Parallel Direction Method of Multipliers

Huahua Wang, Arindam Banerjee, Zhi-Quan Luo
University of Minnesota, Twin Cities
{huwang,banerjee}@cs.umn.edu, luozq@umn.edu

Abstract

We consider the problem of minimizing block-separable (non-smooth) convex functions subject to linear constraints. While the Alternating Direction Method of Multipliers (ADMM) for two-block linear constraints has been intensively studied both theoretically and empirically, in spite of some preliminary work, effective generalizations of ADMM to multiple blocks is still unclear. In this paper, we propose a parallel randomized block coordinate method named Parallel Direction Method of Multipliers (PDMM) to solve optimization problems with multi-block linear constraints. At each iteration, PDMM randomly updates some blocks in parallel, behaving like parallel randomized block coordinate descent. We establish the global convergence and the iteration complexity for PDMM with constant step size. We also show that PDMM can do randomized block coordinate descent on overlapping blocks. Experimental results show that PDMM performs better than state-of-the-art methods in two applications, robust principal component analysis and overlapping group lasso.

1 Introduction

In this paper, we consider the minimization of block-separable convex functions subject to linear constraints, with a canonical form:

$$\min_{\{x_j \in X_j\}} f(x) = \sum_{j=1}^{J} f_j(x_j), \text{ s.t. } Ax = \sum_{j=1}^{J} A^c_j x_j = a,$$

where the objective function $f(x)$ is a sum of $J$ block separable (nonsmooth) convex functions, $A^c_j \in \mathbb{R}^{m \times n_j}$ is the $j$-th column block of $A \in \mathbb{R}^{m \times n}$ where $n = \sum_{j} n_j$, $x_j \in \mathbb{R}^{n_j} \times 1$ is the $j$-th block coordinate of $x$, $X_j$ is a local convex constraint of $x_j$ and $a \in \mathbb{R}^{m \times 1}$. The canonical form can be extended to handle linear inequalities by introducing slack variables, i.e., writing $Ax \leq a$ as $Ax + z = a, z \geq 0$.

A variety of machine learning problems can be cast into the linearly-constrained optimization problem $\{1, 8, 4, 24, 5, 6, 21, 11\}$. For example, in robust Principal Component Analysis (RPCA) $[5]$, one attempts to recover a low rank matrix $L$ and a sparse matrix $S$ from an observation matrix $M$, i.e., the linear constraint is $M = L + S$. Further, in the stable version of RPCA $[29]$, an noisy matrix $Z$ is taken into consideration, and the linear constraint has three blocks, i.e., $M = L + S + Z$. Problem $\{1\}$ can also include composite minimization problems which solve a sum of a loss function and a set of nonsmooth regularization functions. Due to the increasing interest in structural sparsity $[1]$, composite regularizers have become widely used, e.g., overlapping group lasso $[28]$. As the blocks are overlapping in this class of problems, it is difficult to apply block coordinate descent methods for large scale problems $\{16, 18\}$ which assume block-separable. By simply splitting blocks and introducing equality constraints, the composite minimization problem can also formulated as $\{1\}$ $[2]$.

A classical approach to solving $\{1\}$ is to relax the linear constraints using the (augmented) Lagrangian, i.e.,

$$L_\rho(x, y) = f(x) + \langle y, Ax - a \rangle + \frac{\rho}{2} \|Ax - a\|_2^2,$$
where \( \rho \geq 0 \) is called the penalty parameter. We call \( x \) the primal variable and \( y \) the dual variable. Usually leads to primal-dual algorithms which update the primal and dual variables alternatively. While the dual update is simply dual gradient descent, the primal update is to solve a minimization problem of \( (2) \) given \( y \). If \( \rho = 0 \), the primal update can be solved in a parallel block coordinate fashion \([3, 19]\), leading to the dual ascent method. While the dual ascent method can achieve massive parallelism, a careful choice of stepsize and some strict conditions are required for convergence, particularly when \( f \) is nonsmooth. To achieve better numerical efficiency and convergence behavior compared to the dual ascent method, it is favorable to set \( \rho > 0 \) in the augmented Lagrangian \( (2) \) which we call the method of multipliers. However, \( (2) \) is no longer separable and solving entire augmented Lagrangian \( (2) \) exactly is computationally expensive. In \([20]\), randomized block coordinate descent (RBCD) \([16, 18]\) is used to solve \( (2) \) exactly, but leading to a double-loop algorithm along with the dual step. More recent results show \( (2) \) can be solved inexactly by just sweeping the coordinates once using the alternating direction method of multipliers (ADMM) \([12, 2]\). This paper attempts to develop a parallel randomized block coordinate variant of ADMM.

When \( J = 2 \), ADMM has been widely used to solve the augmented Lagrangian \( (2) \) in many applications \([2]\). Encouraged by the success of ADMM with two blocks, ADMM has also been extended to solve the problem with multiple blocks \([15, 14, 10, 7, 13, 7]\). The variants of ADMM can be mainly divided into two categories. The first category considers Gauss-Seidel ADMM (GSADMM) \([15, 14]\), which solves \( (2) \) in a cyclic block coordinate manner. In \([13]\), a back substitution step was added so that the convergence of ADMM for multiple blocks can be proved. In some cases, it has been shown that ADMM might not converge for multiple blocks \([7]\). In \([14]\), a block successive upper bound minimization method of multipliers (BSUMM) is proposed to solve the problem \([1]\). The convergence of BSUMM is established under some fairly strict conditions: (i) certain local error bounds hold; (ii) the step size is either sufficiently small or decreasing. However, in general, Gauss-Seidel ADMM with multiple blocks is not well understood and its iteration complexity is largely open. The second category considers Jacobian variants of ADMM \([26, 10, 17]\), which solves \( (2) \) in a parallel block coordinate fashion. In \([26, 17]\), \( (1) \) is solved by using two-block ADMM with splitting variables (sADMM). \([10]\) considers a proximal Jacobian ADMM (PJADMM) by adding proximal terms. A randomized block coordinate variant of ADMM named RBSUMM was proposed in \([14]\). However, RBSUMM can only randomly update one block. Moreover, the convergence of RBSUMM is established under the same conditions as BSUMM and its iteration complexity is unknown.

In this paper, we propose a parallel randomized block coordinate method named parallel direction method of multipliers (PDMM) which randomly picks up any number of blocks to update in parallel, behaving like randomized block coordinate descent \([16, 18]\). Like the dual ascent method, PDMM solves the primal update in a parallel block coordinate fashion even with the augmentation term. Moreover, PDMM inherits the merits of the method of multipliers and can solve a fairly large class of problems, including nonsmooth functions. Technically, PDMM has three aspects which make it distinct from such state-of-the-art methods. First, if block coordinates of the primal \( x \) is solved exactly, PDMM uses a backward step on the dual update so that the dual variable makes conservative progress. Second, the sparsity of \( A \) and the number of randomized blocks are taken into consideration to determine the step size of the dual update. Third, PDMM can randomly update arbitrary number of primal blocks in parallel. Moreover, we show that sADMM and PJADMM are the two extreme cases of PDMM. The connection between sADMM and PJADMM through PDMM provides better understanding of dual backward step. PDMM can also be used to solve overlapping groups in a randomized block coordinate fashion. Interestingly, the corresponding problem for RBCD \([16, 18]\) with overlapping blocks is still an open problem. We establish the global convergence and \( O(1/T) \) iteration complexity of PDMM with constant step size. We evaluate the performance of PDMM in two applications: robust principal component analysis and overlapping group lasso.

The rest of the paper is organized as follows: We introduce PDMM in Section 2 and establish convergence results in Section 3. We evaluate the performance of PDMM in Section 4 and conclude in Section 5. The technical analysis and detailed proofs are provided in the supplement.

Notations: Assume that \( A \in \mathbb{R}^{m \times n} \) is divided into \( I \times J \) blocks. Let \( A_{i}^{r} \in \mathbb{R}^{m \times n} \) be the \( i \)-th row block of \( A \), \( A_{j}^{c} \in \mathbb{R}^{m \times n} \) be the \( j \)-th column block of \( A \), and \( A_{ij} \in \mathbb{R}^{m_i \times n_j} \) be the \( ij \)-th block of \( A \). Let \( y_i \in \mathbb{R}^{m_i \times 1} \) be the \( i \)-th block of \( y \in \mathbb{R}^{m \times 1} \). Let \( \mathcal{N}(i) \) be a set of nonzero blocks \( A_{ij} \) in the
when the number of blocks is large. Based on this observation, we introduce an intermediate variable
with the explicit dual ascent (4) may lead to too aggressive progress on the dual variable, particularly
dual variable, i.e.,
step on the dual variable, i.e.,

\[ x^{t+1}_j = \arg\min_{x_j \in X_j} L_\rho(x_j, x_{k \neq j}, y^t), \]

\[ y^{t+1} = y^t + \tau \rho (Ax^{t+1} - a). \]

where \( \tau \) is a shrinkage factor for the step size of the dual gradient ascent update. However, empirical
results show that it is almost impossible to make the direct Jacobi updates [3]-[4] to converge even when \( \tau \) is extremely small. [15, 10] also noticed that the direct Jacobi updates may not converge.

To address the problem in (3) and (4), we propose a backward step on the dual update. Moreover,
instead of updating all blocks, the blocks \( x_j \) will be updated in a parallel randomized block coordinate
fashion. We call the algorithm Parallel Direction Method of Multipliers (PDMM). PDMM first
randomly select \( K \) blocks denoted by set \( \mathbb{J}_t \) at time \( t \), then executes the following iterates:

\[ x^{t+1}_{j_t} = \arg\min_{x_{j_t} \in X_{j_t}} L_\rho(x_{j_t}, x_{k \neq j_t}, y^t) + \eta_{j_t} B_{\phi_{j_t}}(x_{j_t}, x^{t}_{j_t}), j_t \in \mathbb{J}_t, \]

\[ y^{t+1}_i = y^t_i + \tau_i \rho (A_i x^{t+1}_i - a_i), \]

\[ y^{t+1}_j = y^{t+1}_i - \nu_i (A_i x^{t+1}_i - a_i), \]

where \( \tau_i > 0, 0 \leq \nu_i < 1, \eta_{j_t} \geq 0, \) and \( B_{\phi_{j_t}}(x_{j_t}, x^{t}_{j_t}) \) is a Bregman divergence. Note \( x^{t+1}_j = (x^{t+1}_{j_t}, x_{k \neq j_t}) \) in (6) and (7), we update all dual blocks. We show that PDMM can also do
randomized dual block coordinate ascent in an extended work [25]. Let \( \tilde{K}_i = \min\{d_i, K\} \). \( \tau_i \) and \( \nu_i \) can take the following values:

\[ \tau_i = \frac{K}{\tilde{K}_i (2J - K)}, \nu_i = 1 - \frac{1}{\tilde{K}_i}. \]

In the \( x_{j_t} \)-update (5), a Bregman divergence is added so that exact PDMM and its inexact variants
are analyzed in an unified framework [23][11]. In particular, if \( \eta_{j_t} = 0 \), (5) is an exact update. If
\( \eta_{j_t} > 0 \), by choosing a suitable Bregman divergence, (5) can be solved by various inexact updates,
often yielding a closed-form for the \( x_{j_t} \) update (see Section 2.1).

To better understand PDMM, we discuss the following three aspects which play roles in choosing \( \tau_i \) and \( \nu_i \): the dual backward step (7), the sparsity of \( A \), and the choice of randomized blocks.

**Dual Backward Step:** We attribute the failure of the Jacobi updates (3)-(4) to the following observation in (3), which can be rewritten as:

\[ x^{t+1}_{j_t} = \arg\min_{x_j \in X_j} f_j(x_j) + \langle y^t + \rho (Ax^t - a), A^*_j x_j \rangle + \frac{\rho}{2} \| A^*_j (x_j - x^t_j) \|_2^2. \]

In the primal \( x_j \) update, the quadratic penalty term implicitly adds full gradient ascent step to the
dual variable, i.e., \( y^t + \rho (Ax^t - a) \), which we call implicit dual ascent. The implicit dual ascent along with
the explicit dual ascent (4) may lead to too aggressive progress on the dual variable, particularly
when the number of blocks is large. Based on this observation, we introduce an intermediate variable
\( \hat{y}^t \) to replace \( y^t \) in (9) so that the implicit dual ascent in (9) makes conservative progress, e.g.,
\( \hat{y}^t + \rho (Ax^t - a) = y^t + (1 - \nu) \rho (Ax^t - a) \), where \( 0 < \nu < 1 \). \( \hat{y}^t \) is the result of a ‘backward step’ on the dual variable, i.e., \( \hat{y}^t = y^t - \nu \rho (Ax^t - a) \).

Moreover, one can show that \( \tau \) and \( \nu \) have also been implicitly used when using two-block ADMM
with splitting variables (sADMM) to solve (1) [17][18][25]. Section 2.2 shows sADMM is a special case
of PDMM. The connection helps in understanding the role of the two parameters \( \tau_i, \nu_i \) in PDMM.
Interestingly, the step sizes \( \tau_i \) and \( \nu_i \) can be improved by considering the block sparsity of \( A \)
and the number of random blocks \( K \) to be updated.

**Sparsity of \( A \):** Assume \( A \) is divided into \( I \times J \) blocks. While \( x_j \) can be updated in parallel,
the matrix multiplication \( Ax \) in the dual update (4) requires synchronization to gather messages
from all block coordinates \( j_t \in \mathbb{J}_t \). For updating the \( i \)-th block of the dual \( y_i \), we need \( A_i x^{t+1}_i = \sum_{j_t \in \mathbb{J}_t} A_{ij_t} x^{t+1}_{j_t} + \sum_{k \neq j_t} A_{ik} x^t_k \) which aggregates “messages” from all \( x_{j_t} \). If \( A_{ij_t} \) is a block of
zeros, there is no “message” from \( x_{jt} \) to \( y_i \). More precisely, \( A_i x^{t+1} = \sum_{j \in J \cap N(i)} A_{ij} x_{jt}^{+1} + \sum_{k \notin J} A_{ik} x_k^t \), where \( N(i) \) denotes a set of nonzero blocks in the \( i \)-th row block \( A_i \). \( N(i) \) can be considered as the set of neighbors of the \( i \)-th dual block \( y_i \), and \( d_i = |N(i)| \) is the degree of the \( i \)-th dual block \( y_i \). If \( A \) is sparse, \( d_i \) could be far smaller than \( J \). According to (5), a low \( d_i \) will lead to bigger step sizes \( \tau_i \) for the dual update and smaller step sizes for the dual backward step (7). Further, as shown in Section 2.3, when using PDMM with all blocks to solve composite minimization with overlapping blocks, PDMM can use \( \tau_i = 0.5 \) which is much larger than \( 1/J \) in sADMM.

**Randomized Blocks:** The number of blocks to be randomly chosen also has the effect on \( \tau_i, \nu_i \). If randomly choosing one block (\( K = 1 \)), then \( \nu_i = 0 \), \( \tau_i = \frac{1}{2} \). The dual backward step (7) vanishes. As \( K \) increases, \( \nu_i \) increases from 0 to 1 and \( \tau_i \) decreases from \( \frac{1}{2} \) to \( 0 \). If updating all blocks (\( K = J \)), \( \tau_i = \frac{1}{J} \), \( \nu_i = 0 \).

PDMM does not necessarily choose any \( K \) combination of \( J \) blocks. The \( J \) blocks can be randomly partitioned into \( J/K \) groups where each group has \( K \) blocks. Then PDMM randomly picks some groups. A simple way is to permute the \( J \) blocks and choose \( K \) blocks cyclically.

### 2.1 Inexact PDMM

If \( \eta_{jt} > 0 \), there is an extra Bregman divergence term in (5), which can serve two purposes. First, choosing a suitable Bregman divergence can lead to an efficient solution for (5). Second, if \( \eta_{jt} \) is sufficiently large, the dual update can use a large step size (\( \tau_i = 1 \)) and the backward step (7) can be removed (\( \nu_i = 0 \)), leading to the same updates as PJADMM [10] (see Section 2.2).

Given a continuously differentiable and strictly convex function \( \psi_{jt} \), its Bregman divergence is defined as

\[
B_{\psi_{jt}}(x_{jt}, x_{jt}^t) = \psi_{jt}(x_{jt}) - \psi_{jt}(x_{jt}^t) - \langle \nabla \psi_{jt}(x_{jt}^t), x_{jt} - x_{jt}^t \rangle,
\]

where \( \nabla \psi_{jt} \) denotes the gradient of \( \psi_{jt} \). Rearranging the terms yields

\[
\psi_{jt}(x_{jt}) - B_{\psi_{jt}}(x_{jt}, x_{jt}^t) = \psi_{jt}(x_{jt}^t) + \langle \nabla \psi_{jt}(x_{jt}^t), x_{jt} - x_{jt}^t \rangle,
\]

which is exactly the linearization of \( \psi_{jt}(x_{jt}) \) at \( x_{jt}^t \). Therefore, if solving (5) exactly becomes difficult due to some problematic terms, we can use the Bregman divergence to linearize these problematic terms so that (5) can be solved efficiently. More specifically, in (5), we can choose \( \phi_{jt} = \varphi_{jt} - \frac{1}{\eta_{jt}} \psi_{jt} \), assuming \( \psi_{jt} \) is the problematic term. Using the linearity of Bregman divergence,

\[
B_{\phi_{jt}}(x_{jt}, x_{jt}^t) = B_{\varphi_{jt}}(x_{jt}, x_{jt}^t) - \frac{1}{\eta_{jt}} B_{\psi_{jt}}(x_{jt}, x_{jt}^t).
\]

For instance, if \( f_{jt} \) is a logistic function, solving (5) exactly requires an iterative algorithm. Setting \( \psi_{jt} = f_{jt} \), \( \phi_{jt} = \frac{1}{2} \| \cdot \|_2^2 \) in (12) and plugging into (5) yield

\[
x_{jt}^{t+1} = \arg \min_{x_{jt} \in X_{jt}} (\nabla f_{jt}(x_{jt}^t), x_{jt}) + (\gamma_{jt}^t, A_j x_{jt}) + \frac{\eta_{jt}}{2} \| A_j x_{jt} + \sum_{k \notin J} A_{ik} x_k^t - a \|_2^2 + \eta_{jt} \| x_{jt} - x_{jt}^t \|_2^2,
\]

which has a closed-form solution. Similarly, if the quadratic penalty term \( \frac{\eta_{jt}}{2} \| A_j x_{jt} + \sum_{k \notin J} A_{ik} x_k^t - a \|_2^2 \) is a problematic term, we can set \( \psi_{jt}(x_{jt}) = \frac{\eta_{jt}}{2} \| A_j x_{jt} + \sum_{k \notin J} A_{ik} x_k^t - a \|_2^2 \) and \( B_{\phi_{jt}}(x_{jt}, x_{jt}^t) = \frac{\eta_{jt}}{2} \| A_j (x_{jt} - x_{jt}^t) \|_2^2 \) can be used to linearize the quadratic penalty term.

In (12), the nonnegativeness of \( B_{\phi_{jt}} \) implies that \( B_{\phi_{jt}} \geq \frac{1}{\eta_{jt}} B_{\psi_{jt}} \). This condition can be satisfied as long as \( \varphi_{jt} \) is more convex than \( \psi_{jt} \). Technically, we assume that \( \varphi_{jt} \) is \( \sigma/\eta_{jt} \)-strongly convex and \( \psi_{jt} \) has Lipschitz continuous gradient with constant \( \sigma \), which has been shown in [23].

### 2.2 Connections to Related Work

Consider the case when all blocks are used in PDMM. There are also two other methods which update all blocks in parallel. If solving the primal updates exactly, two-block ADMM with splitting variables (sADMM) is considered in [17][26]. We show that sADMM is a special case of PDMM when setting \( \tau_i = \frac{1}{J} \) and \( \nu_i = 1 - \frac{1}{J} \) (Appendix B in [25]). If the primal updates are solved inexactly, [10] considers a proximal Jacobian ADMM (PJADMM) by adding proximal terms where
the iteration complexity of RBSUMM is still unknown. In contrast, PDMM converges at a rate of
of RBSUMM requires certain local error bounds to be hold and decreasing step size. Moreover,
In addition to PDMM, RBUSMM \cite{14} can also randomly update one block. The convergence
ADMM, PJADMM introduces more terms and parameters.

2.3 Randomized Overlapping Block Coordinate Descent

Consider the composite minimization problem of a sum of a loss function \( \ell(w) \) and composite regularizers \( g_j(w_j) \):

\[
\min_w \ell(w) + \sum_{j=1}^{L} g_j(w_j),
\]

which considers \( L \) overlapping groups \( w_j \in \mathbb{R}^{b \times 1} \). Let \( J = L + 1, x_J = w \). For \( 1 \leq j \leq L \),
denote \( x_j = w_j, \) then \( x_j = U_j^T x_J, \) where \( U_j \in \mathbb{R}^{b \times L} \) is the columns of an identity matrix and
extracts the coordinates of \( x_J \). Denote \( U = [U_1, \ldots, U_L] \in \mathbb{R}^{n \times (bL)} \) and \( A = [I_L, -U^T] \) where
\( bL \) denotes \( b \times L \). By letting \( f_j(x_j) = g_j(w_j) \) and \( f_j(x_J) = \ell(w), \)
\( \text{(13)} \) can be written as:

\[
\min_x \sum_{j=1}^{J} f_j(x_j) \quad \text{s.t.} \quad Ax = 0.
\]

where \( x = [x_1; \ldots; x_L; x_{L+1}] \in \mathbb{R}^{K \times J} \). \( \text{(14)} \) can be solved by PDMM in a randomized block
coordinate fashion. In \( A \), for \( b \) rows block, there are only two nonzero blocks, i.e., \( d_i = 2 \).
Therefore, \( \tau_i = \frac{K}{2J-K}, \nu_i = 0.5 \). In particular, if \( K = J, \tau_i = \nu_i = 0.5 \). In contrast, sADMM uses
\( \tau_i = 1/J < 0.5, \nu_i = 1 - 1/J > 0.5 \) if \( J \) is larger.

\textbf{Remark 1} (a) ADMM \cite{2} can solve \( \text{(14)} \) where the equality constraint is \( x_j = U_j^T x_J \).

(b) In this setting, Gauss-Seidel ADMM (GSADMM) and BSUMM \cite{14} are the same as ADMM.
BSUMM should converge with constant stepsze \( \rho \) (not necessarily sufficiently small), although the
theory of BSUMM does not include this special case.

3 Theoretical Results

We establish the convergence results for PDMM under fairly simple assumptions:

\textbf{Assumption 1}

(1) \( f_j : \mathbb{R}^n \mapsto \mathbb{R} \cup \{+\infty\} \) are closed, proper, and convex.

(2) A \textit{KKT point of the Lagrangian} \( \rho = 0 \) in \( \text{(2)} \) of Problem \( \text{(1)} \) exists.

\textbf{Assumption 1} is the same as that required by ADMM \cite{2, 22}. Assume that \( \{x_j^*, y^*_j\} \) satisfies
the KKT conditions of the Lagrangian \( \rho = 0 \) in \( \text{(2)} \), i.e.,

\[
-A_j^T y^* \in \partial f_j(x_j^*),
\]

\( A x^* - a = 0. \)

\textbf{(16)} During iterations, \( \text{(16)} \) is satisfied if \( A x^{t+1} = a. \) Let \( f_j(x_j^{t+1}) \in \partial f_j(x_j^{t+1}) \) where \( \partial f_j \) be the
subdifferential of \( f_j. \) For \( x_j^* \in X_j, \) the optimality conditions for the \( x_j \) update \( \text{(5)} \) is

\[
\langle f_j(x_j^{t+1}) + A_j^T [y^{t+1} - f_j(x_j^{t+1})] + \eta_j (\nabla f_j(x_j^{t+1}) - \nabla f_j(x_j^*)) | x_j^{t+1} - x_j^* \rangle \leq 0.
\]

When \( A x^{t+1} = a, y^{t+1} = y^*. \) If \( A_j^T [x_j^{t+1} - x_j^*]) = 0, \) then \( A x^* - a = 0. \) When \( \eta_j \geq 0, \) further
assuming \( B_{\phi_j}(x_j^{t+1}, x_j^*), \) \( \text{(15)} \) will be satisfied. Note \( x_j^* \in X_j \) is always satisfied in \( \text{(5)} \) in
PDMM. Overall, the KKT conditions (15)-(16) are satisfied if the following optimality conditions are satisfied by the iterates:
\[ A x^{t+1} = a, \quad A_j^t (x_j^{t+1} - x_j^t) = 0, \quad (17) \]
\[ B_{\phi_j} (x_j^{t+1}, x_j^t) = 0. \quad (18) \]

The above optimality conditions are sufficient for the KKT conditions. (17) are the optimality conditions for the exact PDMM. (18) is needed only when \( \eta_j > 0 \).

Let \( z_{ij} = A_{ij} x_j \in \mathbb{R}^{m_i \times 1}, z_i^t = [z_{i1}^T, \ldots, z_{iJ}^T]^T \in \mathbb{R}^{m_i \times J} \) and \( z = [(z_1^T)^T, \ldots, (z_J^T)^T]^T \in \mathbb{R}^{Jm \times 1} \). Define the residual of optimality conditions (17)-(18) as
\[ R(x^{t+1}) = \frac{\beta}{2} \| z^{t+1} - z^t \|_2^2 + \sum_{i=1}^l \beta_i \| A_i^t x_i^{t+1} - a_i \|_2^2 + \sum_{j=1}^J \eta_j B_{\phi_j} (x_j^{t+1}, x_j^t). \quad (19) \]

where \( P_i \) is some positive semi-definite matrix and \( \beta_i = \frac{K}{JK_i} \). If \( R(x^{t+1}) \to 0 \), (17)-(18) will be satisfied and thus PDMM converges to the KKT point \( \{ x^*, y^* \} \). Define the current iterate \( v^t = (x_j^t, y_i^t) \) and \( h(v^*, v^t) \) as a distance from \( v^t \) to a KKT point \( v^* = (x_j^*, y_i^*) \):
\[ h(v^*, v^t) = \frac{K}{J} \sum_{i=1}^l \frac{1}{2\tau_i \rho} \| y_i^* - y_i^t \|_2^2 + \tilde{\mathcal{E}}_\rho(x_i^t, y_i^t) + \frac{\beta}{2} \| z^* - z_i^t \|_Q^2 + \sum_{j=1}^J \eta_j B_{\phi_j} (x_j^*, x_j^t), \quad (20) \]

where \( Q \) is a positive semi-definite matrix and \( \tilde{\mathcal{E}}_\rho(x_i^t, y_i^t) \) with \( \gamma_i = \frac{2(J-K)}{K/(2J-K)} + \frac{1}{\tau_i} - \frac{K}{JK_i} \) is
\[ \tilde{\mathcal{E}}_\rho(x_i^t, y_i^t) = f(x_i^t) - f(x_i^*) + \sum_{j=1}^J \left( (y_i^t, A_i^t x_i^t - a_i) + \frac{\gamma_j}{\rho} \| A_i^t x_i^t - a_i \|_2^2 \right). \quad (21) \]

The following Lemma shows that \( h(v^*, v^t) \geq 0 \).

**Lemma 1** Let \( v^t = (x_j^t, y_i^t) \) be generated by PDMM (3)-(7) and \( h(v^*, v^t) \) be defined in (20). Setting \( \nu_i = 1 - \frac{1}{K_i} \) and \( \tau_i = \frac{K}{K/(2J-K)} \), we have
\[ h(v^*, v^t) \geq \frac{\beta}{2} \sum_{i=1}^l \xi_i \| A_i^t x_i^t - a_i \|_2^2 + \frac{\beta}{2} \| z^* - z_i^t \|_Q^2 + \sum_{j=1}^J \eta_j B_{\phi_j} (x_j^*, x_j^t) \geq 0. \quad (22) \]

where \( \xi_i = \frac{(J-K)}{K/(2J-K)} + \frac{1}{\tau_i} - \frac{K}{JK_i} \geq 0 \). Moreover, if \( h(v^*, v^t) = 0 \), then \( A_j^t x_j^t = a_j, z^t = z^* \) and \( B_{\phi_j} (x_j^*, x_j^t) = 0 \). Thus, (15)-(16) are satisfied.

In PDMM, \( y_i^{t+1} \) depends on \( x_i^{t+1} \), which in turn depends on \( y_i^t \). \( x_i^t \) and \( y_i^t \) are independent of \( J_i \). \( x_i^t \) depends on the observed realizations of the random variable \( \xi_{i-1} = \{x_j^1, \ldots, x_j^{i-1}\} \). The following theorem shows that \( h(v^*, v^t) \) decreases monotonically and thus establishes the global convergence of PDMM.

**Theorem 1 (Global Convergence)** Let \( v^t = (x_j^t, y_i^t) \) be generated by PDMM (3)-(7) and \( v^* = (x_j^*, y_i^*) \) be a KKT point satisfying (15)-(16). Setting \( \nu_i = 1 - \frac{1}{K_i} \) and \( \tau_i = \frac{K}{K/(2J-K)} \), we have
\[ 0 \leq \mathbb{E}_{\xi_i} h(v^*, v^{t+1}) \leq \mathbb{E}_{\xi_i} h(v^*, v^t), \quad \mathbb{E}_{\xi_i} R(x^{t+1}) \to 0. \quad (23) \]

The following theorem establishes the iteration complexity of PDMM in an ergodic sense.

**Theorem 2 (Iteration Complexity)** Let \( (x_j^t, y_i^t) \) be generated by PDMM (3)-(7). Let \( \bar{x}^T = \sum_{t=1}^T x_i^t \). Setting \( \nu_i = 1 - \frac{1}{K_i} \) and \( \tau_i = \frac{K}{K/(2J-K)} \), we have
\[ \mathbb{E} f(\bar{x}^T) - f(x^*) \leq \frac{\beta}{K} \left\{ \sum_{i=1}^l \frac{1}{2\tau_i \rho} \| y_i^* \|_2^2 + \tilde{\mathcal{E}}_\rho(x_i^t, y_i^t) + \frac{\beta}{2} \| z^* - z_i^t \|_Q^2 + \sum_{j=1}^J \eta_j B_{\phi_j} (x_j^*, x_j^t) \right\}, \quad \mathbb{E} \sum_{i=1}^l \beta_i \| A_i^t x_i^t - a_i \|_2^2 \leq \frac{\beta}{T} h(v^*, v^0) \].

where \( \beta_i = \frac{K}{JK_i} \), \( Q \) is a positive semi-definite matrix, and the expectation is over \( J_i \).
Figure 1: Comparison of the convergence of PDMM with ADMM methods in RPCA.

Table 1: The best results of PDMM with tuning parameters $\tau_i, \nu_i$ in RPCA.

<table>
<thead>
<tr>
<th></th>
<th>time (s)</th>
<th>iteration</th>
<th>residual($10^{-3}$)</th>
<th>objective (log)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDMM1</td>
<td>118.83</td>
<td>40</td>
<td>3.60</td>
<td>8.07</td>
</tr>
<tr>
<td>PDMM2</td>
<td>137.46</td>
<td>34</td>
<td>5.51</td>
<td>8.07</td>
</tr>
<tr>
<td>PDMM3</td>
<td>147.82</td>
<td>31</td>
<td>6.54</td>
<td>8.07</td>
</tr>
<tr>
<td>GSADMM</td>
<td>163.09</td>
<td>28</td>
<td>6.84</td>
<td>8.07</td>
</tr>
<tr>
<td>RBSUMM</td>
<td>206.96</td>
<td>141</td>
<td>8.55</td>
<td>8.07</td>
</tr>
<tr>
<td>sADMM</td>
<td>731.51</td>
<td>139</td>
<td>9.73</td>
<td>8.07</td>
</tr>
</tbody>
</table>

Remark 2 PDMM converges at the same rate as ADMM and its variants. In Theorem 2, PDMM can achieve the fastest convergence by setting $J = K = 1$, $\tau_i = 1$, $\nu_i = 0$, i.e., the entire matrix $A$ is considered as a single block, indicating PDMM reduces to the method of multipliers. In this case, however, the resulting subproblem may be difficult to solve, as discussed in Section 1. Therefore, the number of blocks in PDMM depends on the trade-off between the number of subproblems and how efficiently each subproblem can be solved.

4 Experimental Results

In this section, we evaluate the performance of PDMM in solving robust principal component analysis (RPCA) and overlapping group lasso (RPCA) and overlapping group lasso 

**RPCA:** RPCA is used to obtain a low rank and sparse decomposition of a given matrix $A$ corrupted by noise:

$$
\min \frac{1}{2} \|X_1\|_F^2 + \gamma_2 \|X_2\|_1 + \gamma_3 \|X_3\|_* \quad \text{s.t.} \quad A = X_1 + X_2 + X_3.
$$

where $A \in \mathbb{R}^{m \times n}$, $X_1$ is a noise matrix, $X_2$ is a sparse matrix and $X_3$ is a low rank matrix. $A = L + S + V$ is generated in the same way as \cite{17}. In this experiment, $m = 1000$, $n = 5000$ and the rank is 100. The number appended to PDMM denotes the number of blocks ($K$) to be chosen in PDMM, e.g., PDMM1 randomly updates one block.

Figure 1 compares the convergence results of PDMM with ADMM methods. In PDMM, $\rho = 1$ and $\tau_i, \nu_i$ are chosen according to \cite{5}, i.e., $(\tau_i, \nu_i) = \{(\frac{1}{5},0), (\frac{1}{5}, \frac{1}{5}), (\frac{1}{5}, \frac{1}{5})\}$ for PDMM1, PDMM2 and PDMM3 respectively. We choose the `best’ results for GSADMM ($\rho = 1$) and RBSUMM ($\rho = 1$, $\alpha = \rho \frac{1}{\sqrt{T_i+16}}$) and sADMM ($\rho = 1$). PDMMs perform better than RBSUMM and sADMM.

Note the public available code of sADMM, does not have dual update, i.e., $\tau_i = 0$. sADMM should be the same as PDMM3 if $\tau_i = \frac{1}{5}$. Since $\tau_i = 0$, sADMM is the slowest algorithm. Without tuning the parameters of PDMM, GSADMM converges faster than PDMM. Note PDMM can run in parallel but GSADMM only runs sequentially. PDMM3 is faster than two randomized version of PDMM since the costs of extra iterations in PDMM1 and PDMM2 have surpassed the savings at each iteration. For the two randomized one block coordinate methods, PDMM1 converges faster than RBSUMM in terms of both the number of iterations and runtime.

**The effect of $\tau_i, \nu_i$:** We tuned the parameter $\tau_i, \nu_i$ in PDMMs. Three randomized methods (RBSUMM, PDMM1 and PDMM2) choose the blocks cyclically instead of randomly. Table 1 compares the ‘best’ results of PDMM with other ADMM methods. In PDMM, $(\tau_i, \nu_i) = \frac{1}{5}$.

{(\(\frac{1}{2}, 0\)), (\(\frac{1}{2}, 1\)), (\(\frac{1}{2}, \frac{1}{2}\))}. GSADMM converges with the smallest number of iterations, but PDMMs can converge faster than GSADMM in terms of runtime. The computation per iteration in GSADMM is slightly higher than PDMM3 because GSADMM updates the sum \(X_1 + X_2 + X_3\) but PDMM3 can reuse the sum. Therefore, if the numbers of iterations of the two methods are close, PDMM3 can be faster than GSADMM. PDMM1 and PDMM2 can be faster than PDMM3. By simply updating one block, PDMM1 is the fastest algorithm and achieves the lowest residual. PDMM3 can be faster than GSADMM, PDMM1 and PDMM2 can be faster than PDMM3. By simply updating one block, PDMM1 is the fastest algorithm and achieves the lowest residual.

**Overlapping Group Lasso:** We consider solving the overlapping group lasso problem \([28]\):

\[
\min_w \frac{1}{2L\lambda} \|Aw - b\|^2 + \sum_{g \in G} d_g \|w_g\|_2.
\]

where \(A \in \mathbb{R}^{m \times n}, w \in \mathbb{R}^{n \times 1}\) and \(w_g \in \mathbb{R}^{b \times 1}\) is the vector of overlapping group indexed by \(g\). \(d_g\) is some positive weight of group \(g \in G\). As shown in Section 2.3, (25) can be rewritten as the form \([14]\). The data is generated in a same way as \([27, 9]\): the elements of \(A\) are sampled from normal distribution, \(b = Ax + \epsilon\) with noise \(\epsilon\) sampled from normal distribution, and \(x_j = (-1)^j \exp(-(j - 1)/100)\). In this experiment, \(m = 5000\), the number of groups is \(L = 100\), and \(d_g = \frac{1}{2}, \lambda = \frac{1}{2}\) in \((25)\). The size of each group is 100 and the overlap is 10. The total number of blocks in PDMM and S-ADMM is \(J = 101\). \(\tau_1, \nu_1\) in PDMM are computed according to \([8]\).

In Figure 2, the first two figures plot the convergence of objective in terms of the number of iterations and time. PDMM uses all 101 blocks and is the fastest algorithm. ADMM is the same as GSADMM in this problem, but is slower than PDMM. Since sADMM does not consider the sparsity, it uses \(\tau_1 = \frac{2}{T+1}, \nu_1 = 1 - \frac{2}{T+1}\), leading to slow convergence. The two accelerated methods, PA-APG \([27]\) and S-APG \([9]\), are slower than PDMM and ADMM.

**The effect of \(K\):** The third figure shows PDMM with different number of blocks \(K\). Although the complexity of each iteration is the lowest when \(K = 1\), PDMM takes much more iterations than other cases and thus takes the longest time. As \(K\) increases, PDMM converges faster and faster. When \(K = 20\), the runtime is already same as using all blocks. When \(K > 21\), PDMM takes less time to converge than using all blocks. The runtime of PDMM decreases as \(K\) increases from 21 to 61. However, the speedup from 61 to 81 is negligible. We tried different set of parameters for RBSUMM \(\rho\sum_{i=0}^{K-1}{L+i\choose i}(0 \leq i \leq 5, \rho = 0.01, 0.1, 1)\) or sufficiently small step size, but could not see the convergence of the objective within 5000 iterations. Therefore, the results are not included here.

### 5 Conclusions

We proposed a randomized block coordinate variant of ADMM named Parallel Direction Method of Multipliers (PDMM) to solve the class of problem of minimizing block-separable convex functions subject to linear constraints. PDMM considers the sparsity and the number of blocks to be updated when setting the step size. We show two existing Jacobian ADMM methods are special cases of PDMM. We also use PDMM to solve overlapping block problems. The global convergence and the iteration complexity are established with constant step size. Experiments on robust PCA and overlapping group lasso show that PDMM is faster than existing methods.

### Acknowledgment

H. W. and A. B. acknowledge the support of NSF via IIS-1447566, IIS-1422557, CCF-1451986, CNS-1314560, IIS-0953274, IIS-1029711, IIS-0916750, and NASA grant NNX12AQ39A. H. W. acknowledges the support of DDF (2013-2014) from the University of Minnesota. A.B. acknowledges support from IBM and Yahoo. Z.Q. Luo is supported in part by the US AFOSR via grant number FA9550-12-1-0340 and the National Science Foundation via grant number DMS-1015346.
References


