A Linearly Convergent Majorized ADMM with Indefinite Proximal Terms for Convex Composite Programming and Its Applications

Ning Zhang,* Jia Wu,† Liwei Zhang ‡

February 7, 2018

Abstract This paper aims to study a majorized alternating direction method of multipliers with indefinite proximal terms (iPADMM) for convex composite optimization problems. We show that the majorized iPADMM for 2-block convex optimization problems converges globally under weaker conditions than those used in the literature and exhibits a linear convergence rate under a local error bound condition. Based on these, we establish the linear rate convergence results for a symmetric Gaussian-Seidel based majorized iPADMM, which is designed for multi-block composite convex optimization problems. Moreover, we apply the majorized iPADMM to solve different types of regularized logistic regression problems. The numerical results on both synthetic and real datasets demonstrate the efficiency of the majorized iPADMM and also illustrate the effectiveness of the introduced indefinite proximal terms.

Keywords Alternating direction method of multiplier · Linear rate convergence · Indefinite proximal term · Logistic regression · Symmetric Gaussian-Seidel

Mathematics Subject Classification 90C25 · 90C30 · 65K05 · 62J12.

1 Introduction

In this paper, we consider the following convex composite optimization problem:

$$\begin{align*}
\min_{y,z} & \quad p(y) + f(y) + q(z) + g(z) \\
\text{s.t.} & \quad \mathcal{A}^*y + \mathcal{B}^*z = c, \\
& \quad y \in \mathcal{Y}, z \in \mathcal{Z},
\end{align*}$$

(1.1)

where $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ are given finite dimensional Euclidean spaces each equipped with an inner product $\langle \cdot, \cdot \rangle$ and its induced norm $\| \cdot \|_2$; $f : \mathcal{Y} \to (-\infty, +\infty)$ and $g : \mathcal{Z} \to (-\infty, +\infty)$ are two convex functions with Lipschitz continuous gradients on $\mathcal{Y}$ and $\mathcal{Z}$, respectively; $p : \mathcal{Y} \to (-\infty, +\infty]$ and

*Future Resilient Systems, Singapore-ETH center; Department of Mathematics, National University of Singapore, Singapore. (matzhn@nus.edu.sg). The Singapore-ETH Centre (SEC) was established as a collaboration between ETH Zurich and National Research Foundation (NRF) Singapore (FI 370074011) under the auspices of the NRF’s Campus for Research Excellence and Technological Enterprise (CREATE) programme.

†School of Mathematical Sciences, Dalian University of Technology, China. (lwzhang@dlt.edu.cn). This author’s research is supported by the National Natural Science Foundation of China under project 11571059 and 11731013.
Let \( q : Z \to (-\infty, +\infty] \) are two closed proper convex (not necessarily smooth) functions; \( A^* : Y \to X \) and \( B^* : Z \to X \) are adjoints of the linear operators \( A : X \to Y \) and \( B : X \to Z \), respectively; and \( c \in X \).

Let \( M : X \to X \) be a self-adjoint linear operator (not necessarily semidefinite positive), denote \( \|x\|^2_M := \langle x, Mx \rangle \). Let \( \sigma > 0 \) be a given parameter. The augmented Lagrangian function of (1.1) is defined by

\[
L_\sigma(y, z; x) = f(y) + p(y) + g(z) + \langle x, A^*x + B^*y - c \rangle + \frac{\sigma}{2}\|A^*x + B^*y - c\|^2.
\]

Consider the following general 2-block ADMM iterative scheme,

\[
\begin{align*}
g^{k+1} &\in \arg\min \left\{ L_\sigma(y, z; x^k) + \frac{1}{2}\|y - y^k\|^2_M \mid y \in Y \right\}, \\
z^{k+1} &\in \arg\min \left\{ L_\sigma(y^{k+1}, z; x^k) + \frac{1}{2}\|z - z^k\|^2_F \mid z \in Z \right\}, \\
x^{k+1} &= x^k + \tau \sigma (A^*y^{k+1} + B^*z^{k+1} - c).
\end{align*}
\]  

It is well known that if \( S = 0 \) and \( T = 0 \), the iterative scheme (1.2) is exactly the classic ADMM designed by Glowinski and Marroco [22] and Gabay and Mercier [21]; if \( S > 0 \), \( T > 0 \) and \( \tau = 1 \), iterative scheme (1.2) reduces to the method of the proximal ADMM introduced by Eckstein [14]; if both \( S \) and \( T \) are self-adjoint positive semidefinite linear operators, \( \tau \in (0, (1 + \sqrt{3})/2) \), iterative scheme (1.2) is known as the semiproximal ADMM (sPADMM) which is proposed by Fazel et al. [17]. To know more about the above mentioned works and their relationships with well known methods, such as proximal point algorithm (PPA) and Douglas-Rachford (DR) splitting method, we refer the readers to [22, 18, 16, 15, 14, 23, 7, 24, 34, 27].

One of the most important motivations behind ADMM is to fully use the separable structures in the problems. In other words, a potential assumption of using ADMM is that each subproblem can be efficiently solved. Generally speaking, if \( f \) or \( g \) is not a quadratic or linear function, the corresponding subproblem does not have closed-form solutions or cannot be solved easily. In order to continue enjoying benefits of the separable structure, Li et al. [34] extended the sPADMM to a majorized ADMM with indefinite proximal terms (iPADMM). Compared with the majorized techniques mentioned in [7, 1], the majorized iPADMM uses the positive semidefinite operators \( \hat{\Sigma}_f \) and \( \hat{\Sigma}_g \) (see (1.5) and (1.6)) instead of the Lipschitz constants of the gradient mappings \( \nabla f \) and \( \nabla g \). The motivation behind this is for the better numerical performance. This will be illustrated in Table 2. Li et al. [34] established the global convergence and the iteration-complexity in the nonergodic sense of the majorized iPADMM, but not the rate of convergence.

In this paper, we further study the convergence and the rate of convergence of the majorized iPADMM presented in [34]. Now, we recall the majorized iPADMM. Since both \( f(\cdot) \) and \( g(\cdot) \) are smooth convex functions, there exist self-adjoint and positive semidefinite linear operators \( \Sigma_f \) and \( \Sigma_g \) such that for any \( y, y' \in Y \) and any \( z, z' \in Z \),

\[
\begin{align*}
f(y) &\geq f(y') + \langle \nabla f(y'), y - y' \rangle + \frac{1}{2}\|y - y'\|^2_{\Sigma_f}, \\
g(z) &\geq g(z') + \langle \nabla g(z'), z - z' \rangle + \frac{1}{2}\|z - z'\|^2_{\Sigma_g}.
\end{align*}
\]  

In addition, by the condition that the gradients \( \nabla f(\cdot) \) and \( \nabla g(\cdot) \) are Lipschitz continuous, we know that there exist self-adjoint and positive semidefinite linear operators \( \hat{\Sigma}_f \geq \Sigma_f \) and \( \hat{\Sigma}_g \geq \Sigma_g \) such
that for any \( y, y' \in \mathcal{Y} \) and any \( z, z' \in \mathcal{Z} \),
\[
f(y) \leq \hat{f}(y, y') := f(y') + \langle \nabla f(y'), y - y' \rangle + \frac{1}{2} \|y - y'\|_\Sigma_f^2, \tag{1.5}
\]
\[
g(z) \leq \hat{g}(z, z') := g(z') + \langle \nabla g(z'), z - z' \rangle + \frac{1}{2} \|z - z'\|_\Sigma_g^2. \tag{1.6}
\]
For any given \((y', z') \in \mathcal{Y} \times \mathcal{Z}\), \( \sigma \in (0, +\infty) \) and any \((y, z, x) \in \mathcal{Y} \times \mathcal{Z} \times \mathcal{X}\), the majorized augmented Lagrangian function is defined as
\[
\hat{L}_\sigma(y, z, x; y', z') := \hat{f}(y, y') + p(y) + \hat{g}(z, z') + q(z) + \langle x, A^* y + B^* z - c \rangle
+n_2 \|A^* y + B^* z - c\|^2, \tag{1.7}
\]
where \( \hat{f} \) and \( \hat{g} \) are defined by (1.5) and (1.6), respectively. Then the majorized iPADMM can be described as follows.

**Algorithm 1** A majorized iPADMM

Let \( \sigma > 0 \) and \( \tau \in (0, (1 + \sqrt{5})/2) \) be given parameters. Let \( S \) and \( T \) be given self-adjoint, possibly indefinite, linear operators defined on \( \mathcal{Y} \) and \( \mathcal{Z} \), respectively. Input \((y^0, z^0, x^0) \in \text{dom } p \times \text{dom } q \times \mathcal{X}\). Set \( k := 0 \).

**Step 1.** Compute
\[
\begin{align*}
    y^{k+1} &\in \arg \min_{y \in \mathcal{Y}} \hat{L}_\sigma(y, z^k, x^k; y^k, z^k) + \frac{1}{2} \|y - y^k\|^2_\Sigma, \tag{1.8a} \\
    &= \arg \min_{y \in \mathcal{Y}} p(y) + \langle y, \nabla f(y^k) + A x^k \rangle + \frac{\sigma}{2} \|A^* y + B^* z^k - c\|^2 + \frac{1}{2} \|y - y^k\|^2_{\tilde{\Sigma}_f + S}, \\
    z^{k+1} &\in \arg \min_{z \in \mathcal{Z}} \hat{L}_\sigma(y^{k+1}, z, x^k; y^k, z^k) + \frac{1}{2} \|z - z^k\|^2_\Sigma, \tag{1.8b} \\
    &= \arg \min_{z \in \mathcal{Z}} q(z) + \langle z, \nabla g(z^k) + B x^k \rangle + \frac{\sigma}{2} \|A^* y^{k+1} + B^* z - c\|^2 + \frac{1}{2} \|z - z^k\|^2_{\tilde{\Sigma}_g + T}, \\
    x^{k+1} &:= x^k + \tau \sigma (A^* y^{k+1} + B^* z^{k+1} - c). \tag{1.8c}
\end{align*}
\]

**Step 2.** If a termination criterion is not met, set \( k := k + 1 \) and go to Step 1.

As mentioned in Han et al. [27], a large amount of literature focus on designing variant forms of ADMM and their applications. The literature on linear rate convergence, especially the Q-linear rate, on the other hand, is relatively sparse. Based on the close connections among DR, PPA and classic ADMM with \( \tau = 1 \), the R-linear rate convergence of ADMM can be established under conditions such as strong monotonicity, error bounds and etc. By using some different technical tools, the local Q-linear rate convergence of sequence \( \{ (z_k, x_k) \} \), generated by the ADMM as defined in Eckstein and Berstekas [15], is established by Han and Yuan [28]. Yang and Han [49] show the Q-linear rate convergence of ADMM for solving a class of convex piecewise linear-quadratic problems. In the same paper, the Q-linear rate convergence of linearized ADMM (L-ADMM) with positive definite proximal operators and \( \tau = 1 \) is also proved. Under assumptions that either \( f(y) + p(y) \) or \( q(z) + q(z) \) is strongly convex with a Lipschitz continuous gradient and others, Deng and Yin [9] provide some scenarios on the R-linear and Q-linear rates of convergence of ADMM. Recently,
based on the easy-to-use convergence theorem in Fazel et al. [17], the Q-linear rate convergence theorem for sPADMM with $\tau \in (0, (1 + \sqrt{5})/2)$ is established in [27] under a calmness condition, which holds automatically for convex composite piecewise-linear programming. To know more about the convergence rate analysis of ADMM, we refer to [27] and the references therein. Inspired by [27], we resolve the Q-linear rate convergence issue for the majorized iPADMM scheme with $\tau \in (0, (1 + \sqrt{5})/2)$, in addition to some improvements on the global convergence analysis.

The main contributions of this paper are threefold. First, we refine the conditions in [34, Theorem 10 (b)] and establish the global convergence of the majorized iPADMM with

$$\frac{1}{2} \Sigma_f + S + \sigma AA^* \succ 0 \& \frac{1}{2} \Sigma_g + T + \sigma BB^* \succ 0$$

being replaced by

$$\frac{1}{2} \tilde{\Sigma}_f + S + \sigma AA^* \succ 0 \& \frac{1}{2} \tilde{\Sigma}_g + T + \sigma BB^* \succ 0.$$  \hfill (1.10)

Note that conditions (1.9) coincide with (1.10) when $f$ and $g$ are linear/quadratic. However, the improvement from (1.10) is significant when $f$ or $g$ is not linear/quadratic. In fact, there are many widely used loss functions in statistical inference and machine learning that are not linear/quadratic, such as the logistic loss function, the multinomial logistic loss function (see e.g. [19]), and the cox loss function (see e.g. [47]).

Second, we build up a general Q-linear rate convergence theorem based on an inequality associated with the iteration sequences generated by majorized iPADMM. Note that the convergence results on sPADMM [27] are no longer applicable due to the indefiniteness of the proximal terms.

Third, inspired by the global convergence and excellent numerical performance of symmetric Gaussian-Seidel based multi-block sPADMM (sGS-sPADMM) for multi-block linearly constrained convex programming, we present a symmetric Gaussian-Seidel based multi-block majorized iPADMM (majorized sGS-iPADMM). The linear rate convergence results for majorized sGS-iPADMM are established through converting it into an equivalent 2-block majorized iPADMM. This is one of the most important motivations that we consider the objective function in (1.1) in the form of $f(y) + p(y)$ and $g(z) + q(z)$. It is worth mentioning that the sGS-sPDMM was initially presented by Li et al. [35]. For more discussions about symmetric Gaussian-Seidel techniques, we refer to [3, 36, 35, 37] and the references therein.

The rest of the paper is organized as follows. In section 2, we give some preliminaries that will be frequently used in other sections. In section 3, we refine the convergence result [34, Theorem 10 (b)] and then establish a general Q-linear rate convergence theorem under a metric subregularity condition. In section 4, we propose a majorized sGS-iPADMM for the multi-block composite optimization problem. Moreover, we show its convergence rate by establishing the relationship between the majorized sGS-iPADMM and the 2-block majorized iPADMM. In section 5, we apply the majorized iPADMM to three types of regularized logistic regression and then present the numerical results. Finally, we give some concluding remarks and future works in section 6.

2 Preliminaries

In this section, we summarize and study some preliminaries that will be used in the subsequent analysis. Let $(\tilde{y}, \tilde{z})$ be the optimal solution of problem (1.1). If there exists $\tilde{x} \in \mathcal{X}$ such that $(\tilde{y}, \tilde{z}, \tilde{x})$
satisfies the following Karush-Kuhn-Tucker (KKT) system:

\[
\begin{cases}
0 \in \partial p(y) + \nabla f(y) + Ax, \\
0 \in \partial q(z) + \nabla g(z) + Bx, \\
A^*y + B^*z - c = 0,
\end{cases}
\]  

(2.1)

then, \((\bar{y}, \bar{z}, \bar{x})\) is called a KKT point of problem (1.1). Denote the set of KKT points by \(\Omega\).

For any convex function \(\theta : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}\), the Moreau-Yosida proximal mapping \(\text{Pr}_\theta(\cdot)\) associated with \(\theta\) is defined by

\[
\text{Pr}_\theta(y) := \arg \min_{y' \in \mathcal{Y}} \left\{ \theta(y') + \frac{1}{2} \|y' - y\|^2 \right\}, \quad \forall y \in \mathcal{Y}.
\]

It is well known that the Moreau-Yosida proximal mapping \(\text{Pr}_\theta(\cdot)\) is globally Lipschitz continuous with modulus one, see e.g. [29, 33].

Denote \(u := (y, z, x) \in U\) with \(U := \mathcal{Y} \times \mathcal{Z} \times \mathcal{X}\). Define the KKT mapping \(\mathcal{R} : U \rightarrow U\) as

\[
\mathcal{R}(u) := \begin{pmatrix}
y - \text{Pr}_p[y - (\nabla f(y) + Ax)] \\
z - \text{Pr}_q[z - (\nabla g(z) + Bx)] \\
c - A^*y - B^*z
\end{pmatrix}.
\]  

(2.2)

From [40], we know that \(u \in \Omega\) if and only if \(\mathcal{R}(u) = 0\).

Let \(\mathcal{F} : \mathcal{X} \rightrightarrows \mathcal{Y}\) be a multivalued mapping. Denote its inverse by \(\mathcal{F}^{-1}\). Define the graph of multivalued function \(\mathcal{F}\) as follows

\[
\text{gph}\mathcal{F} := \{(x, y) \in \mathcal{X} \times \mathcal{Y} | y \in \mathcal{F}(x)\}.
\]

**Definition 2.1.** A multivalued mapping \(\mathcal{F} : \mathcal{X} \rightrightarrows \mathcal{Y}\) is said to be metrically subregular at \((\bar{x}, \bar{y}) \in \text{gph}\mathcal{F}\) with modulus \(\eta > 0\) if there exists a neighborhood \(U\) of \(\bar{x}\) such that

\[
\text{dist}(x, \mathcal{F}^{-1}(\bar{y})) \leq \eta \text{dist}(\bar{y}, \mathcal{F}(x)), \quad \forall x \in U.
\]

The definition of metric subregularity is directly from [13, Definition 3.1]. It is well known that \(\mathcal{F}\) is metrically subregular at \((\bar{x}, \bar{y}) \in \text{gph}\mathcal{F}\) if and only if its inverse multivalued mapping \(\mathcal{F}^{-1}\) is calm (c.f. [50, Definition 2.6], [12, 3.8(3H)]) at \((\bar{y}, \bar{x}) \in \text{gph}\mathcal{F}^{-1}\).

From [42, Proposition 1] and [45], we know that if \(\mathcal{F}\) is piecewise polyhedral or \(\mathcal{F}\) is the subdifferential mapping of a convex piecewise linear-quadratic function, then \(\mathcal{F}\) is metrically subregular at \((\bar{x}, \bar{y}) \in \text{gph}\mathcal{F}\). Till now, numerous works have been done to study the sufficient conditions of calmness of KKT solution mappings, we refer to [10, 8, 36, 25] and the references therein.

In order to establish the linear rate convergence of the majorized iPADMM, we need the metric subregularity of the KKT mapping \(\mathcal{R}\). From the Definition 2.1, the metric subregularity of \(\mathcal{R}\) at \((\bar{u}, 0) \in \text{gph}\mathcal{R}\) with modulus \(\eta > 0\) can be described as: there exists a scalar \(\rho > 0\) such that

\[
\text{dist}(u, \Omega) \leq \eta \|\mathcal{R}(u)\|, \quad \forall u \in \{u \in U : \|u - \bar{u}\| \leq \rho\}.
\]  

(2.3)

The above condition is also referred to as the existence of a local error bound. From [36, Theorem 1 & Remark 1], we know that the KKT mappings corresponding to the Lasso, elastic net Lasso regularized logistic regression models are metrically subregular at \((\bar{u}, 0) \in \text{gph}\mathcal{R}\).
Since for any proper closed convex function $\theta : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$, the subdifferential $\partial \theta(\cdot)$ is a monotone multivalued mapping (see [44]), i.e., for any $y^1, y^2 \in \text{dom}\theta$, it holds that
\[
\langle \zeta^1 - \zeta^2, y^1 - y^2 \rangle \geq 0, \quad \forall \zeta^1 \in \partial \theta(y^1), \forall \zeta^2 \in \partial \theta(y^2).
\] (2.4)

**Proposition 2.1.** [5, Proposition 2.6.5] Let multivalued mapping $F$ be Lipschitz on an open convex set $\mathcal{V}$ in $\mathbb{R}^n$, and let $x$ and $y$ be points in $\mathcal{V}$. Then one has
\[
F(y) - F(x) \in \text{conv} \partial F([x, y])(y - x),
\]
where $\text{conv} \partial F([x, y])(y - x)$ denotes the convex hull of all points of the form $\zeta(y - x)$, where $\zeta \in \partial F(u)$ for some point $u$ in $[x, y]$.

Throughout the subsequent analysis, we always assume that the following two assumptions hold.

**Assumption 2.1.** The KKT system (2.1) has at least one solution, i.e., $\Omega \neq \emptyset$.

**Assumption 2.2.** The two self-adjoint linear operators $S : \mathcal{Y} \to \mathcal{Y}$ and $T : \mathcal{Z} \to \mathcal{Z}$ in majorized iPADMM satisfy
\[
S \succeq -\frac{1}{2} \hat{\Sigma}_f & \quad T \succeq -\frac{1}{2} \hat{\Sigma}_g.
\] (2.5)

**Remark 2.1.** Assumption 2.2 means that the proximal terms $S$ and $T$ cannot be too indefinite as long as $\hat{\Sigma}_f$ and $\hat{\Sigma}_g$ are not very big. Note that $\hat{\Sigma}_f$ and $\hat{\Sigma}_g$ should be chosen as small as possible provided (1.5) and (1.6) are satisfied. For example, when $f$ is a convex quadratic function, we choose $\hat{\Sigma}_f = \Sigma_f = \nabla^2 f$, where $\nabla^2 f$ is the Hessian matrix of $f$.

### 3 Q-Linear rate of convergence of the majorized iPADMM

This section aims to analyze the convergence rate of the majorized iPADMM for solving (1.1). We show that the algorithm achieves a Q-linear rate of convergence under some mild conditions. Before formally stating our main results, we first give some technical results.

#### 3.1 Technical lemmas

For notational convenience, for any $\tau \in (0, +\infty)$, define
\[
s_\tau := \frac{5 - \tau - 3 \min(\tau, \tau^{-1})}{4}, \quad t_\tau := \frac{1 - \tau + \min(\tau, \tau^{-1})}{2},
\]
and two self-adjoint linear operators:
\[
\mathcal{M} := \text{Diag} \left( \hat{\Sigma}_f + S, \hat{\Sigma}_g + T + \sigma BB^*, (\tau \sigma)^{-1}I \right) + s_\tau \sigma E E^*, \quad (3.1)
\]
\[
\mathcal{H} := \text{Diag} \left( \mathcal{H}_f, \mathcal{H}_g, t_\tau (\tau^2 \sigma)^{-1}I \right) + \frac{1}{4} t_\tau \sigma E E^*, \quad (3.2)
\]
where
\[
\mathcal{H}_f := \frac{1}{2} \hat{\Sigma}_f + S, \quad \mathcal{H}_g := \frac{1}{2} \hat{\Sigma}_g + T + 2 t_\tau \sigma BB^*, \quad (3.3)
\]
and $\mathcal{E} : \mathcal{X} \to \mathcal{Y} \times \mathcal{Z} \times \mathcal{X}$ is the linear operator such that its adjoint $\mathcal{E}^*$ satisfies $\mathcal{E}^*(y, z, x) = A^*y + B^*z$. For a given self-adjoint linear operator $\mathcal{G} : \mathcal{X} \to \mathcal{X}$, we denote the largest eigenvalue by $\lambda_{\text{max}}(\mathcal{G})$ and for any $k \geq 0$,

$$r^k := A^*y^k + B^*z^k - c.$$  

The proofs of the lemmas in this subsection are all presented in Appendix for readability.

**Lemma 3.1.** Let $\{u^k := (y^k, z^k, x^k)\}$ be the infinite sequence generated by the majorized iPADMM. Then for any $k \geq 0$,

$$\|R(u^{k+1})\|^2 \leq \|u^{k+1} - u^k\|^2_{\mathcal{H}_0},$$  

(3.4)

where

$$\mathcal{H}_0 := \max\{\kappa_1, \kappa_2, \kappa_3\} \text{Diag} (I, I + \sigma BB^*, (\tau^2\sigma)^{-1}I)$$

with

$$\kappa_1 := 3(\lambda_{\text{max}}(S + \frac{1}{2}\bar{\Sigma}_f) + \frac{1}{2}\lambda_{\text{max}}(\bar{\Sigma}_f))^2,$$

$$\kappa_2 := \max\{2(\lambda_{\text{max}}(T + \frac{1}{2}\bar{\Sigma}_g) + \frac{1}{2}\lambda_{\text{max}}(\bar{\Sigma}_g))^2, 3\sigma\lambda_{\text{max}}(A^*A)\},$$

$$\kappa_3 := \sigma^{-1} + (1 - \tau)^2\sigma(3\lambda_{\text{max}}(A^*A) + 2\lambda_{\text{max}}(BB^*)).$$  

The above lemma is inspired by [27, Lemma 1], but its proof is more complicated due to the majorization techniques. In order to refine the global convergence results of the majorized iPADMM presented in [34, Theorem 10(b)], we need the following lemma.

**Lemma 3.2.** Let $h : \mathcal{X} \to \mathbb{R}$ be a smooth convex function and there is a self-adjoint positive semidefinite linear operator $\mathcal{P}$ such that, for any given $\bar{x} \in \mathcal{X}$,

$$h(x) \leq h(\bar{x}) + \langle \nabla h(\bar{x}), x - \bar{x} \rangle + \frac{1}{2}\|x - \bar{x}\|^2_{\mathcal{P}}, \quad \forall x \in \mathcal{X}. $$  

(3.5)

Then it holds that

$$\langle \nabla h(x) - \nabla h(\bar{x}), y - \bar{x} \rangle \geq -\frac{1}{4}\|x - y\|^2_{\mathcal{P}}, \quad \forall x, y \in \mathcal{X}. $$  

(3.6)

**Lemma 3.3.** Let $\{(y^k, z^k, x^k)\}$ be the infinite sequence generated by the majorized iPADMM. Then, for any $\bar{\omega} := (\bar{y}, \bar{z}, \bar{x}) \in \overline{\Omega}$, $\tau > 0$ and $k \geq 0$, we have

$$\phi_k - \phi_{k+1} \geq \|y^{k+1} - y^k\|^2_{\frac{1}{2}\bar{\Sigma}_{r+S}} + \|z^{k+1} - z^k\|^2_{\frac{1}{2}\bar{\Sigma}_{r+T}} + (1 - \tau)\sigma\|x^{k+1}\|^2 + \sigma\|A^*y^{k+1} + B^*z^{k+1} - c\|^2,$$  

(3.7)

where for any $k \geq 0$,

$$\phi_k := (\tau\sigma)^{-1}\|x^k - \bar{x}\|^2 + \|y^k - \bar{y}\|^2_{\frac{1}{2}\bar{\Sigma}_{r+S}} + \|z^k - \bar{z}\|^2_{\frac{1}{2}\bar{\Sigma}_{r+T+\sigma BB^*}}.$$  

(3.8)

Since the proof of the following lemma is not much different from the one in [34, Theorem 10, Inequality (55)] except for replacing Inequality (33) in [34] by (3.7) in Lemma 3.3, we include an outline in Appendix A.4.
Lemma 3.4. Let \( \{u_k := (y_k, z_k, x_k)\} \) be the infinite sequence generated by the majorized iPADMM. For each \( k \) and any KKT point \( \bar{u} := (\bar{y}, \bar{z}, \bar{x}) \), let \( \phi_k \) be defined in (3.8). Then, for any \( k \geq 1 \), one has
\[
\phi_k + (1 - \min(\tau, \tau^{-1})) \sigma \| r_k \|^2 + \| z_k - z_{k-1} \|^2_{\Sigma_y + T} - \phi_{k+1} - (1 - \min(\tau, \tau^{-1})) \sigma \| r_{k+1} \|^2 + \| z_{k+1} - z_k \|^2_{\Sigma_y + T} \geq t_{k+1} - (1 - \tau) \sigma \| r_{k+1} \|^2,
\]
where
\[
t_{k+1} := \| y^{k+1} - y^k \|^2_{H_f} + \| z^{k+1} - z^k \|^2_{H_g}.
\]

The following lemma plays an important role in establishing the global convergence theorem for the majorized iPDAhMM:

Lemma 3.5. Let \( \tau \in (0, (1 + \sqrt{5})/2) \), \( M \) and \( H \) be defined in (3.1) and (3.2), respectively. Then, \( \frac{1}{2} \Sigma_f + \sigma A A^* > 0 \) & \( \frac{1}{2} \Sigma_g + \sigma B B^* > 0 \) \( \Leftrightarrow H > 0 \) \( \Rightarrow M > 0 \). (3.11)

3.2 Convergence analysis

In this subsection, we investigate the rate of convergence of majorized iPDAhMM for solving (1.1). Inspired by [27, Proposition 4], we first develop a key inequality needed for proving the linear rate convergence for the majorized iPDAhMM.

Proposition 3.1. Let \( \tau \in (0, (1 + \sqrt{5})/2) \) and \( \{u_k := (y^k, z^k, x^k)\} \) be the infinite sequence generated by the majorized iPDAhMM. Then for any KKT point \( \bar{u} := (\bar{y}, \bar{z}, \bar{x}) \) and any \( k \geq 1 \),
\[
\| u^{k+1} - \bar{u} \|^2_M + \| z^{k+1} - z^k \|^2_{\Sigma_y + T} \leq \| u^k - \bar{u} \|^2_M + \| z^k - z_{k-1} \|^2_{\Sigma_y + T} - \| u^{k+1} - u^k \|^2_H.
\]
Consequently, we have for all \( k \geq 1 \),
\[
dist^2_M(u^{k+1}, \bar{u}) + \| z^{k+1} - z^k \|^2_{\Sigma_y + T} \leq \dist^2_M(u^k, \bar{u}) + \| z^k - z_{k-1} \|^2_{\Sigma_y + T} - \| u^{k+1} - u^k \|^2_H.
\]

Proof. By reorganizing the inequality in (3.9), one has
\[
\left[ \phi_{k+1} + s_\tau \sigma \| r_{k+1} \|^2 + \| z_{k+1} - z^k \|^2_{\Sigma_y + T} \right] - \left[ \phi_k + s_\tau \sigma \| r_k \|^2 + \| z^k - z_{k-1} \|^2_{\Sigma_y + T} \right] \leq - \left\{ t_{k+1} \sigma \| r_{k+1} \|^2 + \frac{1}{2} \sigma t_\tau (\| r_{k+1} \|^2 + \| r_k \|^2) + t_{k+1} \right\}.
\]

From definitions of \( x^{k+1} \) and \( (\bar{y}, \bar{z}, \bar{x}) \), we have
\[
r^{k+1} - \bar{r} = (\tau \sigma)^{-1} (x^{k+1} - x^k) = A^*(y^{k+1} - \bar{y}) + B^*(z^{k+1} - \bar{z}),
\]
\[
r^k = A^*(y^k - \bar{y}) + B^*(z^k - \bar{z}),
\]
\[
\| y^{k+1} \|^2 + \| r^{k+1} \|^2 \geq \frac{1}{2} \| A^*(y^{k+1} - y^k) + B^*(z^{k+1} - z^k) \|^2.
\]

Then we can get (3.12) by substituting (3.15) into (3.14). Since (3.12) holds for any \( \bar{u} \in \bar{u} \), we can get (3.13) from the fact that \( \bar{u} \) is a nonempty closed convex set. The proof is completed. \( \square \)
Now we are ready to establish the global convergence and the linear rate of convergence for the majorized iPADMM under a metric subregularity condition of $\mathcal{R}$ at some $(\bar{u}, 0) \in \text{gph} \mathcal{R}$.  

**Theorem 3.1.** Let $\tau \in (0, (1+\sqrt{3})/2)$, the two self-adjoint linear operators $\mathcal{S}$ and $\mathcal{T}$ satisfy (2.5),

\[
\frac{1}{2} \hat{\Sigma}_f + \mathcal{S} + \sigma \mathcal{A} \mathcal{A}^* > 0 \& \frac{1}{2} \hat{\Sigma}_g + \mathcal{T} + \sigma \mathcal{B} \mathcal{B}^* > 0,
\]

and $\{u^k := (y^k, z^k, x^k)\}$ be the infinite sequence generated by the majorized iPADMM. Then, one has the following results.

(a) The sequence $\{(y^k, z^k)\}$ converges to an optimal solution of problem (1.1) and $\{x^k\}$ converges to an optimal solution of the dual of problem (1.1).

(b) Suppose that the sequence $\{(y^k, z^k, x^k)\}$ converges to a KKT point $\bar{u} := (\bar{y}, \bar{z}, \bar{x})$ and the KKT mapping $\mathcal{R}$ is metrically subregular at $(\bar{u}, 0) \in \text{gph} \mathcal{R}$ with modulus $\eta > 0$. Then there exist a positive number $\mu \in (0, 1)$ and an integer $k_0 \geq 1$ such that for all $k \geq k_0$,

\[
\text{dist}^2_{\mathcal{M}}(u^{k+1}, \Omega) + \|z^{k+1} - z^k\|^2_{\hat{\Sigma}_g + \mathcal{T}} \leq \mu \left[ \text{dist}^2_{\mathcal{M}}(u^k, \Omega) + \|z^k - z^{k-1}\|^2_{\hat{\Sigma}_g + \mathcal{T}} \right].
\]

Moreover, there exists a positive number $\tilde{\mu} \in [\mu, 1)$ such that for all $k \geq 1$, 

\[
\text{dist}^2_{\mathcal{M}}(u^{k+1}, \Omega) + \|z^{k+1} - z^k\|^2_{\hat{\Sigma}_g + \mathcal{T}} \leq \tilde{\mu} \left[ \text{dist}^2_{\mathcal{M}}(u^k, \Omega) + \|z^k - z^{k-1}\|^2_{\hat{\Sigma}_g + \mathcal{T}} \right].
\]

**Proof.** We first prove the convergence on the sequences $\{(y^k, z^k)\}$ and $\{x^k\}$. From Lemma 3.5 and (3.16), we know that $\mathcal{H} > 0$, $\mathcal{M} > 0$ and $\hat{\Sigma}_g + \mathcal{T} \succeq 0$. Then it holds from Proposition 3.1 that $\{u^{k+1}\}$ is bounded and

\[
\lim_{k \to \infty} \|u^{k+1} - u^k\| = 0.
\]

Consequently, there is a subsequence $\{u^{k_i}\}$ which converges to a cluster point $u^\infty$. From Lemma 3.1, we know that

\[
\|\mathcal{R}(u^{k_i})\|^2 \leq \|u^{k_i} - u^{k_i-1}\|^2_{\mathcal{H}_0},
\]

where $\mathcal{H}_0 > 0$. Taking limits on both sides of the above inequality, we obtain $\|\mathcal{R}(u^\infty)\| = 0$. Thus, the subsequence $\{u^{k_i}\}$ converges to $u^\infty \in \Omega$. Therefore, the sequence $\|u^{k+1} - u^\infty\|_{\mathcal{M}}^2 + \|z^{k+1} - z^k\|^2_{\hat{\Sigma}_g + \mathcal{T}}$ converges to 0 as $k_i \to \infty$. Since the subsequence is non-increasing and $\|u^{k+1} - u^k\| \to 0$, we have

\[
\lim_{k \to \infty} \|u^k - u^\infty\| = 0.
\]

Therefore, the whole sequence $\{u^k\}$ converges to $u^\infty$. This completes the proof of the result (a).

Next, we prove (b). From (a), we know that the sequence $\{(y^k, z^k, x^k)\}$ generated by the majorized iPADMM converges to a KKT point $\bar{u} = (\bar{y}, \bar{z}, \bar{x})$. Then there exist $\rho > 0$ and an integer $k_0 \geq 1$ such that for all $k \geq k_0$,

\[
\|u^{k+1} - \bar{u}\| \leq \rho.
\]

Therefore, by using Lemma 3.1 and (2.3), we know that for all $k \geq k_0$,

\[
\text{dist}^2(u^{k+1}, \Omega) \leq \eta^2 \|\mathcal{R}(u^{k+1})\|^2 \leq \eta^2 \|u^{k+1} - u^k\|^2_{\mathcal{H}_0}.
\]

(3.19)
The definition of $H$ and the fact that $H > 0$ imply that $Hg > 0$. Then there exists a finite real number $\varrho_1 > 0$ such that $\hat{\Sigma} + T + \sigma B B^* \succ 0$. Thus there exists a finite real number $\varrho_1 > 0$ such that $\hat{\Sigma} + T + \sigma B B^* \succ 0$.

As mentioned in the introduction, the improvement is significant when $f$ or $g$ is not linear/quadratic.

Similarly, there exists a finite real number $\varrho_2 > 0$ such that $H_0 \leq \varrho_2 H$. It follows from (3.19) that for all $k \geq k_0$,

$$\|u^{k+1} - u^k\|_H^2 \geq \varrho_2^{-1} \|u^{k+1} - u^k\|_H^2$$

$$\geq \varrho_2^{-1} \eta^{-2} \text{dist}^2_M(u^{k+1}, \Omega) \geq \varrho_2^{-1} \eta^{-2} \lambda^{-1}_\text{max}(\mathcal{M}) \text{dist}^2_M(u^{k+1}, \Omega).$$

Let $\kappa := \frac{1}{1+\varrho_1 \beta}$ with $\beta := \varrho_2^{-1} \eta^{-2} \lambda^{-1}_\text{max}(\mathcal{M})$. From (3.13), we have that, for all $k \geq k_0$,

$$\left[ \text{dist}^2_M(u^{k+1}, \Omega) + \|z^{k+1} - z^k\|_{\hat{\Sigma} + T}^2 \right] - \left[ \text{dist}^2_M(u^k, \Omega) + \|z^k - z^{k-1}\|_{\hat{\Sigma} + T}^2 \right]$$

$$\leq - \left( (1 - \kappa) \|u^{k+1} - u^k\|_H^2 + \kappa \|u^{k+1} - u^k\|_H^2 \right)$$

(3.21)

$$\leq - \left( (1 - \kappa) \varrho_1^{-1} \|z^{k+1} - z^k\|_{\hat{\Sigma} + T}^2 + \kappa \varrho_2^{-1} \eta^{-2} \lambda^{-1}_\text{max}(\mathcal{M}) \text{dist}^2_M(u^{k+1}, \Omega) \right).$$

Then by reorganizing the above inequality, we know that for all $k \geq k_0$,

$$\left(1 + \kappa \varrho_2^{-1} \eta^{-2} \lambda^{-1}_\text{max}(\mathcal{M}) \right) \text{dist}^2_M(u^{k+1}, \Omega) + (1 - (1 - \kappa) \varrho_1^{-1}) \|z^{k+1} - z^k\|_{\hat{\Sigma} + T}^2$$

$$\leq \text{dist}^2_M(u^k, \Omega) + \|z^k - z^{k-1}\|_{\hat{\Sigma} + T}^2.$$ 

It is easy to check that

$$1 + \kappa \varrho_2^{-1} \eta^{-2} \lambda^{-1}_\text{max}(\mathcal{M}) = 1/\mu,$$

where

$$\mu := \frac{\varrho_1 \beta + 1}{1 + \beta + \varrho_1 \beta} < 1.$$ 

Then we know that inequality (3.17) holds.

By combining (3.17) with Lemma 3.1, (3.13) in Proposition 3.1, we can obtain directly that there exists a positive number $\mu \in [\mu, 1)$ such that (3.18) holds for all $k \geq 1$. This completes the proof.

**Remark 3.1.** We make the following comments.

(a) From the result (a) in Theorem 3.1, we can see that the majorized iPADMM is still globally convergent when the conditions $\frac{1}{2} \Sigma + S + \sigma A A^* > 0$ & $\frac{1}{2} \Sigma + T + \sigma B B^* > 0$ in [34, Theorem 10 (b)] are replaced by

$$\frac{1}{2} \hat{\Sigma} + S + \sigma A A^* > 0 \& \frac{1}{2} \hat{\Sigma} + T + \sigma B B^* > 0.$$

As mentioned in the introduction, the improvement is significant when $f$ or $g$ is not linear/quadratic.

10
(b) By observing the expression of parameter $\mu$ in the proof of result (b) in Theorem 3.1, in order to increase the convergence speed, under the premise of satisfying (2.5) and (3.16), the linear operators $\hat{\Sigma}_f$ and $\hat{\Sigma}_g$ should be chosen such that the majorized functions $\hat{f}$ and $\hat{g}$ are as close to $f$ and $g$ as possible, and the proximal terms $S$ and $T$ should be chosen as close to $-\frac{1}{2}\hat{\Sigma}_f$ and $-\frac{1}{2}\hat{\Sigma}_g$ as possible. For simplicity, we assume that $AA^* > 0$, then one should choose $S = -\frac{1}{2}\hat{\Sigma}_f$. This illustrates the claim mentioned in the introduction that the linear rate convergence results for sPADMM established in [27] are no longer applicable for our majorized iPADMM.

4 Application I: A majorized sGS-iPADMM

Consider the following general multi-block convex composite programming model,

$$\min_{y,z} \quad p(y_1) + f(y_1, \ldots, y_s) + q(z_1) + g(z_1, \ldots, z_t)$$

s.t. \quad $A^*y + B^*z = c,$

$$y \in \mathcal{Y}, \ z \in \mathcal{Z},$$

where $s$ and $t$ are given nonnegative integers, $\mathcal{Y} := \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_s$, $\mathcal{Z} := \mathcal{Z}_1 \times \ldots \times \mathcal{Z}_t$, $f(y_1, \ldots, y_s) := \sum_{i=1}^s f_i(y_i)$, and $g(z_1, \ldots, z_t) := \sum_{j=1}^t g_j(z_j)$.

For $i \in \{1, \ldots, s\}$ and $j \in \{1, \ldots, t\}$, we assume that $f_i : \mathcal{Y}_i \to \mathbb{R}$ and $g_j : \mathcal{Y}_j \to \mathbb{R}$ are convex functions with Lipschitz continuous gradients. Then, there exist positive semidefinite operators $\hat{\Sigma}_{f_i}$ and $\hat{\Sigma}_{g_j}$ such that for given $y'_i \in \mathcal{Y}_i$, $z'_j \in \mathcal{Z}_j$,

$$f_i(y_i) \leq \hat{f}_i(y_i, y'_i) := f_i(y'_i) + \langle \nabla f_i(y'_i), y_i - y'_i \rangle + \frac{1}{2}\|y_i - y'_i\|^2_{\hat{\Sigma}_{f_i}},$$

$$g_j(z_j) \leq \hat{g}_j(z_j, z'_j) := g_j(z'_j) + \langle \nabla g_j(z'_j), z_j - z'_j \rangle + \frac{1}{2}\|z_j - z'_j\|^2_{\hat{\Sigma}_{g_j}}.$$ 

Set

$$\hat{\Sigma}_f := \text{Diag}(\hat{\Sigma}_{f_1}, \ldots, \hat{\Sigma}_{f_s}) \quad \& \quad \hat{\Sigma}_g := \text{Diag}(\hat{\Sigma}_{g_1}, \ldots, \hat{\Sigma}_{g_t}).$$

Then it holds that

$$f(y) \leq \hat{f}(y, y') := f(y') + \langle \nabla f(y'), y - y' \rangle + \frac{1}{2}\|y - y'\|^2_{\hat{\Sigma}_f},$$

$$g(z) \leq \hat{g}(z, z') := g(z') + \langle \nabla g(z'), z - z' \rangle + \frac{1}{2}\|z - z'\|^2_{\hat{\Sigma}_g}.$$ 

For any given parameter $\sigma > 0$, the majorized augmented Lagrangian function $\hat{L}_\sigma(y, z, x; y', z')$ is defined as (1.7) and the sGS based multi-block majorized ADMM with indefinite proximal terms (majorized sGS-iPADMM) is presented in Algorithm 2.
Algorithm 2 A majorized sGS-iPADMM

Let $\sigma > 0$ and $\tau \in (0, (1 + \sqrt{5})/2)$ be given parameters. For $i \in \{1, \ldots, s\}$ and $j \in \{1, \ldots, t\}$, let $S_i$, $T_j$ be given self-adjoint, possibly indefinite, linear operators. Input $(y^0, z^0, x^0) \in \text{dom} \, p \times \text{dom} \, q \times X$. For $k = 0, 1, 2, \ldots$, perform the following steps.

**Step 1a.** (Backward GS sweep) Compute for $i = s, \ldots, 2$,

$$y_i^k = \arg \min_{y_i} \hat{L}_\sigma(y_{\leq i-1}^k, y_i, y_{\geq i+1}^k, z^k, x^k; y^k, z^k) + \frac{1}{2}\|y_i - y_1^k\|_S^2,$$

and $y_1^{k+1} = \arg \min_{y_1} \hat{L}_\sigma(y_1, y_{\geq 2}^k, z^k, x^k; y^k, z^k) + \frac{1}{2}\|y_1 - y_1^k\|_S^2$.

**Step 1b.** (Forward GS sweep) Compute for $i = 2, \ldots, s$,

$$y_i^{k+1} = \arg \min_{y_i} \hat{L}_\sigma(y_{\leq i-1}^{k+1}, y_i, y_{\geq i+1}^k, z^k, x^k; y^k, z^k) + \frac{1}{2}\|y_i - y_i^k\|_S^2.$$  

**Step 1c.** (Backward GS sweep) Compute for $j = t, \ldots, 2$,

$$z_j^k = \arg \min_{z_j} \hat{L}_\sigma(y^{k+1}, z_{\leq j-1}^k, z_j, z_{\geq j+1}^k, x^k; y^k, z^k) + \frac{1}{2}\|z_j - z_j^k\|_{T_j}^2,$$

and $z_1^{k+1} = \arg \min_{z_1} \hat{L}_\sigma(y^{k+1}, z_1, z_{\geq 2}^k, x^k; y^k, z^k) + \frac{1}{2}\|z_1 - z_1^k\|_{T_1}^2$.

**Step 1d.** (Forward GS sweep) Compute for $j = 2, \ldots, t$,

$$z_j^{k+1} = \arg \min_{z_j} \hat{L}_\sigma(y^{k+1}, z_{\leq j-1}^{k+1}, z_j, z_{\geq j+1}^k, x^k; y^k, z^k) + \frac{1}{2}\|z_j - z_j^k\|_{T_j}^2.$$  

**Step 2.** Compute $x^{k+1} = x^k + \tau \sigma(A^*y^{k+1} + B^*z^{k+1} - c)$.

Denote the following two linear operators,

$$\tilde{M} := \frac{1}{2} \hat{\Sigma}_f + \sigma AA^* + \text{Diag}(S_1, \ldots, S_s), \quad \tilde{N} := \frac{1}{2} \hat{\Sigma}_g + \sigma BB^* + \text{Diag}(T_1, \ldots, T_t),$$

where for $i \in \{1, \ldots, s\}$ and $j \in \{1, \ldots, t\}$, $S_i \succeq -\frac{1}{2} \hat{\Sigma}_f_i$ and $T_j \succeq -\frac{1}{2} \hat{\Sigma}_g_j$ are self-adjoint linear operators such that the $i$-th diagonal block operator of $\tilde{M}$ and $j$-th diagonal block operator of $\tilde{N}$ are positive definite, i.e.,

$$\tilde{M}_{ii} = \frac{1}{2} \hat{\Sigma}_{f_i} + S_i + \sigma A_i A_i^* > 0 \quad \& \quad \tilde{N}_{jj} = \frac{1}{2} \hat{\Sigma}_{g_j} + T_j + \sigma B_j B_j^* > 0.$$  

Moreover, define

$$S := \text{Diag}(S_1, \ldots, S_s) + \text{sGS}(\tilde{M}) \quad \& \quad T := \text{Diag}(T_1, \ldots, T_t) + \text{sGS}(\tilde{N}),$$

where $\text{sGS}(\tilde{M}) := \tilde{M}_u \tilde{M}_d^{-1} \tilde{M}_u^*$ and $\text{sGS}(\tilde{N}) := \tilde{N}_u \tilde{N}_d^{-1} \tilde{N}_u^*$. Notations $\tilde{M}_u$ ($\tilde{M}_d$) and $\tilde{N}_u$ ($\tilde{N}_d$) stand for the strictly upper triangular (diagonal) block operators of $\tilde{M}$ and $\tilde{N}$, respectively.
By using the same process in Li et al. [35] (see also Chen et al. [3]), the majorized sGS-iPADMM can be equivalently converted into the following 2-block majorized iPADMM,

\[
\begin{cases}
  y^{k+1} = \arg\min_{y \in Y} p(y) + \langle y, \nabla f(y^k) + A x^k \rangle + \frac{\sigma}{2} \|A^* y + B^* z^k - c\|^2 + \frac{1}{2} \|y - y^k\|_{\tilde{\Sigma}_f + S}^2, \\
  z^{k+1} = \arg\min_{z \in Z} q(z) + \langle z, \nabla g(z^k) + B x^k \rangle + \frac{\sigma}{2} \|A^* y^{k+1} + B^* z - c\|^2 + \frac{1}{2} \|z - z^k\|_{\tilde{\Sigma}_g + T}^2, \\
  x^{k+1} := x^k + \tau \sigma (A^* y^{k+1} + B^* z^{k+1} - c),
\end{cases}
\]

where the operators \(\tilde{\Sigma}_f\) and \(\tilde{\Sigma}_g\) are defined by (4.2), the proximal terms \(S\) and \(T\) are defined by (4.3). It follows from the choice of \(S_i\) and \(T_j\) that Assumption 2.2 holds and

\[
\frac{1}{2} \tilde{\Sigma}_f + S + \sigma A A^* \succ 0 \& \frac{1}{2} \tilde{\Sigma}_g + T + \sigma B B^* \succ 0.
\]

Therefore, directly from Theorem 3.1, we can get the following convergence results for our majorized sGS-iPADMM.

**Proposition 4.1.** Let \(u^k := (y^k, z^k, x^k)\) be the infinite sequence generated by the majorized sGS-iPADMM and proximal terms \(S\) and \(T\) be defined by (4.3). Then, we have the following results.

(a) The sequence \(\{(y^k, z^k)\}\) converges to an optimal solution of problem (4.1) and \(\{x^k\}\) converges to an optimal solution of the dual of problem (4.1).

(b) Suppose that the sequence \(\{(y^k, z^k, x^k)\}\) converges to a KKT point \(\bar{u} := (\bar{y}, \bar{z}, \bar{x})\) and the KKT mapping \(\mathcal{R}\) is metrically subregular at \((\bar{u}, 0) \in \text{gph}\mathcal{R}\). Then the sequence \(\{u^k\}\) is linearly convergent to \(\bar{u}\).

Next we use an example to illustrate the fact that the proximal terms generated by (4.3) could be indefinite.

**Example 4.1.** Consider the following sparse group Lasso (see e.g. [20]),

\[
\min_x \frac{1}{2} \|Dx - d\|^2 + \lambda_1 \|x\|_1 + \lambda_2 \sum_{l=1}^m w_l \|x_{G_l}\|,
\]

where \(\lambda_1, \lambda_2 > 0\), \(l = 1, 2, \ldots, m, w_l > 0\), \(G_l \subseteq \{1, 2, \ldots, n\}\) contains the indices corresponding to the \(l\)-th group of features. For each \(l \in \{1, 2, \ldots, g\}\), define the linear operator \(\mathcal{P}_l\) by \(\mathcal{P}_l x = x_{G_l}\) and \(\mathcal{P} = [\mathcal{P}_1; \mathcal{P}_2; \ldots; \mathcal{P}_g]\). The dual formulation (equivalent minimization form) of (4.4) takes the following form:

\[
\min_{y, z \eta} \frac{1}{2} \|\theta\|^2 + d^T \theta + \mathbb{B}_\infty(\eta) + \mathbb{B}_2(z) \\
\text{s.t.} \quad D^T \theta + \eta + \mathcal{P}^* z = 0,
\]

where \(\mathbb{B}_\infty := \{\eta | \|\eta\|_\infty \leq \lambda_1\}, \quad \mathbb{B}_2 := \{z | \|z_l\|_2 \leq \lambda_2 w_l, \ z_l \in \mathbb{R}|G_l|, \ l = 1, \ldots, m\} \) and \(\mathcal{P}^*\) is the adjoint operator of \(\mathcal{P}\). Let \(y := [\theta; \eta]\) and define

\[
f(y) := \frac{1}{2} \|\theta\|^2 + d^T \theta, \quad p(y) := \mathbb{B}_\infty(\eta), \quad q(z) := \mathbb{B}_2(z), \quad A^* := [D^T; I], \quad B^* := \mathcal{P}^*.
\]
Consequently, the dual problem (4.5) can be reformulated into the framework of (1.1) and the
majorized sGS-iPADMM iterative scheme can be described as below.

\[
\begin{align*}
\theta^{k+1} &= \arg \min_\theta \left( \theta^k + d + Dx^k \right) + \frac{\gamma}{2} \| D^T \theta + \eta^k + \mathcal{P}^* z^k \|^2 + \frac{1}{2} \| \theta - \theta^k \|_I^2 + S_1, \\
\eta^{k+1} &= \arg \min_\eta \left[ \mathcal{B}_\infty (\eta) + \frac{\gamma}{2} \| D^T \theta^{k+1} + \mathcal{P}^* z^k + z^k / \sigma \|^2, \\
\phi^{k+1} &= \arg \min \left\{ \mathcal{B}_1 (z) + \frac{\gamma}{2} \| \mathcal{P}^* z + \eta^{k+1} + D^T \phi^{k+1} + \mathcal{P} x^k / \sigma \|^2, \\
x^{k+1} &= x^k + \tau \sigma (D^T \phi^{k+1} + \mathcal{P}^* z^{k+1} + \eta^{k+1}).
\right. 
\end{align*}
\]

Directly from (4.3), we can take

\[
S = \begin{pmatrix} S_1 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ \sigma D^T & 0 \end{pmatrix} \left( \sigma DD^T + \frac{1}{2} I + S_1 \right)^{-1} \begin{pmatrix} 0 & \sigma D \\ 0 & 0 \end{pmatrix}, \quad T = 0.
\]

For simplicity, we assume that $DD^T \succ 0$. Consequently, one can take $S_1 = -\frac{1}{2} I$ and derive the
following indefinite proximal term,

\[
S = \begin{pmatrix} -\frac{1}{2} I & 0 \\ 0 & \sigma I \end{pmatrix}.
\]

5 Application II: The regularized logistic regression

In this section, we apply the majorized iPADMM to general regularized logistic regression in the
following form,

\[
\min_{y, y_0} f(y, y_0) + \varphi(y), \quad (5.1)
\]

where $f : \mathbb{R}^{n+1} \to \mathbb{R}$ is the logistic loss function and $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a general convex Lasso
regularizer. Specifically, the logistic loss function $f$ takes the following form,

\[
f(y, y_0) = \frac{1}{N} \sum_{i=1}^{N} \log (1 + \exp(-b_i (B_i^T y + y_0))),
\]

where $B_i \in \mathbb{R}^n$ are the predictor variables and $b_i \in \{1, -1\}$ are the responses, $i = 1, \ldots, N$. For
notation convenience, set $\tilde{y} := [y; y_0] \in \mathbb{R}^{n+1}$, $A_i := [-b_i B_i; -b_i] \in \mathbb{R}^{n+1}$ and denote the gradient of $f$ at $\tilde{y} \in \text{dom} f$ by $\nabla f(\tilde{y})$.

Since the gradient $\nabla f$ is Lipschitz continuous on $\text{dom} f$, we know that there exists a positive
semidefinite matrix $\tilde{\Sigma}_f$ such that for any given $\tilde{y} \in \mathbb{R}^{n+1},$

\[
f(\tilde{y}) \leq \hat{f}(\tilde{y} ; \tilde{y}') := f(\tilde{y}') + \langle \nabla f(\tilde{y}'), \tilde{y} - \tilde{y}' \rangle + \frac{1}{2} \| \tilde{y} - \tilde{y}' \|^2_{\tilde{\Sigma}_f}. \quad (5.2)
\]

Elementary calculations show that

\[
\nabla f(\tilde{y}) = \frac{1}{N} \sum_{i=1}^{N} A_i \exp(A_i^T \tilde{y}) \quad \text{and} \quad \nabla^2 f(\tilde{y}) = \frac{1}{N} \sum_{i=1}^{N} A_i A_i^T \frac{\exp(A_i^T \tilde{y})}{(1 + \exp(A_i^T \tilde{y}))^2} \leq \frac{1}{4N} \sum_{i=1}^{N} A_i A_i^T.
\]

14
Therefore, the proximal term can be chosen as
\[ \hat{\Sigma}_f := \frac{1}{4N} AA^T, \quad A := [A_1, \ldots, A_N] \in \mathbb{R}^{(n+1) \times N}. \] (5.3)

In this section, we consider the logistic regression with three types of \( \varphi(\cdot) \), namely, Lasso, fused Lasso and constrained Lasso. Next, we reformulate these three types of regularized logistic regression into the framework of (1.1) and tailor the majorized iPADMM for each of them. Since our main purpose is to test the numerical performance of our majorized iPADMM, we omit the history and development of these regularized logistic regression models. To know more about these models, we refer to [4, 30, 31, 32, 46, 48, 39].

5.1 Fused Lasso logistic regression

The fused Lasso method is introduced by Tibshirani et al. [48] to study the situation that the features have a natural order. In this case, for any given \( \lambda_1 \geq 0, \lambda_2 \geq 0 \), the function \( \varphi(\cdot) \) takes the following form,
\[ \varphi(y) = \lambda_1 \|y\|_1 + \lambda_2 \sum_{i=2}^n |y_i - y_{i-1}|. \] (5.4)
The definition (5.4) contains the Lasso regularizer as a special case if we take \( \lambda_2 = 0 \).

By introducing an auxiliary variable \( z \in \mathbb{R}^n \), we can reformulate the fused Lasso problem into the framework of (1.1), i.e.,
\[ \min_{y, y_0, z} f(y, y_0) + \varphi(z) \quad \text{s.t.} \quad y - z = 0. \] (5.5)
The KKT system can be written as follows,
\[ \nabla f(y, y_0) + [x; 0] = 0, \quad z - \text{Pr}_\varphi(x + z) = 0, \quad y - z = 0. \]

The function \( \tilde{\mathcal{L}}_\sigma(\cdot) \) in (1.7) can be specifically written as
\[ \tilde{\mathcal{L}}_\sigma(y, y_0, z, y'; \tilde{y}') = \hat{f}(\tilde{y}, \tilde{y}') + \varphi(z) + (y - z, x) + \frac{\sigma}{2} \|y - z\|^2. \]

Consequently, the majorized iPADMM scheme for solving (5.5) can be described as follows,
\[
\begin{align*}
(y_{k+1}^{\perp}, y_0^{k+1}) &= \arg \min_{y, y_0} \tilde{\mathcal{L}}_\sigma(y, y_0, z^k, x^k; y^k) + \frac{1}{2} \|\tilde{y} - y^k\|_S^2, \\
z^{k+1} &= \arg \min_z \varphi(z) + \frac{\sigma}{2} \|z - (y^{k+1} + x^k/\sigma)\|^2, \\
x^{k+1} &= x^k + \tau \sigma (y^{k+1} - z^{k+1}),
\end{align*}
\] (5.6)

where \( \tau \in (0, (1 + \sqrt{5})/2) \) and \( S = -\frac{1}{2} \hat{\Sigma}_f + \text{Diag}(0, \sigma r) \) with \( r \) being a given positive number. It is obviously that
\[ \frac{1}{2} \hat{\Sigma}_f + S + \sigma \left( \begin{array}{cc} I & 0 \\ 0 & 0 \end{array} \right) \succ 0. \]
Based on the optimality conditions, we measure the accuracy of an approximate KKT point $(y, y_0, z, x)$ via

$$
\eta_{FL} = \max \{ \eta_P, \eta_D, \eta_C \},
$$

(5.7)

where

$$
\eta_P = \left\{ \frac{\|y - z\|}{1 + \|y\| + \|z\|} \right\}, \quad \eta_D = \frac{\|\nabla f(y, y_0) + [x; 0]\|}{1 + \|\nabla f(y, y_0)\| + \|x\|}, \quad \eta_C := \frac{\|z - \text{Pr}_x(x + z)\|}{1 + \|x\| + \|z\|}.
$$

It is worth mentioning that, the solution $z^{k+1}$ can be obtained by the following result.

**Lemma 5.1.** For any $\lambda_1, \lambda_2 \geq 0$, the optimal solution $z^*$ of

$$
\min_z \lambda_1 \|z\|_1 + \lambda_2 \sum_{i=2}^n |z_i - z_{i-1}| + \frac{1}{2} \|z - v\|^2,
$$

(5.8)

can be described as

$$
z^* = \text{sgn}(z_0) \cdot \max(\|z_0\| - \lambda_1, 0),
$$

where $z_0 := \arg \min_z \lambda_2 \sum_{i=2}^n |z_i - z_{i-1}| + \frac{1}{2} \|z - v\|^2$.

The above result was first shown in [46] by using the subgradient technique and an alternative proof can be found in [39]. Though there is no closed-form expression of $z_0$ when $\lambda_2 > 0$, the algorithms presented in [6] can used to get $z_0$ efficiently. The corresponding code can download from website “https://www.gipsa-lab.grenoble-inp.fr/laurent.condat/software.html”.

By using [36, Theorem 1 & Remark 1] and [11, Theorem 3.1], we know that the KKT mapping $R$ corresponding to (5.5) is metrically subregular at any KKT point for original when $\lambda_2 = 0$. Therefore, directly from Theorem 3.1, we can get the following results.

**Proposition 5.1.** Let $\lambda_2 = 0$ and $\{u^k := (y^k, y_0^k, z^k, x^k)\}$ be the infinite sequence generated by the majorized iPADMM scheme (5.6). Then the sequence $\{u^k\}$ converges linearly to a KKT point of (5.5).

### 5.2 Constrained logistic regression

Inspired by the work of James et al. [31], we consider the function $\varphi$ in the following form,

$$
\varphi(y) = \lambda \|y\|_1 + \mathbb{I}_D(y),
$$

(5.8)

where $D := \{ y \mid Dy \geq d \}$, $D \in \mathbb{R}^{m \times n}$, $d \in \mathbb{R}^m$, function $\mathbb{I}_D(\cdot)$ is an indicator function of convex set $D$. We can rewrite (5.1) as

$$
\min_{y, y_0, u, v} f(y, y_0) + \lambda \|v\|_1 + \mathbb{I}_{\mathbb{R}^m_+}(u)
$$

s.t.

$$
Dy - u = d,
$$

$$
y - v = 0.
$$

(5.9)

The KKT conditions are given by

$$
\nabla f(y, y_0) + [D^T \xi + \zeta; 0] = 0, \quad v - \text{Pr}_{\|\cdot\|_1}(\zeta + v) = 0, \quad u - \Pi_{\mathbb{R}^m_+}(\xi + u) = 0, \quad Du - u = d, \quad y - v = 0,
$$

16
and the majorized augmented Lagrangian function \( \hat{L}_\sigma(\cdot) \) in (1.7) can be specifically written as
\[
\hat{L}_\sigma(y, y_0, u, v, \xi, \zeta; \tilde{y}') = \tilde{f}(\tilde{y}, \tilde{y}') + \lambda \| v \|_1 + \langle \Pi_{\mathcal{F}}(u) - u - d, \xi \rangle + \langle y - v, \zeta \rangle + \frac{\sigma}{2} \| D y - u - d \|^2 + \frac{\sigma}{2} \| y - v \|^2.
\]

Therefore, we can solve (5.9) via the following iterative scheme:
\[
\begin{align*}
(y^{k+1}, y_0^{k+1}) &= \arg \min_{y,y_0} \hat{L}_\sigma(y, y_0, u^k, v^k, \xi^k, \zeta^k; \tilde{y}') \quad \text{subject to} \quad \eta = \max \{ D y^{k+1} - d + \xi^k/\sigma, 0 \}, \\
u^{k+1} &= \arg \min_v \lambda \| v \|_1 + \frac{\sigma}{2} \| v - (y^{k+1} + \xi^k/\sigma) \|^2, \\
\xi^{k+1} &= \xi^k + \tau \sigma (D y^{k+1} - u^{k+1} - d), \\
\zeta^{k+1} &= \zeta^k + \tau \sigma (y^{k+1} - v^{k+1}),
\end{align*}
\]

where \( \tau \in (0, (1 + \sqrt{5})/2) \) and \( \mathcal{S} = -\frac{1}{2} \hat{\Sigma} f + \text{Diag}(0, \sigma r) \) with \( r \) being a given positive number. It is obviously that
\[
\frac{1}{2} \hat{\Sigma} f + \mathcal{S} + \sigma \begin{pmatrix} D^T D + I & 0 \\ 0 & 0 \end{pmatrix} \succ 0.
\]

Based on the optimality conditions, we measure the accuracy of an approximate KKT point \((y, y_0, u, v, \xi, \zeta)\) via
\[
\eta_{CL} = \max \{ \eta_P, \eta_D, \eta_C \},
\]
where
\[
\eta_P = \max \left\{ \frac{\| D y - u - d \|}{1 + \| D y \| + \| u \| + \| d \|}, \frac{\| y - v \|}{1 + \| y \| + \| v \|} \right\}, \quad \eta_D = \frac{\| \nabla f(y, y_0) \| + \| D^T \xi + \zeta \|}{1 + \| \nabla f(y, y_0) \| + \| D^T \xi \| + \| \zeta \|},
\]
\[
\eta_C := \max \left\{ \frac{\| v - \Pi_{\mathcal{F}}(\xi + v) \|}{1 + \| \xi \| + \| v \|}, \frac{\| u - \Pi_{\mathcal{F}}(\xi + u) \|}{1 + \| \xi \| + \| u \|} \right\}.
\]

### 5.3 Numerical experiments

In this subsection, we evaluate the performance of majorized iPADMM for solving Lasso logistic regression, fused Lasso logistic regression (5.5), constrained Lasso logistic regression (5.9), respectively. For all numerical experiments of ADMM-type method, we choose \( \tau = 1.618 \). All computational results are obtained by running Matlab R2015b on Mac OS X 10.10.5 (2.9 GHz Intel Core i5 16GB 1867 MHz DDR3).

Consider the following two self-adjoint linear operators:
\[
\mathcal{S}_0 = \sigma \begin{pmatrix} 0 & 0 \\ 0 & r \end{pmatrix} \quad \text{and} \quad \mathcal{S} = -\frac{1}{2} \hat{\Sigma} f + \sigma \begin{pmatrix} 0 & 0 \\ 0 & r \end{pmatrix}.
\]

Note that the proximal term \( \mathcal{S} \) may not be a positive semidefinite operator. In the subsequent discussions, we call the majorized iPADMM scheme with \( \mathcal{S}_0 \) as “majorized sPADMM”. In all tests, we set \( r = 10^{-6} \) and choose \( \hat{\Sigma} f \) as (5.3) unless otherwise specified.
5.3.1 Lasso logistic regression: LIBSVM dataset

In this part, we apply the majorized iPADMM to Lasso logistic regression and test its performance with data sets: \texttt{a8a}, \texttt{a9a}, \texttt{colon-cancer}, \texttt{duke breast-cancer} (duke-BC). These data sets are obtained from the LIBSVM datasets [2].

Notice that the subproblem corresponding to \((y, y_0)\) can be reformulated as a linear system of equations. In order to exactly and efficiently solve this subproblem, we use different techniques to get the inverse matrix of \(H := \hat{\Sigma}_f + \sigma \text{Diag}(I, 0) + \mathcal{S}\). For \texttt{a8a}, \texttt{a9a}, i.e. \(n \ll N\), \(n\) is moderate, the inverse matrix \(H \in \mathbb{R}^{(n+1) \times (n+1)}\) can be efficiently obtained by MATLAB toolbox. For \texttt{colon-cancer}, \texttt{duke-BC}, i.e. \(n \gg N\), \(N\) is moderate, the Sherman-Morrison-Woodbury formula [26] can be used to get the inverse matrix of \(H\).

Figure 1 shows that the sequence generated by majorized iPADMM converges to a KKT point approximately as a linear rate. This is consistent with our Theorem 3.1. Moreover, we can also observe that the indefinite proximal term can improve the numerical performance. The majorized iPADMM brings about 35\%–60\% reduction in the number of iterations needed as compared with the majorized sPADMM, when one chooses \(\|u^k - \bar{u}\|_M \leq 10^{-6}\).

\[\text{Figure 1: Comparison between the performance of our majorized iPADMM (M-iPADMM) and the majorized sPADMM (M-sPADMM) on datasets a8a, a9a, colon-cancer, and duke-BC. error := \|u^k - \bar{u}\|_M.}\]

5.3.2 Fused lasso logistic regression: LIBSVM dataset

In this part, we apply the majorized iPADMM to fused Lasso logistic regression and test its performance with data sets: \texttt{a8a}, \texttt{a9a}, \texttt{colon-cancer}, \texttt{duke breast-cancer}, \texttt{rcv1.binary}, \texttt{news20.binary}. These data sets are from the LIBSVM datasets [2].
In order to evaluate the numerical performance, we also report majorized sPADMM and a commonly used first order algorithm for solving fused lasso problem: the accelerated proximal gradient algorithm implemented in [39]. This algorithm is called “EFLA” in [39]. In the comparison, we terminate “EFLA” if the objective value \( \text{obj}_E \) obtained by “EFLA” satisfies

\[
\frac{\text{obj}_E - \text{obj}_M}{|\text{obj}_M|} \leq 10\varepsilon,
\]

where \( \text{obj}_M \) is the objective value obtained by majorized iPADMM. To keep things fair, we use the algorithm in [6] to implement the subproblem in EFLA. This makes EFLA faster than the original code in software “SLEP 4.0”.

In this test, we choose the regularization parameters as follows

\[
\lambda_1 = \lambda_2 = \frac{\gamma}{N} \|B^T b\|_\infty,
\]

where \( N \) is the size of sample and \( 0 < \gamma < 1 \). The majorized iPADMM and majorized sPADMM will be terminated when \( \eta_{FL} < 10^{-6} \) or the maximum iteration number 50,000 is reached.

For \texttt{rcv1.binary}, \texttt{news20.binary}, i.e., both \( n \) and \( N \) are large, by taking \( \hat{\Sigma} = \sum_{i=1}^{K} \mu_i P_i P_i^T \), where \( \sqrt{\mu_1}, \ldots, \sqrt{\mu_K} \) are the \( K \) largest singular values of \( A \) and \( P_1, \ldots, P_K \) are the corresponding left-eigenvectors, the Sherman-Morrison-Woodbury formula can also be used to solve the linear system of equations corresponding to \((y,y_0)\) subproblem.

Now we are ready to report the comparison results. From Table 1, we can observe that the numerical performance of majorized iPADMM outperforms the other two methods for all cases. Besides, we can also see that the majorized iPADMM brings about 40%–50% reduction in the number of iterations needed as compared with the majorized sPADMM except cases that \texttt{a8a} and \texttt{a9a} with \( \gamma = 10^{-2} \).

| probName | \( L_C \) | \( \gamma \) | nnz | MiPA | MsPA | E | MiPA | MsPA | E |
|-----------|---------|---------|-----|-----|-----|---|-----|-----|---|
| colon-cancer | 1.9e+04 | 1.0e-02 | 52 | 319 | 542 | 1294 | 0.34 | 0.46 | 0.47 |
| duke-BC 62, 2000 | 1.1e+05 | 1.0e-02 | 61 | 750 | 1271 | 3177 | 1.61 | 2.72 | 3.55 |
| rcv1.binary 20242, 47236 | 4.5e+02 | 1.0e-02 | 123 | 650 | 1099 | 1703 | 6.50 | 11.15 | 14.34 |
| news20.binary 19996, 1355191 | 1.2e+03 | 1.0e-02 | 264 | 795 | 1287 | 2618 | 178.52 | 282.25 | 278.54 |
| colon-cancer | 1.9e+04 | 1.0e-03 | 69 | 519 | 691 | 1294 | 1.59 | 3.05 | 2.22 |
| duke-BC 44, 7129 | 1.1e+05 | 1.0e-03 | 68 | 2287 | 6683 | 16105 | 4.84 | 14.10 | 18.17 |
| rcv1.binary 20242, 47236 | 4.5e+02 | 1.0e-03 | 744 | 2533 | 4835 | 4640 | 25.18 | 49.71 | 38.96 |
| news20.binary 19996, 1355191 | 1.2e+03 | 1.0e-03 | 2973 | 4362 | 3891 | 6576 | 536.45 | 970.01 | 935.00 |

Table 1: Comparison between the performance of majorized iPADMM (MiPA), majorized sPADMM (MsPA), and EFLA (E). “nnz” denotes the number of nonzeros in the solution \( z \) generated by majorized iPADMM. “IterNum” denotes the number of iterations. \( L_C := \lambda_{\text{max}}(B B^T) \).
5.3.3 Constrained logistic regression: synthetic data

This part tests the performance of majorized iPADMM for the constrained logistic regression by using synthetic data. The data $B \in \mathbb{R}^{n \times N}$, $D \in \mathbb{R}^{m \times n}$ and $d \in \mathbb{R}^m$ are generated by the standard normal distribution. This example can also be used to illustrate the claim mentioned in the introduction that our majorized iPADMM with the positive semidefinite operator $\hat{\Sigma}_f$ outperforms that with the Lipschitz constant $\lambda_{\max}((\hat{\Sigma}_f)^T)$ (denoted as L-majorized ADMM).

We compare our majorized iPADMM with majorized sPADMM, L-majorized ADMM with an indefinite proximal term $(\text{Diag}(-\frac{1}{2}\lambda_{\max}(\hat{\Sigma}_f)I, \sigma r))$, and L-majorized sPADMM. All the algorithms will be terminated when $\eta_{CL} < 10^{-5}$ or the maximum iteration number 50,000 is reached. In this test, we choose the regularization parameter $\lambda = \gamma \frac{\|B^Tb\|_{\infty}}{N}$, where $0 < \gamma < 1$.

Table 2 reports the number of iterations, runtime of four different methods. From the table, we can see that our majorized iPADMM outperforms all the other three methods. In each case, the majorized iPADMM can sometimes bring about 40% reduction in the number of iterations needed for convergence as compared with the majorized sPADMM. Note that though the size of each scenario is small, some cases still can not be solved within maximum iteration by using L-majorized iPADMM and L-majorized sPADMM.

Figure 2 shows the relative KKT residual norm $\eta_{CL}$ of the iterates generated by majorized iPADMM and L-majorized iPADMM on two different synthetic data sets. This figure can be used to illustrate the advantage of using the positive semidefinite operator $\hat{\Sigma}_f$ over the Lipschitz constant $\lambda_{\max}(\hat{\Sigma}_f)$.

| N,n,m       | $L_C$ | $\lambda$  | IterNum          | Time (sec)          |
|-------------|-------|------------|------------------|---------------------|
|             |       |           | MiPA | MsPA | LiA | LA | MiPA | MsPA | LiA | LA |
| 2000,500,50 | 2.25  | $10^{-2}$ | 306.7 | 307.3 | 307.5 | 312.7 | 1.5 | 1.5 | 1.5 | 1.5 |
|             |       | $10^{-3}$ | 154.3 | 161.3 | 167.3 | 186.4 | 0.8 | 0.9 | 0.9 | 1.0 |
|             |       | $10^{-4}$ | 115.2 | 119.8 | 128.1 | 149.9 | 0.6 | 0.6 | 0.7 | 0.8 |
| 30,50,20    | 4.59  | $10^{-2}$ | 317.0 | 374.0 | 1917.3 | 3825.7 | 0.1 | 0.1 | 0.3 | 0.5 |
|             |       | $10^{-3}$ | 2014.9 | 3887.8 | 18494.4 | 35075.7 | 0.2 | 0.4 | 1.9 | 3.5 |
|             |       | $10^{-4}$ | 11636.8 | 21473.3 | 49961.2 | 50000.0 | 1.0 | 1.6 | 4.9 | 4.7 |
| 50,200,30   | 9.08  | $10^{-2}$ | 573.3 | 581.8 | 3470.5 | 6734.6 | 0.4 | 0.4 | 1.8 | 3.3 |
|             |       | $10^{-3}$ | 1550.8 | 2303.0 | 27501.7 | 47114.9 | 0.8 | 1.2 | 12.8 | 21.8 |
|             |       | $10^{-4}$ | 6403.4 | 12001.4 | 50000.0 | 50000.0 | 3.2 | 5.7 | 23.2 | 23.2 |

Table 2: Comparison between the performance of majorized iPADMM (MiPA), majorized sPADMM (MsPA), L-majorized iPADMM (LiA), and L-majorized sPADMM (LA); “IterNum” denotes the number of iterations. $L_C := \lambda_{\max}(A^T)/N$. All results are averaged over 10 instances.
Figure 2: Comparison between the performance of majorized iPADMM (M-iPADMM) and L-majorized iPADMM (L-iPADMM) on synthetic data within given iteration numbers. All results are averaged over 10 instances.

6 Conclusion remarks

In this paper, we have established the linear rate convergence of the majorized ADMM with indefinite proximal terms for solving the 2-block linearly constrained convex composite optimization problem under a metric subregularity assumption. Numerical results on three types of regularized logistic regression have been given to evaluate the effectiveness of the 2-block majorized ADMM with indefinite proximal terms. From these results, we can see that, for many cases, the majorized ADMM with indefinite proximal terms can bring about 30%–50% reduction in the number of iterations needed for convergence as compared with the majorized ADMM with semi-proximal terms.

Strongly motivated by the numerical performance of the symmetric Gauss-Seidel based ADMM for solving multi-block convex composite quadratic programming, we also proved the linear rate of convergence of a symmetric Gauss-Seidel based majorized ADMM with indefinite proximal terms by building its equivalence to the 2-block majorized ADMM with specially constructed proximal terms (possibly indefinite). This will extremely facilitate the future exploration of the multi-block general linear/nonlinear models. We leave this topic as our future work.

References

[1] Banert, S., Bot, R.I., Csetnek, E.R.: Fixing and extending some recent results on the ADMM algorithm. arXiv preprint arXiv:1612.05057 (2016).

[2] Chang, C.-C., Lin, C.-J.: A library for support vector machine. Libsvm (2001).

[3] Chen, L., Sun, D.F., Toh K.-C.: An efficient inexact symmetric Gauss-Seidel based majorized ADMM for high-dimensional convex composite conic programming. Mathematical Programming. 161(1-2), 237–270 (2017).

[4] Chuang, H.Y., Lee, E., Liu, Y.T., et al.: Network-based classification of breast cancer metastasis. Molecular Systems Biology. 3(1), 140 (2007).
[5] Clarke, F.H.: *Optimization and nonsmooth analysis*. Wiley, New York (1983).

[6] Condat, L.: A direct algorithm for 1-D total variation denoising. *IEEE Signal Processing Letters*. 20(11), 1054–1057 (2013).

[7] Condat, L.: A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*. 158 (2), 460–479 (2013).

[8] Drusvyatskiy, D., Lewis, A.S.: Error bounds, quadratic growth, and linear convergence of proximal methods. *arXiv preprint arXiv:1602.06661* (2016).

[9] Deng, W., Yin, W.: On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*. 66(3), 889–916 (2016).

[10] Ding, C., Sun D.F., Zhang L.W.: Characterization of the robust isolated calmness for a class of conic programming problems. *SIAM Journal on Optimization*. 27(1), 67–90 (2017).

[11] Dong, Y.: An extension of Luque’s growth condition. *Applied Mathematics Letters*. 22(9), 1390-1393 (2009).

[12] Dontchev, A.L., Rockafellar, R.T.: Implicit functions and solution mappings. Springer, New York (2009).

[13] Dontchev, A.L., Rockafellar, R.T.: Regularity and conditioning of solution mappings in variational analysis. *Set-Valued Analysis*. 12(1), 79–109 (2004).

[14] Eckstein, J.: Some saddle-function splitting methods for convex programming. *Optimization Methods and Software*. 4 (1), 75–83 (1994).

[15] Eckstein, J., Bertsekas, D. P.: On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*. 55(1), 293–318 (1992).

[16] Eckstein, J.: *Splitting methods for monotone operators with applications to parallel optimization*. Doctoral dissertation, Massachusetts Institute of Technology (1989).

[17] Fazel, M., Pong, T.K., Sun, D.F., et al.: Hankel matrix rank minimization with applications to system identification and realization. *SIAM Journal on Matrix Analysis and Applications*. 34(3), 946–977 (2013).

[18] Fortin, M., Glowinski, R.: Chapter III on decomposition-coordination methods using an augmented lagrangian. *Studies in Mathematics and Its Applications*. 15, 97–146 (1983).

[19] Friedman, J., Hastie, T. and Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*. 33(1), 1–22 (2010).

[20] Friedman, J., Hastie, T., Tibshirani, R.: A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736* (2010).
[21] Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications.* 2(1), 17–40 (1976).

[22] Glowinski, R., Marroco, A.: Sur l’approximation, par éléments finis d’ordre un, et la résolution, par pénalisation-dualité d’une classe de problèmes de Dirichlet non linéaires. *Revue française d’automatique, informatique, recherche opérationnelle. Analyse numérique.* 9(2), 41–76 (1975).

[23] Glowinski, R.: *Lectures on numerical methods for non-linear variational problems.* Springer Science & Business Media (2008).

[24] Glowinski, R.: *On alternating direction methods of multipliers: a historical perspective.* Modeling, simulation and optimization for science and technology. Springer Netherlands, 59–82 (2014).

[25] Guo, H.: *The metric subregularity of KKT solution mapping of composite conic programming.* Doctoral dissertation, National University of Singapore (2017).

[26] Golub, G.H., Van Loan, C.F.: *Matrix computations.* Johns Hopkins University Press (1996).

[27] Han, D., Sun, D.F., Zhang, L.W.: Linear rate convergence of the alternating direction method of multipliers for convex composite programming. *Mathematics of Operations Research.* Accepted (2017). https://doi.org/10.1287/moor.2017.0875.

[28] Han, D., Yuan, X.: Local linear convergence of the alternating direction method of multipliers for quadratic programs. *SIAM Journal on Numerical Analysis.* 51(6), 3446–3457 (2013).

[29] Hiriart-Urruty, J.B., Lemaréchal, C.: *Convex analysis and minimization algorithms I: Fundamentals* (Vol. 305). Springer Science & Business Media. (2013).

[30] Jacob, L., Obozinski, G., Vert, J.P.: Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning.* 433–440 (2009).

[31] James, G.M., Paulson, C. Rusmevichientong, P.: Penalized and constrained regression. *Technical report.* (2013).

[32] Jenatton, R., Audibert, J.Y., Bach, F.: Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research.* 12(10), 2777–2824 (2011).

[33] Lemaréchal, C., Sagastizábal, C.: Practical Aspects of the Moreau–Yosida Regularization: Theoretical Preliminaries. *SIAM Journal on Optimization.* 7(2), 367–385 (1997).

[34] Li, M., Sun, D.F., Toh, K.-C.: A majorized ADMM with indefinite proximal terms for linearly constrained convex composite optimization. *SIAM Journal on Optimization.* 26 (2), 922–950 (2016).

[35] Li, X.D., Sun, D.F., Toh, K.-C.: A Schur complement based semi-proximal ADMM for convex quadratic conic programming and extensions. *Mathematical Programming.* 155 (1-2), 333–373 (2016).
[36] Li, X.D., Sun, D.F., Toh, K.-C.: A highly efficient semismooth Newton augmented Lagrangian method for solving Lasso problems. arXiv preprint arXiv:1607.05428v3 (2017).

[37] Li, X.D., Sun, D.F., Toh, K.-C.: A block symmetric Gauss-Seidel decomposition theorem for convex composite quadratic programming and its applications. arXiv preprint arXiv:1703.06629 (2017).

[38] Lichman, M.: UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science (2013).

[39] Liu, J., Yuan, L., Ye, J.: An efficient algorithm for a class of fused lasso problems. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM. 323–332 (2010).

[40] Moreau J.J.: Proximité et dualité dans un espace hilbertien. Bulletin de la Société Mathématique de France. 93, 273–299 (1965).

[41] Nesterov, Y.: Introductory lectures on convex optimization: A basic course (Vol. 87). Springer Science & Business Media. (2013).

[42] Robinson, S.M.: Some continuity properties of polyhedral multifunctions. Mathematical Programming at Oberwolfach. 206–214 (1981).

[43] Rockafellar, R.T., Wets, R.J.B.: Variational analysis (Vol. 317). Springer Science & Business Media. (2009).

[44] Rockafellar, R. T.: Convex analysis, Princeton University Press. (2015).

[45] Sun, J.: On monotropic piecewise quadratic programming. Doctoral dissertation, Department of Mathematics, University of Washington, Seattle. (1986).

[46] Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological). 58(1), 267–288 (1996).

[47] Tibshirani, R.: The lasso method for variable selection in the Cox model. Statistics in Medicine. 16(4), 385–395 (1997).

[48] Tibshirani, R., Saunders, M., Rosset S., et al.: Sparsity and smoothness via the fused lasso, Journal of the Royal Statistical Society: Series B (Statistical Methodology). 67(1), 91–108 (2005).

[49] Yang, W.H., Han D.: Linear convergence of the alternating direction method of multipliers for a class of convex optimization problems. SIAM Journal on Numerical Analysis. 54(2), 625–640 (2016).

[50] Ye, J.J.: Constraint qualifications and necessary optimality conditions for optimization problems with variational inequality constraints. SIAM Journal on Optimization. 10(4), 943–20 (2000).
A Appendix: Proofs of the Lemmas
A.1 Proof of Lemma 3.1

Proof. The optimality condition for (1.8a) is

\[ 0 \in \partial p(y^{k+1}) + Ax^k + \sigma A(A^* y^{k+1} + B^* z^k - c) + (S + \tilde{\Sigma}_f)(y^{k+1} - y^k). \quad (A.1) \]

It follows from (1.8c) that

\[ x^k + \sigma (A^* y^{k+1} + B^* z^k - c) = \tau^{-1}(x^{k+1} - x^k) + x^k - \sigma B^*(z^{k+1} - z^k). \quad (A.2) \]

From Proposition 2.1, we know that there exists \( W_f^k \in \text{conv}\partial f[y^{k+1}, y^k] \) such that

\[ \nabla f(y^k) - \nabla f(y^{k+1}) = W_f^k (y^k - y^{k+1}). \]

Substituting the above equation and (A.2) into (A.1), we get

\[ 0 \in \partial p(y^{k+1}) + \nabla f(y^{k+1}) + A(x^k + \tau^{-1}(x^{k+1} - x^k) - \sigma B^*(z^{k+1} - z^k)) + S^k (y^{k+1} - y^k), \]

where \( S^k := S + \tilde{\Sigma}_f - W_f^k \). Thus, it holds that

\[ y^{k+1} = \text{Pr}_P(y^{k+1} - \nabla f(y^{k+1}) - A[x^k + \tau^{-1}(x^{k+1} - x^k) - \sigma B^*(z^{k+1} - z^k)] + S^k (y^{k+1} - y^k)). \]

Similarly, there exists \( W_g^k \in \text{conv}\partial g(z^{k+1}, z^k) \) such that \( \nabla g(z^k) - \nabla g(z^{k+1}) = W_g^k (z^k - z^{k+1}) \) and

\[ z^{k+1} = \text{Pr}_Q(z^{k+1} - \nabla g(z^{k+1}) - B[\tau^{-1}(x^{k+1} - x^k) + x^k] - T^k(z^{k+1} - z^k)), \]

where \( T^k := T + \tilde{\Sigma}_g - W_g^k \). Since the Moreau-Yosida proximal mappings \( \text{Pr}_P(\cdot) \) and \( \text{Pr}_Q(\cdot) \) are globally Lipschitz continuous, one has that for any \( k \geq 1 \),

\[
\begin{align*}
& \| R(u^{k+1}) \|^2 \\
\leq & \| S^k (y^{k+1} - y^k) - \sigma AB^*(z^{k+1} - z^k) + (\tau^{-1} - 1)A(x^{k+1} - x^k) \|^2 \\
& + \| T^k (z^{k+1} - z^k) + (\tau^{-1} - 1)B(x^{k+1} - x^k) \|^2 + \| (\tau \sigma)^{-1}(x^{k+1} - x^k) \|^2 \\
\leq & 3\lambda_{\max}^2 S^k \| y^{k+1} - y^k \|^2 + 3\sigma \lambda_{\max}(A^* A) \| z^{k+1} - z^k \|^2 + 2\lambda_{\max}(T^k) \| z^{k+1} - z^k \|^2 + (1 - \tau^{-1})^2 [2\lambda_{\max}(B^* B) \| x^{k+1} - x^k \|^2 + 3\lambda_{\max}(A^* A) \| x^{k+1} - x^k \|^2].
\end{align*}
\]

(A.3)

Next, we estimate upper bounds of \( \lambda_{\min}(S^k) \) and \( \lambda_{\max}(T^k) \), respectively. It follows from \( \Sigma_f \preceq W_f^k \preceq \tilde{\Sigma}_f \) and \( \Sigma_g \preceq W_g^k \preceq \tilde{\Sigma}_g \) that

\[ -\frac{1}{2} \tilde{\Sigma}_f \preceq S^k \preceq S + \tilde{\Sigma}_f \quad \text{and} \quad -\frac{1}{2} \tilde{\Sigma}_g \preceq T^k \preceq T + \tilde{\Sigma}_g, \quad \forall k \geq 1. \]

Then

\[ \lambda_{\min}(S^k) \geq -\frac{1}{2} \lambda_{\max}(\tilde{\Sigma}_f) \quad \text{and} \quad \lambda_{\max}(S^k) \leq \lambda_{\max}(S + \frac{1}{2}\tilde{\Sigma}_f) + \frac{1}{2} \lambda_{\max}(\tilde{\Sigma}_f), \]

\[ \lambda_{\min}(T^k) \geq -\frac{1}{2} \lambda_{\max}(\tilde{\Sigma}_g) \quad \text{and} \quad \lambda_{\max}(T^k) \leq \lambda_{\max}(T + \frac{1}{2}\tilde{\Sigma}_g) + \frac{1}{2} \lambda_{\max}(\tilde{\Sigma}_g), \]

and consequently, for any \( k \geq 1 \), one has

\[ \lambda_{\max}(S^k) \leq \lambda_{\max}(S + \frac{1}{2}\tilde{\Sigma}_f) + \frac{1}{2} \lambda_{\max}(\tilde{\Sigma}_f), \lambda_{\max}(T^k) \leq \lambda_{\max}(T + \frac{1}{2}\tilde{\Sigma}_g) + \frac{1}{2} \lambda_{\max}(\tilde{\Sigma}_g). \]

By substituting the above two inequalities into (A.3), we can get (3.4). This completes the proof. \( \square \)

25
A.2 Proof of Lemma 3.2

Proof. For any $\varepsilon \in \mathbb{R}$, define

$$h_\varepsilon(x) := h(x) + \frac{\varepsilon^2}{2} \|x\|^2, \; \forall x \in X.$$ 

Then, similar to the proof of [41, Theorem 2.1.5], one has

$$\langle \nabla h_\varepsilon(x) - \nabla h_\varepsilon(\bar{x}), x - \bar{x} \rangle \geq \|\nabla h_\varepsilon(x) - \nabla h_\varepsilon(\bar{x})\|^2_{(\mathcal{P} + \varepsilon^2 I)^{-1}}, \; \forall \varepsilon \neq 0.$$ 

Consequently, for any $\varepsilon \neq 0$, it holds that

$$\langle \nabla h_\varepsilon(x) - \nabla h_\varepsilon(\bar{x}), y - \bar{x} \rangle \geq \|\nabla h_\varepsilon(x) - \nabla h_\varepsilon(\bar{x})\|^2_{(\mathcal{P} + \varepsilon^2 I)^{-1}} + \langle \nabla h_\varepsilon(x) - \nabla h_\varepsilon(\bar{x}), y - x \rangle$$

$$= \| (\mathcal{P} + \varepsilon^2 I)^{-1/2} (\nabla h_\varepsilon(x) - \nabla h_\varepsilon(\bar{x})) + \frac{1}{2} (\mathcal{P} + \varepsilon^2 I)^{1/2} (y - x) \|_2^2 - \frac{1}{4} \| x - y \|_{(\mathcal{P} + \varepsilon^2 I)}^2$$

$$\geq -\frac{1}{4} \| x - y \|_{(\mathcal{P} + \varepsilon^2 I)}^2.$$ 

This together with the definition of $h$ implies that

$$\langle \nabla h(x) - \nabla h(\bar{x}), y - \bar{x} \rangle + \langle \varepsilon^2(x - \bar{x}), y - \bar{x} \rangle \geq -\frac{1}{4} \| x - y \|_{(\mathcal{P} + \varepsilon^2 I)}^2, \; \forall \varepsilon \neq 0.$$ 

Therefore, by taking limits on both sides of the above inequality for $\varepsilon \to 0$, we complete the proof. $\square$

A.3 Proof of Lemma 3.3

Proof. By the first order optimality conditions of (1.8a) and (1.8b), one has

$$\begin{cases}
0 \in \partial p(y^{k+1}) + \nabla f(y^k) + (\bar{\Sigma}_f + \mathcal{S})(y^{k+1} - y^k) + A(x^k + \sigma(A^*y^{k+1} + B^*z^k - c)), \\
0 \in \partial q(z^{k+1}) + \nabla g(z^k) + (\bar{\Sigma}_g + \mathcal{T})(z^{k+1} - z^k) + B(x^k + \sigma x^{k+1}).
\end{cases} \quad (A.4)$$

Since $(\bar{y}, \bar{z}, \bar{x})$ is a KKT point, it holds that

$$\begin{cases}
0 \in \partial p(\bar{y}) + \nabla f(\bar{y}) + A\bar{x}, \\
0 \in \partial q(\bar{z}) + \nabla g(\bar{z}) + B\bar{x}.
\end{cases} \quad (A.5)$$

It follows from the maximal monotonicity of $\partial p$ that

$$0 \leq \langle -A(x^k + \sigma(A^*y^{k+1} + B^*z^k - c)) + A\bar{x}, y^{k+1} - \bar{y} \rangle + \langle \nabla f(\bar{y}) - \nabla f(y^k), y^{k+1} - \bar{y} \rangle$$

$$- \langle (\bar{\Sigma}_f + \mathcal{S})(y^{k+1} - y^k), y^{k+1} - \bar{y} \rangle.$$ 

Thus, by reorganizing the above inequality and Lemma 3.2, one has

$$\langle \bar{x} - (x^k + \sigma(A^*y^{k+1} + B^*z^k - c)), A^*(y^{k+1} - \bar{y}) \rangle - \langle (\bar{\Sigma}_f + \mathcal{S})(y^{k+1} - y^k), y^{k+1} - \bar{y} \rangle$$

$$\geq \langle \nabla f(y^k) - \nabla f(\bar{y}), y^{k+1} - \bar{y} \rangle \geq -\frac{1}{4} \| y^{k+1} - \bar{y} \|^2_{\bar{\Sigma}_f}. \quad (A.6)$$
Similarly, by using the maximal monotonicity of $\partial q$ and Lemma 3.2, it holds that
\[
\langle \bar{x} - (x^k + \sigma r^{k+1}), B^*(z^{k+1} - z^k) \rangle - \langle (\bar{\Sigma}_g + T)(z^{k+1} - z^k), z^{k+1} - z^k \rangle 
\geq -\frac{1}{4} z^{k+1} - z^k \|^{2}_{\bar{\Sigma}_g}. \tag{A.7}
\]
By adding (A.6) and (A.7) together, we have
\[
\Delta_k + \langle (\bar{\Sigma}_f + S)(y^{k+1} - y^k), \bar{y} - y^{k+1} \rangle + \langle (\bar{\Sigma}_g + T)(z^{k+1} - z^k), \bar{z} - z^{k+1} \rangle 
\geq -\frac{1}{4} y^{k+1} - y^k \|^{2}_{\bar{\Sigma}_f} - \frac{1}{4} z^{k+1} - z^k \|^{2}_{\bar{\Sigma}_g}, \tag{A.8}
\]
where
\[
\Delta_k := \langle \bar{x} - (x^k + \sigma r^{k+1}) + \sigma B^*(z^{k+1} - z^k), A^*(y^{k+1} - \bar{y}) \rangle + \langle \bar{x} - (x^k + \sigma r^{k+1}), B^*(z^{k+1} - \bar{z}) \rangle.
\]
Directly from [34, Equation (47)], it holds that
\[
\Delta_k = (2\sigma \tau)^{-1} \left( \|x^k - \bar{x}\|^2 - \|x^{k+1} - \bar{x}\|^2 \right) + (2\sigma)^{-1} (\tau - 1) \|r^{k+1}\|^2 
- \frac{\sigma}{2} \|A^* y^{k+1} + B^* z^k - c\|^2 + \frac{\sigma}{2} \left( \|B^* z^k - B^* \bar{z}\|^2 + \|B^* z^{k+1} - B^* \bar{z}\|^2 \right). \tag{A.9}
\]
Since for any self-adjoint linear operator $G$, it holds that $\langle u, Gu \rangle = \frac{1}{2} (\|u + v\|_G^2 - \|u\|_G^2 - \|v\|_G^2)$, we have
\[
\langle (\bar{\Sigma}_f + S)(y^{k+1} - y^k), \bar{y} - y^{k+1} \rangle + \langle (\bar{\Sigma}_g + T)(z^{k+1} - z^k), \bar{z} - z^{k+1} \rangle 
= \frac{1}{2} \left( \|y^k - \bar{y}\|^2_{\bar{\Sigma}_f + S} - \|y^{k+1} - \bar{y}\|^2_{\bar{\Sigma}_f + S} \right) - \frac{1}{2} \|y^{k+1} - y^k\|^2_{\bar{\Sigma}_f + S} 
+ \frac{1}{2} \left( \|z^k - \bar{z}\|^2_{\bar{\Sigma}_g + T} - \|z^{k+1} - \bar{z}\|^2_{\bar{\Sigma}_g + T} \right) - \frac{1}{2} \|z^{k+1} - z^k\|^2_{\bar{\Sigma}_g + T}.
\]
This, together with (A.9) and (A.8) implies that the conclusion holds. The proof is completed. \[\square\]

**A.4 Proof of Lemma 3.4**

*Proof.* From [34, Lemma 7], it holds that
\[
\|y^{k+1} - y^k\|^2_{\bar{\Sigma}_f + S} + \|z^{k+1} - z^k\|^2_{\bar{\Sigma}_g + T} + \sigma \|A^* y^{k+1} + B^* z^k - c\|^2 + (1 - \tau) \sigma \|r^{k+1}\|^2 
\geq \|y^{k+1} - y^k\|^2_{\bar{\Sigma}_f + S} + \|z^{k+1} - z^k\|^2_{\bar{\Sigma}_g + T} + \|z^{k+1} - z^k\|^2_{\bar{\Sigma}_g + T} 
+ \min(\tau, 1 + \tau - \tau^2) \sigma (\tau^{-1} \|r^{k+1}\|^2 + \|B^* (z^{k+1} - z^k)\|^2) + (1 - \min(\tau, \tau^{-1})) \sigma (\|r^{k+1}\|^2 - \|r^k\|^2).
\]
This, together with the definition of $\phi_k$ and Lemma 3.3 implies the conclusion. \[\square\]

**A.5 Proof of Lemma 3.5**

*Proof.* First we show that
\[
\frac{1}{2} \bar{\Sigma}_f + S + \sigma A A^* \succ 0 \& \frac{1}{2} \bar{\Sigma}_g + T + \sigma B B^* \succ 0 \Rightarrow \mathcal{H} \succ 0, \mathcal{M} \succ 0.
\]
27
Suppose that $\frac{1}{2} \hat{\Sigma}_f + S + \sigma AA^* \succ 0 \& \frac{1}{2} \hat{\Sigma}_g + T + \sigma BB^* \succ 0$, but there exists a vector $0 \neq d := (d_y, d_z, d_x) \in \mathcal{Y} \times \mathcal{Z} \times \mathcal{X}$ such that $\langle d, \mathcal{H} d \rangle = 0$, by using the definition of $\mathcal{H}$, we have

$$\langle d_y, \mathcal{H}_f d_y \rangle = 0, \langle d_z, \mathcal{H}_g d_z \rangle = 0, \ d_x = 0, \ \mathcal{E}^*(d_y, d_z, d_x) = 0.$$  

Since $t_\tau > 0$ and $\frac{1}{2} \hat{\Sigma}_g + T + \sigma BB^* \succ 0$, we know $d_z = 0$. Consequently, $A^* d_y = 0$. This together with the assumption that $\frac{1}{2} \hat{\Sigma}_f + S + \sigma AA^* \succ 0$ implies $d_y = 0$. This contradiction shows that $\mathcal{H} \succ 0$. We can get $\mathcal{M} \succ 0$ by using the same techniques. For brevity, we omit the details.

Next, we show that $\mathcal{H} \succ 0 \Rightarrow \frac{1}{2} \hat{\Sigma}_f + S + \sigma AA^* \succ 0 \& \frac{1}{2} \hat{\Sigma}_g + T + \sigma BB^* \succ 0$. Since $t_\tau > 0$ and for any $d = (d_y, 0, 0) \in \mathcal{Y} \times \mathcal{Z} \times \mathcal{X}$, we have $\langle d, \mathcal{H} d \rangle = \langle d_y, (\mathcal{H}_f + \frac{1}{4} t_\tau \sigma AA^*) d_y \rangle$, then $\frac{1}{2} \hat{\Sigma}_f + S + \sigma AA^* \succ 0$ by the definition of $\mathcal{H}_f$. Similarly, for any $d = (0, d_z, 0) \in \mathcal{Y} \times \mathcal{Z} \times \mathcal{X}$, $\langle d, \mathcal{H} d \rangle = \langle d_z, (\mathcal{H}_g + \frac{1}{4} t_\tau \sigma BB^*) d_z \rangle$, then we can obtain that $\frac{1}{2} \hat{\Sigma}_g + T + \sigma BB^* \succ 0$. The proof is completed.

\[\blacksquare\]