2018 IEEE Conference on Decision and Control (CDC)
Miami Beach, FL, USA, Dec. 17-19, 2018

An exponentially convergent primal-dual algorithm for nonsmooth composite minimization

Dongsheng Ding, Bin Hu, Neil K. Dhingra, and Mihailo R. Jovanović

Abstract— **We consider a class of nonsmooth convex composite optimization problems, where the objective function is given by the sum of a continuously differentiable convex term and a potentially non-differentiable convex regularizer. In [1], the authors introduced the proximal augmented Lagrangian method and derived the resulting continuous-time primal-dual dynamics that converge to the optimal solution. In this paper, we extend these dynamics from continuous to discrete time via the forward Euler discretization. We prove explicit bounds on the exponential convergence rates of our proposed algorithm with a sufficiently small step size. Since a larger step size can improve the convergence speed, we further develop a linear matrix inequality (LMI) condition which can be numerically solved to provide rate certificates with general step size choices. In addition, we prove that a large range of step size values can guarantee exponential convergence. We close the paper by demonstrating the performance of the proposed algorithm via computational experiments.**

I. Introduction

We consider a class of nonsmooth convex composite optimization problems, where the objective function is the sum of a continuously differentiable convex term and a potentially non-differentiable convex regularizer. This class of problems arises in statistics, machine learning, control, image and signal processing. Two typical examples are the empirical risk minimization problem [2] and the structural optimal control problem [3], [4]. The indicator function, the ℓ_1 norm, and the nuclear norm are commonly used as nonsmooth convex regularizers that enforce constraints, promote sparsity, and induce low-rank structure on optimal solutions.

A common approach for nonsmooth convex composite optimization problems is to use an auxiliary variable to reformulate them as linearly constrained problems that separate the smooth term and the nonsmooth regularizer in the objective function [5]. This facilitates the use of primal-dual methods based on the augmented Lagrangian [6], including the method of multipliers (MM) [6], and the alternating direction method of multipliers (ADMM) [5]. The efficiency of these methods depends on how the nonsmooth primal subproblems are solved. Direct approaches of subgradients [7], or proximal operators [8] have been exploited to derive a large family of customized primal-dual algorithms including the primal-dual subgradient algorithm [9], proximal ADMM [10], and the proximal primal-dual algorithm [11]. Unfortunately, the implementation of such algorithms encounters technical issues such as step size selection [7] and parameter sensitivity [12].

To avoid nonsmooth primal subproblems in MM, a proximal augmented Lagrangian method has been recently developed in [1]. This method exploits the proximal operator associated with the nonsmooth regularizer in the objective function to restrict the augmented Lagrangian to the manifold that corresponds to the explicit minimization over one of primal variables in the nonsmooth term. This constrained augmented Lagrangian is called the proximal augmented Lagrangian and it leads to new efficient algorithms which complement MM and ADMM in solving nonsmooth composite optimization problems.

In this paper, we derive a primal-dual (PD) algorithm from the forward Euler discretization of the continuous-time PD dynamics of the proximal augmented Lagrangian. We demonstrate that the proposed algorithm with a well-chosen step size converges at an exponential rate under standard assumptions. Our main contributions are:

1) We prove explicit bounds on the exponential convergence rates of our proposed algorithm with a sufficiently small step size.
2) Since a large step size can lead to potential improvements in the algorithm performance, we formulate a linear matrix inequality (LMI) condition which can be numerically solved to provide rate certificates for our algorithm with general step size choices.
3) We further prove that a large range of step size values can guarantee the exponential convergence. While our theory for large step sizes proves the exponential convergence, our analysis for small step sizes provides a convergence rate estimate.
4) Finally, we demonstrate performance of our algorithm in solving quadratic optimization problems and show that the large step size choice is preferable for fast convergence.

We note that the convergence rate analysis for the continuous-time PD dynamics in [1] cannot be easily tailored for our discrete-time algorithm. This is consistent with the observations made in [13]–[15] about the difficulty of translating rate bounds from continuous time to discrete time.

Our presentation is organized as follows. In Section II, we formulate the problem, review the proximal augmented Lagrangian and the related continuous-time PD dynamics.

Financial support from the National Science Foundation under award ECCS-1809833 and from the Air Force Office of Scientific Research under award FA9550-16-1-0009 is gratefully acknowledged.

D. Ding and M. R. Jovanović are with the Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089. B. Hu is with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801. N. K. Dhingra is with Numerica Corporation, Fort Collins, CO 80528. E-mails: dongshed@usc.edu, binhu7@illinois.edu, dhin0008@umn.edu, mihailo@usc.edu

In Section III, we present a discretized PD algorithm. We provide various theoretical/numerical tools that can be used to certify the exponential convergence of the proposed algorithm. In Section IV, we provide numerical examples to illustrate the performance of our algorithm.

II. Problem formulation and background

Consider the nonsmooth convex composite optimization problem

$$\underset{x}{\text{minimize}} \quad f(x) + g(Tx) \tag{1}$$

where $x \in \mathbb{R}^n$ is the optimization variable and $T \in \mathbb{R}^{m\times n}$ is a given matrix. We assume that (1) is feasible, that its minimum is finite, and that the matrix T has full row rank.

Assumption 1: The function f is strongly convex with parameter m_f and its gradient is Lipschitz continuous with parameter L_f. The function g is proper, lower semicontinuous, and convex but potentially non-differentiable.

When the matrix T is not diagonal, the sub-gradient and the proximal gradient methods cannot be used to solve (1) directly. A customary approach is to introduce an additional optimization variable $z \in \mathbb{R}^m$,

$$\begin{array}{ll} \underset{x,\, z}{\text{minimize}} & f(x) + g(z) \\ \text{subject to} & Tx - z = 0. \end{array} \tag{2}$$

A. Proximal augmented Lagrangian

The proximal augmented Lagrangian is defined as [1],

$$\mathcal{L}_\mu(x;y) = f(x) + M_{\mu g}(Tx + \mu y) - \tfrac{\mu}{2}\|y\|^2 \tag{3}$$

where x is the primal variable, y is the dual variable, and $\mu > 0$ is the augmented Lagrangian parameter. The proximal augmented Lagrangian is obtained by restricting the augmented Lagrangian associated with (2) along the manifold that results from the explicit minimization over the z-variable [1]. The Moreau envelope is given by

$$M_{\mu g}(v) = g(\mathbf{prox}_{\mu g}(v)) + \tfrac{1}{2\mu}\|\mathbf{prox}_{\mu g}(v) - v\|^2$$

where the proximal operator of the function g is

$$\mathbf{prox}_{\mu g}(v) := \underset{z}{\text{argmin}}\ g(z) + \tfrac{1}{2\mu}\|z - v\|^2$$

and v is a given vector. It is noteworthy that the Moreau envelope is continuously differentiable [8], even when g is not, and its gradient is determined by

$$\nabla M_{\mu g}(v) = \tfrac{1}{\mu}\left(v - \mathbf{prox}_{\mu g}(v)\right).$$

B. Continuous-time primal-dual (PD) dynamics

The continuous differentiability of $\mathcal{L}_\mu(x;y)$ enables the use of PD gradient flow dynamics to compute saddle points of $\mathcal{L}_\mu(x;y)$ [1],

$$\dot{w} = F(w) \tag{4a}$$

where $w := [\, x^T \ y^T \,]^T$ and

$$\begin{aligned} F(w) &:= \begin{bmatrix} -\nabla_x \mathcal{L}_\mu(x;y) \\ \nabla_y \mathcal{L}_\mu(x;y) \end{bmatrix} \\ &= \begin{bmatrix} -(\nabla f(x) + T^T \nabla M_{\mu g}(Tx + \mu y)) \\ \mu\left(\nabla M_{\mu g}(Tx + \mu y) - y\right) \end{bmatrix} \end{aligned} \tag{4b}$$

As shown in [1], when the L_f-smooth term f is m_f-strongly convex, the regularizer g is convex, and T is full row rank, (4) with $\mu \geq L_f - m_f$ are globally exponentially stable with rate ρ that can be calculated explicitly [1, Remark 4].

The implementation of the continuous-time PD dynamics requires temporal discretization. We next utilize explicit forward Euler scheme to obtain a discrete-time version of (4).

III. Proposed algorithm and main results

In this section, we study exponential stability of a PD algorithm (4) with a constant step size α. In Section III-A, we introduce a discrete-time model. In Section III-B, we provide explicit bounds on the exponential decay rate for sufficiently small step sizes. In Section III-C, we allow for larger step sizes and formulate an LMI condition for certifying achievable exponential rates. Finally, in Section III-D, we characterize a range of step size values that guarantee global exponential stability.

A. Discretized primal-dual algorithm

The explicit forward Euler discretization of (4) with a constant step size α yields a discrete-time algorithm

$$w^{k+1} = w^k + \alpha F(w^k) \tag{5a}$$

with $w^k := [\, (x^k)^T \ (y^k)^T \,]^T$. Equivalently, we have,

$$\begin{aligned} x^{k+1} &= x^k - \alpha\left(\nabla f(x^k) + T^T \nabla M_{\mu g}(Tx^k + \mu y^k)\right) \\ y^{k+1} &= y^k + \alpha\mu\left(\nabla M_{\mu g}(Tx^k + \mu y^k) - y^k\right) \end{aligned} \tag{5b}$$

where k is the iteration index.

In what follows, we provide exponential convergence guarantees for (5) under different restrictions on α.

B. Exponential rate for sufficiently small step size

We first provide explicit bounds on the exponential convergence rate for (5) with a sufficiently small step size. Even though Lipschitz continuity of F was proved in [1, Theorem 1], we next derive the expression for the Lipschitz constant of F.

For any w_1 and w_2, we have

$$\begin{aligned} &\|F(w_1) - F(w_2)\| \\ &\quad \leq \|\nabla f(x_1) - \nabla f(x_2)\| + \mu\|y_1 - y_2\| + \\ &(1 + \tfrac{\lambda_m}{\mu})\|\mu\nabla M_{\mu g}(Tx_1 + \mu y_1) - \mu\nabla M_{\mu g}(Tx_2 + \mu y_2)\| \\ &\quad \leq (L_f + \lambda_m + \tfrac{\lambda_m^2}{\mu})\|x_1 - x_2\| + (2\mu + \lambda_m)\|y_1 - y_2\| \end{aligned}$$

where λ_m is the largest eigenvalue value of TT^T. This results follows from the use of triangle inequality and Lipschitz continuity of ∇f and $\nabla M_{\mu g}$. Lipschitz continuity of $\nabla M_{\mu g}$ is due to the firm non-expansiveness of proximal operators [8].

Since $\max(\|x_1 - x_2\|, \|y_1 - y_2\|) \leq \|w_1 - w_2\|$, we conclude that F is Lipschitz continuous with parameter,

$$\nu = L_f + 2\lambda_m + 2\mu + \tfrac{\lambda_m^2}{\mu}. \tag{6}$$

Similar to [16, Lemma 5], Theorem 1 exploits global exponential stability of the continuous-time gradient flow

dynamics (4) [1, Theorem 3], Lipschitz continuity of F, and Assumption 1 .

Theorem 1: Let Assumption 1 hold. Then there is $\alpha > 0$ such that all trajectories of (5) satisfy

$$\|w^k - \bar{w}\| \leq \sqrt{\kappa_p}\, r^k \, \|w^0 - \bar{w}\| \tag{7}$$

where $\bar{w}$ is the equilibrium point of (4),

$$r = \frac{\alpha^2 \nu^2 \kappa_p}{2} + \mathrm{e}^{-\rho\alpha} < 1 \tag{8}$$

ρ is the decay rate of the exponentially stable gradient flow dynamics (4), $P = P^T \succ 0$ is a matrix that certifies exponential stability of (4) with $(w(t) - \bar{w})^T P (w(t) - \bar{w})$, and κ_p is the condition number of P.

Proof: We fix k, and consider (4) with $w(0) = w^k$. According to [1, Theorem 3], when $\mu \geq L_f - m_f$, there exists a positive definite matrix P and a positive rate ρ such that,

$$\|w(t) - \bar{w}\|_P \leq \mathrm{e}^{-\rho t} \|w(0) - \bar{w}\|_P \tag{9}$$

where $\|w\|_P^2 := w^T P w$.

Since the proof is similar to [16, Lemma 5], we omit it here and only mention the existence of $\alpha > 0$. It should be noted that (8) is a continuous function of α. Furthermore, since $r = 1$ for $\alpha = 0$, and since the derivative of r with respect to α is negative for small α, (8) holds. ∎

Remark 1: When α is sufficiently small, we can expand the decay rate r with

$$r = 1 - \rho\alpha + \tfrac{1}{2}(\rho^2 + \nu^2 \kappa_p)\alpha^2 + O(\alpha^3) \tag{10}$$

We can approximate r without the last term. Thus, (8) gives a necessary bound,

$$0 < \alpha < \alpha_0 := \frac{2\rho}{\rho^2 + \nu^2 \kappa_p} \tag{11}$$

However, without knowing κ_p and ρ, this bound cannot be estimated. Therefore, it is difficult to establish the bounds on α to guarantee exponential convergence in this approach. Meanwhile, Theorem 1 does not provide insight into the problems with large step sizes.

C. LMI test for general exponential rate

To complement Theorem 1, we provide a unified LMI condition which can be used to test whether the discretized PD algorithm (5) with any given α converges exponentially at rate r. We build on the framework developed in [12] and analyze (5) as a discrete-time feedback system.

Let $u^k = [\,(u_1^k)^T\ (u_2^k)^T\,]^T$, $\xi^k = [\,(\xi_1^k)^T\ (\xi_2^k)^T\,]^T$, and

$$\begin{aligned}
\xi_1^k &:= x^k \\
\xi_2^k &:= Tx^k + \mu y^k \\
u_1^k &:= \nabla f(x^k) - m_f x^k = \Delta_1(\xi_1^k) \\
u_2^k &:= \mu \nabla M(Tx^k + \mu y^k) = \Delta_2(\xi_2^k).
\end{aligned}$$

PD dynamics (5) can then be represented as a discrete-time linear time-invariant system connected in feedback with a nonlinear block Δ,

$$\begin{aligned}
w^{k+1} &= Aw^k + Bu^k \\
\xi^k &= Cw^k
\end{aligned} \tag{12}$$

with

$$A = \begin{bmatrix} (1-\alpha m_f) I & 0 \\ 0 & (1-\alpha\mu) I \end{bmatrix}$$
$$B = \begin{bmatrix} -\alpha I & -\frac{\alpha}{\mu} T^T \\ 0 & \alpha I \end{bmatrix},\ C = \begin{bmatrix} I & 0 \\ T & \mu I \end{bmatrix}.$$

The input is given by $u^k = \Delta(\xi^k)$, where Δ is a 2×2 block diagonal matrix with the diagonal blocks Δ_1 and Δ_2. The nonlinear block Δ can be characterized using quadratic constraints. The notation $S(m, L)$ in [12] is used to denote functions that are continuously differentiable, strongly convex with parameter m, and have Lipschitz continuous gradients with parameter L.

Note that Δ_1 is the gradient of the convex function $f(\xi_1^k) - (m_f/2)\|\xi_1^k\|^2$, and Δ_2 is the scaled gradient of the convex Moreau envelope, it is not diffcult to show that $\Delta_1 \in S(m_1, L_1)$, where $m_1 = 0, L_1 = L_f - m_f$, and $\Delta_2 \in S(m_2, L_2)$, where $m_2 = 0$ and $L_2 = 1$.

At a stationary point $\bar{w} = \begin{bmatrix} \bar{x}^T & \bar{y}^T \end{bmatrix}^T$ of (12) we have $\bar{\xi}_1 = \bar{x}$, $\bar{\xi}_2 = T\bar{x} + \mu\bar{y}$, $\bar{u}_1 = \Delta_1(\bar{\xi}_1)$, $\bar{u}_2 = \Delta_2(\bar{\xi}_2)$, and apply [12, Proposition 5] to characterize $\Delta_i \in S(m_i, L_i)$ for $i = 1, 2$ via quadratic constraints,

$$\begin{bmatrix} \xi_i^k - \bar{\xi}_i \\ u_i^k - \bar{u}_i \end{bmatrix}^T \begin{bmatrix} -m_i L_i I & (L_i + m_i) I \\ (L_i + m_i) I & -2I \end{bmatrix} \begin{bmatrix} \xi_i^k - \bar{\xi}_i \\ u_i^k - \bar{u}_i \end{bmatrix} \geq 0.$$

The above two quadratic constraints can be combined into

$$(\eta^k - \bar{\eta})^T\ \Pi\ (\eta^k - \bar{\eta}) \geq 0, \tag{13}$$

where $\eta^k = \left[(\xi^k)^T\ (u^k)^T\right]^T$, $\bar{\eta} = \left[(\bar{\xi})^T\ (\bar{u})^T\right]^T$,

$$\Pi = \begin{bmatrix} 0 & \Pi_0 \\ \Pi_0 & -2I \end{bmatrix},\ \Pi_0 = \begin{bmatrix} \hat{L} I & 0 \\ 0 & I \end{bmatrix}.$$

Now, [12, Theorem 4] implies that the PD algorithm (5) with a step size α converges exponentially to the stationary point $\bar{w}$ at a rate r if there exists a positive definite P such that

$$\begin{bmatrix} A^T P A - r^2 P & A^T P B \\ B^T P A & B^T P B \end{bmatrix} + \begin{bmatrix} C^T & 0 \\ 0 & I \end{bmatrix} \Pi \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix} \prec 0. \tag{14}$$

Given (A, B, C) and r, (14) is an LMI condition in P. For a general T, the dimension of the above LMI scales with n. When T is the identity matrix, we can use the argument in [12, Section 4.2] to convert (14) to a 4×4 LMI.

The LMI condition (14) provides a general numerical tool for the convergence rate analysis. It is non-trivial to solve the LMI analytically for general α. In next subsection, we translate (14) into a frequency condition which is then analytically checked to find a large range of step size.

Remark 2: We can further reduce the conservatism in the LMI condition (14) by introducing additional decision variables, See the second to last remark in [12, Section 3.2]. Adding decision variables can lead to useful LMI conditions

for other distributed optimization methods. In [17], this type of LMI conditions have been used to obtain numerical rate bounds for EXTRA [18] and NIDS [19]. However, it is non-trivial to obtain analytical rate bounds via these LMIs.

D. Large step size guaranteeing exponential convergence

In contrast to the analysis in Section III-B, here we establish conditions on step sizes that guarantee exponential convergence. First, we apply the KYP lemma [20] to translate (14) to an equivalent frequency condition.

Lemma 2: Let $G(re^{\mathrm{j}\theta}) = C(re^{\mathrm{j}\theta}I - A)^{-1}B$ be stable for $r \in (0,1)$, where A, B, C are system matrices in (12), let Δ be characterized by a static quadratic constraint Π, and let

$$\begin{bmatrix} G(re^{\mathrm{j}\theta}) \\ I \end{bmatrix}^* \Pi \begin{bmatrix} G(re^{\mathrm{j}\theta}) \\ I \end{bmatrix} \prec 0, \quad \forall\, \theta \in [0,\, 2\pi). \tag{15}$$

Then, the feedback interconnection of G with Δ is exponentially stable with rate r.

Next, we give an upper bound on the step size α to ensure exponential stability of (12) when TT^T is a full rank matrix.

Theorem 3: Let Assumption 1 hold. Then, for the augmented Lagrangian parameter $\mu = L_f - m_f$, the discretized PD algorithm (5) converges exponentially to the optimal solution if the step size α satisfies one of the following conditions:

(i) If $m_f \geq \mu$,

$$0 \;<\; \alpha \;<\; \alpha_1 \;:=\; \tfrac{2}{\mu + m_f + \lambda_m/\mu}. \tag{16}$$

(ii) If $m_f < \mu$,

$$0 \;<\; \alpha \;<\; \min(\alpha_1, \alpha_2) \tag{17}$$

where

$$\alpha_2 \;:=\; \tfrac{a_0}{a_1}\tfrac{2}{1+\sqrt{1-\gamma}},\; \gamma = \tfrac{4\,a_0\,a_2}{a_1^2} \tag{18}$$

and the parameters a_i are given by

$$\begin{aligned} a_2 &= (\mu^2 + \mu m_f - m_f^2)\mu^2 m_f - (\mu^2 - 3\mu m_f + 2m_f^2)\mu\lambda_m \\ a_1 &= 2m_f\big((\mu - m_f)(\lambda_m + \mu m_f) + 2\mu^3\big) > 0 \\ a_0 &= 4m_f\mu^2 > 0. \end{aligned} \tag{19}$$

Proof: We establish the result for system (12) in which the nonlinear block satisfies quadratic inequality (13) with $\mu = \hat{L} = L_f - m_f$. We utilize Lemma 2 to show the exponential convergence with some rate $r \in (0,1)$ by verifying (15) through a series of conditions on α.

Let $\zeta := \cos\theta$. Evaluating the left-hand side of (15) for $\mu = \hat{L}$ and dividing by -2 yield the matrix inequality

$$\begin{bmatrix} a(\zeta)I & b(\zeta)T^T \\ b(\zeta)T & c(\zeta)I + d(\zeta)TT^T \end{bmatrix} \succ 0,\; \forall\, \zeta \in [-1,1] \tag{20a}$$

where

$$\begin{aligned} a(\zeta) &:= 1 + \alpha\,\mu\,h_m(\alpha, r, \zeta) \\ b(\zeta) &:= \alpha\,h_m(\alpha, r, \zeta) \\ c(\zeta) &:= 1 - \alpha\,\mu\,h_\mu(\alpha, r, \zeta) \\ d(\zeta) &:= \tfrac{\alpha}{\mu}\,h_m(\alpha, r, \zeta) \\ h_m(\alpha, r, \zeta) &:= \tfrac{\alpha m_f - 1 + r\zeta}{(\alpha m_f - 1 + r\zeta)^2 + r^2(1-\zeta^2)} \\ h_\mu(\alpha, r, \theta) &:= \tfrac{\alpha\mu - 1 + r\zeta}{(\alpha\mu - 1 + r\zeta)^2 + r^2(1-\zeta^2)}. \end{aligned} \tag{20b}$$

Proving (20a) amounts to establishing:

1) stability of the transfer function $G(re^{\mathrm{j}\theta})$;
2) positive definiteness of the $(1,1)$ block in (20a) via,

$$a(\zeta) \;>\; 0, \quad \forall\, \zeta \in [-1,1]; \tag{21a}$$

3) positive definiteness of the Schur complement [21]

$$c(\zeta)I + \Big(d(\zeta) - \tfrac{b^2(\zeta)}{a(\zeta)}\Big)TT^T \succ 0, \quad \forall\, \zeta \in [-1,1].$$

This condition amounts to checking

$$c(\zeta) + \Big(d(\zeta) - \tfrac{b^2(\zeta)}{a(\zeta)}\Big)\lambda_i \;>\; 0, \quad \forall\, \zeta \in [-1,1] \tag{21b}$$

for each eigenvalue λ_i of TT^T.

Next, we establish the conditions on the step size α such that above three conditions hold for some rate $r \in (0,1)$.

Stability of the transfer function $G(re^{\mathrm{j}\theta})$

The transfer function $G(re^{\mathrm{j}\theta})$ is stable if and only if $|1 - \alpha m_f| < r$ and $|1 - \alpha\mu| < r$. Since r belongs to an open interval $(0,1)$, these conditions hold if $|1 - \alpha m_f| < 1$ and $|1 - \alpha\mu| < 1$ which yield the restriction on α,

$$0 \;<\; \alpha \;<\; \min\big(\tfrac{2}{m_f},\, \tfrac{2}{\mu}\big). \tag{C1}$$

Checking condition (21a)

Since $a(\zeta)$ is a linear fractional function of ζ, it is quasilinear [21]. This implies that (21a) can be established by checking $a(1) > 0$ and $a(-1) > 0$,

$$a(1) \;=\; \tfrac{\alpha(\mu + m_f) - 1 + r}{\alpha m_f - 1 + r} \;>\; 0 \tag{22a}$$

$$a(-1) \;=\; \tfrac{\alpha(\mu + m_f) - 1 - r}{\alpha m_f - 1 - r} \;>\; 0. \tag{22b}$$

From (C1) it follows that the denominator in (22a) is positive and that the denominator in (22b) is negative. Thus, (22a) is satisfied if $1 - \alpha(\mu + m_f) < r < 1$ which clearly holds for all positive α. On the other hand, (22b) holds if $\alpha(\mu + m_f) - 1 < r < 1$. Therefore, condition (21a) holds if

$$0 \;<\; \alpha \;<\; \tfrac{2}{\mu + m_f}. \tag{C2}$$

Clearly, (C2) is more restrictive than (C1).

Checking condition (21b)

The Schur complement (21b) can be written as

$$s(\zeta) \;:=\; 1 + s_1(\zeta) + s_2(\zeta) \;>\; 0, \quad \forall\, \zeta \in [-1,1] \tag{23}$$

where

$$s_1(\zeta) \;:=\; \frac{-\alpha\mu\,(\alpha\mu-1+r\zeta)}{(\alpha\mu-1+r\zeta)^2+r^2(1-\zeta^2)}$$

$$s_2(\zeta) \;:=\; \frac{(\alpha\,\lambda_i/\mu)(\alpha m_f-1+r\zeta)}{(\alpha m_f-1+r\zeta)^2+\alpha\mu(\alpha m_f-1+r\zeta)+r^2(1-\zeta^2)}.$$

Since both s_1 and s_2 are linear fractions of ζ, they are quasilinear functions of ζ. Furthermore, under (C2), their denominators are strictly positive for all $\zeta \in [-1,1]$. Thus, s_1 and s_2 are well-defined for all $\zeta \in [-1,\,1]$. By checking derivatives of s_1 and s_2: $\mathrm{d}s_1/\mathrm{d}\zeta < 0$, $\mathrm{d}s_2/\mathrm{d}\zeta > 0$, we know that s_1 (respectively s_2) is monotonically decreasing (respectively increasing) over the interval $[-1,1]$ and that their extreme values take place at $\zeta = \pm 1$.

To prove exponential convergence with some rate $r \in (0,1)$, we show (21b) for $r = 1$. By continuity of the functions s_1 and s_2 in r, this establishes the existence of $r \in (0,1)$ such that (23) holds. Thus, in the remainder of the proof, we take $r = 1$.

Checking condition (21b) **at** $\zeta = \pm 1$

Evaluating functions s_1 and s_2 at $\zeta = \pm 1$ and $r = 1$ yields,

$$s_1(-1) \;=\; \tfrac{-\alpha\mu}{\alpha\mu-2} \;>\; 0, \qquad s_1(1) \;=\; -1,$$

$$s_2(1) \;=\; \tfrac{\lambda_i/\mu}{\mu+m_f} \;>\; 0, \quad s_2(-1) \;=\; \tfrac{\alpha\lambda_i/\mu}{\alpha(\mu+m_f)-2} \;<\; 0.$$

Clearly, at $\zeta = 1$, condition (23) holds for all values of α. On the other hand, since $s_1(-1) > 0$, the sufficient condition on α for $s(-1) > 0$ is obtained from

$$1 \;+\; s_2(-1) \;=\; 1 \;-\; \tfrac{\alpha\lambda_i/\mu}{2-\alpha(\mu+m_f)} \;>\; 0.$$

Let the largest eigenvalue of TT^T be λ_m. The above condition holds for all i if it holds for $\lambda_i = \lambda_m$. Thus, relative to (C1) and (C2), further restricts the values that the step size α can take,

$$0 \;<\; \alpha \;<\; \tfrac{2}{\mu+m_f+\lambda_m/\mu}. \tag{C3}$$

From (C3), it follows that the quasilinear functions s_1 and s_2 for $r = 1$ satisfy

$$s_1(\zeta) \;>\; -1,\; s_2(\zeta) \;>\; -1,\; \forall\, \zeta \;\in\; (-1,1). \tag{24}$$

Checking condition (21b) **for** $\zeta \in (-1,1)$

We note that $1 + s_1(\zeta)$ is positive for all $\zeta \in (-1,\,1)$ and that $s_2(\zeta)$ is non-negative for $\zeta \in [1-\alpha m_f, 1)$. Thus, we only need to check (23) for $\zeta \in (-1,\,1-\alpha m_f)$. Since s_1 is a decreasing function with $s_1(-1) > 0$ and since s_2 is an increasing function with $1 + s_2(-1) > 0$, under (C3), for any $\zeta \in (-1, 1-\alpha m_f)$, we have

$$\begin{aligned} 1 \;+\; s_1(\zeta) \;+\; s_2(\zeta) \;&>\; 1 \;+\; s_1(1-\alpha m_f) \;+\; s_2(-1) \\ &>\; s_1(1-\alpha m_f) \\ &=\; \tfrac{\alpha\mu(m_f-\mu)}{\alpha(m_f-\mu)^2+m_f(2-\alpha m_f)}. \end{aligned} \tag{25}$$

Since the denominator is always positive, the sign of $s_1(1-\alpha m_f)$ is determined by the sign of $m_f - \mu$. In what follows, we examine the two relevant cases.

1) Case 1: $m_f \geq \mu$: Under (C3) on α, we have $s_1(1-\alpha m_f) \geq 0$. Thus, from (25) we see that $1+s_1(\zeta)+s_2(\zeta) > 0$ holds for all $\zeta \in (-1, 1-\alpha m_f)$ when $m_f \geq \mu$; the left task is to establish conditions on α to ensure $1+s_1(\zeta)+s_2(\zeta) > 0$ for $\zeta \in (-1, 1-\alpha m_f)$ when $m_f < \mu$.

2) Case 2: $m_f < \mu$: We split $(-1,\, 1-\alpha m_f)$ into two intervals $(-1, 1-\alpha\mu]$ and $(1-\alpha\mu, 1-\alpha m_f)$ where $s_1(1-\alpha\mu) = 0$. Using the argument similar to that preceding equation (25), under (C3), for any $\zeta \in (-1, 1-\alpha\mu]$, we have

$$1 \;+\; s_1(\zeta) \;+\; s_2(\zeta) \;>\; 1 \;+\; s_2(-1) \;>\; 0. \tag{26}$$

Since s_1 is decreasing and s_2 is increasing, for any $\zeta \in (1-\alpha\mu, 1-\alpha m_f)$, we have

$$1 + s_1(\zeta) + s_2(\zeta) \;>\; 1 + s_1(1-\alpha m_f) + s_2(1-\alpha\mu). \tag{27}$$

Thus, when $m_f < \mu$, a sufficient condition for the stepsize α to guarantee (23) is given by $1+s_1(1-\alpha m_f)+s_2(1-\alpha\mu) > 0$, which can be simplified into

$$a_2\alpha^2 \;-\; a_1\alpha \;+\; a_0 \;>\; 0 \tag{28}$$

where parameters a_i are given by (19).

The discriminant associated with the quadratic inequality (28) is always positive,

$$\begin{aligned} D \;&=\; a_1^2 \;-\; 4a_0a_2 \\ &=\; 4m_f(\mu \;-\; m_f)^2(\lambda_m^2 m_f + \mu^2 m_f^3 + 2\mu(m_f^2+2\mu^2)\lambda_m) \end{aligned}$$

and the condition (23) is satisfied if

$$0 \;<\; \alpha \;<\; \tfrac{a_0}{a_1}\,\tfrac{2}{1+\sqrt{1-4\,a_0\,|a_2|\,\mathrm{sign}\,(a_2)/a_1^2}} \tag{C4}$$

where

$$\tfrac{a_0}{a_1} \;=\; \tfrac{2}{\mu+m_f+(\mu-m_f)(\lambda_m/\mu+\mu+m_f)/\mu}.$$

Thus, by combining (C3) and (C4) we complete the proof. ∎

The upper bound (16) or (17) in Theorem 3 also holds for $\hat{\mu} > \mu = L_f - m_f$, since we can always choose $\hat{m}_f = m_f$ and $\hat{L}_f = L_f + (\hat{\mu} - \mu)$, and L_f-Lipschitz ∇f is also $\hat{L}_f$-Lipschiz. Meanwhile, the upper bound on α depends on the largest eigenvalue λ_m of the positive definite matrix TT^T.

IV. Computational experiments

We consider a quadratic optimization problem [22, (18)]

$$\begin{array}{ll} \underset{x,\,z}{\text{minimize}} & \frac{1}{2}\,x^TQx \;+\; q^Tx \;+\; g(z) \\ \text{subject to} & Tx \;-\; z \;=\; 0. \end{array} \tag{29}$$

where $x, q \in \mathbb{R}^n$, $T \in \mathbb{R}^{m\times n}$, $Q \in \mathbb{R}^{n\times n}$ is a positive definite matrix, and $g(z)$ is the indicator function as $g(z) = 0$ for $z \leq c$ and $g(z) = +\infty$ otherwise, where $c \in \mathbb{R}^m$.

Denote $f(x) = (1/2)x^TQx + q^Tx$, (29) is a case of the problem (2). We use the proposed PD algorithm (5). We choose L_f and m_f be the largest and the smallest eigenvalue of Q, and $\mu = L_f - m_f$. The gradient of the Moreau envelope $\nabla M_{\mu g}(v_i)$ is given by $\max(0,\,(v_i - c_i)/\mu)$.

We generate problem instances in Matlab. We set $n = m = 10$, $q = 10 \times \mathsf{randn}(n,1)$, and $Q = EE^T + F$,

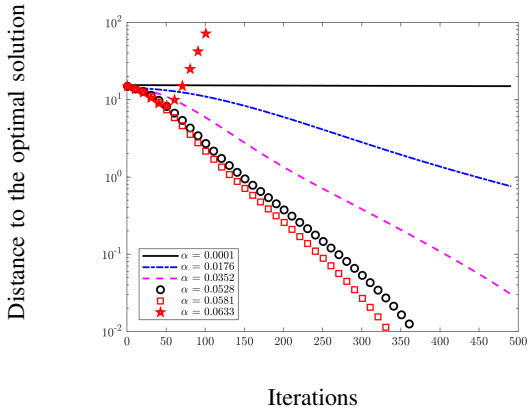

Fig. 1: Problem instance with $L_f = 32.44$ and $m_f = 0.87$.

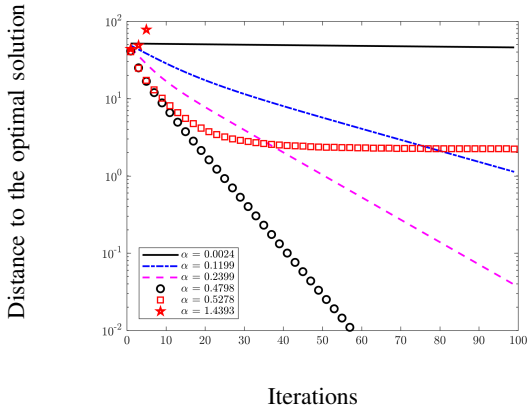

Fig. 2: Problem instance with $L_f = 0.92$ and $m_f = 0.62$.

where $E = \mathsf{randn}(n, n)$ and $F = \mathsf{diag}(\mathsf{exp}(\mathsf{randn}(\mathsf{n}, 1)))$. We choose c as a vector with all ones, and $T = I$. This class of instances have large condition numbers, i.e., $L_f \gg m_f$, which corresponds to the case of $\mu > m_f$. We choose the step size α smaller than $\min(\alpha_1, \alpha_2)$ in (17).

We report one instance as shown in Fig. 1. We take several step sizes near the upper bound: $\min(\alpha_1, \alpha_2) = 0.0528$ (a circle line in Fig. 1). We show the convergence in Fig. 1 by distances of iterations x^k and y^k to the optimal.

For comparison, we test some well-conditioned instances. We rescale the singular values of Q to reduce $L_f - m_f$. One instance is shown in Fig. 2. We choose several step sizes near the upper bound: $\alpha_1 = 0.4798$ using (16) (a line of circles in Fig. 2). We show the convergence in Fig. 2 by distances of iterations x^k and y^k to the optimal.

In Figs. 1 and 2, we show that our algorithm converges exponentially if a step size is selected within upper bounds in Theorem 3. Within these bounds (circle lines in figures), by increasing the step size, the algorithm converges faster and faster. Therefore, a large step size is preferable empirically. It is noted that, black (or solid) lines indicate very slow convergence for small step sizes.

Unfortunately, if we choose larger step sizes beyond upper bounds, the aglorithm can diverge quickly as a star line shown in Figs. 1 and 2. A slightly larger step size over upper bounds may speedup as a square line shown in Fig. 1, but it could be detrimental as shown in Fig. 2. Admittedly, our step size upper bounds are still conservative.

Finally, to show how small analytic step size bounds α_0 in (11) are, we compute rate ρ and P from the matrix inequality condition [1, (16)] using bisection search on ρ. For the above ill- and well-conditioned cases, necessary step size bounds α_0 are 5.7×10^{-16} and 3.6×10^{-4}, respectively, and our upper bounds are significantly larger.

REFERENCES

[1] N. K. Dhingra, S. Z. Khong, and M. R. Jovanović, "The proximal augmented Lagrangian method for nonsmooth composite optimization," *IEEE Trans. Automat. Control*, 2018, doi:10.1109/TAC.2018.2867589.

[2] S. Sra, S. Nowozin, and S. Wright, *Optimization for machine learning*. MIT Press, 2012.

[3] F. Lin, M. Fardad, and M. R. Jovanović, "Design of optimal sparse feedback gains via the alternating direction method of multipliers," *IEEE Trans. Automat. Control*, vol. 58, no. 9, pp. 2426–2431, 2013.

[4] M. R. Jovanović and N. K. Dhingra, "Controller architectures: tradeoffs between performance and structure," *Eur. J. Control*, vol. 30, pp. 76–91, 2016.

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[6] S. Wright and J. Nocedal, "Numerical optimization," *Springer Science*, vol. 35, no. 67-68, p. 7, 1999.

[7] D. P. Bertsekas, *Nonlinear programming*. Athena Scientific, 1999.

[8] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[9] Y. Nesterov, "Primal-dual subgradient methods for convex problems," *Math. Program.*, vol. 120, no. 1, pp. 221–259, 2009.

[10] C. Chen, R. H. Chan, S. Ma, and J. Yang, "Inertial proximal ADMM for linearly constrained separable convex optimization," *SIAM J. Imag. Sci.*, vol. 8, no. 4, pp. 2239–2267, 2015.

[11] M. Hong, D. Hajinezhad, and M. Zhao, "Prox-PDA: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks," in *International Conference on Machine Learning*, 2017, pp. 1529–1538.

[12] L. Lessard, B. Recht, and A. Packard, "Analysis and design of optimization algorithms via integral quadratic constraints," *SIAM J. Optim.*, vol. 26, no. 1, pp. 57–95, 2016.

[13] A. Wibisono, A. Wilson, and M. Jordan, "A variational perspective on accelerated methods in optimization," *Proceedings of the National Academy of Sciences*, p. 201614734, 2016.

[14] A. Wilson, B. Recht, and M. Jordan, "A Lyapunov analysis of momentum methods in optimization," *arXiv preprint arXiv:1611.02635*, 2016.

[15] B. Hu and L. Lessard, "Dissipativity theory for Nesterov's accelerated method," in *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[16] G. Qu and N. Li, "On the exponential stability of primal-dual gradient dynamics," *IEEE Control Syst. Lett.*, vol. 3, no. 1, pp. 43–48, 2019.

[17] A. Sundararajan, B. Hu, and L. Lessard, "Robust convergence analysis of distributed optimization algorithms," in *Communication, Control, and Computing (Allerton), 55th Annual Allerton Conference on*, 2017, pp. 1206–1212.

[18] W. Shi, Q. Ling, G. Wu, and W. Yin, "EXTRA: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[19] Z. Li, W. Shi, and M. Yan, "A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates," *arXiv preprint arXiv:1704.07807*, 2017.

[20] A. Rantzer, "On the Kalman-Yakubovich-Popov lemma," *Syst. Control Lett.*, vol. 28, no. 1, pp. 7–10, 1996.

[21] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[22] E. Ghadimi, A. Teixeira, I. Shames, and M. Johansson, "Optimal parameter selection for the alternating direction method of multipliers (ADMM): quadratic problems," *IEEE Trans. Automat. Control*, vol. 60, no. 3, pp. 644–658, 2015.