumption 3.

3.1.2 ball constraint

Let $X = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \leq r\}$ be a ball of radius $r > 0$ and h be the indicator function on X. Then we have the following result.

Claim 2. *Suppose $\mathbf{A}$ has full row-rank. In addition, there exists a $\hat{\mathbf{x}}$ in the interior of X such that $\mathbf{A}\hat{\mathbf{x}} = \mathbf{b}$. Then there is a constant $v > 0$ such that* (16) *holds.*

3.2 Main Theorems

We give the main convergence results in this subsection. Detailed proofs are provided in the supplementary materials.

Theorem 2 (total complexity of iALM)**.** *Suppose that all conditions in Assumptions 1 through 3 hold. Given $\varepsilon > 0$, then Algorithm 3 with γ_k given in* (14) *needs* $\tilde{O}(\varepsilon^{-3})$ APG *iterations to produce an ε-KKT solution of* (1)*. In addition, if $\mathbf{c}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$, then $\tilde{O}(\varepsilon^{-\frac{5}{2}})$* APG *iterations are needed to produce an ε-KKT solution of* (1)*.*

Remark 5. *For nonconvex-constrained cases, our complexity result $\tilde{O}(\varepsilon^{-3})$ is better than the result $\tilde{O}(\varepsilon^{-4})$ obtained in (Sahin et al., 2019) for an first-order iALM, and it matches the complexity result of a penalty-based FOM by Lin et al. (2019). For the affine equality-constrained case with a composite objective, our result $\tilde{O}(\varepsilon^{-\frac{5}{2}})$ is better than $\tilde{O}(\varepsilon^{-3})$ obtained by Li and Xu (2020) for a first-order iALM and matches the result of the penalty-based FOM by Lin et al. (2019). Numerically, the iALM-based FOM usually significantly outperforms a penalty-based FOM, as shown in our experiments.*

In Theorem 2, we required the dual step size $w_k = w_0 \min\left\{1, \frac{\gamma_k}{\|\mathbf{c}(\mathbf{x}^{k+1})\|}\right\}$, where as in (14),

$$\gamma_k = \frac{(\log 2)^2 \|\mathbf{c}(\mathbf{x}^1)\|}{(k+1)[\log(k+2)]^2}.$$

Numerically, we observed better performance by slightly deviating from this setting. For example, we set $w_k = \frac{1}{\|\mathbf{c}(\mathbf{x}^{k+1})\|}$ in all of our trials. This motivates us to give a more general version of Theorem 2. The following theorem considers $w_k = \frac{O(k^q)}{\mathbf{c}(\mathbf{x}^{k+1})}$ and sacrifices an order of $(\log \varepsilon^{-1})^{q+1}$ in the total complexity compared to Theorem 2.

Theorem 3 (complexity of iALM with general dual step sizes)**.** *In Algorithm 3, for some fixed $q \in \mathbb{Z}_+ \cup \{0\}$ and $M > 0$, let*

$$w_k = \frac{M(k+1)^q}{\|\mathbf{c}(\mathbf{x}^{k+1})\|}, \forall k \geq 0. \quad (17)$$

Assume all other conditions of Theorem 2 hold. Then given $\varepsilon > 0$*, Algorithm 3 with* w_k *given in* (17) *needs* $\tilde{O}(\varepsilon^{-3})$ APG *iterations to produce an* ε*-KKT solution of* (1)*. In addition, if* $\mathbf{c}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$*, then* $\tilde{O}(\varepsilon^{-\frac{5}{2}})$ APG *iterations are needed to produce an* ε*-KKT solution of* (1)*.*

The proof of the above theorem is very similar to the proof of Theorem 2, except we have a nonuniform bound on the dual variable.

Remark 6 (inequality constraints)**.** *Although only equality constraints are considered in* (1)*, our complexity result does not lose generality due to the boundedness of* $\{\mathbf{y}^k\}$*. Suppose we solve a problem with both equality and inequality constraints*

$$\underset{\mathbf{x}}{\text{minimize}}\, f_0(\mathbf{x}),\ \text{s.t. } \mathbf{c}(\mathbf{x}) = \mathbf{0}, \mathbf{d}(\mathbf{x}) \le \mathbf{0}. \tag{18}$$

Introducing a slack variable $\mathbf{s} \ge \mathbf{0}$*, we can have an equivalent formulation*

$$\underset{\mathbf{x},\mathbf{s}\ge\mathbf{0}}{\text{minimize}}\, f_0(\mathbf{x}),\ \text{s.t. } \mathbf{c}(\mathbf{x}) = \mathbf{0},\ \mathbf{d}(\mathbf{x}) + \mathbf{s} = \mathbf{0}. \tag{19}$$

Suppose the conditions required by Theorem 2 hold. Then we can apply Algorithm 3 to (19) *and obtain an* ε*-KKT point* $(\bar{\mathbf{x}}, \bar{\mathbf{s}})$ *with a corresponding multiplier* $(\bar{\mathbf{y}}, \bar{\mathbf{z}})$*, i.e.,*

$$\text{dist}\left(\mathbf{0}, \begin{bmatrix} \partial f_0(\bar{\mathbf{x}}) \\ \mathcal{N}_+(\bar{\mathbf{s}}) \end{bmatrix} + \begin{bmatrix} J_c(\bar{\mathbf{x}})^\top \\ \mathbf{0} \end{bmatrix} \bar{\mathbf{y}} + \begin{bmatrix} J_d(\bar{\mathbf{x}})^\top \\ \mathbf{I} \end{bmatrix} \bar{\mathbf{z}}\right) \le \varepsilon, \tag{20a}$$

$$\|\mathbf{c}(\bar{\mathbf{x}})\|^2 + \|\mathbf{d}(\bar{\mathbf{x}}) + \bar{\mathbf{s}}\|^2 \le \varepsilon^2,\ \bar{\mathbf{s}} \ge \mathbf{0}, \tag{20b}$$

where $\mathcal{N}_+(\mathbf{s})$ *denotes the normal cone of the nonnegative orthant at* $\mathbf{s}$*.*

By (20a) *and the definition of the normal cone, we have* $\|[\bar{\mathbf{z}}]_-\| \le \varepsilon$*. Let* $\hat{\mathbf{z}} = \bar{\mathbf{z}} - [\bar{\mathbf{z}}]_-$*. Then* $\hat{\mathbf{z}} \ge \mathbf{0}$*, and if* $\|J_d(\cdot)\|$ *is uniformly bounded, then it follows from* (20a) *that*

$$\text{dist}\left(\mathbf{0}, \partial f_0(\bar{\mathbf{x}}) J_c(\bar{\mathbf{x}})^\top \bar{\mathbf{y}} + J_d(\bar{\mathbf{x}})^\top \hat{\mathbf{z}}\right) = O(\varepsilon). \tag{21}$$

In addition, from (20b)*, it is straightforward to have* $\|\mathbf{c}(\bar{\mathbf{x}})\|^2 + \|[\mathbf{d}(\bar{\mathbf{x}})]_+\|^2 \le \varepsilon^2$*. Furthermore, notice that if some* $\bar{s}_i = 0$*, then* $|d_i(\bar{\mathbf{x}})| \le \varepsilon$ *from* (20b)*, and if* $\bar{s}_i > 0$*, then* $|\bar{z}_i| \le \varepsilon$ *from* (20a)*. Finally, use the boundedness of* $\mathbf{d}$ *and the fact that* $\|\bar{\mathbf{z}}\| = O(1)$ *is independent of* ε *from the proof of Theorem 2 to have* $|\hat{\mathbf{z}}^\top \mathbf{d}(\bar{\mathbf{x}})| = O(\varepsilon)$*. Therefore,* $\bar{\mathbf{x}}$ *is an* $O(\varepsilon)$*-KKT point of the original problem* (18)*, in terms of primal feasibility, dual feasibility, and the complementarity condition.*

4 NUMERICAL RESULTS

In this section, we conduct experiments to demonstrate the empirical performance of the proposed improved iALM. We consider the nonconvex linearly-constrained quadratic program (LCQP), generalized eigenvalue problem (EV), and clustering problem. We compare our method to the iALM by Sahin et al. (2019) for all three problems, and also to HiAPeM by Li and Xu (2020) for the LCQP problem. All the tests were performed in MATLAB 2019b on a Macbook Pro with 4 cores and 16GB memory. Due to the page limitation, we put some tables with more details in the supplementary materials.

4.1 Nonconvex linearly-constrained quadratic programs (LCQP)

In this subsection, we test the proposed method on solving nonconvex LCQP:

$$\begin{aligned} &\min_{\mathbf{x}\in\mathbb{R}^n} \tfrac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x} + \mathbf{c}^\top \mathbf{x}, \\ &\text{s.t. } \mathbf{A}\mathbf{x} = \mathbf{b},\ x_i \in [l_i, u_i], \forall\, i \in [n], \end{aligned} \tag{22}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$, and $\mathbf{Q} \in \mathbb{R}^{n\times n}$ is symmetric and indefinite (thus the objective is nonconvex). In the test, we generated all data randomly. The smallest eigenvalue of $\mathbf{Q}$ is $-\rho < 0$, and thus the problem is ρ-weakly convex. For all tested instances, we set $l_i = -5$ and $u_i = 5$ for each $i \in [n]$.

We generated two groups of LCQP instances of different sizes. The first group had $m = 10$ and $n = 200$ and the second one $m = 100$ and $n = 1000$. In each group, we generated 10 instances of LCQP with $\rho = 1$. We compared the improved iALM in Algorithm 3 to the iALM by Sahin et al. (2019) and the HiAPeM method by Li and Xu (2020). HiAPeM adopted a hybrid setting ($N_0 = 10, N_1 = 2$) and a pure-penalty setting ($N_0 = 1, N_1 = 10^6$), where N_0 is the number of initial iALM calls, and afterwards, N_1 is roughly the number of penalty method calls before each iALM call). The AL function $\mathcal{L}_\beta(\cdot, \mathbf{y})$ of LCQP is $\|\mathbf{Q} + \beta\mathbf{A}^\top\mathbf{A}\|$-smooth and ρ-weakly convex. We set $\beta_k = \sigma^k\beta_0$ with $\sigma = 3$ and $\beta_0 = 0.01$ for both iALMs. For the subsolver of the iALM by Sahin et al. (2019), we set its step size to $\frac{1}{2\|\mathbf{Q}+\beta_k\mathbf{A}^\top\mathbf{A}\|}$ for the k-th outer iteration, as specified by Ghadimi and Lan (2016). The tolerance was set to $\varepsilon = 10^{-3}$ for all instances. In addition, we set the maximum inner iteration to 10^6 for all methods.

In the top row of Figure 1, we compare the primal residual trajectories of our method and the iALM by Sahin et al. (2019) on one representative instance of (22). Note the dual residuals of both methods are below error tolerance at the end of each outer loop. In Tables 2 and 3, we report, for each method, the primal residual, dual residual, running time (in seconds), and the number of gradient evaluation, shortened as `pres`, `dres`, `time`, and `#Grad`, averaged across all ten trials. The complete tables are given in the appendix.

Table 2: Results by the proposed improved iALM, the iALM by Sahin et al. (2019), and the HiAPeM by Li and Xu (2020) on solving a 1-weakly convex LCQP (22) of size $m = 10$ and $n = 200$.

method	pres	dres	time	#Grad
proposed improved iALM	4.10e-4	5.49e-4	1.44	34294
iALM in (Sahin et al., 2019)	5.26e-4	1.00e-3	11.03	1235210
HiAPeM with $N_0 = 10, N_1 = 2$	1.97e-4	7.53e-4	2.71	172395
HiAPeM with $N_0 = 1, N_1 = 10^6$	3.52e-4	8.20e-4	6.08	493948

Table 3: Results by the proposed improved iALM, the iALM by Sahin et al. (2019), and the HiAPeM by Li and Xu (2020) on solving a 1-weakly convex LCQP (22) of size $m = 100$ and $n = 1000$.

method	pres	dres	time	#Grad
proposed improved iALM	5.57e-4	8.81e-4	135.47	278395
iALM in (Sahin et al., 2019)	4.45e-4	3.37e-3	1782.6	11186171
HiAPeM with $N_0 = 10, N_1 = 2$	3.61e-4	8.16e-4	585.84	3081631
HiAPeM with $N_0 = 1, N_1 = 10^6$	5.46e-4	9.01e-4	991.54	5738336

From the results, we conclude that, to reach an ε-KKT point to the LCQP problem, the proposed improved iALM needs significantly fewer gradient evaluations and takes far less time than all other compared methods.

4.2 Generalized eigenvalue problem

In this subsection, we consider the generalized eigenvalue problem (EV) and compare our method to the iALM by Sahin et al. (2019).

The EV problem is

$$\min_{\mathbf{x}\in\mathbb{R}^n} \mathbf{x}^\top \mathbf{Q}\mathbf{x}, \text{ s.t. } \mathbf{x}^\top \mathbf{B}\mathbf{x} - 1 = 0, \tag{23}$$

where $\mathbf{Q}, \mathbf{B} \in \mathbb{R}^{n\times n}$ are symmetric, and $\mathbf{B}$ is positive definite. In the test, we set $\mathbf{Q} = \frac{1}{2}(\hat{\mathbf{Q}} + \hat{\mathbf{Q}}^\top)$ with the entries of $\hat{\mathbf{Q}}$ independently following from the standard Gaussian $\mathcal{N}(0, 1)$. To ensure $\mathbf{B}$ to be positive definite, we set $\mathbf{B} = \bar{\mathbf{B}} + (\|\bar{\mathbf{B}}\| + 1)\mathbf{I}_{n\times n}$, where $\bar{\mathbf{B}}$ is generated in the same way as $\mathbf{Q}$. The regularity condition in Assumption 3 has been shown for (23) by Sahin et al. (2019). However, we do not have an explicit compact constraint set, i.e., Assumption 2 is violated. Nevertheless, the feasible region of (23) is bounded because of the positive definiteness of $\mathbf{B}$. Hence, the tested methods can still perform well.

Again, we generated two groups of instances of (23), one with $n = 200$ and the other $n = 1000$. Each group consisted of 10 instances. For (23), we were unable to obtain an explicit formula of the smoothness constant L_k and weak convexity constant ρ_k of the AL function $\mathcal{L}_{\beta_k}(\cdot, \mathbf{y}^k)$ for any k. The iALM by Sahin et al. (2019) used the accelerated first-order method by Ghadimi and Lan (2016) as a subroutine. We observed divergence by performing line search to estimate a local smoothness constant. In order to make it converge, we tuned its smoothness constant to $L_k = 2\|\mathbf{Q}\| + 1000 + 100\beta_k$ when $n = 200$, and $L_k = 2\|\mathbf{Q}\| + 100000 + 10000\beta_k$ when $n = 1000$. The weak convexity constant was tuned to $\rho_k = -0.2 \cdot \min(\mathrm{eig}(\mathbf{Q})) + \beta_k$.

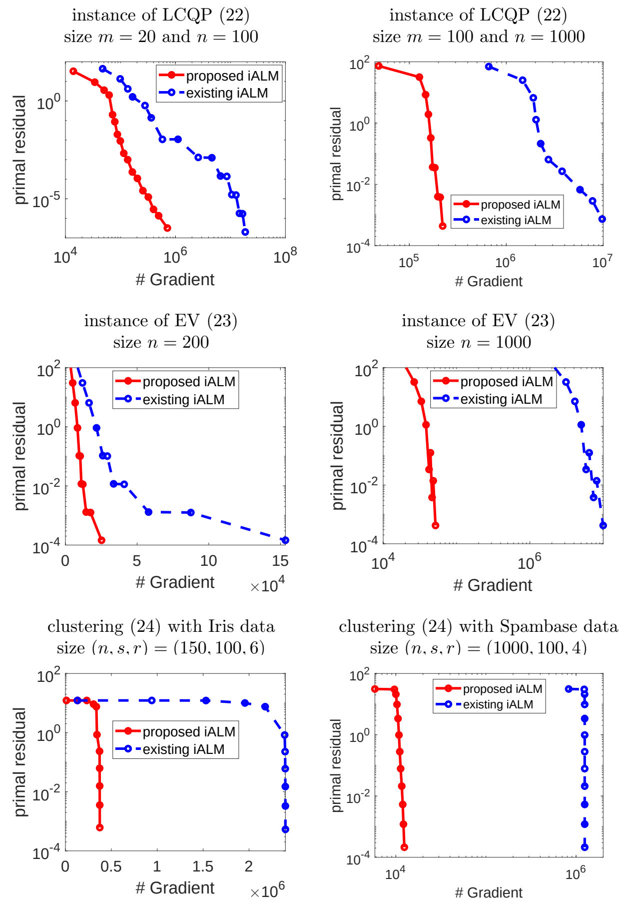

Figure 1: Comparison of the proposed iALM and the existing iALM in (Sahin et al., 2019) on solving the LCQP (first row), EV (middle row), and clustering (last row) problems. Each column shows different problem configurations and dimensions as indicated in the plot title. Each plot shows the primal residual. The markers denote the outer iterations in iALM. Dual residuals for both methods are similar, below a given tolerance ε.

In the middle row of Figure 1, we compare the primal residual trajectories of our method and the iALM by Sahin et al. (2019) on one instance of (23). From the results, we conclude that, to reach an ε-KKT point to the EV problem, the proposed improved iALM takes significantly fewer gradient evaluations (thus shorter time) than the iALM by Sahin et al. (2019), for both small-sized and large-sized instances.

4.3 Clustering problem

In this subsection, we consider the clustering problem proposed in (Sahin et al., 2019) and compare our method to the iALM by Sahin et al. (2019). The clustering problem is formulated as

$$\begin{array}{l} \min_{\mathbf{X}\in C} \sum_{i,j=1}^{n} D_{i,j}\langle \mathbf{x}_i, \mathbf{x}_j\rangle, \\ \text{s.t. } \mathbf{x}_i^\top \sum_{j=1}^{n} \mathbf{x}_j - 1 = 0, \forall\, i = 1,\ldots,n, \end{array} \tag{24}$$

where $\mathbf{X} = [\mathbf{x}_1^\top, \cdots, \mathbf{x}_n^\top]^\top$, C is the intersection of the positive orthant with the Euclidean ball of radius s, and $\mathbf{D}$ is the distance matrix generated by a set of data points $\{\mathbf{z}_i\}_{i=1}^n$ such that $D_{i,j} = \|\mathbf{z}_i - \mathbf{z}_j\|$.

We run two instances, on Iris data set (Dua and Graff, 2017) with $n = 150, s = 100, r = 6$ and Spambase data set (Dua and Graff, 2017) with $n = 1000, s = 100, r = 4$. We tuned the smoothness constant of iALM by Sahin et al. (2019) to $L_k = 80\|\mathbf{D}\| + 1200\beta_k$ in order to have convergence. The weak convexity constant was tuned to $\rho_k = -0.2r \cdot \min(\text{eig}(\mathbf{D})) \cdot \beta_k$.

In the bottom row of Figure 1, we compare the primal residual trajectories of our method and the iALM by Sahin et al. (2019). From the results, we conclude that, to reach an ε-KKT point to the clustering problem, the proposed improved iALM needs significantly fewer gradient evaluations (thus shorter time) than the iALM by Sahin et al. (2019). The advantage of our method is even more significant for the larger-sized instance.

5 CONCLUSION

We have presented an improved iALM for solving nonconvex constrained optimization. Different from existing iALMs, our iALM uses the iPPM to approximately solve each subproblem. Under the same regularity condition as existing works, we explore the better convergence rate of iPPM and the boundedness of AL functions to establish improved complexity results. To reach an ε-KKT solution, our method requires $\tilde{O}(\varepsilon^{-\frac{5}{2}})$ proximal gradient steps for solving nonconvex optimization with affine-equality constraints. The result is slightly worsened to $\tilde{O}(\varepsilon^{-3})$ if the constraints are also nonconvex. Both complexity results are so far the best. Numerically, we demonstrated that the proposed improved iALM could significantly outperform one existing iALM and also one penalty-based FOM.

Acknowledgements

*The authors Pin-Yu Chen, Sijia Liu, Songtao Lu, and Yangyang Xu are listed in alphabetical order.

This work was supported by the Rensselaer-IBM AI Research Collaboration (`http://airc.rpi.edu`), part of the IBM AI Horizons Network (`http://ibm.biz/AIHorizons`).

References

Necdet Serhat Aybat and Garud Iyengar. An augmented lagrangian method for conic convex programming. *arXiv preprint arXiv:1302.6322*, 2013.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1): 183–202, 2009.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2): 1751–1772, 2018.

Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. *SIAM Journal on Optimization*, 21(4):1721–1739, 2011.

Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. On the complexity of finding first-order critical points in constrained nonlinear optimization. *Mathematical Programming*, 144(1-2):93–106, 2014.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2): 59–99, 2016.

Erfan Yazdandoost Hamedani and Necdet Serhat Aybat. A primal-dual algorithm for general convex-concave saddle point problems. *arXiv preprint arXiv:1803.01401*, 2018.

Magnus R Hestenes. Multiplier and gradient methods. *Journal of optimization theory and applications*, 4 (5):303–320, 1969.

Le Thi Khanh Hien, Renbo Zhao, and William B Haskell. An inexact primal-dual smoothing framework for large-scale non-bilinear saddle point problems. *arXiv preprint arXiv:1711.03669*, 2017.

Bo Jiang, Tianyi Lin, Shiqian Ma, and Shuzhong Zhang. Structured nonconvex and nonsmooth optimization: algorithms and iteration complexity analysis. *Computational Optimization and Applications*, 72(1):115–157, 2019.

Naoki Katoh and Toshihide Ibaraki. Resource allocation problems. In *Handbook of combinatorial optimization*, pages 905–1006. Springer, 1998.

Weiwei Kong, Jefferson G Melo, and Renato DC Monteiro. Complexity of a quadratic penalty accelerated inexact proximal point method for solving linearly constrained nonconvex composite programs. *SIAM Journal on Optimization*, 29(4):2566–2593, 2019.

Guanghui Lan and Renato D.C. Monteiro. Iteration-complexity of first-order augmented lagrangian methods for convex programming. *Mathematical Programming*, 155(1-2):511–547, 2016.

Fei Li and Zheng Qu. An inexact proximal augmented lagrangian framework with arbitrary linearly convergent inner solver for composite convex optimization. *arXiv preprint arXiv:1909.09582*, 2019.

Zichong Li and Yangyang Xu. Augmented lagrangian based first-order methods for convex and nonconvex programs: nonergodic convergence and iteration complexity. *arXiv preprint arXiv:2003.08880*, 2020.

Qihang Lin, Selvaprabu Nadarajah, and Negar Soheili. A level-set method for convex optimization with a feasible solution path. *SIAM Journal on Optimization*, 28(4):3290–3311, 2018.

Qihang Lin, Runchao Ma, and Yangyang Xu. Inexact proximal-point penalty methods for constrained non-convex optimization. *arXiv preprint arXiv:1908.11518*, 2019.

Tianyi Lin, Chi Jin, Michael Jordan, et al. Near-optimal algorithms for minimax optimization. *arXiv preprint arXiv:2002.02417*, 2020.

S. Lu, R. Singh, X. Chen, Y. Chen, and M. Hong. Alternating gradient descent ascent for nonconvex min-max problems in robust learning and GANs. In *Proc. of Asilomar Conference on Signals, Systems, and Computers*, pages 680–684, 2019.

Songtao Lu, Meisam Razaviyayn, Bo Yang, Kejun Huang, and Mingyi Hong. Finding second-order stationary points efficiently in smooth nonconvex linearly constrained optimization problems. In *Advances in Neural Information Processing Systems*, 2020a.

Songtao Lu, Ioannis Tsaknakis, Mingyi Hong, and Yongxin Chen. Hybrid block successive approximation for one-sided non-convex min-max problems: Algorithms and applications. *IEEE Transactions on Signal Processing*, 68:3676–3691, 2020b.

Zhaosong Lu and Zirui Zhou. Iteration-complexity of first-order augmented lagrangian methods for convex conic programming. *arXiv preprint arXiv:1803.09941*, 2018.

Jefferson G Melo, Renato DC Monteiro, and Hairong Wang. Iteration-complexity of an inexact proximal accelerated augmented lagrangian method for solving linearly constrained smooth nonconvex composite optimization problems. *Optimization Online*, 2020.

Ion Necoara and Valentin Nedelcu. Rate analysis of inexact dual first-order methods application to dual decomposition. *IEEE Transactions on Automatic Control*, 59(5):1232–1243, 2014.

Valentin Nedelcu, Ion Necoara, and Quoc Tran-Dinh. Computational complexity of inexact gradient augmented lagrangian methods: application to constrained mpc. *SIAM Journal on Control and Optimization*, 52(5):3109–3134, 2014.

Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140 (1):125–161, 2013.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Publisher, 2004.

Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706): 289–337, 1933.

Yuyuan Ouyang, Yunmei Chen, Guanghui Lan, and Eduardo Pasiliao Jr. An accelerated linearized alternating direction method of multipliers. *SIAM Journal on Imaging Sciences*, 8(1):644–681, 2015.

Michael JD Powell. *A method for non-linear constraints in minimization problems.* in Optimization, R. Fletcher Ed., Academic Press, New York, NY, 1969.

Mehmet Fatih Sahin, Ahmet Alacaoglu, Fabian Latorre, Volkan Cevher, et al. An inexact augmented lagrangian framework for nonconvex optimization with nonlinear constraints. In *Advances in Neural Information Processing Systems*, pages 13943–13955, 2019.

Po-Wei Wang and Chih-Jen Lin. Iteration complexity of feasible descent methods for convex optimization. *The Journal of Machine Learning Research*, 15(1): 1523–1548, 2014.

Yangyang Xu. First-order methods for constrained convex programming based on linearized augmented lagrangian function. *arXiv preprint arXiv:1711.08020*, 2017.

Yangyang Xu. Iteration complexity of inexact augmented lagrangian methods for constrained convex programming. *Mathematical Programming, Series A (online first)*, pages 1–46, 2019.

Yangyang Xu. Primal-dual stochastic gradient method for convex programs with many functional constraints. *SIAM Journal on Optimization*, 30(2):1664–1692, 2020.

Hao Yu and Michael J Neely. A primal-dual type algorithm with the $O(1/t)$ convergence rate for large scale constrained convex programs. In *Decision and Control (CDC), 2016 IEEE 55th Conference on*, pages 1900–1905. IEEE, 2016.

Hao Yu and Michael J Neely. A simple parallel algorithm with an $O(1/t)$ convergence rate for general convex programs. *SIAM Journal on Optimization*, 27(2):759–783, 2017.

Jiawei Zhang and Zhiquan Luo. A global dual error bound and its application to the analysis of linearly constrained nonconvex optimization. *arXiv preprint arXiv:2006.16440*, 2020.