SIAM J. OPTIM.
Vol. 27, No. 3, pp. 1459–1484

ACCELERATED FIRST-ORDER PRIMAL-DUAL PROXIMAL METHODS FOR LINEARLY CONSTRAINED COMPOSITE CONVEX PROGRAMMING*

YANGYANG XU†

Abstract. Motivated by big data applications, first-order methods have been extremely popular in recent years. However, naive gradient methods generally converge slowly. Hence, much effort has been made to accelerate various first-order methods. This paper proposes two accelerated methods towards solving structured linearly constrained convex programming, for which we assume composite convex objective that is the sum of a differentiable function and a possibly nondifferentiable one. The first method is the accelerated linearized augmented Lagrangian method (LALM). At each update to the primal variable, it allows linearization to the differentiable function and also the augmented term, and thus it enables easy subproblems. Assuming merely convexity, we show that LALM owns $O(1/t)$ convergence if parameters are kept fixed during all the iterations and can be accelerated to $O(1/t^2)$ if the parameters are adapted, where t is the number of total iterations. The second method is the accelerated linearized alternating direction method of multipliers (LADMM). In addition to the composite convexity, it further assumes two-block structure on the objective. Different from classic alternating direction method of multipliers, our method allows linearization to the objective and also augmented term to make the update simple. Assuming strong convexity on one block variable, we show that LADMM also enjoys $O(1/t^2)$ convergence with adaptive parameters. This result is a significant improvement over that in [Goldstein et. al, *SIAM J. Imag. Sci.*, 7 (2014), pp. 1588–1623], which requires strong convexity on both block variables and no linearization to the objective or augmented term. Numerical experiments are performed on quadratic programming, image denoising, and support vector machine. The proposed accelerated methods are compared to nonaccelerated ones and also existing accelerated methods. The results demonstrate the validity of acceleration and superior performance of the proposed methods over existing ones.

Key words. acceleration, linearization, first-order method, augmented Lagrangian method (ALM), alternating direction method of multipliers (ADMM)

AMS subject classifications. 90C06, 90C25, 68W40, 49M27

DOI. 10.1137/16M1082305

1. Introduction.

In recent years, motivated by applications that involve extremely big data, first-order methods with or without splitting techniques have received tremendous attention in a variety of areas such as statistics, machine learning, data mining, and image processing. Compared to traditional methods like Newton's method, first-order methods only require gradient information instead of the much more expensive Hessian. Splitting techniques can further decompose a single difficult large-scale problem into smaller and easier ones. However, in both theory and practice, first-order methods often converge slowly if no additional techniques are applied. For this reason, lots of effort has been made to accelerate various first-order methods.

In this paper, we consider the linearly constrained problem

$$\min_x F(x), \text{ s.t. } Ax = b, \tag{1.1}$$

where F is a proper closed convex but possibly nondifferentiable function. We allow F to be extended-valued, and thus in addition to the linear constraint, (1.1) can also include the constraint $x \in \mathcal{X}$ if part of F is the indicator function of a convex set $\mathcal{X}$.

*Received by the editors June 29, 2016; accepted for publication (in revised form) March 6, 2017; published electronically July 26, 2017.
http://www.siam.org/journals/siopt/27-3/M108230.html
†Department of Mathematics, University of Alabama, Tuscaloosa, AL 35489 (yangyang.xu@ua.edu).

The augmented Lagrangian method (ALM) [2] is one very popular approach to solve constrained optimization problems like (1.1). Let

$$\mathcal{L}_\beta(x,\lambda) = F(x) - \langle \lambda, Ax-b\rangle + \frac{\beta}{2}\|Ax-b\|^2 \tag{1.2}$$

be the augmented Lagrangian function. Then ALM for (1.1) iteratively performs the updates

$$x^{k+1} \in \arg\min_x \mathcal{L}_\beta(x,\lambda^k), \tag{1.3a}$$

$$\lambda^{k+1} = \lambda^k - \beta(Ax^{k+1}-b). \tag{1.3b}$$

In general, the subproblem (1.3a) may not have a solution or have more than one solutions, and even if a unique solution exists, it could be difficult to find the solution. We will assume certain structures of F and also modify the updates in (1.3) to have well-defined and easier subproblems.

1.1. Linearized ALM for linearly constrained composite convex problems. We first assume the composite convexity structure, i.e., the objective in (1.1) can be written as:

$$F(x) = f(x) + g(x), \tag{1.4}$$

where f is a convex Lipschitz differentiable function, and g is a proper closed convex but possibly nondifferentiable function. Hence, the problem (1.1) reduces to the linearly constrained composite convex programming:

$$\min_x f(x) + g(x), \text{ s.t. } Ax = b. \tag{1.5}$$

Usually, g is simple, such as the indicator function of the nonnegative orthant or ℓ_1-norm, but the smooth term f could be complicated, like the logistic loss function.

Our first modification to the update in (1.3a) is to approximate f by a simple function. Typically, we replace f by a quadratic function that dominates f around x^k, resulting in the linearized ALM as follows:

$$x^{k+1} \in \arg\min_x \left\langle \nabla f(x^k) - A^\top \lambda^k, x\right\rangle + g(x) + \frac{\beta}{2}\|Ax-b\|^2 + \frac{1}{2}\|x-x^k\|_P^2, \tag{1.6a}$$

$$\lambda^{k+1} = \lambda^k - \beta(Ax^{k+1}-b), \tag{1.6b}$$

where the weight matrix P is positive semidefinite (PSD) and can be set according to the Lipschitz constant of ∇f. Choosing appropriate P such as $\eta I - \beta A^\top A$, we can also linearize the augmented term and have a closed form solution if g is simple.

The linearization technique here is not new. It is commonly used in the proximal gradient method, which can be regarded as a special case of (1.6) by removing the linear constraint $Ax = b$. It has also been used in the linearized alternating direction method of multipliers (LADMM) [34] and certain primal-dual methods (e.g., [6, 11, 12]).

Our second modification is to adaptively choose the parameters in the linearized ALM and also linearize f at a point other than x^k to accelerate the convergence of the method. Algorithm 1 summarizes the proposed accelerated linearized ALM. The idea of using three point sequences for acceleration was first adopted in [25], and recently it was used in [34] to accelerate the linearized ADMM.

Algorithm 1. Accelerated linearized augmented Lagrangian method for (1.5).

Initialization: choose $\bar{x}^1 = x^1$ and set $\lambda^1 = 0$.

for $k = 1, 2, \ldots$ **do**

Choose parameters $\alpha_k, \beta_k, \gamma_k$, and P^k and perform updates:

(1.7)
$$\hat{x}^k = (1-\alpha_k)\bar{x}^k + \alpha_k x^k,$$

(1.8)
$$x^{k+1} \in \arg\min_x \langle \nabla f(\hat{x}^k) - A^\top \lambda^k, x\rangle + g(x) + \frac{\beta_k}{2}\|Ax-b\|^2 + \frac{1}{2}\|x-x^k\|^2_{P^k},$$

(1.9)
$$\bar{x}^{k+1} = (1-\alpha_k)\bar{x}^k + \alpha_k x^{k+1},$$

(1.10)
$$\lambda^{k+1} = \lambda^k - \gamma_k(Ax^{k+1} - b).$$

if *A stopping condition is satisfied* **then**

Return $(x^{k+1}, \bar{x}^{k+1}, \lambda^{k+1})$.

1.2. Linearized ADMM for two-block structured problems. In this section, we explore more structures of F. In addition to the composite convexity structure, we assume that the variable x and accordingly the matrix A can be partitioned into two blocks, i.e.,

$$x = (y, z), \quad A = (B, C), \tag{1.11}$$

and the objective can be written as

$$F(x) = h(y) + f(z) + g(z), \tag{1.12}$$

where g and h are proper closed convex but possibly nondifferentiable functions, and f is a convex Lipschitz differentiable function. Hence, the problem (1.1) reduces to the linearly constrained two-block structured problem:

$$\min_{y,z} h(y) + f(z) + g(z), \text{ s.t. } By + Cz = b. \tag{1.13}$$

ADMM [10, 14] is a popular method that explores the two-block structure of (1.13) by alternatingly updating y and z, followed by an update to the multiplier λ. More precisely, it iteratively performs the updates:

$$y^{k+1} \in \arg\min_y \mathcal{L}_\beta(y, z^k, \lambda^k), \tag{1.14a}$$

$$z^{k+1} \in \arg\min_z \mathcal{L}_\beta(y^{k+1}, z, \lambda^k), \tag{1.14b}$$

$$\lambda^{k+1} = \lambda^k - \beta(By^{k+1} + Cz^{k+1} - b), \tag{1.14c}$$

where $\mathcal{L}_\beta$ is given in (1.2) with the notation in (1.11) and (1.12). It can be regarded as an inexact ALM, in the sense that it only finds an approximate solution to (1.3a). If (1.14a) and (1.14b) are run repeatedly before updating λ, a solution to (1.3a) would

Algorithm 2. Accelerated linearized alternating direction method of multipliers for (1.13).

Initialization: choose (y^1, z^1) and set $\lambda^1 = 0$.
for $k = 1, 2, \ldots$ **do**
Choose parameters β_k, γ_k, P^k, and Q^k and perform updates:

(1.15a)
$$y^{k+1} \in \arg\min_y h(y) - \langle \lambda^k, By\rangle + \frac{\beta_k}{2}\|By + Cz^k - b\|^2 + \frac{1}{2}\|y - y^k\|^2_{P^k},$$

(1.15b)
$$z^{k+1} \in \arg\min_z \langle \nabla f(z^k) - C^\top \lambda^k, z\rangle + g(z) + \frac{\beta_k}{2}\|By^{k+1} + Cz - b\|^2 + \frac{1}{2}\|z - z^k\|^2_{Q^k},$$

(1.15c)
$$\lambda^{k+1} = \lambda^k - \gamma_k(By^{k+1} + Cz^{k+1} - b).$$

if *A stopping condition is satisfied* **then**
Return $(y^{k+1}, z^{k+1}, \lambda^{k+1})$.

be found, and thus the above update scheme reduces to that in (1.3). However, one single run of (1.14a) and (1.14b), followed by an update to λ, is sufficient to guarantee the convergence. Thus ADMM is often perferable over ALM for solving the two-block structured problem (1.13) since updating y and z separately could be much cheaper than updating them jointly.

Usually g and h are simple, but the smooth term f in (1.13) could be complicated and thus make the z-update in (1.14b) difficult. We apply the same linearization technique as in (1.6a) to (1.14b) and in addition adaptively choose the parameters to accelerate the method. Algorithm 2 summarizes the accelerated linearized ADMM. If g and h are simple, we can have closed form solutions to (1.15a) and (1.15b) by choosing appropriate P^k and Q^k to linearize the augmented terms.

1.3. Related works. It appears that [27] is the first accelerated gradient method for general smooth convex programming. However, according to the Google citations, the work does not really attract much attention until the late 2010's. One possible reason could be that the problems people encountered before were not too large, so second-order methods could handle them very efficiently. Since 2009, accelerated gradient methods have become extremely popular partly due to [1, 30], which generalize the acceleration idea of [27] to composite convex optimization problems, and due to the increasingly large problems arising in many areas. Both [1, 30] achieve optimal rate for first-order methods, but their acceleration techniques look quite different. The former is essentially based on an extrapolation technique while the latter relies on a sequence of estimate functions with adaptive parameters. The work [36] treats several accelerated methods in a unified way, and [35, 38] study accelerated methods from a continuous-time perspective.

Although the methods in [1, 30] can conceptually handle constrained problems, they require simple projection to the constraint set. Hence, they are not really good choices if we consider the structured linearly constrained problem (1.5) or (1.13).

However, the acceleration idea can still be applied. The ALM method in (1.3) is accelerated in [18] by using an extrapolation technique similar to that in [1] to the multiplier λ. While [18] requires the objective to be smooth, [24] extends it to general convex problems, and [23] further reduces the requirement of exactly solving subproblems by assuming strong convexity of the objective. All these accelerated ALM methods do not consider any linearization to the objective or the augmented term. One exception is [21], which linearizes the augmented term and requires strong convexity of the primal problem in its analysis. Therefore, towards finding a solution to (1.5), they may need to solve difficult subproblems if the smooth term f is complicated.

The extrapolation technique in [1] has also been applied to accelerate the ADMM method in [16] for solving two-block structured problems like (1.13). It requires both h and $f+g$ to be strongly convex, and the extrapolation is performed to the multiplier and the secondly updated block variable. In addition, [16] does not consider linearization to the smooth term f or the augmented term, and hence its applicability is restricted. Although the acceleration is observed empirically in [16] for convex problems, no convergence rate has been shown. A later work [22] accelerates the nonlinearized ADMM by renewing the second updated block variable again after extrapolating the multiplier. It still requires strong convexity on both h and $f+g$. Without assuming any strong convexity to the objective function, [15] accelerates ADMM to have $O(1/t^2)$ convergence rate for a special case of (1.13) with $B = I, C = -I$, and $b = 0$, and [34] achieves partial acceleration on linearized ADMM for solving problems in the form of (1.13). It shows in [34] that the decaying rate related to the gradient Lipschitz constant L_f can be $O(1/t^2)$ while the rate for other parts remains $O(1/t)$, where t is the number of iterations. Without the linear constraint, the result in [34] matches the optimal rate of first-order methods.

Different from the extrapolation technique used in the above-mentioned accelerated ALM and ADMM methods, [34] follows [25] and uses three point sequences and adaptive parameters. Algorithm 1 employs the same idea, and our result indicates that the acceleration to the linearized ALM method is not only applied to the gradient Lipschitz constant but also to other parts, i.e., full acceleration. To gain full acceleration to Algorithm 2, we will require either h or $f+g$ to be strongly convex, which is strictly weaker than that assumed in [16]. This assumption is also made in several accelerated primal-dual methods for solving bilinear saddle-point problems, e.g., [3, 4, 5, 19, 28, 29]. The outstanding work [4] presents a framework of primal-dual method for the problem:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \langle Kx, y\rangle + G(x) - F(y), \tag{1.16}$$

where G and F are both proper closed convex functions, and K is a bounded linear operator. It is shown in [4] that the method has $O(1/t^2)$ convergence if either F or G is strongly convex. As shown in [11], the primal-dual method presented in [4] is a special case of linearized ADMM applied to the dual problem of (1.16) about y. Hence, it can fall into one case of Algorithm 2. However, [4] sets parameters in a different way from what we use to accelerate the more general linearized ADMM method; see the example in section 3.2. On solving (1.16), the Douglas–Rachford splitting method has recently been applied and also accelerated in [3] by assuming one of F and G to be strongly convex. In addition, [7] generalizes the work [4] to multi-block structured problems, and the generalized method still enjoys $O(1/t^2)$ convergence if strong convexity is assumed. Without assuming strong convexity, [5]

proposes a new primal-dual method for the saddle-point problem (1.16) and achieves partial acceleration similar to what is achieved in [34].

Acceleration techniques have also been applied to other types of methods to different problems such as in coordinate descent methods (e.g., [9, 26, 40, 41]) and stochastic approximation methods (e.g., [13, 25]). Extending our discussion to these methods will be out of the scope of this paper. Interested readers are referred to those papers we mention here and the references therein.

1.4. Contributions. We summarize our main contributions below.

- We propose an accelerated linearized ALM method for solving linearly constrained composite convex programming. By linearizing the possibly complicated smooth term in the objective, the method enables easy subproblems. Our acceleration strategy follows [34], which considers an accelerated linearized ADMM method. Different from partial acceleration achieved in [34], we obtain full acceleration and achieve the optimal $O(1/t^2)$ convergence rate by assuming merely convexity.
- We also propose an accelerated linearized ADMM method for solving two-block structured linearly constrained convex programming, where in the objective, one block variable has composite convexity structure. While [16] requires strong convexity on both block variables to achieve $O(1/t^2)$ convergence for nonlinearized ADMM, we only need strong convexity on one of them. Furthermore, linearization is allowed to the smooth term in the objective and also to the augmented Lagrangian term, and thus the method enables much easier subproblems than those for nonlinearized ADMM.
- We test the proposed methods on quadratic programming, total variation regularized image denoising problem, and the elastic net regularized support vector machine. We compare them to nonaccelerated methods and also two other accelerated first-order methods. The numerical results demonstrate the validity of acceleration and the superiority of the proposed accelerated methods over other accelerated ones.

1.5. Outline. The rest of the paper is organized as follows. In section 2, we analyze Algorithm 1 and Algorithm 2 with both fixed and adaptive parameters. Numerical experiments are performed in section 3, and finally section 4 concludes the paper and presents some interesting open questions.

2. Convergence analysis. In this section, we analyze the convergence of Algorithms 1 and 2. Assuming merely convexity, we show that Algorithm 1 with adaptive parameters enjoys a fast convergence with rate $O(1/t^2)$, where t is the number of total iterations. For Algorithm 2, we establish the same order of convergence rate by assuming strong convexity on the z-part.

2.1. Notation and preliminary lemmas. Before proceeding with our analysis, let us introduce some notation and preliminary lemmas.

We denote $\mathcal{X}^*$ as the solution set of (1.1). A point x^* is a solution to (1.1) if there exists λ^* such that the KKT conditions hold:

$$0 \in \partial F(x^*) - A^\top \lambda^*, \tag{2.1a}$$

$$Ax^* - b = 0. \tag{2.1b}$$

Together with the convexity of F, the conditions in (2.1) imply that

$$F(x) - F(x^*) - \langle \lambda^*, Ax - b \rangle \geq 0, \ \forall x. \tag{2.2}$$

For any vector v and any symmetric positive semidefinite matrix W of appropriate size, we define $\|v\|_W^2 = v^\top W v$.

LEMMA 2.1. *For any two vectors u, v and a symmetric positive semidefinite matrix W, we have*

$$2u^\top W v = \|u\|_W^2 + \|v\|_W^2 - \|u - v\|_W^2. \tag{2.3}$$

LEMMA 2.2. *Given a function ϕ and a fixed point $\tilde{x}$, if for any λ it holds that*

$$F(\tilde{x}) - F(x^*) - \langle \lambda, A\tilde{x} - b\rangle \le \phi(\lambda), \tag{2.4}$$

then for any $\rho > 0$ we have

$$F(\tilde{x}) - F(x^*) + \rho\|A\tilde{x} - b\| \le \sup_{\|\lambda\| \le \rho} \phi(\lambda). \tag{2.5}$$

This lemma can be found in [11]. Here we provide a simple proof.

Proof. If $A\tilde{x} = b$, then it is trivial to have (2.5) from (2.4). Otherwise, let $\lambda = -\frac{\rho(A\tilde{x}-b)}{\|A\tilde{x}-b\|}$ in both sides of (2.4) and the result follows by noting

$$\phi\left(-\frac{\rho(A\tilde{x} - b)}{\|A\tilde{x} - b\|}\right) \le \sup_{\|\lambda\| \le \rho} \phi(\lambda). \qquad \square$$

LEMMA 2.3. *For any $\epsilon \ge 0$, if*

$$F(\tilde{x}) - F(x^*) + \rho\|A\tilde{x} - b\| \le \epsilon, \tag{2.6}$$

then we have

$$\|A\tilde{x} - b\| \le \frac{\epsilon}{\rho - \|\lambda^*\|} \text{ and } -\frac{\|\lambda^*\|\epsilon}{\rho - \|\lambda^*\|} \le F(\tilde{x}) - F(x^*) \le \epsilon - \rho\|A\tilde{x} - b\| \le \epsilon, \tag{2.7}$$

where (x^*, λ^*) satisfies the KKT conditions in (2.1), and we assume $\|\lambda^*\| < \rho$.

Proof. From (2.2), we have

$$F(\tilde{x}) - F(x^*) \ge -\|\lambda^*\| \cdot \|A\tilde{x} - b\|,$$

which together with (2.6) implies the first inequality in (2.7). The other inequalities follow immediately. $\square$

2.2. Analysis of the accelerated linearized ALM. In this subsection, we show the convergence of Algorithm 1 under the following assumptions.

ASSUMPTION 2.4. *There exists a point (x^*, λ^*) satisfying the KKT conditions in (2.1).*

ASSUMPTION 2.5. *The function f has Lipschitz continuous gradient with constant L_f, i.e.,*

$$\|\nabla f(x) - \nabla f(\tilde{x})\| \le L_f\|x - \tilde{x}\|, \forall x, \tilde{x}. \tag{2.8}$$

The inequality in (2.8) implies that

$$f(\tilde{x}) \le f(x) + \langle \nabla f(x), \tilde{x} - x\rangle + \frac{L_f}{2}\|\tilde{x} - x\|^2, \forall x, \tilde{x}. \tag{2.9}$$

We first establish a result of running one iteration of Algorithm 1. The proof follows that in [34] and is given in Appendix A.1.

LEMMA 2.6 (One-iteration result). *Let $\{(x^k, \bar{x}^k, \lambda^k)\}_{k\geq 1}$ be the sequence generated from Algorithm 1 with $0 \leq \alpha_k \leq 1$, $\forall k$. Then for any (x, λ) such that $Ax = b$, we have*

(2.10)
$$\begin{aligned}
&\left[F(\bar{x}^{k+1}) - F(x) - \langle \lambda, A\bar{x}^{k+1} - b\rangle\right] - (1-\alpha_k)\left[F(\bar{x}^k) - F(x) - \langle \lambda, A\bar{x}^k - b\rangle\right] \\
&\leq -\frac{\alpha_k}{2}\left[\|x^{k+1} - x\|^2_{P^k} - \|x^k - x\|^2_{P^k} + \|x^{k+1} - x^k\|^2_{P^k}\right] + \frac{\alpha_k^2 L_f}{2}\|x^{k+1} - x^k\|^2 \\
&+ \frac{\alpha_k}{2\gamma_k}\left[\|\lambda^k - \lambda\|^2 - \|\lambda^{k+1} - \lambda\|^2 + \|\lambda^{k+1} - \lambda^k\|^2\right] - \frac{\alpha_k\beta_k}{\gamma_k^2}\|\lambda^{k+1} - \lambda^k\|^2,
\end{aligned}$$

where F is given in (1.4).

Below, we specify the values of the parameters $\alpha_k, \beta_k, \gamma_k$, and P^k and establish the convergence rate of Algorithm 1 through (2.10).

2.2.1. Constant parameters. In this subsection, we fix the parameters α_k, β_k, γ_k, and P^k during all the iterations and show $O(1/t)$ convergence of Algorithm 1. The result is summarized in the following theorem. Note that this result is not totally new. Similar results are indicated by several previous works; see [11, 12] for example. However, this special case seems to be overlooked in the literature. In addition, we notice that our result allows more flexible relation between β and γ. Previous works usually assume $\beta = \gamma$ because they consider problems with at least two block variables.

THEOREM 2.7. *Under Assumptions* 2.4 *and* 2.5, *let $\{(x^k, \bar{x}^k, \lambda^k)\}_{k\geq 1}$ be the sequence generated from Algorithm* 1 *with parameters set to*

$$\forall k: \ \alpha_k = 1, \ \beta_k = \beta > 0, \ \gamma_k = \gamma \in (0, 2\beta), \ P^k = P \succ L_f I. \tag{2.11}$$

Then $\bar{x}^k = x^k$, $\forall k$, and $\{(x^k, \lambda^k)\}_{k\geq 1}$ is bounded and converges to a point $(x^\infty, \lambda^\infty)$ that satisfies the KKT conditions in (2.1). *In addition,*

$$|F(\tilde{x}^{t+1}) - F(x^*)| \leq \frac{1}{2t}\left(\|x^1 - x^*\|^2_P + \frac{\max\{(1+\|\lambda^*\|)^2, 4\|\lambda^*\|^2\}}{\gamma}\right), \tag{2.12a}$$

$$\|A\tilde{x}^{t+1} - b\| \leq \frac{1}{2t}\left(\|x^1 - x^*\|^2_P + \frac{\max\{(1+\|\lambda^*\|)^2, 4\|\lambda^*\|^2\}}{\gamma}\right), \tag{2.12b}$$

where (x^*, λ^*) is any point satisfying the KKT conditions in (2.1), *and*

$$\tilde{x}^{t+1} = \frac{\sum_{k=1}^t x^{k+1}}{t}.$$

Remark 2.8. The results in (2.12) imply that the worst error bound becomes smaller as γ grows bigger. As $\gamma \to \infty$, the bound reduces to $\frac{\|x^1 - x^*\|^2_P}{2t}$. However, numerically a large γ will push the residual $\|A\tilde{x}^{t+1} - b\|$ to zero quickly and make the objective converge to optimal value slowly. This can be explained from our analysis below. As $\gamma \geq \rho^3 \to \infty$, it is easy to see that $\|A\tilde{x}^{t+1} - b\|$ approaches to zero from (2.19) and the lower bound of $F(\tilde{x}^{t+1}) - F(x^*)$ also goes to zero from (2.20) while the upper bound of $F(\tilde{x}^{t+1}) - F(x^*)$ is almost $\frac{\|x^1 - x^*\|^2_P}{2t}$. When γ is not too big, $\|A\tilde{x}^{t+1} - b\|$ is not too small for a medium t, and thus from (2.20) we see that the upper bound of $F(\tilde{x}^{t+1}) - F(x^*)$ can be smaller than $\frac{\|x^1 - x^*\|^2_P}{2t}$. The best value of γ depends on (x^*, λ^*). Since the optimal solution is unknown, practically we need to tune γ. This remark also applies to Theorem 2.9 below.

Proof. It is trivial to have $\bar{x}^k = \hat{x}^k = x^k$ from (1.7) and (1.9) as $\alpha_k = 1$, $\forall k$. With the parameters given in (2.11) and $x = x^*$, the inequality in (2.10) reduces to

(2.13)
$$\begin{aligned} &F(x^{k+1}) - F(x^*) - \langle \lambda, Ax^{k+1} - b\rangle \\ &\quad \le -\frac{1}{2}\left[\|x^{k+1} - x^*\|_P^2 - \|x^k - x^*\|_P^2 + \|x^{k+1} - x^k\|_P^2\right] + \frac{L_f}{2}\|x^{k+1} - x^k\|^2 \\ &\qquad + \frac{1}{2\gamma}\left[\|\lambda^k - \lambda\|^2 - \|\lambda^{k+1} - \lambda\|^2 + \|\lambda^{k+1} - \lambda^k\|^2\right] - \frac{\beta}{\gamma^2}\|\lambda^{k+1} - \lambda^k\|^2. \end{aligned}$$

Let $\lambda = \lambda^*$ in the above inequality, and from (2.2), we have

$$\begin{aligned} &\|x^{k+1} - x^*\|_P^2 + \|x^{k+1} - x^k\|_{P - L_f I}^2 + \frac{1}{\gamma}\|\lambda^{k+1} - \lambda^*\|^2 + \frac{1}{\gamma}\left(\frac{2\beta}{\gamma} - 1\right)\|\lambda^{k+1} - \lambda^k\|^2 \\ &\qquad \le \|x^k - x^*\|_P^2 + \frac{1}{\gamma}\|\lambda^k - \lambda^*\|^2. \end{aligned} \tag{2.14}$$

Since $P \succ L_f I$ and $\gamma < 2\beta$, (2.14) implies the nonincreasing monotonicity of $\{\|x^k - x^*\|_P^2 + \frac{1}{\gamma}\|\lambda^k - \lambda^*\|^2\}$, and thus $\{(x^k, \lambda^k)\}_{k\ge 1}$ must be bounded. Summing (2.14) from $k = 1$ to ∞ gives

$$\sum_{k=1}^{\infty}\left(\|x^{k+1} - x^k\|_{P - L_f I}^2 + \frac{1}{\gamma}\left(\frac{2\beta}{\gamma} - 1\right)\|\lambda^{k+1} - \lambda^k\|^2\right) < \infty,$$

and thus

$$\lim_{k\to\infty}(x^{k+1}, \lambda^{k+1}) - (x^k, \lambda^k) = 0. \tag{2.15}$$

Let $(x^\infty, \lambda^\infty)$ be a limit point of $\{(x^k, \lambda^k)\}_{k\ge 1}$ and assume the subsequence $\{(x^k, \lambda^k)\}_{k\in\mathcal{K}}$ converges to it. From $Ax^{k+1} - b = \frac{1}{\gamma}(\lambda^k - \lambda^{k+1}) \to 0$ as $k \to \infty$, we conclude that

$$Ax^\infty - b = 0. \tag{2.16}$$

In addition, letting $\mathcal{K} \ni k \to \infty$ in (1.8) and using (2.15) gives

$$x^\infty = \arg\min_x \langle \nabla f(x^\infty) - A^\top \lambda^\infty, x\rangle + g(x) + \frac{\beta}{2}\|Ax - b\|^2 + \frac{1}{2}\|x - x^\infty\|_P^2,$$

and thus we have the optimality condition

$$0 \in \nabla f(x^\infty) + \partial g(x^\infty) - A^\top \lambda^\infty + \beta A^\top(Ax^\infty - b).$$

Together with (2.16) this implies

$$0 \in \nabla f(x^\infty) + \partial g(x^\infty) - A^\top \lambda^\infty,$$

and thus $(x^\infty, \lambda^\infty)$ satisfies the KKT conditions in (2.1). Hence, (2.14) still holds if (x^*, λ^*) is replaced by $(x^\infty, \lambda^\infty)$, and we have

$$\|x^{k+1} - x^\infty\|_P^2 + \frac{1}{\gamma}\|\lambda^{k+1} - \lambda^\infty\|^2 \le \|x^k - x^\infty\|_P^2 + \frac{1}{\gamma}\|\lambda^k - \lambda^\infty\|^2.$$

Since $(x^\infty, \lambda^\infty)$ is a limit point of $\{(x^k, \lambda^k)\}_{k\geq 1}$, the above inequality implies the convergence of (x^k, λ^k) to $(x^\infty, \lambda^\infty)$.

To prove (2.12), we sum up (2.13) from $k = 1$ through t and note $P \succ L_f I$ and $\gamma < 2\beta$ to have

$$\sum_{k=1}^{t} \left[F(x^{k+1}) - F(x^*) - \langle \lambda, Ax^{k+1} - b\rangle\right] \leq \frac{1}{2}\|x^1 - x^*\|_P^2 + \frac{1}{2\gamma}\|\lambda^1 - \lambda\|^2,$$

which together with the convexity of F implies

$$F(\tilde{x}^{t+1}) - F(x^*) - \langle \lambda, A\tilde{x}^{t+1} - b\rangle \leq \frac{1}{2t}\|x^1 - x^*\|_P^2 + \frac{1}{2\gamma t}\|\lambda^1 - \lambda\|^2. \tag{2.17}$$

Since $\lambda^1 = 0$, we therefore apply Lemma 2.2 to have

$$F(\tilde{x}^{t+1}) - F(x^*) + \rho\|A\tilde{x}^{t+1} - b\| \leq \frac{1}{2t}\|x^1 - x^*\|_P^2 + \frac{\rho^2}{2\gamma t}, \tag{2.18}$$

$\forall\ \rho > 0$. Letting $\rho > \|\lambda^*\|$ and applying Lemma 2.3, we have

$$\|A\tilde{x}^{t+1} - b\| \leq \frac{1}{2t}\frac{\|x^1 - x^*\|_P^2 + \rho^2/\gamma}{\rho - \|\lambda^*\|} \tag{2.19}$$

$$-\frac{1}{2t}\frac{\|\lambda^*\|(\|x^1 - x^*\|_P^2 + \rho^2/\gamma)}{\rho - \|\lambda^*\|} \leq F(\tilde{x}^{t+1}) - F(x^*) \leq \frac{1}{2t}\|x^1 - x^*\|_P^2 + \frac{\rho^2}{2\gamma t} - \rho\|A\tilde{x}^{t+1} - b\|. \tag{2.20}$$

Now let $\rho = \max\{1 + \|\lambda^*\|, 2\|\lambda^*\|\}$ to have (2.12) and thus complete the proof. ∎

2.2.2. Adaptive parameters. In this subsection, we let the parameters α_k, β_k, γ_k, and P^k be adaptive to the iteration number k and improve the previously established $O(1/t)$ convergence rate to $O(1/t^2)$, which is optimal even without the linear constraint.

THEOREM 2.9. *Under Assumptions* 2.4 *and* 2.5*, let* $\{(x^k, \bar{x}^k, \lambda^k)\}_{k\geq 1}$ *be the sequence generated from Algorithm* 1 *with parameters set to*

$$\forall k:\ \alpha_k = \frac{2}{k+1},\ \gamma_k = k\gamma,\ \beta_k \geq \frac{\gamma_k}{2},\ P^k = \frac{\eta}{k}I, \tag{2.21}$$

where $\gamma > 0$ *and* $\eta \geq 2L_f$*. Then*

$$|F(\bar{x}^{t+1}) - F(x^*)| \leq \frac{1}{t(t+1)}\left(\eta\|x^1 - x^*\|^2 + \frac{\max\{(1 + \|\lambda^*\|)^2, 4\|\lambda^*\|^2\}}{\gamma}\right), \tag{2.22a}$$

$$\|A\bar{x}^{t+1} - b\| \leq \frac{1}{t(t+1)}\left(\eta\|x^1 - x^*\|^2 + \frac{\max\{(1 + \|\lambda^*\|)^2, 4\|\lambda^*\|^2\}}{\gamma}\right), \tag{2.22b}$$

where (x^*, λ^*) *is any point satisfying the KKT conditions in* (2.1)*.*

Proof. With the parameters given in (2.21), we multiply $k(k+1)$ to both sides of (2.10) to have

(2.23)

$$\begin{aligned}
&k(k+1)\left[F(\bar{x}^{k+1}) - F(x) - \langle \lambda, A\bar{x}^{k+1} - b\rangle\right] - k(k-1)\left[F(\bar{x}^{k}) - F(x) - \langle \lambda, A\bar{x}^{k} - b\rangle\right] \\
&\quad \le -\eta\left[\|x^{k+1} - x\|^2 - \|x^k - x\|^2 + \|x^{k+1} - x^k\|^2\right] + \frac{2kL_f}{k+1}\|x^{k+1} - x^k\|^2 \\
&\qquad + \frac{1}{\gamma}\left[\|\lambda^k - \lambda\|^2 - \|\lambda^{k+1} - \lambda\|^2 + \|\lambda^{k+1} - \lambda^k\|^2\right] - \frac{2k\beta_k}{\gamma_k^2}\|\lambda^{k+1} - \lambda^k\|^2 \\
&\quad \le -\eta\left[\|x^{k+1} - x\|^2 - \|x^k - x\|^2\right] + \frac{1}{\gamma}\left[\|\lambda^k - \lambda\|^2 - \|\lambda^{k+1} - \lambda\|^2\right].
\end{aligned}$$

Summing (2.23) from $k = 1$ through t, we have

$$t(t+1)\left[F(\bar{x}^{t+1}) - F(x) - \langle \lambda, A\bar{x}^{t+1} - b\rangle\right] \le \eta\|x^1 - x\|^2 + \frac{1}{\gamma}\|\lambda^1 - \lambda\|^2. \tag{2.24}$$

Letting $x = x^*$ in the above inequality, noting $\lambda^1 = 0$, and then applying Lemmas 2.2 and 2.3 with $\rho = \max\{1 + \|\lambda^*\|, 2\|\lambda^*\|\}$, we obtain the desired results by essentially the same arguments as those at the end of the proof of Theorem 2.7. ∎

Remark 2.10. With a positive definite matrix P^k, the subproblem (1.8) becomes strongly convex and thus has a unique solution. One drawback of Theorem 2.9 is that the setting in (2.21) does not allow linearization to the augmented term. The coexistence of the possibly nonsmooth term g and the augmented term $\|Ax - b\|^2$ can still cause difficult subproblems. In that case, we can solve the subproblem inexactly. Theoretically we are unable to prove the $O(1/t^2)$ rate. However, empirically we still observe fast–convergence even subproblems are solved to a medium accuracy; see the experimental results in section 3.1. To linearize the augmented term and retain $O(1/t^2)$ convergence, we need assume strong convexity of the objective; see Theorem 2.14 below.

2.3. Analysis of the accelerated linearized ADMM. In this subsection, we establish the convergence rate of Algorithm 2. In addition to Assumption 2.4, we make the following assumptions about the objective function of (1.13).

ASSUMPTION 2.11. *The function f has Lipschitz continuous gradient with constant L_f, and f and g have strong convexity modulus μ_f and μ_g that satisfy $\mu_f + \mu_g > 0$ (one of them could be zero).*

Note that without strong convexity, an $O(1/t)$ convergence rate can be shown; see [12, 34] for example. Also note that the $O(1/t^2)$ rate has been established in [16, 22] if both h and $f + g$ are strongly convex and no linearization is performed. In addition, linear convergence of ADMM can be shown if $f + g$ is strongly convex and also Lipschitz differentiable and a certain full-rankness assumption is made to B or C; see [8, 31]. Without strong convexity, [20] establishes the linear convergence of ADMM by assuming a certain local error bound and taking a sufficiently small dual stepsize.

Similar to the analysis in the previous subsection, we first establish the result of running one iteration of Algorithm 2, and its proof is provided in Appendix A.2.

LEMMA 2.12 (One-iteration result). *Let $\{(y^k, z^k, \lambda^k)\}_{k\geq 1}$ be the sequence generated from Algorithm* 2. *Then $\forall\ (y, z, \lambda)$ such that $By + Cz = b$, it holds that*

$$\begin{aligned} F(y^{k+1}, z^{k+1}) - F(y, z) - \langle \lambda, By^{k+1} + Cz^{k+1} - b\rangle \\ \leq -\left\langle \frac{1}{\gamma_k}(\lambda^k - \lambda^{k+1}), \lambda - \lambda^k + \frac{\beta_k}{\gamma_k}(\lambda^k - \lambda^{k+1})\right\rangle \\ + \beta_k \left\langle \frac{1}{\gamma_k}(\lambda^k - \lambda^{k+1}) - C(z^{k+1} - z), C(z^{k+1} - z^k)\right\rangle \\ + \frac{L_f}{2}\|z^{k+1} - z^k\|^2 - \frac{\mu_f}{2}\|z^k - z\|^2 - \frac{\mu_g}{2}\|z^{k+1} - z\|^2 \\ - \langle y^{k+1} - y, P^k(y^{k+1} - y^k)\rangle - \langle z^{k+1} - z, Q^k(z^{k+1} - z^k)\rangle, \end{aligned} \tag{2.25}$$

where F is given in (1.12).

When constant parameters are used in Algorithm 2, one can sum up (2.25) from $k = 1$ through t and use (2.3) to show an $O(1/t)$ convergence result. This has already been established in the literature; see [12], for example. Hence, we state the result here without proof, and note that the result does not require any strong convexity of the objective.

THEOREM 2.13. *Assume the existence of $(x^*, \lambda^*) = (y^*, z^*, \lambda^*)$ satisfying* (2.1) *and the gradient Lipschitz continuity of f. Let $\{(y^k, z^k, \lambda^k)\}_{k\geq 1}$ be the sequence generated from Algorithm* 2 *with parameters set to*

$$\beta_k = \gamma_k = \gamma > 0,\ P^k = P \succeq 0,\ Q^k = Q \succeq L_f I,\ \forall k. \tag{2.26}$$

Then

$$\begin{aligned} &\left|F(\tilde{y}^{t+1}, \tilde{z}^{t+1}) - F(y^*, z^*)\right| \\ &\quad\leq \frac{1}{2t}\left(\frac{\max\{(1 + \|\lambda^*\|)^2, 4\|\lambda^*\|^2\}}{\gamma} + \|y^1 - y^*\|_P^2 + \|z^1 - z^*\|_{Q+\gamma C^\top C}^2\right) \\ &\|B\tilde{y}^{t+1} + C\tilde{z}^{t+1} - b\| \\ &\quad\leq \frac{1}{2t}\left(\frac{\max\{(1 + \|\lambda^*\|)^2, 4\|\lambda^*\|^2\}}{\gamma} + \|y^1 - y^*\|_P^2 + \|z^1 - z^*\|_{Q+\gamma C^\top C}^2\right), \end{aligned}$$

where

$$\tilde{y}^{t+1} = \frac{\sum_{k=1}^t y^{k+1}}{t},\ \tilde{z}^{t+1} = \frac{\sum_{k=1}^t z^{k+1}}{t}.$$

Adapting the parameters, we can accelerate the rate to $O(1/t^2)$ as shown below.

THEOREM 2.14. *Under Assumptions* 2.4 *and* 2.11, *let $\{(y^k, z^k, \lambda^k)\}_{k\geq 1}$ be the sequence generated from Algorithm* 2 *with parameters set to*

$$\beta_k = \gamma_k = (k+1)\gamma,\ \forall k \geq 1, \tag{2.27a}$$

$$P^k = \frac{P}{k+1},\ \forall k \geq 1, \tag{2.27b}$$

$$Q^k = (k+1)\big(Q - \gamma C^\top C\big) + L_f I,\ \forall k \geq 1, \tag{2.27c}$$

where $P \succeq 0$ and $\eta\gamma C^\top C \preceq Q \preceq \frac{\mu_f + \mu_g}{2} I$ with $\eta \geq 1$. Let

$$k_0 = \left\lceil 1 + \frac{2(L_f - \mu_f)}{\mu_f + \mu_g}\right\rceil. \tag{2.28}$$

Then we have

$$\|z^k - z^*\|_Q^2 \le \frac{2\phi_1(y^*, z^*, \lambda^*)}{k(k+k_0)}, \quad \|z^k - z^*\|^2 \le \frac{2\phi_1(y^*, z^*, \lambda^*)}{(k+k_0)(L_f + \mu_f + 2\mu_g)}, \tag{2.29}$$

and

$$|F(\tilde{y}^{t+1}, \tilde{z}^{t+1}) - F(y^*, z^*)| \le \frac{2}{t(t+2k_0+3)} \max_{\|\lambda\| \le \rho} \phi_1(y^*, z^*, \lambda) \tag{2.30a}$$

$$\|B\tilde{y}^{t+1} + C\tilde{z}^{t+1} - b\| \le \frac{2}{t(t+2k_0+3)} \max_{\|\lambda\| \le \rho} \phi_1(y^*, z^*, \lambda), \tag{2.30b}$$

where $\rho = \max\{1 + \|\lambda^*\|, 2\|\lambda^*\|\}$,

$$\tilde{y}^{t+1} = \frac{\sum_{k=1}^t (k+k_0+1) y^{k+1}}{\sum_{k=1}^t (k+k_0+1)}, \quad \tilde{z}^{t+1} = \frac{\sum_{k=1}^t (k+k_0+1) z^{k+1}}{\sum_{k=1}^t (k+k_0+1)},$$

and

$$\begin{aligned} \phi_k(y, z, \lambda) = {} & \frac{k+k_0}{2k}\|y^k - y\|_P^2 + \frac{k+k_0}{2}\left(k\|z^k - z\|_Q^2 + (L_f + \mu_g)\|z^k - z\|^2\right) \\ & + \frac{k+k_0}{2\gamma k}\|\lambda - \lambda^k\|^2. \end{aligned} \tag{2.31}$$

In addition, if $P \succ 0$ *and* $\eta > 1$, *then* $\{(y^k, z^k, \lambda^k)\}_{k \ge 1}$ *is bounded, and*

$$\|By^{k+1} + Cz^{k+1} - b\| \le o\left(\frac{1}{k+1}\right), \tag{2.32a}$$

$$|F(y^{k+1}, z^{k+1}) - F(y^*, z^*)| \le O\left(\frac{1}{k+1}\right). \tag{2.32b}$$

Remark 2.15. Note that if Q is a diagonal matrix in (2.27c), then the augmented term in (1.15b) is also linearized. If $h = 0$ and $B = 0$, the problem (1.13) reduces to (1.5). Therefore, Theorem 2.14 implies that we can further linearize the augmented term in the subproblem of the linearized ALM and still obtain $O(1/t^2)$ convergence if the objective is strongly convex.

Also note that taking $P = 0$ and $Q = \gamma C^\top C$ leads to ADMM with adaptive parameters. Hence, we obtain the same order of convergence rate as that in [16] with strictly weaker conditions.

To show this theorem, we first establish a few inequalities.

PROPOSITION 2.16. *Let* k_0 *be defined in* (2.28). *Then* $\forall\ k \ge 1$,

$$(k+k_0)\big(kQ + (L_f + \mu_g)I\big) \succeq (k+k_0+1)\big((k+1)Q + (L_f - \mu_f)I\big). \tag{2.33}$$

Proof. Expanding the left-hand side of the inequality and using $Q \preceq \frac{\mu_f + \mu_g}{2} I$ and (2.28) shows the result. ∎

PROPOSITION 2.17. *Under the assumptions of Theorem* 2.14*, we have*

$$\begin{aligned}
&F(y^{k+1},z^{k+1})-F(y,z)-\langle\lambda,By^{k+1}+Cz^{k+1}-b\rangle\\
&\quad\le-\frac{1}{2\gamma(k+1)}\left[\|\lambda-\lambda^{k+1}\|^2-\|\lambda-\lambda^k\|^2\right]-\frac{\eta-1}{2\eta\gamma(k+1)}\|\lambda^k-\lambda^{k+1}\|^2\\
&\qquad-\frac{1}{2(k+1)}\left[\|y^{k+1}-y\|_P^2-\|y^k-y\|_P^2+\|y^{k+1}-y^k\|_P^2\right]\\
&\qquad-\frac{1}{2}\left((k+1)\|z^{k+1}-z\|_Q^2+(L_f+\mu_g)\|z^{k+1}-z\|^2\right)\\
&\qquad+\frac{1}{2}\left((k+1)\|z^k-z\|_Q^2+(L_f-\mu_f)\|z^k-z\|^2\right).
\end{aligned}\tag{2.34}$$

Proof. Since $\beta_k=\gamma_k$, we use (2.3) and have from (2.25) that

$$\begin{aligned}
&F(y^{k+1},z^{k+1})-F(y,z)-\langle\lambda,By^{k+1}+Cz^{k+1}-b\rangle\\
&\quad\le-\frac{1}{2\gamma_k}\left[\|\lambda^k-\lambda^{k+1}\|^2+\|\lambda-\lambda^{k+1}\|^2-\|\lambda-\lambda^k\|^2\right]\\
&\qquad+\left\langle\lambda^k-\lambda^{k+1},C(z^{k+1}-z^k)\right\rangle\\
&\qquad-\frac{\gamma_k}{2}\left[\|C(z^{k+1}-z)\|^2-\|C(z^k-z)\|^2+\|C(z^{k+1}-z^k)\|^2\right]\\
&\qquad+\frac{L_f}{2}\|z^{k+1}-z^k\|^2-\frac{\mu_f}{2}\|z^k-z\|^2\\
&\qquad-\frac{\mu_g}{2}\|z^{k+1}-z\|^2-\frac{1}{2}\left[\|y^{k+1}-y\|_{P^k}^2-\|y^k-y\|_{P^k}^2+\|y^{k+1}-y^k\|_{P^k}^2\right]\\
&\qquad-\frac{1}{2}\left[\|z^{k+1}-z\|_{Q^k}^2-\|z^k-z\|_{Q^k}^2+\|z^{k+1}-z^k\|_{Q^k}^2\right].
\end{aligned}\tag{2.35}$$

Note that from the parameter setting, we have

$$\begin{aligned}
\left\langle\lambda^k-\lambda^{k+1},C(z^{k+1}-z^k)\right\rangle-\frac{\gamma_k}{2}\|C(z^{k+1}-z^k)\|^2+\frac{L_f}{2}\|z^{k+1}-z^k\|^2-\frac{1}{2}\|z^{k+1}-z^k\|_{Q^k}^2&\\
=\left\langle\lambda^k-\lambda^{k+1},C(z^{k+1}-z^k)\right\rangle-\frac{k+1}{2}\|z^{k+1}-z^k\|_Q^2&\\
\le\left\langle\lambda^k-\lambda^{k+1},C(z^{k+1}-z^k)\right\rangle-\frac{\eta\gamma_k}{2}\|z^{k+1}-z^k\|_{C^\top C}^2&\\
\le\frac{1}{2\eta\gamma_k}\|\lambda^k-\lambda^{k+1}\|^2.&
\end{aligned}\tag{2.36}$$

Plugging (2.36) and the parameters in (2.27) into (2.35) gives (2.34). ∎

Now we are ready to show Theorem 2.14.

Proof of Theorem 2.14. Letting $(y,z)=(y^*,z^*)$ in (2.34) and rearranging terms gives

$$\begin{aligned}
&\left[F(y^{k+1},z^{k+1})-F(y^*,z^*)-\langle\lambda,By^{k+1}+Cz^{k+1}-b\rangle\right]+\frac{1}{2(k+1)}\|y^*-y^{k+1}\|_P^2\\
&\quad+\frac{1}{2}\left((k+1)\|z^{k+1}-z^*\|_Q^2+(L_f+\mu_g)\|z^{k+1}-z^*\|^2\right)+\frac{1}{2\gamma(k+1)}\|\lambda-\lambda^{k+1}\|^2\\
&\le\frac{1}{2(k+1)}\|y^k-y^*\|_P^2+\frac{1}{2}\left((k+1)\|z^k-z^*\|_Q^2+(L_f-\mu_f)\|z^k-z^*\|^2\right)\\
&\quad+\frac{1}{2\gamma(k+1)}\|\lambda-\lambda^k\|^2-\frac{\eta-1}{2\eta\gamma(k+1)}\|\lambda^k-\lambda^{k+1}\|^2.
\end{aligned}\tag{2.37}$$

Multiplying $k+k_0+1$ to both sides of the above inequality and using notation ϕ_k defined in (2.31), we have

$$\begin{aligned}
&(k+k_0+1)\left[F(y^{k+1},z^{k+1})-F(y^*,z^*)-\langle\lambda,By^{k+1}+Cz^{k+1}-b\rangle\right]+\phi_{k+1}(y^*,z^*,\lambda)\\
&\quad\le\frac{k+k_0+1}{2(k+1)}\|y^k-y^*\|_P^2+\frac{k+k_0+1}{2}\left((k+1)\|z^k-z^*\|_Q^2\right.\\
&\qquad\left.+(L_f-\mu_f)\|z^k-z^*\|^2\right)+\frac{k+k_0+1}{2\gamma(k+1)}\left(\|\lambda-\lambda^k\|^2-\frac{\eta-1}{\eta}\|\lambda^k-\lambda^{k+1}\|^2\right)\\
&\quad\le\frac{k+k_0}{2k}\|y^k-y^*\|_P^2+\frac{k+k_0}{2}\left(k\|z^k-z^*\|_Q^2+(L_f+\mu_g)\|z^k-z^*\|^2\right)\\
&\qquad+\frac{k+k_0}{2\gamma k}\|\lambda-\lambda^k\|^2-\frac{k+k_0+1}{2\gamma(k+1)}\frac{\eta-1}{\eta}\|\lambda^k-\lambda^{k+1}\|^2\\
&\quad=\phi_k(y^*,z^*,\lambda)-\frac{k+k_0+1}{2\gamma(k+1)}\frac{\eta-1}{\eta}\|\lambda^k-\lambda^{k+1}\|^2, \qquad (2.38)
\end{aligned}$$

where in the second inequality, we have used (2.33) and the decreasing monotonicity of $\frac{k+k_0+1}{k+1}$ with respect to k.

Letting $\lambda=\lambda^*$ in (2.38) and using (2.2), we have

$$\phi_{k+1}(y^*,z^*,\lambda^*)\le\phi_k(y^*,z^*,\lambda^*). \qquad (2.39)$$

In addition, note that

$$\begin{aligned}
&F(y^{k+1},z^{k+1})-F(y^*,z^*)-\langle\lambda^*,By^{k+1}+Cz^{k+1}-b\rangle\\
&\quad=F(y^{k+1},z^{k+1})-F(y^*,z^*)-\langle\lambda^*,B(y^{k+1}-y^*)+C(z^{k+1}-z^*)\rangle\\
&\quad=h(y^{k+1})-h(y^*)-\langle B^\top\lambda^*,y^{k+1}-y^*\rangle+(f+g)(z^{k+1})-(f+g)(z^*)\\
&\qquad-\langle C^\top\lambda^*,z^{k+1}-z^*\rangle\\
&\quad\ge\frac{\mu_f+\mu_g}{2}\|z^{k+1}-z^*\|^2,
\end{aligned}$$

where the inequality is from the convexity of h and $f+g$ and also the KKT conditions in (2.1). Hence, from (2.38) and (2.39), it follows that

$$\frac{(\mu_f+\mu_g)(k+k_0+1)}{2}\|z^{k+1}-z^*\|^2+\phi_{k+1}(y^*,z^*,\lambda^*)\le\phi_1(y^*,z^*,\lambda^*),$$

and thus we obtain the results in (2.29). If $P\succ 0$, the above inequality indicates the boundedness of $\{(x^k,y^k,\lambda^k)\}$.

Again, letting $\lambda=\lambda^*$ in (2.38) and summing it from $k=1$ through t, we conclude from (2.2) and (2.39) that

$$\sum_{k=1}^{t}\frac{k+k_0+1}{2\gamma(k+1)}\frac{\eta-1}{\eta}\|\lambda^k-\lambda^{k+1}\|^2\le\phi_1(y^*,z^*,\lambda^*),$$

and thus letting $t\to\infty$, we have $\lambda^k-\lambda^{k+1}\to 0$ from the above inequality as $\eta>1$, and thus (2.32a) follows from the update rule (1.15c). Furthermore, from the boundedness of $\{(y^k,z^k,\lambda^k)\}$, we let $\lambda=0$ in (2.37) to have $F(y^{k+1},z^{k+1})-F(y^*,z^*)\le O(\frac{1}{k+1})$. Using (2.2) and (2.32a), we have $F(y^{k+1},z^{k+1})-F(y^*,z^*)\ge -O(\frac{1}{k+1})$, and thus (2.32b) follows.

Finally, summing (2.38) from $k = 1$ through t and noting $\phi_k \geq 0, \forall k$, we have

$$\sum_{k=1}^{t}(k + k_0 + 1)\left[F(y^{k+1}, z^{k+1}) - F(y^*, z^*) - \langle \lambda, By^{k+1} + Cz^{k+1} - b\rangle\right] \leq \phi_1(y^*, z^*, \lambda).$$

Then by the convexity of F, we have from the above inequality that for any λ,

$$F(\tilde{y}^{t+1}, \tilde{z}^{t+1}) - F(y^*, z^*) - \langle \lambda, B\tilde{y}^{t+1} + C\tilde{z}^{t+1} - b\rangle \leq \frac{\phi_1(y^*, z^*, \lambda)}{\sum_{k=1}^{t}(k + k_0 + 1)}.$$

By Lemmas 2.2 and 2.3 with $\rho = \max\{1 + \|\lambda^*\|, 2\|\lambda^*\|\}$ and the initialization $\lambda^1 = 0$, the above result implies the desired results in (2.30). This completes the proof. ∎

3. Numerical results. In this section, we test the proposed accelerated methods on solving three problems: quadratic programming, total variation regularized image denoising, and elastic net regularized support vector machine. We compare them to nonaccelerated methods and also existing accelerated methods to demonstrate their efficiency.

3.1. Quadratic programming. In this subsection, we test Algorithm 1 on quadratic programming. First, we compare the algorithm with fixed and adaptive parameters, i.e., nonaccelerated ALM and accelerated ALM, on equality constrained quadratic programming (ECQP):

$$\min_x \ F(x) = \frac{1}{2}x^\top Qx + c^\top x, \ \text{s.t.} \ Ax = b. \tag{3.1}$$

Note that ECQP can be solved in a direct way by solving a linear equation (c.f., [32, sect. 16.1]), so ALM may not be the best choice for (3.1). Our purpose of using this simple example is to validate acceleration.

We set the problem size to $m = 20, n = 500$ and generate $A \in \mathbb{R}^{m\times n}, b, c$ and $Q \in \mathbb{R}^{n\times n}$ according to standard Gaussian distribution, where Q is made to be a positive definite matrix. We set the parameters of Algorithm 1 to $\alpha_k = 1, \beta_k = \gamma_k = m$, and $P^k = \|Q\|_2 I, \forall k$ for the nonaccelerated ALM, and $\alpha_k = \frac{2}{k+1}, \beta_k = \gamma_k = mk$, and $P^k = \frac{2\|Q\|_2}{k} I, \forall k$ for the accelerated ALM. Figure 1 plots the objective distance to the optimal value $|F(x) - F(x^*)|$ and the violation of feasibility $\|Ax - b\|$ given by the two methods. We can see that Algorithm 1 with adaptive parameters performs significantly better than with fixed parameters, in both objective and feasibility measures.

Secondly, we test the accelerated linearized augmented Lagrangian method (ALALM) on the nonnegative linearly constrained quadratic programming, which

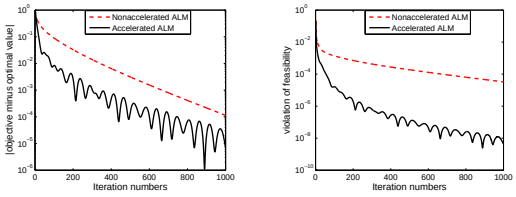

FIG. 1. *Results by the nonaccelerated ALM (Algorithm* 1 *with fixed parameters) and the accelerated ALM (Algorithm* 1 *with adaptive parameters) on solving* (3.1)*. Left: the distance of the objective value to the optimal value* $|F(x) - F(x^*)|$*; right: the violation of feasibility* $\|Ax - b\|$.

is formulated as follows:

$$\min_x \ F(x) = \frac{1}{2}x^\top Qx + c^\top x, \ \text{s.t.} \ Ax = b, \ x \geq 0. \tag{3.2}$$

In the test, we set the problem size to $m = 50$ and $n = 1000$. We let $Q = HH^\top$, where $H \in \mathbb{R}^{n\times(n-100)}$ is generated according to standard Gaussian distribution. Hence, the objective is not strongly convex. The elements of b and c follow identically independent uniform distribution and standard Gaussian distribution, respectively. Thus, $b \geq 0$. The matrix $A \in \mathbb{R}^{m\times n}$ has the form of $[B, I]$ to ensure the feasibility of the problem. We generate B according to both Gaussian and uniform distribution. Note that the uniformly distributed B leads to a more difficult problem.

We set the parameters of Algorithm 1 according to (2.21) with $\gamma = m$, $\eta = 2\|Q\|_2$, and $\beta_k = \gamma_k$, $\forall k$. The most difficult step in Algorithm 1 is (1.8), which does not have a closed form solution with the above setting. We solve the subproblem by the interior-point method to a tolerance `subtol`. Since A only has 50 rows, each step of the interior-point method only needs to solve a 50×50 equation and do some componentwise multiplication. We notice that ALALM converges fast in the beginning but slows down as it approaches the solution. Hence, we also test to restart it after a fixed number of iterations, and in this test, we simply restart it every 50 iterations.

We compare ALALM to FISTA [1], which also has $O(1/t^2)$ convergence rate. At each iteration, FISTA requires a projection to the constraint set of (3.2), and we solve it also by the interior-point method to the tolerance `subtol`. Again, each step of the interior-point method only needs to solve a 50×50 equation and do some componentwise multiplication. We also test restarted FISTA by restarting it every 50 iterations. Note that a restarted FISTA is proposed in [33] by checking the monotonicity of the objective value or gradient norm. However, since subproblems are solved inaccurately, the restart scheme in [33] does not work here.

Figure 2 plots the results corresponding to Gaussian randomly generated matrix B and Figure 3 corresponding to uniformly random B, where the optimal value $F(x^*)$ is obtained by Matlab function `quadprog` with tolerance 10^{-16}. In both figures, `subtol` varies among $\{10^{-6}, 10^{-8}, 10^{-10}\}$. From the figures, we see that both FISTA and ALALM perform better when restarted periodically, and ALALM performs more stably than FISTA to different `subtol`. Even if the subproblems are solved inaccurately only to the tolerance 10^{-6}, the restarted ALALM can still reach almost machine accuracy. However, FISTA can reach an accurate solution only if the subproblems are solved to a high accuracy such as `subtol` $= 10^{-10}$ and B is Gaussian randomly generated.

3.2. Image denoising. In this subsection, we test the accelerated ADMM, i.e., Algorithm 2, on the total variation regularized image denoising problem:

$$\min_X F(X) = \frac{1}{2}\|X - M\|_F^2 + \mu\|\mathcal{D}X\|_1, \tag{3.3}$$

where M is a noisy two-dimensional image, $\mathcal{D}$ is a finite difference operator, and $\|Y\|_1 = \sum_{i,j} |Y_{ij}|$. Replacing $\mathcal{D}X$ by Y, we can write (3.3) equivalently to

$$\min_{X,Y} G(X, Y) = \frac{1}{2}\|X - M\|_F^2 + \mu\|Y\|_1, \ \text{s.t.} \ \mathcal{D}X = Y. \tag{3.4}$$

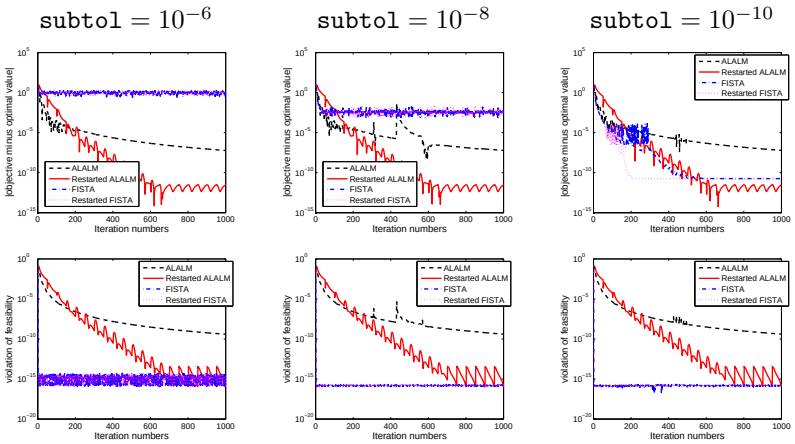

FIG. 2. *Results by FISTA* [1] *and ALALM (Algorithm* 1 *with adaptive parameters) on solving* (3.2) *where* $A = [B, I]$ *and* B *is generated according to standard Gaussian distribution. Subproblems for both methods are solved to a tolerance specified by* `subtol`*. First row: the absolute value of objective value minus the optimal value* $|F(x) - F(x^*)|$*; second row: the violation of feasibility* $\|Ax - b\|$.

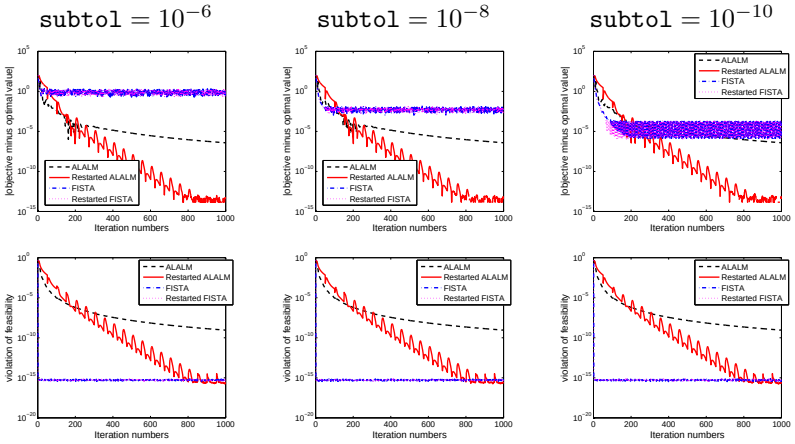

FIG. 3. *Results by FISTA* [1] *and ALALM on solving* (3.2) *where* $A = [B, I]$ *and* B *is generated according to uniform distribution. Subproblems for both methods are solved to a tolerance specified by* `subtol`*. First row: the absolute value of objective value minus the optimal value* $|F(x) - F(x^*)|$*; second row: the violation of feasibility* $\|Ax - b\|$.

Applying Algorithm 2 to (3.4) gives the updates:

(3.5a)
$$Y^{k+1} = \arg\min_Y \mu\|Y\|_1 + \langle \Lambda^k, Y\rangle + \frac{\beta_k}{2}\|Y - \mathcal{D}X\|_F^2 + \frac{1}{2}\|Y - Y^k\|_{P^k}^2,$$

(3.5b)
$$X^{k+1} = \arg\min_X \frac{1}{2}\|X - M\|_F^2 - \langle \Lambda^k, \mathcal{D}X\rangle + \frac{\beta_k}{2}\|Y - \mathcal{D}X\|_F^2 + \frac{1}{2}\|X - X^k\|_{Q^k}^2,$$

(3.5c)
$$\Lambda^{k+1} = \Lambda^k - \gamma_k(\mathcal{D}X^{k+1} - Y^{k+1}).$$

We test the algorithm with four sets of parameters, leading to four different methods listed below:

- nonaccelerated ADMM: $\beta_k = \gamma_k = 10$, $P^k = 0$, $Q^k = 0$, $\forall k$;
- accelerated ADMM: $\beta_k = \gamma_k = \frac{k+1}{2\|\mathcal{D}\|_2^2}$, $P^k = 0$, $Q^k = 0$, $\forall k$;
- nonacclerated linearized ADMM: $\beta_k = \gamma_k = \frac{1}{2\|\mathcal{D}\|_2^2}$, $P^k = 0$, $Q^k = \frac{I}{2} - \frac{\mathcal{D}^\top\mathcal{D}}{2\|\mathcal{D}\|_2^2}$, $\forall k$;
- accelerated linearized ADMM: $\beta_k = \gamma_k = \frac{k+1}{20\|\mathcal{D}\|_2^2}$, $P^k = 0$, $Q^k = \frac{(k+1)I}{20} - \frac{(k+1)\mathcal{D}^\top\mathcal{D}}{20\|\mathcal{D}\|_2^2}$, $\forall k$.

With $P^k = 0$, the solution of (3.5a) can be written analyticly by using the soft thresholding or shrinkage. We assume periodic boundary condition, and thus with $Q^k = 0$, the solution of (3.5b) can be easily obtained by solving a linear system that involves one two-dimensional fast Fourier transform (FFT2), one inverse FFT2, and some componentwise division [37]. For the linearized ADMM, it is easy to write closed form solutions for both X and Y subproblems. We compare Algorithm 2 with the above four settings to the accelerated primal-dual method in [4], which we call the Chambolle–Pock method by using the authors' name. As shown in [11], Chambolle–Pock method is equivalent to linearized ADMM applied to the dual reformulation of (3.3). It iteratively performs the updates:

$$Z^{k+1} = \operatorname*{arg\,min}_{|Z_{ij}|\le 1,\forall i,j} \|Z - Z^k - \sigma_k \mathcal{D}\bar{X}^k\|_F^2, \tag{3.6a}$$

$$X^{k+1} = \arg\min_X \frac{\tau_k}{2\mu}\|X - X^k\|_F^2 + \frac{1}{2}\|X - X^k + \tau_k\mathcal{D}^* Z^{k+1}\|_F^2, \tag{3.6b}$$

$$\bar{X}^{k+1} = X^{k+1} + \theta_k(X^{k+1} - X^k) \tag{3.6c}$$

with $\bar{X}^1 = X^1$, $\tau_1\sigma_1\|\mathcal{D}\|_2^2 \le 1$, and the parameters set to

$$\theta_k = \frac{1}{\sqrt{1+2\gamma\tau_k}},\ \tau_{k+1} = \theta_k\tau_k,\ \sigma_{k+1} = \frac{\sigma_k}{\theta_k}\ \forall k.$$

We set $\tau_1 = \sigma_1 = 1/\|\mathcal{D}\|_2$ and $\gamma = 0.35/\mu$ as suggested in [4].

In this test, we use the cameraman image shown in Figure 4, and we add 10% Gaussian noise. The regularization parameter is set to $\mu = 0.04$. For Algorithm 2, we report the objective value of (3.4) and the violation of feasibility and also the objective value of (3.3), and for the Chambolle–Pock method we only report the objective value of (3.3) since it solves the dual problem and does not guarantee the feasibility of (3.4). Figure 5 plots the results in terms of iteration numbers, where the optimal objective value is obtained by running ADMM to 50,000 iterations. Since the linearized ADMM and Chambolle–Pock methods have lower iteration complexity than the nonlinearized ADMM, we also plot the results in terms of running time. From the figure, we see that Algorithm 2 with adaptive parameters performs significantly better than that with fixed parameters. The Chambolle–Pock method decreases the objective fastest in the beginning, and later the accelerated ADMM with or without linearization catches up and surpasses it.

3.3. Elastic net regularized support vector machine. We test Algorithm 2 on the elastic net regularized support vector machine problem

$$\min_x F(x) = \frac{1}{m}\sum_{i=1}^m [1 - b_i a_i^\top x]_+ + \mu_1\|x\|_1 + \frac{\mu_2}{2}\|x\|^2, \tag{3.7}$$

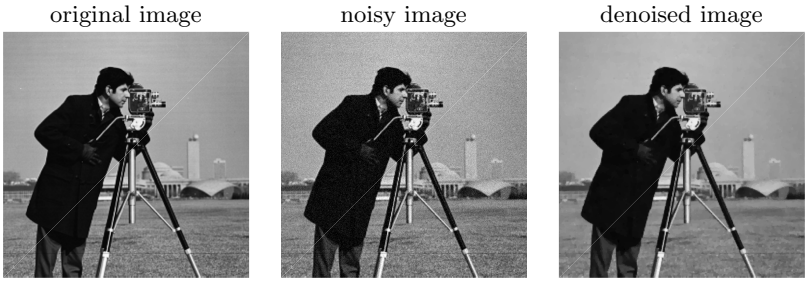

FIG. 4. *The cameraman images. Left: original one; middle: noisy image with* 10% *Gaussian noise, PSNR* = 25.62*; right: denoised image by the accelerated ADMM running to* 200 *iterations, PSNR* = 33.29.

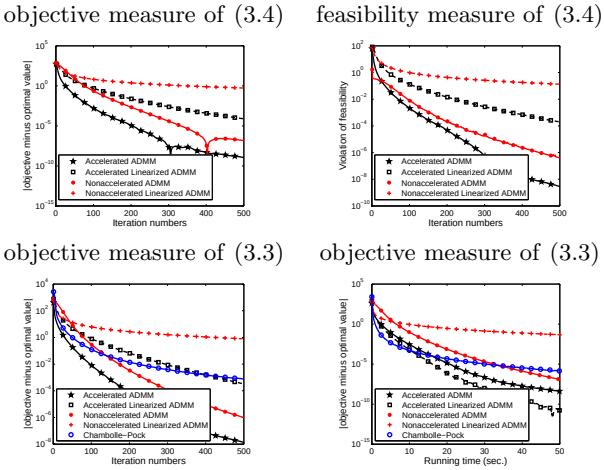

FIG. 5. *Results by Algorithm* 2 *with adaptive parameters (accelerated ADMM) and constant parameters (nonaccelerated ADMM) and also the Chambolle–Pock method on solving* (3.3)*. Top left: the absolute value of objective of* (3.4) *minus optimal value* $|G(X,Y) - G(X^*,Y^*)|$*; top right: the violation of feasibility of* (3.4) $\|\mathcal{D}X - Y\|_F$*; bottom left: the absolute value of objective of* (3.3) *minus optimal value* $|F(X) - F(X^*)|$ *in terms of iteration; bottom right: the absolute value of objective of* (3.3) *minus optimal value* $|F(X) - F(X^*)|$ *in terms of running time.*

where $[c]_+ = \max(0, c)$, $\{(a_i, b_i)\}_{i=1}^m$ are the samples in p-dimensional space, and $b_i \in \{+1, -1\}$ is the label of the ith sample. Let $A = [a_1, \ldots, a_m] \in \mathbb{R}^{p\times m}$ and replace $1 - b_i a_i^\top x$ by $y_i \ \forall\ i$. We obtain the equivalent formulation:

$$\min_x G(x,y) = \frac{1}{m} e^\top [y]_+ + \mu_1 \|x\|_1 + \frac{\mu_2}{2}\|x\|^2, \text{ s.t. } Bx + y = e, \tag{3.8}$$

where e is the vector with all ones, and $B = \mathrm{Diag}(b)A$.

The data are generated in the same way as in [39]. One half of the samples belong to "+1" class and the other to "–1" class. Each sample in "+1" class is generated according to Gaussian distribution $\mathcal{N}(u, \Sigma)$, and each sample in "–1" class follows $\mathcal{N}(-u, \Sigma)$. The mean vector and variance matrix are set to

$$u = \begin{bmatrix} E_{s\times 1} \\ 0_{(p-s)\times 1} \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \rho E_{s\times s} + \rho I_{s\times s} & 0_{s\times(p-s)} \\ 0_{(p-s)\times s} & I_{(p-s)\times(p-s)} \end{bmatrix},$$

where $E_{s\times s}$ is an $s\times s$ matrix with all ones, s is the number of features that are related to classification, and $\rho\in[0,1]$ measures the correlation of the features (the larger it is, the harder the problem is). In the test, we set $m=100$, $p=500$, $s=50,\rho=0.5$, and $\mu_1=\mu_2=0.01$.

Applying Algorithm 2 to (3.8), we iteratively perform the updates:

(3.9a)
$$y^{k+1}=\arg\min_y \frac{1}{m}e^\top[y]_+-\langle\lambda^k,y\rangle+\frac{\beta_k}{2}\|Bx^k+y-e\|^2+\frac{1}{2}\|y-y^k\|^2_{P^k},$$

(3.9b)
$$x^{k+1}=\arg\min_x \mu_1\|x\|_1+\frac{\mu_2}{2}\|x\|^2-\langle\lambda^k,Bx\rangle+\frac{\beta_k}{2}\|Bx+y^{k+1}-e\|^2+\frac{1}{2}\|x-x^k\|^2_{Q^k},$$

(3.9c)
$$\lambda^{k+1}=\lambda^k-\gamma_k(Bx^{k+1}+y^{k+1}-e).$$

Again, we test two sets of parameters. The first one fixes the parameters during all iterations, and the second one adapts the parameters. Since the coexistence of ℓ_1-norm and the least squares term makes (3.9b) difficult to solve, we choose Q^k to cancel the term $x^\top B^\top Bx$, i.e., we linearize the augmented term. Specifically, we set the parameters in the same way as the previous test:

- nonaccelerated linearized ADMM: $\beta_k=\gamma_k=\frac{1}{2\|B\|_2^2}$, $P^k=0$, $Q^k=\frac{I}{2}-\frac{B^\top B}{2\|B\|_2^2}$, $\forall k$;
- accelerated linearized ADMM: $\beta_k=\gamma_k=\frac{\mu_2(k+1)}{20\|B\|_2^2}$, $P^k=0$, $Q^k=\frac{\mu_2(k+1)I}{20}-\frac{\mu_2(k+1)B^\top B}{20\|B\|_2^2}$, $\forall k$.

We also compare the linearized ADMM to the classic ADMM without linearization, which introduces another variable z to split x from the ℓ_1-norm and solves the problem

$$\min_x \frac{1}{m}e^\top[y]_+ +\mu_1\|z\|_1+\frac{\mu_2}{2}\|x\|^2,\ \text{s.t. } Bx+y=e,\ x=z. \tag{3.10}$$

We use the code from [42] to solve (3.10) and tune its parameters as best as we can.

Similar to the previous test, we measure the objective value and feasibility of (3.8) given by the linearized ADMM and the objective value of (3.7) for all three methods. Figure 6 plots the results, where the optimal objective value is obtained by CVX [17] with high precision. From the figure, we see that the accelerated linearized ADMM

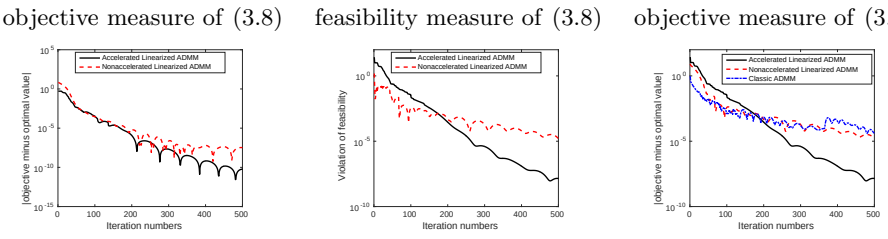

FIG. 6. *Results by Algorithm* 2 *with adaptive parameters (accelerated linearized ADMM) and constant parameters (nonaccelerated linearized ADMM) and also the classic nonlinearized ADMM on solving* (3.7). *Left: the absolute value of objective of* (3.8) *minus optimal value* $|G(x,y)-G(x^*,y^*)|$*; middle: the violation of feasibility of* (3.8) $\|Bx+y-e\|$*; right: the absolute value of objective of* (3.7) *minus optimal value* $|F(x)-F(x^*)|$.

performs significantly better than the nonaccelerated counterpart, and the latter is comparable to the classic nonlinearized ADMM.

4. Conclusions. We have proposed an accelerated linearized augmented Lagrangian method (ALALM) and also an accelerated alternating direction method of multipliers (ALADMM) for solving structured linearly constrained convex programming. We have established $O(1/t^2)$ convergence rate for an ALALM by assuming merely convexity and for an ALADMM by assuming strong convexity to one block variable. Numerical experiments have been performed to demonstrate the validness of acceleration and higher efficiency over existing accelerated methods.

To have the $O(1/t^2)$ convergence rate for the ALALM, our current analysis does not allow linearization to the augmented term, and that may cause great difficulty on solving subproblems if meanwhile we have a complicated nonsmooth term. It is interesting to know whether we can linearize the augmented term and still obtain $O(1/t^2)$ convergence under the same assumptions. We are unable to show this under the setting of Algorithm 1, so it may have to turn to other acceleration techniques. We leave this open question to interested readers.

Appendix A. Technical details of two key lemmas. In this section, we provide detailed proofs of two key lemmas.

A.1. Proof of Lemma 2.6. From (2.9), it follows that

$$f(\bar{x}^{k+1}) \le f(\hat{x}^k) + \langle \nabla f(\hat{x}^k), \bar{x}^{k+1} - \hat{x}^k \rangle + \frac{L_f}{2}\|\bar{x}^{k+1} - \hat{x}^k\|^2.$$

Substituting $\bar{x}^{k+1} = (1-\alpha_k)\bar{x}^k + \alpha_k x^{k+1}$ and also noting $\bar{x}^{k+1} - \hat{x}^k = \alpha_k(x^{k+1} - x^k)$, we have from the above inequality that

$$\begin{aligned} f(\bar{x}^{k+1}) \;&\le\; f(\hat{x}^k) + (1-\alpha_k)\langle \nabla f(\hat{x}^k), \bar{x}^k - \hat{x}^k\rangle + \alpha_k \langle \nabla f(\hat{x}^k), x^{k+1} - \hat{x}^k\rangle \\ &\quad + \frac{\alpha_k^2 L_f}{2}\|x^{k+1} - x^k\|^2 \\ &= (1-\alpha_k)\left[f(\hat{x}^k) + \langle \nabla f(\hat{x}^k), \bar{x}^k - \hat{x}^k\rangle\right] + \alpha_k\left[f(\hat{x}^k) + \langle \nabla f(\hat{x}^k), x - \hat{x}^k\rangle\right] \\ &\quad + \alpha_k \langle \nabla f(\hat{x}^k), x^{k+1} - x\rangle + \frac{\alpha_k^2 L_f}{2}\|x^{k+1} - x^k\|^2 \\ &\le (1-\alpha_k) f(\bar{x}^k) + \alpha_k f(x) + \alpha_k \langle \nabla f(\hat{x}^k), x^{k+1} - x\rangle + \frac{\alpha_k^2 L_f}{2}\|x^{k+1} - x^k\|^2, \end{aligned} \tag{A.1}$$

where the second inequality follows from the convexity of f. Hence,

$$\begin{aligned} &\left[F(\bar{x}^{k+1}) - F(x) - \langle \lambda, A\bar{x}^{k+1} - b\rangle\right] - (1-\alpha_k)\left[F(\bar{x}^k) - F(x) - \langle \lambda, A\bar{x}^k - b\rangle\right] \\ &\quad = \left[f(\bar{x}^k{+}1){-}(1{-}\alpha_k)f(\bar{x}^k){-}\alpha_k f(x)\right] + \left[g(\bar{x}^{k+1}){-}(1{-}\alpha_k)g(\bar{x}^k){-}\alpha_k g(x)\right] \\ &\qquad - \alpha_k\langle \lambda, Ax^{k+1} - b\rangle \\ &\quad \le \alpha_k \langle \nabla f(\hat{x}^k), x^{k+1} - x\rangle + \frac{\alpha_k^2 L_f}{2}\|x^{k+1} - x^k\|^2 + \alpha_k[g(x^{k+1}) - g(x)] \\ &\qquad - \alpha_k\langle \lambda, Ax^{k+1} - b\rangle, \end{aligned} \tag{A.2}$$

where the equality follows from the fact $\bar{x}^{k+1} = (1-\alpha_k)\bar{x}^k + \alpha_k x^{k+1}$, and in the inequality we have used (A.1) and the convexity of g.

On the other hand, from the update rule of x^{k+1}, we have the optimality condition:

$$0 = \nabla f(\hat{x}^k) + \tilde{\nabla} g(x^{k+1}) - A^\top \lambda^k + \beta_k A^\top (Ax^{k+1} - b) + P^k(x^{k+1} - x^k),$$

where $\tilde{\nabla} g(x^{k+1})$ is a subgradient of g at x^{k+1}. Hence, for any x such that $Ax = b$, it holds

(A.3)

$$\begin{aligned}
0 &= \langle x^{k+1} - x, \nabla f(\hat{x}^k) + \tilde{\nabla} g(x^{k+1}) - A^\top \lambda^k + \beta_k A^\top (Ax^{k+1} - b) + P^k(x^{k+1} - x^k)\rangle \\
&\geq \langle x^{k+1} - x, \nabla f(\hat{x}^k) - A^\top \lambda^k + \beta_k A^\top (Ax^{k+1} - b) + P^k(x^{k+1} - x^k)\rangle + g(x^{k+1}) - g(x) \\
&= \left\langle x^{k+1} - x, \nabla f(\hat{x}^k) - A^\top \lambda^k + \frac{\beta_k}{\gamma_k} A^\top (\lambda^k - \lambda^{k+1}) + P^k(x^{k+1} - x^k) \right\rangle + g(x^{k+1}) - g(x) \\
&= \langle x^{k+1} - x, \nabla f(\hat{x}^k)\rangle + g(x^{k+1}) - g(x) + \langle x^{k+1} - x, P^k(x^{k+1} - x^k)\rangle \\
&\quad + \left\langle A(x^{k+1} - x), -\lambda^k + \frac{\beta_k}{\gamma_k}(\lambda^k - \lambda^{k+1}) \right\rangle \\
&= \langle x^{k+1} - x, \nabla f(\hat{x}^k)\rangle + g(x^{k+1}) - g(x) + \langle x^{k+1} - x, P^k(x^{k+1} - x^k)\rangle \\
&\quad + \left\langle Ax^{k+1} - b, \lambda - \lambda^k + \frac{\beta_k}{\gamma_k}(\lambda^k - \lambda^{k+1}) \right\rangle - \langle \lambda, Ax^{k+1} - b\rangle \\
&= \langle x^{k+1} - x, \nabla f(\hat{x}^k)\rangle + g(x^{k+1}) - g(x) - \langle \lambda, Ax^{k+1} - b\rangle + \langle x^{k+1} - x, P^k(x^{k+1} - x^k)\rangle \\
&\quad + \left\langle \frac{1}{\gamma_k}(\lambda^k - \lambda^{k+1}), \lambda - \lambda^k + \frac{\beta_k}{\gamma_k}(\lambda^k - \lambda^{k+1}) \right\rangle,
\end{aligned}$$

where the inequality follows from the convexity of g.

Combining (A.2) and (A.3) together gives

$$\begin{aligned}
&\left[F(\bar{x}^{k+1}) - F(x) - \langle \lambda, A\bar{x}^{k+1} - b\rangle\right] - (1 - \alpha_k)\left[F(\bar{x}^k) - F(x) - \langle \lambda, A\bar{x}^k - b\rangle\right] \\
&\quad \leq \frac{\alpha_k^2 L_f}{2}\|x^{k+1} - x^k\|^2 - \alpha_k \langle x^{k+1} - x, P^k(x^{k+1} - x^k)\rangle \\
&\qquad - \alpha_k \left\langle \frac{1}{\gamma_k}(\lambda^k - \lambda^{k+1}), \lambda - \lambda^k + \frac{\beta_k}{\gamma_k}(\lambda^k - \lambda^{k+1}) \right\rangle.
\end{aligned}$$

Now apply (2.3) to complete the proof.

A.2. Proof of Lemma 2.12. From the update (1.15a), we have the optimality condition

$$0 = \tilde{\nabla} h(y^{k+1}) - B^\top \lambda^k + \beta_k B^\top (By^{k+1} + Cz^k - b) + P^k(y^{k+1} - y^k),$$

where $\tilde{\nabla} h(y^{k+1})$ is a subgradient of h at y^{k+1}. Thus for any y,

(A.4)

$$\begin{aligned}
0 = &\left\langle y^{k+1} - y, \tilde{\nabla} h(y^{k+1}) - B^\top \lambda^k + \beta_k B^\top (By^{k+1} + Cz^k - b) + P^k(y^{k+1} - y^k) \right\rangle \\
\geq\ & h(y^{k+1}) - h(y) + \langle y^{k+1} - y, -B^\top \lambda^k + \beta_k B^\top (By^{k+1} + Cz^k - b) + P^k(y^{k+1} - y^k)\rangle \\
=\ & h(y^{k+1}) - h(y) + \langle y^{k+1} - y, -B^\top \lambda^k + \beta_k B^\top (By^{k+1} + Cz^{k+1} - b) \\
&- \beta_k B^\top C(z^{k+1} - z^k)\rangle + \langle y^{k+1} - y, P^k(y^{k+1} - y^k)\rangle \\
=\ & h(y^{k+1}) - h(y) + \left\langle B(y^{k+1} - y), -\lambda^k + \frac{\beta_k}{\gamma_k}(\lambda^k - \lambda^{k+1}) \right\rangle \\
&- \beta_k \langle B(y^{k+1} - y), C(z^{k+1} - z^k)\rangle + \langle y^{k+1} - y, P^k(y^{k+1} - y^k)\rangle,
\end{aligned}$$

where in the last equality we have used the update rule (1.15c). Similar to (A.1), we have

$$f(z^{k+1}) \le f(z) + \langle \nabla f(z^k), z^{k+1} - z\rangle + \frac{L_f}{2}\|z^{k+1} - z^k\|^2 - \frac{\mu_f}{2}\|z^k - z\|^2. \tag{A.5}$$

From the update rule of z^{k+1}, we have the optimality condition:

$$0 = \tilde{\nabla} g(z^{k+1}) + \nabla f(z^k) - C^\top \lambda^k + \beta_k C^\top (Bx^{k+1} + Cz^{k+1} - b) + Q^k(z^{k+1} - z^k).$$

Hence, for any z, it holds

(A.6)

$$\begin{aligned}
0 &= \left\langle z^{k+1} - z, \tilde{\nabla} g(z^{k+1}) + \nabla f(z^k) - C^\top \lambda^k + \beta_k C^\top (By^{k+1} + Cz^{k+1} - b) + Q^k(z^{k+1} - z^k)\right\rangle \\
&\ge g(z^{k+1}) - g(z) + \frac{\mu_g}{2}\|z^{k+1} - z\|^2 + \langle z^{k+1} - z, \nabla f(z^k)\rangle \\
&\quad + \left\langle z^{k+1} - z, -C^\top \lambda^k + \beta_k C^\top (By^{k+1} + Cz^{k+1} - b) + Q^k(z^{k+1} - z^k)\right\rangle \\
&= g(z^{k+1}) - g(z) + \frac{\mu_g}{2}\|z^{k+1} - z\|^2 + \left\langle z^{k+1} - z, \nabla f(z^k)\right\rangle + \left\langle z^{k+1} - z, Q^k(z^{k+1} - z^k)\right\rangle \\
&\quad + \left\langle C(z^{k+1} - z), -\lambda^k + \frac{\beta_k}{\gamma_k}(\lambda^k - \lambda^{k+1})\right\rangle,
\end{aligned}$$

where the inequality follows from the convexity of g.

Since (y, z) is feasible, summing (A.4), (A.5), and (A.6) gives

$$\begin{aligned}
&F(y^{k+1}, z^{k+1}) - F(y, z) - \langle \lambda, By^{k+1} + Cz^{k+1} - b\rangle \\
&\le -\left\langle B(y^{k+1} - y), -\lambda^k + \frac{\beta_k}{\gamma_k}(\lambda^k - \lambda^{k+1})\right\rangle - \left\langle C(z^{k+1} - z), -\lambda^k + \frac{\beta_k}{\gamma_k}(\lambda^k - \lambda^{k+1})\right\rangle \\
&\quad - \langle \lambda, By^{k+1} + Cz^{k+1} - b\rangle + \beta_k \langle B(y^{k+1} - y), C(z^{k+1} - z^k)\rangle \\
&\quad + \frac{L_f}{2}\|z^{k+1} - z^k\|^2 - \frac{\mu_f}{2}\|z^k - z\|^2 - \frac{\mu_g}{2}\|z^{k+1} - z\|^2 \\
&\quad - \langle y^{k+1} - y, P^k(y^{k+1} - y^k)\rangle - \langle z^{k+1} - z, Q^k(z^{k+1} - z^k)\rangle,
\end{aligned}$$

which implies (2.25) by noting the update rule (1.15c).

REFERENCES

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

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

[3] K. Bredies and H. Sun, *Accelerated Douglas-Rachford methods for the solution of convex-concave saddle-point problems*, preprint, arXiv:1604.06282, (2016).

[4] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, J. Math. Imaging Vision, 40 (2011), pp. 120–145.

[5] Y. Chen, G. Lan, and Y. Ouyang, *Optimal primal-dual methods for a class of saddle point problems*, SIAM J. Optim., 24 (2014), pp. 1779–1814.

[6] L. Condat, *A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms*, J. Optim. Theory Appl., 158 (2013), pp. 460–479.

[7] C. Dang and G. Lan, *Randomized methods for saddle point computation*, preprint, arXiv:1409.8625, 2014.

[8] W. Deng and W. Yin, *On the global and linear convergence of the generalized alternating direction method of multipliers*, J. Sci. Comput., 66 (2016), pp. 889–916.

[9] O. Fercoq and P. Richtárik, *Accelerated, parallel, and proximal coordinate descent*, SIAM J. Optim., 25 (2015), pp. 1997–2023.
[10] D. Gabay and B. Mercier, *A dual algorithm for the solution of nonlinear variational problems via finite element approximation*, Comput. Math. Appl., 2 (1976), pp. 17–40.
[11] X. Gao, Y. Xu, and S. Zhang, *Randomized primal-dual proximal block coordinate updates*, arXiv preprint arXiv:1605.05969, 2016.
[12] X. Gao and S.-Z. Zhang, *First-order algorithms for convex optimization with nonseparable objective and coupled constraints*, J. Oper. Res. Soc. China, (2015), pp. 1–29.
[13] S. Ghadimi and G. Lan, *Accelerated gradient methods for nonconvex nonlinear and stochastic programming*, Math. Prog., 156 (2016), pp. 59–99.
[14] R. Glowinski and A. Marrocco, *Sur l'approximation, par eléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires*, ESAIM: Math. Model. Numer. Anal., 9 (1975), pp. 41–76.
[15] D. Goldfarb, S. Ma, and K. Scheinberg, *Fast alternating linearization methods for minimizing the sum of two convex functions*, Math. Prog., 141 (2013), pp. 349–382.
[16] T. Goldstein, B. O'Donoghue, S. Setzer, and R. Baraniuk, *Fast alternating direction optimization methods*, SIAM J. Imag. Sci., 7 (2014), pp. 1588–1623.
[17] M. Grant, S. Boyd, and Y. Ye, *CVX: Matlab software for disciplined convex programming*, 2008.
[18] B. He and X. Yuan, *On the acceleration of augmented Lagrangian method for linearly constrained optimization*, Optimization online, 3 (2010).
[19] Y. He and R. D. Monteiro, *An accelerated hpe-type algorithm for a class of composite convex-concave saddle-point problems*, SIAM J. Optim., 26 (2016), pp. 29–56.
[20] M. Hong and Z.-Q. Luo, *On the linear convergence of the alternating direction method of multipliers*, Math. Program. Ser. A, 162 (2017), pp. 165–199.
[21] B. Huang, S. Ma, and D. Goldfarb, *Accelerated linearized Bregman method*, J. Sci. Comput., 54 (2013), pp. 428–453.
[22] M. Kadkhodaie, K. Christakopoulou, M. Sanjabi, and A. Banerjee, *Accelerated alternating direction method of multipliers*, in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 497–506.
[23] M. Kang, M. Kang, and M. Jung, *Inexact accelerated augmented Lagrangian methods*, Comput. Optim. Appl., 62 (2015), pp. 373–404.
[24] M. Kang, S. Yun, H. Woo, and M. Kang, *Accelerated Bregman method for linearly constrained ℓ_1–ℓ_2 minimization*, J. Sci. Comput., 56 (2013), pp. 515–534.
[25] G. Lan, *An optimal method for stochastic composite optimization*, Math. Prog., 133 (2012), pp. 365–397.
[26] Q. Lin, Z. Lu, and L. Xiao, *An accelerated proximal coordinate gradient method*, in Adv. Neural Info. Process. Sys., 2014, pp. 3059–3067.
[27] Y. Nesterov, *A method of solving a convex programming problem with convergence rate $O(1/k^2)$*, Soviet Math. Doklady, 27 (1983), pp. 372–376.
[28] Y. Nesterov, *Excessive gap technique in nonsmooth convex minimization*, SIAM J. Optim., 16 (2005), pp. 235–249.
[29] Y. Nesterov, *Smooth minimization of non-smooth functions*, Math. Program. Ser. A, 103 (2005), pp. 127–152.
[30] Y. Nesterov, *Gradient methods for minimizing composite functions*, Math. Program. Ser. B, 140 (2013), pp. 125–161.
[31] R. Nishihara, L. Lessard, B. Recht, A. Packard, and M. Jordan, *A general analysis of the convergence of ADMM*, in Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 2015, pp. 343–352.
[32] J. Nocedal and S. Wright, *Numerical Optimization*, Springer Science & Business Media, New York, 2006.
[33] B. ÓDonoghue and E. Candes, *Adaptive restart for accelerated gradient schemes*, Found. Comput. Math., 15 (2015), pp. 715–732.
[34] Y. Ouyang, Y. Chen, G. Lan, and E. Pasiliao Jr, *An accelerated linearized alternating direction method of multipliers*, SIAM J. Imaging Sci., 8 (2015), pp. 644–681.
[35] W. Su, S. Boyd, and E. Candes, *A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights*, J. Mach. Learn. Res., 17 (2016), pp. 1–43.
[36] P. Tseng, *On Accelerated Proximal Gradient Methods for Convex-Concave Optimization*, preprint, 2008.
[37] Y. Wang, J. Yang, W. Yin, and Y. Zhang, *A new alternating minimization algorithm for total variation image reconstruction*, SIAM J. Imag. Sci., 1 (2008), pp. 248–272.
[38] A. Wibisono, A. C. Wilson, and M. I. Jordan, *A variational perspective on accelerated methods in optimization*, Proc. Natl Acad. Sci. USA, (2016), pp. E7351–E7358.

[39] Y. Xu, I. Akrotirianakis, and A. Chakraborty, *Proximal gradient method for huberized support vector machine*, Pattern Anal. Appl., 19 (2016), pp. 989–1005.

[40] Y. Xu and W. Yin, *A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion*, SIAM J. Imag. Sci., 6 (2013), pp. 1758–1789.

[41] Y. Xu and S. Zhang, *Accelerated primal-dual proximal block coordinate updating methods for constrained convex optimization*, preprint, arXiv:1702.05423, 2017.

[42] G.-B. Ye, Y. Chen, and X. Xie, *Efficient variable selection in support vector machines via the alternating direction method of multipliers.*, in Proceedings of the 14th AISTATS, Ft. Lauderdale, 2011, pp. 832–840.