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 $\ell_1$ 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.
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

\[ \min_{x} \quad f(x) + g(Tx) \tag{1} \]

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 \( z \in \mathbb{R}^m \),

\[ \min_{x,z} \quad f(x) + g(z) \]

subject to \( Tx - z = 0. \tag{2} \]

A. Proximal augmented Lagrangian

The proximal augmented Lagrangian is defined as \( [1] \),

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

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

\[ M_\mu g(v) = g(\text{prox}_{\mu g}(v)) + \frac{1}{2\mu} \|\text{prox}_{\mu g}(v) - v\|^2 \]

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

\[ \nabla M_\mu g(v) = \frac{1}{\mu} (v - \text{prox}_{\mu g}(v)). \]

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

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

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 \( \rho \) that can be calculated explicitly \([1, \text{ Remark 4}]\).

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 \( \alpha \). 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 \( \alpha \) yields a discrete-time algorithm

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

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

\[ x^{k+1} = x^k - \alpha (\nabla f(x^k) + T^T \nabla M_\mu g(Tx^k + \mu y^k)) \]

\[ y^{k+1} = y^k + \alpha \mu (\nabla M_\mu g(Tx^k + \mu y^k) - y^k) \tag{5b} \]

where \( k \) is the iteration index.

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

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 \([1, \text{ Theorem 1}]\), we next derive the expression for the Lipschitz constant of \( F \).

For any \( w_1 \) and \( w_2 \), we have

\[ \|F(w_1) - F(w_2)\| \leq \|\nabla f(x_1) - \nabla f(x_2)\| + \mu \|y_1 - y_2\| 
\]

\[ + (1 + \frac{\lambda_m}{\mu}) \|\mu \nabla M_\mu g(Tx_1 + \mu y_1) - \mu \nabla M_\mu g(Tx_2 + \mu y_2)\| \]

\[ \leq (L_f + \lambda_m + \frac{\lambda_m^2}{\mu}) \|x_1 - x_2\| + (2\mu + \lambda_m) \|y_1 - y_2\| \]

where \( \lambda_m \) is the largest eigenvalue value of \( TT^T \). This results follows from the use of triangle inequality and Lipschitz continuity of \( \nabla f \) 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,

\[ \nu = L_f + 2\lambda_m + 2\mu + \frac{\lambda_m^2}{\mu}. \tag{6} \]

Similar to \([16, \text{ Lemma 5}]\), Theorem 1 exploits global exponential stability of the continuous-time gradient flow.
Theorem 1: Let Assumption 1 hold. Then there is \( \alpha > 0 \) such that all trajectories of (5) satisfy
\[
\|w^k - \bar{w}\| \leq \sqrt{\kappa_p} r^k \|w^0 - \bar{w}\|
\]
where \( \bar{w} \) is the equilibrium point of (4),
\[
r = \frac{\alpha^2 \nu^2 \kappa_p}{2} + e^{-\rho \alpha} < 1
\]
\( \rho \) is the decay rate of the exponentially stable gradient flow dynamics (4), \( P = P^T > 0 \) is a matrix that certifies exponential stability of (4) with \( (w(t) - \bar{w})^T P (w(t) - \bar{w}) \), and \( \kappa_p \) 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 \( \rho \) such that,
\[
\|w(t) - \bar{w}\|_P \leq e^{-\rho t} \|w(0) - \bar{w}\|_P
\]
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 \( \alpha \). Furthermore, since \( r = 1 \) for \( \alpha = 0 \), and since the derivative of \( r \) with respect to \( \alpha \) is negative for small \( \alpha \), (8) holds.

Remark 1: When \( \alpha \) is sufficiently small, we can expand the decay rate \( r \) without the last term. Thus, (8) gives a necessary bound,
\[
0 < \alpha < \alpha_0 := \frac{2\rho}{\rho^2 + \nu^2 \kappa_p}
\]
However, without knowing \( \kappa_p \) and \( \rho \), this bound cannot be estimated. Therefore, it is difficult to establish the bounds on \( \alpha \) 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 \( \alpha \) converges exponentially at rate \( r \). We build on the framework developed in [12] and analyze (5) as a discrete-time feedback system.

Let \( u^k = [(u^k_1)^T (u^k_2)^T]^T \), \( \xi^k = [(\xi^k_1)^T (\xi^k_2)^T]^T \), and
\[
\begin{align*}
\xi^k_1 & := x^k \\
\xi^k_2 & := T x^k + \mu y^k \\
\dot{u}^k_1 & := \nabla f (x^k) - m_f x^k = \Delta_1(\xi^k_1) \\
\dot{u}^k_2 & := \mu \nabla M (T x^k + \mu y^k) = \Delta_2(\xi^k_2)
\end{align*}
\]
PD dynamics (5) can then be represented as a discrete-time linear time-invariant system connected in feedback with a nonlinear block \( \Delta \),
\[
w^{k+1} = Aw^k + Bu^k
\]
with
\[
\begin{align*}
A & = \begin{bmatrix} (1 - \alpha m_f) I & 0 \\ 0 & (1 - \alpha \mu) I \end{bmatrix} \\
B & = \begin{bmatrix} -\alpha I & -\frac{\mu}{\alpha} T^T \\ 0 & \alpha I \end{bmatrix}, \quad C = \begin{bmatrix} I & 0 \\ T & \mu I \end{bmatrix}
\end{align*}
\]
The input is given by \( u^k = \Delta(\xi^k) \), where \( \Delta \) is a \( 2 \times 2 \) block diagonal matrix with the diagonal blocks \( \Delta_1 \) and \( \Delta_2 \). The nonlinear block \( \Delta \) 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 \( \Delta_1 \) is the gradient of the convex function \( f(\xi^k) - (m_f/2) \|\xi^k\|_2^2 \), and \( \Delta_2 \) is the scaled gradient of the convex Moreau envelope, it is not difficult 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 \( \bar{w} = [\bar{x}^T \bar{y}^T]^T \) of (12) we have \( \dot{\xi}_1 = \bar{x}, \dot{\xi}_2 = T \bar{x} + \mu \bar{y} \), \( \bar{u}_1 = \Delta_1(\bar{\xi}_1), \bar{u}_2 = \Delta_2(\bar{\xi}_2) \), and apply [12, Proposition 5] to characterize \( \Delta_1 \in S(m_i, L_i) \) for \( i = 1, 2 \) via quadratic constraints,
\[
\begin{bmatrix} \xi^k - \bar{\xi}_1 \\ u^k_1 - \bar{u}_1 \end{bmatrix}^T \begin{bmatrix} -m_1 L_1 I & (L_i + m_i) I \\ (L_i + m_i) I & -2I \end{bmatrix} \begin{bmatrix} \xi^k - \bar{\xi}_1 \\ u^k_1 - \bar{u}_1 \end{bmatrix} \geq 0.
\]

The above two quadratic constraints can be combined into
\[
(\eta^k - \bar{\eta})^T \Pi (\eta^k - \bar{\eta}) \geq 0,
\]
where \( \eta^k = [(\xi^k)^T (u^k)^T]^T, \bar{\eta} = [(\bar{\xi}^T (\bar{u})^T)^T, \Pi = \begin{bmatrix} 0 & I_0 \\ I_0 & -2I \end{bmatrix}, \Pi_0 = \begin{bmatrix} L I & 0 \\ 0 & I \end{bmatrix}. \)

Now, [12, Theorem 4] implies that the PD algorithm (5) with a step size \( \alpha \) converges exponentially to the stationary point \( \bar{w} \) at a rate \( r \) if there exists a positive definite \( P \) such that
\[
\begin{bmatrix} A^T PA - r^2 P & A^T PB \\ B^T PA & B^T PB \end{bmatrix} + \begin{bmatrix} \Pi & \Pi \end{bmatrix} C \begin{bmatrix} \Pi & \Pi \end{bmatrix} T < 0.
\]
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 \( 4 \times 4 \) 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 \( \alpha \). 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 $G(\exp i\theta) = C(\exp i\theta)I - A)^{-1}B$ be stable for $r \in (0, 1)$, where $A, B, C$ are system matrices in (12), let $\Delta$ be characterized by a static quadratic constraint II, and let

$$
\begin{bmatrix}
G(\exp i\theta) \\
I
\end{bmatrix}^* \Pi \begin{bmatrix}
G(\exp i\theta) \\
I
\end{bmatrix} \prec 0, \quad \forall \theta \in [0, 2\pi).
$$

(15)

Then, the feedback interconnection of $G$ with $\Delta$ is exponentially stable with rate $r$.

Next, we give an upper bound on the step size $\alpha$ 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 $\alpha$ satisfies one of the following conditions:

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

$$
0 < \alpha < \alpha_1 := \frac{2}{\mu + m_f + \lambda_m/\mu^2}.
$$

(ii) If $m_f < \mu$,

$$
0 < \alpha < \min (\alpha_1, \alpha_2)
$$

(17)

where

$$
\alpha_2 := \frac{a_0}{\frac{a_1}{\alpha_1} + \sqrt{\frac{a_1}{\alpha_1} + 4\alpha h}} = \frac{4\alpha h}{a_2}
$$

(18)

and the parameters $a_i$ are given by

$$
a_2 = (\mu^2 + \mu m_f - m_f^2)\mu^2 m_f - (\mu^2 - 3\mu m_f + 2m_f^2)\mu \lambda_m
$$

$$
a_1 = 2m_f((\mu - m_f)(\lambda_m + \mu m_f) + 2\mu^3) > 0
$$

$$
a_0 = 4m_f \mu^2 > 0.
$$

(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 $\alpha$.

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

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

(20a)

where

$$
\begin{align*}
a(\zeta) &:= 1 + \alpha \mu h_m(\alpha, r, \zeta) \\
b(\zeta) &:= \alpha h_m(\alpha, r, \zeta) \\
c(\zeta) &:= 1 - \alpha \mu h_\mu(\alpha, r, \zeta) \\
d(\zeta) &:= \frac{\alpha}{\mu} h_m(\alpha, r, \zeta)
\end{align*}
$$

(20b)

$$
h_{\mu}(\alpha, r, \theta) := \frac{\alpha m_f - 1 + r^2}{(\alpha m_f - 1 + r)^2 + r^2(1 - \zeta^2)}
$$

$$
h_m(\alpha, r, \zeta) := \frac{\lambda_m}{(\alpha m_f - 1 + r^2) + r^2(1 - \zeta^2)}.
$$

Proving (20a) amounts to establishing:

1) stability of the transfer function $G(\exp i\theta)$;
2) positive definiteness of the $(1, 1)$ block in (20a) via,

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

(21a)

3) positive definiteness of the Schur complement [21]

$$
c(\zeta)I + \left( (\hat{d}(\zeta) - \frac{\hat{a}^2(\zeta)}{\hat{a}(\zeta)}) TT^T \right) \lambda_i > 0, \quad \forall \zeta \in [-1, 1],
$$

(21b)

for each eigenvalue $\lambda_i$ of $TT^T$.

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

**Stability of the transfer function $G(\exp i\theta)$**

The transfer function $G(\exp i\theta)$ 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 $\alpha$,

$$
0 < \alpha < \min(\frac{2}{m_f}, \frac{2}{\mu}).
$$

(21c)

**Checking condition (21a)**

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

$$
a(1) = \frac{\alpha(\mu + m_f) - 1 + r}{\alpha m_f - 1 + r} > 0
$$

(22a)

$$
a(-1) = \frac{\alpha(\mu + m_f) - 1 - r}{\alpha m_f - 1 - r} > 0.
$$

(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 $\alpha$. On the other hand, (22b) holds if $\alpha(\mu + m_f) - 1 < r < 1$. Therefore, condition (21a) holds if

$$
0 < \alpha < \frac{2}{\mu + m_f}.
$$

(22c)

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

**Checking condition (21b)**

The Schur complement (21b) can be written as

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

(23)
where
\[ s_1(ζ) := \frac{-αµ(αµ−1 +rζ)}{(αµ−1 +rζ)^2 + r^2(1−ζ^2)} \]
\[ s_2(ζ) := \frac{(αλ_i/µ)(αmf−1 +rζ)}{(αmf−1 +rζ)^2 + αµ(αmf−1 +rζ) + r^2(1−ζ^2)}. \]

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 \( ζ ∈ [-1, 1] \). Thus, \( s_1 \) and \( s_2 \) are well-defined for all \( ζ ∈ [-1, 1] \). By checking derivatives of \( s_1 \) and \( s_2 \): \( ds_1/dζ < 0 \) and \( ds_2/dζ > 0 \), 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 \( ζ = ±1 \).

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

**Checking condition (21b) at \( ζ = ±1 \)**

Evaluating functions \( s_1 \) and \( s_2 \) at \( ζ = ±1 \) and \( r = 1 \) yields,
\[ s_1(-1) = \frac{-αµ}{αµ−2} > 0, \quad s_1(1) = -1, \]
\[ s_2(1) = \frac{λ_i/µ}{µ+mf} > 0, \quad s_2(-1) = \frac{αλ_i/µ}{α(µ+mf)−2} < 0. \]

Clearly, at \( ζ = 1 \), 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
\[ 1 + s_2(−1) = 1 - \frac{2αλ_i/µ}{2(µ+mf)} > 0. \]

Let the largest eigenvalue of \( TT^T \) be \( λ_m \). The above condition holds for all \( i \) if it holds for \( λ_i = λ_m \). Thus, relative to (C1) and (C2), further restricts the values that the step size \( α \) can take,
\[ 0 < α < \frac{2}{µ+mf + λ_m/µ}. \] (C3)

From (C3), it follows that the quasilinear functions \( s_1 \) and \( s_2 \) for \( r = 1 \) satisfy
\[ s_1(ζ) > -1, \quad s_2(ζ) > -1, \quad ∀ \ ζ ∈ (-1, 1). \] (24)

**Checking condition (21b) for \( ζ ∈ (-1, 1) \)**

We note that \( 1 + s_1(ζ) \) is positive for all \( ζ ∈ (-1, 1) \) and that \( s_2(ζ) \) is non-negative for \( ζ ∈ (1 − αmf, 1) \). Thus, we only need to check (23) for \( ζ ∈ (1 - 1, 1 - αmf) \). 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 \( ζ ∈ (−1, 1 − αmf) \), we have
\[ 1 + s_1(ζ) + s_2(ζ) > 1 + s_1(1 − αmf) + s_2(-1) \]
\[ > s_1(1 − αmf) + s_2(-1) \]
\[ = \frac{αµ(λ_i/µ)}{α(λ_i/µ) - 2 + αµ(1 − αmf)} \] (25)

Since the denominator is always positive, the sign of \( s_1(1 − αmf) \) is determined by the sign of \( mf − µ \). In what follows, we examine the two relevant cases.

1) **Case 1**: \( mf ≥ µ \): Under (C3) on \( α \), we have \( s_1(1 − αmf) ≥ 0 \). Thus, from (25) we see that \( 1 + s_1(ζ) + s_2(ζ) > 0 \) holds for all \( ζ ∈ (−1, 1 − αmf) \) when \( mf ≥ µ \); the left task is to establish conditions on \( α \) to ensure \( 1 + s_1(ζ) + s_2(ζ) > 0 \) for \( ζ ∈ (−1, 1 − αmf) \) when \( mf < µ \).

2) **Case 2**: \( mf < µ \): We split \( (−1, 1 − αmf) \) into two intervals \((-1, 1 − αmf)\) and \((1 − αmf, 1)\) where \( s_1(1 − αmf) = 0 \). Using the argument similar to that preceding equation (25), under (C3), for any \( ζ ∈ (−1, 1 − αmf) \), we have
\[ 1 + s_1(ζ) + s_2(ζ) > 1 + s_2(-1) > 0. \] (26)

Since \( s_1 \) is decreasing and \( s_2 \) is increasing, for any \( ζ ∈ (−1, 1 − αmf) \), we have
\[ 1 + s_1(ζ) + s_2(ζ) > 1 + s_1(1 − αmf) + s_2(1 − αmf). \] (27)

Thus, when \( mf < µ \), a sufficient condition for the stepsize \( α \) to guarantee (23) is given by \( 1 + s_1(1 − αmf) + s_2(1 − αmf) > 0 \), which can be simplified into
\[ a_2α^2 - a_1α + a_0 > 0 \] (28)

where parameters \( a_i \) are given by (19).

The discriminant associated with the quadratic inequality (28) is always positive,
\[ D = a_1^2 - 4a_0a_2 \]
\[ = 4mf(μ - mf)^2(λ_m^2mf + μ^2m_f^2 + 2μ(m_f^2 + μ^2)λ_m) \]
and the condition (23) is satisfied if
\[ 0 < α < \frac{α_0}{a_1} + \sqrt{1 - \frac{4a_0}{a_1^2}sign(s_2)^2/α_1^2} \] (C4)
\[ \text{where} \quad \frac{α_0}{a_1} = \frac{μ + mf + (μ - mf)(λ_m^2μ + μ + m_f)}{μ}. \]

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

The upper bound (16) or (17) in Theorem 3 also holds for \( µ > μ = L_f - mf \), since we can always choose \( m_{if} = mf \) and \( L = \bar{f}L + L_f + (\bar{f} - μ) \), and \( L_f \)-Lipschitz \( \nabla f \) is also \( L_f \)-Lipschitz. 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))
\[ \text{minimize}_{x,z} \quad \frac{1}{2} x^T Q x + q^T x + g(z) \] (29)
subject to \( Tx - z = 0 \).

where \( x, q ∈ R^n, T ∈ R^{mxn}, Q ∈ R^{n×n} \) is a positive definite matrix, and \( g(z) \) is the indicator function as \( g(z) = 0 \) for \( z ≤ c \) and \( g(z) = +∞ \) otherwise, where \( c ∈ R^n \).

Denote \( f(x) = (1/2)x^T Q x + q^T x \), (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 \( μ = L_f - m_f \). The gradient of the Moreau envelope \( \nabla M_{y_0}(v_i) \) is given by \( max(0, (v_i - c_i)/μ) \).

We generate problem instances in Matlab. We set \( n = m = 10, q = 10 × \text{randn}(n,1), \) and \( Q = EET + F \),
where $E = \text{randn}(n, n)$ and $F = \text{diag}(\exp(\text{randn}(n, 1)))$.

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 $\alpha$ 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: $\alpha_3 = 0.4798$ 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 algorithm 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 $\alpha_0$ in (11) are, we compute rate $\rho$ and $P$ from the matrix inequality condition [1, (16)] using bisection search on $\rho$. For the above ill- and well-conditioned cases, necessary step size bounds $\alpha_0$ are $5.7 \times 10^{-16}$ and $3.6 \times 10^{-4}$, respectively, and our upper bounds are significantly larger.

References

mented Lagrangian method for nonsmooth composite optimization,”


feedback gains via the alternating direction method of multipliers,”

76–91, 2016.

optimization and statistical learning via the alternating direction method of multipliers,” Foundations and Trends in Machine Learning,


for linearly constrained separable convex optimization,” SIAM J. Imag.

primal-dual algorithm for fast distributed nonconvex optimization and
learning over networks,” in International Conference on Machine


momentum methods in optimization,” arXiv preprint arXiv:1611.02635,
2016.

method,” in Proceedings of the 34th International Conference on


of distributed optimization algorithms,” in Communication, Control,
and Computing (Allerton), 55th Annual Allerton Conference on, 2017,
pp. 1206–1212.

algorithm for decentralized consensus optimization,” SIAM Journal on

with network independent step-sizes and separated convergence rates,”


parameter selection for the alternating direction method of multipli-