Aug-PDG: 带不等式约束凸优化算法的线性收敛性

孟　敏, 李修贤†

(同济大学 电子与信息工程学院 控制科学与工程系, 上海 201804;

同济大学 上海自主智能无人系统科学中心, 上海 201210; 同济大学 上海智能科学与技术中心, 上海 200092)

摘要: 原始–对偶梯度算法广泛应用于求解带约束的凸优化问题, 大部分文献仅证明了该算法的收敛性, 而没有分析其收敛速度. 因此, 本文研究了求解带有不等式约束凸优化的一类离散算法, 即增广原始–对偶梯度算法(Aug-PDG), 证明了Aug-PDG算法在一些较弱的假设条件下可以半全局线性收敛到最优解, 并明确给出了算法中步长的上界. 最后, 数值算例证实了所得理论结果的有效性.

关键词: 凸优化; 非线性约束; 线性收敛; 增广原始–对偶梯度算法

引用格式: 孟敏, 李修贤. Aug-PDG: 带不等式约束凸优化算法的线性收敛性. 控制理论与应用, 2022, 39(10): 1969 – 1977

DOI: 10.7641/CTA.2021.10583

Aug-PDG: linear convergence of convex optimization with inequality constraints

MENG Min, LI Xiu-xian†

(Department of Control Science and Engineering, School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China;
Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University, Shanghai 201210, China;
Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai 200092, China)

Abstract: The primal-dual gradient algorithm has been widely employed for solving constrained optimization problems. While the convergence of this algorithm was proved in most references, it is less investigated whether it is globally linearly convergent. Therefore, this paper studies convergence rate of its variant, i.e., the augmented primal-dual gradient algorithm (Aug-PDG), for handling the convex optimization problem with general convex inequality constraints. Specifically, it is shown that the Aug-PDG can converge semi-globally to the optimizer at a linear rate under some mild assumptions and an explicit bound is provided for the stepsize in this algorithm. Finally, a numerical example is presented to illustrate the effectiveness of the theoretical result.

Key words: convex optimization; nonlinear constraints; linear convergence; augmented primal-dual gradient algorithm

Citation: MENG Min, LI Xiuxian. Aug-PDG: linear convergence of convex optimization with inequality constraints. *Control Theory & Applications*, 2022, 39(10): 1969 – 1977

1 Introduction

This paper deals with the constrained optimization problem formulated as follows:

$$\begin{aligned}&\min_{x\in\mathbb{R}^n}\ f(x),\\&\text{s.t. } g(x)\leqslant 0,\end{aligned}\tag{1}$$

where the objective function $f:\mathbb{R}^n\rightarrow\mathbb{R}$ and $g(x)=(g_1(x)\ g_2(x)\ \cdots\ g_m(x))^{\mathrm{T}}$ with $g_i:\mathbb{R}^n\rightarrow\mathbb{R}$ being convex and continuously differentiable. By resorting to the (or augmented) Lagrangian $L(x,\lambda)$ of problem (1), the corresponding (or augmented) primal-dual gradient algorithm (PDG) (or Aug-PDG) can be designed as

$$\begin{cases}x_{k+1}=x_k-\alpha\nabla_x L(x_k,\lambda_k),\\ \lambda_{k+1}=[\lambda_k+\alpha\nabla_\lambda L(x_k,\lambda_k)]_+,\end{cases}\tag{2}$$

where $\alpha>0$ is a stepsize and $[\cdot]_+$ denotes the projection operator onto the nonnegative orthants component-

Received 2 July 2021; accepted 30 December 2021.
†Corresponding author. E-mail: xli@tongji.edu.cn.
Recommended by Associate Editor: XU Jin-ming.
Supported by the Shanghai Pujiang Program (21PJ1413100), the National Natural Science Foundation of China (62003243, 62103305), the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0100), the Shanghai Municipal Commission of Science and Technology (1951-1132101), the Young Elite Scientist Sponsorship Program by cast of China Association for Science and Technology (YESS20200136) and the Fundamental Research Funds for the Central Universities (22120210096).

wisely. It is known that Eq. (2) can find a saddle point of $L(x,\lambda)$, and thus it has been extensively studied to solve the constrained optimization problem [1].

Optimization has wide applications in the artificial intelligence field such as smart grids [2–3], wireless communication [4], robot systems [5], game theory [6–7], to name just a few. To date, there is a large body of literature on theoretical analysis of asymptotic convergence of various algorithms, including primal-dual gradient-based algorithms, for tackling the optimization problem under different settings [8–18].

In recent decades, researchers have focused on the linear convergence and exponential convergence of primal-dual gradient-based algorithms in discrete-time and continuous-time, respectively. It is well-known that when the objective function is strongly convex and smooth, the gradient decent algorithm for unconstrained convex optimization can achieve global exponential convergence in continuous-time and global linear convergence in discrete-time. In the context of constrained optimization with equality constraints $Ax=b$ or affine inequality constraints $Ax \leqslant b$, PDG is proved to converge globally exponentially in continuous-time setup [19]. A proximal gradient flow was proposed in [20], which can be applied to resolve convex optimization problems with affine inequality constraints and has global exponential convergence when A has full row rank. Local exponential convergence of the primal-dual gradient dynamics can be established with the help of spectral bounds of saddle matrices [21]. Recently, the authors in [22] proved that the Aug-PDGD in continuous-time for optimization with affine equality and inequality constraints achieves global exponential convergence, and the global linear converge of primal-dual gradient optimization (PDGO) in discrete-time was discussed in [23] by contraction theory. It should be noted that the aforementioned works focus on unconstrained optimization or constrained optimization with affine equality and/or affine inequality constraints. For the case with nonlinear inequality constraints, the asymptotic convergence has been extensively studied such as in [24]. However, the linear/exponential convergence for the optimization with nonlinear inequality constraints is seldom investigated in the literature. One exception is the recent work [25], where the authors established a semi-global exponential convergence of continuous-time Aug-PDGD in the sense that the convergence rate depends on the distance from the initial point to the optimal point.

However, [25] concentrates on the continuous-time dynamics. As discrete-time algorithms are easily implemented in practical applications, in this paper, the discrete-time algorithm is addressed for the optimization problem with nonlinear inequality constraints. Theoretical analysis based on a quadratic Lyapunov function that has non-zero off-diagonal terms is first presented to show that the Aug-PDG achieves semi-global linear convergence.

The rest of this paper is organized as follows. Section 2 introduces preliminaries on optimization with nonlinear equality constraints. The main result on the semi-global linear convergence of Aug-PDGA, along with its proof, is presented in Section 3. Section 4 provides a numerical example to illustrate the feasibility of the obtained result. Section 5 makes a brief conclusion.

Notations. Let $\mathbb{R}^m$, $\mathbb{R}^m_+$ and $\mathbb{R}^{m\times n}$ be the sets of m-dimensional real column vectors, m-dimensional nonnegative column vectors and $m \times n$ real matrices, respectively. Define $[x]_+$ to be the component-wise projection of a vector $x \in \mathbb{R}^m$ onto $\mathbb{R}^m_+$. $x \geqslant 0$ for any vector $x \in \mathbb{R}^m$ means that each entry of x is nonnegative. For an integer $n > 0$, denote $[n] := \{1, 2, \cdots, n\}$. I_n is the identity matrix of dimension n. $\mathbf{1}_n$ (resp. $\mathbf{0}_n$) represents an n-dimensional vector with all of its elements being 1 (resp. 0). For a vector or matrix A, A^{T} denotes the transpose of A and $A_{\mathcal{I}}$ is a matrix composed of the rows of A with the indices in $\mathcal{I}$. For real symmetric matrices P and Q, $P \succ (\succeq, \succ, \preceq)\ Q$ means that $P-Q$ is positive (positive semi-, negative, negative semi-) definite, while for two vectors/matrices w, v of the same dimension, $w \leqslant v$ means that each entry of $w-v$ is nonnegative. $\mathrm{diag}\{a_1, a_2, \cdots, a_n\}$ represents a diagonal matrix with a_i, $i \in [n]$, on its diagonal.

2 Preliminaries

Consider problem (1). An augmented Lagrangian associated with problem (1) is introduced as [26]

$$L(x,\lambda) := f(x) + U(x,\lambda), \tag{3}$$

where $x \in \mathbb{R}^n$, $\lambda = (\lambda_1\ \lambda_2\ \cdots\ \lambda_m)^{\mathrm{T}} \in \mathbb{R}^m$, $\rho > 0$ is the penalty parameter, and

$$U(x,\lambda) := \sum_{i=1}^{m} \frac{[\rho g_i(x) + \lambda_i]_+^2 - \lambda_i^2}{2\rho}. \tag{4}$$

It can be verified that $U(x,\lambda)$ is convex in x and concave in λ, and $U(x,\lambda)$ is continuously differentiable, i.e.,

$$\nabla_x U(x,\lambda) = \sum_{i=1}^{m} [\rho g_i(x) + \lambda_i]_+ \nabla g_i(x), \tag{5}$$

$$\nabla_\lambda U(x,\lambda) = \sum_{i=1}^{m} \frac{[\rho g_i(x) + \lambda_i]_+ - \lambda_i}{\rho} e_i, \tag{6}$$

where e_i is an n-dimensional vector with the ith entry being 1 and others 0. Then the Aug-PDG is explicitly written as

$$x_{k+1} = x_k - \alpha \nabla_x L(x_k,\lambda_k) =$$
$$x_k - \alpha \nabla f(x_k) - \alpha \sum_{i=1}^{m} [\rho g_i(x_k) + \lambda_{i,k}]_+ \nabla g_i(x_k), \tag{7a}$$

$$\lambda_{k+1} = \lambda_k + \alpha \nabla_\lambda L(x_k,\lambda_k) =$$
$$\lambda_k + \alpha \sum_{i=1}^{m} \frac{[\rho g_i(x_k) + \lambda_{i,k}]_+ - \lambda_{i,k}}{\rho} e_i, \tag{7b}$$

where $\alpha \in (0, \rho]$ is the stepsize to be specified. Here, the initial conditions are arbitrarily chosen as $x_0 \in \mathbb{R}^n$ and $\lambda_0 \geqslant 0$.

To proceed, the following results are vital for solving the constrained optimization problem.

Lemma 1 For Aug-PDG (7), if $\lambda_0 \geqslant 0$, then $\lambda_k \geqslant 0, \forall k \geqslant 0$.

Proof This result can be proved by mathematical induction, which is omitted here.

Lemma 2 A primal-dual pair (x^*, λ^*) is an equilibrium point of the Aug-PDG (7) if and only if (x^*, λ^*) is a Karush-Kuhn-Tucker (KKT) point of (1).

Proof If a primal-dual pair (x^*, λ^*) is an equilibrium point of the Aug-PDG (7), that is, $x^* = x^* - \alpha\nabla_x L(x^*, \lambda^*)$ and $\lambda^* = \lambda^* + \alpha\nabla_\lambda L(x^*, \lambda^*)$, then $\nabla_x L(x^*, \lambda^*) = 0$ and $\nabla_\lambda L(x^*, \lambda^*) = 0$. $\nabla_\lambda L(x^*, \lambda^*) = 0$ is equivalent to

$$\lambda_i^* = [\rho g_i(x^*) + \lambda_i^*]_+, \text{ for any } i \in [m], \quad (8)$$

which implies $\lambda_i^* \geqslant 0$, $g_i(x^*) \leqslant 0$, and $\lambda_i^* g_i(x^*) = 0$.

For $\nabla_x L(x^*, \lambda^*) = 0$, one equivalently obtains that

$$\nabla f(x^*) + \sum_{i=1}^{m} [\rho g_i(x^*) + \lambda_i^*]_+ \nabla g_i(x^*) = \nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*) = 0.$$

Thus, it can be claimed that the primal-dual pair (x^*, λ^*) is a KKT point.

Conversely, if (x^*, λ^*) is a KKT point of (1), then

$$\begin{aligned} &\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*) = 0, \\ &\lambda_i^* g_i(x^*) = 0, \\ &\lambda^* \geqslant 0, \\ &g_i(x^*) \leqslant 0. \end{aligned}$$

Via a simple computation, $\nabla_x L(x^*, \lambda^*) = 0$ and $\nabla_\lambda L(x^*, \lambda^*) = 0$, which implies that (x^*, λ^*) is an equilibrium point of the Aug-PDG (7). □

3 Main results

In this section, the main result on the linear convergence of the Aug-PDG is presented.

3.1 Convergence results

The following assumptions are essential for deriving the main result.

Assumption 1 The problem (1) has a unique feasible solution x^*, and at x^*, the linear independence constraint qualification (LICQ) holds at x^*, i.e., $\{\nabla g_i(x^*) | i \in \mathcal{I}\}$ is linearly independent, where $\mathcal{I} := \{i \in [m] | g_i(x^*) = 0\}$ is the so-called active set at x^*.

Under Assumption 1, the optimal Lagrangian multiplier λ^* is also unique [27]. Denote by J the Jacobian of $g(x)$ at x^* and $J_{\mathcal{I}}$ the matrix composed of the rows of J with the indices in $\mathcal{I}$. LICQ in Assumption 1 also implies that $J_{\mathcal{I}} J_{\mathcal{I}}^{\mathrm{T}} \succ 0$ [25]. Define

$$\kappa := \lambda_{\min}(J_{\mathcal{I}} J_{\mathcal{I}}^{\mathrm{T}}) > 0 \quad (9)$$

to be the smallest eigenvalue of $J_{\mathcal{I}} J_{\mathcal{I}}^{\mathrm{T}}$.

Assumption 2 The objective function $f(x)$ has a quadratic gradient growth with parameter $\mu > 0$ over $\mathbb{R}^n$, i.e., for any $x \in \mathbb{R}^n$,

$$(\nabla f(x) - \nabla f(x^*))^{\mathrm{T}}(x - x^*) \geqslant \mu \|x - x^*\|^2. \quad (10)$$

The concept of quadratic gradient growth was introduced in [28], which is a relaxation of strong convexity condition for guaranteeing linear convergence of gradient-based optimization algorithms. In fact, the class of functions having quadratic gradient growth include the strongly convex functions as a proper subset and some functions with quadratic gradient growth are even not convex.

Assumption 3 The objective function f is l-smooth over $\mathbb{R}^n$, i.e.,

$$\|\nabla f(x) - \nabla f(y)\| \leqslant l\|x - y\|, \ \forall x, y \in \mathbb{R}^n.$$

For any $i \in [m]$, $g_i(x)$ is L_{gi}-smooth and has bounded gradient, i.e., for some $L_{gi}, B_{gi} > 0$ and any $x, y \in \mathbb{R}^n$, there holds

$$\begin{aligned} &\|\nabla g_i(x) - \nabla g_i(y)\| \leqslant L_{gi}\|x - y\|, \\ &\|\nabla g_i(x)\| \leqslant B_{gi}. \end{aligned}$$

Denote

$$\mathcal{I}^c := [m] \backslash \mathcal{I}, \ L_g := \sqrt{\sum_{i=1}^{m} L_{gi}^2}, \ B_g := \sqrt{\sum_{i=1}^{m} B_{gi}^2}.$$

Under Assumption 3, one can obtain that

$$\|J\| \leqslant B_g, \quad (11)$$

$$\|g(x) - g(y)\| = \sqrt{\sum_{i=1}^{m} (g_i(x) - g_i(y))^2} \leqslant \sqrt{\sum_{i=1}^{m} B_{gi}^2 \|x - y\|^2} = B_g\|x - y\|. \quad (12)$$

Denote

$$d_0 := \sqrt{\|x_0 - x^*\|^2 + \|\lambda_0 - \lambda^*\|^2}.$$

Before giving the main result of this paper, it is convenient to list the following concept similar to that in continuous-time setting [29].

Definition 1 Consider the dynamics $z(t+1) = \phi(z(t))$ with initial point $z(0) = z_0$. Assume that z_e is an equilibrium point satisfying $z_e = \phi(z_e)$. z_e is said to be a semi-global linear stable point if for any $h > 0$, there exist $c > 0$ and $0 < \gamma < 1$ such that for any z_0 satisfying $\|z_0 - z_e\| \leqslant h$, $\|z(t) - z_e\| \leqslant c\gamma^t \|z_0 - z_e\|$, $\forall t \geqslant 0$. z_e is said to be a global linear stable point if c and γ do not depend on h.

Then the main result is presented as follows.

Theorem 1 Under Assumptions 1–3, if the stepsize $0<\alpha<1$ is chosen such that

$$\alpha<\min\{\rho,\frac{2\mu}{b_1+2a_4\delta},\frac{\kappa\delta}{2b_2+4a_5\delta},\frac{1-\pi^*}{2\rho(b_2+2a_5\delta)}\},\tag{13}$$

where $\delta>0$ satisfies

$$\delta<\min\{\frac{\mu}{2a_3},\frac{1-\pi^*}{2\rho(\kappa+8B_g^2+L_g^2(1-\pi^*))},B_g^{-1}\},\tag{14}$$

$\pi^*:=[\rho\max\limits_{i\in\mathcal{I}^c}\{g_i(x^*)\}/(\sqrt{3}d_0)+1]_+^2$, $b_1:=a_1+2B_g^2$, $b_2:=a_2+\dfrac{2}{\rho^2}$, $a_1:=2l+4\theta_1^2$, $a_2:=4B_g^2$, $a_3:=2B_g^2l^2/\kappa+2B_g^2\theta_1^2/\kappa+2B_g^2/(\kappa\rho^2)+\kappa B_g^2\rho^2/4$, $a_4:=B_g^2l^2/2+B_g^2\theta_1^2+2B_g^2$, $a_5:=B_g^4+2/\rho^2$, and $\theta_1:=\rho B_g^2+L_g\|\lambda^*\|$, then the sequences $\{x_k\}$ and $\{\lambda_k\}$ generated by Aug-PDG (7) for the constrained optimization (1) semi-globally converge to the optimal point of the optimization problem (1) at a linear (or exponential) rate. Specifically, it holds that

$$\|x_k-x^*\|^2+\|\lambda_k-\lambda^*\|^2\leqslant 3(1-\gamma)^kd_0^2,\tag{15}$$

where $0<\gamma<1$ satisfies

$$\gamma\leqslant\min\{c_1,c_2,c_3\}\tag{16}$$

with $c_1:=\mu\alpha-a_3\delta\alpha-b_1\alpha^2/2-a_4\delta\alpha^2$, $c_2:=\kappa\delta\alpha/4-b_2\alpha^2/2-a_5\delta\alpha^2$, and $c_3:=\dfrac{\alpha}{2\rho}(1-\pi^*)-(\delta\alpha\kappa+b_2\alpha^2+2a_5\delta\alpha^2)/2-4\alpha\delta B_g^2$.

Proof The proof is postponed to the next subsection. □

Remark 1 The selection of parameters α and δ ensures that c_1,c_2,c_3 are positive, and then $\gamma>0$ can be guaranteed. From Theorem 1, one can see that the convergence rate is related to π^*, where the distance d_0 between the initial point and the optimal one is involved. Therefore, the convergence rate decreases to 0 as d_0 goes to infinity. The rate also changes as (x_k,λ_k) approaches the optimal point. In fact, any (x_k,λ_k) can be regarded as an initial point of the studied algorithm in the sense of running the algorithm from the point (x_k,λ_k). Then, when (x_k,λ_k) approaches the optimal point, π^* is smaller, leading to smaller $1-\gamma$. As a result, the rate is slow at the beginning and then becomes fast when x_k goes to the optimal point. Consequently, Theorem 1 does not guarantee the existence of a global linear convergence, and only semi-global linear convergence can be ensured.

Remark 2 Compared with the most related literature [25], where a continuous-time algorithm, called Aug-PDGD, was studied with a semi-global exponential convergence, a discrete-time algorithm Aug-PDG is analyzed here with a semi-global linear convergence. Although discrete-time algorithms may be obtained by discretizing the continuous-time Aug-PDGD using such as explicit Euler method, it is unclear how to select the sampling stepsize to guarantee the convergence especially in the sense of semi-global convergence. In comparison, an explicit bound on the stepsize α is established here. However, one drawback is that the upper bound of α depends on the bounds of the cost functions, constraint functions and their gradients, as well as the optimal solution. This may be tackled by adaptive methods, which will be our future research of interest.

3.2 Proof of Theorem 1

To prove Theorem 1, intermediate results are first presented as follows.

Lemma 3[22] For any $y,y^*\in\mathbb{R}$, there exists $\xi\in[0,1]$ such that $[y]_+-[y^*]_+=\xi(y-y^*)$. Specifically, ξ can be chosen as

$$\xi=\begin{cases}\dfrac{[y]_+-[y^*]_+}{y-y^*}, & \text{if } y\neq y^*,\\ 0, & \text{if } y=y^*.\end{cases}$$

Lemma 4 Under Assumption 1–3, x_k generated by the Aug-PDG (7) satisfies

$$\begin{aligned}&\|x_{k+1}-x^*\|^2\leqslant\\&(1-2\mu\alpha+a_1\alpha^2)\|x_k-x^*\|^2+a_2\alpha^2\|\lambda_k-\lambda^*\|^2+\\&2\alpha(U(x^*,\lambda_k)-U(x_k,\lambda_k))+\\&2\alpha(U(x_k,\lambda^*)-U(x^*,\lambda^*)).\end{aligned}\tag{17}$$

Proof By iterations in (7), one has that

$$\begin{aligned}&\|x_{k+1}-x^*\|^2=\\&\|x_k-\alpha\nabla_xL(x_k,\lambda_k)-x^*+\alpha\nabla_xL(x^*,\lambda^*)\|^2=\\&\|x_k-x^*\|^2+\alpha^2\|\nabla_xL(x_k,\lambda_k)-\nabla_xL(x^*,\lambda^*)\|^2-\\&2\alpha(\nabla_xL(x_k,\lambda_k)-\nabla_xL(x^*,\lambda^*))^{\mathrm{T}}(x_k-x^*).\end{aligned}\tag{18}$$

Based on

$$\begin{aligned}&\nabla_xL(x_k,\lambda_k)=\nabla f(x_k)+\nabla_xU(x_k,\lambda_k)=\\&\nabla f(x_k)+\sum_{i=1}^m[\rho g_i(x_k)+\lambda_{i,k}]_+\nabla g_i(x_k),\end{aligned}$$

for the second term on the right side of (18), one has

$$\begin{aligned}&\alpha^2\|\nabla_xL(x_k,\lambda_k)-\nabla_xL(x^*,\lambda^*)\|^2\leqslant\\&2\alpha^2\|\nabla f(x_k)-\nabla f(x^*)\|^2+\\&2\alpha^2\|\nabla_xU(x_k,\lambda_k)-\nabla_xU(x^*,\lambda^*)\|^2.\end{aligned}\tag{19}$$

Note that

$$\begin{aligned}&\nabla_xU(x_k,\lambda_k)-\nabla_xU(x^*,\lambda^*)=\\&\sum_{i=1}^m[\rho g_i(x_k)+\lambda_{i,k}]_+\nabla g_i(x_k)-\\&\sum_{i=1}^m[\rho g_i(x^*)+\lambda_i^*]_+\nabla g_i(x^*)=\\&\sum_{i=1}^m([\rho g_i(x_k)+\lambda_{i,k}]_+-[\rho g_i(x^*)+\lambda_i^*]_+)\nabla g_i(x_k)+\\&\sum_{i=1}^m[\rho g_i(x^*)+\lambda_i^*]_+(\nabla g_i(x_k)-\nabla g_i(x^*)),\end{aligned}$$

then

$$\begin{aligned}&\|\nabla_xU(x_k,\lambda_k)-\nabla_xU(x^*,\lambda^*)\|\leqslant\\&\sum_{i=1}^m|[\rho g_i(x_k)+\lambda_{i,k}]_+-\end{aligned}$$

$$[\rho g_i(x^*)+\lambda_i^*]_+|\cdot\|\nabla g_i(x_k)\|+$$
$$\sum_{i=1}^m[\rho g_i(x^*)+\lambda_i^*]_+\|\nabla g_i(x_k)-\nabla g_i(x^*)\|. \quad (20)$$

Define

$$\xi_{i,k}=\frac{[\rho g_i(x_k)+\lambda_{i,k}]_+-[\rho g_i(x^*)+\lambda_i^*]_+}{(\rho g_i(x_k)+\lambda_{i,k})-(\rho g_i(x^*)+\lambda_i^*)}, \quad (21)$$

if $(\rho g_i(x_k)+\lambda_{i,k})-(\rho g_i(x^*)+\lambda_i^*)\neq 0$, and

$$\xi_{i,k}=0, \quad (22)$$

if $(\rho g_i(x_k)+\lambda_{i,k})-(\rho g_i(x^*)+\lambda^*)=0$. Then $\xi_{i,j}\in[0,1]$. It can be obtained from Lemma 3 that

$$[\rho g_i(x_k)+\lambda_{i,k}]_+-[\rho g_i(x^*)+\lambda_i^*]_+=$$
$$\rho\xi_{i,k}(g_i(x_k)-g_i(x^*))+\xi_{i,k}(\lambda_{i,k}-\lambda_i^*). \quad (23)$$

Substituting Eq. (23) into Eq. (20) yields that

$$\|\nabla_x U(x_k,\lambda_k)-\nabla_x U(x^*,\lambda^*)\|\leqslant$$
$$\sum_{i=1}^m|\rho\xi_{i,k}(g_i(x_k)-g_i(x^*))+$$
$$\xi_{i,k}(\lambda_{i,k}-\lambda_i^*)|\|\nabla g_i(x_k)\|+$$
$$\sum_{i=1}^m[\rho g_i(x^*)+\lambda_i^*]_+\|\nabla g_i(x_k)-\nabla g_i(x^*)\|\leqslant$$
$$\sum_{i=1}^m(\rho B_{gi}\|x_k-x^*\|+|\lambda_{i,k}-\lambda_i^*|)B_{gi}+$$
$$\sum_{i=1}^m\lambda_i^*L_{gi}\|x_k-x^*\|\leqslant$$
$$\theta_1\|x_k-x^*\|+B_g\|\lambda_k-\lambda^*\|, \quad (24)$$

where Assumption 3 has been applied to get the second inequality, and the third inequality is derived by $\sum_{i=1}^n a_ib_i\leqslant\sqrt{\sum_{i=1}^n a_i^2}\sqrt{\sum_{i=1}^n b_i^2}$ for any $a_i,b_i\in\mathbb{R}$, $\theta_1=\rho B_g^2+L_g\|\lambda^*\|$ and $B_g=\sqrt{\sum_{i=1}^m B_{gi}^2}$, $L_g=\sqrt{\sum_{i=1}^m L_{gi}^2}$.

By Eqs. (19)–(24)and Assumption 3, one has that

$$\|\nabla_x L(x_k,\lambda_k)-\nabla_x L(x^*,\lambda^*)\|^2=$$
$$\|\nabla f(x_k)-\nabla f(x^*)+\nabla_x U(x_k,\lambda_k)-$$
$$\nabla_x U(x^*,\lambda^*)\|^2\leqslant$$
$$2l^2\|x_k-x^*\|^2+2(\theta_1\|x_k-x^*\|+$$
$$B_g\|\lambda_k-\lambda^*\|)^2\leqslant$$
$$2l^2\|x_k-x^*\|^2+4(\theta_1^2\|x_k-x^*\|^2+$$
$$B_g^2\|\lambda_k-\lambda^*\|^2)=$$
$$a_1\|x_k-x^*\|^2+a_2\|\lambda_k-\lambda^*\|^2, \quad (25)$$

where $a_1=2l^2+4\theta_1^2$ and $a_2=4B_g^2$.

For the third term on the right side of Eq. (18),

$$-2\alpha(\nabla_x L(x_k,\lambda_k)-\nabla_x L(x^*,\lambda^*))^{\mathrm T}(x_k-x^*)=$$
$$-2\alpha(\nabla f(x_k)-\nabla f(x^*))^{\mathrm T}(x_k-x^*)-$$
$$2\alpha(\nabla_x U(x_k,\lambda_k)-\nabla_x U(x^*,\lambda^*))^{\mathrm T}(x_k-x^*)\leqslant$$
$$-2\mu\alpha\|x_k-x^*\|^2+2\alpha(U(x^*,\lambda_k)-U(x_k,\lambda_k))+$$
$$2\alpha(U(x_k,\lambda^*)-U(x^*,\lambda^*)), \quad (26)$$

where the inequality is derived based on Assumption 2 and the convexity of $U(x,\lambda)$ at x, i.e.,

$$U(x,\lambda)-U(x',\lambda)\geqslant(\nabla_x U(x',\lambda))^{\mathrm T}(x-x'), \quad (27)$$

for any $x,x'\in\mathbb{R}^n$. Plugging Eq. (25) and Eq. (26) into Eq. (18), one concludes that Eq. (17) holds. □

Lemma 5 Under Assumptions 1–3, λ_k generated by the Aug-PDG (7) satisfies

$$\|\lambda_{k+1}-\lambda^*\|^2\leqslant(1+\frac{2\alpha^2}{\rho^2})\|\lambda_k-\lambda^*\|^2+$$
$$2B_g^2\alpha^2\|x_k-x^*\|^2+$$
$$2\alpha(U(x_k,\lambda_k)-U(x_k,\lambda^*)). \quad (28)$$

Proof For $\|\lambda_{k+1}-\lambda^*\|^2$, by iteration (7b), one has

$$\|\lambda_{k+1}-\lambda^*\|^2=\|\lambda_k+\alpha\nabla_\lambda U(x_k,\lambda_k)-\lambda^*\|^2=$$
$$\|\lambda_k-\lambda^*\|^2+\alpha^2\|\nabla_\lambda U(x_k,\lambda_k)\|^2+$$
$$2\alpha\nabla_\lambda U(x_k,\lambda_k))^{\mathrm T}(\lambda_k-\lambda^*). \quad (29)$$

Recalling $\nabla_\lambda L(x^*,\lambda^*)=\nabla_\lambda U(x^*,\lambda^*)=0$ and the notation of $\xi_{i,k}$ in Eqs. (21)–(22), it can be obtained that

$$\|\nabla_\lambda U(x_k,\lambda_k)\|^2=$$
$$\|\nabla_\lambda U(x_k,\lambda_k)-\nabla_\lambda U(x^*,\lambda^*)\|^2=$$
$$\|\sum_{i=1}^m\frac{[\rho g_i(x_k)+\lambda_{i,k}]_+-\lambda_{i,k}-[\rho g_i(x^*)+\lambda_i^*]_++\lambda_i^*}{\rho}e_i\|^2=$$
$$\|\sum_{i=1}^m[\xi_{i,k}(g_i(x_k)-g_i(x^*))+$$
$$\frac{1}{\rho}(\xi_{i,k}-1)(\lambda_{i,k}-\lambda_i^*)]e_i\|^2=$$
$$\|\varXi_k(g(x_k)-g(x^*))+\frac{1}{\rho}(\varXi_k-I_m)(\lambda_k-\lambda^*)\|^2\leqslant$$
$$2B_g^2\|x_k-x^*\|^2+\frac{2}{\rho^2}\|\lambda_k-\lambda^*\|^2, \quad (30)$$

where $\varXi_k=\mathrm{diag}\{\xi_{1,k},\cdots,\xi_{m,k}\}$, the inequality is obtained based on Eq. (12) and $\|\varXi_k\|\leqslant 1$, $\|\varXi_k-I_m\|\leqslant 1$ for $\xi_{i,k}\in[0,1]$, $i\in[m]$.

As $U(x,\lambda)$ is concave at λ, one has

$$(\nabla_\lambda U(x_k,\lambda_k))^{\mathrm T}(\lambda_k-\lambda^*)\leqslant$$
$$U(x_k,\lambda_k)-U(x_k,\lambda^*). \quad (31)$$

By Eqs. (30)–(31), it can be derived from Eq. (29) that Eq. (28) holds. □

In what follows, we prove Theorem 1 in detail.

Proof of Theorem 1 Define

$$V_{\delta,k}=\begin{bmatrix}x_k-x^*\\ \lambda_k-\lambda^*\end{bmatrix}^{\mathrm T}Q_\delta\begin{bmatrix}x_k-x^*\\ \lambda_k-\lambda^*\end{bmatrix}, \quad (32)$$

where

$$Q_\delta=\begin{bmatrix}I_n & \delta J^{\mathrm T}\\ \delta J & I_m\end{bmatrix}. \quad (33)$$

As $\delta<\frac{1}{\sqrt2}B_g^{-1}$ from (14), one has $\frac{1}{2}I_{n+m}\preceq Q_\delta$

$\preceq \dfrac{3}{2}I_{n+m}$ by Schur complement. Then, in the following, we first discuss the bound of

$$V_{\delta,k+1} = \|x_{k+1} - x^*\|^2 + \|\lambda_{k+1} - \lambda^*\|^2 + 2\delta(x_{k+1} - x^*)^{\mathrm{T}} J^{\mathrm{T}}(\lambda_{k+1} - \lambda^*).$$

The bounds of $\|x_{k+1} - x^*\|^2$ and $\lambda_{k+1} - \lambda^*$ are given in Lemmas 4 and 5, respectively. It suffices to compute the bound of $2\delta(x_{k+1} - x^*)^{\mathrm{T}} J^{\mathrm{T}}(\lambda_{k+1} - \lambda^*)$.

Note that (x^*, λ^*) is the KKT point of (1), that is,

$$\nabla_x L(x^*, \lambda^*) = 0, \tag{34}$$

$$\nabla_\lambda L(x^*, \lambda^*) = 0. \tag{35}$$

It is easy to obtain that

$$\begin{aligned}&(x_{k+1} - x^*)^{\mathrm{T}} J^{\mathrm{T}}(\lambda_{k+1} - \lambda^*) = \\ &(x_k - x^*)^{\mathrm{T}} J^{\mathrm{T}}(\lambda_k - \lambda^*) - \\ &\alpha(\nabla f(x_k) - \nabla f(x^*))^{\mathrm{T}} J^{\mathrm{T}}(\lambda_k - \lambda^*) - \\ &\alpha(\nabla_x U(x^*, \lambda_k) - \nabla_x U(x^*, \lambda^*))^{\mathrm{T}} J^{\mathrm{T}}(\lambda_k - \lambda^*) - \\ &\alpha(\nabla_x U(x_k, \lambda_k) - \nabla_x U(x^*, \lambda_k))^{\mathrm{T}} J^{\mathrm{T}}(\lambda_k - \lambda^*) + \\ &\alpha(x_{k+1} - x^*)^{\mathrm{T}} J^{\mathrm{T}}(\nabla_\lambda U(x_k, \lambda_k) - \nabla_\lambda U(x^*, \lambda^*)).\end{aligned} \tag{36}$$

By Eq. (24), one has that

$$\|\nabla_x U(x^*, \lambda_k) - \nabla_x U(x^*, \lambda^*)\| \leqslant B_g \|\lambda_k - \lambda^*\|, \tag{37}$$

$$\|\nabla_x U(x_k, \lambda_k) - \nabla_x U(x^*, \lambda_k) \leqslant \theta_1 \|x_k - x^*\|. \tag{38}$$

Similar to Eqs. (21)–(22), define

$$\xi_{i,\lambda} = \begin{cases} \dfrac{[\rho g_i(x^*) + \lambda_{i,k}]_+ - [\rho g_i(x^*) + \lambda_i^*]_+}{\lambda_{i,k} - \lambda^*}, \\ \quad \text{if } \lambda_{i,k} \neq \lambda_i^*, \\ 0,\ \text{if } \lambda_{i,k} = \lambda_i^*, \end{cases} \tag{39}$$

then

$$\begin{aligned}&\nabla_x U(x^*, \lambda_k) - \nabla_x U(x^*, \lambda^*) = \\ &\sum_{i=1}^m ([\rho g_i(x^*) + \lambda_{i,k}]_+ - [\rho g_i(x^*) + \lambda_i^*]_+)\nabla g_i(x^*) = \\ &\sum_{i=1}^m \xi_{i,\lambda}(\lambda_{i,k} - \lambda_i^*)\nabla g_i(x^*) = \\ &J^{\mathrm{T}} \varXi_\lambda (\lambda_k - \lambda^*),\end{aligned} \tag{40}$$

where $\varXi_\lambda := \mathrm{diag}\{\xi_{1,\lambda}, \cdots, \xi_{m,\lambda}\}$. For the last term of Eq. (36), it holds that

$$\begin{aligned}&\alpha(x_{k+1} - x^*)^{\mathrm{T}} J^{\mathrm{T}}(\nabla_\lambda U(x_k, \lambda_k) - \nabla_\lambda U(x^*, \lambda^*)) = \\ &\alpha(x_k - x^*)^{\mathrm{T}} J^{\mathrm{T}}(\nabla_\lambda U(x_k, \lambda_k) - \nabla_\lambda U(x^*, \lambda^*)) - \\ &\alpha^2(\nabla f(x_k) - \nabla f(x^*)) \times \\ &J^{\mathrm{T}}(\nabla_\lambda U(x_k, \lambda_k) - \nabla_\lambda U(x^*, \lambda^*)) - \\ &\alpha^2(\nabla_x U(x_k, \lambda_k) - \nabla_x U(x^*, \lambda^*))^{\mathrm{T}} \times \\ &J^{\mathrm{T}}(\nabla_\lambda U(x_k, \lambda_k) - \nabla_\lambda U(x^*, \lambda^*)) \leqslant \\ &\frac{2B_g^2\alpha}{\kappa\rho^2}\|x_k - x^*\|^2 + \\ &\frac{\kappa\rho^2\alpha}{8}\|\nabla_\lambda U(x_k, \lambda_k) - \nabla_\lambda U(x^*, \lambda^*)\|^2 + \\ &\frac{B_g^2\alpha^2}{2}\|\nabla f(x_k) - \nabla f(x^*)\|^2 + \\ &\frac{B_g^2\alpha^2}{2}\|\nabla_x U(x_k, \lambda_k) - \nabla_x U(x^*, \lambda^*)\|^2 + \\ &\alpha^2\|\nabla_\lambda U(x_k, \lambda_k) - \nabla_\lambda U(x^*, \lambda^*)\|^2.\end{aligned} \tag{41}$$

Therefore, by Eqs. (30)and (37)–(41), Eq. (36) is rewritten as

$$\begin{aligned}&(x_{k+1} - x^*)^{\mathrm{T}} J^{\mathrm{T}}(\lambda_{k+1} - \lambda^*) = \\ &(x_k - x^*)^{\mathrm{T}} J^{\mathrm{T}}(\lambda_k - \lambda^*) - \\ &\alpha(\lambda_k - \lambda^*)^{\mathrm{T}} \varXi_\lambda J J^{\mathrm{T}}(\lambda_k - \lambda^*) - \\ &\alpha(\nabla f(x_k) - \nabla f(x^*))^{\mathrm{T}} J^{\mathrm{T}}(\lambda_k - \lambda^*) - \\ &\alpha(\nabla_x U(x_k, \lambda_k) - \nabla_x U(x^*, \lambda_k))^{\mathrm{T}} J^{\mathrm{T}}(\lambda_k - \lambda^*) + \\ &\alpha(x_{k+1} - x^*)^{\mathrm{T}} J^{\mathrm{T}}(\nabla_\lambda U(x_k, \lambda_k) - \nabla_\lambda U(x^*, \lambda^*)) \leqslant \\ &(x_k - x^*)^{\mathrm{T}} J^{\mathrm{T}}(\lambda_k - \lambda^*) - \\ &\alpha(\lambda_k - \lambda^*)^{\mathrm{T}} \varXi_\lambda J J^{\mathrm{T}}(\lambda_k - \lambda^*) + \\ &\frac{2B_g^2\alpha}{\kappa}\|\nabla f(x_k) - \nabla f(x^*)\|^2 + \frac{\kappa\alpha}{8}\|\lambda_k - \lambda^*\|^2 + \\ &\frac{2B_g^2\alpha}{\kappa}\|\nabla_x U(x_k, \lambda_k) - \nabla_x U(x^*, \lambda_k)\|^2 + \\ &\frac{\kappa\alpha}{8}\|\lambda_k - \lambda^*\|^2 + \frac{2B_g^2\alpha}{\kappa\rho^2}\|x_k - x^*\|^2 + \\ &\frac{\kappa\rho^2\alpha}{8}\|\nabla_\lambda U(x_k, \lambda_k) - \nabla_\lambda U(x^*, \lambda^*)\|^2 + \\ &\frac{B_g^2\alpha^2}{2}\|\nabla f(x_k) - \nabla f(x^*)\|^2 + \\ &\frac{B_g^2\alpha^2}{2}\|\nabla_x U(x_k, \lambda_k) - \nabla_x U(x^*, \lambda^*)\|^2 + \\ &\alpha^2\|\nabla_\lambda U(x_k, \lambda_k) - \nabla_\lambda U(x^*, \lambda^*)\|^2 \leqslant \\ &(x_k - x^*)^{\mathrm{T}} J^{\mathrm{T}}(\lambda_k - \lambda^*) - \\ &\alpha(\lambda_k - \lambda^*)^{\mathrm{T}} \varXi_\lambda J J^{\mathrm{T}}(\lambda_k - \lambda^*) + \\ &(a_3\alpha + a_4\alpha^2)\|x_k - x^*\|^2 + \\ &(\frac{\kappa\alpha}{2} + a_5\alpha^2)\|\lambda_k - \lambda^*\|^2,\end{aligned} \tag{42}$$

where the last inequality is obtained by a simple computation, along with Eqs. (24), (30) and (38).

Define

$$\tilde{\xi}_{i,k} := \begin{cases} [\rho g_i(x^*)/\lambda_{i,k} + 1]_+^2, & i \in \mathcal{I}^c,\ \lambda_{i,k} \neq 0, \\ [\rho g_i(x^*)/d_0 + 1]_+^2, & i \in \mathcal{I}^c,\ \lambda_{i,k} = 0, \\ 1, & i \in \mathcal{I}, \end{cases} \tag{43}$$

where $\tilde{\varXi}_k := \mathrm{diag}\{\tilde{\xi}_{1,\mathrm{k}}, \cdots, \tilde{\xi}_{\mathrm{m,k}}\}$. Then

$$U(x^*, \lambda_k) - U(x^*, \lambda^*) = \frac{1}{2\rho}(\lambda_k - \lambda^*)^{\mathrm{T}}(\tilde{\varXi}_k - I_m)(\lambda_k - \lambda^*). \tag{44}$$

Combining with Eqs. (17) (28) (42) and (44), the bound of $V_{\delta,k+1}$ can be derived that

$$V_{\delta,k+1} =$$

$$\|x_{k+1}-x^*\|^2+\|\lambda_{k+1}-\lambda^*\|^2+$$
$$2\delta(x_{k+1}-x^*)^{\mathrm{T}}J^{\mathrm{T}}(\lambda_{k+1}-\lambda^*)\leqslant$$
$$(1-\gamma)V_{\delta,k}+\begin{bmatrix}x_k-x^*\\ \lambda_k-\lambda^*\end{bmatrix}^{\mathrm{T}}Q\begin{bmatrix}x_k-x^*\\ \lambda_k-\lambda^*\end{bmatrix}, \tag{45}$$

where

$$Q=\begin{bmatrix}Q_1 & \gamma\delta J^{\mathrm{T}}\\ \gamma\delta J & Q_2\end{bmatrix}, \tag{46}$$

with

$$Q_1:=(\gamma-2\mu\alpha+2a_3\delta\alpha+b_1\alpha^2+2a_4\delta\alpha^2)I_n,$$
$$Q_2:=(\gamma+\delta\alpha\kappa+b_2\alpha^2+2a_5\delta\alpha^2)I_m+$$
$$\frac{\alpha}{\rho}(\tilde{\Xi}_k-I_m)-\alpha\delta(\Xi_\lambda JJ^{\mathrm{T}}+JJ^{\mathrm{T}}\Xi_\lambda),$$
$$b_1=a_1+2B_g^2, \text{ and } b_2=a_2+\frac{2}{\rho^2}.$$

If $Q\preceq 0$, then $V_{\delta,k+1}\leqslant(1-\gamma)V_{\delta,k}$, indicating that $V_{\delta,k}\leqslant(1-\gamma)^kV_{\delta,0}$ and $\lambda_{\min}(Q_\delta)(\|x_k-x^*\|^2+\|\lambda_k-\lambda^*\|^2)\leqslant\lambda_{\max}(Q_\delta)(1-\gamma)^k(\|x_0-x^*\|^2+\|\lambda_0-\lambda^*\|^2)$. Therefore,

$$\|x_k-x^*\|^2+\|\lambda_k-\lambda^*\|^2\leqslant$$
$$\frac{\lambda_{\max}(Q_\delta)}{\lambda_{\min}(Q_\delta)}(1-\gamma)^k(\|x_0-x^*\|^2+\|\lambda_0-\lambda^*\|^2). \tag{47}$$

Thus, Eq. (15) holds as $\dfrac{\lambda_{\max}(Q_\delta)}{\lambda_{\min}(Q_\delta)}\leqslant\dfrac{3/2}{1/2}=3$.

Note that

$$\begin{bmatrix}-\gamma I_n & \gamma\delta J^{\mathrm{T}}\\ \gamma\delta J & -\gamma I_m\end{bmatrix}\preceq 0.$$

Hence, to prove $Q\preceq 0$, it suffices to ensure

$$\begin{bmatrix}Q_1+\gamma I_n & \mathbf{0}_{n\times m}\\ \mathbf{0}_{m\times n} & Q_2+\gamma I_m\end{bmatrix}\preceq 0. \tag{48}$$

By Eq. (16), one can obtain that

$$2\gamma-2\mu\alpha+2a_3\delta\alpha+b_1\alpha^2+2a_4\delta\alpha^2\leqslant 0,$$

i.e., $Q_1+\gamma I_m\preceq 0$.

Next, consider $\Theta:=\dfrac{\alpha}{\rho}(\tilde{\Xi}_k-I_m)-\alpha\delta(\Xi_\lambda JJ^{\mathrm{T}}+JJ^{\mathrm{T}}\Xi_\lambda)$ in Q_2. If $\Theta+(2\gamma+\delta\alpha\kappa+b_2\alpha^2+2a_5\delta\alpha^2)I_m\preceq 0$, then $Q_2+\gamma I_m\preceq 0$.

Note that $\xi_{i,\lambda}=\tilde{\xi}_{i,k}=1$ when $i\in\mathcal{I}$. Partition Θ as

$$\Theta=\begin{bmatrix}\Theta_1 & \Theta_3\\ \Theta_3^{\mathrm{T}} & \Theta_2\end{bmatrix}, \tag{49}$$

where

$$\Theta_1:=-2\delta\alpha J_{\mathcal{I}}J_{\mathcal{I}}^{\mathrm{T}},$$
$$\Theta_2:=\frac{\alpha}{\rho}(\tilde{\Xi}_{k,\mathcal{I}^c}-I)-$$
$$\alpha\delta(\Xi_{\lambda,\mathcal{I}^c}J_{\mathcal{I}^c}J_{\mathcal{I}^c}^{\mathrm{T}}+J_{\mathcal{I}^c}J_{\mathcal{I}^c}^{\mathrm{T}}\Xi_{\lambda,\mathcal{I}^c}),$$
$$\Theta_3:=-\delta\alpha J_{\mathcal{I}}J_{\mathcal{I}^c}^{\mathrm{T}}(I+\Xi_{\lambda,\mathcal{I}^c}).$$

Under LICQ in Assumption 1, it can be obtained that $J_IJ_I^{\mathrm{T}}\succeq\kappa I$. Then one can see from Eq. (16) that

$$\Theta_1+(2\gamma+\delta\alpha\kappa+b_2\alpha^2+2a_5\delta\alpha^2)I\preceq-\frac{1}{2}\delta\alpha J_{\mathcal{I}}J_{\mathcal{I}}^{\mathrm{T}}.$$

By Lemma 6 in [22],

$$L_g^2(I-\Xi_{\lambda,\mathcal{I}^c})+\Xi_{\lambda,\mathcal{I}^c}J_{\mathcal{I}^c}J_{\mathcal{I}^c}^{\mathrm{T}}+J_{\mathcal{I}^c}J_{\mathcal{I}^c}^{\mathrm{T}}\Xi_{\lambda,\mathcal{I}^c}\succeq 0,$$

then since $\tilde{\xi}_{i,k}\leqslant\xi_{i,\lambda}$ for $i\in\mathcal{I}^c$, one has

$$\Theta_2\preceq\frac{\alpha}{\rho}(\tilde{\Xi}_{k,\mathcal{I}^c}-I)-\delta\alpha L_g^2(\tilde{\Xi}_{k,\mathcal{I}^c}-I). \tag{50}$$

Denote $\phi=2\gamma+\delta\alpha\kappa+b_2\alpha^2+2a_5\delta\alpha^2$, then one can obtain that

$$\Theta_2+\phi I-\Theta_3^{\mathrm{T}}(\Theta_1+\phi I_m)^{-1}\Theta_3\preceq$$
$$\frac{\alpha}{\rho}(1-\delta\rho L_g^2)(\tilde{\Xi}_{\mathcal{I}^c}-I)+\phi I+$$
$$2\alpha\delta(I+\Xi_{\lambda,\mathcal{I}^c})J_{\mathcal{I}^c}J_{\mathcal{I}}^{\mathrm{T}}(J_{\mathcal{I}}J_{\mathcal{I}})^{-1}J_{\mathcal{I}}J_{\mathcal{I}^c}^{\mathrm{T}}(I+\Xi_{\lambda,\mathcal{I}^c})\preceq$$
$$\frac{\alpha}{\rho}(1-\delta\rho L_g^2)(\tilde{\Xi}_{\mathcal{I}^c}-I)+$$
$$\phi I+2\alpha\delta(I+\Xi_{\lambda,\mathcal{I}^c})J_{\mathcal{I}^c}J_{\mathcal{I}^c}^{\mathrm{T}}(I+\Xi_{\lambda,\mathcal{I}^c})\preceq$$
$$\frac{\alpha}{\rho}(1-\delta\rho L_g^2)(\tilde{\Xi}_{\mathcal{I}^c}-I)+\phi I+8\alpha\delta B_g^2I, \tag{51}$$

where $A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}A\preceq I$ for a full row rank matrix A has been applied in the second inequality.

If for all $i\in\mathcal{I}^c$ and $k\geqslant 0$,

$$\tilde{\xi}_{i,k}\leqslant\pi^*=[\rho\max_{i\in\mathcal{I}^c}\{g_i(x^*)\}/(\sqrt{3}d_0)+1]_+^2, \tag{52}$$

then one can obtain that the sum on the right hand of Eq. (51) is less than or equal to 0 by Eqs. (14) and (16), and then by Schur complement, $Q_2+\gamma I_m\preceq 0$. Thus Theorem 1 is proved.

What we left is to show Eq. (52). Based on Eq. (8) and $g_i(x^*)<0$ for $i\in\mathcal{I}^c$, one has that $\lambda_i^*=0$, $i\in\mathcal{I}^c$. For $k=0$, obviously, for all $i\in\mathcal{I}^c$, $\tilde{\xi}_{i,0}\leqslant\pi^*$, which indicates that Eq. (52) holds for $k=0$, and then $V_{\delta,1}\leqslant V_{\delta,0}$. Therefore, invoking Eq. (47) yields that for all $i\in\mathcal{I}^c$,

$$\tilde{\xi}_{i,1}\leqslant[\rho\max_{i\in\mathcal{I}^c}\{g_i(x^*)\}/(\sqrt{3}d_0)+1]_+^2=\pi^*.$$

Subsequently, by the mathematical induction, Eq. (52) can be proved. The proof is completed. □

4 Example

In this section, an example motivated by applications in power systems [25] is presented to illustrate the feasibility of the discrete-time Aug-PGD (7). Consider the following constrained optimization problem:

$$\begin{aligned}\min_{p_i,q_i\in\mathbb{R}}\ & f(x)=\sum_{i=1}^{n}((p_i-p_{v,i})^2+q_i^2),\\ \text{s.t.}\ & g_i(x)=p_1^2+q_i^2-S_i\leqslant 0,\\ & 0\leqslant p_i\leqslant p_{v,i},\ i\in[n],\end{aligned} \tag{53}$$

where $x=(p_1\ \cdots\ p_n\ q_1\ \cdots\ q_n)^{\mathrm{T}}$ and $p_{v,i},S_i,i\in[n]$ are constants. The problem Eq. (53) along with an affine inequality constraint was considered in [25] but via a continuous-time dynamics Aug-PDGD. The affine inequality constraints can be regarded as special nonlinear constraints. Hence the algorithm Aug-PDG in this paper is applicable to the optimization problem Eq. (53).

Let $n = 10$, $S = (S_1, \cdots, S_n) = (2.7, 1.35, 2.7, 1.35, 2.025, 2.025, 2.7, 2.7, 1.35, 2.025)$ and $p_v = (p_{v,1}, \cdots, p_{v,n}) = 4S$. Choose $\rho = 0.1$. Three cases are simulated, where the initial point (x_0, λ_0) is selected randomly such that the distance from the initial point (x_0, λ_0) to the optimal point (x^*, λ^*) (i.e., d_0) is $0.1\|(x^*, \lambda^*)\|$, $5\|(x^*, \lambda^*)\|$ and $10\|(x^*, \lambda^*)\|$, respectively. The curves of the normalized distance $\dfrac{\|(x_k - x^*, \lambda_k - \lambda^*)\|}{\|(x^*, \lambda^*)\|}$ with respect to the iteration k are shown in Fig. 1 when choosing $\alpha = 0.1$ and in Fig. 2 when choosing $\alpha = 0.05$, where for each case, 10 instances of randomly selected initial points are considered. From Fig. 1, it can be seen that the convergence rates are different for different d_0, and the distance $\|(x_k - x^*, \lambda_k - \lambda^*)\|$ linearly decays on the whole. From Figs. 1 and 2, when the algorithm is convergent by choosing appropriate stepsize α, it can be seen that the convergence rate is smaller if α is smaller. Moreover, for each case, the decreasing rate also changes as (x_k, λ_k) approaches the optimal point. Specifically, the decreasing rates are small at the beginning and then become large when (x_k, λ_k) goes to the optimal point. These observations support the semi-global linear convergence of the Aug-PDG, which is consistent with our theory analysis.

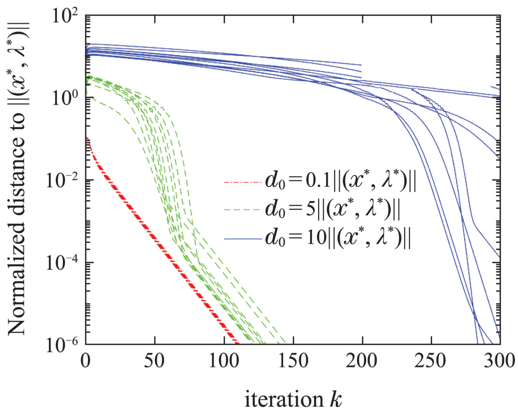

Fig. 1 Simulation of the relative distances to $\|(x^*, \lambda^*)\|$ with respect to iteration k by choosing $\alpha = 0.1$

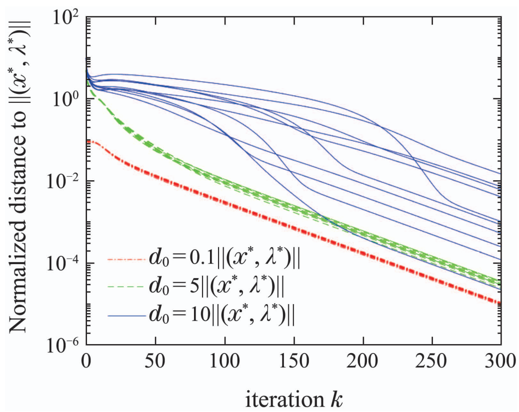

Fig. 2 Simulation of the relative distances to $\|(x^*, \lambda^*)\|$ with respect to iteration k by choosing $\alpha = 0.05$

5 Conclusion

In this paper, the linear convergence of an Aug-PDG in discrete-time for convex optimization with nonlinear inequality constraints has been investigated. Under some mild assumptions, the Aug-PDG has been proved to semi-globally converge at a linear rate, which depends on the distance from the initial point to the optimal point. Future research of interest may be to adopted adaptive methods to determine the upper bound of stepsizes.

References:

[1] RUSZCZYNSKI A. *Nonlinear Optimization*. Princeton: Princeton University Press, 2011.

[2] CHIANG M, LOW S H, CALDERBANK A R, et al. Layering as optimization decomposition: A mathematical theory of network architectures. *Proceedings of the IEEE*, 2007, 95(1): 255 – 312.

[3] ZAHO C, TOPCU U, LI N, et al. Design and stability of load-side primary frequency control in power systems. *IEEE Transactions on Automatic Control*, 2014, 59(5): 1177 – 1189.

[4] CHEN J, LAU V K. Convergence analysis of saddle point problems in time varying wireless systems-control theoretical approach. *IEEE Transactions on Signal Processing*, 2012, 60(1): 443 – 452.

[5] FANG H, SHANG C, CEHN J. An optimization-based shared control framework with applications in multi-robot systems. *Science China Information Sciences*, 2018, 61(1): 1 – 3.

[6] GHARESIFARD B, CORTES J. Distributed convergence to Nash equilibria in two-network zero-sum games. *Automatica*, 2013, 49(6): 1683 – 1692.

[7] PAVEL L. Distributed GNE seeking under partial-decision information over networks via a doubly-augmented operator splitting approach. *IEEE Transactions on Automatic Control*, 2020, 65(4): 1584 – 1597.

[8] POLYAK B. Iterative methods using Lagrange multipliers for solving extremal problems with constraints of the equation type. *USSR Computational Mathematics and Mathematical Physics*, 1970, 10(5): 42 – 52.

[9] GOLSHTEIN E. Generalized gradient method for finding saddle points. *Matekon*, 1974, 10(3): 36 – 52.

[10] NEDIĆ A, OZDAGLAR A. Subgradient methods for saddle point problems. *Journal of Optimization Theory and Applications*, 2009, 142(1): 205 – 228.

[11] FEIJER D, PAGANINI F. Stability of primal-dual gradient dynamics and applications to network optimization. *Automatica*, 2010, 46(12): 1974 – 1981.

[12] CHERUKURI A, MALLADA E, CORTÉS J. Asymptotic convergence of constrained primal-dual dynamics. *Systems & Control Letters*, 2016, 87: 10 – 15.

[13] HONG M, HAJINEZHAD D, ZHAO M. Prox-PDA: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks. *International Conference on Machine Learning*. Sydney, Australia: PMLR, 2017: 1529 – 1538.

[14] LEI Q, YEN I E H, WU C, et al. Doubly greedy primal-dual coordinate descent for sparse empirical risk minimization. *International Conference on Machine Learning*. Sydney, Australia: PMLR, 2017: 2034 – 2042.

[15] LI X, XIE L, HONG Y. Distributed continuous-time nonsmooth convex optimization with coupled inequality constraints. *IEEE Transactions on Control of Network Systems*, 2020, 7(1): 74 – 84.

[16] LI X, FENG G, XIE L. Distributed proximal algorithms for multi-agent optimization with coupled inequality constraints. *IEEE Transactions on Automatic Control*, 2021, 66(33): 1223 – 1230.

[17] XIE Pei, YOU Keyou, HONG Yiguang, et al. A survey of distributed convex optimization algorithms over networks. *Control Theory & Applications*, 2018, 35(7): 918 – 927.
(谢佩, 游科友, 洪奕光, 等. 网络化分布式凸优化算法研究进展. 控制理论与应用, 2018, 35(7): 918 – 927.)

[18] YANG Tao, XU Lei, YI Xinlei, et al. Event-triggered distributed optimization algorithms. *Acta Automatica Sinica*, 2022, 48(1): 133 – 143.
(杨涛, 徐磊, 易新蕾, 等. 基于事件触发的分布式优化算法. 自动化学报, 2022, 48(1): 133 – 143.)

[19] NIEDERÄNDER S K, ALLGÖWER F, CORTÉS J. Exponentially fast distributed coordination for nonsmooth convex optimization. *IEEE the 55th Conference on Decision and Control (CDC)*. Las Vegas, NV, USA: IEEE, 2016: 1036 – 1041.

[20] DHINGRA N K, KHONG S Z, JOVANOVIĆ M R. The proximal augmented Lagrangian method for nonsmooth composite optimization. *IEEE Transactions on Automatic Control*, 2019, 64(7): 2861 – 2868.

[21] SHEN S Q, HUANG T Z, YU J. Eigenvalue estimates for preconditioned nonsymmetric saddle point matrices. *SIAM Journal on Matrix Analysis and Applications*, 2010, 31(5): 2453 – 2476.

[22] QU G, LI N. On the exponential stability of primal-dual gradient dynamics. *IEEE Control Systems Letters*, 2019, 3(1): 43 – 48.

[23] SU Y, SHI Y, SUN C. Contraction analysis on primal-dual gradient optimization. *arXiv preprint*, arXiv:1907.10171.

[24] NEDIĆ A. Random algorithms for convex minimization problems. *Mathematical Programming*, 2011, 129(2): 225 – 253.

[25] TANG Y, QU G, LI N. Semi-global exponential stability of augmented primal-dual gradient dynamics for constrained convex optimization. *Systems & Control Letters*, 2020, 144: 104754.

[26] BERTSEKAS D P. *Constrained Optimization and Lagrange Multiplier Methods*. New York: Academic Press, 2014.

[27] WACHSMUTH G. On LICQ and the uniqueness of Lagrange multipliers. *Operations Research Letters*, 2013, 41(1): 78 – 80.

[28] NECOARA I, NESTEROV Y, GLINEUR F. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 2019, 175(1/2): 69 – 107.

[29] SASTRY S. *Nonlinear Systems: Analysis, Stability, and Control*. New York: Springer Science & Business Media, 2013.

作者简介:

孟 敏 研究员, 博士生导师, 研究方向为分布式博弈与优化、分布式估计与安全控制、布尔网络等, E-mail: mengmin@tongji.edu.cn;

李修贤 研究员, 博士生导师, 研究方向为分布式控制、(在线)优化、算子理论、及其在无人机和无人驾驶中的应用, E-mail: xli@tongji.edu.cn.