J. Oper. Res. Soc. China (2017) 5:131–159
DOI 10.1007/s40305-016-0131-5

First-Order Algorithms for Convex Optimization with Nonseparable Objective and Coupled Constraints

Xiang Gao[1] · **Shu-Zhong Zhang**[1]

Received: 18 February 2016 / Revised: 16 April 2016 / Accepted: 9 May 2016 /
Published online: 18 June 2016

Abstract In this paper, we consider a block-structured convex optimization model, where in the objective the block variables are nonseparable and they are further linearly coupled in the constraint. For the 2-block case, we propose a number of first-order algorithms to solve this model. First, the alternating direction method of multipliers (ADMM) is extended, assuming that it is easy to optimize the augmented Lagrangian function with one block of variables at each time while fixing the other block. We prove that $O(1/t)$ iteration complexity bound holds under suitable conditions, where t is the number of iterations. If the subroutines of the ADMM cannot be implemented, then we propose new alternative algorithms to be called alternating proximal gradient method of multipliers, alternating gradient projection method of multipliers, and the hybrids thereof. Under suitable conditions, the $O(1/t)$ iteration complexity bound is shown to hold for all the newly proposed algorithms. Finally, we extend the analysis for the ADMM to the general multi-block case.

Keywords First-order algorithms · ADMM · Proximal gradient method · Convex optimization · Iteration complexity

This paper is dedicated to Professor Lian-Sheng Zhang in celebration of his 80th birthday.

This paper was supported in part by National Science Foundation (No. CMMI-1462408).

✉ Shu-Zhong Zhang
zhangs@umn.edu

Xiang Gao
gaoxx460@umn.edu

[1] Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN 55455, USA

Mathematics Subject Classification 90C25 · 49M27 · 68Q25

1 Introduction

In this paper, we consider the following model:

$$\begin{array}{ll} \min & f(x,y)+h_1(x)+h_2(y) \\ \text{s.t.} & Ax+By=b, \\ & x\in\mathcal{X},\ y\in\mathcal{Y}, \end{array} \tag{1.1}$$

where $x\in\mathbb{R}^p$, $y\in\mathbb{R}^q$, $A\in\mathbb{R}^{m\times p}$, $B\in\mathbb{R}^{m\times q}$, $b\in\mathbb{R}^m$, $\mathcal{X}$, $\mathcal{Y}$ are closed convex sets, f is a smooth jointly convex function, and h_1, h_2 are (possibly nonsmooth) convex functions. The so-called augmented Lagrangian function for problem (1.1) is

$$\mathcal{L}_\gamma(x,y,\lambda)=f(x,y)+h_1(x)+h_2(y)-\lambda^{\mathrm{T}}(Ax+By-b)+\frac{\gamma}{2}\|Ax+By-b\|^2,$$

where λ is the multiplier.

Many emerging applications from various fields can be formulated as optimization problems in the form of (1.1). For instance, the constrained lasso (classo) problem is described as

$$\begin{array}{ll} \min\limits_{\beta} & \frac{1}{2}\|Y-X\beta\|_2^2+\tau\|\beta\|_1 \\ \text{s.t.} & C\beta\leqslant b, \end{array} \tag{1.2}$$

where $X\in\mathbb{R}^{m\times p}, Y\in\mathbb{R}^m$ are the observed data, and $C\in\mathbb{R}^{n\times p}, b\in\mathbb{R}^n$ are predefined matrix and vector, respectively. The classo problem was first studied by James et al. [1] as a generalization of the lasso problem. By introducing additional linear constraints, [1] shows that many widely used statistical models can be expressed as special cases of (1.2), including the fused lasso, generalized lasso, monotone curve estimation, and so on. In fact, we can partition the variable β into blocks as $\beta=(\beta_1^{\mathrm{T}},\cdots,\beta_K^{\mathrm{T}})^{\mathrm{T}}$ where $\beta_i\in\mathbb{R}^{p_i}$ and partition other matrices and vectors in (1.2) correspondingly. Moreover, if we also introduce another slack variable z, then the classo problem can be cast in the form of (1.1) as follows:

$$\begin{array}{ll} \min\limits_{\beta} & \frac{1}{2}\left\|Y-\sum\limits_{i=1}^{K}X_i\beta_i\right\|_2^2+\tau\sum\limits_{i=1}^{K}\|\beta_i\|_1 \\ \text{s.t.} & \sum\limits_{i=1}^{K}C_i\beta_i+z=b,\ \ z\geqslant 0. \end{array} \tag{1.3}$$

The second example is the demand response (DR) control problem in the smart grid system [2–4]. Basically, it tries to minimize the cost incurred to a utility company which purchases electricity from the electricity market. To achieve this, the utility company controls the power consumption of some appliances of the users. Specifically, the problem can be formulated as

$$\begin{array}{ll}\min\limits_{x,p} & C_t\left(\left(\sum\limits_{i=1}^{n}\Psi_i x_i - p\right)^+\right) + C_s\left(\left(p - \sum\limits_{i=1}^{n}\Psi_i x_i\right)^+\right) + C_d(p) \\ \text{s.t.} & x \geqslant 0,\ \ p \geqslant 0, \\ & x_i \in \mathcal{X}_i,\ i = 1, 2, ..., n,\end{array} \tag{1.4}$$

where $C_t(\cdot)$ and $C_s(\cdot)$ are the cost functions that measure the insufficient and the excessive power bids, respectively, and $C_d(\cdot)$ is the bidding cost function; see [3]. The variable $p \in \mathbb{R}^L$ represents the amount of electricity that the utility company bids from a day-ahead market, and x_i is the control variable of the usage of the appliances of customer i, and $\Psi_i \in \mathbb{R}^{L\times n_i}$ is the matrix of the appliance load profile of customer i, thus $\Psi_i x_i$ is the total electricity consumption of customer i; see [5]. By introducing a new variable $z = \left(\sum\limits_{i=1}^{n}\Psi_i x_i - p\right)^+$, the above problem can be rewritten in the form of (1.1) as

$$\begin{array}{ll}\min\limits_{x,p} & C_t(z) + C_s\left(z + p - \sum\limits_{i=1}^{n}\Psi_i x_i\right) + C_d(p) \\ \text{s.t.} & z + p - \sum\limits_{i=1}^{n}\Psi_i x_i \geqslant 0, \\ & z \geqslant 0,\ \ x \geqslant 0,\ \ p \geqslant 0, x_i \in \mathcal{X}_i,\ i = 1, 2, ..., n.\end{array} \tag{1.5}$$

In fact, optimization problems in the form of (1.1) have many other interesting applications in various areas including signal processing, image processing, machine learning, and statistical learning; see [6,7] and the references therein.

In the case where no coupling term $f(x, y)$ is in the objective, there is a well-known algorithm—Alternating Direction Method of Multipliers (ADMM)—established for solving (1.1). The iterative scheme of the ADMM runs as follows:

$$\begin{cases} x^{k+1} = \text{argmin}_{x\in\mathcal{X}}\, \mathcal{L}_\gamma(x, y^k, \lambda^k), \\ y^{k+1} = \text{argmin}_{y\in\mathcal{Y}}\, \mathcal{L}_\gamma(x^{k+1}, y, \lambda^k), \\ \lambda^{k+1} = \lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b). \end{cases} \tag{1.6}$$

The ADMM is known to be a manifestation of the so-called operator splitting method which can be traced back to 1970s. Considerable amount of early studies on the ADMM can be found in, e.g., [8–11]. The method has gained new momentum in the recent years because of its first-order nature, and its potentials to compute distributively, which are important characteristics for solving very large scale problem instances. For an overview on its recent developments, one is referred to the surveys [12–15] and the references therein. The convergence properties of the ADMM are well known. In fact, its convergence follows from that of the so-called Douglas-Rachford operator splitting method; see [11,16]. However, the rate of convergence was only established recently. In particular, [17,18] show that the ADMM converges at the rate of $O(1/t)$, where t is the number of total iterations. Furthermore, with additional conditions on the objective function or the constraints, the ADMM can be shown to converge linearly; see [19–22]. One extension of the ADMM is to allow multi-block of variables. That extension of the ADMM turns out to perform very well for many instances encountered in practice, compared to other competing first-order alternatives. However, it may fail to converge in general. Specifically, in [23] by constructing a counterexample, the authors show that the ADMM may diverge with 3 blocks of variables. Therefore, it became clear that some additional conditions will be necessary in order to guarantee convergence, other than mere convexity. Indeed, under the strong convexity condition on some parts of the objective and certain assumptions on the constraint matrices, [24–27] show that an $O(1/t)$ convergence rate can still be achieved for the multi-block ADMM. Another direction of study is to consider applications of ADMM to solve nonconvex problems. For instance, [28] shows that the ADMM can converge for some restricted class of nonconvex problems.

The current paper considers the ADMM in the presence of a coupling term f in the objective. Only a handful of papers in the literature considered this model so far, most noticeably [7] and [6]. In [7], the authors consider the general multi-block setting and they propose a upper-bound minimization method of multipliers (BSUMM) approach to cope with the nonseparability of the objective in (1.1); essentially, the nonseparable part of the augmented Lagrangian function is replaced by an upper bound. Under some error bound conditions and a diminishing dual stepsize assumption, the authors are able to show that the iterates produced by the BSUMM algorithm converge to the set of primal-dual optimal solutions. Very recently, Cui et al. [6] consider the problem (1.1) by introducing a quadratic upper bound function for the nonseparable part of augmented Lagrangian function; they show that their algorithm has an $O(1/t)$ convergence rate.

Our contribution In this paper, we study the ADMM and its variants for (1.1). (Some adaptations of the ADMM are particularly relevant if there is a coupling term in the objective, as the minimization subroutines required by the ADMM may become difficult to implement; more discussions on this later.) Instead of using some upper bound approximation (a.k.a. majorization-minimization), we work with the original objective function. In this context, we may extend the ADMM approach directly to solve this more general formulation. It turns out that under the assumptions that the gradient of the coupling function ∇f is Lipschitz continuous and one of h_1 and h_2 is strongly convex, then an $O(1/t)$ convergence rate can still be assured. In some applications, it is difficult or impossible to implement the ADMM iteration, because the augmented Lagrangian function in (1.6) may be difficult to optimize even if the other block of variables and the Lagrangian multipliers are fixed. This motivates us to propose the Alternating Proximal Gradient Method of Multipliers (APGMM), which essentially iterates between proximal gradient methods of each block variables before the multiplier is updated. We show that the APGMM has a convergence rate of $O(1/t)$ if ∇f is Lipschitz continuous. If optimizing the augmented Lagrangian function for one block of variables is easy while optimizing the other block of variables is difficult, then a hybrid between ADMM and APGMM is a natural choice. We show in that case an $O(1/t)$ convergence rate remains valid. What if the gradient proximal subroutines are still too difficult to be implemented? One would then opt to compute the gradient projections. Hence, we propose the Alternating Gradient Projection Method of Multipliers (AGPMM), which replaces the proximal gradient steps in APGMM by the gradient projections. Fortunately, the same $O(1/t)$ iteration bound still holds for such simplifications as well as its ADMM hybrid version. At this stage, all the methods mentioned above are considered in the context of the 2-block model (1.1). In general however, they can be extended to the multi-block model with a coupling term. Similarly, under the Lipschitz continuity of ∇f and the assumptions in [27], an $O(1/t)$ iteration bound still holds for the multi-block model.

The rest of the paper is organized as follows. In Sect. 2, we present some preliminary results and notations. In Sects. 3 and 4, we introduce ADMM, APGMM, AGPMM, and their hybrids. The results on the rate of convergence of these algorithms are presented in the subsections of the same section, while the detailed proofs of the convergence results are presented in Appendix 1. In Sect. 5, we extend our analysis of the ADMM to a general setting with multiple (more than 2) blocks of variables. Finally, we conclude the paper in Sect. 6.

2 Preliminaries

Let us first introduce some notations that will be frequently used in the analysis later. The aggregated primal variables x, y and the primal-dual variables x, y, λ are, respectively, denoted by u and w, and the primal-dual mapping F, namely

$$u := \begin{pmatrix} x \\ y \end{pmatrix},\ w := \begin{pmatrix} x \\ y \\ \lambda \end{pmatrix},\ F(w) := \begin{pmatrix} -A^{\mathrm{T}}\lambda \\ -B^{\mathrm{T}}\lambda \\ Ax + By - b \end{pmatrix}, \tag{2.1}$$

and $h(u) := f(x, y) + h_1(x) + h_2(y)$.

Throughout this paper, we assume f to be smooth and has a Lipschitz continuous gradient; i.e.,

Assumption 2.1 The coupling function f satisfies

$$\|\nabla f(u_2) - \nabla f(u_1)\| \leqslant L\|u_2 - u_1\|,\ \forall u_1, u_2 \in \mathcal{X} \times \mathcal{Y}, \tag{2.2}$$

where L is a Lipschitz constant for ∇f.

For a function f satisfying Assumption 2.1, it is useful to note the following inequalities.

Lemma 2.2 *Suppose that function f satisfies (2.2), then we have*

$$f(u_2) \leqslant f(u_1) + \nabla f(u_1)^{\mathrm{T}}(u_2 - u_1) + \frac{L}{2}\|u_2 - u_1\|^2, \tag{2.3}$$

for any u_1, u_2. In general, if f is also convex then

$$f(u_2) \leqslant f(u_1) + \nabla f(u_3)^{\mathrm{T}}(u_2 - u_1) + \frac{L}{2}\|u_2 - u_3\|^2, \tag{2.4}$$

for any u_1, u_2, u_3.

Inequality (2.3) is well known; see [29]. Moreover, (2.4) can be found as Fact 2 in [30].

For convenience of the analysis, we introduce some matrix notations. Let

$$Q := \begin{pmatrix} G & 0 & 0 \\ 0 & \gamma B^{\mathrm{T}} B & 0 \\ 0 & -B & \frac{1}{\gamma} I_m \end{pmatrix}, \; P := \begin{pmatrix} I_p & 0 & 0 \\ 0 & I_q & 0 \\ 0 & -\gamma B & I_m \end{pmatrix},$$

$$M := \begin{pmatrix} G & 0 & 0 \\ 0 & \gamma B^{\mathrm{T}} B & 0 \\ 0 & 0 & \frac{1}{\gamma} I_m \end{pmatrix}; \tag{2.5}$$

hence, $Q = MP$. Given a sequence $\{w^k\}$, we denote an associated auxiliary sequence to be

$$\tilde{w}^k := \begin{pmatrix} \tilde{x}^k \\ \tilde{y}^k \\ \tilde{\lambda}^k \end{pmatrix} = \begin{pmatrix} x^{k+1} \\ y^{k+1} \\ \lambda^k - \gamma(Ax^{k+1} + By^k - b) \end{pmatrix}. \tag{2.6}$$

Based on (2.6) and (2.5), the relationship between the new sequence $\{\tilde{w}^k\}$ and the original $\{w^k\}$ is

$$w^{k+1} = w^k - P\left(w^k - \tilde{w}^k\right). \tag{2.7}$$

3 Alternating Direction Method of Multipliers

As we discussed earlier, the ADMM can be applied straightforwardly to solve (1.1), assuming that the augmented Lagrangian (with a proximal term) can be optimized for each block of variables, while the other variables are fixed. This gives rise to the following scheme:

```
ADMM
Initialize $x^0 \in \mathcal{X}$, $y^0 \in \mathcal{Y}$ and $\lambda^0$
for $k = 0, 1, \cdots,$ do
    $x^{k+1} = \mathrm{argmin}_{x \in \mathcal{X}}\, \mathcal{L}_\gamma(x, y^k, \lambda^k) + \frac{1}{2}\|x - x^k\|_G^2$;
    $y^{k+1} = \mathrm{argmin}_{y \in \mathcal{Y}}\, \mathcal{L}_\gamma(x^{k+1}, y, \lambda^k) + \frac{1}{2}\|y - y^k\|_H^2$;
    $\lambda^{k+1} = \lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b)$.
end for
```

In the above algorithm, G and H are two pre-specified positive semidefinite matrices. In fact, this algorithm is also known as the G-ADMM and proposed in [20] where no coupled objective function is involved. The main result concerning its convergence and iteration complexity is summarized in the following theorem, whose proof can be found in Appendix 1.

Theorem 3.1 *Suppose that ∇f satisfies Lipschitz condition (2.2), and $h_2(y)$ is strongly convex with parameter $\sigma > 0$, i.e.,*

$$h_2(y) \geqslant h_2(z) + h_2'(z)^{\mathrm{T}}(y - z) + \frac{\sigma}{2}\|y - z\|^2, \tag{3.1}$$

where $h_2'(z) \in \partial h_2(z)$ is a subgradient of $h_2(z)$. Let $\{w^k\}$ be the sequence generated by the ADMM, and $G \succ 0$, $H \succ \left(L + \frac{L^2}{\sigma}\right) I_q$. Then the sequence $\{w^k\}$ generated by the ADMM converges to an optimal solution. Moreover, for any integer $n > 0$ letting

$$\bar{u}_n := \frac{1}{n}\sum_{k=1}^{n} u^k, \tag{3.2}$$

we have

$$\begin{aligned} &h(\bar{u}_t) - h(u^*) + \rho\|A\bar{x}_t + B\bar{y}_t - b\| \\ &\leqslant \frac{1}{2t}\left(\mathrm{dist}(x^0, \mathcal{X}^*)_G^2 + \mathrm{dist}(y^0, \mathcal{Y}^*)_{\hat{H}}^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right), \end{aligned} \tag{3.3}$$

where $\mathcal{X}^* \times \mathcal{Y}^*$ is the optimal solution set, $\mathrm{dist}(x, S)_M := \inf_{y \in S}\|x - y\|_M$, *and* $\hat{H} := \gamma B^{\mathrm{T}} B + H$.

The following lemma shows the connection between different convergence measures.

Lemma 3.2 (Lemma 2.4 in [31]) *Assume that $\rho > 0$, and $\bar{x} \in X$ is an approximate solution of the problem $f^* := \inf\{f(x) : Ax - b = 0, x \in X\}$ where f is convex, satisfying*

$$f(\bar{x}) - f^* + \rho\|A\bar{x} - b\| \leqslant \varepsilon. \tag{3.4}$$

Then, we have

$$\|A\bar{x} - b\| \leqslant \frac{\varepsilon}{\rho - \|\lambda^*\|} \text{ and } f(\bar{x}) - f^* \leqslant \varepsilon,$$

where λ^* is an optimal Lagrange multiplier associated with the constraint $Ax - b = 0$ in the problem $\inf\{f(x) : Ax - b = 0, x \in X\}$, *assuming $\|\lambda^*\| < \rho$.*

In other words, estimation (3.3) in Theorem 3.1 automatically establishes that

$$h(\bar{u}_t) - h(u^*) \leqslant O(1/t) \text{ and } \|A\bar{x}_t + B\bar{y}_t - b\| \leqslant O(1/t).$$

The same applies to all subsequent iteration complexity results presented in this paper.

4 Variants of the ADMM

In some applications, the augmented Lagrangian function may be difficult to minimize for some block of variables while fixing all others. For instance, consider the sparse logistic regression problem (see [32]) given by

$$\min_{x,c}\, l(x, c) + \beta\|x\|_1, \tag{4.1}$$

where $l(x, c) = \frac{1}{m} \sum_{i=1}^{m} \log(1 + \exp(-b_i(a_i^{\mathrm{T}} x + c)))$ and $\{(a_i, b_i),\ i = 1, \cdots, m\}$ is a given training set with m samples $a_1, a_2, \cdots, a_m$ and $b_i \in \{\pm 1\}, i = 1, \cdots, m$ as the binary class labels. To solve this problem within the ADMM framework, we can introduce a new variable z, and reformulate the problem as

$$\begin{aligned} &\min_{x,z,c}\ l(x, c) + \beta \|z\|_1 \\ &\text{s.t.}\quad x - z = 0. \end{aligned} \tag{4.2}$$

As in the ADMM, although the subroutine of solving z is easy, the subroutine of solving (x, c) can be difficult. In order to deal with those types of problems, we propose several variants of ADMM which incorporate different first-order methods into the ADMM framework.

4.1 Alternating Proximal Gradient Method of Multipliers

In this subsection, we consider an approach where we apply *proximal gradient* for each block of variables. The method bears some similarity to the Iterative Shrinkage-Thresholding Algorithm (ISTA) (cf. [33]), although we are dealing with multiple blocks of variables here. We shall call the new method APGMM, presented as follows:

```
APGMM
Initialize $x^0 \in \mathcal{X}$, $y^0 \in \mathcal{Y}$ and $\lambda^0$
for $k = 0, 1, \cdots,$ do
    $x^{k+1} = \mathrm{argmin}_{x \in \mathcal{X}}\ \nabla_x f(x^k, y^k)^{\mathrm{T}}(x - x^k) + h_1(x) + \frac{\gamma}{2}\|Ax + By^k$
            $-b - \frac{1}{\gamma}\lambda^k\|^2 + \frac{1}{2}\|x - x^k\|_G^2;$
    $y^{k+1} = \mathrm{argmin}_{y \in \mathcal{Y}}\ \nabla_y f(x^k, y^k)^{\mathrm{T}}(y - y^k) + h_2(y) + \frac{\gamma}{2}\|Ax^{k+1}$
            $+By - b - \frac{1}{\gamma}\lambda^k\|^2 + \frac{1}{2}\|y - y^k\|_H^2;$
    $\lambda^{k+1} = \lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b).$
end for
```

The convergence property and iteration complexity are summarized in the following theorem, whose proof is in Appendix 1.

Theorem 4.1 *Suppose that ∇f satisfies Lipschitz condition (2.2). Let $\{w^k\}$ be the sequence generated by the APGMM, and $G \succ L I_p$ and $H \succ L I_q$. Then, the sequence $\{w^k\}$ generated by the APGMM converges to an optimal solution. Moreover, with the same notations as before, it holds that $h(\bar{u}_t) - h(u^*) + \rho\|A\bar{x}_t + B\bar{y}_t - b\| \leqslant O(1/t)$.*

4.2 Alternating Gradient Projection Method of Multipliers

Implementing the proximal gradient step may be difficult for some applications. One may wonder if it is possible to further simplify the subroutines. It is therefore natural to consider the simple Gradient Projection method. Namely, for each block of variables, we simply sequentially compute the projection of the gradient of the augmented Lagrangian function before updating the multipliers. The method is depicted as follows:

```
AGPMM
Initialize $x^0 \in \mathcal{X}$, $y^0 \in \mathcal{Y}$ and $\lambda^0$
for $k = 0, 1, \cdots,$ do
    $x^{k+1} = [x^k - \alpha(\nabla_x f(x^k, y^k) + \nabla_x h_1(x^k) - A^{\mathrm{T}}\lambda^k + A^{\mathrm{T}}(Ax^k + By^k - b))]_{\mathcal{X}}$;
    $y^{k+1} = [y^k - \alpha(\nabla_y f(x^k, y^k) + \nabla_y h_2(y^k) - B^{\mathrm{T}}\lambda^k + B^{\mathrm{T}}(Ax^{k+1} + By^k - b))]_{\mathcal{Y}}$;
    $\lambda^{k+1} = \lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b)$.
end for
```

where $[x]_{\mathcal{X}}$ denotes the projection of x onto $\mathcal{X}$, and $[y]_{\mathcal{Y}}$ denotes the projection of y onto $\mathcal{Y}$.

Note here that we used ‘PG’ as acronym for Proximal Gradient, and ‘GP’ as acronym for Gradient Projection. The acronyms are quite similar, and so some attention is needed not to confuse the two! Below we shall present the main convergence and the iteration complexity results for the above method; the proof of the theorem can be found in Appendix 1.

Theorem 4.2 *Suppose that ∇f satisfies Lipschitz condition (2.2). Let w^k be the sequence generated by the AGPMM, and $G := \gamma A^{\mathrm{T}}A + \frac{1}{\alpha}I_p$, $H := \frac{1}{\alpha}I_q - \gamma B^{\mathrm{T}}B$. Moreover, suppose that α is chosen to satisfy $H - 2LI_q \succ 0$ and $G - 2LI_p \succ 0$. Then, the sequence $\{w^k\}$ generated by the AGPMM converges to an optimal solution. Moreover, with the same notations as before, it holds that $h(\bar{u}_t) - h(u^*) + \rho\|A\bar{x}_t + B\bar{y}_t - b\| \leqslant O(1/t)$.*

4.3 Hybrids

Similar to the sparse logistic regression problem in (4.1), there are instances where one part of the block variables is easy to deal with, while the other part is difficult (e.g., the fused logistic regression in [34]). To take advantage of that situation, we propose the following two types of hybrid methods. The first one is to combine ADMM with Proximal Gradient (ADM-PG) in two blocks of variables:

```
ADM-PG
Initialize $x^0 \in \mathcal{X}$, $y^0 \in \mathcal{Y}$ and $\lambda^0$
for $k = 0, 1, \cdots,$ do
    $x^{k+1} = \mathrm{argmin}_{x \in \mathcal{X}}\ \mathcal{L}_\gamma(x, y^k, \lambda^k) + \frac{1}{2}\|x - x^k\|_G^2$ ;
    $y^{k+1} = \mathrm{argmin}_{y \in \mathcal{Y}}\ \nabla_y f(x^{k+1}, y^k)^{\mathrm{T}}(y - y^k) + h_2(y) + \frac{\gamma}{2}\|Ax^{k+1}$
              $+By - b - \frac{1}{\gamma}\lambda^k\|^2 + \frac{1}{2}\|y - y^k\|_H^2$ ;
    $\lambda^{k+1} = \lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b)$.
end for
```

The iteration complexity of the above method is as follows. The proof of the theorem can be found in Appendix 1.

Theorem 4.3 *Suppose that ∇f satisfies Lipschitz condition (2.2). Let w^k be the sequence generated by the ADM-PG, and $G \succ 0, H \succ LI_q$. Then, the sequence $\{w^k\}$ generated by the APGMM converges to an optimal solution. Moreover, with the same notations as before, it holds that $h(\bar{u}_t) - h(u^*) + \rho\|A\bar{x}_t + B\bar{y}_t - b\| \leqslant O(1/t)$.*

Another possible approach is to combine ADMM with Gradient Projection (ADM-GP), which works as follows:

```
ADM-GP
Initialize $x^0 \in \mathcal{X}$, $y^0 \in \mathcal{Y}$ and $\lambda^0$
for $k = 0, 1, \cdots,$ do
    $x^{k+1} = \operatorname{argmin}_{x \in \mathcal{X}} \mathcal{L}_\gamma(x, y^k, \lambda^k) + \frac{1}{2}\|x - x^k\|_G^2$;
    $y^{k+1} = [y^k - \alpha(\nabla_y f(x^{k+1}, y^k) + \nabla_y h_2(y^k) - B^{\mathrm{T}}\lambda^k + B^{\mathrm{T}}(Ax^{k+1} + By^k - b))]_{\mathcal{Y}}$;
    $\lambda^{k+1} = \lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b)$.
end for
```

The main convergence result is as follows, and the proof of the theorem can be found in Appendix 1.

Theorem 4.4 *Let w^k be the sequence generated by the ADM-GP, $G \succ 0$, and $H := \frac{1}{\alpha}I_q - \gamma B^{\mathrm{T}}B$. Moreover, suppose that α is chosen to satisfy $H - LI_q \succ 0$. Then, the sequence $\{w^k\}$ generated by the ADM-GP converges to an optimal solution. Moreover, with the same notations as before, it holds that $h(\bar{u}_t) - h(u^*) + \rho\|A\bar{x}_t + B\bar{y}_t - b\| \leqslant O(1/t)$.*

Remark that for all the algorithms discussed above, besides showing the $O(1/t)$ rate of convergence in the ergodic sense, we have also shown the convergence of the iterates generated by the algorithms. Moreover, for all the variants of ADMM, we do not assume the strong convexity of the objective functions. Another point to note is that AGPMM and ADM-GP can be viewed as special cases of APGMM and ADM-PG, respectively. In fact, one can absorb the separable function h_i into the nonseparable function f and choose matrices G and H appropriately in such a way that APGMM and ADM-PG actually become AGPMM and ADM-GP, respectively.

5 The General Multi-Block Model

Different variations of the ADMM have been a popular subject of study in the recent years. In particular, ADMM has been extended to solve general formulation with multiple blocks of variables; see [27] and the references therein for more information. In this section, we shall discuss the iteration complexity of the ADMM for multi-block optimization with a nonseparable objective function. In particular, the problem that we consider is as follows:

$$\begin{array}{ll} \min & f(x_1, x_2, \cdots, x_n) + \sum\limits_{i=1}^{n} h_i(x_i) \\ \text{s.t.} & A_1x_1 + A_2x_2 + \cdots + A_nx_n = b, \\ & x_i \in \mathcal{X}_i, i = 1, 2, \cdots, n, \end{array} \tag{5.1}$$

where $A_i \in \mathbb{R}^{m\times p_i}$, $b \in \mathbb{R}^m$, $\mathcal{X}_i \subset \mathbb{R}^{p_i}$ are closed convex sets, and f, h_i $i = 1, \cdots, n$, are convex functions. Note that many important applications are in the form of (5.1), e.g., multi-stage stochastic programming. Accordingly, the ADMM algorithm for solving the problem (5.1) is

```
The Multi-block ADMM
Initialize with $x_i^0 \in \mathcal{X}_i, i = 1, \cdots, n$, and $\lambda^0$
for $k = 0, 1, \cdots,$ do
    $x_1^{k+1} = \mathrm{argmin}_{x_1\in\mathcal{X}_1} \mathcal{L}_\gamma(x_1, x_2^k, \cdots, x_n^k, \lambda^k) + \frac{1}{2}\|x_1 - x_1^k\|_{H_1}^2$;
    $x_2^{k+1} = \mathrm{argmin}_{x_2\in\mathcal{X}_2} \mathcal{L}_\gamma(x_1^{k+1}, x_2, x_3^k, \cdots, x_n^k, \lambda^k) + \frac{1}{2}\|x_2 - x_2^k\|_{H_2}^2$;
        $\vdots$
    $x_i^{k+1} = \mathrm{argmin}_{x_i\in\mathcal{X}_i} \mathcal{L}_\gamma(x_1^{k+1}, \cdots, x_{i-1}^{k+1}, x_i, x_{i+1}^k, \cdots, x_n^k, \lambda^k) + \frac{1}{2}\|x_i - x_i^k\|_{H_i}^2$;
        $\vdots$
    $x_n^{k+1} = \mathrm{argmin}_{x_n\in\mathcal{X}_n} \mathcal{L}_\gamma(x_1^{k+1}, \cdots, x_{n-1}^{k+1}, x_n, \lambda^k) + \frac{1}{2}\|x_n - x_n^k\|_{H_n}^2$;
    $\lambda^{k+1} = \lambda^k - \beta(A_1x_1^{k+1} + A_2x_2^{k+1} + \cdots + A_nx_n^{k+1})$.
end for
```

Here, $H_i, i = 1, \cdots, n$, are pre-specified positive semidefinite matrices, γ is the augmented Lagrangian constant, and β is the dual stepsize. In fact, this also generalizes the ADMM in the sense that a proximal term is included in the subproblem. Without the coupled objective function, this algorithm has also been analyzed in [35]. As we will show, an $O(1/t)$ convergence rate of the ADMM can still be achieved even for this general problem. In the following subsection, we sketch a convergence rate analysis highlighting the key components and steps. The details, however, will be omitted for succinctness.

Let us start with the assumptions.

Assumption 5.1 The functions h_i, $i = 2, \cdots, n$, are strongly convex with parameters $\sigma_i > 0$:

$$h_i(y) \geqslant h_i(x) + (y-x)^{\mathrm{T}} h_i^{'}(x) + \frac{\sigma_i}{2}\|y-x\|^2,$$

where $h_i^{'}(x) \in \partial h_i(x)$ is in subdifferential of $h_i(x)$.

Assumption 5.2 The gradient of function $f(x_1, x_2, \cdots, x_n)$ is Lipschitz continuous with parameter $L \geqslant 0$:

$$\|\nabla f(x_1', x_2', \cdots, x_n') - \nabla f(x_1, x_2, \cdots, x_n)\| \leqslant L\|(x_1' - x_1, x_2' - x_2, \cdots, x_n' - x_n)\|$$

for all $(x_1', x_2', \cdots, x_n')$, $(x_1, x_2, \cdots, x_n) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$.

In all the following propositions and theorems, we denote $w^k = (x_1^k, \cdots, x_n^k, \lambda^k)$ to be the iterates generated by ADMM, and $u = (x_1, \cdots, x_n)$.

Proposition 5.3 *Suppose that there are γ, β, and δ satisfying*

$$\frac{n-1}{2}\max_{2\leqslant i\leqslant n}\left\{\lambda_{\max}(A_i^{\mathrm{T}}A_i)\right\}\gamma+\delta\leqslant\min_{2\leqslant i\leqslant n}\sigma_i.$$

Moreover, suppose that the matrices $H_i, i=2,\cdots,n$, satisfy

$$H_i^s:=H_i-\left(L+\frac{(n-i+1)(n+i-2)L^2}{8\delta}\right)I_{p_i}\succeq 0\quad\forall\, 2\leqslant i\leqslant n.$$

Let $(x_1^{k+1},\cdots,x_n^{k+1},\lambda^{k+1})\in\Omega$ be the sequence generated by ADMM. Then, for $u^*=(x_1^*,\cdots,x_n^*)\in\Omega^*$ and $\lambda\in\mathbb{R}^m$, the following inequality holds

$$\begin{aligned}
&h(u^*)-h\left(u^{k+1}\right)+\begin{pmatrix}x_1^*-x_1^{k+1}\\ \vdots\\ x_n^*-x_n^{k+1}\\ \lambda-\lambda^{k+1}\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}-A_1^{\mathrm{T}}\lambda^{k+1}\\ \vdots\\ -A_n^{\mathrm{T}}\lambda^{k+1}\\ \sum\limits_{i=1}^{n}A_ix_i^{k+1}-b\end{pmatrix}\\
&+\frac{\gamma}{2}\sum_{i=2}^{n}\left(\left\|\sum_{j=1}^{i-1}A_jx_j^*+\sum_{j=i}^{n}A_jx_j^k-b\right\|^2-\left\|\sum_{j=1}^{i-1}A_jx_j^*+\sum_{j=i}^{n}A_jx_j^{k+1}-b\right\|^2\right)\\
&+\frac{1}{2\beta}\left(\left\|\lambda-\lambda^k\right\|^2-\left\|\lambda-\lambda^{k+1}\right\|^2\right)+\frac{1}{2}\sum_{i=1}^{n}\left(\|x_i^*-x_i^k\|_{H_i}^2-\|x_i^*-x_i^{k+1}\|_{H_i}^2\right)\\
\geqslant&\left(\frac{\gamma-\beta}{2\beta^2}\right)\|\lambda^k-\lambda^{k+1}\|^2+\frac{1}{2}\sum_{i=1}^{3}\|x_i^{k+1}-x_i^k\|_{H_i^s}^2.
\end{aligned}$$

The following proposition exhibits an important relationship between two consecutive iterates w^k and w^{k+1} from which the convergence readily follows.

Proposition 5.4 *Let w^k be the sequence generated by the ADMM, then*

$$\begin{aligned}
&\frac{\gamma}{2}\sum_{i=2}^{n}\left(\|\mathcal{L}_i(w^*,w^k)\|^2-\|\mathcal{L}_i(w^*,w^{k+1})\|^2\right)+\|w^*-w^k\|_{\hat{\mathcal{M}}}^2\\
&-\|w^*-w^{k+1}\|_{\hat{\mathcal{M}}}^2-\|w^k-w^{k+1}\|_{\mathcal{H}}^2\geqslant 0,
\end{aligned}$$

where $\mathcal{L}_i(w^*,w):=\sum\limits_{j=1}^{i-1}A_jx_j^*+\sum\limits_{j=i}^{n}A_jx_j-b, i=2,\cdots,n$, and

$$\hat{\mathcal{M}}=\mathrm{diag}\left(\frac{1}{2}H_1,\cdots,\frac{1}{2}H_n,\frac{1}{\beta}I_m\right),\mathcal{H}=\mathrm{diag}\left(\frac{1}{2}H_1,\frac{1}{2}H_2^s,\cdots,\frac{1}{2}H_n^s,\frac{\gamma-\beta}{2\beta^2}I_m\right).$$

Propositions 5.3 and 5.4 lead to the following theorem:

Theorem 5.5 *Under the assumptions of Propositions 5.3 and 5.4, and*

$$\mathcal{H} = \operatorname{diag}\left(\frac{1}{2}H_1, \frac{1}{2}H_2^s, \cdots, \frac{1}{2}H_n^s, \frac{\gamma-\beta}{2\beta^2}I_m\right) \succ 0,$$

we conclude that the sequence $\{w^k\}$ generated by the ADMM converges to an optimal solution. Moreover, for any integer $t > 0$, let

$$\bar{w}_t := \frac{1}{t}\sum_{k=0}^{t-1} w^{k+1},$$

and for any $\rho > 0$, we have

$$\begin{aligned} & h(\bar{u}_t) - h(u^*) + \rho\left\|\sum_{i=1}^{n} A_i\bar{x}_t - b\right\| \\ & \leqslant \frac{1}{2t}\left(\gamma\sum_{i=1}^{n-1}\left\|\sum_{j=i+1}^{n} A_j\left(x_j^* - x_j^0\right)\right\|^2 + \sum_{i=1}^{n}\|x_i^* - x_i^0\|_{H_i}^2 + \frac{1}{\beta}\left(\rho + \|\lambda^0\|\right)^2\right). \end{aligned}$$

6 Concluding Remarks

In [36], the following model is considered

$$\min \quad f(x) + g(y) + H(x, y), \tag{6.1}$$

which can be regarded as (1.1) without constraints, and the so-called proximal alternating linearized minimization (PALM) algorithm is proposed. The main focus of [36] is to analyze the convergence of PALM for a class of nonconvex problems based on the Kurdyka–Łojasiewicz property. In that regard, it has an entirely different aim. We note, however, that PALM is similar to APGMM applied to (6.1) when there is no coupling linear constraint. On the linearized gradient part, one noticeable difference is that APGMM operates in a Jacobian fashion, while PALM is Gauss-Seidel. If the computation of gradient is costly, then the Jacobian style is cheaper to implement. As shown in [36], PALM can be extended to allow multiple blocks. Similarly, APGMM is also extendable to solve (5.1). The same is true for the other variations of the ADMM proposed in this paper. It remains a future research topic to establish the convergence rate of such types of first-order algorithms. Other future research topics include the study of first-order algorithms for (1.1) where the objective is nonconvex but satisfies the Kurdyka–Łojasiewicz property. It is also interesting to consider stochastic programming models studied in [31], but now allowing the objective function to be nonseparable.

Appendix: Proofs of the Convergence Theorems

Appendix 1: Proof of Theorem 3.1

We have $F(w) = \begin{pmatrix} 0 & 0 & -A^{\mathrm{T}} \\ 0 & 0 & -B^{\mathrm{T}} \\ A & B & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ \lambda \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ b \end{pmatrix}$, for any w_1 and w_2, and so

$$(w_1 - w_2)^{\mathrm{T}}(F(w_1) - F(w_2)) = 0.$$

Expanding on this identity, we have for any $w^0, w^1, \cdots, w^{t-1}$ and $\bar{w} = \frac{1}{t}\sum\limits_{k=0}^{t-1} w^k$, that

$$(\bar{w} - w)^{\mathrm{T}} F(\bar{w}) = \frac{1}{t}\sum_{k=0}^{t-1}(w^k - w)^{\mathrm{T}} F(w^k). \tag{7.1}$$

We begin our analysis with the following property of the ADMM algorithm.

Proposition 7.1 *Suppose h_2 is strongly convex with parameter $\sigma > 0$. Let $\{\tilde{w}^k\}$ be defined by (2.6), and the matrices Q, M, P be given in (2.5). First of all, for any $w \in \Omega$, we have*

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^{\mathrm{T}} F(\tilde{w}^k) \\ \geqslant\ & (w - \tilde{w}^k)^{\mathrm{T}} Q(w^k - \tilde{w}^k) - \left(\left(\frac{L}{2} + \frac{L^2}{2\sigma} \right) \|y^k - \tilde{y}^k\|^2 \right. \\ & \left. + (y - \tilde{y}^k)^{\mathrm{T}} H(\tilde{y}^k - y^k) \right). \end{aligned} \tag{7.2}$$

Furthermore,

$$\begin{aligned} & (w - \tilde{w}^k)^{\mathrm{T}} Q(w^k - \tilde{w}^k) \\ = & \frac{1}{2}\left(\|w - w^{k+1}\|_M^2 - \|w - w^k\|_M^2 \right) + \frac{1}{2}\|x^k - \tilde{x}^k\|_G^2 + \frac{1}{2\gamma}\|\lambda^k - \tilde{\lambda}^k\|^2. \end{aligned} \tag{7.3}$$

Proof By the optimality condition of the two subproblems in ADMM, we have

$$\begin{aligned} & (x - x^{k+1})^{\mathrm{T}} \Big[\nabla_x f(x^{k+1}, y^k) + h_1'(x^{k+1}) - A^{\mathrm{T}}(\lambda^k - \gamma(Ax^{k+1} + By^k - b)) \\ & + G(x^{k+1} - x^k) \Big] \geqslant 0, \quad \forall x \in \mathcal{X}, \end{aligned}$$

where $h_1'(x^{k+1}) \in \partial h_1(x^{k+1})$, and

$$\begin{aligned} & (y - y^{k+1})^{\mathrm{T}} \Big[\nabla_y f(x^{k+1}, y^{k+1}) + h_2'(y^{k+1}) \\ & -B^{\mathrm{T}}(\lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b)) + H(y^{k+1} - y^k) \Big] \geqslant 0, \quad \forall y \in \mathcal{Y} \end{aligned}$$

where $h_2'(x^{k+1}) \in \partial h_2(x^{k+1})$.

Note that $\tilde{\lambda}^k = \lambda^k - \gamma(Ax^{k+1} + By^k - b)$. The above two inequalities can be rewritten as

$$(x - \tilde{x}^k)^{\mathrm{T}} \left[\nabla_x f(\tilde{x}^k, y^k) + h_1'(\tilde{x}^k) - A^{\mathrm{T}}\tilde{\lambda}^k + G(\tilde{x}^k - x^k)\right] \geqslant 0, \quad \forall x \in \mathcal{X}, \tag{7.4}$$

and

$$\begin{aligned}&(y - \tilde{y}^k)^{\mathrm{T}} \left[\nabla_y f(\tilde{x}^k, \tilde{y}^k) + h_2'(\tilde{y}^k) - B^{\mathrm{T}}\tilde{\lambda}^k + \gamma B^{\mathrm{T}}B(\tilde{y}^k - y^k) + H(\tilde{y}^k - y^k)\right]\\ &\geqslant 0, \quad \forall y \in \mathcal{Y}.\end{aligned} \tag{7.5}$$

Observe the following chain of inequalities

$$\begin{aligned}
&(x - \tilde{x}^k)^{\mathrm{T}}\nabla_x f(\tilde{x}^k, y^k) + (y - \tilde{y}^k)^{\mathrm{T}}\nabla_y f(\tilde{x}^k, \tilde{y}^k)\\
&= (x - \tilde{x}^k)^{\mathrm{T}}\nabla_x f(\tilde{x}^k, y^k) + (y - \tilde{y}^k)^{\mathrm{T}}\nabla_y f(\tilde{x}^k, y^k)\\
&\quad +(y - \tilde{y}^k)^{\mathrm{T}}(\nabla_y f(\tilde{x}^k, \tilde{y}^k) - \nabla_y f(\tilde{x}^k, y^k))\\
&\leqslant (x - \tilde{x}^k)^{\mathrm{T}}\nabla_x f(\tilde{x}^k, y^k) + (y - \tilde{y}^k)^{\mathrm{T}}\nabla_y f(\tilde{x}^k, y^k) + L\|y - \tilde{y}^k\|\|y^k - \tilde{y}^k\|\\
&= (x - \tilde{x}^k)^{\mathrm{T}}\nabla_x f(\tilde{x}^k, y^k) + (y - y^k)^{\mathrm{T}}\nabla_y f(\tilde{x}^k, y^k) + (y^k - \tilde{y}^k)^{\mathrm{T}}\nabla_y f(\tilde{x}^k, y^k)\\
&\quad +L\|y - \tilde{y}^k\|\|y^k - \tilde{y}^k\|\\
&\leqslant f(x, y) - f(\tilde{x}^k, y^k) - (\tilde{y}^k - y^k)^{\mathrm{T}}\nabla_y f(\tilde{x}^k, y^k) + L\|y - \tilde{y}^k\|\|y^k - \tilde{y}^k\|\\
&\quad \text{(from (2.3))}\\
&\leqslant f(x, y) - f(\tilde{x}^k, \tilde{y}^k) + \frac{L}{2}\|y^k - \tilde{y}^k\|^2 + L\|y - \tilde{y}^k\|\|y^k - \tilde{y}^k\|\\
&\leqslant f(x, y) - f(\tilde{x}^k, \tilde{y}^k) + \frac{L}{2}\|y^k - \tilde{y}^k\|^2\\
&\quad +\frac{\sigma}{2}\|y - \tilde{y}^k\|^2 + \frac{L^2}{2\sigma}\|y^k - \tilde{y}^k\|^2.
\end{aligned} \tag{7.6}$$

Since

$$(A\tilde{x}^k + B\tilde{y}^k - b) - B(\tilde{y}^k - y^k) - \frac{1}{\gamma}(\lambda^k - \tilde{\lambda}^k) = 0,$$

we have

$$\begin{aligned}
&(\lambda - \tilde{\lambda}^k)^{\mathrm{T}}\left(A\tilde{x}^k + B\tilde{y}^k - b\right)\\
&= (\lambda - \tilde{\lambda}^k)^{\mathrm{T}}\left(-B(y^k - \tilde{y}^k) + \frac{1}{\gamma}(\lambda^k - \tilde{\lambda}^k)\right).
\end{aligned} \tag{7.7}$$

By the strong convexity of the function $h_2(y)$, we have

$$(y - \tilde{y}^k)^{\mathrm{T}}h_2'(\tilde{y}^k) \leqslant h_2(y) - h_2(\tilde{y}^k) - \frac{\sigma}{2}\|y - \tilde{y}^k\|^2. \tag{7.8}$$

Because of the convexity of $h_1(x)$ and combining (7.8), (7.7), (7.6), (7.5), and (7.4), we have

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + \left(\frac{L}{2} + \frac{L^2}{2\sigma}\right) \|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^{\mathrm{T}} H(\tilde{y}^k - y^k) \\ & + \begin{pmatrix} x - \tilde{x}^k \\ y - \tilde{y}^k \\ \lambda - \tilde{\lambda}^k \end{pmatrix}^{\mathrm{T}} \left[\begin{pmatrix} -A^{\mathrm{T}}\tilde{\lambda}^k \\ -B^{\mathrm{T}}\tilde{\lambda}^k \\ A\tilde{x}^k + B\tilde{y}^k - b \end{pmatrix} - \begin{pmatrix} G(x^k - \tilde{x}^k) \\ \gamma B^{\mathrm{T}} B(y^k - \tilde{y}^k) \\ -B(y^k - \tilde{y}^k) + \frac{1}{\gamma}(\lambda^k - \tilde{\lambda}^k) \end{pmatrix} \right] \geqslant 0 \end{aligned}$$

for any $w \in \Omega$ and $\tilde{w}^k$.

By definition of Q, (7.2) of Proposition 7.1 follows. For (7.3), due to the similarity, we refer to Lemma 3.2 in [17] (noting the matrices Q, P, and M).

The following theorem exhibits an important relationship between two consecutive iterates w^k and w^{k+1} from which the convergence would follow.

Proposition 7.2 *Let w^k be the sequence generated by the ADMM, $\tilde{w}^k$ be defined as in (2.6) and H satisfies $H_s := H - \left(L + \frac{L^2}{\sigma}\right) I_q \succeq 0$. Then the following holds:*

$$\frac{1}{2}\left(\|w^* - w^k\|^2_{\hat{M}} - \|w^* - w^{k+1}\|^2_{\hat{M}}\right) - \frac{1}{2}\|w^k - \tilde{w}^k\|^2_{H_d} \geqslant 0, \tag{7.9}$$

where

$$\hat{H} = \gamma B^{\mathrm{T}} B + H, \;\; \hat{M} = \begin{pmatrix} G & 0 & 0 \\ 0 & \hat{H} & 0 \\ 0 & 0 & \frac{1}{\gamma} I_m \end{pmatrix} \text{ and } H_d = \begin{pmatrix} G & 0 & 0 \\ 0 & H_s & 0 \\ 0 & 0 & \frac{1}{\gamma} I_m \end{pmatrix}. \tag{7.10}$$

Proof It follows from Proposition 7.1 that

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^{\mathrm{T}} F(\tilde{w}^k) \\ \geqslant\; & (w - \tilde{w}^k)^{\mathrm{T}} Q(w^k - \tilde{w}^k) - \left(\left(\frac{L}{2} + \frac{L^2}{2\sigma}\right)\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^{\mathrm{T}} H(\tilde{y}^k - y^k)\right) \\ =\; & \frac{1}{2}\left(\|w - w^{k+1}\|^2_M - \|w - w^k\|^2_M\right) + \frac{1}{2}\|x^k - \tilde{x}^k\|^2_G + \frac{1}{2\gamma}\|\lambda^k - \tilde{\lambda}^k\|^2 \\ & - \left(\left(\frac{L}{2} + \frac{L^2}{2\sigma}\right)\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^{\mathrm{T}} H(\tilde{y}^k - y^k)\right). \end{aligned} \tag{7.11}$$

Note that $H_s := H - (L + \frac{L^2}{\sigma})I_q \succeq 0$, we have the following

$$
\begin{aligned}
&\left(\frac{L}{2} + \frac{L^2}{2\sigma}\right)\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^{\mathrm{T}} H(\tilde{y}^k - y^k) \\
&= \left(\frac{L}{2} + \frac{L^2}{2\sigma}\right)\|y^k - \tilde{y}^k\|^2 + \frac{1}{2}\left(\|y - y^k\|_H^2 - \|y - \tilde{y}^k\|_H^2 - \|y^k - \tilde{y}^k\|_H^2\right) \\
&= \frac{1}{2}\left(\|y - y^k\|_H^2 - \|y - \tilde{y}^k\|_H^2\right) - \frac{1}{2}\|y^k - \tilde{y}^k\|_{H_s}^2. \qquad (7.12)
\end{aligned}
$$

Thus, combining (7.11) and (7.12), we have

$$
\begin{aligned}
&h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^{\mathrm{T}} F(\tilde{w}^k) \\
&\geqslant \frac{1}{2}\left(\|w - w^{k+1}\|_M^2 - \|w - w^k\|_M^2\right) - \frac{1}{2}\left(\|y - y^k\|_H^2 - \|y - \tilde{y}^k\|_H^2\right) \\
&+ \frac{1}{2}\|x^k - \tilde{x}^k\|_G^2 + \frac{1}{2}\|y^k - \tilde{y}^k\|_{H_s}^2 + \frac{1}{2\gamma}\|\lambda^k - \tilde{\lambda}^k\|^2. \qquad (7.13)
\end{aligned}
$$

By the definition of $\hat{M}$ and H_d according to (7.10), it follows from (7.13) that

$$
\begin{aligned}
&h(\tilde{u}^k) - h(u) + (\tilde{w}^k - w)^{\mathrm{T}} F(\tilde{w}^k) \\
&\leqslant \frac{1}{2}\left(\|w - w^k\|_{\hat{M}}^2 - \|w - w^{k+1}\|_{\hat{M}}^2\right) - \frac{1}{2}\|w^k - \tilde{w}^k\|_{H_d}^2. \qquad (7.14)
\end{aligned}
$$

Letting $w = w^*$ in (7.14), we have

$$
\begin{aligned}
&h(\tilde{u}^k) - h(u^*) + (\tilde{w}^k - w^*)^{\mathrm{T}} F(\tilde{w}^k) \\
&\leqslant \frac{1}{2}\left(\|w^* - w^k\|_{\hat{M}}^2 - \|w^* - w^{k+1}\|_{\hat{M}}^2\right) - \frac{1}{2}\|w^k - \tilde{w}^k\|_{H_d}^2. \qquad (7.15)
\end{aligned}
$$

By the monotonicity of F and using the optimality of w^*, we have

$$
\begin{aligned}
&\frac{1}{2}\left(\|w^* - w^k\|_{\hat{M}}^2 - \|w^* - w^{k+1}\|_{\hat{M}}^2\right) - \frac{1}{2}\|w^k - \tilde{w}^k\|_{H_d}^2 \\
&\geqslant h(\tilde{u}^k) - h(u^*) + (\tilde{w}^k - w^*)^{\mathrm{T}} F(\tilde{w}^k) \\
&\geqslant h(\tilde{u}^k) - h(u^*) + (\tilde{w}^k - w^*)^{\mathrm{T}} F(w^*) \\
&\geqslant 0,
\end{aligned}
$$

which completes the proof.

Appendix 2: Proof of Theorem 3.1.

Proof First, according to (7.9), it holds that $\{w^k\}$ is bounded and

$$\lim_{k\to\infty} \|w^k - \tilde{w}^k\|_{H_d} = 0. \tag{7.16}$$

Thus, those two sequences have the same cluster points: For any $w^{k_n} \to w^\infty$, by (7.16) we also have $\tilde{w}^{k_n} \to w^\infty$. Applying inequality (7.2) to $\{w^{k_n}\}$, $\{\tilde{w}^{k_n}\}$ and taking the limit, it yields that

$$h(u) - h(u^\infty) + (w - w^\infty)^{\mathrm{T}} F(w^\infty) \geqslant 0. \tag{7.17}$$

Consequently, the cluster point w^∞ is an optimal solution. Since (7.9) is true for any optimal solution w^*, it also holds for w^∞, and that implies w^k will converge to w^∞.

Recall (7.2) and (7.3) in Proposition 7.1, those would imply that

$$\begin{aligned}
&h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^{\mathrm{T}} F(\tilde{w}^k)\\
&\geqslant (w - \tilde{w}^k)^{\mathrm{T}} Q(w^k - \tilde{w}^k) - \left(\left(\frac{L}{2} + \frac{L^2}{2\sigma}\right)\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^{\mathrm{T}} H(\tilde{y}^k - y^k)\right)\\
&\geqslant \frac{1}{2}\left(\|w - w^{k+1}\|_M^2 - \|w - w^k\|_M^2\right) - \left(\left(\frac{L}{2} + \frac{L^2}{2\sigma}\right)\|y^k - \tilde{y}^k\|^2\right.\\
&\quad \left. + \left(y - \tilde{y}^k\right)^{\mathrm{T}} H(\tilde{y}^k - y^k)\right).
\end{aligned} \tag{7.18}$$

Furthermore, since $H - \left(L + \frac{L^2}{\sigma}\right) I_q \succeq 0$, we have

$$\begin{aligned}
&\left(\frac{L}{2} + \frac{L^2}{2\sigma}\right)\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^{\mathrm{T}} H(\tilde{y}^k - y^k)\\
&= \left(\frac{L}{2} + \frac{L^2}{2\sigma}\right)\|y^k - \tilde{y}^k\|^2 + \frac{1}{2}\left(\|y - y^k\|_H^2 - \|y - \tilde{y}^k\|_H^2 - \|y^k - \tilde{y}^k\|_H^2\right)\\
&\leqslant \frac{1}{2}\left(\|y - y^k\|_H^2 - \|y - \tilde{y}^k\|_H^2\right).
\end{aligned} \tag{7.19}$$

Thus, combining (7.18) and (7.19) leads to

$$\begin{aligned}
&h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^{\mathrm{T}} F(\tilde{w}^k)\\
&\geqslant \frac{1}{2}\left(\|w - w^{k+1}\|_M^2 - \|w - w^k\|_M^2\right)\\
&\quad - \frac{1}{2}\left(\|y - y^k\|_H^2 - \|y - \tilde{y}^k\|_H^2\right).
\end{aligned} \tag{7.20}$$

By the definition of M in (2.5) and denoting $\hat{H} = \gamma B^\top B + H$, (7.20) leads to

$$\begin{aligned} & h(\tilde{u}^k) - h(u) + (\tilde{w}^k - w)^{\mathrm{T}} F(\tilde{w}^k) \\ \leqslant & \frac{1}{2}\left(\|x - x^k\|_G^2 - \|x - x^{k+1}\|_G^2\right) + \frac{1}{2}\left(\|y - y^k\|_{\hat{H}}^2 - \|y - y^{k+1}\|_{\hat{H}}^2\right) \\ & + \frac{1}{2\gamma}\left(\|\lambda - \lambda^k\|^2 - \|\lambda - \lambda^{k+1}\|^2\right). \end{aligned} \tag{7.21}$$

Before proceeding, let us introduce $\bar{w}_n := \frac{1}{n}\sum_{k=0}^{n-1} \tilde{w}^k$. Moreover, recall the definition of $\bar{u}_n$ in (3.2), we have

$$\bar{u}_n = \frac{1}{n}\sum_{k=1}^{n} u^k = \frac{1}{n}\sum_{k=0}^{n-1} \tilde{u}^k.$$

Now, summing the inequality (7.21) over $k = 0, 1, \cdots, t-1$ yields

$$\begin{aligned} & h(\bar{u}_t) - h(u) + (\bar{w}_t - w)^{\mathrm{T}} F(\bar{w}_t) \\ \leqslant & \frac{1}{t}\sum_{k=0}^{t-1} h(\tilde{u}^k) - h(u) + \frac{1}{t}\sum_{k=0}^{t-1} (\tilde{w}^k - w)^{\mathrm{T}} F(\tilde{w}^k) \\ \leqslant & \frac{1}{2t}\left(\|x - x^0\|_G^2 + \|y - y^0\|_{\hat{H}}^2 + \frac{1}{\gamma}\|\lambda - \lambda^0\|^2\right), \end{aligned} \tag{7.22}$$

where the first inequality is due to the convexity of h and (7.1).

Note the above inequality is true for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $\lambda \in \mathbb{R}^m$, hence it is also true for any optimal solution x^*, y^*, and $\mathcal{B}_\rho = \{\lambda : \|\lambda\| \leqslant \rho\}$. As a result,

$$\begin{aligned} & \sup_{\lambda \in \mathcal{B}_\rho} \left\{h(\bar{u}_t) - h(u^*) + (\bar{w}_t - w^*)^{\mathrm{T}} F(\bar{w}_t)\right\} \\ = & \sup_{\lambda \in \mathcal{B}_\rho} \Big\{h(\bar{u}_t) - h(u^*) + (\bar{x}_t - x^*)^{\mathrm{T}}(-A^{\mathrm{T}}\bar{\lambda}_t) \\ & + (\bar{y}_t - y^*)^{\mathrm{T}}(-B^{\mathrm{T}}\bar{\lambda}_t) + (\bar{\lambda}_t - \lambda)^{\mathrm{T}}(A\bar{x}_t + B\bar{y}_t - b)\Big\} \\ = & \sup_{\lambda \in \mathcal{B}_\rho} \left\{h(\bar{u}_t) - h(u^*) + \bar{\lambda}_t^{\mathrm{T}}(Ax^* + By^* - b) - \lambda^{\mathrm{T}}(A\bar{x}_t + B\bar{y}_t - b)\right\} \\ = & \sup_{\lambda \in \mathcal{B}_\rho} \left\{h(\bar{u}_t) - h(u^*) - \lambda^{\mathrm{T}}(A\bar{x}_t + B\bar{y}_t - b)\right\} \\ = & h(\bar{u}_t) - h(u^*) + \rho\|A\bar{x}_t + B\bar{y}_t - b\|, \end{aligned} \tag{7.23}$$

which, combined with (7.22), implies that

$$\begin{aligned}&h(\bar{u}_t) - h(u^*) + \rho\|A\bar{x}_t + B\bar{y}_t - b\|\\ &\leqslant \frac{1}{2t}\left(\|x^* - x^0\|_G^2 + \|y^* - y^0\|_{\hat{H}}^2 + \frac{1}{\gamma}\sup_{\lambda\in\mathcal{B}_\rho}\|\lambda - \lambda^0\|^2\right),\end{aligned}$$

and so by optimizing over $(x^*, y^*) \in \mathcal{X}^* \times \mathcal{Y}^*$, we have

$$\begin{aligned}&h(\bar{u}_t) - h(u^*) + \rho\|A\bar{x}_t + B\bar{y}_t - b\|\\ &\leqslant \frac{1}{2t}\left(\operatorname{dist}(x^0, \mathcal{X}^*)_G^2 + \operatorname{dist}(y^0, \mathcal{Y}^*)_{\hat{H}}^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right).\end{aligned} \tag{7.24}$$

This completes the proof.

Appendix 3: Proof of Theorem 4.1

Similar to the analysis for ADMM, we need the following proposition in the analysis of APGMM.

Proposition 7.3 *Let $\{\tilde{w}^k\}$ be defined by (2.6), and the matrices Q, M, P be given as in (2.5). For any $w \in \Omega$, we have*

$$\begin{aligned}&h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^{\mathrm{T}}F(\tilde{w}^k)\\ &\geqslant (w - \tilde{w}^k)^{\mathrm{T}}Q(w^k - \tilde{w}^k) - \left(\frac{L}{2}(\|x^k - \tilde{x}^k\|^2 + \|y^k - \tilde{y}^k\|^2) + (y - \tilde{y}^k)^{\mathrm{T}}H(\tilde{y}^k - y^k)\right) &(7.25)\\ &= \frac{1}{2}\left(\|w - w^{k+1}\|_M^2 - \|w - w^k\|_M^2\right) + \frac{1}{2}\|x^k - \tilde{x}^k\|_G^2 + \frac{1}{2\gamma}\|\lambda^k - \tilde{\lambda}^k\|^2\\ &\quad - \left(\frac{L}{2}\left(\|x^k - \tilde{x}^k\|^2 + \|y^k - \tilde{y}^k\|^2\right) + (y - \tilde{y}^k)^{\mathrm{T}}H(\tilde{y}^k - y^k)\right). &(7.26)\end{aligned}$$

Proof First, by the optimality condition of the two subproblems in APGMM, we have

$$\begin{aligned}&(x - x^{k+1})^{\mathrm{T}}\Big[\nabla_x f(x^k, y^k) + h_1'(x^{k+1}) - A^{\mathrm{T}}(\lambda^k - \gamma(Ax^{k+1} + By^k - b))\\ &\quad + G(x^{k+1} - x^k)\Big] \geqslant 0, \quad \forall x \in \mathcal{X}\end{aligned}$$

and

$$\begin{aligned}&(y - y^{k+1})^{\mathrm{T}}\Big[\nabla_y f(x^k, y^k) + h_2'(y^{k+1}) - B^{\mathrm{T}}(\lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b))\\ &\quad + H(y^{k+1} - y^k)\Big] \geqslant 0, \quad \forall y \in \mathcal{Y}.\end{aligned}$$

Note that $\tilde{\lambda}^k = \lambda^k - \gamma(Ax^{k+1} + By^k - b)$, and by the definition of $\tilde{w}^k$, the above two inequalities are equivalent to

$$(x - \tilde{x}^k)^{\mathrm{T}}\left[\nabla_x f(x^k, y^k) + h_1'(\tilde{x}^k) - A^{\mathrm{T}}\tilde{\lambda}^k + G(\tilde{x}^k - x^k)\right] \geqslant 0, \quad \forall x \in \mathcal{X} \tag{7.27}$$

and

$$\begin{aligned}(y - \tilde{y}^k)^{\mathrm{T}}\Big[&\nabla_y f(x^k, y^k) + h_2'(\tilde{y}^k) - B^{\mathrm{T}}\tilde{\lambda}^k + \gamma B^{\mathrm{T}}B(\tilde{y}^k - y^k)\\ &+ H(\tilde{y}^k - y^k)\Big] \geqslant 0, \quad \forall y \in \mathcal{Y}.\end{aligned} \tag{7.28}$$

Notice that

$$\begin{aligned}&(x - \tilde{x}^k)^{\mathrm{T}}\nabla_x f(x^k, y^k) + (y - \tilde{y}^k)^{\mathrm{T}}\nabla_y f(x^k, y^k)\\ &= (x - x^k)^{\mathrm{T}}\nabla_x f(x^k, y^k) + (y - y^k)^{\mathrm{T}}\nabla_y f(x^k, y^k) + (x^k - \tilde{x}^k)^{\mathrm{T}}\nabla_x f(x^k, y^k)\\ &\quad +(y^k - \tilde{y}^k)^{\mathrm{T}}\nabla_y f(x^k, y^k)\\ &\leqslant f(x, y) - f(x^k, y^k) - (\tilde{x}^k - x^k)^{\mathrm{T}}\nabla_x f(x^k, y^k) - (\tilde{y}^k - y^k)^{\mathrm{T}}\nabla_y f(x^k, y^k)\\ &\quad \text{(from (2.3))}\\ &\leqslant f(x, y) - f(\tilde{x}^k, \tilde{y}^k) + \frac{L}{2}\left(\|x^k - \tilde{x}^k\|^2 + \|y^k - \tilde{y}^k\|^2\right).\end{aligned} \tag{7.29}$$

Besides, we also have

$$(A\tilde{x}^k + B\tilde{y}^k - b) - B(\tilde{y}^k - y^k) - \frac{1}{\gamma}\left(\lambda^k - \tilde{\lambda}^k\right) = 0.$$

Thus

$$(\lambda - \tilde{\lambda}^k)^{\mathrm{T}}(A\tilde{x}^k + B\tilde{y}^k - b) = (\lambda - \tilde{\lambda}^k)^{\mathrm{T}}\left(-B(y^k - \tilde{y}^k) + \frac{1}{\gamma}(\lambda^k - \tilde{\lambda}^k)\right). \tag{7.30}$$

By the convexity of $h_1(x)$ and $h_2(y)$, combining (7.30), (7.29), (7.28), and (7.27), we have

$$\begin{aligned}&h(u) - h(\tilde{u}^k) + \frac{L}{2}\left(\|x^k - \tilde{x}^k\|^2 + \|y^k - \tilde{y}^k\|^2\right) + (y - \tilde{y}^k)^{\mathrm{T}}H(\tilde{y}^k - y^k)\\ &+ \begin{pmatrix} x - \tilde{x}^k \\ y - \tilde{y}^k \\ \lambda - \tilde{\lambda}^k \end{pmatrix}^{\mathrm{T}}\left[\begin{pmatrix} -A^{\mathrm{T}}\tilde{\lambda}^k \\ -B^{\mathrm{T}}\tilde{\lambda}^k \\ A\tilde{x}^k + B\tilde{y}^k - b \end{pmatrix} - \begin{pmatrix} G(x^k - \tilde{x}^k) \\ \gamma B^{\mathrm{T}}B(y^k - \tilde{y}^k) \\ -B(y^k - \tilde{y}^k) + \frac{1}{\gamma}(\lambda^k - \tilde{\lambda}^k) \end{pmatrix}\right] \geqslant 0\end{aligned}$$

for any $w \in \Omega$ and $\tilde{w}^k$.

By definition of Q, we have shown (7.25) in Proposition 7.3. Equality (7.26) directly follows from (7.3) in Proposition 7.1.

With Proposition 7.3 in place, we can show Theorem 4.1 by exactly following the same steps as in the proof of Theorem 3.1, noting of course the altered assumptions on the matrices G and H. In the meanwhile, we also point out the following proposition which is similar to Proposition 7.2. Since most steps of the proofs are almost identical to that of the previous theorems, we omit the details for succinctness.

Proposition 7.4 *Let w^k be the sequence generated by the APGMM, $\tilde{w}^k$ be as defined in (2.6), and H and G are chosen so as to satisfy $H_s := H - LI_q \succ 0$ and $G_s := G - LI_p \succ 0$. Then the following holds*:

$$\frac{1}{2}\left(\|w^* - w^k\|^2_{\hat{M}} - \|w^* - w^{k+1}\|^2_{\hat{M}}\right) - \frac{1}{2}\|w^k - \tilde{w}^k\|^2_{H_d} \geqslant 0,$$

where

$$\hat{M} = \begin{pmatrix} G & 0 & 0 \\ 0 & \hat{H} & 0 \\ 0 & 0 & \frac{1}{\gamma} I_m \end{pmatrix}, \quad H_d = \begin{pmatrix} G_s & 0 & 0 \\ 0 & H_s & 0 \\ 0 & 0 & \frac{1}{\gamma} I_m \end{pmatrix}$$

and $\hat{H} = \gamma B^{\mathrm{T}} B + H$.

Theorem 4.1 follows from the above propositions.

Appendix 4: Proof of Theorem 4.2

Similar to the analysis for APGMM, we do not need any strong convexity here, but we do need to assume that the gradients $\nabla_x h_1(x)$ and $\nabla_y h_2(y)$ are Lipschitz continuous. Without loss of generality, we further assume that the Lipschitz constant is the same as $\nabla f(x, y)$ which is L, that is,

$$\begin{aligned} &\|\nabla_x h_1(x_2) - \nabla_x h_1(x_1)\| \leqslant L\|x_2 - x_1\|, \ \forall x_1, x_2 \in \mathcal{X}, \\ &\|\nabla_y h_2(y_2) - \nabla_y h_2(y_1)\| \leqslant L\|y_2 - y_1\|, \ \forall y_1, y_2 \in \mathcal{Y}. \end{aligned} \tag{7.31}$$

Proposition 7.5 *Let $\{\tilde{w}^k\}$ be defined by (2.6), and the matrices Q, M, P be as given in (2.5), and $G := \gamma A^{\mathrm{T}} A + \frac{1}{\alpha} I_p$, $H := \frac{1}{\alpha} I_q - \gamma B^{\mathrm{T}} B \succeq 0$. First of all, for any $w \in \Omega$, we have*

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^{\mathrm{T}} F(\tilde{w}^k) \\ \geqslant\ & (w - \tilde{w}^k)^{\mathrm{T}} Q(w^k - \tilde{w}^k) - \Big(L(\|x^k - \tilde{x}^k\|^2 + \|y^k - \tilde{y}^k\|^2) \\ & + (y - \tilde{y}^k)^{\mathrm{T}} H(\tilde{y}^k - y^k) \Big) \qquad (7.32) \\ =\ & \frac{1}{2}\left(\|w - w^{k+1}\|^2_M - \|w - w^k\|^2_M\right) + \frac{1}{2}\|x^k - \tilde{x}^k\|^2_G + \frac{1}{2\gamma}\|\lambda^k - \tilde{\lambda}^k\|^2 \\ & - \Big(L(\|x^k - \tilde{x}^k\|^2 + \|y^k - \tilde{y}^k\|^2) + (y - \tilde{y}^k)^{\mathrm{T}} H(\tilde{y}^k - y^k) \Big). \qquad (7.33) \end{aligned}$$

Proof First, by the optimality condition of the two subproblems in AGPMM, we have

$$(x - x^{k+1})^{\mathrm{T}} \Big[x^{k+1} - x^k + \alpha \Big(\nabla_x f(x^k, y^k) + \nabla_y h_1(x^k) \\ - A^{\mathrm{T}} (\lambda^k - \gamma (A x^k + B y^k - b)) \Big) \Big] \geqslant 0, \quad \forall x \in \mathcal{X}$$

and

$$(y - y^{k+1})^{\mathrm{T}} \Big[y^{k+1} - y^k + \alpha \Big(\nabla_y f(x^k, y^k) + \nabla_y h_2(y^k) \\ - B^{\mathrm{T}} (\lambda^k - \gamma (A x^{k+1} + B y^k - b)) \Big) \Big] \geqslant 0, \quad \forall y \in \mathcal{Y}.$$

Noting $\tilde{\lambda}^k = \lambda^k - \gamma(Ax^{k+1} + By^k - b)$ and the definition of $\tilde{w}^k$, the above two inequalities are, respectively, equivalent to

$$(x - \tilde{x}^k)^{\mathrm{T}} \Big[\nabla_x f(x^k, y^k) + \nabla_x h_2(x^k) - A^{\mathrm{T}} \tilde{\lambda}^k + \gamma A^{\mathrm{T}} A (\tilde{x}^k - x^k) \\ + \frac{1}{\alpha} (\tilde{x}^k - x^k) \Big] \geqslant 0, \quad \forall x \in \mathcal{X} \tag{7.34}$$

and

$$(y - \tilde{y}^k)^{\mathrm{T}} \Big[\nabla_y f(x^k, y^k) + \nabla_y h_2(y^k) - B^{\mathrm{T}} \tilde{\lambda}^k + \frac{1}{\alpha} (\tilde{y}^k - y^k) \Big] \\ \geqslant 0, \quad \forall y \in \mathcal{Y}. \tag{7.35}$$

Similar to Proposition 7.3, we have

$$(x - \tilde{x}^k)^{\mathrm{T}} \nabla_x f(x^k, y^k) + (y - \tilde{y}^k)^{\mathrm{T}} \nabla_y f(x^k, y^k) \\ \text{(from (2.3))} \\ \leqslant f(x, y) - f(\tilde{x}^k, \tilde{y}^k) + \frac{L}{2} \Big(\|x^k - \tilde{x}^k\|^2 + \|y^k - \tilde{y}^k\|^2 \Big). \tag{7.36}$$

Moreover, by (2.4), we have

$$(x - \tilde{x}^k)^{\mathrm{T}} \nabla_x h_1(x^k) \leqslant h_1(x) - h_1(\tilde{x}^k) + \frac{L}{2} \|x^k - \tilde{x}^k\|^2, \\ (y - \tilde{y}^k)^{\mathrm{T}} \nabla_y h_2(y^k) \leqslant h_2(y) - h_2(\tilde{y}^k) + \frac{L}{2} \|y^k - \tilde{y}^k\|^2. \tag{7.37}$$

Besides,

$$(A\tilde{x}^k + B\tilde{y}^k - b) - B(\tilde{y}^k - y^k) - \frac{1}{\gamma} \Big(\lambda^k - \tilde{\lambda}^k \Big) = 0.$$

Thus

$$\begin{aligned}&(\lambda-\tilde{\lambda}^k)^{\mathrm{T}}(A\tilde{x}^k+B\tilde{y}^k-b)\\ &=(\lambda-\tilde{\lambda}^k)^{\mathrm{T}}\left(-B(y^k-\tilde{y}^k)+\frac{1}{\gamma}\left(\lambda^k-\tilde{\lambda}^k\right)\right).\end{aligned}\tag{7.38}$$

Combining (7.38), (7.37), (7.36), (7.35), and (7.34), and noticing that $G:=\gamma A^{\mathrm{T}}A+\frac{1}{\alpha}I_p$, $H:=\frac{1}{\alpha}I_q-\gamma B^{\mathrm{T}}B$, we have, for any $w\in\Omega$ and $\tilde{w}^k$, that

$$\begin{aligned}&h(u)-h(\tilde{u}^k)+L(\|x^k-\tilde{x}^k\|^2+\|y^k-\tilde{y}^k\|^2)+(y-\tilde{y}^k)^{\mathrm{T}}H(\tilde{y}^k-y^k)\\ &+\begin{pmatrix}x-\tilde{x}^k\\ y-\tilde{y}^k\\ \lambda-\tilde{\lambda}^k\end{pmatrix}^{\mathrm{T}}\left\{\begin{pmatrix}-A^{\mathrm{T}}\tilde{\lambda}^k\\ -B^{\mathrm{T}}\tilde{\lambda}^k\\ A\tilde{x}^k+B\tilde{y}^k-b\end{pmatrix}-\begin{pmatrix}G(x^k-\tilde{x}^k)\\ \gamma B^{\mathrm{T}}B(y^k-\tilde{y}^k)\\ -B(y^k-\tilde{y}^k)+\frac{1}{\gamma}(\lambda^k-\tilde{\lambda}^k)\end{pmatrix}\right\}\geqslant 0.\end{aligned}$$

Using the definition of Q, (7.32) follows. In view of (7.3) in Proposition 7.1, equality (7.33) also readily follows.

With Proposition 7.5, similar as before, we can show Theorem 4.2 by following the same approach as in the proof of Theorem 3.1. We skip the details here for succinctness.

Proposition 7.6 *Let w^k be the sequence generated by the AGPMM, $\tilde{w}^k$ be defined in (2.6), and $G:=\gamma A^{\mathrm{T}}A+\frac{1}{\alpha}I_p$, $H:=\frac{1}{\alpha}I_q-\gamma B^{\mathrm{T}}B$. Suppose that α satisfies that $H_s:=H-2LI_q\succ 0$ and $G_s:=G-2LI_p\succ 0$. Then the following holds*

$$\frac{1}{2}\left(\|w^*-w^k\|_{\hat{M}}^2-\|w^*-w^{k+1}\|_{\hat{M}}^2\right)-\frac{1}{2}\|w^k-\tilde{w}^k\|_{H_d}^2\geqslant 0,$$

where

$$\hat{M}=\begin{pmatrix}G&0&0\\ 0&\hat{H}&0\\ 0&0&\frac{1}{\gamma}I_m\end{pmatrix},\quad H_d=\begin{pmatrix}G_s&0&0\\ 0&H_s&0\\ 0&0&\frac{1}{\gamma}I_m\end{pmatrix},$$

and $\hat{H}=\gamma B^{\mathrm{T}}B+H$.

Theorem 4.2 now follows from the above propositions.

Appendix 5: Proofs of Theorems 4.3 and 4.4

Proposition 6.1 *Let $\{\tilde{w}^k\}$ be defined by (2.6), and the matrices Q, M, P be given in (2.5). For any $w \in \Omega$, we have*

$$
\begin{aligned}
&h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^{\mathrm{T}} F(\tilde{w}^k) \\
&\geqslant (w - \tilde{w}^k)^{\mathrm{T}} Q(w^k - \tilde{w}^k) - \left(\frac{L}{2} \|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^{\mathrm{T}} H(\tilde{y}^k - y^k) \right) \\
&= \frac{1}{2} \left(\|w - w^{k+1}\|_M^2 - \|w - w^k\|_M^2 \right) + \frac{1}{2} \|x^k - \tilde{x}^k\|_G^2 + \frac{1}{2\gamma} \|\lambda^k - \tilde{\lambda}^k\|^2 \\
&\quad - \left(\frac{L}{2} \|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^{\mathrm{T}} H(\tilde{y}^k - y^k) \right). \qquad (7.39)
\end{aligned}
$$

Proof First, by the optimality condition of the two subproblems in ADM-PG, we have

$$
\begin{aligned}
&(x - x^{k+1})^{\mathrm{T}} \Big[\nabla_x f(x^{k+1}, y^k) + h_1'(x^{k+1}) - A^{\mathrm{T}}(\lambda^k - \gamma(Ax^{k+1} + By^k - b)) \\
&\quad + G(x^{k+1} - x^k) \Big] \geqslant 0, \quad \forall x \in \mathcal{X}
\end{aligned}
$$

and

$$
\begin{aligned}
&(y - y^{k+1})^{\mathrm{T}} \Big[\nabla_y f(x^{k+1}, y^k) + h_2'(y^{k+1}) - B^{\mathrm{T}}(\lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b)) \\
&\quad + H(y^{k+1} - y^k) \Big] \geqslant 0, \quad \forall y \in \mathcal{Y}.
\end{aligned}
$$

Noting $\tilde{\lambda}^k = \lambda^k - \gamma(Ax^{k+1} + By^k - b)$ and the definition of $\tilde{w}^k$, the above two inequalities are equivalent to

$$
\begin{aligned}
&(x - \tilde{x}^k)^{\mathrm{T}} \Big[\nabla_x f(\tilde{x}^k, y^k) + \nabla_x h_1(\tilde{x}^k) - A^{\mathrm{T}} \tilde{\lambda}^k + G(\tilde{x}^k - x^k) \Big] \\
&\geqslant 0, \quad \forall x \in \mathcal{X} \qquad (7.40)
\end{aligned}
$$

and

$$
\begin{aligned}
&(y - \tilde{y}^k)^{\mathrm{T}} \Big[\nabla_y f(\tilde{x}^k, y^k) + g_2(\tilde{y}^k) - B^{\mathrm{T}} \tilde{\lambda}^k + \gamma B^{\mathrm{T}} B(\tilde{y}^k - y^k) \\
&\quad + H(\tilde{y}^k - y^k) \Big] \geqslant 0, \quad \forall y \in \mathcal{Y}. \qquad (7.41)
\end{aligned}
$$

Moreover,

$$\begin{aligned}
&(x-\tilde{x}^k)^{\mathrm{T}}\nabla_x f(\tilde{x}^k,y^k)+(y-\tilde{y}^k)^{\mathrm{T}}\nabla_y f(\tilde{x}^k,y^k)\\
&=(x-\tilde{x}^k)^{\mathrm{T}}\nabla_x f(\tilde{x}^k,y^k)+(y-y^k)^{\mathrm{T}}\nabla_y f(\tilde{x}^k,y^k)+(y^k-\tilde{y}^k)^{\mathrm{T}}\nabla_y f(\tilde{x}^k,y^k)\\
&\leqslant f(x,y)-f(\tilde{x}^k,y^k)-(\tilde{y}^k-y^k)^{\mathrm{T}}\nabla_y f(\tilde{x}^k,y^k)\\
&\quad \text{(from (2.3))}\\
&\leqslant f(x,y)-f(\tilde{x}^k,\tilde{y}^k)+\frac{L}{2}\|y^k-\tilde{y}^k\|^2. \qquad (7.42)
\end{aligned}$$

Besides,

$$(A\tilde{x}^k+B\tilde{y}^k-b)-B(\tilde{y}^k-y^k)-\frac{1}{\gamma}\left(\lambda^k-\tilde{\lambda}^k\right)=0,$$

and so

$$\begin{aligned}
&(\lambda-\tilde{\lambda}^k)^{\mathrm{T}}(A\tilde{x}^k+B\tilde{y}^k-b)\\
&=(\lambda-\tilde{\lambda}^k)^{\mathrm{T}}\left(-B(y^k-\tilde{y}^k)+\frac{1}{\gamma}(\lambda^k-\tilde{\lambda}^k)\right). \qquad (7.43)
\end{aligned}$$

By the convexity of $h_1(x)$ and $h_2(y)$, combining (7.43), (7.42), (7.41), and (7.40), we have

$$\begin{aligned}
&h(u)-h(\tilde{u}^k)+\frac{L}{2}\|y^k-\tilde{y}^k\|^2+(y-\tilde{y}^k)^{\mathrm{T}}H(\tilde{y}^k-y^k)\\
&+\begin{pmatrix} x-\tilde{x}^k\\ y-\tilde{y}^k\\ \lambda-\tilde{\lambda}^k\end{pmatrix}^{\mathrm{T}}\left[\begin{pmatrix} -A^{\mathrm{T}}\tilde{\lambda}^k\\ -B^{\mathrm{T}}\tilde{\lambda}^k\\ A\tilde{x}^k+B\tilde{y}^k-b\end{pmatrix}-\begin{pmatrix} G(x^k-\tilde{x}^k)\\ \gamma B^{\mathrm{T}}B(y^k-\tilde{y}^k)\\ -B(y^k-\tilde{y}^k)+\frac{1}{\gamma}(\lambda^k-\tilde{\lambda}^k)\end{pmatrix}\right]\geqslant 0
\end{aligned}$$

for any $w\in\Omega$ and $\tilde{w}^k$.

By similar derivations as in the proofs for Proposition 7.5, (7.39) follows.

With Proposition 6.1 in place, we can prove Theorem 4.3 similarly as in the proof of Theorem 3.1. We skip the details here for succinctness.

For ADM-GP, we do not need strong convexity, but we do need to assume that the gradient $\nabla_y h_2(y)$ of $h_2(y)$ is Lipschitz continuous. Without loss of generality, we further assume that the Lipschitz constant of $\nabla_y h_2(y)$ is the same as $\nabla f(x,y)$ which is L:

$$\|\nabla_y h_2(y_2)-\nabla_y h_2(y_1)\|\leqslant L\|y_2-y_1\|,\ \forall\, y_1,y_2\in\mathcal{Y}. \qquad (7.44)$$

Proposition 6.2 *Let $\{\tilde{w}^k\}$ be defined by (2.6), and the matrices Q, M, P be given in (2.5), and $H := \frac{1}{\alpha} I_q - \gamma B^{\mathrm{T}} B \succeq 0$. For any $w \in \Omega$, we have*

$$\begin{aligned}
&h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^{\mathrm{T}} F(\tilde{w}^k) \\
&\geqslant (w - \tilde{w}^k)^{\mathrm{T}} Q(w^k - \tilde{w}^k) - \left(L\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^{\mathrm{T}} H(\tilde{y}^k - y^k)\right) \\
&= \frac{1}{2}\left(\|w - w^{k+1}\|_M^2 - \|w - w^k\|_M^2\right) + \frac{1}{2}\|x^k - \tilde{x}^k\|_G^2 + \frac{1}{2\gamma}\|\lambda^k - \tilde{\lambda}^k\|^2 \\
&\quad - \left(L\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^{\mathrm{T}} H(\tilde{y}^k - y^k)\right). \qquad (7.45)
\end{aligned}$$

Proof By the optimality condition of the two subproblems in ADMM, we have

$$\begin{aligned}
&(x - x^{k+1})^{\mathrm{T}}\Big[\nabla_x f(x^{k+1}, y^k) + h_1'(x^{k+1}) - A^{\mathrm{T}}(\lambda^k - \gamma(Ax^{k+1} + By^k - b)) \\
&\quad + G(x^{k+1} - x^k)\Big] \geqslant 0, \qquad \forall x \in \mathcal{X}
\end{aligned}$$

and

$$\begin{aligned}
&(y - y^{k+1})^{\mathrm{T}}\Big[y^{k+1} - y^k + \alpha\Big(\nabla_y f(x^{k+1}, y^k) + \nabla_y h_2(y^k) - B^{\mathrm{T}}(\lambda^k - \gamma(Ax^{k+1} \\
&\quad + By^k - b))\Big)\Big] \geqslant 0, \qquad \forall y \in \mathcal{Y}.
\end{aligned}$$

Noting $\tilde{\lambda}^k = \lambda^k - \gamma(Ax^{k+1} + By^k - b)$ and the definition of $\tilde{w}^k$, the above two inequalities are equivalent to

$$(x - \tilde{x}^k)^{\mathrm{T}}\Big[\nabla_x f(\tilde{x}^k, y^k) + h_1'(\tilde{x}^k) - A^{\mathrm{T}}\tilde{\lambda}^k + G(\tilde{x}^k - x^k)\Big] \geqslant 0, \quad \forall x \in \mathcal{X} \quad (7.46)$$

and

$$\begin{aligned}
&(y - \tilde{y}^k)^{\mathrm{T}}\left[\nabla_y f(\tilde{x}^k, y^k) + \nabla_y h_2(y^k) - B^{\mathrm{T}}\tilde{\lambda}^k + \frac{1}{\alpha}(\tilde{y}^k - y^k)\right] \\
&\geqslant 0, \quad \forall y \in \mathcal{Y}. \qquad (7.47)
\end{aligned}$$

Therefore,

$$\begin{aligned}
&(x - \tilde{x}^k)^{\mathrm{T}}\nabla_x f(\tilde{x}^k, y^k) + (y - \tilde{y}^k)^{\mathrm{T}}\nabla_y f(\tilde{x}^k, y^k) \\
&= (x - \tilde{x}^k)^{\mathrm{T}}\nabla_x f(\tilde{x}^k, y^k) + (y - y^k)^{\mathrm{T}}\nabla_y f(\tilde{x}^k, y^k) + (y^k - \tilde{y}^k)^{\mathrm{T}}\nabla_y f(\tilde{x}^k, y^k) \\
&\leqslant f(x, y) - f(\tilde{x}^k, y^k) - (\tilde{y}^k - y^k)^{\mathrm{T}}\nabla_y f(\tilde{x}^k, y^k) \\
&\leqslant f(x, y) - f(\tilde{x}^k, \tilde{y}^k) + \frac{L}{2}\|y^k - \tilde{y}^k\|^2. \qquad (7.48)
\end{aligned}$$

Moreover, by (2.4), we have

$$(y - \tilde{y}^k)^{\mathrm{T}}\nabla_y h_2(y^k) \leqslant h_2(y) - h_2(\tilde{y}^k) + \frac{L}{2}\|y^k - \tilde{y}^k\|^2. \qquad (7.49)$$

Since

$$A\tilde{x}^k + B\tilde{y}^k - b - B(\tilde{y}^k - y^k) - \frac{1}{\gamma}\left(\lambda^k - \tilde{\lambda}^k\right) = 0,$$

we have

$$\begin{aligned}&(\lambda - \tilde{\lambda}^k)^{\mathrm{T}}(A\tilde{x}^k + B\tilde{y}^k - b)\\ &= (\lambda - \tilde{\lambda}^k)^{\mathrm{T}}\left(-B(y^k - \tilde{y}^k) + \frac{1}{\gamma}(\lambda^k - \tilde{\lambda}^k)\right).\end{aligned} \tag{7.50}$$

By the convexity of $h_1(x)$, combining (7.50), (7.49), (7.48), (7.47), (7.46), and noticing $H := \frac{1}{\alpha}I_q - \gamma B^{\mathrm{T}}B$ for any $w \in \Omega$ and $\tilde{w}^k$, we have

$$\begin{aligned}&h(u) - h(\tilde{u}^k) + L\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^{\mathrm{T}}H(\tilde{y}^k - y^k)\\ &+\begin{pmatrix} x - \tilde{x}^k \\ y - \tilde{y}^k \\ \lambda - \tilde{\lambda}^k \end{pmatrix}^{\mathrm{T}} \left\{ \begin{pmatrix} -A^{\mathrm{T}}\tilde{\lambda}^k \\ -B^{\mathrm{T}}\tilde{\lambda}^k \\ A\tilde{x}^k + B\tilde{y}^k - b \end{pmatrix} - \begin{pmatrix} G(x^k - \tilde{x}^k) \\ \gamma B^{\mathrm{T}}B(y^k - \tilde{y}^k) \\ -B(y^k - \tilde{y}^k) + \frac{1}{\gamma}(\lambda^k - \tilde{\lambda}^k) \end{pmatrix} \right\} \geqslant 0.\end{aligned}$$

As a result, (7.45) follows.

The proof of Theorem 4.4 follows a similar line of derivations as in the proof of Theorem 3.1, and so we omit the details here.

References

[1] James, G.M., Paulson, C., Rusmevichientong, P.: The constrained lasso. Technical report, University of Southern California (2013)

[2] Alizadeh, M., Li, X., Wang, Z., Scaglione, A., Melton, R.: Demand-side management in the smart grid: information processing for the power switch. IEEE Sig. Process. Mag. **29**(5), 55–67 (2012)

[3] Chang, T.-H., Alizadeh, M., Scaglione A.: Coordinated home energy management for real-time power balancing. In: IEEE Power and Energy Society General Meeting, pp. 1–8 (2012)

[4] Li, N., Chen, L., Low, S.H.: Optimal demand response based on utility maximization in power networks. In: IEEE Power and Energy Society General Meeting, pp. 1–8 (2011)

[5] Paatero, J.V., Lund, P.D.: A model for generating household electricity load profiles. Int. J. Ener. Res. **30**(5), 273–290 (2006)

[6] Cui, Y., Li, X., Sun, D., Toh, K.-C.: On the convergence properties of a majorized ADMM for linearly constrained convex optimization problems with coupled objective functions. arXiv:1502.00098 (2015)

[7] Hong, M., Chang, T.-H., Wang, X., Razaviyayn, M., Ma, S., Luo, Z.-Q.: A block successive upper bound minimization method of multipliers for linearly constrained convex optimization. arXiv:1401.7079 (2014)

[8] Bertsekas, D.P., Tsitsiklis, J.N.: Parallel and Distributed Computation: Numerical Methods, vol. 23. Prentice Hall, Englewood Cliffs (1989)

[9] Douglas, J., Rachford, H.: On the numerical solution of heat conduction problems in two and three space variables. Trans. Am. Math. Soc. **82**(2), 421–439 (1956)

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

[11] Eckstein, J., Bertsekas, D.: On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. **55**(1–3), 293–318 (1992)

[12] 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), 1–122 (2011)

[13] Feng, C., Xu, H., Li, B.: An alternating direction method approach to cloud traffic management. arXiv:1407.8309 (2014)

[14] Scheinberg, K., Ma, S., Goldfarb, D.: Sparse inverse covariance selection via alternating linearization methods. In: Advances in neural information processing systems, pp. 2101–2109 (2010)

[15] Yin, W., Osher, S., Goldfarb, D., Darbon, J.: Bregman iterative algorithms for l_1-minimization with applications to compressed sensing. SIAM J. Imag. Sci. **1**(1), 143–168 (2008)

[16] Glowinski, R., Le Tallec, P.: Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics, vol. 9. SIAM, Philadelphia (1989)

[17] He, B., Yuan, X.: On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. SIAM J. Numer. Anal. **50**(2), 700–709 (2012)

[18] Monteiro, R.D., Svaiter, B.F.: Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. SIAM J. Optim. **23**(1), 475–507 (2013)

[19] Boley, D.: Local linear convergence of the alternating direction method of multipliers on quadratic or linear programs. SIAM J. Optim. **23**(4), 2183–2207 (2013)

[20] Deng, W., and Yin, W.: On the global and linear convergence of the generalized alternating direction method of multipliers. J. Sci. Comput. 1–28 (2012)

[21] Hong, M., Luo, Z.: On the linear convergence of the alternating direction method of multipliers. arXiv:1208.3922 (2012)

[22] Lin, T., Ma, S., Zhang, S.: On the global linear convergence of the ADMM with multi-block variables. arXiv:1408.4266 (2014)

[23] Chen, C., He, B., Ye, Y., Yuan, X.: The direct extension of admm for multi-block convex minimization problems is not necessarily convergent. Math. Program. **155**(1–2), 57–79 (2016)

[24] Deng, W., Lai, M.-J., Peng, Z., Yin, W.: Parallel multi-block admm with o (1/k) convergence. arXiv:1312.3040 (2013)

[25] He, B., Hou, L., Yuan, X.: On full jacobian decomposition of the augmented lagrangian method for separable convex programming. SIAM J. Optim. **25**(4), 2274–2312 (2015)

[26] He, B., Tao, M., Yuan, X.: Convergence rate and iteration complexity on the alternating direction method of multipliers with a substitution procedure for separable convex programming. Optimization Online (2012)

[27] Lin, T., Ma, S., Zhang, S.: Iteration complexity analysis of multi-block ADMM for a family of convex minimization without strong convexity. J. Sci. Comput. pp. 1–30 (2016)

[28] Hong, M., Luo, Z.-Q., Razaviyayn, M.: Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3836–3840 (2015)

[29] Ortega, J., Rheinboldt, W.: Iterative Solution of Nonlinear Equations in Several Variables. Classics in applied mathematics, vol. 30. SIAM, Philadelphia (2000)

[30] Drori, Y., Sabach, S., Teboulle, M.: A simple algorithm for a class of nonsmooth convex-concave saddle-point problems. Oper. Res. Lett. **43**(2), 209–214 (2015)

[31] Gao, X., Jiang, B., Zhang, S.: On the information-adaptive variants of the ADMM: an iteration complexity perspective. Optimization Online (2014)

[32] Liu, J., Chen, J., Ye, J.: Large-scale sparse logistic regression. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 547–556. ACM (2009)

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

[34] Lin, T., Ma, S., Zhang, S.: An extragradient-based alternating direction method for convex minimization. Foundations of Computational Mathematics, pp. 1–25 (2015)

[35] Robinson, D.P., Tappenden, R.E.: A flexible admm algorithm for big data applications. arXiv:1502.04391 (2015)

[36] Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. **146**, 459–494 (2014)