Journal of the Operations Research Society of China (2022) 10:1–52
https://doi.org/10.1007/s40305-021-00368-3

A Survey on Some Recent Developments of Alternating Direction Method of Multipliers

De-Ren Han[1]

Received: 1 February 2020 / Revised: 8 September 2021 / Accepted: 9 September 2021 /
Published online: 1 January 2022

Abstract
Recently, alternating direction method of multipliers (ADMM) attracts much attentions from various fields and there are many variant versions tailored for different models. Moreover, its theoretical studies such as rate of convergence and extensions to nonconvex problems also achieve much progress. In this paper, we give a survey on some recent developments of ADMM and its variants.

Keywords Alternating direction method of multipliers · Global convergence · Rate of convergence · Nonconvex optimization

Mathematics Subject Classification 90C30 · 90C33 · 65K05

1 Introduction

In this paper, we survey the developments of the alternating direction method of multipliers (ADMM) and its variants for solving the minimization problem with linear constrains and a separable objective function which is the sum of many individual functions without coupled variables:

$$\min \left\{ \sum_{i=1}^{m} \theta_i(x_i) \;\Big|\; \sum_{i=1}^{m} A_i x_i = b,\ x_i \in \mathcal{X}_i,\ i = 1, \cdots, m \right\}, \tag{1}$$

This work is supported by the National Natural Science Foundation of China (Nos. 11625105 and 12131004).

✉ De-Ren Han
handr@buaa.edu.cn

[1] LMIB of the Ministry of Education, School of Mathematical Sciences, Beihang University, Beijing 100191, China

where $\theta_i : \mathbb{R}^{n_i} \to \mathbb{R} \cup \{\infty\}$ are closed proper functions; $A_i \in \mathbb{R}^{l\times n_i}$; $\mathcal{X}_i \subseteq \mathbb{R}^{n_i}$ are closed and convex nonempty sets; $b \in \mathbb{R}^l$; and $\sum_{i=1}^m n_i = n$. As a linearly constrained optimization problem, though the model (1) is special, it is rich enough to characterize many optimization problems arising from various application fields, e.g., the image alignment problem in [1], the robust principal component analysis model with noisy and incomplete data in [2], the latent variable Gaussian graphical model selection in [3], the quadratic discriminant analysis model in [4] and the quadratic conic programming in [5]; just list a few.

We now give some concrete application models.

ℓ_1-norm minimization: In some applications such as statistics, machine learning, and signal processing, one wants to find a 'sparse' solution from the given data. Let $b \in \mathbb{R}^l$ denote the observed signal and we know that it comes from a linear transformation $A \in \mathbb{R}^{l\times n}$ and $l << n$. The task is to find the sparsest solution, i.e., the vector that contains as many zero elements as possible and satisfies the equation $Ax = b$. Let $\|y\|_0$ denote the number of nonzero elements of the vector y, then we can formulate the problem as

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax - b\|^2 + \mu\|x\|_0, \tag{2}$$

where $\mu > 0$ is a scalar regularization parameter that is usually chosen by cross-validation. Introducing a new variable, we can reformulate (2) as

$$\min_{x\in\mathbb{R}^n} \left\{ \frac{1}{2}\|Ax - b\|^2 + \mu\|y\|_0 \;\middle|\; x = y \right\}, \tag{3}$$

which is a special case of (1) with $m = 2$.

Since the zero norm is discontinuous and nonconvex, researchers usually replace it with its convex hull, the ℓ_1-norm. Then, (2) and (3) can be relaxed to

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax - b\|^2 + \mu\|x\|_1, \tag{4}$$

and

$$\min_{x\in\mathbb{R}^n} \left\{ \frac{1}{2}\|Ax - b\|^2 + \mu\|y\|_1 \;\middle|\; x = y \right\}, \tag{5}$$

respectively. The model (4) is just the well-known *lasso* [6].

A generalization of the above model is that it is not the solution itself but its linear transformation is required to be sparse, and the optimization model is

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax - b\|^2 + \mu\|Fx\|_0, \tag{6}$$

where F is an arbitrary linear transformation. Again, after introducing an auxiliary variable, we get a special case of (1) with $m = 2$,

$$\min_{x\in\mathbb{R}^n} \left\{ \frac{1}{2}\|Ax-b\|^2 + \mu\|y\|_0 \;\middle|\; Fx = y \right\};$$

and replacing the zero norm with its convex hull, the ℓ_1-norm, we obtain its relaxed model

$$\min_{x\in\mathbb{R}^n} \left\{ \frac{1}{2}\|Ax-b\|^2 + \mu\|y\|_1 \;\middle|\; Fx = y \right\}, \tag{7}$$

which is called *generalized lasso*. When F is the difference matrix

$$F_{ij} = \begin{cases} 1, & j = i+1, \\ -1, & j = i, \\ 0, & \text{otherwise}, \end{cases}$$

then $\|Fx\|_1$ is the total variation of x [7], which finds wide applications in image processing.

Other ℓ_1-norm minimization models that can be reformulated into (1) include basis pursuit [8], Huber function fitting [9], group lasso [10], etc.

Matrix completion: In some applications such as the movie ratings in the Netflix problem, part of the data (elements of a matrix) is unaccessible, and the task is filling in the missing entries of a partially observed matrix. That is, given a ratings matrix in which each entry (i, j) represents the rating of movie j by customer i if customer i has watched movie j and is otherwise missing, we would like to predict the remaining entries. A property that helps to accomplish the task is that the preferred matrix is low rank, or its rank is known as priori; otherwise the hidden entries could be assigned arbitrary values.

Let M be the matrix to be recovered and let Ω be the set of locations corresponding to the observed entries ($(i, j) \in \Omega$ if M_{ij} is observed). The optimization model is [11]

$$\min_{x\in\mathbb{R}^{l\times n}} \left\{ \text{rank}(x) \;\middle|\; x_{ij} = M_{ij}, \text{ for } (i, j) \in \Omega \right\}. \tag{8}$$

As (5) to (2), we can also relax and reformulate (8) to the convex separable problem

$$\min_{x,y\in\mathbb{R}^{l\times n}} \left\{ \|x\|_* \;\middle|\; x = y, y_{ij} = M_{ij}, \text{ for } (i, j) \in \Omega \right\}, \tag{9}$$

where $\|x\|_*$ denotes the nuclear norm of the matrix x which is defined as the sum of its singular values. Then, we obtain a special case of (1) for $m = 2$ and with matrix variables.

Robust principal component analysis: Given part of the elements of a data matrix which is the superposition of a low rank matrix and a sparse matrix, the robust principal component analysis (RPCA) is to recover each component individually [12]. Moreover, the given data may be corrupted by noises. As in the matrix completion

example, let Ω be the set of locations corresponding to the given entries, and let $P_\Omega : \mathbb{R}^{l\times n} \to \mathbb{R}^{l\times n}$ be the orthogonal projection onto the span of matrices vanishing outside of Ω, i.e., the ij-th entry of $P_\Omega(x)$ is x_{ij} if $(i, j) \in \Omega$ and zero otherwise. The optimization model for the robust principal component analysis problem is

$$\min_{x\in\mathbb{R}^{l\times n}} \left\{ \operatorname{rank}(x) + \tau_1\|y\|_0 + \tau_2\|P_\Omega(z)\|_F^2 \;\middle|\; x + y + z = M \right\}, \tag{10}$$

where M is the given data, $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. Relaxing the rank and the zero norm with their convex hull, we obtain a relaxation [2]

$$\min_{x\in\mathbb{R}^{l\times n}} \left\{ \|x\|_* + \tau_1\|y\|_1 + \tau_2\|P_\Omega(z)\|_F^2 \;\middle|\; x + y + z = M \right\}. \tag{11}$$

Both (10) and (11) fall into the framework of (1) with $m = 3$.

Note that the original application models (2), (6), (8) and (10) contain discrete terms such as the ℓ_0-norm and the rank function, and they are nonconvex optimization problems. Solving these optimization problems are usually NP-hard (nondeterministic polynomial hard). Peoples thus heuristically turn to solving their relaxation problems (5), (7), (9) and (11). Fortunately, under suitable conditions, the relaxed problem and the original one share the same solutions [11,13].

The relaxation problems (5), (7), (9) and (11) are convex optimization problems, and there are many state-of-the-art solvers for solving them. In particular, the problems can be further reformulated as a linear programming (LP) or a semidefinite programming (SDP) and the interior point algorithm and Newton's methods can solve them. Nevertheless, the rapid increase in the dimension brings great challenge to the solvers [14]. However, the hardness induced by the high dimension can be alleviated by utilizing the problems' structure. In all these models, the objective function is the sum of several individual functions, and the constraints are linear equalities. We call these optimization models, and the uniform (1), separable optimization problems with linear constraints.

Among the methods that take advantage of the separable structure of the model (1), the alternating direction method of multipliers (ADMM) attracts much attentions. For solving these modern application problems arising from big data and artificial intelligence, ADMM performs reasonably well. Though the individual component functions can be nonsmooth, the subproblems in ADMM are very easy to solve, and they usually even possess closed-form solution. This makes the method relatively simple in application. Moreover, in these applications, extremely high accuracy is not required, and as a consequence, the slow 'tail convergence' in ADMM is not a serious impact.

With the rapid development of ADMM, there have been several survey papers. In particular:

- In [15], the authors gave a thorough survey of ADMM, mainly from the viewpoint of the applications from statistics and machine learning. Essentially, it is these applications from big data and artificial intelligence that make the renaissance of ADMM.

- In [16], the author delineated the origin of ADMM from a historical point of view. ADMM originates from numerical partial differential equations [17–19] developed it and extended it to variational inequality problems.
- In [20], the authors pointed out that while ADMM can be regarded as an inexact application of the well-known augmented Lagrangian method (ALM), it is helpful to the convergence analysis. They then suggested an accessible version of the 'operator splitting' version of the ADMM convergence proof.

Though these papers survey ADMM from different point of view, there are still very large part of the recent developments that are not mentioned, e.g., the extension of the classical two-block case to the multi-block case; the rate of convergence; and the extension to the nonconvex case. In this paper, we give a thorough survey on these aspects of the developments of ADMM.

ADMM is an augmented Lagrangian-based method. Introducing a Lagrange multiplier λ to the linear constraint, we can write down the augmented Lagrangian function associated with (1)

$$\mathcal{L}_\beta(x_1, \cdots, x_m, \lambda) = \sum_{i=1}^{m} \theta_i(x_i) - \langle \lambda, \sum_{i=1}^{m} A_i x_i - b \rangle + \frac{\beta}{2} \| \sum_{i=1}^{m} A_i x_i - b \|^2,$$

where $\beta > 0$ is a parameter.

Throughout, the solution set of (1) is assumed to be nonempty. Moreover, we use the notation $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_m$ and $x = (x_1, x_2, \cdots, x_m) \in \mathbb{R}^n$, and denote by x_{-i} the subvector of x excluding only x_i, i.e.,

$$x_{-i} := (x_1, x_2, \cdots, x_{i-1}, x_{i+1}, \cdots, x_m) \in \mathbb{R}^{n-n_i}, \quad i = 1, \cdots, m.$$

Similarly, denote by $\mathcal{X}_{-i}$ the subset of $\mathcal{X}$ excluding only $\mathcal{X}_i$, i.e.,

$$\mathcal{X}_{-i} := \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \mathcal{X}_m \subseteq \mathbb{R}^{n-n_i}, \quad i = 1, \cdots, m.$$

In the alternating direction method of multipliers for solving (1), the variable x_1 plays only an intermediate role and is not involved in the execution of the main recursion. Therefore, the input for executing the next iterate of ADMM is only x_{-1}^k and λ^k, and in the convergence analysis, we usually use the sequence $\{(x_{-1}^k, \lambda^k)\}$. For notation convenience, let $v = (x_{-1}, \lambda)$ and $w = (x, \lambda)$. The variables with superscript such as v^*, v^k, w^* and w^k can be defined similarly.

For any vector x, $\min\{0, x\}$ is a vector with the same dimension of x, whose ith element is x_i if $x_i < 0$ and 0 otherwise. The Euclidean norm is denoted by $\|\cdot\|$. For any convex function $\theta : \mathcal{X} \to (-\infty, +\infty]$, we use $\operatorname{dom}\theta$ to denote its effective domain, i.e., $\operatorname{dom}\theta := \{x \in \mathcal{X} : \theta(x) < \infty\}$; $\operatorname{epi}\theta$ to denote its epigraph, i.e., $\operatorname{epi}\theta := \{(x, t) \in \mathcal{X} \times \mathbb{R} : \theta(x) \leqslant t\}$; and $\theta^* : \mathcal{X} \to (-\infty, +\infty]$ represents its Fenchel conjugate, i.e., $\theta^*(y) = \sup_x\{y^\top x - \theta(x)\}$. The multivalued mapping $\partial\theta : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is called the subdifferential of θ and its element $\xi \in \partial\theta(x)$ is a

subgradient of θ satisfying

$$\theta(z) \geqslant \theta(x) + \langle \xi, z - x \rangle, \ \forall z.$$

For a set C in $\mathbb{R}^n$, ri(C) denotes the relative interior of C; and P_C denotes the Euclidean projector onto C, which maps a vector in $\mathbb{R}^n$ onto the nearest point of C if it is closed.

The rest of the paper is organized as follows. In the next section, we give a simple review of the augmented Lagrangian method due to the fact that (one of the main understanding of ADMM), ADMM is an approximate application of ALM, using one path of block coordinate minimization to approximately minimize the augmented Lagrangian per iteration. We then review the classical ADMM in Sect. 3, where we divide it into four subsections focusing on four different factors. The first one is deriving ADMM from the Douglas–Rachford splitting method (DRSM) which helps understand the convergence analysis, and the other three factors will impact the efficiency of the implementation of ADMM, i.e., selection of the penalty parameter, easier subproblems from splitting, and approximate solution of the subproblems. In Sect. 4, we review the results on the rate of convergence of ADMM, including the sublinear rate and the linear rate of convergence. In Sect. 5, we review the extensions and variants of ADMM. We first present the counter example from [21], which shows that for the multi-block case, although the direct extension of ADMM converges and performs well in many applications, it does not converge for the convex separable problem of the form (1) when $m \geqslant 3$. We then review two development directions for the multi-block case, i.e., conditions that guarantee the convergence and simple variants of the algorithm such as a correction-step. In Sect. 6, we review the recent develop of ADMM for solving (1) where some of the component functions θ_i are nonconvex. Usually, it is regarded that when there is a nonconvex component, the minimization problem (1) is harder than its convex counterpart. We show that for some special cases, e.g., when $m = 2$ and the model is 'strongly+weakly' convex[1], it is the same to the total convex case. For the general model when there is no strongly convex component, we can also have some results with the aid of the Kurdyka–Lojasiewicz inequality [22,23] or(and) the error bound conditions. In Sect. 7, we list some future research questions (topics) and conclude the paper.

2 The Augmented Lagrangian Method

The classical augmented Lagrangian method (ALM) [24,25] for solving the linearly constrained optimization problem

$$\min \left\{ \theta(x) \,\middle|\, Ax = b \right\}, \tag{12}$$

[1] A function f is 'strongly' convex with modulus $\alpha_1 > 0$ if $f - \frac{\alpha_1}{2}\|\cdot\|^2$ is a convex function; it is 'weakly' convex with modulus $\alpha_2 > 0$ if $f + \frac{\alpha_2}{2}\|\cdot\|^2$ is convex.

where $\theta : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a closed proper convex function, $A \in \mathbb{R}^{l \times n}$, and $b \in \mathbb{R}^l$. The iterative scheme of (ALM) is

$$\begin{cases} x^{k+1} = \arg\min_x \mathcal{L}_\beta(x, \lambda^k), \\ \lambda^{k+1} = \lambda^k - \tau\beta(Ax^{k+1} - b), \end{cases} \tag{13}$$

where $\tau \in (0, 2)$ is a parameter and

$$\mathcal{L}_\beta(x, \lambda) = \theta(x) - \langle \lambda, Ax - b \rangle + \frac{\beta}{2}\|Ax - b\|^2,$$

is the augmented Lagrangian function, $\lambda \in \mathbb{R}^l$ is the Lagrange multiplier associated with the linear equality constraint, and $\beta > 0$ is a penalty parameter.

Let

$$d_\beta(\lambda) = \inf_x \mathcal{L}_\beta(x, \lambda)$$

denote the augmented dual function of (12). Then, d_β is a differentiable function and

$$\nabla d_\beta(\lambda) = -(Ax(\lambda) - b),$$

where $x(\lambda)$ is the optimal solution of the following problem (with parameter λ)

$$x(\lambda) \in \arg\min_x \ \mathcal{L}_\beta(x, \lambda). \tag{14}$$

Here and throughout the paper, we use 'arg min' to denote the solution set of the minimization problem. Moreover, ∇d_β is Lipschitz continuous with constant $1/\beta$. Hence, the augmented Lagrangian method can be viewed as a gradient method for the augmented dual problem

$$\max_\lambda \ d_\beta(\lambda).$$

The differentiability of d_β is based on the fact that, although the solution $x(\lambda)$ is usually not unique, Ax is a constant over $X(\lambda)$, where $X(\lambda)$ denotes the set of optimal solutions for (14). The following two lemmas are Lemmas 2.1-2.2 in [26], respectively. The proof is mainly based on the Danskin's Theorem [27, Prop. 4.5.1].

Lemma 1 *For any $\lambda \in \mathbb{R}^l$, Ax is a constant over $X(\lambda)$. Thus, the dual function $d_\beta(\cdot)$ is differentiable everywhere and*

$$\nabla d_\beta(\lambda) = -(Ax(\lambda) - b),$$

where $x(\lambda) \in X(\lambda)$ is any minimizer of (14).

Lemma 2 *∇d_β is Lipschitz continuous with constant $1/\beta$. That is,*

$$\|\nabla d_\beta(\lambda_1) - \nabla d_\beta(\lambda_2)\| \leqslant \frac{1}{\beta}\|\lambda_1 - \lambda_2\|, \quad \forall \lambda_1, \lambda_2 \in \mathbb{R}^l.$$

The optimality condition for (12) are the primal feasibility and dual feasibility

$$Ax^* - b = 0, \quad A^\top \lambda^* \in \partial\theta(x^*).$$

Assuming that strong duality holds, the optimal value of the primal and the dual problems are the same. Once we get the dual optimal solution λ^*, the original solution can be obtained by solving the optimization problem

$$\min_x \mathcal{L}_\beta(x, \lambda^*).$$

A predecessor of the augmented Lagrangian method is the dual ascent method. The Lagrangian function for (12) is

$$\mathcal{L}(x, \lambda) = \theta(x) - \langle \lambda, Ax - b\rangle,$$

and the dual function is

$$d(\lambda) = \inf_x \ \mathcal{L}(x, \lambda) = -\theta^*(A^\top \lambda) + b^\top \lambda,$$

where θ^* is the conjugate of θ [28]. The iterative scheme of the dual ascent method is

$$\begin{cases} x^{k+1} = \arg\min_x \mathcal{L}(x, \lambda^k), \\ \lambda^{k+1} = \lambda^k - \alpha_k(Ax^{k+1} - b), \end{cases}$$

where $\alpha_k > 0$ is a step size.

The augmented Lagrangian method is usually more efficient (less cpu time in finding an approximate solution) and robust (performance that does not heavily depend on parameters such as the initial point, the step size) than the dual ascent method. However, the dual ascent method has its advantage that it can lead to a decentralized algorithm in the primal subproblems, e.g., the objective function is the sum of some component functions. However, it is not an easy task to have a 'full' decentralized manner in both primal and dual variables; the interested reader is referred to [29].

3 The Alternating Direction Method of Multipliers

The classical alternating direction method of multipliers (ADMM) is for solving the linearly constrained convex optimization problem with two blocks of variables and functions

$$\min \ \left\{ \theta_1(x_1) + \theta_2(x_2) \ \middle| \ A_1x_1 + A_2x_2 = b \right\}, \tag{15}$$

where $\theta_i : \mathbb{R}^{n_i} \to \mathbb{R} \cup \{\infty\}$ are closed proper convex functions, $A_i \in \mathbb{R}^{l \times n_i}$, $i = 1, 2$, and $b \in \mathbb{R}^l$.

Definition 1 A KKT point for problem (15) is some $(x_1^*, x_2^*, \lambda^*) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \mathbb{R}^l$ such that

1. x_1^* minimizes $\theta_1(x_1) - \langle \lambda^*, A_1 x_1 \rangle$ with respect to x_1;
2. x_2^* minimizes $\theta_2(x_2) - \langle \lambda^*, A_2 x_2 \rangle$ with respect to x_2;
3. $A_1 x_1^* + A_2 x_2^* = b$.

Let $\lambda \in \mathbb{R}^l$ be the Lagrange multiplier, and let $\beta > 0$ be a penalty parameter. The augmented Lagrangian function of (15) is

$$\mathcal{L}_\beta(x_1, x_2, \lambda) = \theta_1(x_1) + \theta_2(x_2) - \langle \lambda, A_1 x_1 + A_2 x_2 - b \rangle + \frac{\beta}{2} \|A_1 x_1 + A_2 x_2 - b\|^2, \quad (16)$$

and the iterative scheme of ADMM is

$$\begin{cases} x_1^{k+1} = \arg\min\limits_{x_1} \mathcal{L}_\beta(x_1, x_2^k, \lambda^k), \\ x_2^{k+1} = \arg\min\limits_{x_2} \mathcal{L}_\beta(x_1^{k+1}, x_2, \lambda^k), \\ \lambda^{k+1} = \lambda^k - \tau\beta(A_1 x_1^{k+1} + A_2 x_2^{k+1} - b), \end{cases} \quad (17)$$

where $\tau \in \left(0, \frac{1+\sqrt{5}}{2}\right)$. In most part of the paper, we omit the parameter τ for succinctness of the discussion (τ is fixed to be 1); though in applications, a larger τ is much advisable. ADMM was originally proposed by Glowinski and Marrocco [17], and Gabay and Mercier [18]. Recently, due to its great success in solving the optimization problems arising from machine learning, statistics, artificial intelligence, ADMM gets more and more attentions, and there have several survey papers from different point of view [15,16,20].

One can certainly group the two blocks of variables and functions into a single variable and a single function, and then use the augmented Lagrangian method to solve it, with the following iterative scheme:

$$\begin{cases} (x_1^{k+1}, x_2^{k+1}) = \arg\min\limits_{x_1, x_2} \mathcal{L}_\beta(x_1, x_2, \lambda^k), \\ \lambda^{k+1} = \lambda^k - \beta(A_1 x_1^{k+1} + A_2 x_2^{k+1} - b). \end{cases} \quad (18)$$

However, such a scheme does not exploit the separable structure of the problem (15), and usually, the (x_1, x_2)-minimization problem has to be iteratively solved. Obviously, the scheme (17) is capable of exploiting the properties of θ_1 and θ_2 individually, making the subproblems much easier and sometimes easy enough to have closed-form solutions.

Comparing the (x_1, x_2)-minimization problem in (18) and (17), we can understand the ADMM as a Gauss–Seidel implementation for solving (18), approximately with a single iterative [16]. On the other hand, it can also be understood as the Douglas–Rachford splitting method applying to the dual problem of (15) [30].

3.1 Deriving ADMM from Douglas–Rachford Splitting

There are several ways deriving ADMM from the Douglas–Rachford splitting method and here we adopt that in [31]. Consider the problem (15) where θ_1 and θ_2 are proper closed convex functions. Its dual function is

$$\begin{aligned} d(\lambda) &= \inf_{x_1,x_2} \theta_1(x_1) + \theta_2(x_2) - \lambda^\top (A_1 x_1 + A_2 x_2 - b) \\ &= \inf_{x_1} \{\theta_1(x_1) - \lambda^\top A_1 x_1\} + \inf_{x_2} \{\theta_2(x_2) - \lambda^\top A_2 x_2\} + \lambda^\top b \\ &= -\sup_{x_1} \{-\theta_1(x_1) + \lambda^\top A_1 x_1\} - \sup_{x_2} \{-\theta_2(x_2) + \lambda^\top A_2 x_2\} + \lambda^\top b \\ &= -\theta_1^*(A_1^\top \lambda) - \theta_2^*(A_2^\top \lambda) + \lambda^\top b, \end{aligned}$$

and the dual problem is

$$\max_{\lambda} \ d(\lambda) = -\theta_1^*(A_1^\top \lambda) - \theta_2^*(A_2^\top \lambda) + \lambda^\top b, \tag{19}$$

where $\lambda \in \mathbb{R}^l$ is the dual variable. From the Fermat's rule, if λ^* is an optimal solution of the dual problem (19), then

$$0 \in \partial d(\lambda^*). \tag{20}$$

Let 'ri,' 'dom,' and '$\circ$' denote the relative interior of a set, the domain of a function, and the composition of two operators, respectively. Assuming that

$$\mathrm{ri}(\mathrm{dom}(\theta_1 \circ A_1^\top)) \cap \mathrm{ri}(\mathrm{dom}\ (\theta_2 \circ A_2^\top)) \neq \varnothing,$$

then it follows from [28, Thm. 23.8] and [28, Thm. 23.9] that

$$\partial d(\lambda) = -A_1 \partial \theta_1^*(A_1^\top \lambda) - A_2 \partial \theta_2^*(A_2^\top \lambda) + b. \tag{21}$$

Problem (20)–(21) is a special case of the zero finding problem

$$0 \in F(\lambda^*) = (F_1 + F_2)(\lambda^*), \tag{22}$$

where

$$F_1(\lambda) = A_1 \partial \theta_1^*(A_1^\top \lambda) - b \ \text{ and } \ F_2(\lambda) = A_2 \partial \theta_2^*(A_2^\top \lambda). \tag{23}$$

For any $\alpha > 0$, we have [30]

$$0 \in (F_1 + F_2)(\lambda) \iff z = (1/2I + 1/2 R_{\alpha F_1} R_{\alpha F_2})(z), \ \lambda = J_{\alpha F_2}(z), \tag{24}$$

where for a maximal monotone operator $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, $J_{\alpha T} = (I + \alpha T)^{-1}$ denotes the resolvent of T, and $R_{\alpha T} = 2J_{\alpha T} - I$ denotes its Cayley operator (reflection operator).

The iteration scheme of Douglas–Rachford splitting method (DRSM) [32] for solving (22) is

$$\begin{aligned}\lambda^{k+1} &= J_{\alpha F_2}(z^k),\\ v^{k+1} &= J_{\alpha F_1}(2\lambda^{k+1} - z^k),\\ z^{k+1} &= z^k + v^{k+1} - \lambda^{k+1}.\end{aligned}$$

Evaluating the resolvent involves a minimization step. Using the structure of F_2 in (23), let $\lambda^{k+1} = J_{\alpha F_2}(z^k)$, and $x_2^{k+1} \in \partial\theta_2^*(A_2^\top\lambda^{k+1})$, then

$$\lambda^{k+1} = J_{\alpha F_2}(z^k) \Longleftrightarrow \lambda^{k+1} = z^k - \alpha A_2 x_2^{k+1}.$$

Recall that for any proper closed convex function f and any vector x, a vector $\xi \in \partial f(x)$ is equivalent to $x \in \partial f^*(\xi)$ [28, Thm 23.5]. Consequently,

$$\begin{aligned}x_2^{k+1} \in \partial\theta_2^*(A_2^\top\lambda^{k+1}) &\Longleftrightarrow A_2^\top\lambda^{k+1} \in \partial\theta_2(x_2^{k+1})\\ &\Longleftrightarrow 0 \in \partial\theta_2(x_2^{k+1}) - A_2^\top z^k + \alpha A_2^\top A_2 x_2^{k+1}\\ &\Longleftrightarrow x_2^{k+1} = \arg\min_{x_2}\left\{\theta_2(x_2) - (z^k)^\top A_2 x_2 + \frac{\alpha}{2}\|A_2x_2\|^2\right\}.\end{aligned}$$

Similarly, for $v^{k+1} = J_{\alpha F_1}(2\lambda^{k+1} - z^k)$, let $\bar{x}_1^{k+1} \in \partial\theta_1^*(A_1^\top v^{k+1})$, then we get

$$\begin{aligned}\bar{x}_1^{k+1} &= \arg\min_{x_1}\left\{\theta_1(x_1) - (z^k - 2\alpha A_2 x_2^{k+1})^\top(A_1x_1 - b) + \frac{\alpha}{2}\|A_1x_1 - b\|^2\right\},\\ v^{k+1} &= z^k - \alpha(A_1\bar{x}_1^{k+1} - b) - 2\alpha A_2 x_2^{k+1}.\end{aligned}$$

Making these explicit together, we obtain

$$\begin{cases} x_2^{k+1} = \arg\min_{x_2}\left\{\theta_2(x_2) - (z^k)^\top A_2x_2 + \frac{\alpha}{2}\|A_2x_2\|^2\right\},\\ \lambda^{k+1} = z^k - \alpha A_2 x_2^{k+1},\\ \bar{x}_1^{k+1} = \arg\min_{x_1}\left\{\theta_1(x_1) - (z^k + 2\alpha A_2x_2^{k+1})^\top(A_1x_1 - b) + \frac{\alpha}{2}\|A_1x_1 - b\|^2\right\},\\ v^{k+1} = z^k - \alpha(A_1\bar{x}_1^{k+1} - b) - 2\alpha A_2x_2^{k+1},\\ z^{k+1} = z^k - \alpha(A_1\bar{x}_1^{k+1} + A_2x_2^{k+1} - b).\end{cases}$$

Removing λ^{k+1} and v^{k+1} and then substituting $z^k = \lambda^k - \alpha(A_1\bar{x}_1^k - b)$, we have

$$\begin{cases} x_2^{k+1} = \arg\min_{x_2}\left\{\theta_2(x_2) - (\lambda^k)^\top A_2x_2 + \frac{\alpha}{2}\|A_1\bar{x}_1^k + A_2x_2 - b\|^2\right\},\\ \lambda^{k+1} = \lambda^k - \alpha(A_1\bar{x}_1^k + A_2x_2^{k+1} - b),\\ \bar{x}_1^{k+1} = \arg\min_{x_1}\left\{\theta_1(x_1) - (\lambda^{k+1})^\top(A_1x_1 - b) + \frac{\alpha}{2}\|A_1x_1 + A_2x_2^{k+1} - b\|^2\right\}.\end{cases}$$

Finally, we swap the order to get the correct dependency and substitute $\bar{x}_1^k = x_1^{k+1}$ to get the alternating direction method of multipliers:

$$\begin{cases} x_1^{k+1} = \arg\min_{x_1} \left\{\theta_1(x_1) - (\lambda^k)^\top A_1 x_1 + \frac{\alpha}{2}\|A_1 x_1 + A_2 x_2^k - b\|^2\right\}, \\ x_2^{k+1} = \arg\min_{x_2} \left\{\theta_2(x_2) - (\lambda^k)^\top A_2 x_2 + \frac{\alpha}{2}\|A_1 x_1^{k+1} + A_2 x_2 - b\|^2\right\}, \\ \lambda^{k+1} = \lambda^k - \alpha(A_1 x_1^{k+1} + A_2 x_2^{k+1} - b). \end{cases}$$

From the above subsection, we can see that the classical alternating direction method of multipliers for solving (15) can be viewed as an application of the Douglas–Rachford splitting method (24) to solving the optimality condition of the dual (19). As a consequence, the global convergence of ADMM can be obtained directly from the convergence result of the Douglas–Rachford splitting method [30, Thm. 8]. Here, we summarize the convergence of ADMM in a similar way of [20, Prop. 4.6].

Theorem 1 *Consider the problem* (15) *where* θ_1 *and* θ_2 *are proper closed convex functions. Let*

$$d_1(\lambda) := \min_{x_1 \in \mathbb{R}^{n_1}} \{\theta_1(x_1) - \langle \lambda, A_1 x_1\rangle\}, \; d_2(\lambda) := \min_{x_1 \in \mathbb{R}^{n_2}} \{\theta_2(x_2) - \langle \lambda, A_2 x_2\rangle\}. \quad (25)$$

Suppose that all subgradients of the functions d_1 *and* d_2 *at each point* $\lambda \in \mathbb{R}^l$ *take the form* $A_1\bar{x}_1$ *and* $A_2\bar{x}_2$*, respectively, where* $\bar{x}_i$, $i = 1, 2$ *attain the stated minimum over* x_i *in* (25)*. Let the constant* $\beta > 0$ *be given and there exists a KKT point for problem* (15)*. Then, the sequences* $\{x_1^k\} \subset \mathbb{R}^{n_1}$, $\{x_2^k\} \subset \mathbb{R}^{n_2}$, *and* $\{\lambda^k\} \subset \mathbb{R}^l$ *conforming to the recursions* (17) *converge, i.e.,* $\lambda^k \to \lambda^\infty$, $A_1 x_1^k \to A_1 x_1^\infty$, $A_2 x_2^k \to A_2 x_2^\infty$, *where* $(x_1^\infty, x_2^\infty, \lambda^\infty)$ *is some KKT point for problem* (15)*.*

3.2 Selection of the Penalty Parameter

The penalty parameter β in the augmented Lagrangian function plays an important role for the methods, both ALM and ADMM. Experience on applications has shown that, if the fixed penalty β is chosen too small or too large, the efficiency of the methods can be degraded significantly. In fact, there is no 'optimal' fixed parameter. For ALM, Rockafellar proposed that it should be varied along the iteration [33,34]. For ADMM, He and Yang [35] suggest to choose β_k either in an increasing manner or in a decreasing manner; [36] took a decreasing sequence of penalty symmetric positive-definite matrices; and in [37,38], the authors designed a *self-adaptive* strategy for choosing the parameter, i.e., the parameter can be increased and can be decreased according to some rules. The adjusting rule in [37,38] is:

$$\beta^{k+1} := \begin{cases} (1+\mu)\beta^k, & \text{if } \|r^k\| > \nu_1\|s^k\|, \\ \beta^k/(1+\mu), & \text{if } \|r^k\| < \nu_2\|s^k\|, \\ \beta^k, & \text{otherwise}, \end{cases} \quad (26)$$

where μ, ν_1, ν_2 are positive parameters, and

$$r^k = A_1 x_1^k + A_2 x_2^k - b$$

and

$$s^k = \beta^k A_1^\top A_2 (x_2^{k+1} - x_2^k)$$

denote the primal error and the dual error associated with the iteration, respectively.

The philosophy behind this 'simple scheme that often works well' [15] is as follows. The optimality conditions for (15) are primal feasibility

$$0 = A_1 x_1^* + A_2 x_2^* - b, \tag{27}$$

and dual feasibility

$$0 \in \partial\theta_1(x_1^*) - A_1^\top \lambda^*,$$

and

$$0 \in \partial\theta_2(x_2^*) - A_2^\top \lambda^*. \tag{28}$$

From the ADMM iteration scheme (17) we can see that

$$0 \in \partial\theta_1(x_1^{k+1}) - A_1^\top \lambda^{k+1} - \beta_k A_1^\top A_2 (x_2^k - x_2^{k+1}),$$

and

$$0 \in \partial\theta_2(x_2^{k+1}) - A_2^\top \lambda^{k+1}.$$

Hence, we can see that

$$\begin{pmatrix} r^{k+1} \\ s^{k+1} \\ 0 \end{pmatrix} \in \begin{pmatrix} A_1 x_1^{k+1} + A_2 x_2^{k+1} - b \\ \partial\theta_1(x_1^{k+1}) - A_1^\top \lambda^{k+1} \\ \partial\theta_2(x_2^{k+1}) - A_2^\top \lambda^{k+1} \end{pmatrix},$$

and the primal residual and the dual residual of the iterative scheme (17) to the optimality conditions (27)–(28) are r^{k+1} and s^{k+1}. Since ADMM is a primal–dual algorithm, it will be desirable that the primal residual and the dual residual behave in a coherent way. The adjust rule (26) is exactly that when comparing to the dual residual the primal residual is too large, it increases the penalty parameter; when comparing to the dual residual the primal residual is too small, it decreases the penalty parameter; otherwise, there is no need of adjustment for the parameter.

We have the following convergence results on the alternating direction methods of multipliers with self-adaptive parameter selecting strategy [37, Theorem 4.1].

Theorem 2 *Consider the optimization problem* (15) *and let* $\{x_1^k, x_2^k, \lambda^k\}$ *be the sequence generated by the self-adaptive ADMM,*

$$\begin{cases} x_1^{k+1} = \arg\min\limits_{x_1} \mathcal{L}_{\beta_k}(x_1, x_2^k, \lambda^k), \\ x_2^{k+1} = \arg\min\limits_{x_2} \mathcal{L}_{\beta_k}(x_1^{k+1}, x_2, \lambda^k), \\ \lambda^{k+1} = \lambda^k - \tau\beta_k(A_1x_1^{k+1} + A_2x_2^{k+1} - b), \end{cases} \tag{29}$$

where $\tau \in \left(0, \frac{1+\sqrt{5}}{2}\right)$, *and* $\{\beta^k\}$ *is selected according to the strategy* (26). *Then,* $\lambda^k \to \lambda^\infty$, $A_1x_1^k \to A_1x_1^\infty$, $A_2x_2^k \to A_2x_2^\infty$, *where* $(x_1^\infty, x_2^\infty, \lambda^\infty)$ *is some KKT point of* (15).

3.3 Easier Subproblems

One main reason for the renaissance of ADMM is that when applying to the modern application models, the subproblems are easy to solve, and in fact, in many cases, they possess closed-form solutions. For example, for solving the ℓ_1-norm minimization problem (5), the iterative scheme is

$$\begin{aligned} x^{k+1} &:= (A^\top A + \beta I)^{-1}(A^\top b + \beta(y^k - \lambda^k)), \\ y^{k+1} &:= \mathcal{S}_{\nu/\beta}(x^{k+1} + \lambda^k), \\ \lambda^{k+1} &:= \lambda^k - (x^{k+1} - y^{k+1}), \end{aligned}$$

where for $\alpha > 0$, $\mathcal{S}_\alpha$ is the soft thresholding operator and is defined as

$$S_\alpha(x) := \begin{cases} x - \alpha, & \text{if } x > \alpha, \\ 0, & \text{if } |x| \leqslant \alpha, \\ x + \alpha, & \text{if } x < -\alpha. \end{cases}$$

The x-minimization is a system of linear equations, and the coefficient matrix is symmetric and positive definite. Hence, it can be solved easily; particularly, when A has some circulant structure as those in signal/image processing, the system can be solved via fast Fourier transform and the computation cost is very low. However, when we have to solve the subproblems via iterative algorithms, we need to balance the cost on the subproblems solving and the outer iteration. That is, we need design variant of the classical ADMM with easier subproblems.

In applications, there are usually simple constraints for the variables, e.g., in image processing, the value of pixels should be bounded in [0, 255] or [0, 1], and we have the following optimization model [39]:

$$\min \left\{ \theta_1(x_1) + \theta_2(x_2) \;\middle|\; A_1x_1 + A_2x_2 = b,\ x_1 \in \mathcal{X}_1,\ x_2 \in \mathcal{X}_2 \right\}, \tag{30}$$

where $\mathcal{X}_i$ are closed convex subsets of $\mathbb{R}^{n_i}$, $i = 1, 2$. Certainly, we can move the constraints to the objective function via the indicator function for a given closed

convex set S

$$\delta_S(x) = \begin{cases} 0, & \text{if } x \in S, \\ \infty, & \text{otherwise.} \end{cases}$$

Nevertheless, the corresponding subproblems are still constrained optimization problems, which usually exclude the possibility of having closed-form solutions, even for the case that both the function θ_i and A_i are simple enough, e.g., when θ_i is the ℓ_1-norm and A_i is the identity. In [40], the authors designed a new variant ADMM-type method, in which the main cost is to solve the following optimization problems

$$\begin{cases} x_1^{k+1} = \arg\min\limits_{x_1 \in \mathbb{R}^{n_1}} \left\{ \theta_1(x_1) + \dfrac{1}{2\beta} \left\| x_1 - s_1^k \right\|^2 \right\}, \\ x_2^{k+1} = \arg\min\limits_{x_2 \in \mathbb{R}^{n_2}} \left\{ \theta_2(x_2) + \dfrac{1}{2\beta} \left\| x_2 - s_2^k \right\|^2 \right\}, \end{cases} \tag{31}$$

where $s_i^k := x_i^k + \beta\xi_i^k - \gamma\alpha_k e_i(x_i^k, \lambda^k))$, $\xi_i^k = (1/\beta)(s_i^{k-1} - x_i^k) \in \partial\theta_i(x_i^k)$, and

$$e_i(x_i^k, \lambda^k) = x_i^k - P_{\mathcal{X}_i}[x_i^k - \beta(\xi_i^k - A_i^\top \lambda^k)],$$

α_k is the step size that computed according to some formula to ensure convergence, which can also be chosen smaller than a threshold.

Comparing the iterative scheme (31) with the ADMM scheme (17), we can find that

1. No matter if there is a constraint $\mathcal{X}_i$ for the ith-minimization problem, and no matter what the constraint matrix A_i is, the ith-minimization problem in (31) is always the evaluation of the proximal operator of the component function θ_i at a given s_i^k. This will be particularly advantageous when both the function θ_i and the set $\mathcal{X}_i$ are simple in the sense that evaluation their individual proximal operator (projection onto the set) is easy, while evaluation the proximal operator jointly is difficult.
2. The iterative scheme (31) is essentially coming from the application of the Douglas–Rachford splitting method [32] to the optimality condition for (15), following with a simple decouple skill. Recall that ADMM is the application of the Douglas–Rachford splitting method [32] to the dual of (15), and (31) provides an alternative for solving (15). The numerical results reported in [40] show the advantage of (31) over ADMM, especially when the subproblems of ADMM do not possess closed-form solutions while (31) has.

The following result shows that the alternating direction methods of multipliers with easier subproblems is convergent [40, Thm. 4.1].

Theorem 3 *Consider the optimization problem* (30) *and let* $\{x_1^k, x_2^k, \lambda^k\}$ *be the sequence generated by* (31)*. Then,* $\{(x_1^k, x_2^k, \lambda^k)\}$ *converges to a solution of* (30)*.*

***Remark* 1** Besides the numerical advantage listed above, the customized decomposition algorithm also has some advantages from the theoretical point of view. In

particular, as stated in Theorem 3, the 'whole' sequence $\{(x_1^k, x_2^k, \lambda^k)\}$ converges to a solution $\{(x_1^\infty, x_2^\infty, \lambda^\infty)\}$.

3.4 Approximate Solutions of the Subproblems

The slow 'tail convergence' in ADMM is used to be criticized by some researchers. However, in recent applications, the large-scale optimization problems are not required to be solved to obtain solutions with extremely high accuracy. Hence, there is no need to get solutions with high accuracy per iteration. On the same time, when the subproblems are solved numerically, we must design some stopping criteria to exit the inner iteration. Only after this, ADMM can be a practical algorithm. Although it is a long history adopting approximate solutions of subproblems [41], the first inexact ADMM comes from [38]. In [38], besides selecting the penalty parameter in a self-adaptive way, they designed an accuracy criterion, under which the algorithm still converges.

To get an inexact ADMM, the authors [38] modified the subproblems in (16) by adding a quadratic term as done in [42], and the iterative scheme is

$$\begin{cases} x_1^{k+1} \approx \arg\min\limits_{x_1} \mathcal{L}_\beta(x_1, x_2^k, \lambda^k) + \frac{1}{2}\|x_1 - x_1^k\|^2_{H_1^k}, \\ x_2^{k+1} \approx \arg\min\limits_{x_2} \mathcal{L}_\beta(x_1^{k+1}, x_2, \lambda^k) + \frac{1}{2}\|x_2 - x_2^k\|^2_{H_2^k}, \\ \lambda^{k+1} = \lambda^k - \beta(A_1 x_1^{k+1} + A_2 x_2^{k+1} - b). \end{cases} \tag{32}$$

The accuracy criterion is

$$\|x_i^{k+1} - \tilde{x}_i^{k+1}\| \leqslant \nu_{i,k}. \tag{33}$$

The scaler sequences $\{\nu_{i,k}\}$ are nonnegative sequences satisfying $\sum_{k=1}^\infty \nu_{i,k} < +\infty$, $H_i^k \succ \alpha I$ with $\alpha > 0$, and $\tilde{x}_i^{k+1}$ are the exact solutions of the corresponding subproblems.

At the first glance, it seems the accuracy criterion (33) is impractical, since it involves the exact solution $\tilde{x}_i^{k+1}$. However, since the objective functions in (33) are strongly convex with modulus $r_k > \alpha > 0$, according to Lemma 1 in [38], we have that for x_1-minimization problem

$$\|x_1^{k+1} - \tilde{x}_1^{k+1}\|^2 \leqslant 2\mu\langle E_{1,\mu}(x_1^{k+1}), f_{1,k}(x_1^{k+1})\rangle - \|E_{1,\mu}(x_1^{k+1})\|^2, \quad \forall\, \mu \geqslant r_k^{-1}, \tag{34}$$

where

$$E_{1,\mu}(x_1^{k+1}) = x_1^{k+1} - P_{\mathcal{X}_1}[x_1^{k+1} - \mu f_{1,k}(x_1^{k+1})], \tag{35}$$

and $\xi_1^{k+1} \in \partial\theta_1(x_1^{k+1})$,

$$f_{1,k}(x_1^{k+1}) := \xi_1^{k+1} - A_1^\top \lambda^k + \beta A_1^\top (A_1 x_1^{k+1} + A_2 x_2^k - b) + H_1^k(x_1^{k+1} - x_1^k). \tag{36}$$

For the exact solution $\tilde{x}_1^{k+1}$, replacing x_1^{k+1} with $\tilde{x}_1^{k+1}$ in (35) and $\xi_1^{k+1} \in \partial\theta_1(x_1^{k+1})$ with $\tilde{\xi}_1^{k+1} \in \partial\theta_1(\tilde{x}_1^{k+1})$ in (36), we have

$$E_{1,\mu}(\tilde{x}_1^{k+1}) = 0, \qquad \mu\langle E_{1,\mu}(\tilde{x}_1^{k+1}), f_{1,k}(\tilde{x}^{k+1})\rangle - \|E_{1,\mu}(\tilde{x}_1^{k+1})\|^2 = 0$$

and recall that (see Eq. (2.2) in [38])

$$2\mu\langle E_{1,\mu}(x_1^{k+1}), f_{1,k}(x_1^{k+1})\rangle - \|E_{1,\mu}(x_1^{k+1})\|^2 \geqslant \|E_{1,\mu}(x_1^{k+1})\|^2.$$

We can take $\mu \geqslant \alpha^{-1}$ and find x_1^{k+1} such that

$$2\mu\langle E_{1,\mu}(x_1^{k+1}), f_{1,k}(x_1^{k+1})\rangle - \|E_{1,\mu}(x_1^{k+1})\|^2 \leqslant \nu_{1,k}^2. \tag{37}$$

This guarantees that $\|x_1^{k+1} - \tilde{x}_1^{k+1}\| \leqslant \nu_{1,k}$. Note that there is no $\tilde{x}_1^{k+1}$ in (37). Inequality (37) provides a practical and achievable condition of satisfying (33). Following the same discussion, a similar condition of satisfying (33) can be established for the x_2-minimization problem.

The inequality (34) is an error bound inequality [43] in the sense that the left is the distance between a point x_1^{k+1} and a solution point $\tilde{x}_1^{k+1}$, and the right-hand side is a function of x_1^{k+1}. This is due to the quadratic terms added to the objective functions, which makes the objective function being strongly convex. Besides this property, another important effect is to improve the condition number of the problem. Recall that the convergence property of the numerical algorithms for solving the optimization problems, from the steepest gradient method, the conjugate gradient method, to the Newton's type methods, all depend on the condition number. From this point of view, the quadratic term makes more state-of-the-art solvers to be available.

Note that we also have freedom in choosing the matrix H_k, and in [38], it suggests that we choose it in a self-adaptive way as (26). Note however, in some applications as those listed in the introduction part, the original objective function θ_i is easy enough in the sense that it possesses the separable structure, e.g., the ℓ_1-norm, which is the sum of the absolute of its entries, while the coefficient matrix A_i may not possess any structure. In other words, the difficulty in solving the ADMM subproblems is caused by the augmented quadratic term. In this case, we can select $H_i^k = \mu_{i,k} I - \beta A_i^\top A_i$, leading to the proximal operator evaluating subproblems:

$$\begin{cases} x_1^{k+1} = \underset{x_1 \in \mathbb{R}^{n_1}}{\arg\min} \left\{ \theta_1(x_1) + \dfrac{\mu_{1,k}}{2} \left\| x_1 - (x_1^k + (1/\mu_{1,k}) A_1^\top \bar{\lambda}^k) \right\|^2 \right\}, \\ x_2^{k+1} = \underset{x_2 \in \mathbb{R}^{n_2}}{\arg\min} \left\{ \theta_2(x_2) + \dfrac{\mu_{2,k}}{2} \left\| x_2 - (x_2^k + (1/\mu_{2,k}) A_2^\top \hat{\lambda}^k) \right\|^2 \right\}, \\ \lambda^{k+1} = \lambda^k - \beta(A_1 x_1^{k+1} + A_2 x_2^{k+1} - b), \end{cases} \tag{38}$$

where

$$\bar{\lambda}^k = \lambda^k - \beta(A_1 x_1^k + A_2 x_2^k - b),$$

and

$$\hat{\lambda}^k = \lambda^k - \beta(A_1 x_1^{k+1} + A_2 x_2^k - b).$$

The iterative scheme (38) is called *linearized alternating direction method of multipliers* [44–46], whose main computational task is the proximal operator evaluating, and has shown its advantage in solving modern optimization problems arising from statistics [47].

The technique of linearizing the quadratic term of the augmented Lagrangian function to reduce the computational cost in solving the subproblems can be dated back to the work in [48], where the authors proposed the following iterative scheme

$$\begin{cases} x_1^{k+1} = \arg\min\limits_{x_1 \in \mathbb{R}^{n_1}} \left\{ \theta_1(x_1) + \dfrac{\mu_{1,k}}{2} \left\| x_1 - (x_1^k + (1/\mu_{1,k}) A_1^\top \bar{\lambda}^k) \right\|^2 \right\}, \\ x_2^{k+1} = \arg\min\limits_{x_2 \in \mathbb{R}^{n_2}} \left\{ \theta_2(x_2) + \dfrac{\mu_{2,k}}{2} \left\| x_2 - (x_2^k + (1/\mu_{2,k}) A_2^\top \bar{\lambda}^k) \right\|^2 \right\}, \\ \lambda^{k+1} = \lambda^k - \beta(A_1 x_1^{k+1} + A_2 x_2^{k+1} - b). \end{cases} \tag{39}$$

Comparing the two augmented Lagrangian-based splitting schemes (38) and (39), we can find that the only difference between them is that (38) solves the two minimization problems in a Gauss–Seidel manner, while (39) solves them in a Jacobian manner. Some variants of (39) can be found in [49].

In some applications, each component function θ_i itself is composed with a smooth convex function and a simple nonsmooth function, i.e., $\theta_i = \vartheta_i + \iota_i$ where ϑ_i is a smooth convex function and ι_i is a nonsmooth convex function. The optimization model (15) is

$$\min \left\{ \vartheta_1(x_1) + \iota_1(x_1) + \vartheta_2(x_2) + \iota_2(x_2) \;\middle|\; A_1 x_1 + A_2 x_2 = b \right\}. \tag{40}$$

There are two smooth parts in the x_i-minimization problem in (17), ϑ_i and the quadratic term. One can choose to either linearize one of them, or linearize both of them, depending on the problem data's structure. If one linearizes both of them, the resulting linearized ADMM has the following iterative scheme:

$$\begin{cases} x_1^{k+1} = \arg\min\limits_{x_1 \in \mathbb{R}^{n_1}} \left\{ \iota_1(x_1) + \dfrac{\mu_{1,k}}{2} \left\| x_1 - (x_1^k + (1/\mu_{1,k})[\nabla \vartheta_1(x_1^k) + A_1^\top \bar{\lambda}^k]) \right\|^2 \right\}, \\ x_2^{k+1} = \arg\min\limits_{x_2 \in \mathbb{R}^{n_2}} \left\{ \iota_2(x_2) + \dfrac{\mu_{2,k}}{2} \left\| x_2 - (x_2^k + (1/\mu_{2,k})[\nabla \vartheta_2(x_2^k) + A_2^\top \hat{\lambda}^k]) \right\|^2 \right\}, \\ \lambda^{k+1} = \lambda^k - \beta(A_1 x_1^{k+1} + A_2 x_2^{k+1} - b), \end{cases} \tag{41}$$

and the x_i-minimization problem is simply the evaluating of the proximal operator of the nonsmooth convex function ι_i. Also, the minimization subproblems can be solved

in a parallel manner,

$$\begin{cases} x_1^{k+1} = \arg\min\limits_{x_1 \in \mathbb{R}^{n_1}} \left\{ \iota_1(x_1) + \dfrac{\mu_{1,k}}{2} \left\| x_1 - (x_1^k + (1/\mu_{1,k})[\nabla \vartheta_1(x_1^k) + A_1^\top \bar{\lambda}^k]) \right\|^2 \right\}, \\ x_2^{k+1} = \arg\min\limits_{x_2 \in \mathbb{R}^{n_2}} \left\{ \iota_2(x_2) + \dfrac{\mu_{2,k}}{2} \left\| x_2 - (x_2^k + (1/\mu_{2,k})[\nabla \vartheta_2(x_2^k) + A_2^\top \bar{\lambda}^k]) \right\|^2 \right\}, \\ \lambda^{k+1} = \lambda^k - \beta(A_1 x_1^{k+1} + A_2 x_2^{k+1} - b). \end{cases} \tag{42}$$

The inequality (33) and the *summable* requirement on the error parameter sequences $\{\nu_{i,k}\}$ (or squares summable [50]) makes it not always handy in applications. Actually, one will prefer to *relative* error criteria. The relative criterion was introduced in [51] for finding the zero of a maximal monotone operator, which includes convex optimization and monotone variational inequality problems as special cases. Using an approximate solution under relative accuracy criterion, the next iterate is generated by a projection onto a hyperplane. Numerical algorithms with relative accuracy criteria are further studied in [52–54]. For the problem (1) with one block, i.e., $m = 1$, [55] proposed a practical relative error criterion for augmented Lagrangian method (13). Exploiting the relationship between the ADMM and both the proximal point algorithm and Douglas–Rachford splitting method for maximal monotone operators, recently, [56,57] proposed two new inexact ADMMs. In [56], a summable criterion and a relative criterion were presented. The relative criterion is presented in [56,57], while it is restrictive in the sense that it allows only one of the two subproblems to be minimized approximately, which covers commonly encountered special cases such as lasso. Most recently, [58] proposed two relative criteria for ADMM. In the first one, it was also restricted that only one subproblem can be solved approximately, and to improve the numerical performance, the parameter τ can be larger. In the second one, both subproblems were allowed to be solved approximately, while the dual step size parameter τ was restricted less than 1.

4 Rate of Convergence

With the popular of the research on the numerical variants and applications of the alternating direction method of multipliers, theoretical analysis on its convergence behavior, in addition of global convergence, rate of convergence attracts more and more attentions.

4.1 Sublinear Rate

For first-order-based method, a measure for the convergence speed is the iteration complexity, $O(1/t)$, $O(1/t^2)$ and so on, where t is the iteration counter [59]. A worst-case $O(1/t)$ convergence rate means that the solution accuracy under certain criteria is of the order $O(1/t)$ after t iterations of the iterative scheme, or equivalently, that it requires at most $O(1/\varepsilon)$ iterations to find a solution to an accuracy of ε. For the convex optimization problem (15), ADMMs convergence rate appears to be at most

sublinear, and [60,61] showed it has a worst-case iteration complexity of $O(1/t)$ in an ergodic sense[2].

There are several different measure functions in the literature. Here, we introduced two of them. The first one is as follows:

Definition 2 $\tilde{w}$ is an ε-solution of (15), if

$$g(\tilde{w}) \leqslant \varepsilon,$$

where

$$g(\tilde{w}) = \sup_{w \in \Omega \cap \mathcal{D}_{\tilde{w}}} \{\theta(\tilde{\boldsymbol{x}}) - \theta(\boldsymbol{x}) + \langle \tilde{w} - w, F(w) \rangle\}. \tag{43}$$

Here, we use the notation

$$\theta(x) = \theta_1(x_1) + \theta_2(x_2), \ \Omega := \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \mathbb{R}^l,$$

and

$$\mathcal{D}_{\tilde{w}} := \{w \mid \|w - \tilde{w}\| \leqslant \delta\}$$

with $\delta > 0$ being a scalar. Moreover,

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad w = \begin{pmatrix} x_1 \\ x_2 \\ \lambda \end{pmatrix}, \quad F(w) = \begin{pmatrix} -A_1^\top \lambda \\ -A_2^\top \lambda \\ A_1 x_1 + A_2 x_2 - b \end{pmatrix}. \tag{44}$$

The above definition was used in [60]. It arises essentially from characterizing the optimality condition of (15) with a variational inequality (VI), i.e., finding

$$w^* := (x_1^*, x_2^*, \lambda^*) \in \Omega := \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \mathbb{R}^l$$

such that

$$\mathrm{VI}(\Omega, F, \theta) : \theta(x) - \theta(x^*) + \langle w - w^*, F(w^*) \rangle \geqslant 0, \quad \forall\, w \in \Omega, \tag{45}$$

Let $\{w^k\}$ be the sequence generated by the ADMM (17), and define

$$\tilde{w}^k := \begin{pmatrix} x_1^{k+1} \\ x_2^{k+1} \\ \lambda^k - \beta(A_1 x_1^{k+1} + A_2 x_2^k - b) \end{pmatrix} \quad \text{and} \quad H := \begin{pmatrix} 0 & 0 & 0 \\ 0 & \beta A_2^\top A_2 & 0 \\ 0 & 0 & \frac{1}{\beta} I_m \end{pmatrix}.$$

[2] The convergence rate is in the 'ergodic' sense means that the approximate solution with a certain accuracy is found based on all k iterates.

[60] proved that

$$\theta(\tilde{x}_t) - \theta(x) + \langle \tilde{w}_t - w, F(w)\rangle \leqslant \frac{1}{2(t+1)}\|w - w^0\|_H^2, \quad \forall\, w \in \Omega, \tag{46}$$

where

$$\tilde{w}_t = \frac{1}{t+1}\sum_{k=0}^{t} \tilde{w}^k.$$

Then, to obtain the conclusion, set

$$d := \sup\{\|w - w^0\|_H \mid w \in \mathcal{D}_{\tilde{w}_t}\};$$

consequently, for a given accuracy tolerance $\varepsilon > 0$, it holds that after at most $\lceil \frac{d^2}{2\varepsilon} - 1 \rceil$ iterations,

$$g(\tilde{w}_t) \leqslant \varepsilon. \tag{47}$$

There are two main differences between (47) and (45). The first one is that the inequality has been relaxed from zero to a nonnegative parameter ε; and the second one is that there is an artificial constraint $\mathcal{D}_{\tilde{w}_t}$. Although (47) reduces to (45) when $\varepsilon = 0$, the added constraint $\mathcal{D}_{\tilde{w}_t}$ means that they are vastly different when $\varepsilon > 0$. In fact, if there is no constraint $\mathcal{D}_{\tilde{w}_t}$, the function defined by (43) is exactly the dual-gap function for the variational inequality characterization (45) [43, Eq.(2.3.13)] satisfying

$$0 \leqslant g(w) \leqslant \infty, \ \forall w \in \Omega,$$

and if $g(w) = 0$, then w is a solution for (45). Hence, g can be used as an approximate function for variational inequality characterization [43]. The measure (43), due to the restrict $\mathcal{D}_{\tilde{w}_t}$, may cause the measure too loose [62].

We now introduce the second measure. Since (15) is a constrained optimization problem, a natural way measuring an approximate solution is that both the objective function value and the constraint validation [62,63].

Definition 3 Consider the constrained optimization problem

$$\min \theta(u), \quad u \in \Omega.$$

We say that u^* is an ε-optimal solution $\tilde{u}$ if

$$\theta(\tilde{u}) - \theta(u^*) \leqslant \varepsilon, \quad \text{and} \quad d \leqslant \varepsilon, \quad \text{where} \quad d := \inf\{\|\tilde{u} - u\| \mid u \in \Omega\}, \tag{48}$$

where the first inequality represents the objective function approximation and the second one represents the constraint violation.

Cai and Han [62] considered the sublinear rate of convergence of the generalized ADMM (G-ADMM) proposed in [30]:

$$\begin{cases} x^{k+1} = \arg\min \left\{ \theta_1(x_1) - x_1^{\mathbf{A}^\top} A_1^\top \lambda^k + \dfrac{\beta}{2} \|A_1 x_1 + A_2 x_2^k - b\|^2 \right\}, & \text{(49a)} \\ x_2^{k+1} = \arg\min \left\{ \theta_2(x_1) - x_2^{\mathbf{A}^\top} A_2^\top \lambda^k + \dfrac{\beta}{2} \|\alpha A_1 x_1^{k+1} - (1-\alpha)(A_2 x_2^k - b) + A_2 x_2 - b\|^2 \right\}, & \text{(49b)} \\ \lambda^{k+1} = \lambda^k - \beta(\alpha A_1 x_1^{k+1} - (1-\alpha)(A_2 x_2^k - b) + A_2 x_2^{k+1} - b), & \text{(49c)} \end{cases}$$

where the parameter $\alpha \in (0, 2)$ is an acceleration factor. It is usually suggested that we take $\alpha \in (1, 2)$. Note that the original ADMM scheme (17) is a special case of the G-ADMM scheme (49a)–(49c) with $\alpha = 1$. Using these measures, it is proved that G-ADMM needs at most $\lceil \frac{\omega}{\varepsilon} \rceil$ iterations to obtain an ε-optimal solution, where ω is

$$\omega := \frac{1}{2} \sup_{\|\lambda\| \leqslant \rho} \left\| \begin{matrix} y^0 - \hat{y} \\ \lambda^0 - \lambda \end{matrix} \right\|^2_{H_{\alpha\beta}} \quad \text{and} \quad H_{\alpha\beta} := \begin{pmatrix} \frac{\beta}{\alpha} B^{\mathbf{A}^\top} B & \frac{1-\alpha}{\alpha} B^{\mathbf{A}^\top} \\ \frac{1-\alpha}{\alpha} B & \frac{1}{\alpha\beta} I_m \end{pmatrix}. \tag{50}$$

Comparing with results in [60], the number of iterations in the above results required is dependent of the distance between the initial point w^0 and the solution set, which is meaningful since usually the number of iterations should be larger when the initial point is further away from the solution set and smaller when it is nearer to the solution set.

The results in [60,62] are in the ergodic sense, since they used the information from the whole iterative sequence. If we can establish the monotonic decrease in the measure sequence, the same results under the non-ergodic sense can be obtained, i.e., the measure only uses the current iterative information. Moreover, the complexity is essentially $o(1/t)$ instead of $O(1/t)$ [64]. For more results on the sublinear convergence of ADMM, we can refer to [65–67].

4.2 Linear Rate

Compared with the large amount of literature mainly being devoted to the applications of the ADMM, there is a much smaller number of papers targeting the linear rate, in particular the Q-linear rate, convergence analysis. Recall that the classical ADMM for the two-block model (15) can be viewed as the application of the Douglas–Rachford splitting method applied to the dual problem of (15), and the Douglas–Rachford splitting method can be viewed as a special application of the proximal point algorithm (PPA) for certain maximal monotone operator. As a consequence, one can derive the corresponding R-linear rate convergence of the ADMM from [68] on the Douglas–Rachford splitting method with a globally Lipschitz continuous and strongly monotone operator, and [33,41,69] on the convergence rates of the PPAs under various error bound conditions imposed on certain splitting operator. Using these connections. An early work [70] proved the global R-linear convergence rate of the ADMM when it is applied to linear programming.

Recently, more and more interesting developments on the linear convergence rate of the ADMM were established.

For convex quadratic programming, [71] provided a local linear convergence result under the conditions of the uniqueness of the optimal solutions to both the primal and dual problems and the strict complementarity. The quadratic programming problem considered in [71] is

$$\min \ \frac{1}{2}x^\top Qx + q^\top x \ \text{ s.t. } \ Cx = c, \ x \geqslant 0, \tag{51}$$

and to use ADMM, it was reformulated into

$$\min \ \frac{1}{2}x^\top Qx + q^\top x \ \text{ s.t. } \ Cx = c, \ x = z, \ z \geqslant 0, \tag{52}$$

which is a special case of the separable quadratic programming problem. A direct using of the classical ADMM yields the procedure

1. Set

$$x^{k+1} = \arg\min_x \frac{1}{2}x^\top Qx + q^\top x - x^\top \lambda^k + \frac{\beta}{2}\|x - z^k\|^2 \text{ s.t. } Cx = c; \tag{53}$$

2. Set

$$z^{k+1} = \arg\min_z z^\top \lambda^k + \frac{\beta}{2}\|x^{k+1} - z\|^2 \text{ s.t. } z \geqslant 0; \tag{54}$$

3. Set

$$\lambda^{k+1} = \lambda^k - \beta(x^{k+1} - z^{k+1}).$$

The first step (53) is equivalent to a well-conditioned saddle-point problem, which is well-studied in linear algebra society. The second step (54) is essentially the orthogonal projection onto the nonnegative orthant,

$$z^{k+1} = \max\{0, x^{k+1} - \lambda^k/\beta\}.$$

These special structures in the algorithm enable to use tools from numerical linear algebra to establish the local linear convergence of ADMM under the conditions mentioned above [71].

In [72], the authors established the local linear rate convergence of the generalized ADMM in the sense of [30], which does not assume the restrictive conditions such as uniqueness and strict complementarity of the solution. In the following, we introduce their results for the classical ADMM. The quadratic programming problem considered there is

$$\min \ \frac{1}{2}x^\top Qx + q^\top x + \frac{1}{2}y^\top Ry + r^\top y \ \text{ s.t. } \ Ax + By = b, \ x \in \mathcal{X}, y \in \mathcal{Y}, \tag{55}$$

where Q and R are symmetric positive semidefinite matrices in $\mathbb{R}^{n\times n}$, and $\mathbb{R}^{m\times m}$, respectively; $A \in \mathbb{R}^{l\times n}$ and $B \in \mathbb{R}^{l\times m}$ are two given matrices, and $q \in \mathbb{R}^n$, $r \in \mathbb{R}^m$, and $b \in \mathbb{R}^l$ are given vectors. $\mathcal{X} = \{x | Cx = c,\ x \in \mathbb{R}^n(\mathbb{R}^n_+)\}$ and $\mathcal{Y} = \{y | Dy = d,\ y \in \mathbb{R}^m(\mathbb{R}^m_+)\}$ are two polyhedral sets. (55) includes (52) as a special case with $\mathcal{X} = \{x | Cx = c,\ x \in \mathbb{R}^n\}$; $\mathcal{Y} = \{y | y \in \mathbb{R}^m_+\}$; R and r are zero matrix and zero vector, respectively; and $A = -B = I$. The first key result established in [72] is the following lemma.

Lemma 3 *The sequence* $\{w^k := (x^k, y^k, \lambda^k)\}$ *generated by the ADMM satisfies*

$$\begin{aligned}\left\| \begin{matrix} y^{k+1} - y^* \\ \lambda^{k+1} - \lambda^* \end{matrix} \right\|^2_H &\leqslant \left\| \begin{matrix} y^k - y^* \\ \lambda^k - \lambda^* \end{matrix} \right\|^2_H - \frac{1}{\beta} \left\| \begin{matrix} y^k - y^{k+1} \\ \lambda^k - \lambda^{k+1} \end{matrix} \right\|^2_G \\ &\leqslant \left\| \begin{matrix} y^k - y^* \\ \lambda^k - \lambda^* \end{matrix} \right\|^2_H - \frac{1}{\beta} \left\| \begin{matrix} y^k - y^{k+1} \\ \lambda^k - \lambda^{k+1} \end{matrix} \right\|^2_H, \end{aligned} \tag{56}$$

where

$$G := \begin{pmatrix} \beta^2 B^\top B & \beta B^\top \\ \beta B & I_m \end{pmatrix} \quad \text{and} \quad H := \begin{pmatrix} \beta B^\top B & 0 \\ 0 & \frac{1}{\beta} I_m \end{pmatrix}. \tag{57}$$

Based on (56), the main task is to bound the first term or the second term by the third one, i.e., bound $\left\| \begin{matrix} y^{k+1} - y^* \\ \lambda^{k+1} - \lambda^* \end{matrix} \right\|^2_H$ or $\left\| \begin{matrix} y^k - y^* \\ \lambda^k - \lambda^* \end{matrix} \right\|^2_H$ by $\left\| \begin{matrix} y^k - y^{k+1} \\ \lambda^k - \lambda^{k+1} \end{matrix} \right\|^2_H$.

To this end, recall the residual function of the first-order optimality condition for (55)

$$e(w, \beta) := \begin{pmatrix} e_{\mathcal{X}}(w, \beta) \\ e_{\mathcal{Y}}(w, \beta) \\ e_{\Lambda}(w, \beta) \end{pmatrix} := \begin{pmatrix} x - P_{\mathcal{X}}[x - \beta(Qx + q - A^\top \lambda)] \\ y - P_{\mathcal{Y}}[y - \beta(Ry + d - B^\top \lambda)] \\ Ax + By - b \end{pmatrix},$$

and the following error bound result [43,73].

Lemma 4 *Let* Ω *be a polyhedral set. Then, there exists scalars* $\varepsilon > 0$ *and* $\tau > 0$ *such that*

$$d(w, \Omega^*) \leqslant \tau \|e(w, 1)\| \tag{58}$$

for all $w \in \Omega$ *with* $\|e(w, \beta)\| \leqslant \varepsilon$, *where* Ω^* *is the set of primal dual solutions of the first-order optimality condition for* (55), $d(w, \Omega^*)$ *is the distance from* w *to the set* Ω^*.

A basic property for the residual is that for a given $w \in \Omega$, the magnitude $\|e(w, \beta)\|$ is increasing with β, while $\|e(w, \beta)\|/\beta$ is decreasing with β (see [74] for a simple proof). That is, for any $\tilde{\beta} \geqslant \beta > 0$ and $w \in \Omega$,

$$\|e(w, \tilde{\beta})\| \geqslant \|e(w, \beta)\|$$

and

$$\frac{\|e(w, \tilde{\beta})\|}{\tilde{\beta}} \leqslant \frac{\|e(w, \beta)\|}{\beta}.$$

As a consequence, for any fixed $\gamma > 0$,

$$\|e(w, 1)\| \leqslant \max\{\gamma, 1/\gamma\}\|e(w, \gamma)\|.$$

Thus, (58) holds for any fixed $\gamma > 0$, i.e.,

$$d(w, \Omega^*) \leqslant \tau\|e(w, \gamma)\|.$$

Take $\gamma = 1$ is only for the purpose of simplicity.

Based on this lemma, we have the following result.

Lemma 5 *Let $\{w^k := (x^k, y^k, \lambda^k)\}$ be the sequence generated by ADMM. Then,*

$$\|e(w^{k+1}, 1)\|^2 \leqslant \left\| \begin{matrix} y^k - y^{k+1} \\ \lambda^k - \lambda^{k+1} \end{matrix} \right\|^2_{G_2}, \tag{59}$$

where

$$G_2 := \begin{pmatrix} \beta^2 B^\top A A^\top B & 0 \\ 0 & \frac{1}{\beta^2} I \end{pmatrix}.$$

Using (58) and (59), one can easily bound the distance between the iterates and the solution set with the two consecutive iterates. Hence, the local linear rate of convergence of ADMM for solving quadratic programming is established [72, Thm. 3.2].

Theorem 4 *Let $\{w^k\}$ be the sequence generated by the ADMM scheme and denote $\{v^k = (y^k, \lambda^k)\}$. When the iterative w^k is close enough to Ω^* such that $\|e(w^k, 1)\| \leqslant \varepsilon$ is satisfied, we have*

$$\text{dist}^2_{H_\alpha}(v^{k+1}, \Omega^*) \leqslant \frac{1}{1+\xi} \cdot \text{dist}^2_H(v^k, \Omega^*), \tag{60}$$

where

$$\xi := \frac{\lambda_{\min}(H)}{\beta\tau^2\lambda_{\max}(H)\lambda_{\max}(G_2)} > 0.$$

Recently, [75] showed that the local linear rate result in [72] can be globalized under a slightly more general setting for the ADMM and a linearized ADMM. The key is that instead of the local error bound result, a new 'global' error bound for piecewise linear multifunction can be used.

Lemma 6 [76, Thm. 3.3] *Let F be a piecewise linear multifunction. For any $\kappa > 0$, there exists $\eta > 0$ such that*

$$\text{dist}\,(x, F^{-1}(0)) \leqslant \eta\ \text{dist}\,(0, F(x)), \qquad \forall \|x\| < \kappa.$$

A multi-valued mapping F is piecewise linear if its graph $\mathrm{gph}\, F := \{(x, y)|y \in F(x)\}$ is the union of finitely many polyhedral sets. One important class of piecewise linear multi-valued mappings is the subdifferential of convex piecewise linear-quadratic functions. A closed proper convex function θ is said to be piecewise linear-quadratic if $\mathrm{dom}\,\theta$ is the union of finitely many polyhedral sets and on each of these polyhedral sets, θ is either an affine or a quadratic function. In [77], it showed that a closed proper convex function θ is piecewise linear-quadratic if and only if the graph of $\partial\theta$ is piecewise polyhedral; see [78] for a complete proof and its extensions.

Most recently, [79] considered the model (32) and under a calmness condition only, it provides a global Q-linear rate convergence analysis for the ADMM. Here, the definition of calmness is taken from [80, Sec. 3.8(3H)], which says that a multifunction $F : \mathcal{X} \rightrightarrows \mathcal{Y}$ is calm at $(x^0, y^0) \in \mathrm{gph}\, F$ with modulus $\kappa_0 \geqslant 0$ if there exist a neighborhood V of x^0 and a neighborhood W of y^0 such that

$$F(x) \cap W \subseteq F(x^0) + \kappa_0 \|x - x^0\| \boldsymbol{B}_{\mathcal{Y}}, \quad \forall\, x \in V,$$

where $\boldsymbol{B}_{\mathcal{Y}}$ denotes the unit ball in the Euclidean space $\mathcal{Y}$.

Furthermore, it is well-known, e.g., [80, Thm. 3H.3], that for any $(x^0, y^0) \in \mathrm{gph}\, F$, the mapping F is calm at x^0 for y^0 if and only if F^{-1}, the inverse mapping of F, is metrically subregular at y^0 for x^0, i.e., there exist a constant $\kappa_0' \geqslant 0$, a neighborhood W of y^0, and a neighborhood V of x^0 such that

$$\mathrm{dist}(y, F(x^0)) \leqslant \kappa_0' \,\mathrm{dist}(x^0, F^{-1}(y) \cap V), \quad \forall\, y \in W. \tag{61}$$

In [65], the authors provided a number of scenarios on the linear rate convergence for the ADMM and ADMM with quadratic terms (similar to (32) but allowing H_i^k to be positive semidefinite, denoted by sPADMM) under the assumptions that either θ_1 or θ_2 is strongly convex with a Lipschitz continuous gradient, and the boundedness condition on the generated iteration sequence and others. [65] also made a detailed comparison between their linear rate convergence result and that of [68]. Other interesting results can be found in [81,82], etc.

5 Extensions and Variants

As shown in the introduction, modern applications arise a lot of problems that naturally (or after a simple reformulation) possess the structure as (1), where each component function θ_i represents one natural property of the system. The most important intrinsic character of the model is that each component function θ_i is simple enough in the sense that its proximal operator is easy to evaluate or even possesses closed-form solution, while the composition of any two (or more) of them is difficult. Hence, in the numerical algorithm, dealing them one by one, in a Gauss–Seidel manner or a Jacobi manner, is a fundamental choice. For the case that there is just one, or there are two component functions, the classical augmented Lagrangian method and the alternating direction method of multipliers, not only possess beautiful theoretical

convergence results, but also exhibit good performance in many applications. While for the model (1) with $m \geqslant 3$, the situation is totally different.

One certainly can treat the well-structured problem (1) using the ALM scheme (13), and can also regroup the variables and component functions into two blocks and then treat the resulting model using the ADMM scheme (17). However, these two schemes treat the problem on a generic purpose and ignore completely, or at least partially, the favorable separable structure in (1). Thus, this straightforward application of ALM or ADMM to (1) is not recommended. On the other hand, (13) and (17) provide us the possibility of developing customized algorithms with consideration of the specific structure of (1). Taking a close look at the minimization problem in (13) and (17), we find that the minimization tasks over the variables x_i's are coupled only by the quadratic term $\frac{\beta}{2}\left\|\sum_{i=1}^{m} A_i x_i - b\right\|^2$ or the quadratic term of partial regroup of them. Therefore, we can split the minimization subproblem in (13) and (17) into m easier and smaller subproblems by applying a Gauss–Seidel or Jacobian decomposition to this quadratic term. With the alternating direction method of multipliers (17) in mind, one may naturally, incline to extend the scheme to the general case of (1) with $m \geqslant 3$, obtaining the scheme

$$\begin{cases} x_i^{k+1} = \arg\min\limits_{x_i \in \mathcal{X}_i} \Big\{\theta_i(x_i) - (\lambda^k)^\top \big(\sum\limits_{j=1}^{i-1} A_j x_j^{k+1} + A_i x_i + \sum\limits_{j=i+1}^{m} A_j x_j^k - b\big) \\ \qquad\qquad + \frac{\beta}{2}\big\|\sum\limits_{j=1}^{i-1} A_j x_j^{k+1} + A_i x_i + \sum\limits_{j=i+1}^{m} A_j x_j^k - b\big\|^2\Big\},\ i = 1, \cdots, m; \\ \lambda^{k+1} = \lambda^k - \beta\Big(\sum\limits_{i=1}^{m} A_i x_i^{k+1} - b\Big). \end{cases} \tag{62}$$

Though the numerical efficiency of (62) was verified empirically [1,2], the theoretical convergence was only partially understood. During the last decade, research was mostly focused on the following three topics about decomposition methods for (1):

1. Does the scheme (62), the direct extension of ADMM for $m \geqslant 3$, converge for the convex case?
2. If the answer is 'no', then under what additional conditions does it converge?
3. Can the iterate generated by (62) be slightly twisted such that the convergence can be guaranteed?

5.1 A Counter Example

A simple example against the convergence of the heuristic extension of ADMM (62) was reported in [21], where they consider the special cases where $\theta_i \equiv 0$ for $i = 1, 2, 3$, and the problem (1) reduces to solving the linear homogeneous equation

$$A_1 x_1 + A_2 x_2 + A_3 x_3 = 0, \tag{63}$$

where $A_i \in \mathbb{R}^3$ and the matrix $[A_1, A_2, A_3]$ is nonsingular and $x_i \in \mathbb{R}$. With these settings, the problem (63) has a unique solution $x_1 = x_2 = x_3 = 0$. When applying the iterative scheme (62) to (63), the recursion is

$$x_1^{k+1} = \frac{1}{A_1^\top A_1}\left(-A_1^\top A_2 x^k - A_1^\top A_3 x_3^k + A_1^\top \lambda^k\right),$$

and

$$\begin{pmatrix} x_2^{k+1} \\ x_3^{k+1} \\ \lambda^{k+1} \end{pmatrix} = M \begin{pmatrix} x_2^{k} \\ x_3^{k} \\ \lambda^{k} \end{pmatrix}, \tag{64}$$

where $M = L^{-1}R$ and

$$L = \begin{pmatrix} A_2^\top A_2 & 0 & 0_{1\times 3} \\ A_3^\top A_2 & A_3^\top A_3 & 0_{1\times 3} \\ A_2 & A_3 & I_{3\times 3} \end{pmatrix},$$

and

$$R = \begin{pmatrix} 0 & -A_2^\top A_3 & A_2^\top \\ 0 & 0 & A_3^\top \\ 0_{3\times 1} & 0_{3\times 1} & I_{3\times 3} \end{pmatrix} - \frac{1}{A_1^\top A_1}\begin{pmatrix} A_2^\top A_1 \\ A_3^\top A_1 \\ A_1 \end{pmatrix}(-A_1^\top A_2, -A_1^\top A_3, A_1^\top).$$

After getting the reformulation (64), the task to get a counterexample is then to set the conceret data A_i for $i = 1, 2, 3$, such that the spectral radius of M, denoted by $\rho(M)$, is larger than 1.

The concrete example constructed in [21] is

$$A = (A_1, A_2, A_3) = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 2 \end{pmatrix},$$

for which the associated matrix M is

$$M = \frac{1}{162}\begin{pmatrix} 144 & -9 & -9 & -9 & 18 \\ 8 & 157 & -5 & 13 & -8 \\ 64 & 122 & 122 & -58 & -64 \\ 56 & -35 & -35 & 91 & -56 \\ -88 & -26 & -26 & -62 & 88 \end{pmatrix}$$

and

$$\rho(M) = 1.028 > 1.$$

A concrete example against the convergence of the heuristic extension of the alternating direction method of multipliers to $m \geqslant 3$ is thus presented.

5.2 Conditions Guaranteeing Convergence

For the general convex case, people cannot manage to prove the convergence of the direct extension scheme (62) for $m \geqslant 3$. (In fact, the example in the last subsection shows its divergence.) Prior to the work [21], people try to prove the convergence of (62) by imposing some additional conditions on the problems' data.

In [21], the authors also gave some conditions on the constraint matrices that guarantee the convergence of (62). For example, when $A_1^\top A_2 = 0$, or $A_1^\top A_3 = 0$, $A_2^\top A_3 = 0$, they proved that the sequence generated by (62) converges and the worst convergence complexity is $O(1/t)$. Since for these cases the model is essentially a two-block model and (62) is in some sense equivalent to (17), the convergence is long-established.

Most work on conditions that guarantee convergence of ADMM for (1) considers the properties of the objective functions θ_i. A strictly stronger condition than convexity is strong convexity. For the general case $m \geqslant 3$, the first sufficient convergence condition for (62) was given in [83], where it states that:

Theorem 5 *Suppose that for all $i = 1, \cdots, m$, θ_i are strongly convex with constant $\mu_i > 0$ and*

$$\beta < \min_{1 \leqslant i \leqslant m} \left\{ \frac{2\mu_i}{3(m-1)\|A_i\|^2} \right\}. \tag{65}$$

Then, the sequence $\{\boldsymbol{x}^k\}$ generated by (62) converges to a solution of (1) and $\{\lambda^k\}$ converges to an optimal multiplier.

In [21], it gave the strongly convex minimization problem with three variables:

$$\begin{aligned} &\min\ 0.05x_1^2 + 0.05x_2^2 + 0.05x_3^2 \\ &\text{s.t.} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 0, \end{aligned}$$

for which the iteration scheme of the direct extension of ADMM (62) can also be reformulated as the recursion (64). For $\beta = 1$, the spectral radius of the involved matrix is 1.008 7, indicating that the direct extension recursion does not always converge for strongly convex optimization problems, justifying the necessary of choosing a suitable penalty parameter β, as (65).

The requirement that *all* the component functions θ_i, $i = 1, \cdots, m$, are strongly convex is too restrictive, excluding the potential applications of ADMM (62). Some studies then try to weaken this restriction. In [84,85], the condition was relaxed and only $m - 1$ functions in the objective are required to be strongly convex to ensure the convergence of (62). Note that in ADMM, even for the classical case that $m = 2$, the first component does not appear in the convergence analysis, and it is usually regarded as an auxiliary part for the algorithm. From this point of view, it is not surprising that the strong convexity requirement can be relaxed from all components to $m - 1$ components. However, assuming the strong convexity for $m-1$ functions still excludes most of the applications that can be efficiently solved by the scheme (62). Thus, these

conditions are only of theoretical interests and they are usually too strict to be satisfied by the mentioned applications.

Understanding the gap between the empirical efficiency (less cputime cost than other algorithms) of (62) and the lack of theoretical conditions that can both ensure the convergence of (62) and be satisfied by some applications of the abstract model (1) motivated further results. For notation simplicity, the results are for the case $m = 3$ and can be easily extended to the general case $m > 3$. In [5,86], the authors independently showed that the convergence of (62) can be ensured when one function in the objective of (1) is strongly convex, the penalty parameter β is appropriately restricted and some assumptions on the linear operators A_i's hold—some conditions that hold for some concrete applications of (1). The assumption on the penalty parameter β in [5] is determined through checking the positive definiteness of some operators because it targets a setting more general than (62) with larger step sizes for updating λ and semi-proximal terms for regularizing the x_i-subproblems. While in [86], it also established the convergence rate for the scheme (62), including the worst-case convergence rate measured by the iteration complexity and the globally linear convergence rate in asymptotical sense under some additional assumptions. See also [26,84,85,87] for some convergence rate analysis for (62).

5.3 Correction Step for Convergence

Besides additional conditions for ensuring the convergence of the direct extension of ADMM for the general case $m \geqslant 3$, a parallel line is slightly twisting the iterate. Observing the high numerical efficiency of (62) in practice, the slighter of the twist, the more desirable in keeping the nice property of the algorithm.

This line of study is motivated by the observation that though the point generated by the direct extension of ADMM (62) is not qualified as the next iterate, it provides useful information, utilizing which judiciously can construct convergent and efficient algorithms.

Recall that one way to prove the convergence of the classical alternating direction method of multipliers is based on the *contraction* property of the iterates, or the *Fejér monotonicity* of the iterates with respect to the solution set. The *Fejér monotonicity* [88, Def. 5.1] of a sequence with respect to a closed convex set Ω says that for two consecutive point v^{k+1} and v^k, we have

$$\mathrm{dist}^2(v^{k+1}, \Omega) \leqslant \mathrm{dist}^2(v^k, \Omega) - c_0\|v^k - v^{k+1}\|^2, \tag{66}$$

where

$$\mathrm{dist}(v, \Omega) := \inf_{u\in\Omega} \|u - v\|$$

denotes the distance between a point v and Ω. Once we can prove that the ADMM-type step (62) can generate a point, denoted by $\tilde{x}^k$, and based on which a descent direction $d(\tilde{x}^k)$ of the implicit measure function $\mathrm{dist}(v, \Omega)$ can be constructed, then we can

generate the next iterate x^{k+1} via the simple principle

$$x^{k+1} = x^k - \alpha_k d(\tilde{x}^k), \tag{67}$$

where α_k is a step size, which can be a constant (setting down prior to the start of the iterative process according to some rule) or be computed along the iterate. Such a correction step can not only guarantee that the whole generated sequence converges to a solution of the problem, but also provide new freedom in the manner of generating the auxiliary point $\tilde{x}^k$.

In [89], an ADMM with Gaussian back substitution was proposed. Let $\alpha \in (0, 1)$, and let

$$M = \begin{pmatrix} \beta A_2^\top A_2 & 0 & \cdots & \cdots & 0 \\ \beta A_3^\top A_2 & \beta A_3^\top A_3 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \beta A_m^\top A_2 & \beta A_m^\top A_3 & \cdots & \beta A_m^\top A_m & 0 \\ 0 & 0 & \cdots & \cdots & \frac{1}{\beta} I_m \end{pmatrix},$$

and

$$Q = \begin{pmatrix} \beta A_2^\top A_2 & \beta A_2^\top A_3 & \cdots & \beta A_2^\top A_m & \beta A_2^\top \\ \beta A_3^\top A_2 & \beta A_3^\top A_3 & \ldots & \beta A_3^\top A_m & \beta A_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \beta A_m^\top A_2 & \beta A_m^\top A_3 & \cdots & \beta A_m^\top A_m & \beta A_m^\top \\ A_2 & A_3 & \cdots & A_m & \frac{1}{\beta} I_m \end{pmatrix}.$$

Then, define

$$H = \mathrm{diag}(\beta A_2^\top A_2, \beta A_3^\top A_3, \cdots, \beta A_m^\top A_m, \frac{1}{\beta} I_m).$$

Suppose that for all $i = 2, \cdots, m$, $A_i^\top A_i$ is nonsingular. Consequently, M is nonsingular and H is symmetric and positive definite. With the given iterate w^k, and let $\tilde{w}^k$ be generated by the ADMM scheme (62), then the new iterate w^{k+1} is generated as follows:

$$v^{k+1} = v^k - \alpha M^{-\mathbf{A}^\top} H(v^k - \tilde{v}^k), \tag{68}$$

and

$$x_1^{k+1} = \tilde{x}_1^k.$$

The construction of the matrices M and Q is to guarantee that the generated sequence converges to a solution of the problem, based on the fact that for any $v^* \in \mathcal{V}^*$, $-M^{-\mathbf{A}^\top} H(v^k - \tilde{v}^k)$ is a descent direction of the implicit function $\mathrm{dist}_G^2(v, \mathcal{V}^*)$ with

$G = MH^{-1}M^\top$ at the point $v = v^k$. Based on this fact, the step size can be chosen dynamically [90]

$$\alpha_k^* = \frac{\|v^k - \tilde{v}^k\|_H^2 + \|v^k - \tilde{v}^k\|_Q^2}{2\|v^k - \tilde{v}^k\|_H^2}$$

and the Gaussian back substitution procedure can be modified accordingly into the following form:

$$H^{-1}M^\top(v^{k+1} - v^k) = \gamma\alpha_k^*(v^k - \tilde{v}^k),$$

where $\gamma \in (0, 2)$ is a relaxation constant.

From the implementation point of view, note that the matrix M is a lower triangular matrix and H is a diagonal matrix. As a consequence, the matrix $M^{-T}H$ in (68) is an upper triangular matrix. The procedure can be performed in an inverse order by updating λ, x_m, x_{m-1}, $\cdots$, x_2, x_1 sequently, and this is where the name 'Gaussian back substitution' comes from.

The main cost in the back substitution step is in the computation of the inverse of the matrices $A_i^\top A_i$. Hence, the algorithm has its potential advantages in the case that $A_i^\top A_i$ is easy to invert, especially for the case that $A_i^\top A_i = I$, as for the robust principal component analysis problem (11).

Many efforts were devoted to simplifying the correction step. For the case that $m = 3$, [91] proposed a new correction scheme. Since the numerical experiments from various applications indicate the direct extension of ADMM is efficient, we should twist the component variables as few as possible. The prediction step and the correction step of the method in [91] are

S1. *Prediction step*: Generate the predictor $\tilde{w}^k = (\tilde{x}^k, \tilde{y}^k, \tilde{z}^k, \tilde{\lambda}^k)$ via

$$\begin{cases} \tilde{x}^k = \arg\min\{f(x) - (\lambda^k)^\top Ax + \frac{\beta}{2}\|Ax + By^k + \zeta^k - b\|^2 \mid x \in \mathcal{X}\}, & \text{(69a)} \\ \tilde{y}^k = \arg\min\{g(y) - (\lambda^k)^\top By + \frac{\beta}{2}\|A\tilde{x}^k + By + \zeta^k - b\|^2 \mid y \in \mathcal{Y}\}, & \text{(69b)} \\ \tilde{z}^k = \arg\min\{h(z) - (\lambda^k)^\top Cz + \frac{\beta}{2}\|A\tilde{x}^k + B\tilde{y}^k + Cz - b\|^2 \mid z \in \mathcal{Z}\}, & \text{(69c)} \\ \tilde{\lambda}^k = \lambda^k - \beta(A\tilde{x}^k + B\tilde{y}^k + C\tilde{z}^k - b); & \text{(69d)} \end{cases}$$

and

S2. *Correction step*: Generate the new iterate $(x^{k+1}, y^{k+1}, \zeta^{k+1}, \lambda^{k+1})$ by

$$\begin{cases} x^{k+1} = \tilde{x}^k, \\ y^{k+1} = \tilde{y}^k, \\ \zeta^{k+1} = \zeta^k - \alpha(\zeta^k - C\tilde{z}^k + By^k - B\tilde{y}^k), \\ \lambda^{k+1} = \lambda^k - \alpha(\lambda^k - \tilde{\lambda}^k), \end{cases} \tag{70}$$

respectively.

Note that the subproblems (69a)–(69c) in the prediction step are slightly different from the ADMM scheme (62) in the quadratic terms of the augmented Lagrangian function. In (69a)–(69c), an auxiliary variable ζ was introduced, which took the place of Cz. Hence, the computational cost on solving the subproblems is the same as (62). However, it provides great benefit in the correction step. In the correction step, the x part and the y part were kept, and only the z part and the dual variable were twisted. Moreover, the cost is very low, and it does not involve any additional matrix-vector product. The rule behind the introduction of the auxiliary variable ζ is simple, which is just from the observation that in the x and the y subproblem, what we essentially need is the information of By^k and Cz^k, but not y^k and z^k. Hence, in the procedure, we provide the information of By^k and Cz^k via the variable y^k and ζ^k, respectively.

For the general separable model (1), [26] proposed the following algorithm:

$$\begin{cases} x_i^{k+1} = \arg\min\limits_{x_i \in \mathcal{X}_i} \Big\{\theta_i(x_i) - (\lambda^k)^\top \big(\sum\limits_{j=1}^{i-1} A_j x_j^{k+1} + A_i x_i + \sum\limits_{j=i+1}^{m} A_j x_j^k - b\big) \\ \qquad\qquad + \frac{\beta}{2} \big\| \sum\limits_{j=1}^{i-1} A_j x_j^{k+1} + A_i x_i + \sum\limits_{j=i+1}^{m} A_j x_j^k - b \big\|^2 \Big\}, \ i = 1, \cdots, m, \\ \lambda^{k+1} = \lambda^k - \tau\beta \Big(\sum\limits_{i=1}^{m} A_i x_i^{k+1} - b \Big), \end{cases} \tag{71}$$

where $\tau > 0$ is sufficiently small. Denote the point generated by (62) by $\tilde{w}^k$, the new iterate w^{k+1} generated by (71) can be viewed as one that generated by

$$\begin{cases} \boldsymbol{x}^{k+1} = \tilde{\boldsymbol{x}}^k, \\ \lambda^{k+1} = \lambda^k - \tau(\lambda^k - \tilde{\lambda}^k), \end{cases}$$

and as a consequence, the algorithm can be viewed as a prediction-correction method that only twists the dual variable λ. Assuming an error bound condition and some others, [26] provided a linear rate convergence of (71). From the theoretical point of view, this constitutes important progress on understanding the convergence and the linear rate of convergence of the ADMM, while from the computational point of view, this is far from being satisfactory as in practical implementations one always prefers a larger step length for achieving numerical efficiency.

The correction step can not only guarantee the convergence of the ADMM, but also provide the freedom on generating the predictor. The predictor in ADMM scheme (62) is generated in a Gaussian–Seidel manner, and an alternative is to generate them simultaneously, leading to Jacobi-type splitting methods. In [92], a parallel splitting augmented Lagrangian method was proposed, whose prediction step and correction step are

S1. *Prediction step* Generate the predictor $\tilde{w}^k = (\tilde{x}^k, \tilde{\lambda}^k)$ via solving the following convex programs for $i = 1, \cdots, m$ (possibly in parallel):

$$\tilde{x}_i^k = \arg\min_{x_i \in \mathcal{X}_i} \left\{ \theta_i(x_i) - \langle \lambda^k, A_i x_i \rangle + \frac{\beta}{2} \left\| \sum_{j=1}^{i-1} A_j x_j^k + A_i x_i + \sum_{j=i+1}^{m} A_j x_j^k - b \right\|^2 \right\},$$

and

$$\tilde{\lambda}^k = \lambda^k - \beta \Big(\sum_{i=1}^{m} A_i \tilde{x}_i^k - b \Big).$$

S2. *Convex combination step to generate the new iterate*

$$w^{k+1} = w^k - \gamma \alpha_k (w^k - \tilde{w}^k),$$

where

$$\begin{aligned}
\alpha_k &:= \frac{\varphi(x^k, \tilde{x}^k)}{\psi(x^k, \tilde{x}^k)}, \\
\varphi(x^k, \tilde{x}^k) &:= \sum_{i=1}^{m} \| A_i x_i^k - A_i \tilde{x}_i^k \|^2 + \Big\| \sum_{i=1}^{m} A_i x_i^k - b \Big\|^2, \\
\psi(x^k, \tilde{x}^k) &:= (m+1) \Big(\sum_{i=1}^{m} \| A_i x_i^k - A_i \tilde{x}_i^k \|^2 \Big) + \Big\| \sum_{i=1}^{m} A_i \tilde{x}_i^k - b \Big\|^2.
\end{aligned}$$

The benefit introduced by the parallelization is great [92]; further results also indicated when there are many blocks of variables the advantage of Jacobi-type splitting methods becomes more obvious [93–96].

Computing the step size α_k just involves matrix-vector production, which is simple. Nevertheless, in some applications the prediction step is also low cost. In this case, we need to further simplify the correction step, e.g., adopting a constant step. Essentially, it was proved in [92, Lem. 3.6] that the step size α_k computed is uniformly bounded below from zero; i.e., there is $\alpha_{\min} := 1/(3m+1)$, such that for all $k \geqslant 0$, $\alpha_k \geqslant \alpha_{\min}$. Consequently, a constant step size $\alpha \in (0, \alpha_{\min})$ is possible.

Extending the customized Douglas–Rachford splitting method (31) to the multi-block case, [97] proposed the following distributed splitting method:

Find $(x^{k+1}, \lambda^{k+1}) \in \Omega$ *such that*

$$\begin{cases}
\bar{\lambda}^k = \lambda^k - \beta \left(\sum_{i=1}^m A_i x_i^k - b \right), \\
x_i^{k+1} = \arg\min\limits_{x_i \in \mathbb{R}^{n_i}} \left\{ \theta_i(x_i) + \frac{1}{2\beta} \left\| x_i - \omega_i^k \right\|^2 \right\}, \quad (i = 1, 2, \cdots, m), \\
\lambda^{k+1} = \lambda^k - \gamma \alpha_k \left(\widetilde{\boldsymbol{e}}(x^k, \beta) - \beta \sum_{i=1}^m A_i \boldsymbol{e}_i(x_i^k, \bar{\lambda}^k) \right),
\end{cases}$$

where $\omega_i^k = x_i^k + \beta\xi_i^k - \gamma\alpha_k \boldsymbol{e}_i(x_i^k, \bar{\lambda}^k)$ *with* $\xi_i^k \in \partial\theta_i(x_i^k)$ *and*

$$\begin{cases} \alpha_k := \frac{\varphi(x^k,\bar{\lambda}^k,\beta)}{\psi(x^k,\bar{\lambda}^k,\beta)}, \\ \varphi(x^k,\bar{\lambda}^k,\beta) := \frac{1}{2}\left\|E(x^k,\bar{\lambda}^k,\beta)\right\|^2 + \frac{1}{2}\left\|\widetilde{\boldsymbol{e}}(x^k,\beta)\right\|^2 - \sum_{i=1}^m \beta\widetilde{\boldsymbol{e}}(x^k,\beta)^\top A_i \boldsymbol{e}_i(x_i^k,\bar{\lambda}^k), \\ \psi(x^k,\bar{\lambda}^k,\beta) := \sum_{i=1}^m \left\|\boldsymbol{e}_i(x_i^k,\bar{\lambda}^k)\right\|^2 + \left\|\widetilde{\boldsymbol{e}}(x^k,\beta) - \beta\sum_{i=1}^m A_i\boldsymbol{e}_i(x_i^k,\bar{\lambda}^k)\right\|^2. \end{cases}$$

Here

$$\begin{cases} \boldsymbol{e}_i(x_i,\lambda) := x_i - P_{\mathcal{X}_i}\left\{x_i - \beta(\xi_i - A_i^\top\lambda)\right\}, \quad (i = 1,2,\cdots,m), \\ E(\boldsymbol{x},\lambda,\beta) := (\,\boldsymbol{e}_1(x_1,\lambda),\ \cdots,\boldsymbol{e}_m(x_m,\lambda),\ \widetilde{\boldsymbol{e}}(x,\beta)\,), \\ \widetilde{\boldsymbol{e}}(\boldsymbol{x},\beta) := \beta\left(\sum_{i=1}^m A_i x_i - b\right). \end{cases}$$

Combining the idea of Gauss–Seidel splitting and Jacobian splitting, several partial parallel splitting methods were proposed. For simplicity, let $m = 3$. [98] proposed to solve the first subproblem and then to solve the second and the third in a parallel manner; [99] presented a partial parallel splitting method and adopted a relaxation step with low computational cost, in which the variables were updated in the order $x_1 \to \lambda \to (x_2, x_3)$, rather than $x_1 \to (x_2, x_3) \to \lambda$; [91] proposed to just corrected the last primal variable x_3 and the multiplier to keep the efficiency of the splitting method, and such an idea was exploited in [100] for image processing.

6 Solving Nonconvex Problems

Consider again the two-block model (1) with $m = 2$. For the case that both components of the objective function are convex, the convergence of ADMM is well-understood, both for the global convergence, the sublinear convergence measured by iteration complexity, and linear convergence. When one or both of the component objective functions are nonconvex, the convergence analysis for ADMM is much more challenging, and it is only partially understood. On the other hand, as stated in the introduction, the applications usually involve models (1) where at least one component function is nonconvex. This phenomenon inspires the recent interest in studying convergence of ADMM for nonconvex model (1).

6.1 Convergence under KL Properties

A very important technique to prove the convergence for nonconvex optimization problems relies on the assumption that the objective functions satisfies the Kurdyka–Lojasiewicz inequality. The importance of Kurdyka–Lojasiewicz (KL) inequality is due to the fact that many functions satisfy this inequality. In particular, when the function belongs to some functional classes, e.g., semi-algebraic (such as $\|\cdot\|_p^p$, $p \in [0,1]$ is a rational number), real subanalytic, log-exp (see also [101–104] and references therein), it is often elementary to check that such an inequality holds. The inequality

was established in the pioneering and fundamental works [22,23], and it was recently extended to nonsmooth functions in [103,104].

Definition 4 ([101] *Kurdyka–Lojasiewicz inequality*) Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a proper lower semicontinuous function. For $-\infty < \eta_1 < \eta_2 \leqslant +\infty$, set

$$[\eta_1 < f < \eta_2] = \{x \in \mathbb{R}^n : \eta_1 < f(x) < \eta_2\}.$$

We say that function f has the KL property at $x^* \in \operatorname{dom} \partial f$ if there exist $\eta \in (0, +\infty]$, a neighborhood U of x^*, and a continuous concave function $\varphi : [0, \eta) \to \mathbb{R}_+$, such that

(i) $\varphi(0) = 0$;
(ii) φ is C^1 on $(0, \eta)$ and continuous at 0;
(iii) $\varphi'(s) > 0, \forall s \in (0, \eta)$;
(iv) for all x in $U \cap [f(x^*) < f < f(x^*) + \eta]$, the Kurdyka–Lojasiewicz inequality holds:

$$\varphi'(f(x) - f(x^*)) d(0, \partial f(x)) \geqslant 1.$$

Here and in the following, $\partial f(x)$ denotes the limiting-subdifferential, or simply the subdifferential, of a proper lower semicontinuous function. Let us list some definitions of subdifferential calculus [78,105].

Definition 5 Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function.

(i). The Fréchet subdifferential, or regular subdifferential, of f at $x \in \operatorname{dom} f$, written as $\hat{\partial} f(x)$, is the set of vectors $x^* \in \mathbb{R}^n$ that satisfy

$$\lim_{y \neq x} \inf_{y \to x} \frac{f(y) - f(x) - \langle x^*, y - x \rangle}{\|y - x\|} \geqslant 0.$$

When $x \notin \operatorname{dom} f$, we set $\hat{\partial} f(x) = \varnothing$.

(ii). The limiting subdifferential, or simply the subdifferential, of f at $x \in \operatorname{dom} f$, written as $\partial f(x)$, is defined as follows:

$$\partial f(x) = \{x^* \in \mathbb{R}^n : \exists x_n \to x,\ f(x_n) \to f(x),\ x_n^* \in \hat{\partial} f(x_n), \text{ with } x_n^* \to x^*\}.$$

Definition 6 ([102] *Kurdyka–Lojasiewicz function*) If f satisfies the KL property at each point of $\operatorname{dom} \partial f$, then f is called a KL function.

Using Kurdyka–Lojasiewicz property, convergence of some optimization methods was established. Some latest references are gradient-related methods [106], proximal point algorithm [107–109], nonsmooth subgradient descent method [110], etc., (see also [111,112]). In the field of ADMM studies, relying on the assumption that the Kurdyka–Lojasiewicz inequality holds for certain functions, [113,114] proved convergence of variant ADMM for (15) with some special structures. [113] considered a special case of problem (15) with $A_2 = -I$ and $b = 0$, i.e., the problem

$$\min \left\{ \theta_1(x_1) + \theta_2(x_2) \;\middle|\; A_1 x_1 - x_2 = 0 \right\}, \tag{72}$$

and proposed a variant of ADMM by regularizing the second subproblem. Their recursion is

$$\begin{cases} x_2^{k+1} \in \arg\min\{\mathcal{L}_\beta(x_1^k, x_2, \lambda^k)\}, \\ x_1^{k+1} \in \arg\min\{\mathcal{L}_\beta(x_1, x_2^{k+1}, \lambda^k) + \Delta_\phi(x_1, x_1^k)\}, \\ \lambda^{k+1} = \lambda^k - \beta(A_1 x_1^{k+1} - x_2^{k+1}), \end{cases} \tag{73}$$

where the function ϕ is strictly convex and Δ_ϕ is the Bregman distance with respect to ϕ,

$$\Delta_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y\rangle.$$

When $\phi = 0$, the algorithm (73) reduces to the classic ADMM (17). Assuming that the penalty parameter is chosen sufficiently large; and one of the component function, θ_1, is twice continuously differentiable with uniformly bounded Hessian, Li and Pong proved that if the sequence generated by the algorithm (73) has a cluster point, then it gives a stationary point of the nonconvex problem (72), i.e., a point (x_1, x_2, λ) that satisfies

$$\begin{cases} 0 \in \partial\theta_1(x_1) - A_1^\top \lambda, \\ 0 \in \partial\theta_2(x_2) + \lambda, \\ 0 = A_1 x_1 - x_2. \end{cases}$$

Reference [114] proposed to regularize both subproblems, resulting the following variant ADMM

$$\begin{cases} x_2^{k+1} \in \arg\min\{\mathcal{L}_\beta(x_1^k, x_2, \lambda^k) + \Delta_\psi(x_2, x_2^k)\}, \\ x_1^{k+1} \in \arg\min\{\mathcal{L}_\beta(x_1, x_2^{k+1}, \lambda^k) + \Delta_\phi(x_1, x_1^k)\}, \\ \lambda^{k+1} = \lambda^k - \beta(A_1 x_1^{k+1} - A_2 x_2^{k+1}). \end{cases} \tag{74}$$

Under the assumptions that A_1 is full row rank, either $\mathcal{L}_\beta(x_1, x_2, \lambda)$ with respect to x_1 or ϕ is μ_1 strongly convex with $\mu_1 > 0$, and some other mild conditions, they proved the sequence generated by (74) converges to a stationary point.

Reference [115] considered solving the model (72) with the classical ADMM (17), where they made the following assumption:

Assumption 1 Consider the classical ADMM (17) solving the model (72). Suppose that $\theta_1 : \mathcal{R}^n \to \mathbb{R} \cup \{\infty\}$ is a proper lower semicontinuous function, $\theta_2 : \mathbb{R}^m \to \mathbb{R}$ is a continuously differentiable function with $\nabla\theta_2$ being Lipschitz continuous with modulus $L > 0$, and suppose that the augmented Lagrangian function associated to the model (72) is a KL function. Moreover, suppose following two items for the parameters hold:

- $\beta > 2L$, then $\delta = \frac{\beta - L}{2} - \frac{L^2}{\beta} > 0$,
- $A_1^\top A_1 \succeq \mu I$ for some $\mu > 0$.

Note that the conditions assumed in Assumption 1 are much weaker than those in [113,114].

Measuring the iterates $\{w^k = (x_1^k, x_2^k, \lambda^k)\}$ by the augmented Lagrangian function $\mathcal{L}_\beta$ associated with the model (72), it was first proved that

$$\mathcal{L}_\beta(w^{k+1}) \leqslant \mathcal{L}_\beta(w^k) - \delta\|x_2^{k+1} - x_2^k\|^2.$$

Note that if we can establish that

$$\mathcal{L}_\beta(w^{k+1}) \leqslant \mathcal{L}_\beta(w^k) - \delta\|w^{k+1} - w^k\|^2, \tag{75}$$

then we get the 'sufficient' decrease in the iterate $\{w^k\}$ measured by $\mathcal{L}_\beta$. Thanks to the Lipschitz continuity of $\nabla\theta_2$ and the optimality condition for the x_2-subproblem, we can reach the aim with judicious choice of the penalty parameter ensuring that $\frac{\beta-L}{2} - \frac{L^2}{\beta} > 0$. Then with the aid of the KL property, the convergence and the rate of convergence of ADMM can be established, and we summarize them in the following two theorems.

Theorem 6 *Suppose that Assumption 1 holds. Let $\{w^k = (x^k, y^k, \lambda^k)\}$ be the sequence generated by the ADMM procedure (17) which is assumed to be bounded. Suppose that θ_1 and θ_2 are semi-algebraic functions, then $\{w^k\}$ has finite length, that is*

$$\sum_{k=0}^{+\infty} \|w^{k+1} - w^k\| < +\infty, \tag{76}$$

and as a consequence, $\{w^k\}$ converges to a critical point of $\mathcal{L}_\beta(\cdot)$.

Theorem 7 (Convergence rate) *Suppose that Assumption 1 holds. Let $\{w^k = (x^k, y^k, \lambda^k)\}$ be the sequence generated by the ADMM procedure (17) and converges to $\{w^* = (x^*, y^*, \lambda^*)\}$. Assume that $\mathcal{L}_\beta(\cdot)$ has the KL property at (x^*, y^*, λ^*) with $\varphi(s) = cs^{1-\theta}$, $\theta \in [0, 1)$, $c > 0$. Then, the following estimations hold*:

(i) *If $\theta = 0$, the sequence $\{w^k = (x^k, y^k, \lambda^k)\}$ converges in a finite number of steps.*

(ii) *If $\theta \in (0, \frac{1}{2}]$, there exist $c > 0$ and $\tau \in [0, 1)$, such that*

$$\|(x^k, y^k, \lambda^k) - (x^*, y^*, \lambda^*)\| \leqslant c\tau^k.$$

(iii) *If $\theta \in (\frac{1}{2}, 1)$, there exists $c > 0$, such that*

$$\|(x^k, y^k, \lambda^k) - (x^*, y^*, \lambda^*)\| \leqslant ck^{\frac{\theta-1}{2\theta-1}}.$$

The key role played by the KL property [113–115] is as follows. After establishing the sufficient descent of the augmented Lagrange function (75), one can prove that if $w^{k+1} \neq w^k$, then for any $\eta > 0$, and $\varepsilon > 0$, there exist $\bar{k} > 0$, such that for any $k > \bar{k}$,

$$d(w^k, S(w^0)) < \varepsilon, \quad \mathcal{L}_\beta(w^*) < \mathcal{L}_\beta(w^k) < \mathcal{L}_\beta(w^*) + \eta,$$

where $w^* \in S(w^0)$ and $S(w^0)$ is the set of cluster points of the sequence $\{w^k\}$ generated by the ADMM (17) (respectively, (73) or (74)). Applying the uniformized KL property [116], we conclude that for any $k > \bar{k}$

$$\varphi'(\mathcal{L}_\beta(w^k) - \mathcal{L}_\beta(w^*))d(0, \partial\mathcal{L}_\beta(w^k)) \geqslant 1.$$

Moreover, the concavity of φ means

$$\begin{aligned}&\varphi\ (\mathcal{L}_\beta(w^k) - \mathcal{L}_\beta(w^*)) - \varphi(\mathcal{L}_\beta(w^{k+1}) - \mathcal{L}_\beta(w^*)) \geqslant\\ &\varphi'\ (\mathcal{L}_\beta(w^k) - \mathcal{L}_\beta(w^*))(\mathcal{L}_\beta(w^k) - \mathcal{L}_\beta(w^{k+1})).\end{aligned}$$

A simple further argument establishes (76) and the reader is referred to [113–115] for more details.

Convergence of ADMM for nonconvex optimization problems without using KL property, but with similar property was established recently, e.g.,

- Wang et al. [117] proposed the concept of restricted prox-regularity (see Definition 2 in [117]): For a lower semicontinuous function f, let $M \in \mathbb{R}_+$, $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, and let

  $$S_M := \{x \in \mathrm{dom}(f) : \ \|d\| > M \text{ for all } d \in \partial f(x)\}.$$

  Then, f is called restricted prox-regular if, for any $M > 0$ and bounded set $T \subseteq \mathrm{dom}(f)$, there exists $\gamma > 0$ such that

  $$\begin{aligned}&f(x) + \frac{\gamma}{2}\|x - y\|^2 \geqslant\\ &\quad f(x) + \langle d, y - x\rangle,\ \forall y \in T\backslash S_M,\ y \in T,\ d \in \partial f(x),\ \|d\| \leqslant M.\end{aligned}$$

  Under the conditions that the objective function is restricted prox-regular and it is also coercive on the feasible set, it proved that the generated sequence has a cluster point which is solution of (72).
- Jia et al. [118] proved that if certain error bound condition like (58) holds, the sequence generated by the ADMM for solving the nonconvex optimization problem (72) converges locally with a linear rate.
- Jiang et al. [119] considered a proximal ADMM for solving linearly constrained nonconvex optimization models. Under the assumption that all the block variables except for the last block were updated with the proximal ADMM, while the last block was updated either with a gradient step or a majorization–minimization step, they showed the iterative sequence converges to a stationary point.
- Zhang et al. [120] proposed a proximal ADMM for solving the linearly constrained nonconvex differentiable optimization problems, in which the authors introduced a "smoothed" sequence of primal iterates, and added to the augmented Lagrangian function an extra quadratic term. Under some additional conditions such as strict complementarity, Lipschitz continuity of the gradient of the objective function, they proved the iterative sequence converges to a stationary point and established its rate of convergence.

Sufficient conditions guaranteeing the boundedness of the generated iterate sequence $\{w^k\}$ are also presented; see Lemma 3.5 in [115].

In [121], the authors considered the general separable optimization problem with linear equality constraint (1) where one or more components are nonconvex, i.e.,

$$\begin{aligned} \min \ & \sum_{i=1}^{m} \theta_i(x_i) \\ \text{s.t.} \ & A_1x_1 + A_2x_2 + \cdots + A_{m-1}x_{m-1} + x_m = b, \end{aligned}$$

where $\theta_1 : \mathbb{R}^{n_1} \to \mathbb{R} \cup \{\infty\}$ is a proper lower semicontinuous function, $\theta_i : \mathbb{R}^{n_i} \to \mathbb{R}$, $i = 2, \cdots, m-1$ and $\theta_m : \mathbb{R}^s \to \mathbb{R}$ are continuous differentiable functions with $\nabla\theta_i$ being Lipschitz continuous with modulus $L_i > 0$, $A_i \in \mathbb{R}^{s\times n_i}$, $i = 1, \cdots, m-1$ is a given matrix and $b \in \mathbb{R}^s$ is a vector. Set $L := \max\limits_{2\leqslant i\leqslant m} L_i$ and assume the following two items hold:

(i) $\beta > \max\{2L, \frac{L}{\mu}\}$;
(ii) $A_1^\top A_1 \succeq \mu I$, $A_2^\top A_2 \succeq \mu I$ for some $\mu > 0$.

Then, similar results as those for the two-block can be established.

Guo et al. [122] then gave another extension of the result in [115], where they considered the nonconvex optimization problem

$$\begin{aligned} \min \ & \theta_1(x_1) + \theta_2(x_2) + H(x_1, x_2) \\ \text{s.t.} \ & A_1x_1 + x_2 = b, \end{aligned} \tag{77}$$

where $\theta_1 : \mathbb{R}^{n_1} \to \mathbb{R}\cup\{\infty\}$ is a proper lower semicontinuous function, $\theta_2 : \mathbb{R}^{n_2} \to \mathbb{R}$ is a continuously differentiable function whose gradient $\nabla\theta_2$ is Lipschitz continuous with constant $L_2 > 0$, $H : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}$ is a smooth function, $A_1 \in \mathbb{R}^{n_2\times n_1}$ is a given matrix, and $b \in \mathbb{R}^{n_2}$ is a vector.

Let $\beta > 0$ be a given parameter and the augmented Lagrangian function for problem (77) is

$$\mathcal{L}_\beta(x_1, x_2, \lambda) := \theta_1(x_1) + \theta_2(x_2) + H(x_1, x_2) - \lambda^\top(A_1x_1 + x_2 - b) + \frac{\beta}{2}\|A_1x_1 + x_2 - b\|^2,$$

where λ is the Lagrangian multiplier associated with the linear constraints. Based on alternately optimizing the augmented Lagrangian function $\mathcal{L}_\beta(\cdot)$ for one variable but with the others fixed, the alternating direction method of multipliers generates the iterative sequence with the following recursion:

$$\begin{cases} x_1^{k+1} \in \arg\min\limits_{x_1}\{\mathcal{L}_\beta(x_1, x_2^k, \lambda^k)\}, \\ x_2^{k+1} \in \arg\min\limits_{x_2}\{\mathcal{L}_\beta(x_1^{k+1}, x_2, \lambda^k)\}, \\ \lambda^{k+1} = \lambda^k - \beta(A_1x_1^{k+1} + x_2^{k+1} - b). \end{cases} \tag{78}$$

Assume that $\theta_1 : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a weakly convex function with constant $\omega > 0$ and assume the following items hold:

(i) All the component functions are bound below, i.e.,

$$\inf_{(x,y)\in\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}} H(x_1, x_2) > -\infty, \quad \inf_{x_1\in\mathbb{R}^{n_1}} \theta_1(x_1) > -\infty, \quad \inf_{x_2\in\mathbb{R}^{n_2}} \theta_2(x_2) > -\infty;$$

(ii) For any fixed x_1, the partial gradient $\nabla_{x_2} H(x_1, x_2)$ is globally Lipschitz with constant $L_2(x_1)$, that is

$$\|\nabla_{x_2} H(x_1, \bar{x}_2) - \nabla_{x_2} H(x_1, \hat{x}_2)\| \leqslant L_2(x_1)\|\bar{x}_2 - \hat{x}_2\|, \ \ \forall \bar{x}_2, \hat{x}_2 \in \mathbb{R}^{n_2};$$

For any fixed x_2, the partial gradient $\nabla_{x_1} H(x_1, x_2)$ is globally Lipschitz with constant $L_3(x_2)$, that is

$$\|\nabla_{x_1} H(\bar{x}_1, x_2) - \nabla_{x_1} H(\hat{x}_1, x_2)\| \leqslant L_3(x_2)\|\bar{x}_1 - \hat{x}_1\|, \ \ \forall \bar{x}_1, \hat{x}_1 \in \mathbb{R}^{n_1};$$

(iii) There exist $L_2, L_3 > 0$ such that

$$\sup\{L_2(x_1^k) : k \geqslant 0\} \leqslant L_2, \quad \sup\{L_3(x_2^k) : k \geqslant 0\} \leqslant L_3;$$

(iv) ∇H is Lipschitz continuous on bounded subsets of $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$. In other words, for each bounded subset $B_1 \times B_2 \subseteq \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, there exists $M > 0$ such that for all $(x_i, y_i) \in B_1 \times B_2, \ i = 1, 2$:

$$\begin{aligned} &\| \ (\nabla_x H(x_1, y_1) - \nabla_x H(x_2, y_2), \nabla_y H(x_1, y_1) - \nabla_y H(x_2, y_2))\| \\ &\leqslant M\|(x_1 - x_2, y_1 - y_2)\|; \end{aligned}$$

(v) $A^\top A \succeq \mu I$ for some $\mu > 0$;

(vi) The penalty parameter β satisfies $\beta > \max\{\beta_1, \beta_2\}$, where

$$\beta_1 := \frac{(L_3 + \omega) + \sqrt{(L_3 + \omega)^2 + 16\mu M^2}}{2\mu}$$

and

$$\beta_2 := \frac{(L_1 + L_2) + \sqrt{(L_1 + L_2)^2 + 16(L_1^2 + M^2)}}{2}.$$

The convergence result is as follows:

Theorem 8 *Let $\{w^k\}$ be the sequence generated by the ADMM (78) which is assumed to be bounded. Suppose that $\mathcal{L}_\beta(\cdot)$ is a KL function, then $\{w^k\}$ has finite length, that is*

$$\sum_{k=0}^{+\infty} \|w^{k+1} - w^k\| < \infty,$$

and as a consequence, we have $\{w^k\}$ *converges to a critical point of* $\mathcal{L}_\beta(\cdot)$.

***Remark* 2** Note that most algorithms mentioned above cannot solve the general nonconvex, constrained optimization problem (1), even for $m = 2$ (15). They rely on the assumption that their exists only one nonsmooth component, and the others are smooth with Lipschitz gradients. Another assumption is that A_1 is of full row rank or $\text{Image}(A_2) \subseteq \text{Image}(A_1)$, which is not satisfied in many practical problems. Therefore, the convergence of ADMM for many nonconvex problems is still unknown.

6.2 Convergence with Special Structures

Studies on special classes of nonconvex optimization problems were performed recently, e.g., [123] proved the convergence of ADMM for consensus problems. In this subsection, we focused on the problems having the following structure:

$$\min_{x\in\mathbb{R}^n} f(x) + g(x). \tag{79}$$

Recall that ADMM is the application of the Douglas–Rachford splitting algorithm to the dual of (15), which has the form of (79), and the convergence has been well-studied [30] when both the two component functions are convex. When there is a nonconvex component function, the study is in its infancy. Recently, the convergence of DRSM for solving (79) was established [124–126] with the assumption that $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ are proper and lower semicontinuous, f is strongly convex with constant $\mu > 0$, and g is semiconvex with constant $\omega > 0$.

The model (79) arises frequently in the big data and artificial intelligence fields, where $f(x)$ represents a data-fidelity term and $g(x)$ is a sparsity-driven penalty term. Usually, weak convexity can often reduce bias in nonzero estimates (which is a serious problem for various convex penalty terms), see, e.g., [127–129] for some applications in sparse signal recovery applications. In particular, it was proved that several popular penalty functions are weakly convex [130].

1. The smoothly clipped absolute deviation (SCAD) penalty [131]: Let

$$g_\lambda(\theta) := \begin{cases} \lambda|\theta|, & |\theta| \leqslant \lambda, \\ \frac{-\theta^2+2a\lambda|\theta|-\lambda^2}{2(a-1)}, & \lambda < |\theta| \leqslant a\lambda, \\ \frac{(a+1)\lambda^2}{2}, & |\theta| > a\lambda, \end{cases} \tag{80}$$

where $a > 2$ and $\lambda > 0$. Then SCAD is given by $\sum_{i=1}^n g_\lambda(|x_i|)$ with

$$g_\lambda(|x_i|) := \begin{cases} \lambda|x_i|, & |x_i| \leqslant \lambda, \\ \frac{-x_i^2+2a\lambda|x_i|-\lambda^2}{2(a-1)}, & \lambda < |x_i| \leqslant a\lambda, \\ \frac{(a+1)\lambda^2}{2}, & |x_i| > a\lambda. \end{cases}$$

It was proved that g_λ defined by (80) is semiconvex with modulus $1/(a-1)$ [130, Thm. 3.1].

2. The minimax concave penalty (MCP) proposed in [132]: Let

$$P_\gamma(|\theta|;\lambda) := \begin{cases} \lambda|\theta| - \frac{1}{2\gamma}\theta^2, & |\theta| < \lambda\gamma, \\ \frac{\lambda^2\gamma}{2}, & |\theta| \geqslant \lambda\gamma, \end{cases} \tag{81}$$

with $\gamma > 0$ and $\lambda > 0$. For $x \in \mathbb{R}^n$, the minimax concave penalty (MCP) is given by $\sum_{i=1}^n P_\gamma(|x_i|;\lambda)$ with

$$P_\gamma(|x_i|;\lambda) := \begin{cases} \lambda|x_i| - \frac{1}{2\gamma}|x_i|^2, & |x_i| < \lambda\gamma, \\ \frac{\lambda^2\gamma}{2}, & |x_i| \geqslant \lambda\gamma. \end{cases}$$

It was proved that P_γ defined by (81) is semiconvex with modulus $1/\gamma$ [130, Thm. 3.2].

3. Smoothed surrogate of the ℓ_p-norm: For any $0 < p < 1$ and $\varepsilon > 0$, the function $\sum_{i=1}^n (|x_i| + \varepsilon)^p$ is semiconvex with constant $p(1-p)\varepsilon^{p-2}$ [130, Thm. 3.3].

The classical convergence results [32,68] were built on the Krasnosel'skiĭ–Mann theorem [88, Thm. 5.14], which states that if D is a nonempty closed convex subset of $\mathbb{R}^n$ and $T : D \to D$ is a nonexpansive operator such that $\mathrm{Fix}(T) \neq \varnothing$, then starting with $z_0 \in D$, the iterative scheme $z_{k+1} := (1-\alpha)z_k + \alpha T(z_k)$ for $0 < \alpha < 1$ generates $\{z_k\}$ converging to a point in $\mathrm{Fix}(T)$. Recall that the DRSM (24) for solving (79) takes the form ((24) is a special case with $\alpha = 1/2$)

$$z_{k+1} = ((1-\alpha)I + \alpha R_{\mu f} R_{\mu g})(z_k), \tag{82}$$

where $R_{\mu f} := 2\mathrm{prox}_{\mu f} - I$ and $R_{\mu g} := 2\mathrm{prox}_{\mu g} - I$ are the reflection operators of f and g; and $\mathrm{prox}_{\mu f}$ and $\mathrm{prox}_{\mu g}$ are their proximal operators, respectively, where for a proper lower semicontinuous function f and a scalar $\mu > 0$,

$$\mathrm{prox}_{\mu f}(x) \in \arg\min_y \left\{ f(y) + \frac{1}{2\mu}\|x-y\|^2 \right\}.$$

When f is convex, then it follows from [88, Prop. 12.27] that $\mathrm{prox}_{\lambda f}$ is a single-valued mapping which is also firmly nonexpansive[3]. As a consequence, the reflection operator $R_{\mu f} := 2\mathrm{prox}_{\mu f} - I$ is nonexpansive [88, Prop. 4.2]. Then, according to [88, Prop. 4.21], if both f and g are convex, $R_{\mu f} R_{\mu g}$ is nonexpansive. The convergence of (82) is then a direct consequence of the Krasnosel'skiĭ–Mann theorem.

[3] Let D be a nonempty subset of $\mathbb{R}^n$ and let $T : D \to \mathbb{R}^n$. Then T is firmly nonexpansive if

$$\|Tx - Ty\|^2 + \|(I-T)x - (I-T)y\|^2 \leqslant \|x-y\|^2, \ \forall x, y \in D.$$

If either f or g is nonconvex, the situation is totally changed since the above argument is not correct due to the fact that the reflection operator of the nonconvex mapping may be expansive. To ensure the nonexpansiveness of $R_{\mu f}R_{\mu g}$ such that the Krasnosel'skiĭ–Mann theorem is still compliant, one needs further conditions on the data.

In [124], it established the convergence of DRSM under the further assumption that f is strongly convex with constant $\nu > 0$, g is semiconvex with constant $\omega > 0$, and $\nu = \omega$. Moreover, f is assumed to be second-order differentiable with ∇f being Lipschitz continuous with constant $\sigma > 0$. One may argue that since the whole model is convex, we can set $\tilde{f}(x) := f(x) - \frac{\nu}{2}\|x\|^2$ and $\tilde{g}(x) := g(x) + \frac{\nu}{2}\|x\|^2$ and convert (79) to

$$\min \quad \tilde{f}(x) + \tilde{g}(x) \tag{83}$$

and the DRSM (82) is transformed into

$$z_{k+1} = ((1-\alpha)I + \alpha R_{\mu\tilde{f}}R_{\mu\tilde{g}})(z_k) \tag{84}$$

and the convergence of (84) is well-understood. Unfortunately, although theoretically the model (83) is equivalent to (79), the numerical behavior of DRSM applying to them is different; results tested on a simple example reported in [124] indicated that in some cases, one prefers the original one (82) to the reformulation (84).

In [133], the DRSM is applied to a feasibility reformulation of the minimization of the sum of a proper lower semicontinuous function (not necessarily convex) and a weakly convex function whose gradient is required to be Lipschitz continuous, and both the functions are required to satisfy the Kurdyka–Lojasiewicz inequality. It should be mentioned that the strong convexity implies the Kurdyka–Lojasiewicz inequality, see [101,116], while there is no differentiability requirement on g in [124].

The differentiability on f required in [124] was removed in [125]; alternately, it assumed that f is strongly convex with constant ν and g is semiconvex with constant ω, and $\nu > \omega$. The scalar relationship $\nu > \omega$ means the whole objective function in (79) is strongly convex and thus it satisfies the Kurdyka–Lojasiewicz inequality. But the assumption on g is weaker than [133] because there is no differentiability assumption on g. The justification of considering this situation is illustrated via the joint image denoising and sharpening problem in [134], where the optimization model is

$$\min_{x\in\mathbb{R}^n} \quad \frac{c}{2}\|x-b\|^2 + \iota_{[0,1]}(x) + \|\nabla x\|_{2,1} - \frac{\nu}{2}\|\nabla x\|_{2,2}^2. \tag{85}$$

In (85), $x \in \mathbb{R}^n$ is the vector representation of a digital image to be recovered, $b \in \mathbb{R}^n$ is an observed image; $\nabla := (\nabla_1, \nabla_2) : \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^n$ denotes a discrete gradient operator where $\nabla_1 : \mathbb{R}^n \to \mathbb{R}^n$ and $\nabla_2 : \mathbb{R}^n \to \mathbb{R}^n$ are the standard finite difference with periodic boundary conditions in the horizontal and vertical directions, respectively; $\|\nabla x\|_{2,1}$ is the total variational regularization term [7] to preserve sharp edges; $-\frac{\nu}{2}\|\nabla x\|_{2,2}^2$ is a sharpening/edge enhancement term aiming at removing a blur if the blur is assumed to follow a diffusion process; and

$$\iota_{[0,1]}(x) := \begin{cases} 0, & 0 \leqslant x \leqslant 1; \\ \infty, & \text{else.} \end{cases}$$

The definitions of $\|\cdot\|_{2,1} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ and $\|\cdot\|_{2,2} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ are given respectively by

$$\|y\|_{2,1} := \sum_{i,j=1}^{n} \sqrt{|y^1_{i,j}|^2 + |y^2_{i,j}|^2}, \quad \forall y = (y^1, y^2) \in \mathbb{R}^n \times \mathbb{R}^n;$$

and

$$\|y\|_{2,2} := \sqrt{\sum_{i,j=1}^{n} |y^1_{i,j}|^2 + |y^2_{i,j}|^2}, \quad \forall y = (y^1, y^2) \in \mathbb{R}^n \times \mathbb{R}^n.$$

The main results of [125] are based on another important result, i.e., Fejér monotone theorem [88, Thm. 5.5], which states that in $\mathbb{R}^n$, if a sequence $\{x^k\}$ is Fejér monotone with respect to a nonempty subset D and every cluster point of $\{x^k\}$ belongs to D, then $\{x^k\}$ converges to a point in D. To prove the Fejér monotone of the sequence generated by (82), it first established in [125] that for a proper lower semicontinuous strongly convex function f with constant $\nu > 0$ and for any $x, y \in \mathbb{R}^n$ and $\mu > 0$, we have

$$\|R_{\mu f}(x) - R_{\mu f}(y)\|^2 \leqslant \|x - y\|^2 - 4\nu\mu\|\text{prox}_{\mu f}(x) - \text{prox}_{\mu f}(y)\|^2. \tag{86}$$

On the same time, for a proper lower semicontinuous function $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ which is weakly convex with constant $\omega > 0$, we have that for any $x, y \in \mathbb{R}^n$ and $0 < \mu < \frac{1}{\omega}$,

$$\|R_{\mu g}(x) - R_{\mu g}(y)\|^2 \leqslant \|x - y\|^2 + 4\omega\mu\|\text{prox}_{\mu g}(x) - \text{prox}_{\mu g}(y)\|^2. \tag{87}$$

Based on (86) and (87), though we cannot obtain the nonexpansiveness of $R_{\mu f}R_{\mu g}$ or $R_{\mu g}R_{\mu f}$, we can prove that if $\nu > \omega$ and $0 < \mu < \frac{(1-\alpha)(\nu-\omega)}{\nu\omega}$, then there exists $\eta_1 > 0$ and $\eta_2 > 0$ such that the sequence $\{z_k\}$ generated by the DRSM (82) satisfies

$$\|z_{k+1} - z^*\|^2 \leqslant \|z_k - z^*\|^2 - \eta_1\|z_{k+1} - z_k\|^2 - \eta_2\|\text{prox}_{\mu g}(z_k) - \text{prox}_{\mu g}(z^*)\|^2, \tag{88}$$

where z^* is a fixed point of $R_{\mu f}R_{\mu g}(z^*)$, i.e., $z^* = R_{\mu f}R_{\mu g}(z^*)$. Thus, the convergence of the sequence $\{z_k\}$ generated by the DRSM (82) follows immediately from (88). Moreover, $\{\text{prox}_{\mu g}(z_k)\}$ converges to the unique solution of model (79).

Besides the global convergence of the DRSM (82) for the 'strongly+weakly' convex model (79), [125] also establishes its sublinear rate and linear rate of convergence, under additional mild conditions, and we summarize them below.

Theorem 9 *Let $\{z_k\}$ be the sequence generated by the DRSM (83) with $\nu > \omega$ and $0 < \mu < \frac{(1-\alpha)(\nu-\omega)}{\nu\omega}$. Then, there exists $\eta_1 > 0$ such that*

$$\|\widetilde{e}(z_k, \mu)\|^2 \leqslant \frac{d^2(z_0, \text{Fix}(\widetilde{T}_{DR}))}{(k+1)\eta_1\alpha^2}, \quad \forall k \geqslant 0.$$

Moreover, it holds that

$$\|\widetilde{e}(z_k, \mu)\|^2 = o(1/k), \quad as\ k \to \infty.$$

That is, the rate of asymptotic regularity for $\widetilde{T}_{DR}$ *is* $o(1/\sqrt{k})$*, where* $\widetilde{T}_{DR} := ((1-\alpha)I + \alpha R_{\mu f} R_{\mu g})$ *is called a Douglas–Rachford operator*; $\mathrm{Fix}(\widetilde{T}_{DR})$ *is the set of fixed point of* $\widetilde{T}_{DR}$; *and*

$$\widetilde{e}(z, \mu) := z - R_{\mu f} R_{\mu g}(z).$$

Theorem 10 *Let* $\{z_k\}$ *be the sequence generated by the DRSM* (83) *with* $\nu > \omega$ *and* $0 < \mu < \frac{(1-\alpha)(\nu-\omega)}{\nu\omega}$; z^* *a fixed point of* $Fix(\widetilde{T}_{DR})$*. Assume that* $\widetilde{e}(\cdot, \mu) = (I - R_{\mu f} R_{\mu g})(\cdot)$ *is metrically subregular at* z^* *for* 0 *with neighborhood* $N(z^*)$ *of* z^* *and modulus* $\kappa > 0$*. Take sufficiently small* $r > 0$ *such that* $B(z^*, r) \subseteq N(z^*)$*. Then for any starting point* $z_0 \in B(z^*, r)$ *and for any* $k \geqslant 0$*, we have*

$$d(z_{k+1}, \mathrm{Fix}(\widetilde{T}_{DR})) \leqslant \sqrt{1 - \frac{\eta_1 \alpha^2}{\kappa^2}} \cdot d(z_k, \mathrm{Fix}(\widetilde{T}_{DR})).$$

That is, the DRSM (83) *converges to* $\mathrm{Fix}(\widetilde{T}_{DR})$ *linearly.*

[126] then further weaken the requirement $\nu > \omega$ in [125] to $\nu = \omega$; and the objective function is only convex but not strongly convex. The cost is that f is further assumed to be continuously differentiable such that ∇f is Lipschitz continuous with constant $L > 0$. Nonetheless, the assumption in [126] is still weaker than that in [124]. Besides, they also allowed that the reflection (proximal) operators to be evaluated approximately, which is important from numerical point of view.

7 Conclusions

In this paper, we gave a survey of the popular numerical algorithm, the alternating direction method of multiplier, for solving large-scale structured optimization problems, i.e., the optimization problem whose objective function is the sum of some component functions and each of them only depends on its own variable; the total variables are coupled via a linear equality. The classical ADMM is for solving the special case that there are only two block of variables, which can be dated back to the Douglas–Rachford splitting method, and interests on it rose up again in the last decade, along with the requirement for solving the general case with multi-block of variables. Though there has explosive growth of the results as we have surveyed, there are still some tasks to complete.

1. Improvements of the complexity. Several sublinear rate of convergence of ADMM, measured by certain merit functions, were established to be $O(1/k)$. The existing works related to how to derive a worst-case $O(1/k^2)$ convergence rate for the ADMM either require stronger assumptions or are eligible only for some variants of the original ADMM scheme [135,136].

2. Linear rate of convergence under weaker conditions. Similar for works on $O(1/k^2)$ convergence rate for the ADMM, works for linear rate convergence of the ADMM also depend on additional assumptions which are only satisfied by special cases such as linear programming, piecewise linear-quadratic programming.
3. More efficient results on multi-block case. Currently, there are lots of works on solving the general case (1) with $m \geqslant 3$. However, due to its complication there are still great demands for further developments of efficient 'customized' methods.
4. More results for solving the nonconvex case. The development of ADMM for solving (1) involving nonconvex component functions is still in its infancy; and the current works are very limited. For example, the results [115] is for the special case that θ_2 is smooth and $A_2 = -I$, while in applications, θ_2 is nonsmooth and A_2 is even not a square matrix.

References

[1] Peng, Y., Ganesh, A., Wright, J., Xu, W., Ma, Y.: RASL: robust alignment by sparse and low-rank decomposition for linearly correlated images. IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2233–2246 (2012)
[2] Tao, M., Yuan, X.: Recovering low-rank and sparse components of matrices from incomplete and noisy observations. SIAM J. Optim. **21**(1), 57–81 (2011)
[3] Chandrasekaran, V., Parrilo, P.A., Willsky, A.S.: Latent variable graphical model selection via convex optimization. Ann. Stat. **40**(4), 1935–1967 (2012)
[4] McLachlan, G.: Discriminant Analysis and Statistical Pattern Recognition, vol. 544. Wiley (2004)
[5] Li, M., Sun, D., Toh, K.-C.: A convergent 3-block semi-proximal ADMM for convex minimization problems with one strongly convex block. Asia-Pac. J. Oper. Res. **32**(04), 1550024 (2015)
[6] Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc.: Ser. B (Methodol.) **58**(1), 267–288 (1996)
[7] Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Phys. D **60**(1–4), 259–268 (1992)
[8] Bruckstein, A.M., Donoho, D.L., Elad, M.: From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Rev. **51**(1), 34–81 (2009)
[9] Huber, P.J.: Robust Statistics. Springer (2011)
[10] Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. Roy. Stat. Soc.: Ser. B (Stat. Methodol.) **68**(1), 49–67 (2006)
[11] Candès, E.J., Recht, B.: Exact matrix completion via convex optimization. Found. Comput. Math. **9**(6), 717–772 (2009)
[12] Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? J. ACM (JACM) **58**(3), 1–37 (2011)
[13] Donoho, D.L.: De-noising by soft-thresholding. IEEE Trans. Inf. Theory **41**(3), 613–627 (1995)
[14] Donoho, D.L.: High-dimensional data analysis: the curses and blessings of dimensionality. AMS Math. Chall. Lect. 1–32, 375 (2000)

[15] Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011)
[16] Glowinski, R.: On alternating direction methods of multipliers: a historical perspective. In: Modeling, Simulation and Optimization for Science and Technology, pp. 59–82. Springer (2014)
[17] Glowinski, R., Marroco, A.: Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires", Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique **9**(R2), 41–76 (1975)
[18] Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comput. Math. Appl. **2**(1), 17–40 (1976)
[19] Gabay, D.: Applications of the method of multipliers to variational inequalities. Stud. Math. Appl. **15**, 299–331 (1983)
[20] Eckstein, J., Yao, W.: Understanding the convergence of the alternating direction method of multipliers: theoretical and computational perspectives. Pac. J. Optim. **11**(4), 619–644 (2015)
[21] Chen, C., He, B., Ye, Y., Yuan, X.: The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. Math. Program. **155**(1–2), 57–79 (2016)
[22] Kurdyka, K.: On gradients of functions definable in *o*-minimal structures. Annales de l'institut Fourier **48**, 769–783 (1998)
[23] Lojasiewicz, S.: Une propriété topologique des sous-ensembles analytiques réels, les Équations auxdérivées partielles. Les Éditions aux Dérivées Partielles **117**, 87–89 (1963)
[24] Hestenes, M.R.: Multiplier and gradient methods. J. Optim. Theory Appl. **4**(5), 303–320 (1969)
[25] Powell, M.J.: A method for nonlinear constraints in minimization problems. In: Fletcher, R. (ed.) Optimization, pp. 283–298. Academic Press (1969)
[26] Hong, M., Luo, Z.-Q.: On the linear convergence of the alternating direction method of multipliers. Math. Program. **162**(1–2), 165–199 (2017)
[27] Bertsekas, D.P., Nedi, A., Ozdaglar, A.E.: Convex Analysis and Optimization. Athena Scientific (2003)
[28] Rockafellar, R.T.: Convex Analysis. Princeton University Press (2015)
[29] Zhang, J., Ge, S., Chang, T.-H., Luo, Z.-Q.: Decentralized non-convex learning with linearly coupled constraints. arXiv:2103.05378 (2021)
[30] Eckstein, J., Bertsekas, D.P.: On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. **55**(1–3), 293–318 (1992)
[31] Ryu, E.K., Boyd, S.: Primer on monotone operator methods. Appl. Comput. Math. **15**(1), 3–43 (2016)
[32] Douglas, J., Rachford, H.: On the numerical solution of heat conduction problems in two and three space variables. Trans. Am. Math. Soc. **82**(2), 421–439 (1956)
[33] Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM J. Control. Optim. **14**(5), 877–898 (1976)
[34] Rockafellar, R.: Monotone operators and augmented Lagrangian methods in nonlinear programming. In: Nonlinear Programming 3, pp. 1–25. Elsevier (1978)
[35] He, B., Yang, H.: Some convergence properties of a method of multipliers for linearly constrained monotone variational inequalities. Oper. Res. Lett. **23**(3–5), 151–161 (1998)
[36] Kontogiorgis, S., Meyer, R.R.: A variable-penalty alternating directions method for convex optimization. Math. Program. **83**(1–3), 29–53 (1998)
[37] He, B., Yang, H., Wang, S.: Alternating direction method with self-adaptive penalty parameters for monotone variational inequalities. J. Optim. Theory Appl. **106**(2), 337–356 (2000)
[38] He, B., Liao, L.-Z., Han, D., Yang, H.: A new inexact alternating directions method for monotone variational inequalities. Math. Program. **92**(1), 103–118 (2002)
[39] Chan, R., Tao, M., Yuan, X.: Constrained total variation deblurring models and fast algorithms based on alternating direction method of multipliers. SIAM J. Imag. Sci. **6**(1), 680–697 (2013)
[40] Han, D., He, H., Yang, H., Yuan, X.: A customized Douglas–Rachford splitting algorithm for separable convex minimization with linear constraints. Numer. Math. **127**(1), 167–200 (2014)
[41] Rockafellar, R.T.: Augmented Lagrangians and applications of the proximal point algorithm in convex programming. Math. Oper. Res. **1**(2), 97–116 (1976)
[42] Eckstein, J., Fukushima, M.: Some reformulations and applications of the alternating direction method of multipliers. In: Large Scale Optimization, pp. 115–134. Springer (1994)

[43] Facchinei, F., Pang, J.S.: Finite-Dimensional Variational Inequalities and Complementarity Problems. Springer Science & Business Media, Berlin (2003)
[44] Lin, Z., Liu, R., Su, Z.: Linearized alternating direction method with adaptive penalty for low-rank representation. In: Advances in Neural Information Processing Systems, pp. 612–620 (2011)
[45] Xu, M., Wu, T.: A class of linearized proximal alternating direction methods. J. Optim. Theory Appl. **151**(2), 321–337 (2011)
[46] Yang, J., Yuan, X.: Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization. Math. Comput. **82**(281), 301–329 (2013)
[47] Wang, X., Yuan, X.: The linearized alternating direction method of multipliers for Dantzig selector. SIAM J. Sci. Comput. **34**(5), A2792–A2811 (2012)
[48] Chen, G., Teboulle, M.: A proximal-based decomposition method for convex minimization problems. Math. Program. **64**(1–3), 81–101 (1994)
[49] Shefi, R., Teboulle, M.: Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization. SIAM J. Optim. **24**(1), 269–297 (2014)
[50] Han, D., He, B.: A new accuracy criterion for approximate proximal point algorithms. J. Math. Anal. Appl. **263**(2), 343–354 (2001)
[51] Solodov, M.V., Svaiter, B.F.: A hybrid projection-proximal point algorithm. J. Convex Anal. **6**(1), 59–70 (1999)
[52] Solodov, M.V., Svaiter, B.F.: A hybrid approximate extragradient-proximal point algorithm using the enlargement of a maximal monotone operator. Set-Valued Anal. **7**(4), 323–345 (1999)
[53] Han, D.: A new hybrid generalized proximal point algorithm for variational inequality problems. J. Global Optim. **26**(2), 125–140 (2003)
[54] He, B., Yang, Z., Yuan, X.: An approximate proximal-extragradient type method for monotone variational inequalities. J. Math. Anal. Appl. **300**(2), 362–374 (2004)
[55] Eckstein, J., Silva, P.J.: A practical relative error criterion for augmented Lagrangians. Math. Program. **141**(1–2), 319–348 (2013)
[56] Eckstein, J., Yao, W.: Approximate ADMM algorithms derived from Lagrangian splitting. Comput. Optim. Appl. **68**(2), 363–405 (2017)
[57] Eckstein, J., Yao, W.: Relative-error approximate versions of Douglas–Rachford splitting and special cases of the ADMM. Math. Program. **170**(2), 417–444 (2018)
[58] Xie, J.: On inexact ADMMs with relative error criteria. Comput. Optim. Appl. **71**(3), 743–765 (2018)
[59] Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course, vol. 87. Springer Science & Business Media, Berlin (2013)
[60] He, B., Yuan, X.: On the $O(1/n)$ convergence rate of the Douglas–Rachford alternating direction method. SIAM J. Numer. Anal. **50**(2), 700–709 (2012)
[61] Monteiro, R.D., Svaiter, B.F.: Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. SIAM J. Optim. **23**(1), 475–507 (2013)
[62] Cai, X., Han, D.: $O(1/t)$ complexity analysis of the generalized alternating direction method of multipliers. Sci. China Math. **62**(4), 795–808 (2019)
[63] Ouyang, Y., Chen, Y., Lan, G., Pasiliao, E., Jr.: An accelerated linearized alternating direction method of multipliers. SIAM J. Imag. Sci. **8**(1), 644–681 (2015)
[64] He, B., Yuan, X.: On non-ergodic convergence rate of Douglas–Rachford alternating direction method of multipliers. Numer. Math. **130**(3), 567–577 (2015)
[65] Deng, W., Yin, W.: On the global and linear convergence of the generalized alternating direction method of multipliers. J. Sci. Comput. **66**(3), 889–916 (2016)
[66] Gao, X., Jiang, B., Zhang, S.: On the information-adaptive variants of the ADMM: an iteration complexity perspective. J. Sci. Comput. **76**(1), 327–363 (2018)
[67] Gonçalves, M.L., Melo, J.G., Monteiro, R.D.: Improved pointwise iteration-complexity of a regularized ADMM and of a regularized non-Euclidean HPE framework. SIAM J. Optim. **27**(1), 379–407 (2017)
[68] Lions, P.-L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. **16**(6), 964–979 (1979)
[69] Luque, F.J.: Asymptotic convergence analysis of the proximal point algorithm. SIAM J. Control. Optim. **22**(2), 277–293 (1984)
[70] Eckstein, J.: Splitting Methods for Monotone Operators with Applications to Parallel Optimization. PhD thesis, Massachusetts Institute of Technology (1989)

[71] Boley, D.: Local linear convergence of ADMM on quadratic or linear programs. SIAM J. Optim. **23**(4), 2183–2207 (2013)
[72] Han, D., Yuan, X.: Local linear convergence of the alternating direction method of multipliers for quadratic programs. SIAM J. Numer. Anal. **51**(6), 3446–3457 (2013)
[73] Luo, Z.-Q., Tseng, P.: Error bounds and convergence analysis of feasible descent methods: a general approach. Ann. Oper. Res. **46**(1), 157–178 (1993)
[74] Zhu, T., Yu, Z.: A simple proof for some important properties of the projection mapping. Math. Inequal. Appl. **7**, 453–456 (2004)
[75] Yang, W.H., Han, D.: Linear convergence of the alternating direction method of multipliers for a class of convex optimization problems. SIAM J. Numer. Anal. **54**(2), 625–640 (2016)
[76] Zheng, X.Y., Ng, K.F.: Metric subregularity of piecewise linear multifunctions and applications to piecewise linear multiobjective optimization. SIAM J. Optim. **24**(1), 154–174 (2014)
[77] Sun, J.: On monotropic piecewise quadratic programming (network, algorithm, convex programming, decomposition method) Ph.D. Dissertation. University of Washington, USA. Order Number: AAI8706680 (1986)
[78] Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis, vol. 317. Springer Science & Business Media, Berlin (2009)
[79] Han, D., Sun, D., Zhang, L.: Linear rate convergence of the alternating direction method of multipliers for convex composite programming. Math. Oper. Res. **43**(2), 622–637 (2017)
[80] Dontchev, A.L., Rockafellar, R.T.: Implicit Functions and Solution Mappings, vol. 543. Springer (2009)
[81] Chang, T.-H., Hong, M., Wang, X.: Multi-agent distributed optimization via inexact consensus ADMM. IEEE Trans. Signal Process. **63**(2), 482–497 (2014)
[82] Shi, W., Ling, Q., Yuan, K., Wu, G., Yin, W.: On the linear convergence of the ADMM in decentralized consensus optimization. IEEE Trans. Signal Process. **62**(7), 1750–1761 (2014)
[83] Han, D., Yuan, X.: A note on the alternating direction method of multipliers. J. Optim. Theory Appl. **155**(1), 227–238 (2012)
[84] Chen, C., Shen, Y., You, Y.: On the convergence analysis of the alternating direction method of multipliers with three blocks. Abstr. Appl. Anal. **2013**,(2013)
[85] Lin, T., Ma, S., Zhang, S.: On the sublinear convergence rate of multi-block ADMM. J. Oper. Res. Soc. China **3**(3), 251–274 (2015)
[86] Cai, X., Han, D., Yuan, X.: On the convergence of the direct extension of ADMM for three-block separable convex minimization models with one strongly convex function. Comput. Optim. Appl. **66**(1), 39–73 (2017)
[87] Lin, T., Ma, S., Zhang, S.: On the global linear convergence of the ADMM with multiblock variables. SIAM J. Optim. **25**(3), 1478–1497 (2015)
[88] Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces, vol. 408. Springer (2011)
[89] He, B., Tao, M., Yuan, X.: Alternating direction method with Gaussian back substitution for separable convex programming. SIAM J. Optim. **22**(2), 313–340 (2012)
[90] Ye, C., Yuan, X.: A descent method for structured monotone variational inequalities. Optim. Methods Softw. **22**(2), 329–338 (2007)
[91] Han, D., Yuan, X., Zhang, W., Cai, X.: An ADM-based splitting method for separable convex programming. Comput. Optim. Appl. **54**(2), 343–369 (2013)
[92] Han, D., Yuan, X., Zhang, W.: An augmented Lagrangian based parallel splitting method for separable convex minimization with applications to image processing. Math. Comput. **83**(289), 2263–2291 (2014)
[93] He, B.: Parallel splitting augmented Lagrangian methods for monotone structured variational inequalities. Comput. Optim. Appl. **42**(2), 195–212 (2009)
[94] Wang, K., Han, D., Xu, L.: A parallel splitting method for separable convex programs. J. Optim. Theory Appl. **159**(1), 138–158 (2013)
[95] He, B., Hou, L., Yuan, X.: On full Jacobian decomposition of the augmented Lagrangian method for separable convex programming. SIAM J. Optim. **25**(4), 2274–2312 (2015)
[96] Deng, W., Lai, M.-J., Peng, Z., Yin, W.: Parallel multi-block ADMM with $o(1/k)$ convergence. J. Sci. Comput. **71**(2), 712–736 (2017)
[97] He, H., Han, D.: A distributed Douglas–Rachford splitting method for multi-block convex minimization problems. Adv. Comput. Math. **42**(1), 27–53 (2016)

[98] Cao, C., Han, D., Xu, L.: A new partial splitting augmented Lagrangian method for minimizing the sum of three convex functions. Appl. Math. Comput. **219**(10), 5449–5457 (2013)
[99] Hou, L., He, H., Yang, J.: A partially parallel splitting method for multiple-block separable convex programming with applications to robust PCA. Comput. Optim. Appl. **63**(1), 273–303 (2016)
[100] Han, D., Kong, W., Zhang, W.: A partial splitting augmented Lagrangian method for low patch-rank image decomposition. J. Math. Imag. Vis. **51**(1), 145–160 (2015)
[101] Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka–Lojasiewicz inequality. Math. Oper. Res. **35**(2), 438–457 (2010)
[102] Attouch, H., Bolte, J., Svaiter, B.F.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss–Seidel methods. Math. Program. **137**(1–2), 91–129 (2013)
[103] Bolte, J., Daniilidis, A., Lewis, A.: The Lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM J. Optim. **17**(4), 1205–1223 (2007)
[104] Bolte, J., Daniilidis, A., Lewis, A., Shiota, M.: Clarke subgradients of stratifiable functions. SIAM J. Optim. **18**(2), 556–572 (2007)
[105] Mordukhovich, B.S.: Variational Analysis and Generalized Differentiation I: Basic Theory, vol. 330. Springer Science & Business Media, Berlin (2006)
[106] Absil, P.-A., Mahony, R., Andrews, B.: Convergence of the iterates of descent methods for analytic cost functions. SIAM J. Optim. **16**(2), 531–547 (2005)
[107] Attouch, H., Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Math. Program. **116**(1–2), 5–16 (2009)
[108] Bolte, J., Daniilidis, A., Ley, O., Mazet, L.: Characterizations of Lojasiewicz inequalities: subgradient flows, talweg, convexity. Trans. Am. Math. Soc. **362**(6), 3319–3363 (2010)
[109] Merlet, B., Pierre, M.: Convergence to equilibrium for the backward Euler scheme and applications. Commun. Pure Appl. Anal. **9**(3), 685–702 (2010)
[110] Noll, D.: Convergence of non-smooth descent methods using the Kurdyka–Lojasiewicz inequality. J. Optim. Theory Appl. **160**(2), 553–572 (2014)
[111] Chouzenoux, E., Pesquet, J.-C., Repetti, A.: Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function. J. Optim. Theory Appl. **162**(1), 107–132 (2014)
[112] Frankel, P., Garrigos, G., Peypouquet, J.: Splitting methods with variable metric for Kurdyka–Lojasiewicz functions and general convergence rates. J. Optim. Theory Appl. **165**(3), 874–900 (2015)
[113] Li, G., Pong, T.K.: Global convergence of splitting methods for nonconvex composite optimization. SIAM J. Optim. **25**(4), 2434–2460 (2015)
[114] Wang, F., Xu, Z., Xu, H.-K.: Convergence of bregman alternating direction method with multipliers for nonconvex composite problems. arXiv:1410.8625 (2014)
[115] Guo, K., Han, D., Wu, T.: Convergence of alternating direction method for minimizing sum of two nonconvex functions with linear constraints. Int. J. Comput. Math. **94**(8), 1653–1669 (2017)
[116] Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. **146**(1–2), 459–494 (2014)
[117] Wang, Y., Yin, W., Zeng, J.: Global convergence of ADMM in nonconvex nonsmooth optimization. J. Sci. Comput. **78**(1), 29–63 (2019)
[118] Jia, Z., Gao, X., Cai, X., Han, D.: Local linear convergence of the alternating direction method of multipliers for nonconvex separable optimization problems. J. Optim. Theory Appl. **188**, 1–25 (2021)
[119] Jiang, B., Lin, T., Ma, S., Zhang, S.: Structured nonconvex and nonsmooth optimization: algorithms and iteration complexity analysis. Comput. Optim. Appl. **72**(1), 115–157 (2019)
[120] Zhang, J., Luo, Z.-Q.: A proximal alternating direction method of multiplier for linearly constrained nonconvex minimization. SIAM J. Optim. **30**(3), 2272–2302 (2020)
[121] Guo, K., Han, D., Wang, D.Z., Wu, T.: Convergence of ADMM for multi-block nonconvex separable optimization models. Front. Math. China **12**(5), 1139–1162 (2017)
[122] Guo, K., Han, D., Wu, T.: Convergence of ADMM for optimization problems with nonseparable nonconvex objective and linear constraints. Pac. J. Optim. **14**(3), 489–506 (2018)
[123] Hong, M., Luo, Z.-Q., Razaviyayn, M.: Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. SIAM J. Optim. **26**(1), 337–364 (2016)

[124] Bayram, I., Selesnick, I.W.: The Douglas–Rachford algorithm for weakly convex penalties. arXiv preprint arXiv:1511.03920 (2015)
[125] Guo, K., Han, D., Yuan, X.: Convergence analysis of Douglas–Rachford splitting method for 'strongly + weakly' convex programming. SIAM J. Numer. Anal. **55**(4), 1549–1577 (2017)
[126] Guo, K., Han, D.: A note on the Douglas–Rachford splitting method for optimization problems involving hypoconvex functions. J. Glob. Optim. **72**(3), 431–441 (2018)
[127] Bayram, I.: Penalty functions derived from monotone mappings. IEEE Signal Process. Lett. **22**(3), 265–269 (2014)
[128] Chen, L., Gu, Y.: The convergence guarantees of a non-convex approach for sparse recovery. IEEE Trans. Signal Process. **62**(15), 3754–3767 (2014)
[129] Selesnick, I.W., Bayram, I.: Sparse signal estimation by maximally sparse convex optimization. IEEE Trans. Signal Process. **62**(5), 1078–1092 (2014)
[130] Guo, K., Yuan, X., Zeng, S.: Convergence analysis of ISTA and FISTA for 'strongly + semi' convex programming (2016)
[131] Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**(456), 1348–1360 (2001)
[132] Zhang, C.-H.: Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. **38**(2), 894–942 (2010)
[133] Li, G., Pong, T.K.: Douglas–Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems. Math. Program. **159**(1–2), 371–401 (2016)
[134] Mollenhoff, T., Strekalovskiy, E., Moeller, M., Cremers, D.: The primal-dual hybrid gradient method for semiconvex splittings. SIAM J. Imag. Sci. **8**(2), 827–857 (2015)
[135] Goldstein, T., O'Donoghue, B., Setzer, S., Baraniuk, R.: Fast alternating direction optimization methods. SIAM J. Imag. Sci. **7**(3), 1588–1623 (2014)
[136] Tian, W., Yuan, X.: An alternating direction method of multipliers with a worst-case $O(1/n^2)$ convergence rate. Math. Comput. **88**(318), 1685–1713 (2019)