# ON THE CONVERGENCE RATE OF THE BI-ALTERNATING DIRECTION METHOD OF MULTIPLIERS

*Guoqiang Zhang, Richard Heusdens and W. B. Kleijn*

Department of Intelligent Systems
Delft University of Technology
Delft, the Netherlands Email: {g.zhang-1,r.heusdens,w.b.kleijn}@tudelft.nl

**ABSTRACT**

In this paper, we analyze the convergence rate of the bi-alternating direction method of multipliers (BiADMM). Differently from ADMM that optimizes an augmented Lagrangian function, BiADMM optimizes an augmented primal-dual Lagrangian function. The new function involves both the objective functions and their conjugates, thus incorporating more information of the objective functions than the augmented Lagrangian used in ADMM. We show that BiADMM has a convergence rate of $\mathcal{O}(K^{-1})$ ($K$ denotes the number of iterations) for general convex functions. We consider the lasso problem as an example application. Our experimetal results show that BiADMM outperforms not only ADMM, but fast-ADMM as well.

***Index Terms*—** Distributed optimization, alternating direction method of multipliers, bi-alternating direction of multipliers

## 1. INTRODUCTION

Consider a decomposable optimization problem with a linear equality constraint

$$\min_{x,z} f(x) + g(z) \quad \text{subject to} \quad Ax + Bz = c, \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R} \bigcup \{\infty\}$ and $g : \mathbb{R}^m \to \mathbb{R} \bigcup \{\infty\}$ are closed, proper and convex functions and $(A, B, c) \in (\mathbb{R}^{q\times n}, \mathbb{R}^{q\times m}, \mathbb{R}^q)$. Optimization of the above problem has received considerable attention in computer science and engineering [1]. Typical applications that involve (1) include network resource allocation [2], compressive sensing [3], channel coding [4] and distributed computation in sensor networks [5]. The main research challenge is how to reach the optimal solution of (1) efficiently by exploiting the decomposable structure of the objective function.

In the literature, the dual-ascent method, proposed in the mid-1960s [6, 7, 8], is a classic approach for solving (1). The method iteratively approaches the saddle point of the Lagrangian function by alternating updates of the primal variables $(x, z)$ and the Lagrange multipliers (dual variables). However, the convergence of the dual-ascent method requires strong assumptions on the objective function [9, 1] like strong convexity of $f(x)$ and $g(z)$, making it less useful in practical applications.

The method of multipliers was introduced to bring in robustness to the dual ascent algorithm. The method of multipliers optimizes an augmented Lagrangian function where a quadratic penalty function is introduced. The introduction of the penalty function, however, prevents the method for parallel updates of the primal variables. ADMM solves this problem by alternately updating the primal variables in a Gauss-Seidel procedure [10, 11]. The convergence analysis of ADMM has been studied extensively in a series of papers [12, 13, 14, 15]. It was found that ADMM is guaranteed to convergence under very mild conditions. A thorough review on ADMM has been provided in [1] by Boyd et al. In the last few years, research interest has moved to find out the convergence rates of ADMM for objective functions with different functional properties (e.g., strongly convex or not) [16, 17].

In [18], we have proposed the bi-alternating direction method multipliers (BiADMM). Compared to ADMM, the new method optimizes a different constructed function that involves both $(f(x), g(z))$ and their conjugates [19]. Our main motivation was to have the function carry more information about $(f(x), g(z))$ than the augmented Lagrangian function does for ADMM, and therefore make BiADMM more efficient. We note that in [18], the optimal solution of (1) is found by minimizing the newly constructed function. Later on, we noticed that such a construction makes it difficult to characterize the convergence rate of the algorithm. In this paper, we construct the function in a different way in order to facilitate the convergence-rate analysis, which we refer to as the *augmented primal-dual Lagrangian function*. In particular, the new function is constructed such that the optimal solution of (1) is computed by reaching a saddle point.

In this work, we first construct the augmented primal-dual Lagrangian function. After that we analyze the convergence rate of BiADMM for the newly constructed function. We show that for closed, proper and convex functions, BiADMM has a convergence rate of $\mathcal{O}(K^{-1})$, where $K$ represents the number of iterations. We then apply BiADMM to the lasso problem to test its efficiency. Experimental results show that for the lasso problem, BiADMM outperforms both ADMM and fast-ADMM considerably.

## 2. BI-ALTERNATING DIRECTION OF MULTIPLIERS

In this section, we first construct the augmented primal-dual Lagrangian function for (1). Similarly to that of [18], BiADMM follows directly from optimizing the new function.

### 2.1. Constructing augmented bi-conjugate function

We consider the problem (1) where the two functions $f(x)$ and $g(z)$ are closed, proper and convex. The Lagrangian function associated with (1) is defined by

$$L_p(x, z, \delta) = f(x) + g(z) + \delta^T(c - Ax - Bz),$$

This work was supported by the COMMIT program, The Netherlands.

where $\delta$ is a Lagrangian multiplier (dual variable) and the subscript $p$ indicates that $L_p$ is the Lagrangian of the primal problem. The Lagrangian function is a convex function of $(x, z)$ for fixed $\delta$, and a concave function of $\delta$ for fixed $(x, z)$. Throughout the rest of the paper, we will make the following (common) assumption:

**Assumption 1.** *There exists a saddle point $(x^*, z^*, \delta^*)$ to the Lagrangian function $L_p(x, z, \delta)$ such that for all $(x, z) \in (\mathbb{R}^n, \mathbb{R}^m)$ and $\delta \in \mathbb{R}^q$ we have*

$$L_p(x^*, z^*, \delta) \leq L_p(x^*, z^*, \delta^*) \leq L_p(x, z, \delta^*).$$

The Lagrangian dual problem associated with the primal problem (1) can be expressed as

$$\max_{\delta} \ -f^*(A^T\delta) - g^*(B^T\delta) + \delta^T c, \tag{2}$$

where $f^*(\cdot), g^*(\cdot)$, are the conjugate functions of $f(\cdot), g(\cdot)$, respectively, satisfying Frenchel's inequalities

$$f(x) + f^*(A^T\lambda) \geq \lambda^T Ax \quad \text{for all} \ \ x, \lambda, \tag{3a}$$

$$g(z) + g^*(B^T\delta) \geq \delta^T Bz \quad \text{for all} \ \ z, \delta. \tag{3b}$$

In order to decouple the joint optimization of the two conjugate functions, we introduce an auxiliary variable $\lambda$ and reformulate the dual problem as

$$\max_{\delta,\lambda} \ -f^*(A^T\lambda) - g^*(B^T\delta) + \lambda^T c \quad \text{subject to} \quad \lambda = \delta. \tag{4}$$

We can construct a Lagrangian function for the dual problem (4), which takes the form

$$L_d(y, \delta, \lambda) = -f^*(A^T\lambda) - g^*(B^T\delta) + \lambda^T c + y^T(\delta - \lambda),$$

where the Lagrange multiplier $y = Bz$, which follows from the fact that at a saddle point of $L_d$ we have $0 \in \partial_\delta L_d(z^*, \delta^*, \lambda^*) = -\partial_\delta g^*(B^T\delta^*) + y^*$. On the other hand, Frenchel's inequality (3b) must hold with equality so that $0 \in \partial_\delta g^*(B^T\delta^*) - Bz^*$. Note that $L_d(z, \delta, \lambda)$ is convex in $z$ for fixed $(\delta, \lambda)$, and concave in $(\delta, \lambda)$ for fixed $z$.

Given the primal and dual Lagrangian, we define the *augmented primal-dual Lagrangian* function as

$$\begin{aligned} \mathcal{L}_\rho(x, z, \delta, \lambda) &= L_p(x, z, \delta) + L_d(z, \delta, \lambda) + h_\rho(x, z, \delta, \lambda) \\ &= f(x) + g(z) - f^*(A^T\lambda) - g^*(B^T\delta) \\ &\quad + \delta^T(c - Ax) + \lambda^T(c - Bz) + h_\rho(x, z, \delta, \lambda), \end{aligned} \tag{5}$$

where

$$h_\rho(x, z, \delta, \lambda) = \frac{\rho}{2}\|c - Ax - Bz\|^2 - \frac{1}{2\rho}\|\lambda - \delta\|^2,$$

where the parameter $\rho > 0$. The quadratic function $h_\rho(x, z, \delta, \lambda)$ is imposed in (5) in order to implicitly enforce the equality constraints described in (1) and (4). The particular arrangement of the parameter $\rho$ in $h_\rho(x, z, \delta, \lambda)$ facilities the convergence analysis (see Section 3). The function $\mathcal{L}_\rho(x, z, \delta, \lambda)$ is convex in $(x, z)$ for $(\delta, \lambda)$ fixed, and concave in $(\delta, \lambda)$ for $(x, z)$ fixed.

Similar to $L_p$, we have a saddle point theorem for $\mathcal{L}_\rho$ which states that $(x^*, z^*)$ solves the primal problem if and only if $(x^*, z^*, \delta^*, \lambda^*)$ is a saddle point of $\mathcal{L}_\rho(x, z, \delta, \lambda)$. To prove this result, we need the following lemma.

**Lemma 1.** *If $(x^*, z^*, \delta^*)$ is a saddle point of $L_p(x, z, \delta)$, then $(z^*, \delta^*, \delta^*)$ is a saddle point of $L_d(z, \delta, \lambda)$.*

*Proof.* If $(x^*, z^*, \delta^*)$ is a saddle point of $L_p(x, z, \delta)$, $(x^*, z^*)$ solves the primal problem and $\delta^*$ the dual problem, so that $\lambda^* = \delta^*$. □

**Theorem 1** (Saddle point theorem)**.** *If $(x^*, z^*)$ solves the primal problem, $\exists(\delta^*, \lambda^*)$ such that $(x^*, z^*, \delta^*, \lambda^*)$ is a saddle point of $\mathcal{L}(x, z, \delta, \lambda)$. Conversely, if $(x^*, z^*, \delta^*, \lambda^*)$ is a saddle point of $\mathcal{L}(x, z, \delta, \lambda)$, then $(x^*, z^*)$ solves the primal problem.*

*Proof.* If $(x^*, z^*)$ solves the primal problem, then there exists $\delta^*$ such that $(x^*, z^*, \delta^*)$ is a saddle point of $L_p(x, z, \delta)$ and thus $(z^*, \delta^*, \delta^*)$ a saddle point of $L_d(z, \delta, \lambda)$ by Lemma 1. Hence we have

$$\begin{aligned} &\mathcal{L}_\rho(x^*, z^*, \delta, \lambda) \\ &= L_p(x^*, z^*, \delta) + L_d(z^*, \delta, \lambda) + h_\rho(x^*, z^*, \delta, \lambda) \\ &\leq L_p(x^*, z^*, \delta^*) + L_d(z^*, \delta^*, \lambda^*) + h_\rho(x^*, z^*, \delta^*, \lambda^*) \\ &= \mathcal{L}_\rho(x^*, z^*, \delta^*, \lambda^*) \\ &\leq L_p(x, z, \delta^*) + L_d(z, \delta^*, \lambda^*) + h_\rho(x, z, \delta^*, \lambda^*) \\ &= \mathcal{L}_\rho(x, z, \delta^*, \lambda^*). \end{aligned}$$

Conversely, suppose $(x^*, z^*, \delta^*)$ is a saddle point of $L_p(x, z, \delta)$. Firstly, we use the saddle point $(x^*, z^*, \delta^*)$ to show that any point $(\hat{x}, \hat{z}, \hat{\delta}, \hat{\lambda})$ such that $A\hat{x} + B\hat{z} \neq c$ or $\hat{\delta} \neq \hat{\lambda}$ is not a saddle point of $\mathcal{L}_\rho$. This is because for the considered point, at least one of the following two strict inequality holds due to the function $h_\rho$:

$$\mathcal{L}_\rho(x^*, z^*, \hat{\delta}, \hat{\lambda}) < \mathcal{L}_\rho(x^*, z^*, \delta^*, \delta^*)$$

$$\mathcal{L}_\rho(\hat{x}, \hat{z}, \delta^*, \delta^*) > \mathcal{L}_\rho(x^*, z^*, \delta^*, \delta^*)$$

Based on the above result, we conclude that the optimality conditions for $(x^*, z^*, \delta^*, \lambda^*)$ being a saddle point of $\mathcal{L}_\rho$ are given by $\lambda^* = \delta^*, Ax^* + Bz^* = c, 0 \in \partial_x \mathcal{L}_\rho(x^*, z^*, \delta^*, \lambda^*) = \partial_x f(x^*) - A^T\lambda^* = \partial_x L_p(x^*, z^*, \delta^*)$, and $0 \in \partial_z \mathcal{L}_\rho(x^*, z^*, \delta^*, \lambda^*) = \partial_z g(z^*) - B^T\delta^* = \partial_z L_p(x^*, z^*, \delta^*)$, from which we conclude that $(x^*, z^*, \delta^*)$ is a saddle point of $L_p(x, z, \delta)$, and thus $(x^*, z^*)$ solves the primal problem. □

**Remark 1.** *Intuitively speaking, due to the presence of the conjugate functions $(f^*(\cdot), g^*(\cdot))$, $\mathcal{L}_\rho$ carries more information about the functions $(f(\cdot), g(\cdot))$ than the original augmented Lagrangian does for ADMM. As a result, if the parameter $\rho$ is set properly, BiADMM should converge faster than ADMM. The experimental results in Section 4 confirm this conjecture.*

### 2.2. Alternating optimization

Given the augmented primal-dual Lagrangian $\mathcal{L}_\rho$, we introduce our BiADMM in the following. The procedure is similar to our earlier work presented in [18].

For notational convenience, let $w = (x^T, z^T, \delta^T, \lambda^T)^T$, and we will refer to the augmented primal-dual Lagrangian as $\mathcal{L}_\rho(w)$. We optimize $\mathcal{L}_\rho(w)$ by performing a Gauss-Seidel iteration. Each time we optimize the function over some variables in $w$ while keeping all the others fixed. After each iteration, every variable receives a new estimate. Note that fixing $(z, \delta)$ (or equivalently, $(x, \lambda)$), the function $\mathcal{L}_\rho(w)$ is decoupled w.r.t. $x$ and $\lambda$ (or equivalently, $z$ and $\delta$)). One natural scheme for updating the estimates at iteration $k + 1$

is, therefore,

$$(\hat{x}_{k+1}, \hat{\lambda}_{k+1}) = \arg\min_x \max_\lambda \mathcal{L}_\rho(x, \hat{z}_k, \hat{\delta}_k, \lambda) \quad (6a)$$

$$(\hat{z}_{k+1}, \hat{\delta}_{k+1}) = \arg\min_z \max_\delta \mathcal{L}_\rho(\hat{x}_{k+1}, z, \delta, \hat{\lambda}_{k+1}). \quad (6b)$$

At iteration $k+1$, we denote $\hat{w}_{k+\frac{1}{2}} = (\hat{x}_{k+1}^T, \hat{z}_k^T, \hat{\delta}_k^T, \hat{\lambda}_{k+1}^T)^T$ and $\hat{w}_{k+1} = (\hat{x}_{k+1}^T, \hat{z}_{k+1}^T, \hat{\delta}_{k+1}^T, \hat{\lambda}_{k+1}^T)^T$. The quantity $\hat{w}_{k+\frac{1}{2}}$ represents an intermediate estimate of $w^*$ at iteration $k+1$. In addition, we consider designing the stopping criterion for the iterates (6a)-(6b). To do so, we define the objective function

$$p(w) = f(x) + g(z) + f^*(A^T\lambda) + g^*(B^T\delta) - \lambda^T c.$$

One can easily show that $p(w^*) = 0$.

### 2.3. Comparison to fast-ADMM

In this subsection, we briefly discuss the fast-ADMM, first proposed in [20] (which was originally named as *Symmetric Alternating Direction Augmented Lagrangian Method*). Our main motivation is to point out the relationship between BiADMM and fast-ADMM.

The augmented Lagrangian function for the primal problem (1) takes form of [20]

$$L_{p,\rho}(x, z, \delta) = L_p(x, z, \delta) + \frac{\rho}{2}\|c - Ax - Bz\|^2, \quad (7)$$

where $\rho > 0$. Given (7), fast-ADMM updates the estimate $(\hat{x}_{k+1}, \hat{z}_{k+1}, \hat{\delta}_{k+1})$ at iteration $k+1$ as follows [20]:

$$\hat{x}_{k+1} = \arg\min_x(L_{p,\rho}(x, \hat{z}_k, \hat{\delta}_k)) \quad (8a)$$

$$\hat{\delta}_{k+\frac{1}{2}} = \hat{\delta}_k + \rho(M\hat{x}_{k+1} - \hat{z}_k) \quad (8b)$$

$$\hat{z}_{k+1} = \arg\min_z(L_{p,\rho}(\hat{x}_{k+1}, z, \hat{\delta}_{k+\frac{1}{2}})) \quad (8c)$$

$$\hat{\delta}_{k+1} = \hat{\delta}_{k+\frac{1}{2}} + \rho(M\hat{x}_{k+1} - \hat{z}_{k+1}). \quad (8d)$$

As opposed to the updates (8a)-(8d), ADMM does not have the intermediate update (8b) for $\delta$. Instead, $\hat{\delta}_{k+1}$ is computed only after both $\hat{x}_{k+1}$ and $\hat{z}_{k+1}$ are computed. Since fast-ADMM captures more recent information of $\hat{x}$ and $\hat{z}$, it naturally accelerates ADMM.

By inspection of the updates for BiADMM and fast-ADMM, we conclude that both methods involve four computations at each iteration. The update (8b) corresponds to the computation of $\hat{\lambda}$ in (6a). Note that with (fast)ADMM the $\delta$ update is a gradient-ascent step, whereas with BiADMM the $\delta$ and $\lambda$ updates are obtained by coordinate ascent.

## 3. CONVERGENCE ANALYSIS

In this section, we show that BiADMM has a convergence rate of $\mathcal{O}(1/K)$ for general closed, proper and convex functions. The main mathematical tool that we will use in our proof is the variational inequality (VI), which is widely applied in the convergence analysis of ADMM [17, 21]. We have the following result.

**Theorem 2.** *Define* $F^T(w) = \left(-\delta^T A, -\lambda^T B, (Ax - c)^T, (Bz)^T\right)$. *Let* $\bar{w}_K = \frac{1}{K}\sum_{k=1}^K \hat{w}_k$. *We have*

$$0 \le p(\bar{w}_K) + (\bar{w}_K - w^*)^T F(\bar{w}_K) \le \mathcal{O}(K^{-1}). \quad (9)$$

In order to prove this result, we need the VI corresponding to (5), which we present in the lemma below.

**Lemma 2.** *Let* $w^* = (x^*, z^*, \delta^*, \lambda^*)$ *denote a saddle point of* $\mathcal{L}_\rho(w)$. *Then*

$$p(w) + (w - w^*)^T F(w) \ge 0,$$

*where equality holds if and only if*

$$\begin{array}{l} 0 \in \partial_x f(x) - A^T\lambda^* \\ 0 \in \partial_z g(z) - B^T\delta^* \\ 0 \in \partial_\lambda f^*(A^T\lambda) - Ax^* \\ 0 \in \partial_\delta g^*(B^T\delta) - Bz^* \end{array}. \quad (10)$$

*Proof.* Given $w^*$, we have

$$\begin{aligned} &p(w) + (w - w^*)^T F(w) \\ &\quad = f(x) + g(z) + f^*(A^T\lambda) + g^*(B^T\delta) + {\delta^*}^T c \\ &\qquad - {\delta^*}^T Ax - {\lambda^*}^T Bz - \delta^T(c - Ax^*) - \lambda^T(c - Bz^*) \\ &\quad = f(x) + g(z) + f^*(A^T\lambda) + g^*(B^T\delta) + {\delta^*}^T c \\ &\qquad - {\lambda^*}^T Ax - {\delta^*}^T Bz - \lambda^T Ax^* - \delta^T Bz^*, \end{aligned} \quad (11)$$

where the last equality holds since $Ax^* + Bz^* = c$ and $\delta^* = \lambda^*$. Using Frenchel's inequalities (3a) and (3b), we conclude that

$$\begin{aligned} -{\lambda^*}^T Ax &\ge -f(x) - f^*(A^T\lambda^*), \\ -{\delta^*}^T Bz &\ge -g(z) - g^*(B^T\delta^*), \\ -\lambda^T Ax^* &\ge -f(x^*) - f^*(A^T\lambda), \\ -\delta^T Bz^* &\ge -g(z^*) - g^*(B^T\delta), \end{aligned} \quad (12)$$

from which we conclude, using (11), that

$$p(w) + (w - w^*)^T F(w) \ge -p(w^*) = 0,$$

where equality holds if and only if we have equality in (12) and thus if and only if (10) holds. □

We are now in the position to prove Theorem 2.

*Proof of Theorem 2.* From (6a)-(6b), the VIs for $\hat{w}_{k+1}$ are given by: $\forall w \in \mathbb{R}^{n+m+2q}$

$$\begin{aligned} 0 \le\ & f(x) - f(\hat{x}_{k+1}) \\ & - \left(\hat{\delta}_k + \rho(c - A\hat{x}_{k+1} - B\hat{z}_k)\right)^T A(x - \hat{x}_{k+1}) \end{aligned} \quad (13a)$$

$$\begin{aligned} 0 \le\ & g(z) - g(\hat{z}_{k+1}) \\ & - \left(\hat{\lambda}_{k+1} + \rho(c - A\hat{x}_{k+1} - B\hat{z}_{k+1})\right)^T B(z - \hat{z}_{k+1}) \end{aligned} \quad (13b)$$

$$\begin{aligned} 0 \le\ & g^*(B^T\delta) - g^*(B^T\hat{\delta}_{k+1}) \\ & - \left(c - A\hat{x}_{k+1} + (1/\rho)(\hat{\lambda}_{k+1} - \hat{\delta}_{k+1})\right)^T (\delta - \hat{\delta}_{k+1}) \end{aligned} \quad (13c)$$

$$\begin{aligned} 0 \le\ & f^*(A^T\lambda) - f^*(A^T\hat{\lambda}_{k+1}) \\ & - \left(c - B\hat{z}_k - (1/\rho)(\hat{\lambda}_{k+1} - \hat{\delta}_k)\right)^T (\lambda - \hat{\lambda}_{k+1}). \end{aligned} \quad (13d)$$

Adding (13a)-(13c) and substituting $w = w^*$ yields

$$\begin{aligned}
&p(\hat{w}_{k+1}) - p(w^*) + (\hat{w}_{k+1} - w^*)^T F(\hat{w}_{k+1}) \\
&\quad \leq \frac{1}{\rho}\left(\rho A x^* + \lambda^* - (\rho A\hat{x}_{k+1} + \hat{\lambda}_{k+1})\right)^T \\
&\quad\quad \cdot \left((\rho B\hat{z}_k - \hat{\delta}_k - \rho c) - (\rho B\hat{z}_{k+1} - \hat{\delta}_{k+1} - \rho c)\right) \\
&\quad\quad - \rho\|A\hat{x}_{k+1} + B\hat{z}_{k+1} - c\|^2 - (1/\rho)\|\hat{\lambda}_{k+1} - \hat{\delta}_{k+1}\|^2 \\
&\quad = \frac{1}{2\rho}\|\rho(Ax^* + B\hat{z}_k - c) + (\lambda^* - \hat{\delta}_k)\|^2 \\
&\quad\quad - \frac{1}{2\rho}\|\rho(Ax^* + B\hat{z}_{k+1} - c) + (\lambda^* - \hat{\delta}_{k+1})\|^2 \\
&\quad\quad - \frac{1}{2\rho}\|\rho(A\hat{x}_{k+1} + B\hat{z}_{k+1} - c) - (\hat{\lambda}_{k+1} - \hat{\delta}_{k+1})\|^2 \\
&\quad\quad - \frac{1}{2\rho}\|\rho(A\hat{x}_{k+1} + B\hat{z}_k - c) + (\hat{\lambda}_{k+1} - \hat{\delta}_k)\|^2. \qquad (14)
\end{aligned}$$

Since both $p(w)$ and $(w - w^*)F(w)$ are convex functions of $w$, summing (14) over $k$ and applying Jensen's inequality yields

$$\begin{aligned}
p(\bar{w}_K) - p(w^*) &+ (\bar{w}_K - w^*)^T F(\bar{w}_K) \\
&\leq \frac{1}{2\rho K}\|\rho(x^* + B\hat{z}_0 - c) + (\lambda^* - \hat{\delta}_0)\|_2^2 \\
&= \mathcal{O}(K^{-1}).
\end{aligned}$$

□

## 4. APPLICATION TO LASSO PROBLEM

In this section, we consider solving the lasso problem [22] by using BiADMM. This example confirms that BiADMM is more efficient than ADMM and fast-ADMM.

The lasso problem originates from bioinformatics and machine learning, and can be expressed as

$$\max_{\delta,\lambda}\left(-\frac{1}{2}\|M\lambda - b\|_2^2 - \alpha\|\delta\|_1\right) \quad \text{subject to} \quad \lambda = \delta, \qquad (15)$$

where $M$ is a $n \times q$ matrix ($q > n$), and $\alpha > 0$ is a regularization parameter. Note that the above problem formulation is of the form (2). The corresponding dual problem can be formulated as

$$\min_{x,z}\left(\frac{1}{2}\|x\|_2^2 + b^T x + I_S(z)\right) \quad \text{subject to} \quad M^T x = z, \qquad (16)$$

where $I_S(z)$ is the the indicator function on $S = \{|z| \preceq \alpha\}$ and is the conjugate function of $\alpha\|\delta\|_1$. The symbol $\preceq$ denotes componentwise inequality. In practice, one can either solve (15) or (16) by using ADMM or fast-ADMM.

### 4.1. Experimental results

In the experiment, we set $(n, q) = (60, 100)$ and $\alpha = 1.1$. The elements in $(M, b)$ were generated randomly from a normal Gaussian distribution. Both ADMM and fast-ADMM were applied to the minimization problem (16). We mainly investigated the number of iterations needed for each algorithm under a particular error criterion.

The convergence results are displayed in Figure 1. Each point in the figure for a particular $\rho$ is obtained by averaging over 200 realizations of $(M, b)$. For each realization of $(M, b)$, all the three algorithms share the same initialization of $\hat{w}_0$. In order to terminate the iterations, we define an error criterion at iteration $k+1$ as

$$\epsilon_{k+1} = \frac{1}{2}\left(|p(\hat{w}_{k+\frac{1}{2}})| + |p(\hat{w}_{k+1})|\right).$$

Hence, convergence of BiADMM implies $\epsilon_k = 0$ as $k \to \infty$. In particular, we set the threshold for $\epsilon_k \leq 10^{-5}$ for stopping the algorithms. For ADMM and fast-ADMM, the component $\hat{\lambda}$ in $\hat{w}$ was replaced with $\hat{\delta}$ when computing $\epsilon_k$.

By inspection of Figure 1, we conclude that BiADMM converges faster than ADMM and fast-ADMM on average. This phenomenon may be due to the fact that the augmented primal-dual Lagrangian function $\mathcal{L}_\rho(w)$ is more informative about $(f(\cdot), g(\cdot))$ than the augmented primal Lagrangian function $L_{p,\rho}(x, z, \delta)$, making BiADMM more efficient. Further, one observes that the parameter $\rho$ has a big impact on the convergence speeds of the three algorithms. The optimal $\rho$ values are roughly the same for the three algorithms, which is around $\rho = 0.12$. This suggests that in practice, the parameter $\rho$ has to be set properly to gain convergence efficiency.

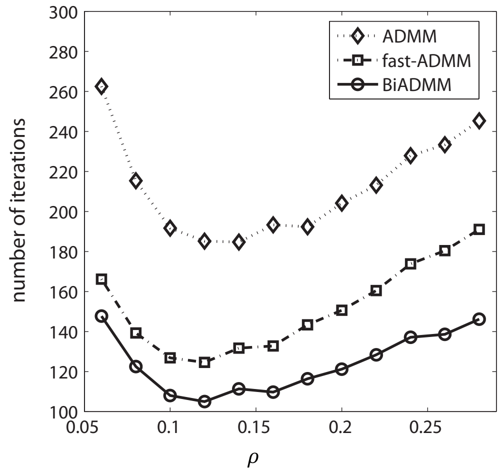


**Fig. 1**. Convergence comparison of ADMM, fast-ADMM and Bi-AMM for $0.06 \leq \rho \leq 0.28$.

## 5. CONCLUSION

In this paper, we have analyzed the convergence rate of BiADMM. To facilitate the analysis, we construct the augmented primal-dual Lagrangian function. We have shown that for general closed, proper and convex functions, BiADMM possesses a convergence rate of $\mathcal{O}(K^{-1})$. Experimental results demonstrate that BiADMM outperforms both ADMM and fast-ADMM for the lasso problem.

## 6. REFERENCES

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *In Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[2] D. P. Palomar and M. Chiang, "A Tutorial on Decomposition Methods for Network Utility Maximization," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, 2006.

[3] Joo F. C. Mota, Joo M. F. Xavier, P. M. Q. Aguiar, and M. Pschel, "Distributed Basis Pursuit," *IEEE Trans. on Signal Processing*, vol. 60, no. 4, pp. 1942–1956, 2012.

[4] S. Barman and X. Liu and S. Draper and B. Recht, "Decomposition Methods for Large Scale LP Decoding," arXiv:1204.0556 [cs.IT], 2012.

[5] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized Gossip Algorithms," *In Foundations and Trends in Machine Learning*, vol. 52, no. 6, pp. 2508–530, 2006.

[6] G. B. Dantzig., *Linear Programming and Extensions*, RAND Corporation, 1963.

[7] H. Everett, "Generalized Lagrange Multiplier Method for Solving Problems of Optimum Allocation of Resources," *Operations Research*, vol. 11, no. 3, pp. 399417, 1963.

[8] J. F. Benders, "Partitioning Procedures for Solving Mixed-Variables Programming Problems," *Numerische Mathematik*, p. 4:238252, 1962.

[9] D. P. Bertsekas and J. N. Tsitsikis, *Parallel and Distributed Computation: Numerical Methods*, Belmont, MA: Athena Scientific, 1997.

[10] R. Glowinski and A. Marrocco, "Sur lapproximation par elements finis dordre un, et la resolution par penalisation-dualite dune classe de problems de Dirichlet nonlineaires," *Revue Franccaise d'Automatique, Informatique, et Recherche Operationelle*, vol. 9, pp. 41–76, 1975.

[11] D. Gabay and B. Mercier, "A Dual Algorithm for the Solution of Nonlinear Variational Problems via Finite-Element Approximations," *Computers and Mathematics with Applications*, vol. 2, pp. 17–40, 1976.

[12] D. Gabay, "Applications of the Method of Multipliers to Variational Inequalities," *in M. Fortin and R. Glowinski, editors, Augmented Lagrangian Methods: Applications to the Solution of Boundary-Value Problems*, 1983.

[13] P. Tseng, "Applications of a Splitting Algorithm to Decomposition in Convex Programming and Variational Inequalities," *SIAM Journal on Control and Optimization*, vol. 29, no. 1, pp. 119–138, 1991.

[14] R. Glowinski and P. Le Tallec, "Augmented Lagrangian Methods for the Solution of Variational Problems," Technical Report 2965, 1987.

[15] J. Eckstein and M. Fukushima, "Some Reformulations and Applications of the Alternating Direction Method of Multipliers," *Large Scale Optimization: State of the Art*, pp. 119–138, 1993.

[16] X. M. Yuan B. S. He and X. M. Xu, "On the O(1/n) Convergence Rate of Douglas–Rachford Alternating Direction Method," *SIAM J. Numer. Anal.*, vol. 50, no. 2, pp. 700709, 2012.

[17] H. Wang and A. Banerjee, "Online Alternating Direction Method," in *Proc. International Conference on Machine Learning*, June 2012.

[18] G. Zhang and R. Heusdens, "Bi-Alternating Direction of Multipliers," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013, pp. 3317–3321.

[19] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[20] D. Goldfard, S. Ma, and K. Scheinberg, "Fast Alternating Linearization Methods for Minmiming the Sum of Two Convex Functions," *Math. Progrm.*, pp. 141:349–382, 2013.

[21] W. Deng and W. Yin, "On the Global and Linear Convergence of the Generalized Alternating Direction Method of Multipliers," Technical Report TR12-14, Rice University CAAM, 2012.

[22] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, vol. 58, no. 1, pp. 267288, 1996.