Math. Prog. Comp.
DOI 10.1007/s12532-015-0078-2

FULL LENGTH PAPER

Generalized alternating direction method of multipliers: new theoretical insights and applications

Ethan X. Fang · Bingsheng He · Han Liu · Xiaoming Yuan

Received: 27 February 2014 / Accepted: 15 January 2015

Abstract Recently, the alternating direction method of multipliers (ADMM) has received intensive attention from a broad spectrum of areas. The generalized ADMM (GADMM) proposed by Eckstein and Bertsekas is an efficient and simple acceleration scheme of ADMM. In this paper, we take a deeper look at the linearized version of GADMM where one of its subproblems is approximated by a linearization strategy. This linearized version is particularly efficient for a number of applications arising from different areas. Theoretically, we show the worst-case $\mathcal{O}(1/k)$ convergence rate measured by the iteration complexity (k represents

Bingsheng He: This author was supported by the NSFC Grant 11471156.

Xiaoming Yuan: This author was supported by the Faculty Research Grant from HKBU: FRG2/13-14/061 and the General Research Fund from Hong Kong Research Grants Council: 203613.

E. X. Fang · H. Liu
Department of Operations Research and Financial Engineering,
Princeton University, Princeton, NJ 08544, USA
e-mail: xingyuan@princeton.edu

H. Liu
e-mail: hanliu@princeton.edu

B. He
International Centre of Management Science and Engineering,
and Department of Mathematics, Nanjing University,
Nanjing 210093, China
e-mail: hebma@nju.edu.cn

X. Yuan (✉)
Department of Mathematics, Hong Kong Baptist University,
Kowloon, Hong Kong
e-mail: xmyuan@hkbu.edu.hk

Published online: 06 February 2015

the iteration counter) in both the ergodic and a nonergodic senses for the linearized version of GADMM. Numerically, we demonstrate the efficiency of this linearized version of GADMM by some rather new and core applications in statistical learning. Code packages in Matlab for these applications are also developed.

Keywords Convex optimization · Alternating direction method of multipliers · Convergence rate · Variable selection · Discriminant analysis · Statistical learning

Mathematics Subject Classification 90C25 · 90C06 · 62J05

1 Introduction

A canonical convex optimization model with a separable objective function and linear constraints is:

$$\min\{f_1(\mathbf{x}) + f_2(\mathbf{y}) \mid \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} = \mathbf{b},\ \mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}\}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{n\times n_1}$, $\mathbf{B} \in \mathbb{R}^{n\times n_2}$, $\mathbf{b} \in \mathbb{R}^n$, and $\mathcal{X} \subset \mathbb{R}^{n_1}$ and $\mathcal{Y} \subset \mathbb{R}^{n_2}$ are closed convex nonempty sets, $f_1 : \mathbb{R}^{n_1} \to \mathbb{R}$ and $f_2 : \mathbb{R}^{n_2} \to \mathbb{R}$ are convex but not necessarily smooth functions. Throughout our discussion, the solution set of (1) is assumed to be nonempty, and the matrix $\mathbf{B}$ is assumed to have full column rank.

The motivation of discussing the particular model (1) with separable structures is that each function f_i might have its own properties, and we need to explore these properties effectively in algorithmic design in order to develop efficient numerical algorithms. A typical scenario is where one of the functions represents some data-fidelity term, and the other is a certain regularization term—we can easily find such an application in many areas such as inverse problem, statistical learning and image processing. For example, the famous least absolute shrinkage and selection operator (LASSO) model introduced in [44] is a special case of (1) where f_1 is the ℓ_1-norm term for promoting sparsity, f_2 is a least-squares term multiplied by a trade-off parameter, $\mathbf{A} = \mathbf{I}_{n\times n}$, $\mathbf{B} = -\mathbf{I}_{n\times n}$, $\mathbf{b} = \mathbf{0}$, $\mathcal{X} = \mathcal{Y} = \mathbb{R}^n$.

To solve (1), a benchmark is the alternating direction method of multipliers (ADMM) proposed originally in [24] which is essentially a splitting version of the augmented Lagrangian method in [34,42]. The iterative scheme of ADMM for solving (1) reads as

$$\begin{aligned}
\mathbf{x}^{t+1} &= \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}} \left\{ f_1(\mathbf{x}) - \mathbf{x}^T\mathbf{A}^T\gamma^t + \frac{\rho}{2}\|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y}^t - \mathbf{b}\|^2 \right\},\\
\mathbf{y}^{t+1} &= \operatorname*{argmin}_{\mathbf{y}\in\mathcal{Y}} \left\{ f_2(\mathbf{y}) - \mathbf{y}^T\mathbf{B}^T\gamma^t + \frac{\rho}{2}\|\mathbf{A}\mathbf{x}^{t+1} + \mathbf{B}\mathbf{y} - \mathbf{b}\|^2 \right\},\\
\gamma^{t+1} &= \gamma^t - \rho\left(\mathbf{A}\mathbf{x}^{t+1} + \mathbf{B}\mathbf{y}^{t+1} - \mathbf{b}\right),
\end{aligned} \tag{2}$$

where $\gamma \in \mathbb{R}^n$ is the Lagrangian multiplier; $\rho > 0$ is a penalty parameter, and $\|\cdot\|$ is the Euclidean 2-norm. An important feature of ADMM is that the functions f_1 and f_2 are treated individually; thus the decomposed subproblems in (2) might be significantly easier than the original problem (1). Recently, the ADMM has received wide attention from a broad spectrum of areas because of its easy implementation and impressive efficiency. We refer to [6,13,23] for excellent review papers for the history and applications of ADMM.

In [21], the ADMM was explained as an application of the well-known Douglas-Rachford splitting method (DRSM) in [36] to the dual of (1); and in [14], the DRSM was further explained as an application of the proximal point algorithm (PPA) in [37]. Therefore, it was suggested in [14] to apply the acceleration scheme in [25] for the PPA to accelerate the original ADMM (2). A generalized ADMM (GADMM for short) was thus proposed:

$$\begin{aligned}
\mathbf{x}^{t+1} &= \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmin}} \left\{ f_1(\mathbf{x}) - \mathbf{x}^T\mathbf{A}^T\gamma^t + \frac{\rho}{2}\|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y}^t - \mathbf{b}\|^2 \right\},\\
\mathbf{y}^{t+1} &= \underset{\mathbf{y}\in\mathcal{Y}}{\operatorname{argmin}} \left\{ f_2(\mathbf{y}) - \mathbf{y}^T\mathbf{B}^T\gamma^t + \frac{\rho}{2}\|\alpha\mathbf{A}\mathbf{x}^{t+1} + (1-\alpha)(\mathbf{b} - \mathbf{B}\mathbf{y}^t) + \mathbf{B}\mathbf{y} - \mathbf{b}\|^2 \right\},\\
\gamma^{t+1} &= \gamma^t - \rho\Big(\alpha\mathbf{A}\mathbf{x}^{t+1} + (1-\alpha)(\mathbf{b} - \mathbf{B}\mathbf{y}^t) + \mathbf{B}\mathbf{y}^{t+1} - \mathbf{b}\Big),
\end{aligned} \tag{3}$$

where the parameter $\alpha \in (0, 2)$ is a relaxation factor. Obviously, the generalized scheme (3) reduces to the original ADMM scheme (2) when $\alpha = 1$. Preserving the main advantage of the original ADMM in treating the objective functions f_1 and f_2 individually, the GADMM (3) enjoys the same easiness in implementation while can numerically accelerate (2) with some values of α, e.g., $\alpha \in (1, 2)$. We refer to [2,8,12] for empirical studies of the acceleration performance of the GADMM.

It is necessary to discuss how to solve the decomposed subproblems in (2) and (3). We refer to [41] for the ADMM's generic case where no special property is assumed for the functions f_1 and f_2, and thus the subproblems in (2) must be solved approximately subject to certain inexactness criteria in order to ensure the convergence for inexact versions of the ADMM. For some concrete applications such as those arising in sparse or low-rank optimization models, one function (say, f_1) is nonsmooth but well-structured (More mathematically, the resolvent of ∂f_1 has a closed-form representation), and the other function f_2 is smooth and simple enough so that the $\mathbf{y}$-subproblem is easy (e.g., when f_2 is the least-squares term). For such a case, instead of discussing a generic strategy to solve the $\mathbf{x}$-subproblem in (2) or (3) approximately, we prefer to seek some particular strategies that can take advantage of the speciality of f_1 effectively. More accurately, when f_1 is a special function such as the ℓ_1-norm or nuclear-norm function arising often in applications, we prefer linearizing the quadratic term of the $\mathbf{x}$-subproblem in (2) or (3) so that the linearized $\mathbf{x}$-subproblem has a closed-form solution (amounting to estimating the resolvent of ∂f_1) and thus no inner iteration is required. The efficiency of this linearization strategy for the ADMM has been well illustrated in different literatures, see e.g., [52] for image reconstruction problems, [48] for the Dantzig Selector model, and [50] for some low-rank optimization models. Inspired by these applications, we thus consider the linearized version of the GADMM

("L-GADMM" for short) where the $\mathbf{x}$-subproblem in (3) is linearized:

$$\begin{aligned}
\mathbf{x}^{t+1} &= \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmin}} \left\{ f_1(\mathbf{x}) - \mathbf{x}^T\mathbf{A}^T\gamma^t + \frac{\rho}{2}\|\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{y}^t-\mathbf{b}\|^2 + \frac{1}{2}\|\mathbf{x}-\mathbf{x}^t\|_{\mathbf{G}}^2 \right\},\\
\mathbf{y}^{t+1} &= \underset{\mathbf{y}\in\mathcal{Y}}{\operatorname{argmin}} \left\{ f_2(\mathbf{y}) - \mathbf{y}^T\mathbf{B}^T\gamma^t + \frac{\rho}{2}\|\alpha\mathbf{A}\mathbf{x}^{t+1} + (1-\alpha)(\mathbf{b}-\mathbf{B}\mathbf{y}^t) + \mathbf{B}\mathbf{y} - \mathbf{b}\|^2 \right\},\\
\gamma^{t+1} &= \gamma^t - \rho\Big(\alpha\mathbf{A}\mathbf{x}^{t+1} + (1-\alpha)(\mathbf{b}-\mathbf{B}\mathbf{y}^t) + \mathbf{B}\mathbf{y}^{t+1} - \mathbf{b}\Big),
\end{aligned} \tag{4}$$

where $\mathbf{G} \in \mathbb{R}^{n_1\times n_1}$ is a symmetric positive definite matrix. Note that we use the notation $\|\mathbf{x}\|_{\mathbf{G}}$ to denote the quantity $\sqrt{\mathbf{x}^T\mathbf{G}\mathbf{x}}$. Clearly, if $\mathcal{X} = \mathbb{R}^{n_1}$ and we choose $\mathbf{G} = \tau\mathbf{I}_{n_1} - \rho\mathbf{A}^T\mathbf{A}$ with the requirement $\tau > \rho\|\mathbf{A}^T\mathbf{A}\|_2$, where $\|\cdot\|_2$ denotes the spectral norm of a matrix, the $\mathbf{x}$-subproblem in (4) reduces to estimating the resolvent of ∂f_1:

$$\mathbf{x}^{t+1} = \underset{\mathbf{x}\in\mathbb{R}^{n_1}}{\operatorname{argmin}} \left\{ f_1(\mathbf{x}) + \frac{\tau}{2}\left\|\mathbf{x} - \frac{1}{\tau}\Big((\tau\mathbf{I}_{n_1} - \rho\mathbf{A}^T\mathbf{A})\mathbf{x}^t - \rho\mathbf{A}^T\mathbf{B}\mathbf{y}^t + \mathbf{A}^T\gamma^t + \rho\mathbf{A}^T\mathbf{b}\Big)\right\|^2 \right\},$$

which has a closed-form solution for some cases such as $f_1 = \|\mathbf{x}\|_1$. The scheme (4) thus includes the linearized version of ADMM (see e.g. [48,50,52]) as a special case with $\mathbf{G} = \tau\mathbf{I}_{n_1} - \rho\mathbf{A}^T\mathbf{A}$ and $\alpha = 1$.

The convergence analysis of ADMM has appeared in earlier literatures, see e.g., [20,22,29,30]. Recently, it also becomes popular to estimate ADMM's worst-case convergence rate measured by the iteration complexity (see e.g., [39,40] for the rationale of measuring the convergence rate of an algorithm by means of its iteration complexity). In [31], a worst-case $\mathcal{O}(1/k)$ convergence rate in the ergodic sense was established for both the original ADMM scheme (2) and its linearized version (i.e., the special case of (4) with $\alpha = 1$), and then a stronger result in a nonergodic sense was proved in [32]. We also refer to [33] for an extension of the result in [32] to the DRSM for the general problem of finding a zero point of the sum of two maximal monotone operators, [11] for the linear convergence of the ADMM under additional stronger assumptions, and [5,27] for the linear convergence of ADMM for the special case of (1) where both f_1 and f_2 are quadratic functions.

This paper aims at further studying the L-GADMM (4) both theoretically and numerically. Theoretically, we shall establish the worst-case $\mathcal{O}(1/k)$ convergence rate in both the ergodic and a nonergodic senses for L-GADMM. This is the first worst-case convergence rate for L-GADMM, and it includes the results in [8,32] as special cases. Numerically, we apply the L-GADMM (4) to solve some rather new and core applications arising in statistical learning. The acceleration effectiveness of embedding the linearization technique with the GADMM is thus verified.

The rest of this paper is organized as follows. We summarize some preliminaries which are useful for further analysis in Sect. 2. Then, we derive the worst-case convergence rate for the L-GADMM (4) in the ergodic and a nonergodic senses in Sects. 3 and 4, respectively. In Sect. 5, we apply the L-GADMM (4) to solve some statistical learning applications and verify its numerical efficiency. Finally, we make some conclusions in Sect. 6.

2 Preliminaries

First, as well known in the literature (see, e.g. [30,31]), solving (1) is equivalent to solving the following variational inequality (VI) problem: Finding $\mathbf{w}^* = (\mathbf{x}^*, \mathbf{y}^*, \gamma^*) \in \Omega := \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^n$ such that

$$f(\mathbf{u}) - f(\mathbf{u}^*) + (\mathbf{w} - \mathbf{w}^*)^T F(\mathbf{w}^*) \geq 0, \quad \forall \mathbf{w} \in \Omega, \tag{5}$$

where $f(\mathbf{u}) = f_1(\mathbf{x}) + f_2(\mathbf{y})$ and

$$\mathbf{u} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \\ \gamma \end{pmatrix}, \quad F(\mathbf{w}) = \begin{pmatrix} -\mathbf{A}^T\gamma \\ -\mathbf{B}^T\gamma \\ \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} - \mathbf{b} \end{pmatrix}. \tag{6}$$

We denote by VI(Ω, F, f) the problem (5, 6). It is easy to see that the mapping $F(\mathbf{w})$ defined in (6) is affine with a skew-symmetric matrix; it is thus monotone:

$$(\mathbf{w}_1 - \mathbf{w}_2)^T \big(F(\mathbf{w}_1) - F(\mathbf{w}_2)\big) \geq 0, \quad \forall \mathbf{w}_1, \mathbf{w}_2 \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^n.$$

This VI reformulation will provide significant convenience for theoretical analysis later. The solution set of (5), denoted by Ω^*, is guaranteed to be nonempty under our nonempty assumption on the solution set of (1).

Then, we define two auxiliary sequences for the convenience of analysis. More specifically, for the sequence $\{\mathbf{w}^t\}$ generated by the L-GADMM (4), let

$$\widetilde{\mathbf{w}}^t = \begin{pmatrix} \widetilde{\mathbf{x}}^t \\ \widetilde{\mathbf{y}}^t \\ \widetilde{\gamma}^t \end{pmatrix} = \begin{pmatrix} \mathbf{x}^{t+1} \\ \mathbf{y}^{t+1} \\ \gamma^t - \rho(\mathbf{A}\mathbf{x}^{t+1} + \mathbf{B}\mathbf{y}^t - \mathbf{b}) \end{pmatrix} \quad \text{and} \quad \widetilde{\mathbf{u}}^t = \begin{pmatrix} \widetilde{\mathbf{x}}^t \\ \widetilde{\mathbf{y}}^t \end{pmatrix}. \tag{7}$$

Note that, by the definition of γ^{t+1} in (4), we get

$$\gamma^t - \gamma^{t+1} = -\rho\mathbf{B}\left(\mathbf{y}^t - \mathbf{y}^{t+1}\right) + \rho\alpha\left(\mathbf{A}\mathbf{x}^{t+1} + \mathbf{B}\mathbf{y}^t - \mathbf{b}\right).$$

Plugging the identities $\rho(\mathbf{A}\mathbf{x}^{t+1} + \mathbf{B}\mathbf{y}^t - \mathbf{b}) = \gamma^t - \widetilde{\gamma}^t$ and $\mathbf{y}^{t+1} = \widetilde{\mathbf{y}}^t$ [see (7)] into the above equation, it holds that

$$\gamma^t - \gamma^{t+1} = -\rho\mathbf{B}\left(\mathbf{y}^t - \widetilde{\mathbf{y}}^t\right) + \alpha\left(\gamma^t - \widetilde{\gamma}^t\right). \tag{8}$$

Then we have

$$\mathbf{w}^t - \mathbf{w}^{t+1} = \mathbf{M}\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right), \tag{9}$$

where $\mathbf{M}$ is defined as

$$\mathbf{M} = \begin{pmatrix} \mathbf{I}_{n_1} & 0 & 0 \\ 0 & \mathbf{I}_{n_2} & 0 \\ 0 & -\rho\mathbf{B} & \alpha\mathbf{I}_n \end{pmatrix}. \tag{10}$$

For notational simplicity, we define two matrices that will be used later in the proofs:

$$\mathbf{H}=\begin{pmatrix}\mathbf{G} & 0 & 0\\ 0 & \frac{\rho}{\alpha}\mathbf{B}^T\mathbf{B} & \frac{1-\alpha}{\alpha}\mathbf{B}^T\\ 0 & \frac{1-\alpha}{\alpha}\mathbf{B} & \frac{1}{\alpha\rho}\mathbf{I}_n\end{pmatrix} \quad \text{and} \quad \mathbf{Q}=\begin{pmatrix}\mathbf{G} & 0 & 0\\ 0 & \rho\mathbf{B}^T\mathbf{B} & (1-\alpha)\mathbf{B}^T\\ 0 & -\mathbf{B} & \frac{1}{\rho}\mathbf{I}_n\end{pmatrix}. \tag{11}$$

It is easy to verify that

$$\mathbf{Q}=\mathbf{H}\mathbf{M}. \tag{12}$$

3 A worst-case $\mathcal{O}(1/k)$ convergence rate in the ergodic sense

In this section, we establish a worst-case $\mathcal{O}(1/k)$ convergence rate in the ergodic sense for the L-GADMM (4). This is a more general result than that in [8] which focuses on the original GADMM (3) without linearization.

We first prove some lemmas. The first lemma is to characterize the accuracy of the vector $\widetilde{\mathbf{w}}^t$ to a solution point of VI(Ω, F, f).

Lemma 1 *Let the sequence* $\{\mathbf{w}^t\}$ *be generated by the L-GADMM* (4) *and the associated sequence* $\{\widetilde{\mathbf{w}}^t\}$ *be defined in* (7)*. Then we have*

$$f(\mathbf{u})-f\left(\widetilde{\mathbf{u}}^t\right)+\left(\mathbf{w}-\widetilde{\mathbf{w}}^t\right)^T F\left(\widetilde{\mathbf{w}}^t\right)\geq\left(\mathbf{w}-\widetilde{\mathbf{w}}^t\right)^T\mathbf{Q}\left(\mathbf{w}^t-\widetilde{\mathbf{w}}^t\right),\quad\forall\mathbf{w}\in\Omega, \tag{13}$$

where $\mathbf{Q}$ *is defined in* (11)*.*

Proof This lemma is proved by deriving the optimality conditions for the minimization subproblems in (4) and performing some algebraic manipulation. By deriving the optimality condition of the $\mathbf{x}$-subproblem of (4), as shown in [31], we have

$$\begin{aligned}&f_1(\mathbf{x})-f_1(\mathbf{x}^{t+1})+(\mathbf{x}-\mathbf{x}^{t+1})^T\Big(-\mathbf{A}^T[\gamma^t-\rho(\mathbf{A}\mathbf{x}^{t+1}+\mathbf{B}\mathbf{y}^t-\mathbf{b})]-\mathbf{x}^t)\Big)\\ &\geq 0,\forall\mathbf{x}\in\mathcal{X}.\end{aligned} \tag{14}$$

Using $\widetilde{\mathbf{x}}^t$ and $\widetilde{\gamma}^t$ defined in (7, 14) can be rewritten as

$$f_1(\mathbf{x})-f_1(\widetilde{\mathbf{x}}^t)+\left(\mathbf{x}-\widetilde{\mathbf{x}}^t\right)^T\left[-\mathbf{A}^T\widetilde{\gamma}^t+\mathbf{G}\left(\widetilde{\mathbf{x}}^t-\mathbf{x}^t\right)\right]\geq 0,\quad\forall\mathbf{x}\in\mathcal{X}. \tag{15}$$

Similarly, deriving the optimality condition for the $\mathbf{y}$-subproblem in (4), we have

$$f_2(\mathbf{y})-f_2\left(\mathbf{y}^{t+1}\right)+\left(\mathbf{y}-\mathbf{y}^{t+1}\right)^T\left(-\mathbf{B}^T\gamma^{t+1}\right)\geq 0,\quad\forall\mathbf{y}\in\mathcal{Y}. \tag{16}$$

From (8), we get

$$\gamma^{t+1}=\widetilde{\gamma}^t-\rho\mathbf{B}\left(\widetilde{\mathbf{y}}^t-\mathbf{y}^t\right)-(1-\alpha)\left(\widetilde{\gamma}^t-\gamma^t\right). \tag{17}$$

Substituting (17) into (16) and using the identity $\mathbf{y}^{t+1} = \widetilde{\mathbf{y}}^t$, we obtain that

$$f_2(\mathbf{y}) - f_2\left(\widetilde{\mathbf{y}}^t\right) + \left(\mathbf{y} - \widetilde{\mathbf{y}}^t\right)^T \left[-\mathbf{B}^T \widetilde{\gamma}^t + \rho \mathbf{B}^T \mathbf{B}\left(\widetilde{\mathbf{y}}^t - \mathbf{y}^t\right) + (1-\alpha)\mathbf{B}^T\left(\widetilde{\gamma}^t - \gamma^t\right)\right] \geq 0, \forall \mathbf{y} \in \mathcal{Y}. \tag{18}$$

Meanwhile, the third row of (7) implies that

$$\left(\mathbf{A}\widetilde{\mathbf{x}}^t + \mathbf{B}\widetilde{\mathbf{y}}^t - \mathbf{b}\right) - \mathbf{B}\left(\widetilde{\mathbf{y}}^t - \mathbf{y}^t\right) + \frac{1}{\rho}\left(\widetilde{\gamma}^t - \gamma^t\right) = 0. \tag{19}$$

Combining (15, 18) and (19), we get

$$f(\mathbf{u}) - f(\widetilde{\mathbf{u}}^t) + \begin{pmatrix} \mathbf{x} - \widetilde{\mathbf{x}}^t \\ \mathbf{y} - \widetilde{\mathbf{y}}^t \\ \gamma - \widetilde{\gamma}^t \end{pmatrix}^T \left\{ \begin{pmatrix} -\mathbf{A}^T\widetilde{\gamma}^t \\ -\mathbf{B}^T\widetilde{\gamma}^t \\ \mathbf{A}\widetilde{\mathbf{x}}^t + \mathbf{B}\widetilde{\mathbf{y}}^t - \mathbf{b} \end{pmatrix} + \begin{pmatrix} \mathbf{G}(\widetilde{\mathbf{x}}^t - \mathbf{x}^t) \\ \rho\mathbf{B}^T\mathbf{B}(\widetilde{\mathbf{y}}^t - \mathbf{y}^t) + (1-\alpha)\mathbf{B}^T(\widetilde{\gamma}^t - \gamma^t) \\ -\mathbf{B}(\widetilde{\mathbf{y}}^t - \mathbf{y}^t) + \frac{1}{\rho}(\widetilde{\gamma}^t - \gamma^t) \end{pmatrix} \right\} \geq 0, \quad \forall \mathbf{w} \in \Omega. \tag{20}$$

By the definition of F in (6) and $\mathbf{Q}$ in (11, 20) can be written as

$$f(\mathbf{u}) - f(\widetilde{\mathbf{u}}^t) + \left(\mathbf{w} - \widetilde{\mathbf{w}}^t\right)^T \left[F(\widetilde{\mathbf{w}}^t) + \mathbf{Q}\left(\widetilde{\mathbf{w}}^t - \mathbf{w}^t\right)\right] \geq 0, \ \forall \mathbf{w} \in \Omega.$$

The assertion (13) is proved. □

Recall the VI characterization (5, 6) of the model (1). Thus, according to (13), the accuracy of $\widetilde{\mathbf{w}}^t$ to a solution of VI(Ω, F, f) is measured by the quantity $(\mathbf{w} - \widetilde{\mathbf{w}}^t)^T\mathbf{Q}(\mathbf{w}^t - \widetilde{\mathbf{w}}^t)$. In the next lemma, we further explore this term and express it in terms of some quadratic terms, with which it becomes more convenient to estimate the accuracy of $\widetilde{\mathbf{w}}^t$ and thus to estimate the convergence rate for the scheme (4). Note that the matrix $\mathbf{B}$ is of full column rank and the matrix $\mathbf{G}$ is positive definite. Thus, the matrix $\mathbf{H}$ defined in (11) is positive definite for $\alpha \in (0, 2)$ and $\rho > 0$; and recall that we use the notation

$$\|\mathbf{w} - \mathbf{v}\|_{\mathbf{H}} = \sqrt{(\mathbf{w} - \mathbf{v})^T\mathbf{H}(\mathbf{w} - \mathbf{v})}$$

for further analysis.

Lemma 2 *Let the sequence $\{\mathbf{w}^t\}$ be generated by the L-GADMM (4) and the associated sequence $\{\widetilde{\mathbf{w}}^t\}$ be defined in (7). Then for any $\mathbf{w} \in \Omega$, we have*

$$\begin{aligned} &\left(\mathbf{w} - \widetilde{\mathbf{w}}^t\right)^T \mathbf{Q}\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right) \\ &= \frac{1}{2}\left(\|\mathbf{w} - \mathbf{w}^{t+1}\|_{\mathbf{H}}^2 - \|\mathbf{w} - \mathbf{w}^t\|_{\mathbf{H}}^2\right) + \frac{1}{2}\|\mathbf{x}^t - \widetilde{\mathbf{x}}^t\|_{\mathbf{G}}^2 + \frac{2-\alpha}{2\rho}\|\gamma^t - \widetilde{\gamma}^t\|^2. \end{aligned} \tag{21}$$

Proof Using $\mathbf{Q} = \mathbf{HM}$ and $\mathbf{M}(\mathbf{w}^t - \widetilde{\mathbf{w}}^t) = (\mathbf{w}^t - \mathbf{w}^{t+1})$ [see (9)], it follows that

$$\left(\mathbf{w}-\widetilde{\mathbf{w}}^t\right)^T \mathbf{Q}\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right) = \left(\mathbf{w}-\widetilde{\mathbf{w}}^t\right)^T \mathbf{HM}\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right) = \left(\mathbf{w} - \widetilde{\mathbf{w}}^t\right)^T \mathbf{H}\left(\mathbf{w}^t - \mathbf{w}^{t+1}\right). \tag{22}$$

For the vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$ in the same space and a matrix $\mathbf{H}$ with appropriate dimensionality, we have the identity

$$(\mathbf{a}-\mathbf{b})^T\mathbf{H}(\mathbf{c}-\mathbf{d}) = \frac{1}{2}\left(\|\mathbf{a}-\mathbf{d}\|_{\mathbf{H}}^2 - \|\mathbf{a}-\mathbf{c}\|_{\mathbf{H}}^2\right) + \frac{1}{2}\left(\|\mathbf{c}-\mathbf{b}\|_{\mathbf{H}}^2 - \|\mathbf{d}-\mathbf{b}\|_{\mathbf{H}}^2\right).$$

In this identity, we take

$$\mathbf{a} = \mathbf{w}, \quad \mathbf{b} = \widetilde{\mathbf{w}}^t, \quad \mathbf{c} = \mathbf{w}^t \quad \text{and} \quad \mathbf{d} = \mathbf{w}^{t+1},$$

and plug them into the right-hand side of (22). The resulting equation is

$$\begin{aligned}&\left(\mathbf{w} - \widetilde{\mathbf{w}}^t\right)^T \mathbf{H}\left(\mathbf{w}^t - \mathbf{w}^{t+1}\right)\\ &= \frac{1}{2}\left(\|\mathbf{w}-\mathbf{w}^{t+1}\|_{\mathbf{H}}^2 - \|\mathbf{w}-\mathbf{w}^t\|_{\mathbf{H}}^2\right) + \frac{1}{2}\left(\|\mathbf{w}^t - \widetilde{\mathbf{w}}^t\|_{\mathbf{H}}^2 - \|\mathbf{w}^{t+1} - \widetilde{\mathbf{w}}^t\|_{\mathbf{H}}^2\right).\end{aligned}$$

The remaining task is to prove

$$\|\mathbf{w}^t - \widetilde{\mathbf{w}}^t\|_{\mathbf{H}}^2 - \|\mathbf{w}^{t+1} - \widetilde{\mathbf{w}}^t\|_{\mathbf{H}}^2 = \|\mathbf{x}^t - \widetilde{\mathbf{x}}^t\|_{\mathbf{G}}^2 + \frac{2-\alpha}{\rho}\|\gamma^t - \widetilde{\gamma}^t\|^2. \tag{23}$$

By the definition of $\mathbf{H}$ given in (11), we have

$$\begin{aligned}\|\mathbf{w}^t - \widetilde{\mathbf{w}}^t\|_{\mathbf{H}}^2 &= \|\mathbf{x}^t - \widetilde{\mathbf{x}}^t\|_{\mathbf{G}}^2 + \frac{1}{\alpha\rho}\big(\|\rho\mathbf{B}(\mathbf{y}^t - \widetilde{\mathbf{y}}^t)\|^2 + \|\gamma^t - \widetilde{\gamma}^t\|^2\\ &\quad +2\rho(1-\alpha)(\mathbf{y}^t - \widetilde{\mathbf{y}}^t)^T\mathbf{B}^T(\gamma^t - \widetilde{\gamma}^t)\big)\\ &= \|\mathbf{x}^t - \widetilde{\mathbf{x}}^t\|_{\mathbf{G}}^2 + \frac{1}{\alpha\rho}\|\rho\mathbf{B}(\mathbf{y}^t - \widetilde{\mathbf{y}}^t) + (1-\alpha)(\gamma^t - \widetilde{\gamma}^t)\|^2\\ &\quad +\frac{2-\alpha}{\rho}\|\gamma^t - \widetilde{\gamma}^t\|^2.\end{aligned} \tag{24}$$

On the other hand, we have by (17) and the definition that $\widetilde{\mathbf{u}}^t = \mathbf{u}^{t+1}$,

$$\begin{aligned}\|\mathbf{w}^{t+1} - \widetilde{\mathbf{w}}^t\|_{\mathbf{H}}^2 &= \frac{1}{\alpha\rho}\|\gamma^{t+1} - \widetilde{\gamma}^t\|^2\\ &= \frac{1}{\alpha\rho}\|\rho\mathbf{B}\left(\mathbf{y}^t - \widetilde{\mathbf{y}}^t\right) + (1-\alpha)\left(\gamma^t - \widetilde{\gamma}^t\right)\|^2.\end{aligned} \tag{25}$$

Subtracting (25) from (24), performing some algebra yields (23), and the proof is completed. □

Lemmas 1 and 2 actually reassemble a simple proof for the convergence of the L-GADMM (4) from the perspectives of contraction type methods.

Theorem 1 *The sequence* $\{\mathbf{w}^t\}$ *generated by the L-GADMM* (4) *satisfies that for all* $\mathbf{w}^* \in \Omega^*$

$$\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_{\mathbf{H}}^2 \leq \|\mathbf{w}^t - \mathbf{w}^*\|_{\mathbf{H}}^2 - \left(\|\mathbf{x}^t - \mathbf{x}^{t+1}\|_{\mathbf{G}}^2 + \frac{2-\alpha}{\alpha^2}\|\mathbf{v}^t - \mathbf{v}^{t+1}\|_{\mathbf{H}_0}^2\right), \quad (26)$$

where

$$\mathbf{v} = \begin{pmatrix} \mathbf{y} \\ \gamma \end{pmatrix} \quad \text{and} \quad \mathbf{H}_0 = \begin{pmatrix} \rho\mathbf{B}^T\mathbf{B} & 0 \\ 0 & \frac{1}{\rho}\mathbf{I}_n \end{pmatrix}. \quad (27)$$

Proof Set $\mathbf{w} = \mathbf{w}^*$ in the assertion of Lemma 2, we get

$$\begin{aligned} &\|\mathbf{w}^t - \mathbf{w}^*\|_{\mathbf{H}}^2 - \|\mathbf{w}^{t+1} - \mathbf{w}^*\|_{\mathbf{H}}^2 \\ &= \|\mathbf{x}^t - \widetilde{\mathbf{x}}^t\|_{\mathbf{G}}^2 + \frac{2-\alpha}{\rho}\|\gamma^t - \widetilde{\gamma}^t\|^2 + 2\left(\widetilde{\mathbf{w}}^t - \mathbf{w}^*\right)^T \mathbf{HM}\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right). \end{aligned}$$

On the other hand, by using (13) and the monotonicity of F, we have

$$\left(\widetilde{\mathbf{w}}^t - \mathbf{w}^*\right)^T \mathbf{HM}\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right) \geq f(\widetilde{\mathbf{u}}^t) - f(\mathbf{u}^*) + \left(\widetilde{\mathbf{w}}^t - \mathbf{w}^*\right)^T F(\widetilde{\mathbf{w}}^t) \geq 0.$$

Consequently, we obtain

$$\|\mathbf{w}^t - \mathbf{w}^*\|_H^2 - \|\mathbf{w}^{t+1} - \mathbf{w}^*\|_{\mathbf{H}}^2 \geq \|\mathbf{x}^t - \widetilde{\mathbf{x}}^t\|_{\mathbf{G}}^2 + \frac{2-\alpha}{\rho}\|\gamma^t - \widetilde{\gamma}^t\|^2. \quad (28)$$

Since $\widetilde{\mathbf{y}}^t = \mathbf{y}^{t+1}$, it follows from (9) that

$$\gamma^t - \widetilde{\gamma}^t = \frac{1}{\alpha}\big(\rho\mathbf{B}(\mathbf{y}^t - \mathbf{y}^{t+1}) + (\gamma^t - \gamma^{t+1})\big).$$

Substituting it into (28), we obtain

$$\begin{aligned} &\|\mathbf{w}^t - \mathbf{w}^*\|_{\mathbf{H}}^2 - \|\mathbf{w}^{t+1} - \mathbf{w}^*\|_{\mathbf{H}}^2 \\ &\geq \|\mathbf{x}^t - \widetilde{\mathbf{x}}^t\|_{\mathbf{G}}^2 + \frac{(2-\alpha)}{\alpha^2\rho}\|\rho\mathbf{B}\left(\mathbf{y}^t - \mathbf{y}^{t+1}\right) + \left(\gamma^t - \gamma^{t+1}\right)\|^2. \end{aligned} \quad (29)$$

Note that (16) is true for any integer $t \geq 0$. Thus we have

$$f_2(\mathbf{y}) - f_2(\mathbf{y}^t) + \left(\mathbf{y} - \mathbf{y}^t\right)^T\left(-\mathbf{B}^T\gamma^t\right) \geq \mathbf{0}, \quad \forall\, \mathbf{y} \in \mathcal{Y}. \quad (30)$$

Setting $\mathbf{y} = \mathbf{y}^t$ and $\mathbf{y} = \mathbf{y}^{t+1}$ in (16) and (30), respectively, we get

$$f_2(\mathbf{y}^t) - f_2(\mathbf{y}^{t+1}) + \left(\mathbf{y}^t - \mathbf{y}^{t+1}\right)^T\left(-\mathbf{B}^T\gamma^{t+1}\right) \geq 0$$

and

$$f_2(\mathbf{y}^{t+1}) - f_2(\mathbf{y}^t) + \left(\mathbf{y}^{t+1} - \mathbf{y}^t\right)^T \left(-\mathbf{B}^T \gamma^t\right) \geq 0.$$

Adding the above two inequalities yields

$$\left(\gamma^t - \gamma^{t+1}\right)^T \mathbf{B} \left(\mathbf{y}^t - \mathbf{y}^{t+1}\right) \geq 0. \tag{31}$$

Substituting it in (29) and using $\widetilde{\mathbf{x}}^t = \mathbf{x}^{t+1}$, we obtain

$$\begin{aligned} &\|\mathbf{w}^t - \mathbf{w}^*\|_{\mathbf{H}}^2 - \|\mathbf{w}^{t+1} - \mathbf{w}^*\|_{\mathbf{H}}^2 \\ &\geq \|\mathbf{x}^t - \mathbf{x}^{t+1}\|_{\mathbf{G}}^2 + \frac{2-\alpha}{\alpha^2} \left(\rho \|\mathbf{B}\left(\mathbf{y}^t - \mathbf{y}^{t+1}\right)\|^2 + \frac{1}{\rho}\|\gamma^t - \gamma^{t+1}\|^2\right), \end{aligned}$$

and the assertion of this theorem follows directly. □

Remark 1 Since the matrix $\mathbf{H}$ defined in (11) is positive definite, the assertion (26) implies that the sequence $\{\mathbf{w}^t\}$ generated by the L-GADMM (4) is contractive with respect to Ω^* (according to the definition in [4]). Thus, the convergence of $\{\mathbf{w}^t\}$ can be trivially derived by applying the standard technique of contraction methods.

Remark 2 For the special case where $\alpha = 1$, i.e., the L-GADMM (4) reduces to the split inexact Uzawa method in [52,53], due to the structure of the matrix $\mathbf{H}$ [see (11)], the inequality (26) is simplified to

$$\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_{\mathbf{H}}^2 \leq \|\mathbf{w}^t - \mathbf{w}^*\|_{\mathbf{H}}^2 - \|\mathbf{w}^t - \mathbf{w}^{t+1}\|_{\mathbf{H}}^2, \quad \forall\, \mathbf{w}^* \in \Omega^*.$$

Moreover, when $\alpha = 1$ and $\mathbf{G} = \mathbf{0}$, i.e., the L-GADMM (4) reduces to the ADMM (2), the inequality (26) becomes

$$\|\mathbf{v}^{t+1} - \mathbf{v}^*\|_{\mathbf{H}_0}^2 \leq \|\mathbf{v}^t - \mathbf{v}^*\|_{\mathbf{H}_0}^2 - \|\mathbf{v}^t - \mathbf{v}^{t+1}\|_{\mathbf{H}_0}^2, \quad \forall\, \mathbf{v}^* \in \mathcal{V}^*,$$

where $\mathbf{v}$ and $\mathbf{H}_0$ are defined in (27). This is exactly the contraction property of the ADMM (2) analyzed in the appendix of [6].

Now, we are ready to establish a worst-case $\mathcal{O}(1/k)$ convergence rate in the ergodic sense for the L-GADMM (4). Lemma 2 plays an important role in the proof.

Theorem 2 *Let* $\mathbf{H}$ *be given by* (11) *and* $\{\mathbf{w}^t\}$ *be the sequence generated by the L-GADMM* (4). *For any integer* $k > 0$, *let* $\widehat{\mathbf{w}}_k$ *be defined by*

$$\widehat{\mathbf{w}}_k = \frac{1}{k+1} \sum_{t=0}^{k} \widetilde{\mathbf{w}}^t, \tag{32}$$

where $\widetilde{\mathbf{w}}^t$ *is defined in* (7). *Then, we have* $\widehat{\mathbf{w}}_k \in \Omega$ *and*

$$f(\widehat{\mathbf{u}}_k) - f(\mathbf{u}) + (\widehat{\mathbf{w}}_k - \mathbf{w})^T F(\mathbf{w}) \leq \frac{1}{2(k+1)} \|\mathbf{w} - \mathbf{w}^0\|_{\mathbf{H}}^2, \quad \forall \mathbf{w} \in \Omega.$$

Proof First, because of (7) and $\mathbf{w}^t \in \Omega$, it holds that $\widetilde{\mathbf{w}}^t \in \Omega$ for all $t \geq 0$. Thus, together with the convexity of $\mathcal{X}$ and $\mathcal{Y}$, (84) implies that $\widehat{\mathbf{w}}_k \in \Omega$. Second, due to the monotonicity, we have

$$\left(\mathbf{w} - \widetilde{\mathbf{w}}^t\right)^T F(\mathbf{w}) \geq \left(\mathbf{w} - \widetilde{\mathbf{w}}^t\right)^T F\left(\widetilde{\mathbf{w}}^t\right), \tag{33}$$

thus Lemmas 1 and 2 imply that for all $\mathbf{w} \in \Omega$,

$$f(\mathbf{u}) - f(\widetilde{\mathbf{u}}^t) + (\mathbf{w} - \widetilde{\mathbf{w}}^t)^T F(\mathbf{w}) + \frac{1}{2}\|\mathbf{w} - \mathbf{w}^t\|_{\mathbf{H}}^2 \geq \frac{1}{2}\|\mathbf{w} - \mathbf{w}^{t+1}\|_{\mathbf{H}}^2. \tag{34}$$

Summing the inequality (34) over $t = 0, 1, \ldots, k$, we obtain that for all $\mathbf{w} \in \Omega$

$$(k+1) f(\mathbf{u}) - \sum_{t=0}^{k} f(\widetilde{\mathbf{u}}^t) + \left((k+1)\mathbf{w} - \sum_{t=0}^{k} \widetilde{\mathbf{w}}^t\right)^T F(\mathbf{w}) + \frac{1}{2}\|\mathbf{w} - \mathbf{w}^0\|_{\mathbf{H}}^2 \geq 0.$$

Using the notation of $\widehat{\mathbf{w}}_t$, it can be written as for all $\mathbf{w} \in \Omega$

$$\frac{1}{k+1}\sum_{t=0}^{k} f(\widetilde{\mathbf{u}}^t) - f(\mathbf{u}) + (\widehat{\mathbf{w}}_k - \mathbf{w})^T F(\mathbf{w}) \leq \frac{1}{2(k+1)}\|\mathbf{w} - \mathbf{w}^0\|_{\mathbf{H}}^2. \tag{35}$$

Since $f(\mathbf{u})$ is convex and

$$\widehat{\mathbf{u}}_k = \frac{1}{k+1}\sum_{t=0}^{k} \widetilde{\mathbf{u}}^t,$$

we have that

$$f(\widehat{\mathbf{u}}_k) \leq \frac{1}{k+1}\sum_{t=0}^{k} f(\widetilde{\mathbf{u}}^t).$$

Substituting it into (35), the assertion of this theorem follows directly. □

For an arbitrary substantial compact set $\mathcal{D} \subset \Omega$, we define

$$D = \sup\left\{\|\mathbf{w} - \mathbf{w}^0\|_{\mathbf{H}} \mid \mathbf{w} \in \mathcal{D}\right\},$$

where $\mathbf{w}^0 = (\mathbf{x}^0, \mathbf{y}^0, \gamma^0)$ is the initial point. After k iterations of the L-GADMM (4), we can find a $\widehat{\mathbf{w}}_k \in \Omega$ such that

$$\sup_{\mathbf{w}\in\mathcal{D}}\left\{f(\widehat{\mathbf{u}}_k) - f(\mathbf{u}) + (\widehat{\mathbf{w}}_k - \mathbf{w})^T F(\mathbf{w})\right\} \leq \frac{D^2}{2k}.$$

A worse-case $\mathcal{O}(1/k)$ convergence rate in the ergodic sense is thus proved for the L-GADMM (4).

4 A worst-case $\mathcal{O}(1/k)$ convergence rate in a nonergodic sense

In this section, we prove a worst-case $\mathcal{O}(1/k)$ convergence rate in a nonergodic sense for the L-GADMM (4). The result includes the assertions in [32] for the ADMM and its linearized version as special cases.

For the right-hand-side of the inequality (21) in Lemma 2, the first term $\frac{1}{2}(\|\mathbf{w}-\mathbf{w}^{t+1}\|_{\mathbf{H}}^2-\|\mathbf{w}-\mathbf{w}^{t}\|_{\mathbf{H}}^2)$ is already in the form of the difference to $\mathbf{w}$ of two consecutive iterates, which is ideal for performing recursively algebraic reasoning in the proof of convergence rate (see theorems later). Now we have to deal with the last two terms in the right-hand side of (21) for the same purpose. In the following lemma, we find a bound in the term of $\|\mathbf{w}^t-\mathbf{w}^{t+1}\|_{\mathbf{H}}^2$ for the sum of these two terms.

Lemma 3 *Let the sequence $\{\mathbf{w}^t\}$ be generated by the L-GADMM (4) and the associated sequence $\{\widetilde{\mathbf{w}}^t\}$ be defined in (7). Then we have*

$$\|\mathbf{x}^t-\widetilde{\mathbf{x}}^t\|_{\mathbf{G}}^2+\frac{2-\alpha}{\rho}\|\gamma^t-\widetilde{\gamma}^t\|^2\geq c_\alpha\|\mathbf{w}^t-\mathbf{w}^{t+1}\|_{\mathbf{H}}^2, \tag{36}$$

where

$$c_\alpha=\min\left\{\frac{2-\alpha}{\alpha},1\right\}>0. \tag{37}$$

Proof Similar to (24), it follows from $\widetilde{\mathbf{x}}^t=\mathbf{x}^{t+1}$ and the definition of $\mathbf{H}$ that

$$\begin{aligned}\|\mathbf{w}^t-\mathbf{w}^{t+1}\|_{\mathbf{H}}^2=\|\mathbf{x}^t-\widetilde{\mathbf{x}}^t\|_{\mathbf{G}}^2+\frac{1}{\alpha\rho}\Big(&\|\rho\mathbf{B}\left(\mathbf{y}^t-\mathbf{y}^{t+1}\right)\|^2+\|\gamma^t-\gamma^{t+1}\|^2\\&+2(1-\alpha)\rho\left(\mathbf{y}^t-\mathbf{y}^{t+1}\right)^T\mathbf{B}^T\left(\gamma^t-\gamma^{t+1}\right)\Big).\end{aligned} \tag{38}$$

From (31, 8) and the assumption $\alpha\in(0,2)$, we have

$$\begin{aligned}&\left\|\rho\mathbf{B}\left(\mathbf{y}^t-\mathbf{y}^{t+1}\right)\right\|^2+\|\gamma^t-\gamma^{t+1}\|^2+2(1-\alpha)\rho\left(\mathbf{y}^t-\mathbf{y}^{t+1}\right)^T\mathbf{B}^T\left(\gamma^t-\gamma^{t+1}\right)\\&\leq\left\|\rho\mathbf{B}\left(\mathbf{y}^t-\mathbf{y}^{t+1}\right)\right\|^2+\|\gamma^t-\gamma^{t+1}\|^2+2\rho\left(\mathbf{y}^t-\mathbf{y}^{t+1}\right)^T\mathbf{B}^T\left(\gamma^t-\gamma^{t+1}\right)\\&=\left\|\rho\mathbf{B}(\mathbf{y}^t-\mathbf{y}^{t+1})+\left(\gamma^t-\gamma^{t+1}\right)\right\|^2\\&=\alpha^2\|\gamma^t-\widetilde{\gamma}^t\|^2.\end{aligned} \tag{39}$$

Using (38) and (39), we have

$$\|\mathbf{w}^t-\mathbf{w}^{t+1}\|_{\mathbf{H}}^2\leq\|\mathbf{x}^t-\widetilde{\mathbf{x}}^t\|_{\mathbf{G}}^2+\frac{\alpha}{\rho}\|\gamma^t-\widetilde{\gamma}^t\|^2.$$

Since $\alpha \in (0, 2)$, we get $(2-\alpha)/\alpha > 0$, $1 \geq c_\alpha > 0$ and

$$\begin{aligned}
&\|\mathbf{x}^t - \widetilde{\mathbf{x}}^t\|_{\mathbf{G}}^2 + \frac{2-\alpha}{\rho}\|\gamma^t - \widetilde{\gamma}^t\|^2 \\
&\geq \min\left\{\frac{2-\alpha}{\alpha}, 1\right\}\left(\|\mathbf{x}^t - \widetilde{\mathbf{x}}^t\|_{\mathbf{G}}^2 + \frac{\alpha}{\rho}\|\gamma^t - \widetilde{\gamma}^t\|^2\right) \\
&\geq c_\alpha \|\mathbf{w}^t - \mathbf{w}^{t+1}\|_{\mathbf{H}}^2.
\end{aligned}$$

The proof is completed. □

With Lemmas 1, 2 and 3, we can find a bound of the accuracy of $\widetilde{\mathbf{w}}^t$ to a solution point of VI(Ω, F, f) in term of some quadratic terms. We show it in the next theorem.

Theorem 3 *Let the sequence* $\{\mathbf{w}^t\}$ *be generated by the L-GADMM* (4) *and the associated sequence* $\{\widetilde{\mathbf{w}}^t\}$ *be defined in* (7)*. Then for any* $\mathbf{w} \in \Omega$*, we have*

$$\begin{aligned}
&f(\mathbf{u}) - f(\widetilde{\mathbf{u}}^t) + \left(\mathbf{w} - \widetilde{\mathbf{w}}^t\right)^T F(\mathbf{w}) \\
&\geq \frac{1}{2}\left(\|\mathbf{w} - \mathbf{w}^{t+1}\|_{\mathbf{H}}^2 - \|\mathbf{w} - \mathbf{w}^t\|_{\mathbf{H}}^2\right) + \frac{c_\alpha}{2}\|\mathbf{w}^t - \mathbf{w}^{t+1}\|_{\mathbf{H}}^2,
\end{aligned} \tag{40}$$

where $\mathbf{H}$ *and* $c_\alpha > 0$ *are defined in* (11) *and* (37)*, respectively.*

Proof Using the monotonicity of $F(\mathbf{w})$ [see (33)] and replacing the right-hand side term in (13) with the equality (21), we obtain that

$$\begin{aligned}
&f(\mathbf{u}) - f(\widetilde{\mathbf{u}}^t) + \left(\mathbf{w} - \widetilde{\mathbf{w}}^t\right)^T F(\mathbf{w}) \\
&\geq \frac{1}{2}\left(\|\mathbf{w} - \mathbf{w}^{t+1}\|_{\mathbf{H}}^2 - \|\mathbf{w} - \mathbf{w}^t\|_{\mathbf{H}}^2\right) + \frac{1}{2}\left(\|\mathbf{x}^t - \widetilde{\mathbf{x}}^t\|_{\mathbf{G}}^2 + \frac{2-\alpha}{\rho}\|\gamma^t - \widetilde{\gamma}^t\|^2\right).
\end{aligned}$$

The assertion (40) follows by plugging (36) into the above inequality immediately. □

Now we show an important inequality for the scheme (4) by using Lemmas 1, 2 and 3. This inequality immediately shows the contraction of the sequence generated by (4), and based on this inequality we can establish its worst-case $\mathcal{O}(1/k)$ convergence rate in a nonergodic sense.

Theorem 4 *The sequence* $\{\mathbf{w}^t\}$ *generated by the L-GADMM* (4) *satisfies*

$$\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_{\mathbf{H}}^2 \leq \|\mathbf{w}^t - \mathbf{w}^*\|_{\mathbf{H}}^2 - c_\alpha\|\mathbf{w}^t - \mathbf{w}^{t+1}\|_{\mathbf{H}}^2, \quad \forall \mathbf{w}^* \in \Omega^*, \tag{41}$$

where $\mathbf{H}$ *and* $c_\alpha > 0$ *are defined in* (11) *and* (37)*, respectively.*

Proof Setting $\mathbf{w} = \mathbf{w}^*$ in (40), we get

$$\begin{aligned}
&f(\mathbf{u}^*) - f(\widetilde{\mathbf{u}}^t) + (\mathbf{w}^* - \widetilde{\mathbf{w}}^t)^T F(\mathbf{w}^*) \\
&\geq \frac{1}{2}\left(\|\mathbf{w}^* - \mathbf{w}^{t+1}\|_{\mathbf{H}}^2 - \|\mathbf{w}^* - \mathbf{w}^t\|_{\mathbf{H}}^2\right) + \frac{c_\alpha}{2}\|\mathbf{w}^t - \mathbf{w}^{t+1}\|_{\mathbf{H}}^2.
\end{aligned}$$

On the other hand, since $\widetilde{\mathbf{w}}^t \in \Omega$, and $\mathbf{w}^* \in \Omega^*$, we have

$$0 \geq f(\mathbf{u}^*) - f(\widetilde{\mathbf{u}}^t) + \left(\mathbf{w}^* - \widetilde{\mathbf{w}}^t\right)^T F(\mathbf{w}^*).$$

From the above two inequalities, the assertion (41) is proved. □

Remark 3 Theorem 4 also explains why the relaxation parameter α is restricted into the interval $(0, 2)$ for the L-GADMM (4). In fact, if $\alpha \leq 0$ or $\alpha \geq 2$, then the constant c_α defined in (37) is only non-positive and the inequality (41) does not suffice to ensure the strict contraction of the sequence generated by (4); it is thus difficult to establish its convergence. We will empirically verify the failure of convergence for the case of (4) with $\alpha = 2$ in Sect. 5.

To establish a worst-case $\mathcal{O}(1/k)$ convergence rate in a nonergodic sense for the scheme (4), we first have to mention that the term $\|\mathbf{w}^t - \mathbf{w}^{t+1}\|_{\mathbf{H}}^2$ can be used to measure the accuracy of an iterate. This result has been proved in some literatures such as [32] for the original ADMM.

Lemma 4 *For a given $\mathbf{w}^t$, let $\mathbf{w}^{t+1}$ be generated by the L-GADMM (4). When $\|\mathbf{w}^t - \mathbf{w}^{t+1}\|_{\mathbf{H}}^2 = 0$, $\widetilde{\mathbf{w}}^t$ defined in (7) is a solution to (5).*

Proof By Lemma 1 and (22), it implies that for all $\mathbf{w} \in \Omega$,

$$f(\mathbf{u}) - f\left(\widetilde{\mathbf{u}}^t\right) + \left(\mathbf{w} - \widetilde{\mathbf{w}}^t\right)^T F\left(\widetilde{\mathbf{w}}^t\right) \geq \left(\mathbf{w} - \widetilde{\mathbf{w}}^t\right)^T \mathbf{H}\left(\mathbf{w}^t - \mathbf{w}^{t+1}\right). \tag{42}$$

As $\mathbf{H}$ is positive definite, the right-hand side of (42) vanishes if $\|\mathbf{w}^{t+1} - \mathbf{w}^t\|_{\mathbf{H}}^2 = 0$, since we have $\mathbf{H}(\mathbf{w}^{t+1} - \mathbf{w}^t) = 0$ whenever $\|\mathbf{w}^{t+1} - \mathbf{w}^t\|_{\mathbf{H}}^2 = 0$. The assertion is proved. □

Now, we are ready to establish a worst-case $\mathcal{O}(1/k)$ convergence rate in a nonergodic sense for the scheme (4). First, we prove some lemmas.

Lemma 5 *Let the sequence $\{\mathbf{w}^t\}$ be generated by the L-GADMM (4) and the associated $\{\widetilde{\mathbf{w}}^t\}$ be defined in (7); the matrix $\mathbf{Q}$ be defined in (11). Then, we have*

$$\left(\widetilde{\mathbf{w}}^t - \widetilde{\mathbf{w}}^{t+1}\right)^T \mathbf{Q}\left[\left(\mathbf{w}^t - \mathbf{w}^{t+1}\right) - \left(\widetilde{\mathbf{w}}^t - \widetilde{\mathbf{w}}^{t+1}\right)\right] \geq 0. \tag{43}$$

Proof Setting $\mathbf{w} = \widetilde{\mathbf{w}}^{t+1}$ in (13), we have

$$f\left(\widetilde{\mathbf{u}}^{t+1}\right) - f(\widetilde{\mathbf{u}}^t) + \left(\widetilde{\mathbf{w}}^{t+1} - \widetilde{\mathbf{w}}^t\right)^T F\left(\widetilde{\mathbf{w}}^t\right) \geq \left(\widetilde{\mathbf{w}}^{t+1} - \widetilde{\mathbf{w}}^t\right)^T \mathbf{Q}\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right). \tag{44}$$

Note that (13) is also true for $t := t + 1$ and thus for all $\mathbf{w} \in \Omega$,

$$f(\mathbf{u}) - f\left(\widetilde{\mathbf{u}}^{t+1}\right) + \left(\mathbf{w} - \widetilde{\mathbf{w}}^{t+1}\right)^T F\left(\widetilde{\mathbf{w}}^{t+1}\right) \geq \left(\mathbf{w} - \widetilde{\mathbf{w}}^{t+1}\right)^T \mathbf{Q}\left(\mathbf{w}^{t+1} - \widetilde{\mathbf{w}}^{t+1}\right).$$

Setting $\mathbf{w} = \widetilde{\mathbf{w}}^t$ in the above inequality, we obtain

$$f\left(\widetilde{\mathbf{u}}^t\right) - f\left(\widetilde{\mathbf{u}}^{t+1}\right) + \left(\widetilde{\mathbf{w}}^t - \widetilde{\mathbf{w}}^{t+1}\right)^T F\left(\widetilde{\mathbf{w}}^{t+1}\right) \geq \left(\widetilde{\mathbf{w}}^t - \widetilde{\mathbf{w}}^{t+1}\right)^T \mathbf{Q}\left(\mathbf{w}^{t+1} - \widetilde{\mathbf{w}}^{t+1}\right). \tag{45}$$

Adding (44) and (45) and using the monotonicity of F, we get (43) immediately. The proof is completed. □

Lemma 6 *Let the sequence* $\{\mathbf{w}^t\}$ *be generated by the L-GADMM* (4) *and the associated* $\{\widetilde{\mathbf{w}}^t\}$ *be defined in* (7)*; the matrices* $\mathbf{M}$, $\mathbf{H}$ *and* $\mathbf{Q}$ *be defined in* (10) *and* (11)*. Then, we have*

$$\begin{aligned} &\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right)^T \mathbf{M}^T \mathbf{H} \mathbf{M}\left[\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right) - \left(\mathbf{w}^{t+1} - \widetilde{\mathbf{w}}^{t+1}\right)\right] \\ &\quad \geq \frac{1}{2}\left\|\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right) - \left(\mathbf{w}^{t+1} - \widetilde{\mathbf{w}}^{t+1}\right)\right\|^2_{(\mathbf{Q}^T + \mathbf{Q})}. \end{aligned} \tag{46}$$

Proof Adding the equation

$$\begin{aligned} &\left[\left(\mathbf{w}^t - \mathbf{w}^{t+1}\right) - \left(\widetilde{\mathbf{w}}^t - \widetilde{\mathbf{w}}^{t+1}\right)\right]^T \mathbf{Q}\left[\left(\mathbf{w}^t - \mathbf{w}^{t+1}\right) - \left(\widetilde{\mathbf{w}}^t - \widetilde{\mathbf{w}}^{t+1}\right)\right] \\ &= \frac{1}{2}\left\|\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right) - \left(\mathbf{w}^{t+1} - \widetilde{\mathbf{w}}^{t+1}\right)\right\|^2_{(\mathbf{Q}^T + \mathbf{Q})} \end{aligned}$$

to both sides of (43), we get

$$\begin{aligned} &\left(\mathbf{w}^t - \mathbf{w}^{t+1}\right)^T \mathbf{Q}\left[\left(\mathbf{w}^t - \mathbf{w}^{t+1}\right) - \left(\widetilde{\mathbf{w}}^t - \widetilde{\mathbf{w}}^{t+1}\right)\right] \\ &\quad \geq \frac{1}{2}\left\|\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right) - \left(\mathbf{w}^{t+1} - \widetilde{\mathbf{w}}^{t+1}\right)\right\|^2_{(\mathbf{Q}^T + \mathbf{Q})}. \end{aligned} \tag{47}$$

Using [see (9)]

$$\mathbf{w}^t - \mathbf{w}^{t+1} = \mathbf{M}\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right) \qquad \text{and} \qquad \mathbf{Q} = \mathbf{H}\mathbf{M},$$

to the term $\mathbf{w}^t - \mathbf{w}^{t+1}$ in the left-hand side of (47), we obtain

$$\begin{aligned} &\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right)^T \mathbf{M}^T \mathbf{H} \mathbf{M}\left[\left(\mathbf{w}^t - \mathbf{w}^{t+1}\right) - \left(\widetilde{\mathbf{w}}^t - \widetilde{\mathbf{w}}^{t+1}\right)\right] \\ &\quad \geq \frac{1}{2}\left\|\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right) - \left(\mathbf{w}^{t+1} - \widetilde{\mathbf{w}}^{t+1}\right)\right\|^2_{(\mathbf{Q}^T + \mathbf{Q})}, \end{aligned}$$

and the lemma is proved. □

We then prove that the sequence $\{\|\mathbf{w}^t - \mathbf{w}^{t+1}\|_{\mathbf{H}}\}$ is monotonically non-increasing.

Theorem 5 *Let the sequence* $\{\mathbf{w}^t\}$ *be generated by the L-GADMM* (4) *and the matrix* $\mathbf{H}$ *be defined in* (11)*. Then, we have*

$$\|\mathbf{w}^{t+1} - \mathbf{w}^{t+2}\|_{\mathbf{H}}^2 \leq \|\mathbf{w}^t - \mathbf{w}^{t+1}\|_{\mathbf{H}}^2. \tag{48}$$

Proof Setting $\mathbf{a} = \mathbf{M}(\mathbf{w}^t - \widetilde{\mathbf{w}}^t)$ and $\mathbf{b} = \mathbf{M}(\mathbf{w}^{t+1} - \widetilde{\mathbf{w}}^{t+1})$ in the identity

$$\|\mathbf{a}\|_{\mathbf{H}}^2 - \|\mathbf{b}\|_{\mathbf{H}}^2 = 2\mathbf{a}^T\mathbf{H}(\mathbf{a} - \mathbf{b}) - \|\mathbf{a} - \mathbf{b}\|_{\mathbf{H}}^2,$$

we obtain that

$$\begin{aligned}
&\|\mathbf{M}\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right)\|_{\mathbf{H}}^2 - \left\|\mathbf{M}\left(\mathbf{w}^{t+1} - \widetilde{\mathbf{w}}^{t+1}\right)\right\|_{\mathbf{H}}^2 \\
&= 2\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right)\mathbf{M}^T\mathbf{H}\mathbf{M}\left[\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right) - \left(\mathbf{w}^{t+1} - \widetilde{\mathbf{w}}^{t+1}\right)\right] \\
&\quad - \left\|\mathbf{M}\left[\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right) - \left(\mathbf{w}^{t+1} - \widetilde{\mathbf{w}}^{t+1}\right)\right]\right\|_{\mathbf{H}}^2.
\end{aligned} \tag{49}$$

Inserting (46) into the first term of the right-hand side of (49), we obtain that

$$\begin{aligned}
&\|\mathbf{M}\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right)\|_{\mathbf{H}}^2 - \|\mathbf{M}\left(\mathbf{w}^{t+1} - \widetilde{\mathbf{w}}^{t+1}\right)\|_{\mathbf{H}}^2 \\
&\geq \left\|\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right) - \left(\mathbf{w}^{t+1} - \widetilde{\mathbf{w}}^{t+1}\right)\right\|_{(\mathbf{Q}^T + \mathbf{Q})}^2 \\
&\quad - \left\|\mathbf{M}\left[\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right) - \left(\mathbf{w}^{t+1} - \widetilde{\mathbf{w}}^{t+1}\right)\right]\right\|_{\mathbf{H}}^2 \\
&= \left\|\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right) - \left(\mathbf{w}^{t+1} - \widetilde{\mathbf{w}}^{t+1}\right)\right\|_{(\mathbf{Q}^T + \mathbf{Q}) - \mathbf{M}^T\mathbf{H}\mathbf{M}}^2 \\
&\geq 0,
\end{aligned} \tag{50}$$

where the last inequality is by the fact that $(\mathbf{Q}^T + \mathbf{Q}) - \mathbf{M}^T\mathbf{H}\mathbf{M}$ is positive definite. This is derived from the following calculation:

$$\begin{aligned}
&\left(\mathbf{Q}^T + \mathbf{Q}\right) - \mathbf{M}^T\mathbf{H}\mathbf{M} = \left(\mathbf{Q}^T + \mathbf{Q}\right) - \mathbf{M}^T\mathbf{Q} \\
&= \begin{pmatrix} 2\mathbf{G} & 0 & 0 \\ 0 & 2\rho\mathbf{B}^T\mathbf{B} & -\alpha\mathbf{B}^T \\ 0 & -\alpha\mathbf{B} & \frac{2}{\rho}\mathbf{I}_n \end{pmatrix} - \begin{pmatrix} \mathbf{I}_{n_1} & 0 & 0 \\ 0 & \mathbf{I}_{n_2} & -\rho\mathbf{B}^T \\ 0 & 0 & \alpha\mathbf{I}_n \end{pmatrix} \begin{pmatrix} \mathbf{G} & 0 & 0 \\ 0 & \rho\mathbf{B}^T\mathbf{B} & (1-\alpha)\mathbf{B}^T \\ 0 & -\mathbf{B} & \frac{1}{\rho}\mathbf{I}_n \end{pmatrix} \\
&= \begin{pmatrix} 2\mathbf{G} & 0 & 0 \\ 0 & 2\rho\mathbf{B}^T\mathbf{B} & -\alpha\mathbf{B}^T \\ 0 & -\alpha\mathbf{B} & \frac{2}{\rho}\mathbf{I}_n \end{pmatrix} - \begin{pmatrix} \mathbf{G} & 0 & 0 \\ 0 & 2\rho\mathbf{B}^T\mathbf{B} & -\alpha\mathbf{B}^T \\ 0 & -\alpha\mathbf{B} & \frac{\alpha}{\rho}\mathbf{I}_n \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{G} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{2-\alpha}{\rho}\mathbf{I}_n \end{pmatrix}.
\end{aligned}$$

As we have assumed that $\mathbf{G}$ is positive definite. It is clear that $(\mathbf{Q}^T + \mathbf{Q}) - \mathbf{M}^T\mathbf{H}\mathbf{M}$ is positive definite. By (50), we have shown that

$$\left\|\mathbf{M}\left(\mathbf{w}^{t+1} - \widetilde{\mathbf{w}}^{t+1}\right)\right\|_{\mathbf{H}}^2 \leq \|\mathbf{M}\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right)\|_{\mathbf{H}}^2.$$

Recall the fact that $\mathbf{w}^t - \mathbf{w}^{t+1} = \mathbf{M}(\mathbf{w}^t - \widetilde{\mathbf{w}}^t)$. The assertion (48) follows immediately. □

We now prove the worst-case $\mathcal{O}(1/k)$ convergence rate in a nonergodic sense for the L-GADMM (4).

Theorem 6 *Let* $\{\mathbf{w}^t\}$ *be the sequence generated by the L-GADMM* (4) *with* $\alpha \in (0, 2)$. *Then we have*

$$\|\mathbf{w}^k - \mathbf{w}^{k+1}\|_{\mathbf{H}}^2 \leq \frac{1}{(k+1)c_\alpha}\|\mathbf{w}^0 - \mathbf{w}^*\|_{\mathbf{H}}^2, \quad \forall k \geq 0, \mathbf{w}^* \in \Omega^*, \tag{51}$$

where $\mathbf{H}$ *and* $c_\alpha > 0$ *are defined in* (11) *and* (37), *respectively.*

Proof It follows from (41) that

$$c_\alpha \sum_{t=0}^{k} \|\mathbf{w}^t - \mathbf{w}^{t+1}\|_{\mathbf{H}}^2 \leq \|\mathbf{w}^0 - \mathbf{w}^*\|_{\mathbf{H}}^2, \quad \forall \mathbf{w}^* \in \Omega^*. \tag{52}$$

As shown in Theorem 5, $\{\|\mathbf{w}^t - \mathbf{w}^{t+1}\|_{\mathbf{H}}^2\}$ is non-increasing. Therefore, we have

$$(k+1)\|\mathbf{w}^k - \mathbf{w}^{k+1}\|_{\mathbf{H}}^2 \leq \sum_{t=0}^{k} \|\mathbf{w}^t - \mathbf{w}^{t+1}\|_{\mathbf{H}}^2. \tag{53}$$

The assertion (51) follows from (52) and (53) immediately. □

Notice that Ω^* is convex and closed. Let $\mathbf{w}^0$ be the initial iterate and $d := \inf\{\|\mathbf{w}^0 - \mathbf{w}^*\|_{\mathbf{H}} \mid \mathbf{w}^* \in \Omega^*\}$. Then, for any given $\epsilon > 0$, Theorem 6 shows that the L-GADMM (4) needs at most $\lceil d^2/(c_\alpha\epsilon) - 1\rceil$ iterations to ensure that $\|\mathbf{w}^k - \mathbf{w}^{k+1}\|_{\mathbf{H}}^2 \leq \epsilon$. It follows from Lemma 4 that $\widetilde{\mathbf{w}}^k$ is a solution of VI(Ω, F, f) if $\|\mathbf{w}^k - \mathbf{w}^{k+1}\|_{\mathbf{H}}^2 = 0$. A worst-case $\mathcal{O}(1/k)$ convergence rate in a nonergodic sense for the L-GADMM (4) is thus established in Theorem 6.

5 Numerical experiments

In the literature, the efficiency of both the original ADMM (2) and its linearized version has been very well illustrated. The emphasis of this section is to further verify the acceleration performance of the L-GADMM (4) over the linearized version of the original ADMM. That is, we want to further show that for the scheme (4), the general

case with $\alpha \in (1, 2)$ could lead to better numerical result than the special case with $\alpha = 1$. The examples we will test include some rather new and core applications in statistical learning area. All codes were written in Matlab 2012a, and all experiments were run on a Macbook Pro with an Intel 2.9 GHz i7 Processor and 16 GB Memory.

5.1 Sparse linear discriminant analysis

In this and the upcoming subsection, we apply the scheme (4) to solve two rather new and challenging statistical learning models, i.e. the linear programming discriminant rule in [7] and the constrained LASSO model in [35]. The efficiency of the scheme (4) will be verified. In particular, the acceleration performance of (4) with $\alpha \in (1, 2)$ will be demonstrated.

We consider the problem of binary classification. Let $\{(\mathbf{x}_i, y_i), i = 1, ..., n\}$ be n samples drawn from a joint distribution of $(\boldsymbol{X}, Y) \in \mathbb{R}^d \times \{0, 1\}$. Given a new data $\mathbf{x} \in \mathbb{R}^d$, the goal of the problem is to determine the associated value of Y. A common way to solve this problem is the *linear discriminant analysis* (LDA), see e.g. [1]. Assuming Gaussian conditional distributions with a common covariance matrix, i.e., $(\boldsymbol{X}|Y = 0) \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ and $(\boldsymbol{X}|Y = 1) \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, let the prior probabilities be $\pi_0 = \mathbb{P}(Y = 0)$ and $\pi_1 = \mathbb{P}(Y = 1)$. Denote $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ as the precision matrix; $\boldsymbol{\mu}_a = (\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)/2$ and $\boldsymbol{\mu}_d = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$. Under the normality assumption and when $\pi_0 = \pi_1 = 0.5$, by Bayes' rule, we have a linear classifier

$$\ell(\mathbf{x}; \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) := \begin{cases} 1 & \text{if } (\mathbf{x} - \boldsymbol{\mu}_a)^T \boldsymbol{\Omega} \boldsymbol{\mu}_d > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{54}$$

As the covariance matrix $\boldsymbol{\Sigma}$ and the means $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$ are unknown, given n_0 samples of $(\boldsymbol{X}|Y = 0)$ and n_1 samples of $(\boldsymbol{X}|Y = 1)$ respectively, a natural way to build a linear classifier is plugging sample means $\widehat{\boldsymbol{\mu}}_0$, $\widehat{\boldsymbol{\mu}}_1$ and sample inverse covariance matrix $\widehat{\boldsymbol{\Sigma}}^{-1}$ into (54) to have $\ell(\mathbf{x}; \widehat{\boldsymbol{\mu}}_0, \widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\Sigma}})$. However, in the high-dimensional case, where $d > n$, plug-in rule does not work as $\widehat{\boldsymbol{\Sigma}}$ is not of full rank.

To resolve this problem, we consider the linear programming discriminant (LPD) rule proposed in [7] which assumes the sparsity directly on the discriminant direction $\boldsymbol{\beta} = \boldsymbol{\Omega}\boldsymbol{\mu}_d$ instead of $\boldsymbol{\mu}_d$ or $\boldsymbol{\Omega}$, which is formulated as following: We first estimate sample mean and sample covariance matrix $\widehat{\boldsymbol{\mu}}_0$, $\widehat{\boldsymbol{\Sigma}}_0$ of $(\boldsymbol{X}|Y = 0)$ and $\widehat{\boldsymbol{\mu}}_1$, $\widehat{\boldsymbol{\Sigma}}_1$ of $(\boldsymbol{X}|Y = 1)$ respectively. Let $\widehat{\boldsymbol{\Sigma}} = \frac{n_0}{n}\widehat{\boldsymbol{\Sigma}}_0 + \frac{n_1}{n}\widehat{\boldsymbol{\Sigma}}_1$, where n_0, n_1 are the sample sizes of $(\boldsymbol{X}|Y = 0)$ and $(\boldsymbol{X}|Y = 1)$ respectively, and $n = n_0 + n_1$. The LPD model proposed in [7] is

$$\begin{aligned} &\min \|\boldsymbol{\beta}\|_1 \\ &\text{s.t.} \quad \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \widehat{\boldsymbol{\mu}}_d\|_\infty \le \lambda, \end{aligned} \tag{55}$$

where λ is a tuning parameter; $\widehat{\boldsymbol{\mu}}_d = \widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0$, and $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are the ℓ_1 and ℓ_∞ norms in Euclidean space, respectively.

Clearly, the model (55) is essentially a linear programming problem, thus interior point methods are applicable. See e.g., [7] for a primal-dual interior point method. However, it is well known that interior point methods are not efficient for large-scale problems because the involved systems of equations which are solved by Newton type methods are too computationally demanding for large-scale cases.

Here, we apply the L-GADMM (4) to solve (55). To use (4), we first reformulate (55) as

$$\begin{aligned} &\min \|\boldsymbol{\beta}\|_1 \\ &\text{s.t.} \quad \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \widehat{\boldsymbol{\mu}}_d = \mathbf{y}, \\ &\qquad \mathbf{y} \in \mathcal{Y} := \{\mathbf{y} : \|\mathbf{y}\|_\infty \le \lambda\}, \end{aligned} \tag{56}$$

which is a special case of the model (1), and thus the scheme (4) is applicable.

We then elaborate on the resulting subproblems when (4) is applied to (56). First, let us see the application of the original GADMM scheme (3) without linearization to (56):

$$\boldsymbol{\beta}^{t+1} = \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ \|\boldsymbol{\beta}\|_1 + \frac{\rho}{2} \left\| \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \widehat{\boldsymbol{\mu}}_d - \mathbf{y}^t - \frac{\gamma^t}{\rho} \right\|^2 \right\}, \tag{57}$$

$$\mathbf{y}^{t+1} = \underset{\mathbf{y}\in\mathcal{Y}}{\operatorname{argmin}} \left\{ \frac{\rho}{2} \left\| \mathbf{y} - \left(\alpha\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^{t+1} + (1-\alpha)\left(\widehat{\boldsymbol{\mu}}_d + \mathbf{y}^t\right) - \widehat{\boldsymbol{\mu}}_d\right) + \frac{\gamma^t}{\rho} \right\|^2 \right\}, \tag{58}$$

$$\gamma^{t+1} = \gamma^t - \rho\left(\alpha\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^{t+1} + (1-\alpha)(\mathbf{y}^t + \widehat{\boldsymbol{\mu}}_d) - \widehat{\boldsymbol{\mu}}_d - \mathbf{y}^{t+1}\right). \tag{59}$$

For the $\boldsymbol{\beta}$-subproblem (57), since $\widehat{\boldsymbol{\Sigma}}$ is not a full rank matrix in the model (55) (in high-dimensional setting, the rank of $\widehat{\boldsymbol{\Sigma}}$ is much smaller than d), it has no closed-form solution. As described in (4), we choose $\mathbf{G} = \tau\mathbf{I}_d - \rho\widehat{\boldsymbol{\Sigma}}^T\widehat{\boldsymbol{\Sigma}}$ and consider the following linearized version of the $\boldsymbol{\beta}$-subproblem (57):

$$\boldsymbol{\beta}^{t+1} = \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ \|\boldsymbol{\beta}\|_1 + \rho(\boldsymbol{\beta} - \boldsymbol{\beta}^t)^T\mathbf{v}^t + \frac{\tau}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^t\|^2) \right\}, \tag{60}$$

where $\mathbf{v}^t := \widehat{\boldsymbol{\Sigma}}^T\left(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^t - \widehat{\boldsymbol{\mu}}_d - \mathbf{y}^t - \frac{\gamma^t}{\rho}\right)$ is the gradient of the quadratic term $\frac{1}{2}\|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \widehat{\boldsymbol{\mu}}_d - \mathbf{y}^t - \frac{\gamma^t}{\rho}\|^2$ at $\boldsymbol{\beta} = \boldsymbol{\beta}^t$. It is seen that (60) is equivalent to

$$\boldsymbol{\beta}^{t+1} = \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ \|\boldsymbol{\beta}\|_1 + \frac{\rho\tau}{4} \left\| \boldsymbol{\beta} - \left(\boldsymbol{\beta}^t - \frac{2}{\tau}\mathbf{v}^t\right) \right\|^2 \right\}. \tag{61}$$

Let $\text{shrinkage}(\mathbf{u}, \eta) := \text{sign}(\mathbf{u}) \cdot \max\left(0, |\mathbf{u}| - \eta\right)$ be the soft shrinkage operater, where $\text{sign}(\cdot)$ is the sign function. The closed-form solution of (61) is given by

$$\boldsymbol{\beta}^{t+1} = \text{shrinkage}\left(\boldsymbol{\beta}^t - \frac{2}{\tau}\mathbf{v}^t, \frac{2}{\rho\tau}\right). \tag{62}$$

For the $\mathbf{y}$-subproblem (58), by simple calculation, we get its solution is

$$\mathbf{y}^{t+1} = \min\left(\max\left(\alpha\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^{t+1} + (1-\alpha)(\mathbf{y}^t + \widehat{\boldsymbol{\mu}}_d) - \widehat{\boldsymbol{\mu}}_d - \frac{\gamma^t}{\rho}, -\lambda\right), \lambda\right). \tag{63}$$

Overall, the application of the L-GADMM (4) to the model (55) is summarized in Algorithm 1.

Algorithm 1 Solving sparse LPD (55) by the L-GADMM (4)

Initialize $\boldsymbol{\beta}^0$, $\mathbf{y}^0$, γ^0, τ, ρ.
while Stopping criterion is not satisfied. **do**
Compute $\boldsymbol{\beta}^{t+1}$ by (62).
Compute $\mathbf{y}^{t+1}$ by (63).
Update γ^{t+1} by (59)
end while

To implement Algorithm 1, we use the stopping criterion described in [6]: Let the primal residual at the t-th iteration be $r^t = \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^t - \widehat{\boldsymbol{\mu}}_d - \mathbf{y}^t\|$ and the dual residual be $s^t = \|\rho\widehat{\boldsymbol{\Sigma}}(\mathbf{y}^k - \mathbf{y}^{k-1})\|$; let the tolerance of the primal and dual residual at the t-th iteration be $\epsilon^{pri} = \sqrt{d}\epsilon + \epsilon \max\left(\|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^t\|, \|\mathbf{y}^t\|, \|\widehat{\boldsymbol{\mu}}_d\|\right)$ and $\epsilon^{dua} = \sqrt{d}\epsilon + \epsilon\|\widehat{\boldsymbol{\Sigma}}\gamma^t\|$, respectively, where ϵ is chosen differently for different applications; then the iteration is stopped when $r^t < \epsilon^{pri}$ and $s^t < \epsilon^{dua}$.

5.1.1 Simulated data

We first test some synthetic dataset. Following the settings in [7], we consider three schemes and for each scheme, we take $n_0 = n_1 = 150$, $d = 400$ and set $\boldsymbol{\mu}_0 = \mathbf{0}$, $\boldsymbol{\mu}_1 = (1, ..., 1, 0, ..., 0)^T$, $\lambda = 0.15$ for Scheme 1, and $\lambda = 0.1$ for Schemes 2 and 3, where the number of 1's in $\boldsymbol{\mu}_1$ is $s = 10$. The details of the schemes to be tested is listed below:

- [Scheme 1]: $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ where $\boldsymbol{\Sigma}_{jj} = 1$ for $1 \le j \le d$ and $\boldsymbol{\Sigma}_{jk} = 0.5$ for $j \ne k$.
- [Scheme 2]: $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, where $\boldsymbol{\Sigma}_{jk} = 0.6^{|j-k|}$ for $1 \le j, k \le d$.
- [Scheme 3]: $\boldsymbol{\Omega} = (\mathbf{B} + \delta\mathbf{I})/(1 + \delta)$, where $\mathbf{B} = (b_{jk})_{d\times d}$ with independent $b_{jk} = b_{kj} = 0.5 \times Ber(0.2)$ for $1 \le j, k \le s$, $i \ne j$; $b_{jk} = b_{kj} = 0.5$ for $s + 1 \le j < k \le d$; $b_{jj} = 1$ for $1 \le j \le d$, where $Ber(0.2)$ is a Bernoulli random variable whose value is taken as 1 with the probability 0.2 and 0 with the probability 0.8, and $\delta = \max(-\Lambda_{\min}(\mathbf{B}), 0) + 0.05$, where $\Lambda_{\min}(\mathbf{W})$ is the smallest eigenvalue of the matrix $\mathbf{W}$, to ensure the positive definiteness of $\boldsymbol{\Omega}$.

For Algorithm 1, we set the parameters $\rho = 0.05$ and $\tau = 2.1\|\widehat{\boldsymbol{\Sigma}}^T\widehat{\boldsymbol{\Sigma}}\|_2$. These values are tuned via experiments. The starting points are that $\boldsymbol{\beta}^0 = \mathbf{0}$, $\mathbf{y}^0 = \mathbf{0}$ and $\gamma^0 = \mathbf{1}$. We set $\epsilon = 5 \times 10^{-4}$ for the stopping criterion. Note that, as described by [7], we add $\delta\mathbf{I}_d$ to the sample covariance matrix to avoid the ill conditionness, where $\delta = 10^{-12}$.

Since synthetic dataset is considered, we repeat each scheme ten times and report the averaged numerical performance. In particular, we plot the evolutions of the number of iterations and computing time in seconds with respect to different values of α in the interval [1.0, 1.9] with an equal distance of 0.1, and we further choose finer grids in the interval [1.91, 1.98] with an equal distance of 0.01 in Fig. 1.[1] To see the performance of Algorithm 1 with $\alpha \in [1, 2)$ clearly, for the last simulation case of Scheme 2, we plot the evolutions of objective function value, primal residual and dual residual with

[1] As well known in [2,8,14,25], $\alpha \in (1, 2)$ usually results in acceleration for the GADMM. We thus do not report the numerical result when $\alpha \in (0, 1)$.

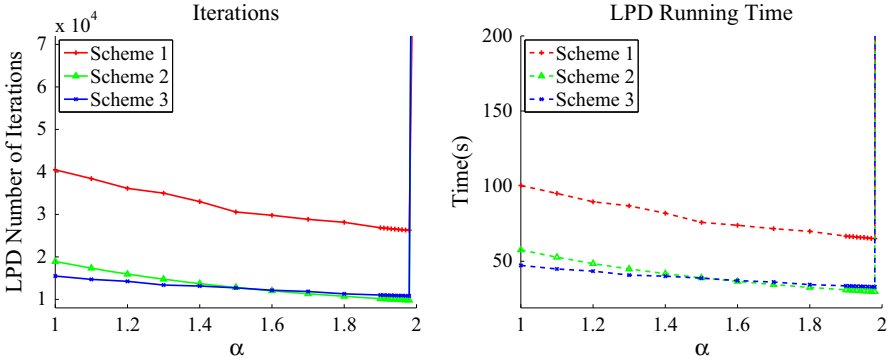

Fig. 1 Algorithm 1: evolution of number of iterations and computing time in seconds w.r.t. different values of α' for synthetic dataset.

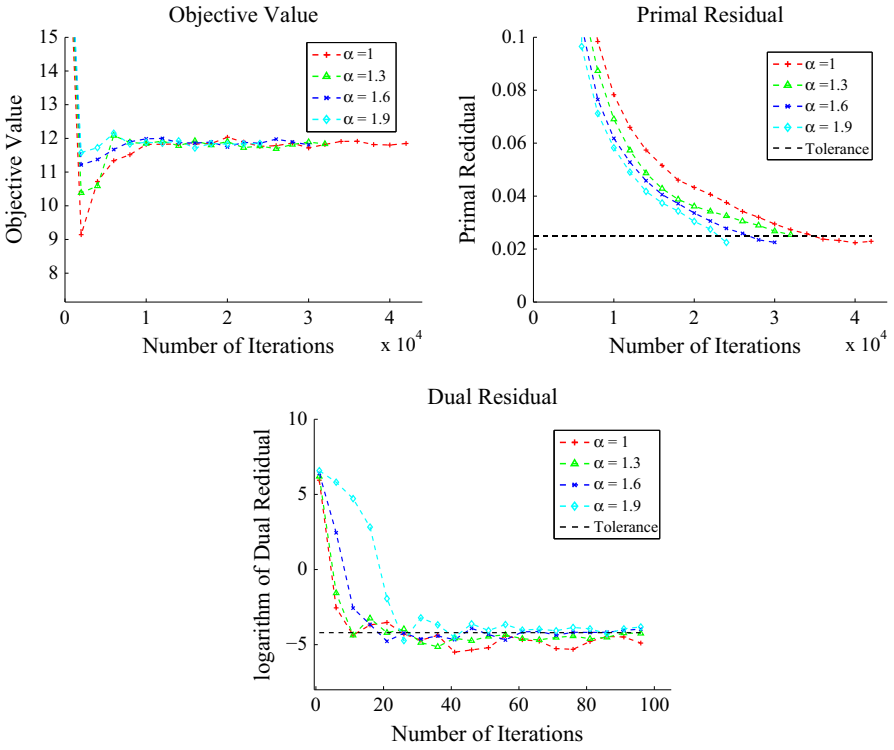

Fig. 2 Algorithm 1: evolution of objective value, primal residual and dual residual w.r.t. some values of α for Scheme 2

respect to iterations for different values of α in Fig. 2. The acceleration performance of the relaxation factor $\alpha \in [1, 2)$, especially when α is close to 2, over the case where $\alpha = 1$ is clearly demonstrated. Also, we see that when α is close to 2 (e.g., $\alpha = 1.9$),

the objective value and the primal residual decrease faster than the cases where α is close to 1.

As we have mentioned, the $\boldsymbol{\beta}$-subproblem (57) has no closed-form solution; and Algorithm 1 solves it by linearizing its quadratic term and thus only solves the $\boldsymbol{\beta}$-subproblem by one iteration because the linearized $\boldsymbol{\beta}$-subproblem has a closed-form solution. This is actually very important to ensure the efficiency of ADMM-like method for some particular cases of (1), as emphasized in [48]. One may be curious in comparing with the case where the $\boldsymbol{\beta}$-subproblem (57) is solved iteratively by a generic, rather than the special linearization strategy. Note that the $\boldsymbol{\beta}$-subproblem (57) is a standard ℓ_1-ℓ_2 model, and we can simply apply the ADMM scheme (2) to solve it iteratively by introducing an auxiliary variable $\mathbf{v} = \boldsymbol{\beta}$ and thus reformulating it as a special case of (1). This case is denoted by "$ADMM^2$" when we report numerical results because the ADMM is used both externally and internally. As analyzed in [14,41], to ensure the convergence of $ADMM^2$, the accuracy of solving the $\boldsymbol{\beta}$-subproblem (57) should keep increasing. Thus, in the implementation of ADMM for the subproblem (57), we gradually decrease the tolerance for the inner problem from $\epsilon = 5 \times 10^{-2}$ to $\epsilon = 5 \times 10^{-4}$. Specifically, we take $\epsilon = 5 \times 10^{-2}$ when $\min(r^t/\epsilon^{pri}, s^t/\epsilon^{dua}) > 50$; $\epsilon = 5 \times 10^{-3}$ when $10 < \max(r^t/\epsilon^{pri}, s^t/\epsilon^{dua}) < 50$; and $\epsilon = 5 \times 10^{-4}$ when $\max(r^t/\epsilon^{pri}, s^t/\epsilon^{dua}) < 10$. We further set the maximal iteration numbers as 1,000 and 40,000 for the inner and outer loops executed by the $ADMM^2$, respectively.

We compare Algorithm 1 and $ADMM^2$ for their averaged performances. In Table 1, we list the averaged computing time in seconds and the objective function values for Algorithm 1 with $\alpha = 1$ and 1.9, and $ADMM^2$. Recall Algorithm 1 with $\alpha = 1$ is the linearized version of the original ADMM (2) and $ADMM^2$ is the case where the $\boldsymbol{\beta}$-subproblem (57) is solved iteratively by the original ADMM scheme. The data in this table shows that achieving the same level of objective function values, Algorithm 1 with $\alpha = 1.9$ is faster than Algorithm 1 with $\alpha = 1$ (thus, the acceleration performance of the GADMM with $\alpha \in (1, 2)$ is illustrated); and Algorithm 1 with either $\alpha = 1$ or $\alpha = 1.9$ is much faster than $ADMM^2$ (thus the necessity of linearization for

Table 1 Numerical comparison between the averaged performance of Algorithm 1 and $ADMM^2$

	Alg. 1 ($\alpha = 1$)	Alg. 1 ($\alpha = 1.9$)	$ADMM^2$
Scheme 1			
CPU time(s)	100.54	72.25	141.86
Objective value	2.1826	2.1998	2.2018
Violations	0.0081	0.0073	0.0182
Scheme 2			
CPU time(s)	54.70	28.33	481.44
Objective value	14.3276	14.3345	14.2787
Violations	0.0099	0.0098	0.0183
Scheme 3			
CPU time(s)	49.67	27.94	223.15
Objective value	2.4569	2.4537	2.4912
Violations	0.0078	0.0080	0.0176

Table 2 Numerical comparison between Algorithm 1 and $ADMM^2$ for microarray dataset

	Alg. 1 ($\alpha = 1$)	Alg. 1 ($\alpha = 1.9$)	$ADMM^2$
Training error	1/60	1/60	1/60
Testing error	3/60	3/60	3/60
CPU time(s)	201.17	167.61	501.95
Objective value	24.67	24.85	24.81
Violation	0.0121	0.0121	0.0241

the $\boldsymbol{\beta}$-subproblem (57) is clearly demonstrated). We also list the averaged primal feasibility violations of the solutions generated by the algorithms in the table. The violation is defined as $\|\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\mu}}_d - \mathbf{w}\|$, where $\widehat{\boldsymbol{\beta}}$ is the output solution and $\mathbf{w} = \min\left(\max\left(\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\mu}}_d, -\lambda\right), \lambda\right) \in \mathbb{R}^d$. It is clearly seen in the table that Algorithm 1 achieves better feasibility than $ADMM^2$.

5.1.2 Real dataset

In this section, we compare Algorithm 1 with $ADMM^2$ on a real dataset of microarray dataset in [38]. The dataset contains 13,182 microarray samples from Affymetrixs HGU133a platform. The raw data contains 2,711 tissue types (e.g., lung cancers, brain tumors, Ewing tumor etc.). In particular, we select 60 healthy samples and 60 samples from those with breast cancer. We use the first 1,000 genes to conduct the experiments.

It is believed that different tissues are associated with different sets of genes and microarray data have been heavily adopted to classify tissues, see e.g. [28,47]. Our aim is to classify those tissues of breast cancer from those healthy tissues. We randomly select 60 samples from each group to be our training set and use another 60 samples from each group as testing set. The tuning parameter λ is chosen by five-fold cross validation as described in [7].

We set $\rho = 1$ and $\tau = 2\|\widehat{\boldsymbol{\Sigma}}^T\widehat{\boldsymbol{\Sigma}}\|_2$ for Algorithm 1; and take the tolerance $\epsilon = 10^{-3}$ in the stopping criterion of Algorithm 1. The starting points are $\boldsymbol{\beta}^0 = \mathbf{0}$, $\mathbf{z}^0 = \mathbf{0}$ and $\gamma^0 = \mathbf{0}$. The comparison among Algorithm 1 with $\alpha = 1$ and $\alpha = 1.9$ and $ADMM^2$ is shown in Table 2; where the pair "a/b" means there are "a" errors out of "b" samples when iteration is terminated. It is seen that Algorithm 1 outperforms $ADMM^2$ in both accuracy and efficiency; also the case where $\alpha = 1.9$ accelerates Algorithm 1 with $\alpha = 1$.

5.2 Constrained lasso

In this subsection, we apply the L-GADMM (4) to the constrained LASSO model proposed recently in [35].

Consider the standard linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon,$$

where $\mathbf{X} \in \mathbb{R}^{n\times d}$ is a design matrix; $\boldsymbol{\beta} \in \mathbb{R}^d$ is a vector of regression coefficients; $\mathbf{y} \in \mathbb{R}^n$ is a vector of observations, and $\epsilon \in \mathbb{R}^n$ is a vector of random noises. In high-dimensional setting where the number of observations is much smaller than the number of regression coefficients, $n \ll d$, the traditional least-squares method does not perform well. To overcome this difficulty, just like the sparse LDA model (55), certain sparsity conditions are assumed for the linear regression model. With the sparsity assumption on the vector of regression coefficients $\boldsymbol{\beta}$, we have the following model

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + p_\lambda(\boldsymbol{\beta}),$$

where $p_\lambda(\cdot)$ is a penalization function. Different penalization functions have been proposed in the literature, such as the LASSO in [44] where $p_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$, the SCAD [16], the adaptive LASSO [54], and the MCP [51]. Inspired by significant applications such as portfolio selection [18] and monotone regression [35], the constrained LASSO (CLASSO) model was proposed recently in [35]:

$$\begin{array}{ll} \min_{\boldsymbol{\beta}} & \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 \\ s.t. & \mathbf{A}\boldsymbol{\beta} \le \mathbf{b}, \end{array} \tag{64}$$

where $\mathbf{A} \in \mathbb{R}^{m\times d}$ and $\mathbf{b} \in \mathbb{R}^m$ with $m < d$ in some applications like portfolio selection [18]. It is worth noting that many statistical problems such as the fused LASSO [45] and the generalized LASSO [46] can be formulated as the form of (64). Thus it is important to find efficient algorithms for solving (64). Although (64) is a quadratic programming problem that can be solved by interior point methods theoretically, again for high-dimensional cases the application of interior point methods is numerically inefficient because of its extremely expensive computation. Note that existing algorithms for LASSO can not be extended to solve (64) trivially.

In fact, introducing a slack variable to the inequality constraints in (64), we reformulate (64) as

$$\min_{\boldsymbol{\beta},\mathbf{z}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 \quad \text{s.t.} \quad \mathbf{A}\boldsymbol{\beta} - \mathbf{z} = \mathbf{0},\ \mathbf{z} \le \mathbf{b}, \tag{65}$$

which is a special case of the model (1) and thus the L-GADMM (4) is applicable. More specifically, the iterative scheme of (4) with $\mathbf{G} = \tau\mathbf{I}_d - \rho(\mathbf{A}^T\mathbf{A})$ for solving (65) reads as

$$\boldsymbol{\beta}^{t+1} = \operatorname*{argmin}_{\boldsymbol{\beta}} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 + \rho(\boldsymbol{\beta} - \boldsymbol{\beta}^t)^T\mathbf{u}^t + \frac{\tau}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^t\|^2 \right\}, \tag{66}$$

$$\mathbf{z}^{t+1} = \operatorname*{argmin}_{\mathbf{z}\le\mathbf{b}} \left\{ \frac{\rho}{2} \left\| \alpha\mathbf{A}\boldsymbol{\beta}^{t+1} + (1-\alpha)\mathbf{z}^t - \mathbf{z} - \frac{\gamma^t}{\rho} \right\|^2 \right\}, \tag{67}$$

$$\gamma^{t+1} = \gamma^t - \rho\Big(\alpha\mathbf{A}\boldsymbol{\beta}^{t+1} + (1-\alpha)(\mathbf{b} - \mathbf{z}^t) + \mathbf{z}^{t+1} - \mathbf{b}\Big), \tag{68}$$

where $\mathbf{u}^t = \mathbf{A}^T\left(\mathbf{A}\boldsymbol{\beta}^t - \mathbf{z}^t - \frac{\gamma^t}{\rho}\right)$.

Now, we delineate the subproblems (66–68). First, the $\boldsymbol{\beta}$-subproblem (66) is equivalent to

$$\boldsymbol{\beta}^{t+1} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{2}\boldsymbol{\beta}^T \left(\mathbf{X}^T\mathbf{X} + \tau\mathbf{I}_d\right)\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{w}^t + \lambda\|\boldsymbol{\beta}\|_1 \right\}, \tag{69}$$

where $\mathbf{w}^t = \mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^t - \rho\mathbf{u}^t$, and $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ is the identity matrix. Deriving the optimality condition of (69), the solution of (69), $\boldsymbol{\beta}^{t+1}$ has to satisfy the following equation:

$$\mathbf{C}\boldsymbol{\beta}^{t+1} = \text{shrinkage}\left(\mathbf{w}^t, \lambda\right),$$

where $\mathbf{C} = \left(\mathbf{X}^T\mathbf{X} + \tau\mathbf{I}_d\right)$. As $\mathbf{X}^T\mathbf{X}$ is positive semidefinite and $\mathbf{I}_d$ is of full rank with all positive eigenvalues, we have that $\mathbf{C}$ is invertible. We have a closed-form solution for (66) that

$$\boldsymbol{\beta}^{t+1} = \mathbf{C}^{-1} \times \text{shrinkage}\left(\mathbf{w}^t, \lambda\right). \tag{70}$$

Then, for the $\mathbf{z}$-subproblem (67), its closed-form solution is given by

$$\mathbf{z}^{t+1} = \min\left(\mathbf{b}, \alpha\mathbf{A}\boldsymbol{\beta}^{t+1} + (1-\alpha)\mathbf{z}^t - \frac{\gamma^t}{\rho}\right). \tag{71}$$

Overall, the application of the L-GADMM (4) to the constrained CLASSO model (64) is summarized in Algorithm 2.

We consider two schemes to test the efficiency of Algorithm 2.

– [Scheme 1]: We first generate an $n \times d$ matrix $\mathbf{X}$ with independent standard Gaussian entries where $n = 100$ and $d = 400$, and we standardize the columns of $\mathbf{X}$ to have unit norms. After that, we set the coefficient vector $\boldsymbol{\beta} = (1, \ldots, 1, 0, \ldots 0)^T$ with the first $s = 5$ entries to be 1 and the rest to be 0. Next, we generate a $m \times d$ matrix $\mathbf{A}$ with independent standard Gaussian entries where $m = 100$, and we generate $\mathbf{b} = \mathbf{A}\boldsymbol{\beta} + \widehat{\boldsymbol{\epsilon}}$ where the entries of $\widehat{\epsilon}$ are independent random variables uniformly distributed in $[0, \sigma]$, and the vector of observations $\mathbf{y}$ is generated by $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_d)$ where $\sigma = 0.1$. We set $\lambda = 1$ in the model (64).

Algorithm 2 Solving Constrained LASSO (64) by the L-GADMM (4)

Initialize $\boldsymbol{\beta}^0, \mathbf{z}^0, \gamma^0, \tau, \rho$.
while Stopping criterion is not satisfied. **do**
Compute $\boldsymbol{\beta}^{t+1}$ by (70).
Compute $\mathbf{z}^{t+1}$ by (71).
Update γ^{t+1} by (68).
end while

– [Scheme 2]: We follow the same procedures of Scheme 1, but we change σ to be 0.3

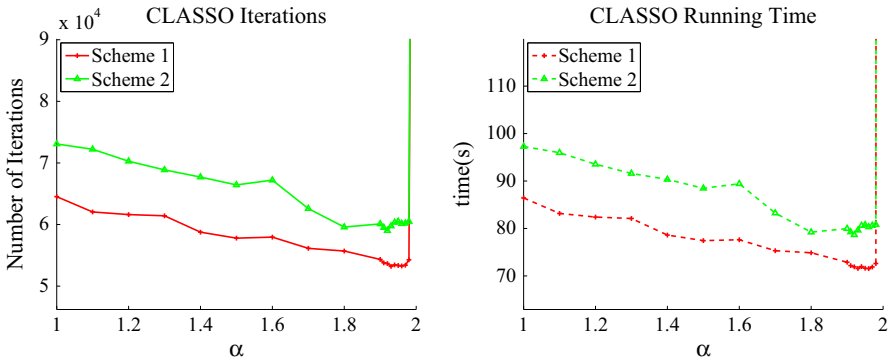

Fig. 3 Algorithm 2: evolution of number of iterations and computing time in seconds w.r.t. different values of α for synthetic dataset

To implement Algorithm 2, we use the stopping criterion proposed by [48]: Let the primal and dual residual at the t-th iteration be $p^t = \|\mathbf{A}\boldsymbol{\beta}^t - \mathbf{z}^t\|$ and $d^t = \|\gamma^t - \gamma^{t+1}\|$, respectively; the tolerance of primal and dual residuals are set by the criterion described in [6] where $\epsilon^{pri} = \sqrt{d}\epsilon + \epsilon \max\left(\|\mathbf{A}\boldsymbol{\beta}^t\|, \|\mathbf{z}^t\|, \|\mathbf{b}\|\right)$ and $\epsilon^{dua} = \sqrt{m}\epsilon + \epsilon\|\mathbf{A}^T\gamma\|$; then the iteration is stopped when $p^t < \epsilon^{pri}$ and $d^t < \epsilon^{dua}$. We choose $\epsilon = 10^{-4}$ for our experiments. We set $\rho = 10^{-3}$ and $\tau = 20\|\mathbf{A}^T\mathbf{A}\|_2$ and the starting iterate as $\boldsymbol{\beta}^0 = \mathbf{0}$, $\mathbf{z}^0 = \mathbf{0}$ and $\gamma^0 = \mathbf{0}$.

Since synthetic dataset is considered, we repeat the simulation ten times and report the averaged numerical performance of Algorithm 2. In Fig. 3, we plot the evolutions of number of iterations and computing time in seconds with different values of $\alpha \in [1, 2]$ for Algorithm 2 with α chosen the same way as we did in the previous subsection. The acceleration performance when $\alpha \in (1, 2)$ over the case where $\alpha = 1$ is clearly shown by the curves in this figure. For example, the case where $\alpha = 1.9$ is about 30 % faster than the case where $\alpha = 1$. We also see that when $\alpha = 2$, Algorithm 2 does not converge after 100,000 iterations. This coincides with the failure of strict contraction of the sequence generated by Algorithm 2, as we have mentioned in Remark 3.

Like Sect. 5.1.1, we also compare the averaged performance of Algorithm 2 with the application of the original ADMM (2) where the resulting $\boldsymbol{\beta}$-subproblem is solved iteratively by the ADMM. The penalty parameter is set as 1, and the ADMM is implemented to solve the $\boldsymbol{\beta}$-subproblem, whose tolerance ϵ is gradually decreased from 10^{-2} to 10^{-4} obeying the same rule as that mentioned in Sect. 5.1.1. We further set the maximal iteration numbers to be 1,000 and 10,000 for inner and outer loops executed by $ADMM^2$. Again, this case is denoted by "$ADMM^2$". In Table 3, we list the computing time in seconds to achieve the same level of objective function values for three cases: Algorithm 2 with $\alpha = 1$, Algorithm 2 with $\alpha = 1.9$, and $ADMM^2$. We see that Algorithm 2 with either $\alpha = 1$ or $\alpha = 1.9$ is much faster than $ADMM^2$; thus the necessity of linearization when ADMM-like methods are applied to solve the constrained LASSO model (64) is illustrated. Also, the acceleration performance of the GADMM with $\alpha \in (1, 2)$ is demonstrated as Algorithm 2 with $\alpha = 1.9$ is faster than Algorithm 2 with $\alpha = 1$. We also report the averaged primal feasibility violations

Table 3 Numerical comparison of the averaged performance between Algorithm 2 and $ADMM^2$

	Alg. 2 ($\alpha = 1$)	Alg. 2 ($\alpha = 1.9$)	$ADMM^2$
Scheme 1			
CPU time(s)	88.03	71.01	166.17
Objective value	5.4679	5.4678	5.4823
Violations	0.0033	0.0031	0.2850
Scheme 2			
CPU time(s)	97.19	82.49	165.90
Objective value	5.3861	5.3864	5.6364
Violations	0.0032	0.0032	0.1158

of the solutions generated by each algorithm. The violation is defined as $\|\mathbf{A}\widehat{\boldsymbol{\beta}} - \mathbf{w}\|$, where $\widehat{\boldsymbol{\beta}}$ is the output solution and $\mathbf{w} = \min\left(\mathbf{b}, \mathbf{A}\widehat{\boldsymbol{\beta}}\right) \in \mathbb{R}^m$, respectively. It is seen in the table that Algorithm 2 achieves better feasibility than $ADMM^2$.

5.3 Dantzig selector

Last, we test the Dantzig Selector model proposed in [9]. In [48], this model has been suggested to be solved by the linearized version of ADMM which is a special case of (4) with $\alpha = 1$. We now test this example again to show the acceleration performance of (4) with $\alpha \in (1, 2)$.

The Dantzig selector model in [9] deals with the case where the the number of observations is much smaller than the number of regression coefficients, i.e. $n \ll d$. In particular, the Dantzig selector model is

$$\begin{array}{ll} \min & \|\boldsymbol{\beta}\|_1 \\ s.t. & \|\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} - \mathbf{y})\|_\infty \le \delta, \end{array} \tag{72}$$

where $\delta > 0$ is a tuning parameter, and $\|\cdot\|_\infty$ is the infinity norm.

As elaborated in [48], the model (72) can be formulated as

$$\begin{array}{ll} \min & \|\boldsymbol{\beta}\|_1 \\ s.t. & \mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} - \mathbf{y}) - \mathbf{x} = 0, \\ & \boldsymbol{\beta} \in \mathbb{R}^p, \mathbf{x} \in \Omega := \{\mathbf{x} : \|\mathbf{x}\|_\infty \le \delta\}, \end{array} \tag{73}$$

where $\mathbf{x} \in \mathbb{R}^d$ is an auxiliary variable. Obviously, (73) is a special case of (1) and thus the L-GADMM (4) is applicable. In fact, applying (4) with $\mathbf{G} = \mathbf{I}_d - \tau\|(\mathbf{X}^T\mathbf{X})^T(\mathbf{X}^T\mathbf{X})\|_2$ to (73), we obtain the subproblems as the following:

$$\boldsymbol{\beta}^{t+1} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\left\{\|\boldsymbol{\beta}\|_1 + \frac{\rho}{2}\left(2(\mathbf{v}^t)^T(\boldsymbol{\beta} - \boldsymbol{\beta}^t) + \frac{\tau}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^t\|^2\right)\right\}, \tag{74}$$

$$\mathbf{x}^{t+1} = \underset{\mathbf{x}\in\Omega}{\operatorname{argmin}} \left\{ \frac{\rho}{2} \left\| \mathbf{x} - \alpha \mathbf{X}^T(\mathbf{X}\boldsymbol{\beta}^{t+1} - \mathbf{y}) - (1-\alpha)\mathbf{x}^t + \frac{\gamma^t}{\rho} \right\|^2 \right\}, \tag{75}$$

$$\gamma^{t+1} = \gamma^t - \rho(\alpha \mathbf{X}^T(\mathbf{X}\boldsymbol{\beta}^{t+1} - \mathbf{y}) + (1-\alpha)\mathbf{x}^t - \mathbf{x}^{t+1}), \tag{76}$$

where $\mathbf{v}^t := (\mathbf{X}^T\mathbf{X})^T[\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta}^t - \mathbf{y}) - \mathbf{x}^t - \frac{\gamma^t}{\rho}]$. Note that $\tau \geq 2\|(\mathbf{X}\mathbf{X}^T)\mathbf{X}^T\mathbf{X}\|_2$ is required to ensure the convergence, see [48] for the detailed proof.

The $\boldsymbol{\beta}$-subproblem (74) can be rewritten as

$$\boldsymbol{\beta}^{t+1} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \|\boldsymbol{\beta}\|_1 + \frac{\rho\tau}{4} \left\| \boldsymbol{\beta} - \left(\boldsymbol{\beta}^t - \frac{2}{\tau}\mathbf{v}^t \right) \right\|^2 \right\},$$

whose closed-form solution is given by

$$\boldsymbol{\beta}^{t+1} = \text{shrinkage}\left(\boldsymbol{\beta}^t - \frac{2}{\tau}\mathbf{v}^t, \frac{2}{\rho\tau} \right). \tag{77}$$

Moreover, the solution of the $\mathbf{x}$-subproblem (75) is given by

$$\mathbf{x}^{t+1} = \min\left\{ \max\left\{ \alpha\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta}^{t+1} - \mathbf{y}) + (1-\alpha)\mathbf{x}^t - \frac{\gamma^t}{\rho}, -\delta \right\}, \delta \right\}. \tag{78}$$

Overall, the application of the L-GADMM (4) to the Dantzig Selector model (72) is summarized in Algorithm 3.

Algorithm 3 Solving Dantzig Selector (72) by the L-GADMM (4)

Given $\mathbf{X}$, $\mathbf{y}$, δ.
Initialize $\boldsymbol{\beta}^0$, $\mathbf{x}^0$, γ^0, τ, ρ.
while Stopping criteria is not satisfied. **do**
  Compute $\boldsymbol{\beta}^{t+1}$ by (77).
  Compute $\mathbf{x}^{t+1}$ by (78).
  Update γ^{t+1} by (76).
end while

5.3.1 Synthetic dataset

We first test some synthetic dataset for the Dantizg Selector model (72). We follow the simulation setup in [9] to generate the design matrix $\mathbf{X}$ whose columns all have the unit norm. Then, we randomly choose a set S of cardinality s. $\boldsymbol{\beta}$ is generated by

$$\boldsymbol{\beta}_i = \begin{cases} \xi_i(1 + |a_i|) & \text{if } i \in S; \\ 0 & \text{otherwise,} \end{cases}$$

where $\xi \sim U(-1, 1)$(i.e., the uniform distribution on the interval $(-1, 1)$) and $a_i \sim N(0, 1)$. At last, $\mathbf{y}$ is generated by $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$ with $\epsilon \sim N(0, \sigma^2\mathbf{I})$.

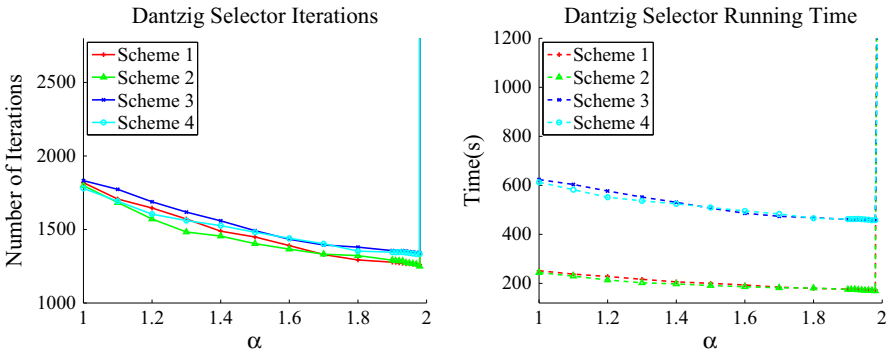

Fig. 4 Algorithm 3: evolution of number of iterations and computing time w.r.t. different values of α for synthetic dataset

We consider four schemes as listed below:

- [Scheme 1]: $(n, d, s) = (720, 2560, 80)$, $\sigma = 0.03$.
- [Scheme 2]: $(n, d, s) = (720, 2560, 80)$, $\sigma = 0.05$.
- [Scheme 3]: $(n, d, s) = (1440, 5120, 160)$, $\sigma = 0.03$.
- [Scheme 4]: $(n, d, s) = (1440, 5120, 160)$, $\sigma = 0.05$.

We take $\delta = \sigma\sqrt{2\log d}$.

To implement Algorithm 3, we set the penalty parameter $\rho = 0.1$ and $\tau = 2.5\|(\mathbf{X}\mathbf{X}^T)\mathbf{X}^T\mathbf{X}\|_2$ respectively. Again, we use the stopping criterion described in [6]: Let the primal and dual residual at the t-th iteration be $r^t = \|\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}^t - \mathbf{x}^t - \mathbf{X}^T\mathbf{y}\|$ and $s^t = \|\rho\mathbf{X}^T\mathbf{X}(\gamma^t - \gamma^{t-1})\|$, respectively; let the tolerance of the primal and dual residual at the t-th iteration be $\epsilon^{pri} = \sqrt{d}\epsilon + \epsilon \max\left(\|\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}^t\|, \|\mathbf{x}^t\|, \|\mathbf{X}^T\mathbf{y}\|\right)$ and $\epsilon^{dua} = \sqrt{n}\epsilon + \epsilon\|\mathbf{X}^T\mathbf{X}\gamma^t\|$, respectively; then the iteration is stopped when $r^t < \epsilon^{pri}$ and $s^t < \epsilon^{dua}$ simultaneously. We choose $\epsilon = 10^{-4}$ for our experiments. We set the starting points as $\boldsymbol{\beta}^0 = \mathbf{0}$, $\mathbf{x}^0 = \mathbf{0}$ and $\gamma^0 = \mathbf{0}$.

In Fig. 4, we plot the evolutions of number of iterations and computing time in seconds with respect to different values of $\alpha \in [1, 2]$. We choose the values of $\alpha \in [1, 2]$ for Algorithm 3 with α chosen the same way as we did in the previous subsections. The curves in Fig. 4 show the acceleration performance of Algorithm 3 with $\alpha \in (1, 2)$ clearly. Also, if $\alpha = 2$, Algorithm 3 does not converge after 10,000 iterations.

Like previous sections, we compare the averaged performance of our method with $ADMM^2$ which solves the $\boldsymbol{\beta}$-subproblem iteratively by ADMM for all of the four schemes. In the implementation of using ADMM to solve the $\boldsymbol{\beta}$-subproblem, we use the same stopping criterion as described in the previous sections, and the tolerance ϵ is gradually decreased from 10^{-2} to 10^{-4} obeying the same rule as the rule in Sect. 5.1.1. We further set the maximal iteration numbers to be 1,000 and 10,000 for inner and outer loops executed by $ADMM^2$, respectively. Also, we set the penalty parameter for inner loop as 1. For the outer loop, we set the tolerance $\epsilon = 10^{-4}$ for both Algorithm 3 and $ADMM^2$. In Table 4, we present the computing time in seconds to achieve the same level of objective values for three cases: Algorithm 3 with $\alpha = 1$, Algorithm 3

Table 4 Numerical comparison of the averaged performance between Algorithm 3 and $ADMM^2$

	Alg. 3 ($\alpha = 1$)	Alg. 3 ($\alpha = 1.9$)	$ADMM^2$
Scheme 1			
CPU time(s)	215.12	154.7	1,045.62
Objective value	60.6777	60.6821	60.8028
Violation	0.0048	0.0050	0.1464
Scheme 2			
CPU time(s)	213.80	147.29	1,184.98
Objective value	54.7704	54.7729	54.7915
Violation	0.0049	0.0052	0.1773
Scheme 3			
CPU time(s)	613.20	462.41	5,514.81
Objective value	124.0258	124.0208	124.2314
Violation	0.0072	0.0072	0.2361
Scheme 4			
CPU time(s)	607.80	463.58	5,357.55
Objective value	111.8513	111.8489	112.0675
Violation	0.0072	0.0071	0.2137

Table 5 Numerical comparison between Algorithm 3 and $ADMM^2$ for microarray dataset

	Alg. 3 ($\alpha = 1$)	Alg. 3 ($\alpha = 1.9$)	$ADMM^2$
Training error	0/45	0/45	0/45
Testing error	0/37	0/37	0/37
CPU time(s)	1,006.68	532.96	2,845.27
Objective value	17.596	17.579	17.771
Violation	0.0131	0.0131	0.0285

with $\alpha = 1.9$, and $ADMM^2$. It is seen that Algorithm 3 with either $\alpha = 1$ or $\alpha = 1.9$ is much faster than $ADMM^2$, and Algorithm 3 with $\alpha = 1.9$ is faster than Algorithm 3 with $\alpha = 1$; we have thus demonstrated the necessity of linearization when ADMM-like methods are applied to solve the Dantzig selector model (64) and the acceleration performance of GADMM with $\alpha \in (1, 2)$. Also, we list the averaged primal feasibility violation of each algorithm. The violation is defined as $\|\mathbf{X}^T(\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{y}) - \mathbf{w}\|$, where $\widehat{\boldsymbol{\beta}}$ is the output solution and $\mathbf{w} = \min\left(\max\left(\mathbf{X}^T(\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{y}), -\delta\right), \delta\right) \in \mathbb{R}^d$. Algorithm 3 achieves better feasibility than $ADMM^2$ as illustrated in the table.

5.3.2 Real dataset

Then, we test the Dantizg Selector model (72) for a real dataset. In particular, we test the same dataset as in Sect. 5.1.2. We look at another disease: Ewing's sarcoma which has drawn much attention in the literature, see e.g. [26]. We have 27 samples diagnosed

with Ewing's sarcoma. Then, we randomly select 55 healthy samples. Next, we select 30 healthy samples and 15 samples diagnosed with Ewing's sarcoma as training set, and let the remaining 42 samples be the testing set. We use the first 2,000 genes to conduct the analysis. Thus, this dataset corresponds to the model (72) with $n = 45$ and $d = 2{,}000$. In the implementation, the parameter δ in (72) is chosen by five-fold cross validation, and we set $\rho = 0.02$ and $\tau = 9\|(\mathbf{X}^T\mathbf{X})^T\mathbf{X}^T\mathbf{X}\|_2$ for Algorithm 3, and we use the same stopping criterion with $\epsilon = 4 \times 10^{-4}$. The comparison between Algorithm 3 and $ADMM^2$ is listed in Table 5. From this table, we see that Algorithm 3 with either $\alpha = 1$ or $\alpha = 1.9$ outperforms $ADMM^2$ significantly—it requires much less computing time to achieve the same level of objective function values with similar primal feasibility and attains the same level of training or testing error. In addition, the acceleration performance of the case where $\alpha = 1.9$ over the case where $\alpha = 1$ is again demonstrated for Algorithm 3.

6 Conclusion

In this paper, we take a deeper look at the linearized version of the generalized alternating direction method of multiplier (ADMM) and establish its worst-case $\mathcal{O}(1/k)$ convergence rate in both the ergodic and a nonergodic senses. This result subsumes some existing results established for the original ADMM and generalized ADMM schemes; and it provides accountable and novel theoretical support to the numerical efficiency of the generalized ADMM. Further, we apply the linearized version of the generalized ADMM to solve some important statistical learning applications; and enlarge the application range of the generalized ADMM. Finally we would mention that the worst-case $\mathcal{O}(1/k)$ convergence rate established in this paper amounts to a sublinear speed of convergence. If certain conditions are assumed (e.g., some error bound conditions) or the model under consideration has some special properties, it is possible to establish the linear convergence rate for the linearized version of the generalized ADMM by using similar techniques in, e.g., [5,11,27]. We omit the detail of analysis and only focus on the convergence rate analysis from the iteration complexity perspective in this paper.

7 Appendices

We show that our analysis in Sects. 3 and 4 can be extended to the case where both the $\mathbf{x}$- and $\mathbf{y}$-subproblems in (3) are linearized. The resulting scheme, called doubly linearized version of the GADMM ("DL-GADMM" for short), reads as

$$\begin{aligned}
\mathbf{x}^{t+1} &= \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmin}} \left\{ f_1(\mathbf{x}) - \mathbf{x}^T\mathbf{A}^T\gamma^t + \frac{\rho}{2}\|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y}^t - \mathbf{b}\|^2 + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^t\|^2_{\mathbf{G}_1} \right\}, \\
\mathbf{y}^{t+1} &= \underset{\mathbf{y}\in\mathcal{Y}}{\operatorname{argmin}} \left\{ f_2(\mathbf{y}) - \mathbf{y}^T\mathbf{B}^T\gamma^t + \frac{\rho}{2} \Big\| \alpha\mathbf{A}\mathbf{x}^{t+1} \right. \\
&\quad \left. + (1-\alpha)(\mathbf{b} - \mathbf{B}\mathbf{y}^t) + \mathbf{B}\mathbf{y} - \mathbf{b} \Big\|^2 + \frac{1}{2}\|\mathbf{y} - \mathbf{y}^t\|^2_{\mathbf{G}_2} \right\},
\end{aligned}$$

$$\gamma^{t+1} = \gamma^t - \rho\Big(\alpha\mathbf{A}\mathbf{x}^{t+1} + (1-\alpha)(\mathbf{b} - \mathbf{B}\mathbf{y}^t) + \mathbf{B}\mathbf{y}^{t+1} - \mathbf{b}\Big), \tag{79}$$

where the matrices $\mathbf{G}_1 \in \mathbb{R}^{n_1 \times n_1}$ and $\mathbf{G}_2 \in \mathbb{R}^{n_2 \times n_2}$ are both symmetric and positive definite.

For further analysis, we define two matrices, which are analogous to $\mathbf{H}$ and $\mathbf{Q}$ in (11), respectively, as

$$\begin{aligned}
\mathbf{H}_2 &= \begin{pmatrix} \mathbf{G}_1 & 0 & 0 \\ 0 & \frac{\rho}{\alpha}\mathbf{B}^T\mathbf{B} + \mathbf{G}_2 & \frac{1-\alpha}{\alpha}\mathbf{B}^T \\ 0 & \frac{1-\alpha}{\alpha}\mathbf{B} & \frac{1}{\alpha\rho}\mathbf{I}_n \end{pmatrix}, \\
\mathbf{Q}_2 &= \begin{pmatrix} \mathbf{G}_1 & 0 & 0 \\ 0 & \rho\mathbf{B}^T\mathbf{B} + \mathbf{G}_2 & (1-\alpha)\mathbf{B}^T \\ 0 & -\mathbf{B} & \frac{1}{\rho}\mathbf{I}_n \end{pmatrix}.
\end{aligned} \tag{80}$$

Obviously, we have

$$\mathbf{Q}_2 = \mathbf{H}_2\mathbf{M}, \tag{81}$$

where $\mathbf{M}$ is defined in (10). Note that the equalities (8) and (9) still hold.

7.1 A worst-case $\mathcal{O}(1/k)$ convergence rate in the ergodic sense for (79)

We first establish a worst-case $\mathcal{O}(1/k)$ convergence rate in the ergodic sense for the DL-GADMM (79). Indeed, using the relationship (81), the resulting proof is nearly the same as that in Sect. 3 for the L-GADMM (4). We thus only list two lemmas (analogous to Lemmas 1 and 2) and one theorem (analogous to Theorem 2) to demonstrate a worst-case $\mathcal{O}(1/k)$ convergence rate in the ergodic sense for (79), and omit the details of proofs.

Lemma 7 *Let the sequence $\{\mathbf{w}^t\}$ be generated by the DL-GADMM (79) with $\alpha \in (0, 2)$ and the associated sequence $\{\widetilde{\mathbf{w}}^t\}$ be defined in (7). Then we have*

$$f(\mathbf{u}) - f(\widetilde{\mathbf{u}}^t) + \left(\mathbf{w} - \widetilde{\mathbf{w}}^t\right)^T F(\widetilde{\mathbf{w}}^t) \geq \left(\mathbf{w} - \widetilde{\mathbf{w}}^t\right)^T \mathbf{Q}_2 \left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right), \quad \forall \mathbf{w} \in \Omega, \tag{82}$$

where $\mathbf{Q}_2$ is defined in (80).

Lemma 8 *Let the sequence $\{\mathbf{w}^t\}$ be generated by the DL-GADMM (79) with $\alpha \in (0, 2)$ and the associated sequence $\{\widetilde{\mathbf{w}}^t\}$ be defined in (7). Then for any $\mathbf{w} \in \Omega$, we have*

$$\begin{aligned}
&\left(\mathbf{w} - \widetilde{\mathbf{w}}^t\right)^T \mathbf{Q}_2 \left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right) \\
&= \frac{1}{2}\left(\|\mathbf{w} - \mathbf{w}^{t+1}\|_{\mathbf{H}_2}^2 - \|\mathbf{w} - \mathbf{w}^t\|_{\mathbf{H}_2}^2\right) + \frac{1}{2}\|\mathbf{x}^t - \widetilde{\mathbf{x}}^t\|_{\mathbf{G}_1}^2 \\
&\quad + \frac{1}{2}\|\mathbf{y}^t - \widetilde{\mathbf{y}}^t\|_{\mathbf{G}_2}^2 + \frac{2-\alpha}{2\rho}\|\gamma^t - \widetilde{\gamma}^t\|^2.
\end{aligned} \tag{83}$$

Theorem 7 *Let* $\mathbf{H}_2$ *be given by* (80) *and* $\{\mathbf{w}^t\}$ *be the sequence generated by the DL-GADMM* (79) *with* $\alpha \in (0, 2)$. *For any integer* $k > 0$, *let* $\widehat{\mathbf{w}}_k$ *be defined by*

$$\widehat{\mathbf{w}}_k = \frac{1}{k+1} \sum_{t=0}^{k} \widetilde{\mathbf{w}}^t, \tag{84}$$

where $\widetilde{\mathbf{w}}^t$ *is defined in* (7). *Then,* $\widehat{\mathbf{w}}_k \in \Omega$ *and*

$$f(\widehat{\mathbf{u}}_k) - f(\mathbf{u}) + (\widehat{\mathbf{w}}_k - \mathbf{w})^T F(\mathbf{w}) \leq \frac{1}{2(k+1)} \|\mathbf{w} - \mathbf{w}^0\|^2_{\mathbf{H}_2}, \quad \forall \mathbf{w} \in \Omega.$$

7.2 A worst-case $\mathcal{O}(1/k)$ convergence rate in a nonergodic sense for (79)

Next, we prove a worst-case $\mathcal{O}(1/k)$ convergence rate in a nonergodic sense for the DL-GADMM (79). Note that Lemma 4 still holds by replacing $\mathbf{H}$ with $\mathbf{H}_2$. That is, if $\|\mathbf{w}^t - \mathbf{w}^{t+1}\|^2_{\mathbf{H}_2} = 0$, $\widetilde{\mathbf{w}}^t$ defined in (7) is an optimal solution point to (5). Thus, for the sequence $\{\mathbf{w}^t\}$ generated by the DL-GADMM (79), it is reasonable to measure the accuracy of an iterate by $\|\mathbf{w}^t - \mathbf{w}^{t+1}\|^2_{\mathbf{H}_2}$.

Proofs of the following two lemmas are analogous to those of Lemmas 5 and 6, respectively. We thus omit them.

Lemma 9 *Let the sequence* $\{\mathbf{w}^t\}$ *be generated by the DL-GADMM* (79) *with* $\alpha \in (0, 2)$ *and the associated* $\{\widetilde{\mathbf{w}}^t\}$ *be defined in* (7); *the matrix* $\mathbf{Q}_2$ *be defined in* (80). *Then, we have*

$$\left(\widetilde{\mathbf{w}}^t - \widetilde{\mathbf{w}}^{t+1}\right)^T \mathbf{Q}_2 \left[\left(\mathbf{w}^t - \mathbf{w}^{t+1}\right) - \left(\widetilde{\mathbf{w}}^t - \widetilde{\mathbf{w}}^{t+1}\right)\right] \geq 0.$$

Lemma 10 *Let the sequence* $\{\mathbf{w}^t\}$ *be generated by the DL-GADMM* (79) *with* $\alpha \in (0, 2)$ *and the associated* $\{\widetilde{\mathbf{w}}^t\}$ *be defined in* (7); *the matrices* $\mathbf{M}$, $\mathbf{H}_2$, $\mathbf{Q}_2$ *be defined in* (10) *and* (80). *Then, we have*

$$\begin{aligned} &\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right)^T \mathbf{M}^T \mathbf{H}_2 \mathbf{M} \left[\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right) - \left(\mathbf{w}^{t+1} - \widetilde{\mathbf{w}}^{t+1}\right)\right] \\ &\geq \frac{1}{2} \left\| \left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right) - \left(\mathbf{w}^{t+1} - \widetilde{\mathbf{w}}^{t+1}\right) \right\|^2_{(\mathbf{Q}_2^T + \mathbf{Q}_2)}. \end{aligned}$$

Based on the above two lemmas, we see that the sequence $\{\|\mathbf{w}^t - \mathbf{w}^{t+1}\|_{\mathbf{H}_2}\}$ is monotonically non-increasing. That is, we have the following theorem.

Theorem 8 *Let the sequence* $\{\mathbf{w}^t\}$ *be generated by the DL-GADMM* (79) *and the matrix* $\mathbf{H}_2$ *be defined in* (80). *Then, we have*

$$\|\mathbf{w}^{t+1} - \mathbf{w}^{t+2}\|^2_{\mathbf{H}_2} \leq \|\mathbf{w}^t - \mathbf{w}^{t+1}\|^2_{\mathbf{H}_2}.$$

Note that for the DL-GADMM (79), the $\mathbf{y}$-subproblem is also proximally regularized, and we can not extend the inequality (31) to this new case. This is indeed the main difficulty for proving a worst-case $\mathcal{O}(1/k)$ convergence rate in a nonergodic sense for the DL-GADMM (79). A more elaborated analysis is needed. Let us show one lemma first to bound the left-hand side in (31).

Lemma 11 *Let* $\{\mathbf{y}^t\}$ *be the sequence generated by the DL-GADMM* (79) *with* $\alpha \in (0, 2)$. *Then, we have*

$$\left(\mathbf{y}^t - \mathbf{y}^{t+1}\right)\mathbf{B}^T\left(\gamma^t - \gamma^{t+1}\right) \geq \frac{1}{2}\|\mathbf{y}^t - \mathbf{y}^{t+1}\|^2_{\mathbf{G}_2} - \frac{1}{2}\|\mathbf{y}^{t-1} - \mathbf{y}^t\|^2_{\mathbf{G}_2}. \tag{85}$$

Proof It follows from the optimality condition of the $\mathbf{y}$-subproblem in (79) that

$$f_2(\mathbf{y}) - f_2(\mathbf{y}^{t+1}) + \left(\mathbf{y} - \mathbf{y}^{t+1}\right)^T\left[-\mathbf{B}^T\gamma^{t+1} + \mathbf{G}_2(\mathbf{y}^{t+1} - \mathbf{y}^t)\right] \geq 0, \quad \forall \mathbf{y} \in \mathcal{Y}. \tag{86}$$

Similarly, we also have,

$$f_2(\mathbf{y}) - f_2(\mathbf{y}^t) + (\mathbf{y} - \mathbf{y}^t)^T\left[-\mathbf{B}^T\gamma^t + \mathbf{G}_2(\mathbf{y}^t - \mathbf{y}^{t-1})\right] \geq 0, \quad \forall \mathbf{y} \in \mathcal{Y}. \tag{87}$$

Setting $\mathbf{y} = \mathbf{y}^t$ in (86) and $\mathbf{y} = \mathbf{y}^{t+1}$ in (87), and summing them up, we have

$$\begin{aligned}\left(\mathbf{y}^t - \mathbf{y}^{t+1}\right)\mathbf{B}^T\left(\gamma^t - \gamma^{t+1}\right) &\geq (\mathbf{y}^{t+1} - \mathbf{y}^t)\mathbf{G}_2(\mathbf{y}^{t+1} - \mathbf{y}^t + \mathbf{y}^{t-1} - \mathbf{y}^t)\\ &\geq \|\mathbf{y}^t - \mathbf{y}^{t+1}\|^2_{\mathbf{G}_2} - \frac{1}{2}\|\mathbf{y}^t - \mathbf{y}^{t+1}\|^2_{\mathbf{G}_2} - \frac{1}{2}\|\mathbf{y}^{t-1} - \mathbf{y}^t\|^2_{\mathbf{G}_2}\\ &= \frac{1}{2}\|\mathbf{y}^t - \mathbf{y}^{t+1}\|^2_{\mathbf{G}_2} - \frac{1}{2}\|\mathbf{y}^{t-1} - \mathbf{y}^t\|^2_{\mathbf{G}_2},\end{aligned}$$

where the second inequality holds by the fact that $\mathbf{a}^T\mathbf{b} \geq -\frac{1}{2}(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2)$. The assertion (85) is proved. □

Two more lemmas should be proved in order to establish a worst-case $\mathcal{O}(1/k)$ convergence rate in a nonergodic sense for the DL-GADMM (79).

Lemma 12 *The sequence* $\{\mathbf{w}^t\}$ *generated by the DL-GADMM* (79) *with* $\alpha \in (0, 2)$ *and the associated* $\{\widetilde{\mathbf{w}}^t\}$ *be defined in* (7), *then we have*

$$c_\alpha\left(\|\mathbf{x}^t - \widetilde{\mathbf{x}}^t\|^2_{\mathbf{G}_1} + \|\mathbf{y}^t - \widetilde{\mathbf{y}}^t\|^2_{\mathbf{G}_2} + \frac{\alpha}{\rho}\|\gamma^t - \widetilde{\gamma}^t\|^2\right) \leq \|\mathbf{w}^t - \widetilde{\mathbf{w}}^t\|^2_{\mathbf{Q}_2^T + \mathbf{Q}_2 - \mathbf{M}^T\mathbf{H}_2\mathbf{M}} \tag{88}$$

where c_α *is defined in* (37).

Proof By the definition of $\mathbf{Q}_2$, $\mathbf{M}$ and $\mathbf{H}_2$, we have

$$\begin{aligned}
&\|\mathbf{w}^t - \widetilde{\mathbf{w}}^t\|^2_{\mathbf{Q}_2^T+\mathbf{Q}_2-\mathbf{M}^T\mathbf{H}_2\mathbf{M}} \\
&= \|\mathbf{x}^t - \widetilde{\mathbf{x}}^t\|^2_{\mathbf{G}_1} + \|\mathbf{y}^t - \widetilde{\mathbf{y}}^t\|^2_{\mathbf{G}_2} + \frac{2-\alpha}{\rho}\|\gamma^t - \widetilde{\gamma}^t\|^2 \\
&\geq \min\left\{\frac{2-\alpha}{\alpha}, 1\right\}\left(\|\mathbf{x}^t - \widetilde{\mathbf{x}}^t\|^2_{\mathbf{G}_1} + \|\mathbf{y}^t - \widetilde{\mathbf{y}}^t\|^2_{\mathbf{G}_2} + \frac{\alpha}{\rho}\|\gamma^t - \widetilde{\gamma}^t\|^2\right),
\end{aligned}$$

which implies the assertion (88) immediately. □

In the next lemma, we refine the bound of $(\mathbf{w} - \widetilde{\mathbf{w}}^t)^T\mathbf{Q}_2(\mathbf{w}^t - \widetilde{\mathbf{w}}^t)$ in (82). The refined bound consists of the terms $\|\mathbf{w} - \mathbf{w}^{t+1}\|^2_{\mathbf{H}_2}$ recursively, which is favorable for establishing a worst-case $\mathcal{O}(1/k)$ convergence rate in a nonergodic sense for the DL-GADMM (79).

Lemma 13 *Let* $\{\mathbf{w}^t\}$ *be the sequence generated by the DL-GADMM* (79) *with* $\alpha \in (0, 2)$. *Then,* $\widetilde{\mathbf{w}}^t \in \Omega$ *and*

$$\begin{aligned}
f(\mathbf{u}) - f(\mathbf{u}^t) + (\mathbf{w} - \widetilde{\mathbf{w}})^T F(\mathbf{w}) \geq \frac{1}{2}\left(\|\mathbf{w} - \mathbf{w}^{t+1}\|^2_{\mathbf{H}_2} - \|\mathbf{w} - \mathbf{w}^t\|^2_{\mathbf{H}_2}\right) \\
+ \frac{1}{2}\|\mathbf{w}^t - \widetilde{\mathbf{w}}^t\|^2_{\mathbf{Q}_2^T+\mathbf{Q}_2-\mathbf{M}^T\mathbf{H}_2\mathbf{M}}, \quad \forall \mathbf{w} \in \Omega,
\end{aligned} \tag{89}$$

where $\mathbf{M}$ *is defined in* (10), *and* $\mathbf{H}_2$ *and* $\mathbf{Q}_2$ *are defined in* (80).

Proof By the identity $\mathbf{Q}_2(\mathbf{w}^t - \widetilde{\mathbf{w}}^t) = \mathbf{H}_2(\mathbf{w}^t - \mathbf{w}^{t+1})$, it holds that

$$\left(\mathbf{w} - \widetilde{\mathbf{w}}^t\right)^T \mathbf{Q}_2 \left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right) = \left(\mathbf{w} - \widetilde{\mathbf{w}}^t\right)^T \mathbf{H}_2 \left(\mathbf{w}^t - \mathbf{w}^{t+1}\right), \quad \forall \mathbf{w} \in \Omega.$$

Setting $\mathbf{a} = \mathbf{w}$, $\mathbf{b} = \widetilde{\mathbf{w}}^t$, $\mathbf{c} = \mathbf{w}^t$ and $\mathbf{d} = \mathbf{w}^{t+1}$ in the identity

$$(\mathbf{a} - \mathbf{b})^T\mathbf{H}_2(\mathbf{c} - \mathbf{d}) = \frac{1}{2}\left(\|\mathbf{a}-\mathbf{d}\|^2_{\mathbf{H}_2} - \|\mathbf{a}-\mathbf{c}\|^2_{\mathbf{H}_2}\right) + \frac{1}{2}\left(\|\mathbf{c}-\mathbf{b}\|^2_{\mathbf{H}_2} - \|\mathbf{d} - \mathbf{b}\|^2_{\mathbf{H}_2}\right),$$

we have

$$\begin{aligned}
&2\left(\mathbf{w} - \widetilde{\mathbf{w}}^t\right)\mathbf{Q}_2\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right) \\
&= \|\mathbf{w} - \mathbf{w}^{t+1}\|^2_{\mathbf{H}_2} - \|\mathbf{w} - \mathbf{w}^t\|^2_{\mathbf{H}_2} + \|\mathbf{w}^t - \widetilde{\mathbf{w}}^t\|^2_{\mathbf{H}_2} - \|\mathbf{w}^{t+1} - \widetilde{\mathbf{w}}^t\|^2_{\mathbf{H}_2}.
\end{aligned} \tag{90}$$

Meanwhile, we have

$$\begin{aligned}
\|\mathbf{w}^t - \widetilde{\mathbf{w}}^t\|^2_{\mathbf{H}_2} - \|\mathbf{w}^{t+1} - \widetilde{\mathbf{w}}^t\|^2_{\mathbf{H}_2} &= \|\mathbf{w}^t - \widetilde{\mathbf{w}}^t\|^2_{\mathbf{H}_2} - \|(\mathbf{w}^t - \widetilde{\mathbf{w}}^t) - \left(\mathbf{w}^t - \mathbf{w}^{t+1}\right)\|^2_{\mathbf{H}_2} \\
&= \|\mathbf{w}^t - \widetilde{\mathbf{w}}^t\|^2_{\mathbf{H}_2} - \|(\mathbf{w}^t - \widetilde{\mathbf{w}}^t) - \mathbf{M}(\mathbf{w}^t - \widetilde{\mathbf{w}}^t)\|^2_{\mathbf{H}_2} \\
&= \left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right)(2\mathbf{H}_2\mathbf{M} - \mathbf{M}^T\mathbf{H}\mathbf{M})\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right) \\
&= \left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right)(\mathbf{Q}_2^T + \mathbf{Q}_2 - \mathbf{M}^T\mathbf{H}\mathbf{M})\left(\mathbf{w}^t - \widetilde{\mathbf{w}}^t\right),
\end{aligned}$$

where the last equality comes from the identity $\mathbf{Q}_2 = \mathbf{H}_2\mathbf{M}$.

Substituting the above identity into (90), we have, for all $\mathbf{w} \in \Omega$,

$$2\left(\mathbf{w}-\widetilde{\mathbf{w}}^t\right)\mathbf{Q}_2\left(\mathbf{w}^t-\widetilde{\mathbf{w}}^t\right) = \|\mathbf{w}-\mathbf{w}^{t+1}\|_{\mathbf{H}_2}^2 - \|\mathbf{w}-\mathbf{w}^t\|_{\mathbf{H}_2}^2 + \|\mathbf{w}^t-\widetilde{\mathbf{w}}^t\|_{\mathbf{Q}_2^T+\mathbf{Q}_2-\mathbf{M}^T\mathbf{H}_2\mathbf{M}}^2$$

Plugging this identity into (82), our claim follows immediately. □

Then, we show the boundedness of the sequence $\{\mathbf{w}^t\}$ generated by the DL-GADMM (79), which essentially implies the convergence of $\{\mathbf{w}^t\}$.

Theorem 9 *Let $\{\mathbf{w}^t\}$ be the sequence generated by the DL-GADMM (79) with $\alpha \in (0, 2)$. Then, it holds that*

$$\sum_{t=0}^{\infty} \|\mathbf{w}^t - \widetilde{\mathbf{w}}^t\|_{\mathbf{Q}_2^T+\mathbf{Q}_2-\mathbf{M}^T\mathbf{H}_2\mathbf{M}}^2 \leq \|\mathbf{w}^0 - \mathbf{w}^*\|_{\mathbf{H}_2}^2, \tag{91}$$

where $\mathbf{H}_2$ is defined in (80).

Proof Setting $\mathbf{w} = \mathbf{w}^*$ in (89), we have

$$\begin{aligned} f(\mathbf{u}^*) - f(\mathbf{u}^t) + \left(\mathbf{w}^* - \widetilde{\mathbf{w}}^t\right)^T F(\mathbf{w}^*) \geq \frac{1}{2}\left(\|\mathbf{w}^* - \mathbf{w}^{t+1}\|_{\mathbf{H}_2}^2 - \|\mathbf{w}^* - \mathbf{w}^t\|_{\mathbf{H}_2}^2\right) \\ + \frac{1}{2}\|\mathbf{w}^t - \widetilde{\mathbf{w}}^t\|_{\mathbf{Q}_2^T+\mathbf{Q}_2-\mathbf{M}^T\mathbf{H}_2\mathbf{M}}^2. \end{aligned}$$

Then, recall (5), we have

$$\|\mathbf{w}^t - \widetilde{\mathbf{w}}^t\|_{\mathbf{Q}_2^T+\mathbf{Q}_2-\mathbf{M}^T\mathbf{H}_2\mathbf{M}}^2 \leq \|\mathbf{w}^t - \mathbf{w}^*\|_{\mathbf{H}_2}^2 - \|\mathbf{w}^{t+1} - \mathbf{w}^*\|_{\mathbf{H}_2}^2.$$

It is easy to see that $\mathbf{Q}_2^T + \mathbf{Q}_2 - \mathbf{M}^T\mathbf{H}_2\mathbf{M} \succeq \mathbf{0}$. Thus, it holds

$$\sum_{t=0}^{\infty} \|\mathbf{w}^t - \widetilde{\mathbf{w}}^t\|_{\mathbf{Q}_2^T+\mathbf{Q}_2-\mathbf{M}^T\mathbf{H}_2\mathbf{M}}^2 \leq \|\mathbf{w}^0 - \mathbf{w}^*\|_{\mathbf{H}_2}^2,$$

which completes the proof. □

Finally, we establish a worst-case $\mathcal{O}(1/k)$ convergence rate in a nonergodic sense for the DL-GADMM (79).

Theorem 10 *Let the sequence $\{\mathbf{w}^t\}$ be generated by the scheme DL-GADMM (79) with $\alpha \in (0, 2)$. It holds that*

$$\|\mathbf{w}^k - \mathbf{w}^{k+1}\|_{\mathbf{H}_2}^2 = \mathcal{O}(1/k). \tag{92}$$

Proof By the definition of $\mathbf{H}_2$ in (80), we have

$$\begin{aligned}\|\mathbf{w}^t-\mathbf{w}^{t+1}\|_{\mathbf{H}_2}^2 &= \|\mathbf{x}^t-\widetilde{\mathbf{x}}^t\|_{\mathbf{G}_1}^2+\|\mathbf{y}^t-\widetilde{\mathbf{y}}^t\|_{\mathbf{G}_2}^2+\frac{1}{\alpha\rho}\left(\|\rho\mathbf{B}\left(\mathbf{y}^t-\mathbf{y}^{t+1}\right)\|^2\right.\\ &\quad\left.+\|\gamma^t-\gamma^{t+1}\|^2+2(1-\alpha)\rho\left(\mathbf{y}^t-\mathbf{y}^{t+1}\right)^T\mathbf{B}^T\left(\gamma^t-\gamma^{t+1}\right)\right)\\ &= \|\mathbf{x}^t-\widetilde{\mathbf{x}}^t\|_{\mathbf{G}_1}^2+\|\mathbf{y}^t-\widetilde{\mathbf{y}}^t\|_{\mathbf{G}_2}^2+\frac{\alpha}{\rho}\|\gamma^t-\widetilde{\gamma}^t\|^2\\ &\quad-2\left(\mathbf{y}^t-\mathbf{y}^{t+1}\right)^T\mathbf{B}^T\left(\gamma^t-\gamma^{t+1}\right). \end{aligned}\tag{93}$$

Using (85, 88, 91) and (93), we obtain

$$\begin{aligned}\sum_{t=1}^k\|\mathbf{w}^t-\mathbf{w}^{t+1}\|_{\mathbf{H}_2}^2 &\le \frac{1}{c_\alpha}\sum_{t=1}^k\|\mathbf{w}^t-\widetilde{\mathbf{w}}^t\|_{\mathbf{Q}_2^T+\mathbf{Q}_2-\mathbf{M}^T\mathbf{H}_2\mathbf{M}}^2\\ &\quad+\sum_{t=1}^k\left(\|\mathbf{y}^{t-1}-\mathbf{y}^t\|_{\mathbf{G}_2}^2-\|\mathbf{y}^t-\mathbf{y}^{t+1}\|_{\mathbf{G}_2}^2\right)\\ &\le \frac{1}{c_\alpha}\|\mathbf{w}^0-\mathbf{w}^*\|_{\mathbf{H}_2}^2+\|\mathbf{y}^0-\mathbf{y}^1\|_{\mathbf{G}_2}^2.\end{aligned}$$

By Theorem 8, the sequence $\{\|\mathbf{w}^t-\mathbf{w}^{t+1}\|_{\mathbf{H}_2}^2\}$ is non-increasing. Thus, we have

$$\begin{aligned}k\|\mathbf{w}^k-\mathbf{w}^{k+1}\|_{\mathbf{H}_2}^2 &\le \sum_{t=1}^k\|\mathbf{w}^t-\mathbf{w}^{t+1}\|_{\mathbf{H}_2}^2\\ &\le \frac{1}{c_\alpha}\|\mathbf{w}^0-\mathbf{w}^*\|_{\mathbf{H}_2}^2+\|\mathbf{y}^0-\mathbf{y}^1\|_{\mathbf{G}_2}^2,\end{aligned}$$

and the assertion (92) is proved. □

Recall that for the sequence $\{\mathbf{w}^t\}$ generated by the DL-GADMM (79), it is reasonable to measure the accuracy of an iterate by $\|\mathbf{w}^t-\mathbf{w}^{t+1}\|_{\mathbf{H}_2}^2$. Thus, Theorem 10 demonstrates a worst-case $\mathcal{O}(1/k)$ convergence rate in a nonergodic sense for the DL-GADMM (79).

References

1. Anderson, T.W.: An introduction to multivariate statistical analysis, 3rd edn. Wiley (2003)
2. Bertsekas, D.P.: Constrained optimization and Lagrange multiplier methods. Academic Press, New York (1982)
3. Bickel, P.J., Levina, E.: Some theory for Fisher's linear discriminant function, naive Bayes', and some alternatives when there are many more variables than observations. Bernoulli **6**, 989–1010 (2004)
4. Blum, E., Oettli, W.: Mathematische Optimierung. Grundlagen und Verfahren. Ökonometrie und Unternehmensforschung. Springer, Berlin (1975)

5. Boley, D.: Local linear convergence of ADMM on quadratic or linear programs. SIAM J. Optim. **23**(4), 2183–2207 (2013)
6. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**, 1–122 (2011)
7. Cai, T.T., Liu, W.: A Direct estimation approach to sparse linear discriminant analysis. J. Amer. Stat. Assoc. **106**, 1566–1577 (2011)
8. Cai, X., Gu, G., He, B., Yuan, X.: A proximal point algorithm revisit on alternating direction method of multipliers. Sci. China Math. **56**(10), 2179–2186 (2013)
9. Candès, E.J., Tao, T.: The Dantzig selector: statistical estimation when p is much larger than n. Ann. Stat. **35**, 2313–2351 (2007)
10. Clemmensen, L., Hastie, T., Witten, D., Ersbøll, B.: Sparse discriminant analysis. Technometrics **53**, 406–413 (2011)
11. Deng, W., Yin, W.: On the global and linear convergence of the generalized alternating direction method of multipliers. Manuscript (2012)
12. Eckstein, J.: Parallel alternating direction multiplier decomposition of convex programs. J. Optim. Theory Appli. **80**(1), 39–62 (1994)
13. Eckstein, J., Yao, W.: Augmented Lagrangian and alternating direction methods for convex optimization: a tutorial and some illustrative computational results. RUTCOR Research Report RRR 32–2012 (2012)
14. Eckstein, J., Bertsekas, D.: On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. **55**, 293–318 (1992)
15. Fan, J., Fan, Y.: High dimensional classification using features annealed independence rules. Ann. Stat. **36**, 2605–2037 (2008)
16. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**, 1348–1360 (2001)
17. Fan, J., Feng, Y., Tong, X.: A road to classification in high dimensional space: the regularized optimal affine discriminant. J. R. Stat. Soc. Series B Stat. Methodol. **74**, 745–771 (2012)
18. Fan, J., Zhang, J., Yu, K.: Vast portfolio selection with gross-exposure constraints. J. Am. Stat. Assoc. **107**, 592–606 (2012)
19. Fazeland, M., Hindi, H., Boyd, S.: A rank minimization heuristic with application to minimum order system approximation. Proc. Am. Control Conf. (2001)
20. Fortin, M., Glowinski, R.: Augmented Lagrangian methods: applications to the numerical solutions of boundary value problems Stud. Math. Appl. 15. NorthHolland, Amsterdam (1983)
21. Gabay, D.: Applications of the method of multipliers to variational inequalities, Augmented Lagrange Methods: applications to the solution of boundary-valued problems. Fortin, M. Glowinski, R. eds. North Holland pp. 299–331 (1983)
22. Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite-element approximations. Comput. Math. Appl. **2**, 17–40 (1976)
23. Glowinski, R.: On alternating directon methods of multipliers: a historical perspective. Springer Proceedings of a Conference Dedicated to J. Periaux (to appear)
24. Glowinski, R., Marrocco, A.: Approximation par éléments finis d'ordre un et résolution par pénalisation-dualité d'une classe de problèmes non linéaires. R.A.I.R.O., R2, pp. 41–76 (1975)
25. Gol'shtein, E.G., Tret'yakov, N.V.: Modified Lagrangian in convex programming and their generalizations. Math. Program. Study **10**, 86–97 (1979)
26. Grier, H.E., Krailo, M.D., Tarbell, N.J., Link, M.P., Fryer, C.J., Pritchard, D.J., Gebhardt, M.C., Dickman, P.S., Perlman, E.J., Meyers, P.A.: Addition of ifosfamide and etoposide to standard chemotherapy for Ewing's sarcoma and primitive neuroectodermal tumor of bone. New Eng. J. Med. 348, 694–701 (2003)
27. Han, D., Yuan, X.: Local linear convergence of the alternating direction method of multipliers for quadratic programs. SIAM J. Numer. Anal. **51**(6), 3446–3457 (2013)
28. Hans, C.P., Weisenburger, D.D., Greiner, T.C., Gascone, R.D., Delabie, J., Ott, G., M'uller-Hermelink, H., Campo, E., Braziel, R., Elaine, S.: Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray. Blood 103, 275–282 (2004)
29. He, B., Liao, L.-Z., Han, D.R., Yang, H.: A new inexact alternating directions method for monotone variational inequalities. Math. Program. **92**, 103–118 (2002)
30. He, B., Yang, H.: Some convergence properties of a method of multipliers for linearly constrained monotone variational inequalities. Oper. Res. Let. **23**, 151–161 (1998)

31. He, B., Yuan, X.: On the $O(1/n)$ convergence rate of Douglas-Rachford alternating direction method. SIAM J. Numer. Anal. **50**, 700–709 (2012)
32. He, B., Yuan, X.: On non-ergodic convergence rate of Douglas–Rachford alternating direction method of multipliers. Numerische Mathematik (to appear)
33. He, B., Yuan, X.: On convergence rate of the Douglas–Rachford operator splitting method. Math. Program (to appear)
34. Hestenes, M.R.: Multiplier and gradient methods. J. Optim. Theory Appl. **4**, 302–320 (1969)
35. James, G.M., Paulson, C., Rusmevichientong, P.: The constrained LASSO. Manuscript (2012)
36. Lions, P.L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operator. SIAM J. Numer. Anal. **16**, 964–979 (1979)
37. Martinet, B.: Regularisation, d'inéquations variationelles par approximations succesives. Rev. Francaise d'Inform. Recherche Oper. **4**, 154–159 (1970)
38. McCall, M.N., Bolstad, B.M., Irizarry, R.A.: Frozen robust multiarray analysis (fRMA). Biostatistics **11**, 242–253 (2010)
39. Nemirovsky, A.S., Yudin, D.B.: Problem complexity and method efficiency in optimization, Wiley-Interscience series in discrete mathematics. Wiley, New York (1983)
40. Nesterov, Y.E.: A method for solving the convex programming problem with convergence rate $O(1/k^2)$. Dokl. Akad. Nauk SSSR **269**, 543–547 (1983)
41. Ng, M.K., Wang, F., Yuan, X.: Inexact alternating direction methods for image recovery. SIAM J. Sci. Comput. **33**(4), 1643–1668 (2011)
42. Powell, M.J.D.: A method for nonlinear constraints in minimization problems. In: Fletcher, R. (ed.) Optimization. Academic Press (1969)
43. Shao, J., Wang, Y., Deng, X., Wang, S.: Sparse linear discriminant analysis by thresholding for high dimensional data. Ann. Stat. **39**, 1241–1265 (2011)
44. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Series B Stat. Methodol. **58**, 267–288 (1996)
45. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. J. R. Stat. Soc. Series B Stat. Methodol. **67**, 91–108 (2005)
46. Tibshirani, R.J., Taylor, J.: The solution path of the generalized lasso. Ann. Stat. **39**, 1335–1371 (2011)
47. Wang, L., Zhu, J., Zou, H.: Hybrid huberized support vector machines for microarray classification and gene selection. Bioinformatics **24**, 412–419 (2008)
48. Wang, X., Yuan, X.: The linearized alternating direction method of multipliers for Dantzig Selector. SIAM J. Sci. Comput. **34**, 2782–2811 (2012)
49. Witten, D.M., Tibshirani, R.: Penalized classification using Fisher's linear discriminant. J. R. Stat. Soc. Series B Stat. Methodol. **73**, 753–772 (2011)
50. Yang, J., Yuan, X.: Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization. Math. Comput. **82**, 301–329 (2013)
51. Zhang, C.-H.: Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. **38**, 894–942 (2010)
52. Zhang, X.Q., Burger, M., Osher, S.: A unified primal-dual algorithm framework based on Bregman iteration. J. Sci. Comput. **6**, 20–46 (2010)
53. Zhang, X.Q., Burger, M., Bresson, X., Osher, S.: Bregmanized nonlocal regularization for deconvolution and sparse reconstruction. SIAM J. Imag. Sci. **3**(3), 253–276 (2010)
54. Zou, H.: The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. **101**, 1418–1429 (2006)