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

Xiang Gao * Shuzhong Zhang †

May 9, 2015

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* (APGMM), *alternating gradient projection method of multipliers* (AGPMM), 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.

Mathematics Subject Classification: 90C25, 49M27, 68Q25.

*Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN 55455, USA. Email: gaoxx460@umn.edu

†Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN 55455, USA. Email: zhangs@umn.edu

1 Introduction

In this paper we consider the following model:

$$\begin{array}{ll} \min & f(x,y)+h_1(x)+h_2(y) \\ \text{s.t.} & Ax+By=b, \\ & x\in\mathcal{X}, y\in\mathcal{Y} \end{array} \tag{1}$$

where $x\in\mathbf{R}^p$, $y\in\mathbf{R}^q$, $A\in\mathbf{R}^{m\times p}$, $B\in\mathbf{R}^{m\times q}$, $b\in\mathbf{R}^m$, $\mathcal{X},\mathcal{Y}$ 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) is

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

where λ is the multiplier.

Many problems of interest in various areas including signal processing, image processing, machine learning and statistical learning, can be formulated in the form of (1); see [21, 9] and the references therein. When the coupling term $f(x,y)$ is absent from the objective, then there is a popular algorithm known as the *Alternating Direction Method of Multipliers* (ADMM) for solving (1), with the following iterative scheme:

$$\left\{\begin{array}{l} x^{k+1}=\arg\min_{x\in\mathcal{X}}\mathcal{L}_\gamma(x,y^k,\lambda^k) \\ y^{k+1}=\arg\min_{y\in\mathcal{Y}}\mathcal{L}_\gamma(x^{k+1},y,\lambda^k) \\ \lambda^{k+1}=\lambda^k-\gamma(Ax^{k+1}+By^{k+1}-b). \end{array}\right. \tag{2}$$

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. [12, 14, 13, 3]. 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 [6, 31, 30, 15] and the references therein. The ADMM and its variants are known to have superior performances over existing traditional nonlinear programming algorithms in practice; see [2, 7, 28].

If the coupling term f is absent from the objective, then 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 [17, 14]. However, the rate of convergence was only established recently. In particular, [20, 27] 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 [11, 22, 4, 24]. 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 [8] the authors show

by example 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 convexity. Indeed, under the strong convexity condition on some parts of the objective and certain assumptions on the constraint matrices, [10, 19, 18, 25] show that an $O(1/t)$ convergence rate can still be achieved for the multi-block ADMM. Another direction of study is to consider application of ADMM to solve non-convex problems. For instance, [23] 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 [21] and [9]. In [21], the authors consider the general multi-block setting and they propose a BSUMM approach to cope with the non-separability of the objective in (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. [9] consider the problem (1) by introducing a quadratic upper-bound function for the non-separable part of augmented Lagrangian function; they show that their algorithm has an $O(1/t)$ convergence rate. In this paper, we study the ADMM and its variants for (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; see 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 algorithm 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 (2) 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 that 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). 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 [25], an $O(1/t)$ iteration bound still holds for the multi-block model.

The rest of the paper is organized as follows. In Section 2, 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 A. In Section 3, 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 Section 4.

2 The Algorithms

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

$$u := \begin{pmatrix} x \\ y \end{pmatrix}, \quad w := \begin{pmatrix} x \\ y \\ \lambda \end{pmatrix}, \quad F(w) := \begin{pmatrix} -A^\top \lambda \\ -B^\top \lambda \\ Ax + By - b \end{pmatrix}, \tag{3}$$

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*

$$\|\nabla f(u_2) - \nabla f(u_1)\| \leq L\|u_2 - u_1\|, \ \forall\, u_1, u_2 \in \mathcal{X} \times \mathcal{Y}, \tag{4}$$

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 2.1 *Suppose that function f satisfies (4), then we have*

$$f(u_2) \leq f(u_1) + \nabla f(u_1)^\top (u_2 - u_1) + \frac{L}{2}\|u_2 - u_1\|^2, \tag{5}$$

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

$$f(u_2) \leq f(u_1) + \nabla f(u_3)^\top (u_2 - u_1) + \frac{L}{2}\|u_2 - u_3\|^2, \tag{6}$$

for any u_1, u_2, u_3.

Proof. Inequality (5) is well known; see ([29]). In the sequel we shall focus on (6). Since f is convex, we have

$$\begin{aligned} (u_1 - u_2)^\top \nabla f(u_3) &= (u_1 - u_3)^\top \nabla f(u_3) + (u_3 - u_2)^\top \nabla f(u_3) \\ &\leq f(u_1) - f(u_3) - (u_2 - u_3)^\top \nabla f(u_3). \end{aligned} \tag{7}$$

By (5), we have

$$f(u_2) - f(u_3) - \frac{L}{2}\|u_3 - u_2\|^2 \leq (u_2 - u_3)^\top \nabla f(u_3). \tag{8}$$

Combining (7) and (8) leads to

$$\begin{aligned}(u_1 - u_2)^\top \nabla f(u_3) &\leq f(u_1) - f(u_3) - \left(f(u_2) - f(u_3) - \frac{L}{2}\|u_3 - u_2\|^2\right) \\ &= f(u_1) - f(u_2) + \frac{L}{2}\|u_3 - u_2\|^2.\end{aligned}$$

□

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

$$Q := \begin{pmatrix} G & 0 & 0 \\ 0 & \gamma B^\top B & 0 \\ 0 & -B & \frac{1}{\gamma} I_m \end{pmatrix}, \quad P := \begin{pmatrix} I_p & 0 & 0 \\ 0 & I_q & 0 \\ 0 & -\gamma B & I_m \end{pmatrix}, \quad M := \begin{pmatrix} G & 0 & 0 \\ 0 & \gamma B^\top B & 0 \\ 0 & 0 & \frac{1}{\gamma} I_m \end{pmatrix}; \tag{9}$$

hence, $Q = MP$. Suppose the sequence $\{w^k\}$ is generated by an algorithm, we introduce an auxiliary sequence:

$$\tilde{w}^k := \begin{pmatrix} \tilde{x}^k \\ \tilde{y}^k \\ \tilde{\lambda}^k \end{pmatrix} = \begin{pmatrix} x^{k+1} \\ y^{k+1} \\ \lambda^k - \gamma(Ax^{k+1} + By^k - b) \end{pmatrix}. \tag{10}$$

Based on (10) and (9), the relationship between the new sequence $\{\tilde{w}^k\}$ and the original $\{w^k\}$ is

$$w^{k+1} = w^k - P(w^k - \tilde{w}^k). \tag{11}$$

2.1 Alternating Direction Method of Multipliers

As we discussed earlier, the ADMM can be applied straightforwardly to solve (1), assuming that the augmented Lagrangian (with a proximal term) can be optimized for each block of variables, while other variables are fixed. This gives rise to the following scheme:

ADMM

Initialize $x^0 \in \mathcal{X}, y^0 \in \mathcal{Y}$ and λ^0

for $k = 0, 1, \cdots,$ **do**

$x^{k+1} = \arg\min_{x \in \mathcal{X}} \mathcal{L}_\gamma(x, y^k, \lambda^k) + \frac{1}{2}\|x - x^k\|_G^2$;

$y^{k+1} = \arg\min_{y \in \mathcal{Y}} \mathcal{L}_\gamma(x^{k+1}, y, \lambda^k) + \frac{1}{2}\|y - y^k\|_H^2$;

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

end for

In the above algorithm, G and H are two pre-specified positive semidefinite matrices. The main result concerning its convergence and iteration complexity are summarized in the following theorem, whose proof can be found in Appendix A.1.

Theorem 2.2 *Suppose that ∇f satisfies Lipschitz condition (4), and $h_2(y)$ is strongly convex with parameter $\sigma > 0$, i.e.*

$$h_2(y) \geq h_2(z) + h_2'(z)^\top (y - z) + \frac{\sigma}{2}\|y - z\|^2 \tag{12}$$

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

$$\bar{u}_n := \frac{1}{n}\sum_{k=1}^{n} u^k, \tag{13}$$

we have

$$h(\bar{u}_t) - h(u^*) + \rho\|A\bar{x}_t + B\bar{y}_t - b\| \leq \frac{1}{2t}\left(\mathrm{dist}(x^0, \mathcal{X}^*)_G^2 + \mathrm{dist}(y^0, \mathcal{Y}^*)_{\hat{H}}^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right), \tag{14}$$

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

We quote Lemma 2.4 in [16] as follows:

> Assume that $\rho > 0$, and $\tilde{x} \in X$ is an approximate solution of the problem $f^* := \inf\{f(x) : Ax - b = 0, x \in X\}$ where f is convex, satisfying
>
> $$f(\bar{x}) - f^* + \rho\|A\bar{x} - b\| \leq \epsilon. \tag{15}$$
>
> Then, we have
>
> $$\|A\bar{x} - b\| \leq \frac{\epsilon}{\rho - \|\lambda^*\|} \text{ and } f(\bar{x}) - f^* \leq \epsilon$$
>
> 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 (14) in Theorem 2.2 automatically establishes that

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

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

2.2 Alternating Proximal Gradient Method of Multipliers

In some applications, the augmented Lagrangian function may be difficult to minimize for some block of variables, while fixing all others. 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 (ISTA) Algorithm (cf. [1]), although we are dealing with multiple blocks of variables here. We shall call the new method *Alternating Proximal Gradient Method of Multipliers* (APGMM), presented as follows:

APGMM

Initialize $x^0 \in \mathcal{X}, y^0 \in \mathcal{Y}$ and λ^0

for $k = 0, 1, \cdots,$ **do**

$x^{k+1} = \arg\min_{x\in\mathcal{X}} \nabla_x f(x^k, y^k)^\top (x - x^k) + h_1(x) + \frac{\gamma}{2}\|Ax + By^k - b - \frac{1}{\gamma}\lambda^k\|^2 + \frac{1}{2}\|x - x^k\|_G^2$;

$y^{k+1} = \arg\min_{y\in\mathcal{Y}} \nabla_y f(x^k, y^k)^\top (y - y^k) + h_2(y) + \frac{\gamma}{2}\|Ax^{k+1} + By - b - \frac{1}{\gamma}\lambda^k\|^2 + \frac{1}{2}\|y - y^k\|_H^2$;

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

end for

The convergence property and iteration complexity are summarized in the following theorem, whose proof is in Appendix A.2.

Theorem 2.3 *Suppose that ∇f satisfies Lipschitz condition (4). Let $\{w^k\}$ be the sequence generated by the APGMM, and $G \succ LI_p$ and $H \succ LI_q$. Then, the sequence $\{w^k\}$ generated by the APGMM converges to an optimal solution. Moreover, for any integer $n > 0$, letting*

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

it holds that

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

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

2.3 Alternating Gradient Projection Method of Multipliers

Implementing proximal gradient step may still be difficult for some instances of applications. It is therefore natural to further simplify the step to *Gradient Projection*. 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 λ^0

for $k = 0, 1, \cdots,$ **do**

$x^{k+1} = [x^k - \alpha(\nabla_x f(x^k, y^k) + \nabla_x h_1(x^k) - A^\top\lambda^k + A^\top(Ax^k + By^k - b))]_{\mathcal{X}}$;

$y^{k+1} = [y^k - \alpha(\nabla_y f(x^k, y^k) + \nabla_y h_2(y^k) - B^\top\lambda^k + B^\top(Ax^{k+1} + By^k - b))]_{\mathcal{Y}}$;

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

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 A.3.

Theorem 2.4 *Suppose that ∇f satisfies Lipschitz condition (4). Let w^k be the sequence generated by the AGPMM, and $G := \gamma A^\top A + \frac{1}{\alpha} I_p$, $H := \frac{1}{\alpha} I_q - \gamma B^\top B$. 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. For any integer $n > 0$, letting*

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

it holds that

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

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

2.4 Hybrids

There are instances where one part of the block variables is easy to deal with, while the other part is difficult, e.g. [26]. 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 in two blocks of variables:

ADM-PG

Initialize $x^0 \in \mathcal{X}, y^0 \in \mathcal{Y}$ and λ^0

for $k = 0, 1, \cdots,$ **do**

$x^{k+1} = \arg\min_{x \in \mathcal{X}} \mathcal{L}_\gamma(x, y^k, \lambda^k) + \frac{1}{2}\|x - x^k\|_G^2$;

$y^{k+1} = \arg\min_{y \in \mathcal{Y}} \nabla_y f(x^{k+1}, y^k)^\top (y - y^k) + h_2(y) + \frac{\gamma}{2}\|Ax^{k+1} + By - b - \frac{1}{\gamma}\lambda^k\|^2 + \frac{1}{2}\|y - y^k\|_H^2$;

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

end for

The iteration complexity of the above method is as follows. The proof of the theorem can be found in Appendix A.4.

Theorem 2.5 *Suppose that ∇f satisfies Lipschitz condition (4). 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. For any integer $n > 0$, letting*

$$\bar{u}_n := \frac{1}{n}\sum_{k=1}^{n} u^k, \tag{16}$$

it holds that

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

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

Another possible approach is to combine ADMM with Gradient Projection, which works as follows:

ADM-GP

Initialize $x^0 \in \mathcal{X}, y^0 \in \mathcal{Y}$ and λ^0

for $k = 0, 1, \cdots,$ **do**

$x^{k+1} = \arg\min_{x\in\mathcal{X}} \mathcal{L}_\gamma(x, y^k, \lambda^k) + \frac{1}{2}\|x - x^k\|_G^2$;

$y^{k+1} = [y^k - \alpha(\nabla_y f(x^{k+1}, y^k) + \nabla_y h_2(y^k) - B^\top\lambda^k + B^\top(Ax^{k+1} + By^k - b))]_{\mathcal{Y}}$;

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

end for

The main convergence result is as follows, and the proof of the theorem can be found in Appendix A.4.

Theorem 2.6 *Let w^k be the sequence generated by the ADM-GP, $G \succ 0$ and $H := \frac{1}{\alpha}I_q - \gamma B^\top B$. 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. For any integer $n > 0$, letting*

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

it holds that

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

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

3 The General Multi-Block Model

Different variations of the ADMM have been a popular subject of study in the recent years, and the ADMM has been extended to solve general formulation with multiple blocks of variables; see [25] 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 non-separable objective function. In particular, the problem that we consider is as follows:

$$\begin{array}{ll} \min & f(x_1, x_2, \ldots, x_n) + \sum_{i=1}^{n} h_i(x_i) \\ \text{s.t.} & A_1x_1 + A_2x_2 + \cdots + A_nx_n = b, \\ & x_i \in \mathcal{X}_i, i = 1, 2, \ldots, n \end{array} \tag{17}$$

where $A_i \in \mathbf{R}^{m\times p_i}$, $b \in \mathbf{R}^m$, $\mathcal{X}_i \subset \mathbf{R}^{p_i}$ are closed convex sets, and f, h_i $i = 1, \ldots, n$, are convex closed functions. Note that many important applications are in the form of (17), e.g. multi-stage stochastic programming. Accordingly, the ADMM algorithm for solving the problem (17) is:

The Multi-block ADMM

Initialize with $x_i^0 \in \mathcal{X}_i, i = 1, \ldots, n$, and λ^0

for $k = 0, 1, \cdots$, **do**

$x_1^{k+1} = \arg\min_{x_1\in\mathcal{X}_1} \mathcal{L}_\gamma(x_1, x_2^k, \ldots, x_n^k, \lambda^k) + \frac{1}{2}\|x_1 - x_1^k\|^2_{H_1}$;

$x_2^{k+1} = \arg\min_{x_2\in\mathcal{X}_2} \mathcal{L}_\gamma(x_1^{k+1}, x_2, x_3^k \ldots, x_n^k, \lambda^k) + \frac{1}{2}\|x_2 - x_2^k\|^2_{H_2}$;

$\vdots$

$x_i^{k+1} = \arg\min_{x_i\in\mathcal{X}_i} \mathcal{L}_\gamma(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i, x_{i+1}^k \ldots, x_n^k, \lambda^k) + \frac{1}{2}\|x_i - x_i^k\|^2_{H_i}$;

$\vdots$

$x_n^{k+1} = \arg\min_{x_n\in\mathcal{X}_n} \mathcal{L}_\gamma(x_1^{k+1}, \ldots, x_{n-1}^{k+1}, x_n, \lambda^k) + \frac{1}{2}\|x_n - x_n^k\|^2_{H_n}$;

$\lambda^{k+1} = \lambda^k - \beta(A_1x_1^{k+1} + A_2x_2^{k+1} + \cdots + A_nx_n^{k+1})$.

end for

---

where H_i, $i = 1, \ldots, n$, are pre-specified positive semidefinite matrices, γ is the augmented Lagrangian constant, and β is the dual stepsize. An $O(1/t)$ convergence rate of the ADMM can still be shown to hold 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 3.1 *The functions h_i, $i = 2, \ldots, n$, are strongly convex with parameters $\sigma_i > 0$:*

$$h_i(y) \geq h_i(x) + (y - x)^\top h_i'(x) + \frac{\sigma_i}{2}\|y - x\|^2,$$

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

Assumption 3.2 *The gradient of function $f(x_1, x_2, \ldots, x_n)$ is Lipschitz continuous with parameter $L \geq 0$:*

$$\|\nabla f(x_1', x_2', \ldots, x_n') - \nabla f(x_1, x_2, \ldots, x_n)\| \leq L\|(x_1' - x_1, x_2' - x_2, \ldots, x_n' - x_n)\|$$

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

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

Proposition 3.1 *Suppose that there are γ, β and δ satisfying*

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

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

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

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

$$\begin{aligned}
& h(u^*) - h(u^{k+1}) + \begin{pmatrix} x_1^* - x_1^{k+1} \\ \vdots \\ x_n^* - x_n^{k+1} \\ \lambda - \lambda^{k+1} \end{pmatrix}^\top \begin{pmatrix} -A_1^\top \lambda^{k+1} \\ \vdots \\ -A_n^\top \lambda^{k+1} \\ \sum_{i=1}^n A_i x_i^{k+1} - b \end{pmatrix} \\
& + \frac{\gamma}{2} \sum_{i=2}^n \left(\left\| \sum_{j=1}^{i-1} A_j x_j^* + \sum_{j=i}^n A_j x_j^k - b \right\|^2 - \left\| \sum_{j=1}^{i-1} A_j x_j^* + \sum_{j=i}^n A_j x_j^{k+1} - b \right\|^2 \right) \\
& + \frac{1}{2\beta} \left(\left\| \lambda - \lambda^k \|^2 - \|\lambda - \lambda^{k+1} \right\|^2 \right) + \frac{1}{2} \sum_{i=1}^n \left(\|x_i^* - x_i^k\|_{H_i}^2 - \|x_i^* - x_i^{k+1}\|_{H_i}^2 \right) \\
\geq & \left(\frac{\gamma - \beta}{2\beta^2} \right) \|\lambda^k - \lambda^{k+1}\|^2 + \frac{1}{2} \sum_{i=1}^3 \|x_i^{k+1} - x_i^k\|_{H_i^s}^2.
\end{aligned}$$

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

Proposition 3.2 *Let w^k be the sequence generated by the ADMM, then*

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

where $\mathcal{L}_i(w^*, w) := \sum_{j=1}^{i-1} A_j x_j^* + \sum_{j=i}^{n} A_j x_j - b$, $i = 2, ..., n$, *and*

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

Propositions 3.1 and 3.2 lead to the following theorem:

Theorem 3.3 *Under the assumptions of Propositions 3.1 and 3.2, and*

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

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*

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

4 Concluding Remarks

In [5], the following model is considered

$$\min \quad f(x) + g(y) + H(x, y), \tag{18}$$

which can be regarded as (1) without constraints, and the so-called *proximal alternating linearized minimization* (PALM) algorithm is proposed. The main focus of [5] 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 (18) 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 [5], PALM can be extended to allow multiple blocks. Similarly, APGMM is also extendable to solve (17). 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) where the objective is non-convex but satisfies the Kurdyka-Łojasiewicz property. It is also interesting to consider stochastic programming models studied in [16], but now allowing the objective function to be non-separable.

References

[1] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. 6

[2] D. P. Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014. 2

[3] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and distributed computation: numerical methods*, volume 23. Prentice hall Englewood Cliffs, NJ, 1989. 2

[4] D. Boley. Local linear convergence of the alternating direction method of multipliers on quadratic or linear programs. *SIAM Journal on Optimization*, 23(4):2183–2207, 2013. 2

[5] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146:459–494, 2014. 12

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011. 2

[7] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004. 2

[8] C. Chen, B. He, Y. Ye, and X. Yuan. The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. *Optimization Online*, 2013. 2

[9] Y. Cui, X. Li, D. Sun, and K.-C. Toh. On the convergence properties of a majorized ADMM for linearly constrained convex optimization problems with coupled objective functions. *arXiv preprint arXiv:1502.00098*, 2015. 2, 3

[10] W. Deng, M. Lai, and W. Yin. On the $o(1/k)$ convergence and parallelization of the alternating direction method of multipliers. *arXiv preprint arXiv:1312.3040*, 2013. 3

[11] W. Deng and W. Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. 2012. 2

[12] J. Douglas and H. Rachford. On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the American mathematical Society*, 82(2):421–439, 1956. 2

[13] J. Eckstein. *Splitting methods for monotone operators with applications to parallel optimization*. PhD thesis, Massachusetts Institute of Technology, 1989. 2

[14] J. Eckstein and D. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318, 1992. 2

[15] C. Feng, H. Xu, and B. Li. An alternating direction method approach to cloud traffic management. *arXiv preprint arXiv:1407.8309*, 2014. 2

[16] X. Gao, B. Jiang, and S. Zhang. On the information-adaptive variants of the ADMM: an iteration complexity perspective. 2014. 6, 12

[17] R. Glowinski and P. Le Tallec. *Augmented Lagrangian and operator-splitting methods in nonlinear mechanics*, volume 9. SIAM, 1989. 2

[18] B. He, L. Hou, and X. Yuan. On full Jacobian decomposition of the augmented Lagrangian method for separable convex programming. 2013. 3

[19] B. He, M. Tao, and X. Yuan. Convergence rate and iteration complexity on the alternating direction method of multipliers with a substitution procedure for separable convex programming. *Math. Oper. Res., under revision*, 2012. 3

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

[21] M. Hong, T.-H. Chang, X. Wang, M. Razaviyayn, S. Ma, and Z.-Q. Luo. A block successive upper bound minimization method of multipliers for linearly constrained convex optimization. *arXiv preprint arXiv:1401.7079*, 2014. 2, 3

[22] M. Hong and Z. Luo. On the linear convergence of the alternating direction method of multipliers. *arXiv preprint arXiv:1208.3922*, 2012. 2

[23] M. Hong, Z.-Q. Luo, and M. Razaviyayn. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. *arXiv preprint arXiv:1410.1390*, 2014. 3

[24] T. Lin, S. Ma, and S. Zhang. On the global linear convergence of the ADMM with multi-block variables. *arXiv preprint arXiv:1408.4266*, 2014. 2

[25] T. Lin, S. Ma, and S. Zhang. Iteration complexity analysis of multi-block ADMM for a family of convex minimization without strong convexity. *Optimization Online, 2015-04-4860*, 2015. 3, 10

[26] S. Ma and S. Zhang. An extragradient-based alternating direction method for convex minimization. *arXiv preprint arXiv:1301.6308*, 2013. 8

[27] R. Monteiro and B. Svaiter. Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. *Optimization-online preprint*, 2713:1, 2010. 2

[28] A. Nedic and A. Ozdaglar. Cooperative distributed multi-agent. *Convex Optimization in Signal Processing and Communications*, page 340, 2010. 2

[29] J. Ortega and W. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Classics Appl. Math. 30, SIAM, Philadelphia, 2000. 4

[30] K. Scheinberg, S. Ma, and D. Goldfarb. Sparse inverse covariance selection via alternating linearization methods. In *Advances in neural information processing systems*, pages 2101–2109, 2010. 2

[31] W. Yin, S. Osher, D. Goldfarb, and J. Darbon. Bregman iterative algorithms for l_1-minimization with applications to compressed sensing. *SIAM Journal on Imaging Sciences*, 1(1):143–168, 2008. 2

A Proofs of the Convergence Theorems

A.1 Proof of Theorem 2.2

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

$$(w_1 - w_2)^\top (F(w_1) - F(w_2)) = 0.$$

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

$$(\bar{w} - w)^\top F(\bar{w}) = \frac{1}{t}\sum_{k=0}^{t-1} (w^k - w)^\top F(w^k). \tag{19}$$

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

Proposition A.1 *Suppose h_2 is strongly convex with parameter $\sigma > 0$. Let $\{\tilde{w}^k\}$ be defined by (10), and the matrices Q, M, P be given in (9). First of all, for any $w \in \Omega$, we have*

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\ \geq\ & (w - \tilde{w}^k)^\top Q(w^k - \tilde{w}^k) - \left(\left(\frac{L}{2} + \frac{L^2}{2\sigma} \right) \|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k) \right). \end{aligned} \tag{20}$$

Furthermore,

$$(w - \tilde{w}^k)^\top Q(w^k - \tilde{w}^k) = \frac{1}{2}\left(\|w - w^{k+1}\|_M^2 - \|w - w^k\|_M^2 \right) + \frac{1}{2}\|x^k - \tilde{x}^k\|_G^2 + \frac{1}{2\gamma}\|\lambda^k - \tilde{\lambda}^k\|^2. \tag{21}$$

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

$$\begin{aligned} & (x - x^{k+1})^\top \left[\nabla_x f(x^{k+1}, y^k) + h_1'(x^{k+1}) - A^\top(\lambda^k - \gamma(Ax^{k+1} + By^k - b)) + G(x^{k+1} - x^k) \right] \\ \geq\ & 0 \quad \forall x \in \mathcal{X}, \end{aligned}$$

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

$$\begin{aligned} & (y - y^{k+1})^\top \left[\nabla_y f(x^{k+1}, y^{k+1}) + h_2'(y^{k+1}) - B^\top(\lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b)) + H(y^{k+1} - y^k) \right] \\ \geq \; & 0 \quad \forall y \in \mathcal{Y} \end{aligned}$$

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

$$(x - \tilde{x}^k)^\top \left[\nabla_x f(\tilde{x}^k, y^k) + h_1'(\tilde{x}^k) - A^\top \tilde{\lambda}^k + G(\tilde{x}^k - x^k) \right] \geq 0 \quad \forall x \in \mathcal{X}, \tag{22}$$

and

$$(y - \tilde{y}^k)^\top \left[\nabla_y f(\tilde{x}^k, \tilde{y}^k) + h_2'(\tilde{y}^k) - B^\top \tilde{\lambda}^k + \gamma B^\top B(\tilde{y}^k - y^k) + H(\tilde{y}^k - y^k) \right] \geq 0 \quad \forall y \in \mathcal{Y}. \tag{23}$$

Observe the following chain of inequalities

$$\begin{aligned} & (x - \tilde{x}^k)^\top \nabla_x f(\tilde{x}^k, y^k) + (y - \tilde{y}^k)^\top \nabla_y f(\tilde{x}^k, \tilde{y}^k) \\ = \; & (x - \tilde{x}^k)^\top \nabla_x f(\tilde{x}^k, y^k) + (y - \tilde{y}^k)^\top \nabla_y f(\tilde{x}^k, y^k) + (y - \tilde{y}^k)^\top (\nabla_y f(\tilde{x}^k, \tilde{y}^k) - \nabla_y f(\tilde{x}^k, y^k)) \\ \leq \; & (x - \tilde{x}^k)^\top \nabla_x f(\tilde{x}^k, y^k) + (y - \tilde{y}^k)^\top \nabla_y f(\tilde{x}^k, y^k) + L\|y - \tilde{y}^k\| \|y^k - \tilde{y}^k\| \\ = \; & (x - \tilde{x}^k)^\top \nabla_x f(\tilde{x}^k, y^k) + (y - y^k)^\top \nabla_y f(\tilde{x}^k, y^k) + (y^k - \tilde{y}^k)^\top \nabla_y f(\tilde{x}^k, y^k) + L\|y - \tilde{y}^k\| \|y^k - \tilde{y}^k\| \\ \leq \; & f(x, y) - f(\tilde{x}^k, y^k) - (\tilde{y}^k - y^k)^\top \nabla_y f(\tilde{x}^k, y^k) + L\|y - \tilde{y}^k\| \|y^k - \tilde{y}^k\| \\ \overset{(5)}{\leq} \; & f(x, y) - f(\tilde{x}^k, \tilde{y}^k) + \frac{L}{2}\|y^k - \tilde{y}^k\|^2 + L\|y - \tilde{y}^k\| \|y^k - \tilde{y}^k\| \\ \leq \; & f(x, y) - f(\tilde{x}^k, \tilde{y}^k) + \frac{L}{2}\|y^k - \tilde{y}^k\|^2 + \frac{\sigma}{2}\|y - \tilde{y}^k\|^2 + \frac{L^2}{2\sigma}\|y^k - \tilde{y}^k\|^2. \end{aligned} \tag{24}$$

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

$$(\lambda - \tilde{\lambda}^k)^\top \left(A\tilde{x}^k + B\tilde{y}^k - b \right) = (\lambda - \tilde{\lambda}^k)^\top \left(-B(y^k - \tilde{y}^k) + \frac{1}{\gamma}(\lambda^k - \tilde{\lambda}^k) \right). \tag{25}$$

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

$$(y - \tilde{y}^k)^\top h_2'(\tilde{y}^k) \leq h_2(y) - h_2(\tilde{y}^k) - \frac{\sigma}{2}\|y - \tilde{y}^k\|^2. \tag{26}$$

Because of the convexity of $h_1(x)$ and combining (26), (25), (24), (23) and (22), we have

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

for any $w \in \Omega$ and $\tilde{w}^k$. By definition of Q, (20) of Proposition A.1 follows. For (21), due to the similarity, we refer to Lemma 3.2 in [20] (noting the matrices Q, P and M). $\square$

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

Proposition A.2 *Let w^k be the sequence generated by the ADMM, $\tilde{w}^k$ be defined as in (10) and H satisfy $H_s := H - \left(L + \frac{L^2}{\sigma}\right) I_q \succeq 0$. Then the following holds*

$$\frac{1}{2}\left(\|w^* - w^k\|^2_{\hat{M}} - \|w^* - w^{k+1}\|^2_{\hat{M}}\right) - \frac{1}{2}\|w^k - \tilde{w}^k\|^2_{H_d} \geq 0, \tag{27}$$

where

$$\hat{H} = \gamma B^\top B + H, \quad \hat{M} = \begin{pmatrix} G & 0 & 0 \\ 0 & \hat{H} & 0 \\ 0 & 0 & \frac{1}{\gamma} I_m \end{pmatrix}, \quad \text{and } H_d = \begin{pmatrix} G & 0 & 0 \\ 0 & H_s & 0 \\ 0 & 0 & \frac{1}{\gamma} I_m \end{pmatrix}. \tag{28}$$

Proof. It follows from Proposition A.1 that

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\ \geq\ & (w - \tilde{w}^k)^\top Q(w^k - \tilde{w}^k) - \left(\left(\frac{L}{2} + \frac{L^2}{2\sigma}\right)\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k)\right) \\ =\ & \frac{1}{2}\left(\|w - w^{k+1}\|^2_M - \|w - w^k\|^2_M\right) + \frac{1}{2}\|x^k - \tilde{x}^k\|^2_G + \frac{1}{2\gamma}\|\lambda^k - \tilde{\lambda}^k\|^2 \\ & - \left(\left(\frac{L}{2} + \frac{L^2}{2\sigma}\right)\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k)\right). \end{aligned} \tag{29}$$

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

$$\begin{aligned} & \left(\frac{L}{2} + \frac{L^2}{2\sigma}\right)\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k) \\ =\ & \left(\frac{L}{2} + \frac{L^2}{2\sigma}\right)\|y^k - \tilde{y}^k\|^2 + \frac{1}{2}\left(\|y - y^k\|^2_H - \|y - \tilde{y}^k\|^2_H - \|y^k - \tilde{y}^k\|^2_H\right) \\ =\ & \frac{1}{2}\left(\|y - y^k\|^2_H - \|y - \tilde{y}^k\|^2_H\right) - \frac{1}{2}\|y^k - \tilde{y}^k\|^2_{H_s}. \end{aligned} \tag{30}$$

Thus, combining (29) and (30) we have

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\ \geq\ & \frac{1}{2}\left(\|w - w^{k+1}\|^2_M - \|w - w^k\|^2_M\right) - \frac{1}{2}\left(\|y - y^k\|^2_H - \|y - \tilde{y}^k\|^2_H\right) \\ & + \frac{1}{2}\|x^k - \tilde{x}^k\|^2_G + \frac{1}{2}\|y^k - \tilde{y}^k\|^2_{H_s} + \frac{1}{2\gamma}\|\lambda^k - \tilde{\lambda}^k\|^2. \end{aligned} \tag{31}$$

By the definition of $\hat{M}$ and H_d according to (28), it follows from (31) that

$$h(\tilde{u}^k) - h(u) + (\tilde{w}^k - w)^\top F(\tilde{w}^k) \leq \frac{1}{2}\left(\|w - w^k\|^2_{\hat{M}} - \|w - w^{k+1}\|^2_{\hat{M}}\right) - \frac{1}{2}\|w^k - \tilde{w}^k\|^2_{H_d}. \tag{32}$$

Letting $w = w^*$ in (32) we have

$$h(\tilde{u}^k) - h(u^*) + (\tilde{w}^k - w^*)^\top F(\tilde{w}^k) \leq \frac{1}{2}\left(\|w^* - w^k\|^2_{\hat{M}} - \|w^* - w^{k+1}\|^2_{\hat{M}}\right) - \frac{1}{2}\|w^k - \tilde{w}^k\|^2_{H_d}. \quad (33)$$

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

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

which completes the proof. □

Proof of Theorem 2.2.

Proof. First, according to (27), it holds that $\{w^k\}$ is bounded and

$$\lim_{k\to\infty} \|w^k - \tilde{w}^k\|_{H_d} = 0. \quad (34)$$

Thus, those two sequences have the same cluster points: For any $w^{k_n} \to w^\infty$, by (34) we also have $\tilde{w}^{k_n} \to w^\infty$. Applying inequality (20) to $\{w^{k_n}\}, \{\tilde{w}^{k_n}\}$ and taking the limit, it yields that

$$h(u) - h(u^\infty) + (w - w^\infty)^\top F(w^\infty) \geq 0. \quad (35)$$

Consequently, the cluster point w^∞ is an optimal solution. Since (27) is true for any optimal solution w^*, it also holds for w^∞, and that implies w^k will converge to w^∞.

Recall (20) and (21) in Proposition A.1, those would imply that:

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\ \geq\ & (w - \tilde{w}^k)^\top Q(w^k - \tilde{w}^k) - \left(\left(\frac{L}{2} + \frac{L^2}{2\sigma}\right)\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k)\right) \\ \geq\ & \frac{1}{2}\left(\|w - w^{k+1}\|^2_M - \|w - w^k\|^2_M\right) - \left(\left(\frac{L}{2} + \frac{L^2}{2\sigma}\right)\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k)\right). \end{aligned} \quad (36)$$

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

$$\begin{aligned} & \left(\frac{L}{2} + \frac{L^2}{2\sigma}\right)\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k) \\ =\ & \left(\frac{L}{2} + \frac{L^2}{2\sigma}\right)\|y^k - \tilde{y}^k\|^2 + \frac{1}{2}\left(\|y - y^k\|^2_H - \|y - \tilde{y}^k\|^2_H - \|y^k - \tilde{y}^k\|^2_H\right) \\ \leq\ & \frac{1}{2}\left(\|y - y^k\|^2_H - \|y - \tilde{y}^k\|^2_H\right). \end{aligned} \quad (37)$$

Thus, combining (36) and (37) leads to

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\ \geq\ & \frac{1}{2}\left(\|w - w^{k+1}\|_M^2 - \|w - w^k\|_M^2\right) - \frac{1}{2}\left(\|y - y^k\|_H^2 - \|y - \tilde{y}^k\|_H^2\right). \end{aligned} \tag{38}$$

By the definition of M in (9) and denoting $\hat{H} = \gamma B^\top B + H$, (38) leads to

$$\begin{aligned} & h(\tilde{u}^k) - h(u) + (\tilde{w}^k - w)^\top F(\tilde{w}^k) \\ \leq\ & \frac{1}{2}\left(\|x - x^k\|_G^2 - \|x - x^{k+1}\|_G^2\right) + \frac{1}{2}\left(\|y - y^k\|_{\hat{H}}^2 - \|y - y^{k+1}\|_{\hat{H}}^2\right) \\ & + \frac{1}{2\gamma}\left(\|\lambda - \lambda^k\|^2 - \|\lambda - \lambda^{k+1}\|^2\right). \end{aligned} \tag{39}$$

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 (16), 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 (39) over $k = 0, 1, \ldots, t-1$ yields

$$\begin{aligned} & h(\bar{u}_t) - h(u) + (\bar{w}_t - w)^\top F(\bar{w}_t) \\ \leq\ & \frac{1}{t}\sum_{k=0}^{t-1} h(\tilde{u}^k) - h(u) + \frac{1}{t}\sum_{k=0}^{t-1} (\tilde{w}^k - w)^\top F(\tilde{w}^k) \\ \leq\ & \frac{1}{2t}\left(\|x - x^0\|_G^2 + \|y - y^0\|_{\hat{H}}^2 + \frac{1}{\gamma}\|\lambda - \lambda^0\|^2\right), \end{aligned} \tag{40}$$

where the first inequality is due to the convexity of h and (19).

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

$$\begin{aligned} & \sup_{\lambda \in \mathcal{B}_\rho} \left\{ h(\bar{u}_t) - h(u^*) + (\bar{w}_t - w^*)^\top F(\bar{w}_t) \right\} \\ =\ & \sup_{\lambda \in \mathcal{B}_\rho} \left\{ h(\bar{u}_t) - h(u^*) + (\bar{x}_t - x^*)^\top (-A^\top \bar{\lambda}_t) + (\bar{y}_t - y^*)^\top (-B^\top \bar{\lambda}_t) + (\bar{\lambda}_t - \lambda)^\top (A\bar{x}_t + B\bar{y}_t - b) \right\} \\ =\ & \sup_{\lambda \in \mathcal{B}_\rho} \left\{ h(\bar{u}_t) - h(u^*) + \bar{\lambda}_t^\top (Ax^* + By^* - b) - \lambda^\top (A\bar{x}_t + B\bar{y}_t - b) \right\} \\ =\ & \sup_{\lambda \in \mathcal{B}_\rho} \left\{ h(\bar{u}_t) - h(u^*) - \lambda^\top (A\bar{x}_t + B\bar{y}_t - b) \right\} \\ =\ & h(\bar{u}_t) - h(u^*) + \rho\|A\bar{x}_t + B\bar{y}_t - b\|, \end{aligned} \tag{41}$$

which, combined with (40), implies that

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

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

$$h(\bar{u}_t) - h(u^*) + \rho\|A\bar{x}_t + B\bar{y}_t - b\| \le \frac{1}{2t}\left(\mathrm{dist}(x^0, \mathcal{X}^*)_G^2 + \mathrm{dist}(y^0, \mathcal{Y}^*)_H^2 + \frac{1}{\gamma}\left(\rho + \|\lambda^0\|\right)^2\right). \quad (42)$$

This completes the proof. □

A.2 Proof of Theorem 2.3

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

Proposition A.3 *Let $\{\tilde{w}^k\}$ be defined by (10), and the matrices Q, M, P be given as in (9). For any $w \in \Omega$, we have*

$$\begin{aligned}
& h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\
\ge\ & (w - \tilde{w}^k)^\top Q(w^k - \tilde{w}^k) - \left(\frac{L}{2}(\|x^k - \tilde{x}^k\|^2 + \|y^k - \tilde{y}^k\|^2) + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k)\right) \quad (43)\\
=\ & \frac{1}{2}\left(\|w - w^{k+1}\|_M^2 - \|w - w^k\|_M^2\right) + \frac{1}{2}\|x^k - \tilde{x}^k\|_G^2 + \frac{1}{2\gamma}\|\lambda^k - \tilde{\lambda}^k\|^2 \\
& - \left(\frac{L}{2}\left(\|x^k - \tilde{x}^k\|^2 + \|y^k - \tilde{y}^k\|^2\right) + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k)\right). \quad (44)
\end{aligned}$$

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

$$\begin{aligned}
& (x - x^{k+1})^\top \left[\nabla_x f(x^k, y^k) + h_1'(x^{k+1}) - A^\top(\lambda^k - \gamma(Ax^{k+1} + By^k - b)) + G(x^{k+1} - x^k)\right] \\
\ge\ & 0, \quad \forall x \in \mathcal{X},
\end{aligned}$$

and

$$\begin{aligned}
& (y - y^{k+1})^\top \left[\nabla_y f(x^k, y^k) + h_2'(y^{k+1}) - B^\top(\lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b)) + H(y^{k+1} - y^k)\right] \\
\ge\ & 0, \quad \forall y \in \mathcal{Y}.
\end{aligned}$$

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

$$(x - \tilde{x}^k)^\top \left[\nabla_x f(x^k, y^k) + h_1'(\tilde{x}^k) - A^\top\tilde{\lambda}^k + G(\tilde{x}^k - x^k)\right] \ge 0 \quad \forall x \in \mathcal{X}, \quad (45)$$

and

$$(y - \tilde{y}^k)^\top \left[\nabla_y f(x^k, y^k) + h_2'(\tilde{y}^k) - B^\top\tilde{\lambda}^k + \gamma B^\top B(\tilde{y}^k - y^k) + H(\tilde{y}^k - y^k)\right] \ge 0 \quad \forall y \in \mathcal{Y}. \quad (46)$$

Notice that

$$\begin{aligned}
& (x - \tilde{x}^k)^\top \nabla_x f(x^k, y^k) + (y - \tilde{y}^k)^\top \nabla_y f(x^k, y^k) \\
=\ & (x - x^k)^\top \nabla_x f(x^k, y^k) + (y - y^k)^\top \nabla_y f(x^k, y^k) + (x^k - \tilde{x}^k)^\top \nabla_x f(x^k, y^k) + (y^k - \tilde{y}^k)^\top \nabla_y f(x^k, y^k) \\
\le\ & f(x, y) - f(x^k, y^k) - (\tilde{x}^k - x^k)^\top \nabla_x f(x^k, y^k) - (\tilde{y}^k - y^k)^\top \nabla_y f(x^k, y^k) \\
\overset{(5)}{\le}\ & f(x, y) - f(\tilde{x}^k, \tilde{y}^k) + \frac{L}{2}\left(\|x^k - \tilde{x}^k\|^2 + \|y^k - \tilde{y}^k\|^2\right). \quad (47)
\end{aligned}$$

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

$$(\lambda - \tilde{\lambda}^k)^\top (A\tilde{x}^k + B\tilde{y}^k - b) = (\lambda - \tilde{\lambda}^k)^\top \left(-B(y^k - \tilde{y}^k) + \frac{1}{\gamma}(\lambda^k - \tilde{\lambda}^k)\right). \tag{48}$$

By the convexity of $h_1(x)$ and $h_2(y)$, combining (48), (47), (46) and (45), we have

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

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

By definition of Q, we have shown (43) in Proposition A.3. Equality (44) directly follows from (21) in Proposition A.1. □

With Proposition A.3 in place, we can show Theorem 2.3 by exactly following the same steps as in the proof of Theorem 2.2, 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 A.2. Since most steps of the proofs are almost identical to that of the previous theorems, we omit the details for succinctness.

Proposition A.4 *Let w^k be the sequence generated by the APGMM, and $\tilde{w}^k$ be as defined in (10), 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*

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

where

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

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

Theorem 2.3 follows from the above propositions.

A.3 Proof of Theorem 2.4

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,

$$\begin{aligned}&\|\nabla_x h_1(x_2) - \nabla_x h_1(x_1)\| \le L\|x_2 - x_1\|,\ \forall\, x_1, x_2 \in \mathcal{X},\\ &\|\nabla_y h_2(y_2) - \nabla_y h_2(y_1)\| \le L\|y_2 - y_1\|,\ \forall\, y_1, y_2 \in \mathcal{Y}.\end{aligned} \tag{49}$$

Proposition A.5 *Let $\{\tilde{w}^k\}$ be defined by (10), and the matrices Q, M, P be as given in (9), and $G := \gamma A^\top A + \frac{1}{\alpha} I_p$, $H := \frac{1}{\alpha} I_q - \gamma B^\top B \succeq 0$. First of all, for any $w \in \Omega$, we have*

$$\begin{aligned}& h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k)\\ \ge\ & (w - \tilde{w}^k)^\top Q(w^k - \tilde{w}^k) - \left(L(\|x^k - \tilde{x}^k\|^2 + \|y^k - \tilde{y}^k\|^2) + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k)\right) \qquad (50)\\ =\ & \frac{1}{2}\left(\|w - w^{k+1}\|_M^2 - \|w - w^k\|_M^2\right) + \frac{1}{2}\|x^k - \tilde{x}^k\|_G^2 + \frac{1}{2\gamma}\|\lambda^k - \tilde{\lambda}^k\|^2\\ & - \left(L(\|x^k - \tilde{x}^k\|^2 + \|y^k - \tilde{y}^k\|^2) + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k)\right). \qquad (51)\end{aligned}$$

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

$$\begin{aligned}& (x - x^{k+1})^\top \left[x^{k+1} - x^k + \alpha\left(\nabla_x f(x^k, y^k) + \nabla_y h_1(x^k) - A^\top(\lambda^k - \gamma(Ax^k + By^k - b))\right)\right]\\ \ge\ & 0 \quad \forall x \in \mathcal{X},\end{aligned}$$

and

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

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 respectively equivalent to

$$(x - \tilde{x}^k)^\top \left[\nabla_x f(x^k, y^k) + \nabla_x h_2(x^k) - A^\top\tilde{\lambda}^k + \gamma A^\top A(\tilde{x}^k - x^k) + \frac{1}{\alpha}(\tilde{x}^k - x^k)\right] \ge 0 \quad \forall x \in \mathcal{X}, \tag{52}$$

and

$$(y - \tilde{y}^k)^\top \left[\nabla_y f(x^k, y^k) + \nabla_y h_2(y^k) - B^\top\tilde{\lambda}^k + \frac{1}{\alpha}(\tilde{y}^k - y^k)\right] \ge 0 \quad \forall y \in \mathcal{Y}. \tag{53}$$

Similar to Proposition A.3, we have

$$\begin{aligned}& (x - \tilde{x}^k)^\top \nabla_x f(x^k, y^k) + (y - \tilde{y}^k)^\top \nabla_y f(x^k, y^k)\\ \overset{(5)}{\le}\ & f(x, y) - f(\tilde{x}^k, \tilde{y}^k) + \frac{L}{2}\left(\|x^k - \tilde{x}^k\|^2 + \|y^k - \tilde{y}^k\|^2\right).\end{aligned} \tag{54}$$

Moreover, by (6) we have

$$\begin{aligned}&(x - \tilde{x}^k)^\top \nabla_x h_1(x^k) \le h_1(x) - h_1(\tilde{x}^k) + \frac{L}{2}\|x^k - \tilde{x}^k\|^2\\ &(y - \tilde{y}^k)^\top \nabla_y h_2(y^k) \le h_2(y) - h_2(\tilde{y}^k) + \frac{L}{2}\|y^k - \tilde{y}^k\|^2.\end{aligned} \tag{55}$$

Besides,

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

Thus

$$(\lambda - \tilde{\lambda}^k)^\top (A\tilde{x}^k + B\tilde{y}^k - b) = (\lambda - \tilde{\lambda}^k)^\top \left(-B(y^k - \tilde{y}^k) + \frac{1}{\gamma}\left(\lambda^k - \tilde{\lambda}^k\right)\right). \tag{56}$$

Combining (56), (55), (54), (53), and (52), and noticing that $G := \gamma A^\top A + \frac{1}{\alpha} I_p$, $H := \frac{1}{\alpha} I_q - \gamma B^\top B$, we have, for any $w \in \Omega$ and $\tilde{w}^k$, that

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

Using the definition of Q, (50) follows. In view of (21) in Proposition A.1, equality (51) also readily follows. □

With Proposition A.5, similar as before, we can show Theorem 2.4 by following the same approach as in the proof of Theorem 2.2. We skip the details here for succinctness.

Proposition A.6 *Let w^k be the sequence generated by the AGPMM, $\tilde{w}^k$ be defined in (10) and $G := \gamma A^\top A + \frac{1}{\alpha} I_p$, $H := \frac{1}{\alpha} I_q - \gamma B^\top B$. Suppose that α satisfies that $H_s := H - 2LI_q \succ 0$ and $G_s := G - 2LI_p \succ 0$. Then the following holds*

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

where

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

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

Theorem 2.4 now follows from the above propositions.

A.4 Proofs of Theorems 2.5 and 2.6

Proposition A.7 *Let $\{\tilde{w}^k\}$ be defined by (10), and the matrices Q, M, P be given in (9). For any $w \in \Omega$, we have*

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\ \geq\ & (w - \tilde{w}^k)^\top Q(w^k - \tilde{w}^k) - \left(\frac{L}{2}\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k)\right) \\ =\ & \frac{1}{2}\left(\|w - w^{k+1}\|_M^2 - \|w - w^k\|_M^2\right) + \frac{1}{2}\|x^k - \tilde{x}^k\|_G^2 + \frac{1}{2\gamma}\|\lambda^k - \tilde{\lambda}^k\|^2 \\ & - \left(\frac{L}{2}\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k)\right). \end{aligned} \tag{57}$$

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

$$\begin{aligned} & (x - x^{k+1})^\top \left[\nabla_x f(x^{k+1}, y^k) + h_1'(x^{k+1}) - A^\top(\lambda^k - \gamma(Ax^{k+1} + By^k - b)) + G(x^{k+1} - x^k)\right] \\ \geq\ & 0 \quad \forall x \in \mathcal{X}, \end{aligned}$$

and

$$\begin{aligned} & (y - y^{k+1})^\top \left[\nabla_y f(x^{k+1}, y^k) + h_2'(y^{k+1}) - B^\top(\lambda^k - \gamma(Ax^{k+1} + By^{k+1} - b)) + H(y^{k+1} - y^k)\right] \\ \geq\ & 0 \quad \forall y \in \mathcal{Y}. \end{aligned}$$

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

$$(x - \tilde{x}^k)^\top \left[\nabla_x f(\tilde{x}^k, y^k) + \nabla_x h_1(\tilde{x}^k) - A^\top\tilde{\lambda}^k + G(\tilde{x}^k - x^k)\right] \geq 0 \quad \forall x \in \mathcal{X}, \tag{58}$$

and

$$(y - \tilde{y}^k)^\top \left[\nabla_y f(\tilde{x}^k, y^k) + g_2(\tilde{y}^k) - B^\top\tilde{\lambda}^k + \gamma B^\top B(\tilde{y}^k - y^k) + H(\tilde{y}^k - y^k)\right] \geq 0 \quad \forall y \in \mathcal{Y}. \tag{59}$$

Moreover,

$$\begin{aligned} & (x - \tilde{x}^k)^\top \nabla_x f(\tilde{x}^k, y^k) + (y - \tilde{y}^k)^\top \nabla_y f(\tilde{x}^k, y^k) \\ =\ & (x - \tilde{x}^k)^\top \nabla_x f(\tilde{x}^k, y^k) + (y - y^k)^\top \nabla_y f(\tilde{x}^k, y^k) + (y^k - \tilde{y}^k)^\top \nabla_y f(\tilde{x}^k, y^k) \\ \leq\ & f(x, y) - f(\tilde{x}^k, y^k) - (\tilde{y}^k - y^k)^\top \nabla_y f(\tilde{x}^k, y^k) \\ \overset{(5)}{\leq}\ & f(x, y) - f(\tilde{x}^k, \tilde{y}^k) + \frac{L}{2}\|y^k - \tilde{y}^k\|^2. \end{aligned} \tag{60}$$

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

$$(\lambda - \tilde{\lambda}^k)^\top (A\tilde{x}^k + B\tilde{y}^k - b) = (\lambda - \tilde{\lambda}^k)^\top \left(-B(y^k - \tilde{y}^k) + \frac{1}{\gamma}(\lambda^k - \tilde{\lambda}^k) \right). \tag{61}$$

By the convexity of $h_1(x)$ and $h_2(y)$, combining (61), (60), (59), and (58), we have

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

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

By similar derivations as in the proofs for Proposition A.5, (57) follows. □

With Proposition A.7 in place, we can prove Theorem 2.5 similarly as in the proof of Theorem 2.2. 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:

$$\|\nabla_y h_2(y_2) - \nabla_y h_2(y_1)\| \leq L\|y_2 - y_1\|, \ \forall\, y_1, y_2 \in \mathcal{Y}. \tag{62}$$

Proposition A.8 *Let $\{\tilde{w}^k\}$ be defined by (10), and the matrices Q, M, P be given in (9), and $H := \frac{1}{\alpha} I_q - \gamma B^\top B \succeq 0$. For any $w \in \Omega$, we have*

$$\begin{aligned} & h(u) - h(\tilde{u}^k) + (w - \tilde{w}^k)^\top F(\tilde{w}^k) \\ \geq\ & (w - \tilde{w}^k)^\top Q(w^k - \tilde{w}^k) - \left(L\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k) \right) \\ =\ & \frac{1}{2}\left(\|w - w^{k+1}\|_M^2 - \|w - w^k\|_M^2 \right) + \frac{1}{2}\|x^k - \tilde{x}^k\|_G^2 + \frac{1}{2\gamma}\|\lambda^k - \tilde{\lambda}^k\|^2 \\ & - \left(L\|y^k - \tilde{y}^k\|^2 + (y - \tilde{y}^k)^\top H(\tilde{y}^k - y^k) \right). \end{aligned} \tag{63}$$

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

$$\begin{aligned} & (x - x^{k+1})^\top \left[\nabla_x f(x^{k+1}, y^k) + h_1'(x^{k+1}) - A^\top(\lambda^k - \gamma(Ax^{k+1} + By^k - b)) + G(x^{k+1} - x^k) \right] \\ \geq\ & 0, \qquad \forall x \in \mathcal{X} \end{aligned}$$

and

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

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

$$(x - \tilde{x}^k)^\top \left[\nabla_x f(\tilde{x}^k, y^k) + h_1'(\tilde{x}^k) - A^\top \tilde{\lambda}^k + G(\tilde{x}^k - x^k)\right] \geq 0 \quad \forall x \in \mathcal{X}, \tag{64}$$

and

$$(y - \tilde{y}^k)^\top \left[\nabla_y f(\tilde{x}^k, y^k) + \nabla_y h_2(y^k) - B^\top \tilde{\lambda}^k + \frac{1}{\alpha}(\tilde{y}^k - y^k)\right] \geq 0 \quad \forall y \in \mathcal{Y}. \tag{65}$$

Therefore,

$$\begin{aligned}
& (x - \tilde{x}^k)^\top \nabla_x f(\tilde{x}^k, y^k) + (y - \tilde{y}^k)^\top \nabla_y f(\tilde{x}^k, y^k) \\
= \; & (x - \tilde{x}^k)^\top \nabla_x f(\tilde{x}^k, y^k) + (y - y^k)^\top \nabla_y f(\tilde{x}^k, y^k) + (y^k - \tilde{y}^k)^\top \nabla_y f(\tilde{x}^k, y^k) \\
\leq \; & f(x, y) - f(\tilde{x}^k, y^k) - (\tilde{y}^k - y^k)^\top \nabla_y f(\tilde{x}^k, y^k) \\
\leq \; & f(x, y) - f(\tilde{x}^k, \tilde{y}^k) + \frac{L}{2}\|y^k - \tilde{y}^k\|^2.
\end{aligned} \tag{66}$$

Moreover, by (6), we have

$$(y - \tilde{y}^k)^\top \nabla_y h_2(y^k) \leq h_2(y) - h_2(\tilde{y}^k) + \frac{L}{2}\|y^k - \tilde{y}^k\|^2. \tag{67}$$

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

$$(\lambda - \tilde{\lambda}^k)^\top (A\tilde{x}^k + B\tilde{y}^k - b) = (\lambda - \tilde{\lambda}^k)^\top \left(-B(y^k - \tilde{y}^k) + \frac{1}{\gamma}(\lambda^k - \tilde{\lambda}^k)\right). \tag{68}$$

By the convexity of $h_1(x)$, combining (68), (67), (66), (65), (64), and noticing $H := \frac{1}{\alpha} I_q - \gamma B^\top B$ for any $w \in \Omega$ and $\tilde{w}^k$ we have

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

As a result, (63) follows. □

The proof of Theorem 2.6 follows a similar line of derivation as in the proof of Theorem 2.2, and we omit the details here.