Journal of Machine Learning Research 21 (2020) 1-75 Submitted 8/18; Revised 4/20; Published 4/20

Discerning the Linear Convergence of ADMM for Structured Convex Optimization through the Lens of Variational Analysis

Xiaoming Yuan XMYUAN@HKU.HK
Shangzhi Zeng ZENGSZ@CONNECT.HKU.HK
Department of Mathematics
The University of Hong Kong
Hong Kong SAR, China

Jin Zhang ZHANGJ9@SUSTECH.EDU.CN
Corresponding author
Department of Mathematics
SUSTech International Center for Mathematics
Southern University of Science and Technology
National Center for Applied Mathematics Shenzhen
Shenzhen, Guangdong, China

Editor: Mark Schmidt

Abstract

Despite the rich literature, the linear convergence of alternating direction method of multipliers (ADMM) has not been fully understood even for the convex case. For example, the linear convergence of ADMM can be empirically observed in a wide range of applications arising in statistics, machine learning, and related areas, while existing theoretical results seem to be too stringent to be satisfied or too ambiguous to be checked and thus why the ADMM performs linear convergence for these applications still seems to be unclear. In this paper, we systematically study the local linear convergence of ADMM in the context of convex optimization through the lens of variational analysis. We show that the local linear convergence of ADMM can be guaranteed without the strong convexity of objective functions together with the full rank assumption of the coefficient matrices, or the full polyhedricity assumption of their subdifferential; and it is possible to discern the local linear convergence for various concrete applications, especially for some representative models arising in statistical learning. We use some variational analysis techniques sophisticatedly; and our analysis is conducted in the most general proximal version of ADMM with Fortin and Glowinski's larger step size so that all major variants of the ADMM known in the literature are covered.

Keywords: Convex programming, variational analysis, alternating direction method of multipliers, linear convergence, calmness, metric subregularity, machine learning, statistics.

1. Introduction

We consider the convex minimization problem with linear constraints and an objective function in form of the sum of two functions without coupled variables:

$$\begin{aligned} \min_{x,y} \quad & f(x)+g(y) \\ s.t. \quad & Ax+By=b, \end{aligned} \tag{1}$$

where $A \in \mathbb{R}^{m\times n_1}$ and $B \in \mathbb{R}^{m\times n_2}$ are two given matrices, $x \in \mathbb{R}^{n_1}, y \in \mathbb{R}^{n_2}$, and $f : \mathbb{R}^{n_1} \to (-\infty,\infty]$ and $g : \mathbb{R}^{n_2} \to (-\infty,\infty]$ are convex, proper, lower semicontinuous functions. We work in finite dimensional Euclidean spaces composed by column vectors, where $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product and $\|\cdot\|$ denotes the Euclidean norm.

The abstract model (1) is general enough to capture a number of applications arising in areas such as statistical learning, image processing, computer vision, and distributed optimization, in which one of the functions in the objective is a data fidelity term and the other one is a regularization term.

To solve Problem (1), the alternating direction method of multipliers (ADMM) proposed in (Chan and Glowinski, 1978; Glowinski and Marroco, 1975) becomes a benchmark solver because of its features of easy implementability, competitive numerical performance and wide applicability in various areas. The ADMM has been receiving attention from a broad spectrum of areas; and various variants have been well studied in the literature. We refer to (Boyd et al., 2011; Eckstein and Yao, 2015; Glowinski, 2014) for some review papers. Even though the original ADMM is of our core interest, to capture its various variants simultaneously, as Han et al. (2017); Li et al. (2016), we study the so-called proximal version of ADMM with positive semidefinite regularization terms for updating the primal variables (x, y) and Fortin and Glowinski's larger step size (see Fortin and Glowinski, 1983) for updating the dual variable λ, as shown below.

Algorithm 1: Proximal ADMM with Fortin and Glowinski's larger step size for (1)

Initial $\beta > 0$, $G_1 \succeq 0$, $G_2 \succeq 0$, $\gamma \in (0, \frac{1+\sqrt{5}}{2})$, and choose value $x^0 \in \mathbb{R}^{n_1}$, $y^0 \in \mathbb{R}^{n_2}$, $\lambda^0 \in \mathbb{R}^m$.

for $k = 0, 1, 2, \ldots$ **do**

$$\begin{aligned} x^{k+1} &\in \arg\min_x \{f(x) - \langle\lambda^k, Ax + By^k - b\rangle + \frac{\beta}{2}\|Ax + By^k - b\|^2 + \frac{1}{2}\|x - x^k\|^2_{G_1}\}, \\ y^{k+1} &\in \arg\min_y \{g(y) - \langle\lambda^k, Ax^{k+1} + By - b\rangle + \frac{\beta}{2}\|Ax^{k+1} + By - b\|^2 + \frac{1}{2}\|y - y^k\|^2_{G_2}\}, \\ \lambda^{k+1} &= \lambda^k - \gamma\beta(Ax^{k+1} + By^{k+1} - b), \end{aligned} \tag{2}$$

end

Algorithm 1 is abbreviated as PADMM-FG hereafter for succinctness. Note that a number of variants of the ADMM, whose theoretical and algorithmic interests have been studied individually in the literature, can be recovered by the PADMM-FG (2). Obviously, the original ADMM in (Chan and Glowinski, 1978; Glowinski and Marroco, 1975) is the

special case of (2) with $G_1 = 0$, $G_2 = 0$ and $\gamma = 1$; the ADMM variant with Fortin and Glowinski's larger step size in (Fortin and Glowinski, 1983) is the special case of (2) with $G_1 = 0$, $G_2 = 0$ and $\gamma \in (0, \frac{1+\sqrt{5}}{2})$; the linearized version of ADMM studied in (Esser et al., 2010; Wang and Yuan, 2012; Yang and Yuan, 2013) is the special case of (2) where $G_1 = rI - \beta A^T A$ with $r > \beta\|A^T A\|$, $G_2 = 0$ and $\gamma = 1$; and more generally, the proximal version of ADMM in (He et al., 2002) is the special case of (2) with $G_1 \succ 0$, $G_2 \succ 0$ and $\gamma = 1$. Note that, as in (Han et al., 2017; Li et al., 2016), it is by-default assumed that $\beta A^T A + G_1 \succ 0$ and $\beta B^T B + G_2 \succ 0$ if the general PADMM-FG (2) is studied. We refer to, for example, (Lin et al., 2015; Wang and Yuan, 2012; Yang and Yuan, 2013), for various applications of the linearized ADMM in areas such as statistical learning and computer vision, and (Glowinski and Le Tallec, 1989; He et al., 2011; Sun and Zhang, 2010; Wen et al., 2010) for numerical acceleration performance of Fortin and Glowinski's larger step size $\gamma \in (0, \frac{1+\sqrt{5}}{2})$. Furthermore, as found in (Gabay, 1983), the original ADMM is equivalent to the application of the general Douglas-Rachford splitting method (DRSM) proposed in (Douglas and Rachford, 1956; Lions and Mercier, 1979) to a stationary system to the dual of Problem (1); and as analyzed in (Esser et al., 2010; Shefi, 2015), if the special case of Problem (1) with $B = -I$ and $b = 0$ is considered, then the linearized ADMM turns out to be highly relevant to the so-called primal-dual hybrid gradient (PDHG) studied in (Chambolle and Pock, 2011) for saddle-point problems. Finally, it is worthy mentioning that in some existing literatures such as (Gabay and Mercier, 1976; Tao and Yuan, 2018b), convergence of the original ADMM with $\gamma \in (0, 2)$ has been discussed for some special cases of the model (1) with quadratic or linearity assumptions on the functions f and/or g. But here we focus on the generic case of f and g in the model (1) and thus do not discuss the possibility of $\gamma \in (0, 2)$ for the PADMM-FG (2); more rationales can be referred to (Glowinski, 2013; Tao and Yuan, 2018a).

Our primary purposes are: (1) to discuss the nonergodic local linear convergence of the original ADMM, the linearized ADMM and the general PADMM-FG (2) through the lens of variational analysis; and (2) to show that it is possible to discern the linear convergence behaviors as well as the exact convergence rates of the PADMM-FG (2) for various concrete applications in statistical and machine learning problems of recent interest. We also deepen our discussion via the dual perspective and show, as byproducts, how to discern the linear convergence of other methods which are highly relevant to various variants of the ADMM, including the DRSM in the general operator form and the PDHG for saddle-point problems.

1.1. State-of-the-art

Under some mild conditions such as the non-emptiness of the solution set of Problem (1), convergence properties have been well studied in earlier literature for the original ADMM and its variants; (see, e.g. Eckstein and Bertsekas, 1990, 1992; Fortin and Glowinski, 1983; Gabay and Mercier, 1976; Glowinski and Marroco, 1975; Glowinski and Le Tallec, 1989; He and Yang, 1998; Lions and Mercier, 1979). Recently, in (He and Yuan, 2012a, 2015; Monteiro and Svaiter, 2013), the worst-case $O(1/k)$ sublinear convergence rate measured by the iteration complexity has been established for the original ADMM and the linearized ADMM in both ergodic and nonergodic senses, where k is the iteration counter. The linear convergence of ADMM has also been discussed in the literature under further assumptions

beyond convexity, typically smoothing and strong convexity assumptions on objective functions; see, e.g., (Davis and Yin, 2017; Deng and Yin, 2016; Nishihara et al., 2015). Below we try to summarize some representative scenarios in which the global linear convergence of the PADMM-FG or its special cases is known.[1]

(S1) If f (Resp. g) is strongly convex, and differentiable with a Lipschitz continuous gradient, together with full row rank condition of the coefficient matrix A (Resp. B), then the sequence $\{(x^k, By^k, \lambda^k)\}$ (or $\{(Ax^k, y^k, \lambda^k)\}$) generated by the PADMM-FG (2) with $G_1 \succ 0, G_2 \succeq 0$ (Resp. $G_1 \succeq 0, G_2 \succ 0$) converges linearly; (see, e.g. Deng and Yin, 2016; Giselsson and Boyd, 2016).

(S2) If both f and g are strongly convex, and differentiable with Lipschitz continuous gradients, then the sequence $\{(x^k, y^k, \lambda^k)\}$ generated by the PADMM-FG (2) with $\gamma = 1$ converges linearly; (see, e.g. Deng and Yin, 2016).

(S3) If f (Resp. g) is strongly convex, g (Resp. f) is differentiable with a Lipschitz continuous gradient, together with full row rank condition of the coefficient matrix B (Resp. A), then the sequence $\{(Ax^k, By^k, \lambda^k)\}$ generated by the original ADMM converges linearly; (see, e.g. Davis and Yin, 2017).

The strong convexity of objective functions and the full row-rank assumption of coefficient matrices, however, can be barely satisfied simultaneously for applications. Below, we show by a very simple application in machine learning that these scenarios (S1)-(S3) might be too restrictive.

Example 1 *Regularization methods for simultaneous variable selection and estimation in linear regression and more general contexts have received intense interest recently. In particular, the least absolute shrinkage and selection operator (LASSO) which was proposed in (Tibshirani et al., 2005) has been used extensively in high-dimensional statistics and machine learning. It uses the squared error and an ℓ_1-norm regularizer which induces sparsity in the solution. As a result, the features which have non-zero coefficients can be easily selected.*

Consider the LASSO model and its dual form as illustrative examples:

$$\min_{\mathbb{x}\in\mathbb{R}^m} \tfrac{1}{2}\|L\mathbb{x} - \mathbb{b}\|^2 + \nu\|\mathbb{x}\|_1 \qquad \begin{array}{ll} \min_{\mathbb{y}\in\mathbb{R}^l} & \tfrac{1}{2}\|\mathbb{y}\|^2 - \mathbb{b}^T\mathbb{y} \\ s.t. & \|L^T\mathbb{y}\|_\infty \le \nu, \end{array}$$

where $L \in \mathbb{R}^{l\times m}$ with $l \ll m$, $\mathbb{b} \in \mathbb{R}^l$, $\nu > 0$ and $\|\mathbb{x}\|_1$ is the ℓ_1-norm defined as $\sum_{i=1}^m |x_i|$. By introducing an auxiliary variable $\mathbb{z} = \mathbb{x}$ for the nonsmooth ℓ_1-norm regularizer, the LASSO model can be reformulated as a special case of Problem (1). Certainly, unless L is of full column rank , an assumption contradicting with the purpose of variable selection, the objective function of the reformulated problem is not strongly convex. On the other hand, in some practices one may employ the ADMM to solve the dual form so as to ensure the desired strong convexity in the objective function. But, in this case, by introducing an auxiliary variable $\mathbb{z} = L^T\mathbb{y}$ for the $\ell_\infty-$norm ball constraint, in general L^T does not meet the full row rank assumption.

1. For succinctness, several scenarios discussed in (Davis and Yin, 2017; Deng and Yin, 2016) are not included because they seem to be less practical to find applications so far.

Hence, even for the LASSO case, regardless of the degeneracy of L, the well observed linear convergence of ADMM cannot be justified by scenarios (S1)-(S3). More theories are urged to justify the repertoire of known practical instances that can be efficiently solved by the ADMM and its variants with linear convergence.

This observation has well motivated another line of analysis for studying the linear convergence of the ADMM, apart from the strong convexity assumption on the objective function and/or the full row-rank assumption on coefficient matrices, but on the metric subregularity, calmness, or error bound, that relates the distance of a point to the solution set to a certain optimality residual function.

At the core of variational analysis, the concepts of metric subregularity and calmness have been playing an important role in various optimization topics.

Definition 1 *A set-valued map $\Psi : \mathbb{R}^n \rightrightarrows \mathbb{R}^q$ is said to be metrically subregular at $(\bar{u}, \bar{v}) \in \mathrm{gph}\,(\Psi)$ if, for some $\epsilon > 0$, there exists $\kappa \geq 0$ such that*

$$\mathrm{dist}\left(u, \Psi^{-1}(\bar{v})\right) \leq \kappa \mathrm{dist}\left(\bar{v}, \Psi(u)\right), \quad \forall u \in \mathbb{B}_\epsilon(\bar{u}),$$

where $dist(d, \mathcal{D}) := \inf\{\|d - d'\| \,\big|\, d' \in \mathcal{D}\}$ for a given subset $\mathcal{D}$ and vector d in the same space, and $\mathbb{B}_\epsilon(\bar{u}) := \{u : \|u - \bar{u}\| < \epsilon\}$.

Definition 2 *A set-valued map $\Phi : \mathbb{R}^q \rightrightarrows \mathbb{R}^n$ is said to be calm (or pseudo upper-Lipschitz continuous, (see Rockafellar and Wets, 2009; Ye and Ye, 1997)) around $(\bar{p}, \bar{x}) \in gph\Phi$ if there exist a neighborhood $\mathbb{B}_{\epsilon_1}(\bar{p})$ of $\bar{p}$, a neighborhood $\mathbb{B}_{\epsilon_2}(\bar{x})$ of $\bar{x}$ and $\kappa \geq 0$ such that*

$$\Phi(p) \cap \mathbb{B}_{\epsilon_2}(\bar{x}) \subseteq \Phi(\bar{p}) + \kappa \|p - \bar{p}\| \overline{\mathbb{B}}, \quad \forall p \in \mathbb{B}_{\epsilon_1}(\bar{p}), \tag{3}$$

where $\overline{\mathbb{B}}$ denotes the closed unit ball centered at the origin.

The calmness property is a Lipschitz-like property of the corresponding perturbed set-valued map. It is well-known that a set-valued map is calm if and only if its inverse map is metrically subregular.

To elucidate on applications to the convergence rate analysis for the ADMM and its variants, we define the set-valued map $T_{KKT} : \mathbb{R}^{n_1+n_2+m} \rightrightarrows \mathbb{R}^{n_1+n_2+m}$, which is associated with the Karush-Kuhn-Tucker (KKT) system of Problem (1), as the following:

$$T_{KKT}(x, y, \lambda) := \begin{pmatrix} \partial f(x) - A^T\lambda \\ \partial g(y) - B^T\lambda \\ Ax + By - b \end{pmatrix}. \tag{4}$$

Obviously, any (x, y, λ) satisfying $0 \in T_{KKT}(x, y, \lambda)$ is a KKT point. In terms of T_{KKT}, we may define the KKT residue $\mathrm{Res}(x, y, \lambda)$ as

$$\mathrm{Res}(x, y, \lambda) = \mathrm{dist}\left(0, T_{KKT}(x, y, \lambda)\right) \tag{5}$$

and use $\mathrm{Res}(x, y, \lambda)$ to measure the optimality of the iterate (x, y, λ). In (Yang and Han, 2016), linear convergence of the linearized ADMM is established under the metric subregularity of T_{KKT}. For the special case of the PADMM-FG (2) with $G_2 = 0$ and $\gamma = 1$, it is known that the second part of the KKT system, i.e., $0 \in \partial g(y^k) - B^T\lambda^k$, holds for all

iterates (x^k, y^k, λ^k) . Very recently, in (Liu et al., 2018), the authors first take advantage of this observation to improve the results in (Yang and Han, 2016). In particular, let

$$\Omega_g := \{(x, y, \lambda) \mid 0 \in \partial g(y) - B^T\lambda\},$$

it is shown in (Liu et al., 2018) that the linear convergence of the PADMM-FG with $G_2 = 0$ and $\gamma = 1$ is guaranteed under the metric subregularity of T_{KKT} over the set Ω_g. In the literature, in addition to T_{KKT}, other KKT mappings have been defined as well for studying the linear convergence of the ADMM and its variants. For instance, based on the so-called natural map (see Facchinei and Pang, 2007, page 83) in terms of the Moreau-Yosida proximal mapping, the following mapping is used in (Han et al., 2017; Han and Yuan, 2013):

$$T^p_{KKT}(x, y, \lambda) = \begin{pmatrix} x - \mathrm{Prox}_f(x + A^T\lambda) \\ y - \mathrm{Prox}_g(y + B^T\lambda) \\ Ax + By - b \end{pmatrix}, \tag{6}$$

where Prox_h is the proximal mapping associated with the function h, i.e.,

$$\mathrm{Prox}_h(a) := \arg\min_{t\in\mathbb{R}^n} \left\{ h(t) + \frac{1}{2}\|t - a\|^2 \right\}.$$

Obviously, any (x, y, λ) such that $0 = T^p_{KKT}(x, y, \lambda)$ is also a KKT point. In (Han et al., 2017), the linear convergence of PADMM-FG is proved when T^p_{KKT} is metrically subregular. Indeed, it is proved in (Liu et al., 2018) via a perturbation perspective that, despite the different forms in notation, the metric subregularity conditions of T_{KKT} and T^p_{KKT} are essentially equivalent; it is further justified in (Liu et al., 2018) that the metric subregularity of T_{KKT} is more advantageous than that of T^p_{KKT} in sense of analyzing the linear convergence for the ADMM and its variant.

Recall that a set-valued mapping is called a polyhedral multifunction if its graph is the union of finitely many convex polyhedra; (see, e.g., Robinson, 1975). Obviously, if both f and g are piecewise linear-quadratic functions,[2] then the desired metric subregularity of T_{KKT} (as well as T^p_{KKT}) follows immediately from (Robinson, 1980, Proposition 1). As a consequence, the local linear convergence of PADMM-FG is an immediate assertion in the following setting of full polyhedricity.

(S4) If Problem (1) satisfies the full polyhedricity, i.e., both f and g fall into the category of convex piecewise linear-quadratic functions, then

- $\{(Ax^k, By^k, \lambda^k)\}$ generated by the original ADMM converges linearly, (see, e.g., Aspelmeier et al., 2016; Liu et al., 2018);
- $\{(x^k, By^k, \lambda^k)\}$ generated by the linearized ADMM converges linearly,(see, e.g., Liu et al., 2018; Yang and Han, 2016);
- $\{(x^k, y^k, \lambda^k)\}$ generated by PADMM-FG with $\beta A^T A + G_1 \succ 0$ and $\beta B^T B + G_2 \succ 0$ converges linearly, (see, e.g., Han et al., 2017).

2. A function $\phi : \mathbb{R}^n \to \mathbb{R}$ is called piecewise linear-quadratic if dom ϕ can be represented as the union of finitely many polyhedral sets, relative to each of which $\phi(x)$ is given by an expression of the form $\frac{1}{2}\langle x, Lx\rangle + \langle \mathfrak{a}, x\rangle + \mathfrak{b}$ for some scalar $\mathfrak{b} \in \mathbb{R}$, vector $\mathfrak{a} \in \mathbb{R}^n$, and symmetric matrix $L \in \mathbb{R}^n \times \mathbb{R}^n$. ϕ is a convex piecewise linear-quadratic function if and only if $\partial\phi$ is a polyhedral multifunction.

It is notable that the desired metric subregularity above is trivially satisfied by the polyhedral case such as the LASSO model. But, the given condition seems too ambiguous to be checked for a wide range of applications to be shown soon.

1.2. Motivating Examples in Machine Learning

Below we show some concrete applications in machine learning and statistics to which the ADMM and its variants are usually applied as solution schemes, while they are not the cases for which any of the scenarios (S1)-(S4) is effective.

Example 2 *(Boyd et al., 2011, Section 11.2) Variable selection in ℓ_1 regularized logistic regression (ℓ_1 RLR):*

$$\min_x \quad \sum_j \left(\log\left(1+e^{L_j^T x}\right) - \mathbb{b}_j L_j^T x\right) + \mu\|x\|_1,$$

with L_j the $j-$th row of $L \in \mathbb{R}^{l\times m}$, $\mathbb{b}_j \in \{0,1\}$, a regularization parameter μ and a convex polyhedral regularizer $\|x\|_1$. Obviously, it can be reformulated as a special case of Problem (1) so that the ADMM and its variants can be applied:

$$\begin{aligned} \min_{x,y} \quad & \sum_j \left(\log\left(1+e^{L_j^T x}\right) - \mathbb{b}_j L_j^T x\right) + \mu\|y\|_1 \\ s.t. \quad & x = y. \end{aligned} \tag{7}$$

Obviously, for this example, in general the strong convexity does not hold in the objective. It is easy to see that assumptions S(1)-S(4) do not hold.

Example 3 *(James et al., 2013) In spite of the success of unconstrained variable selection models, e.g., LASSO and RLR, they still suffer from limited information induced by the regularizer. To address these issues, the constrained models have been proposed in order to incorporate more informative data. In particular, the penalized and constrained (PAC) regression for computing the penalized coefficient paths on high-dimensional generalized linear model:* [3]

$$\begin{aligned} \min_{x,y,z} \quad & \sum_j \left(-\log\left(L_j^T x\right) + \mathbb{b}_j L_j^T x\right) + \mu\|y\|_1 + \delta_{\mathbb{R}^{l_2}_+}(z) \\ s.t. \quad & x = y,\ \ \mathbb{C}x + z = \mathbb{d}, \end{aligned} \tag{8}$$

where $L \in \mathbb{R}^{l_1\times m}$ is the design matrix covariates, $\mathbb{b} \in \mathbb{R}^{l_1}_+$ is the response vector, $\mathbb{C} \in \mathbb{R}^{l_2\times m}$, $\mu \in \mathbb{R}_+$ and $\mathbb{d} \in \mathbb{R}^{l_2}$ are predefined matrices and vectors. It is easy to see that, in general assumptions S(1)-S(4) do not hold for this example.

The variable selection and estimation in high-dimensional regression with compositional covariates proposed in (Lin et al., 2014) perfectly fall into the constrained regression model (8). Compositional data, which consist of the proportions or percentages of a composition, appear frequently in a wide range of applications; examples include geochemical compositions

3. Hereafter, for succinctness, for the examples to be presented, we directly show the reformulations in form of Problem (1) with auxiliary variables, instead of the original models without constraints.

of rocks in geology, household patterns of expenditure in economics, species compositions of biological communities in ecology, and topic compositions of documents in machine learning; see, e.g., (James et al., 2013; Lin et al., 2014) for the details. Owing to the special nature of compositional data that the components of a composition must sum to unity, the usual unconstrained linear regression model is inappropriate. To this end, a regularized linear log-contrast model that respects the unique features of compositional data has been formulated in (Lin et al., 2014) as a constrained convex optimization in the form of (8).

Another important problem in microbiome analysis is to identify the bacterial taxa that are associated with a response, where the microbiome data are summarized as the composition of the bacterial taxa at different taxonomic levels. (Shi et al., 2016) considers regression analysis with compositional data as covariates. Inspired by the modeling procedure in (Lin et al., 2014), (Shi et al., 2016) also proposed linear models with a set of linear constraints on the regression coefficients, in order to satisfy the subcompositional coherence of the results.

Example 4 *(Yuan and Lin (2006) and Boyd et al. (2011, Sections 11.3)) For high dimensional supervised learning problems where the predictor variables were divided into different groups, for example in gene expression data these groups may be gene pathways, or factor level indicators in categorical data, rather than just sparsity in the selected variable, people would like a solution which uses only a few of the groups. In 2006, (Yuan and Lin, 2006) introduced the group LASSO in order to allow predefined groups of covariates to be selected into or out of a model together, so that all the members of a particular group are either included or not included; the problem is*

$$\begin{aligned} \min_{x,y} \quad & \frac{1}{2}\|Lx - \mathbb{b}\|^2 + \mu \sum_{J\in\mathcal{J}} \omega_J \|y_J\| \\ s.t. \quad & x = y, \end{aligned} \tag{9}$$

where $\omega_J \geq 0$, $\mathcal{J}$ is a partition of $\{1,\ldots,n\}$. This criterion exploits the non-differentiability of $\|x_J\|$ at $x_J = 0$; setting groups of coefficients to exactly 0. The sparsity of the solution is determined by the magnitude of the tuning parameter μ. If the size of each group is 1, this gives us exactly the regular LASSO solution. In general, $\sum_{J\in\mathcal{J}} \omega_J\|y_J\|$ is not a convex piecewise linear-quadratic function unless it degenerates to the ℓ_1 regularizer.

Example 5 *(Friedman et al., 2010; Zhou et al., 2010) While the group LASSO gives a sparse set of groups, if it includes a group in the model then all coefficients in the group will be nonzero. Sometimes people would like both sparsity of groups and within each group, for example if the predictors are genes people would like to identify particularly important genes in pathways of interest. Toward this end, (Friedman et al., 2010) proposed the sparse-group LASSO model:*

$$\begin{aligned} \min_{x,y} \quad & \frac{1}{2}\|Lx - \mathbb{b}\|^2 + \mu\|y\|_1 + \sum_{J\in\mathcal{J}} \omega_J \|y_J\| \\ s.t. \quad & x = y, \end{aligned} \tag{10}$$

where $\mu \geq 0$, $\omega_J \geq 0$ and $\mathcal{J}$ is a partition of $\{1,\ldots,n\}$. This model improves the group LASSO regularizer for the case where there is a possibility of within-group sparsity. Obviously, the regularizer is not convex piecewise linear-quadratic except the case $\mathcal{J} = \{\{1\},\{2\},\ldots,\{n\}\}$.

Example 6 *(Tseng, 2010) The image denoising using total variation (TV) regularization:*

$$\begin{array}{ll} \min\limits_{x,y} & \frac{1}{2}\|Lx - \mathbb{b}\|^2 + \delta_{\mathbb{B}}(y) \\ s.t. & x = y, \end{array} \tag{11}$$

where L is the adjoint of the discrete (via finite difference) gradient mapping.

The ADMM and its variants have been shown to perform linear convergence for these applications (see, e.g., Boyd et al., 2011, Sections 11.2,11.3 for ADMM applications on the ℓ_1 regularized logistic regression and the group LASSO), and as shown, existing results fail to explain the linear convergence. We are thus motivated to answer the following questions:

Besides scenarios (S1) - (S4), is it still possible that the PADMM-FG (2) converges linearly for practical applications; and if yes, how can we discern the linear convergence?

We answer these questions affirmatively, and show particularly how to discern the linear convergence of PADMM-FG (2) for a wide range of applications that are important in statistical learning.

1.3. Setting for Discussion

We present the assumptions under which our analysis will be carried on. Throughout, to avoid triviality, the following nonemptiness assumption is required.

Assumption 1.1 (Standing assumption) *The optimal KKT solution set of Problem (1) is nonempty.*

Instead of the general case of (1) without any structure, and as motivated by various applications including those listed before, we focus on some structured cases of Problem (1) and make the following assumptions regarding the structure of Problem (1).

Assumption 1.2 (Structured assumption of f) *A convex function $f : \mathbb{R}^n \to (-\infty, \infty]$ is said to satisfy the structured assumption if f is a function in form of*

$$f(x) = h(Lx) + \langle q, x\rangle,$$

where L is some $m \times n$ matrix, q is some vector in $\mathbb{R}^n$, and $h : \mathbb{R}^m \to (-\infty, \infty]$ is a convex proper lsc function with the following properties:

(i) h is essentially locally strongly convex, i.e., for any compact and convex subset $\mathbb{K} \subset$ dom ∂h, h is strongly convex on $\mathbb{K}$;

(ii) h is essentially differentiable, i.e., int(dom h) *is nonempty, h is differentiable on* int(dom h)*, and $\lim_{k\to\infty} |\nabla h(\alpha_k)| = \infty$ for any sequence $\{\alpha_k\}_{k=1}^{\infty}$ converging to a boundary point of* int(dom h)*, and ∇h is locally Lipschitz continuous on* int(dom h)*;*

(iii) range $(L) \cap$ int(dom h) $\neq \varnothing$*, where* range $(L) \subseteq \mathbb{R}^m$ *denotes the range of matrix L.*

Some commonly used loss functions in statistical learning such as linear regression, logistic regression and likelihood estimation under Poisson noise all satisfy Assumption 1.2. We summarize these cases in Table 1, where $b_1 \in \mathbb{R}^m$, $b_2 \in \{0,1\}^m$ and $b_3 \in \mathbb{R}^m_+$ are parameters. Indeed, Part (iii) in Assumption 1.2 fulfills if dom h is an open set and dom $h \neq \varnothing$. It is easy to check that, all the smooth parts of the objective functions in Examples 2-6 satisfy Assumption 1.2 equipped with the scenarios in Table 1 automatically.

Loss function	Linear regression	Logistic regression	Likelihood estimation
$h(y)$	$\frac{1}{2}\|y - b_1\|$	$\sum_{i=1}^m \log(1+e^{y_i}) - \langle b_2, y\rangle$	$-\sum_{i=1}^m \log(y_i) + \langle b_3, y\rangle$

Table 1: Some commonly used loss functions h

Assumption 1.3 (Structured polyhedricity assumption) *Problem (1) is said to satisfy the structured polyhedricity assumption if f meets Assumption 1.2 and g is convex piecewise linear-quadratic function.*

Assumption 1.4 (Structured subregularity assumption) *Problem (1) is said to satisfy the structured subregularity assumption at a KKT point $(\bar{x}, \bar{y}, \bar{\lambda})$ if f meets Assumption 1.2, $\partial(g^*(B^T\lambda))$ is calm at the reference point $(\bar{\lambda}, -A\bar{x})$ and*

$$\hat{\Omega}_x(p) := \{x \mid p = Lx - L\bar{x},\ 0 \in \partial(g^*(B^T\bar{\lambda})) - Ax\}$$

is calm at $(0, \bar{x})$.

We next make some comments on Assumptions 1.3 and 1.4.

Remark 3 *It is worthwhile mentioning that the structured polyhedricity assumption, which is in general stronger than the structured subregularity assumption, is satisfied by some applications such as the aforementioned RLR model (7) and PAC model (8). Moreover, the structured subregularity assumption is satisfied at any KKT point $(\bar{x}, \bar{y}, \bar{\lambda})$*

- *if g represents the $\ell_{1,q}$-norm regularizer with $q \in [1,2]$;*
- *if g represents the sparse-group LASSO regularizer;*
- *if g represents the indicator function of a ball constraint, i.e., $g = \delta_{\mathbb{B}}(\cdot)$ and $B^T\bar{\lambda} \neq 0$.*

More details will be presented in subsection 3.4.

1.4. Contributions in Discerning Local Nonergodic Linear Convergence of ADMM

As mentioned, our first purpose is to discuss the local nonergodic linear convergence of the PADMM-FG (2) through the lens of variational analysis. We next present the first contribution in establishing the local linear convergence of original ADMM, linearized ADMM and

the general PADMM-FG separately. To measure the optimality of an iterate for Problem (1), we define two indicators: the objective function value

$$\mathrm{Val}(x, y) = f(x) + g(y)$$

and the feasibility of constraints

$$\mathrm{Fea}(x, y) = \|Ax + By - b\|.$$

For notation simplicity, hereafter, for a generated iterate (x^k, y^k, λ^k), we denote the KKT residue $\mathrm{Res}(x^k, y^k, \lambda^k)$ defined in (5) by Res^k, the objective function value $\mathrm{Val}(x^k, y^k)$ by Val^k, and the feasibility of constraints $\mathrm{Fea}(x^k, y^k)$ by Fea^k, respectively. Below we present the main results to be obtained, and further summarize them in Table 2.

- **Discerning the local nonergodic linear convergence of original ADMM**. For the original ADMM where $G_1 = 0$, $G_2 = 0$ and $\gamma = 1$ in (2), if Problem (1) satisfies the structured polyhedricity assumption, then we derive the linear convergence in the following senses:
  - the KKT residues sequence $\{\mathrm{Res}^k\}$ converges linearly;
  - the sequence of objective function value and constraint feasibility pairs $\{\mathrm{Val}^k, \mathrm{Fea}^k\}$ converges linearly;
  - the sequence $\{\lambda^k\}$ converges linearly.
- **Discerning the local nonergodic linear convergence of linearized ADMM**. For the linearized ADMM where $G_1 = rI - \beta A^T A$ with $r > \beta\|A^T A\|$, $G_2 = 0$ and $\gamma = 1$ in (2), if one of the following assumptions is satisfied:
  1. Problem (1) satisfies the structured polyhedricity assumption;
  2. Problem (1) satisfies the structured subregularity assumption and A is of full row rank;

  then we derive the linear convergence in the following senses:
  - the KKT residues sequence $\{\mathrm{Res}^k\}$ converges linearly;
  - the sequence of objective function value and constraint feasibility pairs $\{\mathrm{Val}^k, \mathrm{Fea}^k\}$ converges linearly;
  - the sequences $\{(x^k, \lambda^k)\}$ converges linearly.
- **Discerning the local nonergodic linear convergence of the general PADMM-FG**. For the general PADMM-FG with $\beta A^T A + G_1 \succ 0$ and $\beta B^T B + G_2 \succ 0$, if one of the following assumptions is satisfied:
  1. Problem (1) satisfies the structured polyhedricity assumption;
  2. Problem (1) satisfies the structured subregularity assumption, A is of full row rank and B is of full column rank;

  then we derive the local nonergodic linear convergence in the following senses:

- the KKT residues sequence $\{\text{Res}^k\}$ converges linearly;
- the sequence of objective function value and constraint feasibility pairs $\{\text{Val}^k, \text{Fea}^k\}$ converges linearly;
- the sequences $\{(x^k, y^k, \lambda^k)\}$ converges linearly.

Algorithmic setting	Regularity beyond convexity				Linear convergence
	Structured Polyhedricity	Structured Subregularity	Full row rank of A	Full column rank of B	
$\gamma = 1, G_1 = 0, G_2 = 0$	✓	-	-	-	$\{\lambda^k; \text{Res}^k; \text{Val}^k, \text{Fea}^k\}$
$\gamma = 1$ with $r > \beta\|A^TA\|$, $G_2 = 0$, $G_1 = rI - \beta A^TA$	✓	-	-	-	$\{(x^k, \lambda^k); \text{Res}^k; \text{Val}^k, \text{Fea}^k\}$
	-	✓	✓	-	$\{(x^k, \lambda^k); \text{Res}^k; \text{Val}^k, \text{Fea}^k\}$
$\gamma \in (0, \frac{1+\sqrt{5}}{2})$, $\beta A^TA + G_1 \succ 0, \beta B^TB + G_2 \succ 0$	✓	-	-	-	$\{(x^k, y^k, \lambda^k); \text{Res}^k; \text{Val}^k, \text{Fea}^k\}$
	-	✓	✓	✓	$\{(x^k, y^k, \lambda^k); \text{Res}^k; \text{Val}^k, \text{Fea}^k\}$

Table 2: Summary of local nonergodic linear convergence for the PADMM-FG (2)

1.5. Contributions to Machine Learning Applications

As mentioned, our second purpose is to discern the linear convergence behaviors as well as the exact convergence rates of the PADMM-FG (2) for applications. Our results in discerning local linear convergence of ADMM variants are established under two key assumptions, i.e., the structured polyhericity assumption and the structured subregularity assumption. These two conditions are actually application-driven, although they seem technical. We shall develop new techniques that can be used to verify the key assumptions for an array of concrete models in machine learning and statistics, and hence establish ADMM linear convergence together with our first contribution. Our second contribution is then summarized as ADMM linear convergence in concrete machine learning applications, with a particular interest in the mentioned ℓ_1 RLR, PAC, group LASSO and sparse-group LASSO models, which can not be covered by Scenarios (S1)-(S4).

Assuming the structured assumption of f in Assumption 1.2, we then present our results systematically according to different application-driven scenarios of g.

Scenario 1. When ∂g is a polyhedral multifunction. This setting covers the following sparse learning regularizers for feature selection in high-dimensional data analysis.

- $g(x)$ represents the ℓ_1 regularizer which performs as the sparsity-inducing norms to force the coefficients of non-important features to be zero.

In the high-dimensional data, the highly correlated features widely exist. However, the ℓ_1 regularized sparse learning methods tend to arbitrarily select only one of them. Consequently, estimation can be unstable, and the resultant model difficult to interpret. The grouping of features, on the other hand, is highly beneficial in learning with high-dimensional data. It reduces the variance in the estimation and improves the stability of feature selection, leading to improved generalization (see, e.g., Zhong and Kwok, 2012). We list some representative groups-keeping regularizers which are also covered in this category:

- the elastic net regularizer (see, e.g., Zou and Hastie, 2005), encourages highly correlated covariates to have similar regression coefficients.

- the fused LASSO regularizer (see, e.g., Tibshirani et al., 2005), directly enforces the successive feature coefficients to be similar by the regularizer, if the features are ordered in some meaningful way.
- the octagonal selection and clustering algorithm for regression (OSCAR) (see, e.g., Bondell and Reich, 2008), uses the pairwise ℓ_∞ norm to encourage the equality of coefficients for highly correlated features.

The definitions of the listed polyhedral convex regularizers are summarized in Table 3, where μ, μ_1 and μ_2 are given nonnegative parameters.

Regularizers	ℓ_1-norm	elastic net	fused LASSO	OSCAR
$g(x)$	$\mu\|x\|_1$	$\mu_1\|x\|_1 + \mu_2\|x\|^2$	$\mu_1\|x\|_1 + \mu_2 \sum_i \|x_i - x_{i+1}\|$	$\mu_1\|x\|_1 + \mu_2 \sum_{i<j} \max\{\|x_i\|, \|x_j\|\}$

Table 3: Polyhedral convex regularizers

Scenario 2. If ∂g is metrically subregular (not necessarily a polyhedral multifunction). This setting covers two classes of important groups-keeping regularizers, where, different from those mentioned polyhedral groups-keeping regularizers, some prior information about the group structure of the underlying solution is assumed to be known in advance:

- the $\ell_{1,q}$-norm regularizer with $q \in [1, 2]$ (see, e.g., Fornasier and Rauhut, 2008; Kowalski, 2009). In particular, $g(y) = \sum_{J\in\mathcal{J}} \omega_J \|y_J\|_q$, where $\omega_J \geq 0$, $\mathcal{J}$ is a partition of $\{1, \ldots, n\}$ and $\|\cdot\|_q$ denotes the ℓ_q-norm, i.e., $\|x\|_q := \Big(\sum_{i=1}^n |x_i|^q\Big)^{\frac{1}{q}}$. When $q = 2$, the $\ell_{1,q}$-norm reduces to the group LASSO regularizer.
- the sparse-group LASSO regularizer, has been widely applied to different areas, such as text processing, bioinformatics, signal interpretation, and object tracking.

Motivated by a substantial number of practical applications in statistics and machine learning, we shall achieve our second contribution within the analysis framework

"structured assumption of f" + "application-driven scenarios of g".

In Table 4, we further specify our second contribution and list the conditions that can be used to discern the linear convergence of various specific cases of the PADMM-FG (2). In this table, "oADMM", "lADMM" and "pADMM" stand for the original ADMM, the linearized ADMM and the general PADMM-FG (2) with the conditions $\beta A^T A + G_1 \succ 0$ and $\beta B^T B + G_2 \succ 0$, respectively. This table serves as a "dictionary" for looking up to the linear convergence when the ADMM and its variants are employed to solve a number of popular applications.

Remark 4 *In Table 4, we mark with a thick line box the full polyhedricity case where* $f(x) = \frac{1}{2}\|Lx - \mathbb{b}\|^2$ *and* g *is convex piecewise linear-quadratic. For this full polyhedricity case, the linear convergence of various cases of the PADMM-FG (2) has been studied in the literature, (see, e.g., Aspelmeier et al., 2016; Han et al., 2017; Liu et al., 2018; Yang and Han, 2016). For other applications in this table, it seems to be the first time to obtain the linear convergence of various cases of the PADMM-FG (2).*

f(x) \ g(y)	$\mu\|y\\|_1$	condition	$\mu_1\|y\\|_1+\mu_2\|y\\|^2$	condition	$\mu_1\|y\\|_1+\mu_2\sum_i\|y_i-y_{i+1}\|$	condition	$\mu_1\|y\\|_1+\mu_2\sum_{i<j}\max\{\|y_i\|,\|y_j\|\}$	condition	$\sum_{J\in\mathcal{J}}\omega_J\|y_J\\|_q$ $(1\le q\le 2)$	condition	$\mu\|y\\|_1+\sum_{J\in\mathcal{J}}\omega_J\|y_J\\|$	condition
$\frac{1}{2}\|Lx-\mathbb{b}\|^2$		condition		condition		condition		condition		condition		condition
	oADMM $\{\lambda^k\}$	-	oADMM $\{\lambda^k\}$	-	oADMM $\{\lambda^k\}$	-	oADMM $\{\lambda^k\}$	-	oADMM $\{Ax^k,y^k,\lambda^k\}$	full row rank A; full column rank B	oADMM $\{Ax^k,y^k,\lambda^k\}$	full row rank A; full column rank B
	lADMM $\{(x^k,\lambda^k)\}$	-	lADMM $\{(x^k,\lambda^k)\}$	-	lADMM $\{(x^k,\lambda^k)\}$	-	lADMM $\{(x^k,\lambda^k)\}$	-	lADMM $\{(x^k,\lambda^k)\}$	full row rank A	lADMM $\{(x^k,\lambda^k)\}$	full row rank A
	pADMM $\{(x^k,y^k,\lambda^k)\}$	-	pADMM $\{(x^k,y^k,\lambda^k)\}$	-	pADMM $\{(x^k,y^k,\lambda^k)\}$	-	pADMM $\{(x^k,y^k,\lambda^k)\}$	-	pADMM $\{(x^k,y^k,\lambda^k)\}$	full row rank A; full column rank B	pADMM $\{(x^k,y^k,\lambda^k)\}$	full row rank A; full column rank B
$\sum_j\left(\log\left(1+e^{L_j^Tx}\right)-\mathbb{b}_jL_j^Tx\right)$		condition		condition		condition		condition		condition		condition
	oADMM $\{\lambda^k\}$	-	oADMM $\{\lambda^k\}$	-	oADMM $\{\lambda^k\}$	-	oADMM $\{\lambda^k\}$	-	oADMM $\{Ax^k,y^k,\lambda^k\}$	full row rank A; full column rank B	oADMM $\{Ax^k,y^k,\lambda^k\}$	full row rank A; full column rank B
	lADMM $\{(x^k,\lambda^k)\}$	-	lADMM $\{(x^k,\lambda^k)\}$	-	lADMM $\{(x^k,\lambda^k)\}$	-	lADMM $\{(x^k,\lambda^k)\}$	-	lADMM $\{(x^k,\lambda^k)\}$	full row rank A	lADMM $\{(x^k,\lambda^k)\}$	full row rank A
	pADMM $\{(x^k,y^k,\lambda^k)\}$	-	pADMM $\{(x^k,y^k,\lambda^k)\}$	-	pADMM $\{(x^k,y^k,\lambda^k)\}$	-	pADMM $\{(x^k,y^k,\lambda^k)\}$	-	pADMM $\{(x^k,y^k,\lambda^k)\}$	full row rank A; full column rank B	pADMM $\{(x^k,y^k,\lambda^k)\}$	full row rank A; full column rank B
$\sum_j\left(-\log\left(L_j^Tx\right)+\mathbb{b}_jL_j^Tx\right)$		condition		condition		condition		condition		condition		condition
	oADMM $\{\lambda^k\}$	-	oADMM $\{\lambda^k\}$	-	oADMM $\{\lambda^k\}$	-	oADMM $\{\lambda^k\}$	-	oADMM $\{Ax^k,y^k,\lambda^k\}$	full row rank A; full column rank B	oADMM $\{Ax^k,y^k,\lambda^k\}$	full row rank A; full column rank B
	lADMM $\{(x^k,\lambda^k)\}$	-	lADMM $\{(x^k,\lambda^k)\}$	-	lADMM $\{(x^k,\lambda^k)\}$	-	lADMM $\{(x^k,\lambda^k)\}$	-	lADMM $\{(x^k,\lambda^k)\}$	full row rank A	lADMM $\{(x^k,\lambda^k)\}$	full row rank A
	pADMM $\{(x^k,y^k,\lambda^k)\}$	-	pADMM $\{(x^k,y^k,\lambda^k)\}$	-	pADMM $\{(x^k,y^k,\lambda^k)\}$	-	pADMM $\{(x^k,y^k,\lambda^k)\}$	-	pADMM $\{(x^k,y^k,\lambda^k)\}$	full row rank A; full column rank B	pADMM $\{(x^k,y^k,\lambda^k)\}$	full row rank A; full column rank B

Table 4: Summary of applications with guaranteed local linear convergence for ADMM and its variants

a. For all cases listed in this table, the KKT residues sequence $\{\text{Res}^k\}$, the objective values sequence $\{\text{Val}^k\}$ and the constraint feasibility sequence $\{\text{Fea}^k\}$ converge linearly.

Remark 5 *Even for the full polyhedricity case, in (Liu et al., 2018; Yang and Han, 2016), the linear convergence of* $\{(Ax^k, By^k, \lambda^k)\}$ *and* $\{(x^k, By^k, \lambda^k)\}$ *is proved under the assumption of the convergence of the sequence* $\{(x^k, y^k, \lambda^k)\}$ *generated by the original ADMM and the linearized ADMM, respectively. In our analysis, instead of assuming the convergence of the sequence* $\{(x^k, y^k, \lambda^k)\}$*, we only delineate the sequences that are known to be convergent. In particular, for the full polyhedricity case, we prove the linear convergence of the sequence* $\{\lambda^k\}$ *generated by the original ADMM and the linear convergence of the sequence* $\{(x^k, \lambda^k)\}$ *generated by the linearized ADMM. It is trivial to deduce that, under similar convergence assumptions on the sequence* $\{(x^k, y^k, \lambda^k)\}$ *as those in (Liu et al., 2018; Yang and Han, 2016), the linear convergence of* $\{(Ax^k, By^k, \lambda^k)\}$ *and* $\{(x^k, By^k, \lambda^k)\}$ *can also be obtained for the original ADMM and the linearized ADMM, respectively.*

Remark 6 *In (Han et al., 2017), it is noticed that the metric subregularity of the mapping* T^p_{KKT} *at a KKT point can be used for the sake of proving the linear convergence of the PADMM-FG (2). It is known that the metric subregularity condition is indeed a pointwise condition. Therefore, in general, it is too ambiguous to be checked when the reference point is unknown. What is more meaningful and challenging is finding out appropriate methodologies that can verify the required metric subregularity so as to discern the linear convergence for various concrete applications. This issue is out of the scope of (Han et al., 2017). Through the lens of variational analysis, we shall show that the metric subregularity condition is not just conceptual, but also verifiable for a wide range of applications arising in statistical learning. Hence the empirically observed linear convergence of a number of algorithms is tightly proved with rigorous mathematics; and the understanding of linear convergence of ADMM and its variants is significantly enhanced.*

1.6. Insights

Although the original ADMM and the linearized ADMM are special cases of the general PADMM-FG (2), we conduct linear convergence analysis separately as shown hierarchically in the last subsection, rather than just for the general PADMM-FG (2) as a whole. Generically speaking, it is because treating all variants in the most general form of PADMM-FG (2) will result in the loss of some special properties owned by the special cases of the original ADMM and the linearized ADMM. Indeed, we shall show that individual treatments on the original ADMM and the linearized ADMM enable us to take advantage of their special algorithmic structures more effectively and thus to derive some specific properties. This is a striking feature of our study that leads to some new results for the original ADMM and the linearized ADMM.

We understand that, because the sequence $\{(x^k, y^k, \lambda^k)\}$ generated by the PADMM-FG (2) with $\beta A^T A + G_1 \succ 0, \beta B^T B + G_2 \succ 0$ converges to a KKT point $(\bar{x}, \bar{y}, \bar{\lambda})$, as long as the KKT mapping T_{KKT} is metrically subregular at $(\bar{x}, \bar{y}, \bar{\lambda}, 0))$, the convergence rate of $\{(x^k, y^k, \lambda^k)\}$ is indeed linear. This can be seen in Proposition 57. On the other hand, instead of $\{(x^k, y^k, \lambda^k)\}$, the original ADMM generates the convergent sequence $\{(Ax^k, By^k, \lambda^k)\}$ while the linearized ADMM generates $\{(x^k, By^k, \lambda^k)\}$. Naturally, we focus on the convergence rate analysis in terms of the sequences $\{\lambda^k\}$ for the original ADMM, and $\{(x^k, \lambda^k)\}$ for the linearized ADMM, respectively. In our recent work (Wang et al., 2018), we introduce the perturbation analysis technique for analyzing the convergence of an

algorithm, by appropriately constructing an iteration-tailored perturbed solution set-valued map and defining a perturbing parameter as the difference of two consecutive iterates of the algorithm under investigation. Particularly, in this paper we adopt this technique for the sequence $\{\lambda^k\}$ of the original ADMM, $\{(x^k, \lambda^k)\}$ of the linearized ADMM and $\{(x^k, y^k, \lambda^k)\}$ of the general PADMM-FG (2), respectively, and accordingly induce different perturbed solution set-valued maps. More details of the difference between those perturbed solution set-valued maps will be delineated in Sections 2 and 3.

- **Insight into algorithmic structure**. Therefore, our first insight is that the original ADMM, the linearized ADMM and the general PADMM-FG (2) should be treated independently. Then, our main purpose becomes verifying calmness/metric subregularity of the set-valued maps induced by the perturbation analysis technique which guarantee the desired linear convergence. Our analysis is conducted case by case using different techniques, because of the significant differences of the original ADMM, the linearized ADMM and the general PADMM-FG (2). This analyzing framework seems novel as it is quite distinct from those in the literature regarding ADMM linear convergence.

In addition to the need of considering the algorithmic structure, it is commonly known that the model's structure should be fully considered when studying the convergence of a particular algorithm applied to solve the model under investigation. Our analysis is also based on the understanding that the verification of required subregularity conditions should be conducted in accordance with the model's special structure. Indeed, it is well-known that the verification of subregularity conditions for practical application problems is usually a challenging task. In the literature, there are various criteria proposed in the generic context by following standard variational analysis, for ensuring the metric subregularity, (see, e.g., Gfrerer, 2011, 2013; Gfrerer and Ye, 2017; Guo et al., 2013; Henrion et al., 2002; Henrion and Outrata, 2005; Ye and Ye, 1997; Ye and Zhang, 2013). But it seems there is very little discussion on how to define some model-tailored subregularity conditions that can inherently make use of the model's structures for the study of linear convergence of the ADMM and its variants. This fact limits the application of various existing work, including (Aspelmeier et al., 2016; Han et al., 2017; Liang et al., 2017; Valkonen, 2014, 2017), to the theoretical explanation of the linear convergence of the ADMM and its variants for some of the mentioned models, see, e.g., the motivating examples 2-5.

- **Insight into model structure**. Therefore, our second insight is that the model's structure should be well exploited to initiate new criteria for verifying different types of metric subregularity conditions that can both ensure the linear convergence of the ADMM and its variants and be easily verified by an array of concrete machine learning applications including those listed in Table 4. The new criterion differs significantly from those discussed in standard variational analysis, which seems to be novel in the literature.

Motivated by the mentioned insights, we employ perturbation analysis techniques to identify appropriate forms of the metric subregularity for different cases of the PADMM-FG (2), and then penetrate the model's structures to re-characterize the desired subregularity

conditions step by step to find more verifiable characterizations. The employment of calm intersection theorem on the re-characterized perturbed solution map allows us to adpot Robinson's celebrated result (Robinson, 1981, Proposition 1) and hence calculate the error bound modulus. Through this roadmap, we uncover the fact that the required calmness conditions can be indeed verified and the linear convergence of the ADMM and its variants can be discerned by a number of applications including those shown in Table 4.

1.7. Outline

The remaining part of the paper is organized as following. In Section 2, we focus on discerning the linear convergence of the original ADMM. In particular, we derive the linear convergence of the DRSM by studying the dual problem of Problem (1) and then convert the result to the linear convergence of the original ADMM. In Section 3, we study the linear convergence of the linearized ADMM for various cases. Particularly, by examining the dual problem of (1), we show how to discern the linear convergence of the PDHG and then convert the result to the linear convergence of the linearized ADMM. Then, we discuss the general PADMM-FG (2) in Section 4 and give some concluding remarks in Section 5.

1.8. Symbols and Notations

The index of notations as well as the corresponding descriptions that will be used in this paper are listed in Appendix O.

2. Linear Convergence of the Original ADMM

In this section, we shall show that, under Assumption 1.3, i.e., the structured polyhedricity assumption, the original ADMM converges linearly in sense of the sequences $\{\lambda^k\}$, $\{\mathrm{Res}^k\}$ and $\{\mathrm{Val}^k, \mathrm{Fea}^k\}$.

2.1. Roadmap of Analysis

We first recall the well-known equivalence between the original ADMM and the DRSM. This relationship indicates that we just need to show the linear convergence for one of these two methods. We first concentrate on the linear convergence of the DRSM. As illustrated in Remark 13, using the perturbation analysis techniques, we introduce the set-valued mappings $\mathcal{T}_1$ defined in (13) and $\mathcal{T}_2$ defined in (14) that are tailored for the iterative scheme of the DRSM. In Proposition 12 and Theorem 14, under the calmness of $\mathcal{T}_1$ and $\mathcal{T}_2$, we derive the linear convergence of the DRSM. Furthermore, to verify the calmness of $\mathcal{T}_1$, one subtle step is probing the characterization of the calmness of $\mathcal{T}_1$ in terms of the calmness of the set-valued map Γ_{DR} defined in (17). Taking full advantage of Assumption 1.2, we notice that the structure of Γ_{DR} helps us investigate the calmness of $\mathcal{T}_1$ and it is easily verified when g is a convex piecewise linear-quadratic function, according to Robinson's celebrated result (Robinson, 1981, Proposition 1). Similarly, we re-characterize the calmness of $\mathcal{T}_2$ with the calmness of the structured $\tilde{\Gamma}_{DR}$ defined in (21). With the re-characterization in terms of Γ_{DR} and $\tilde{\Gamma}_{DR}$, it then turns out to be easy to verify the calmness of $\mathcal{T}_1$ and $\mathcal{T}_2$ under the structured polyhedricity assumption.
To present our analysis more clearly, let us show the roadmap of this section in Figure 1.

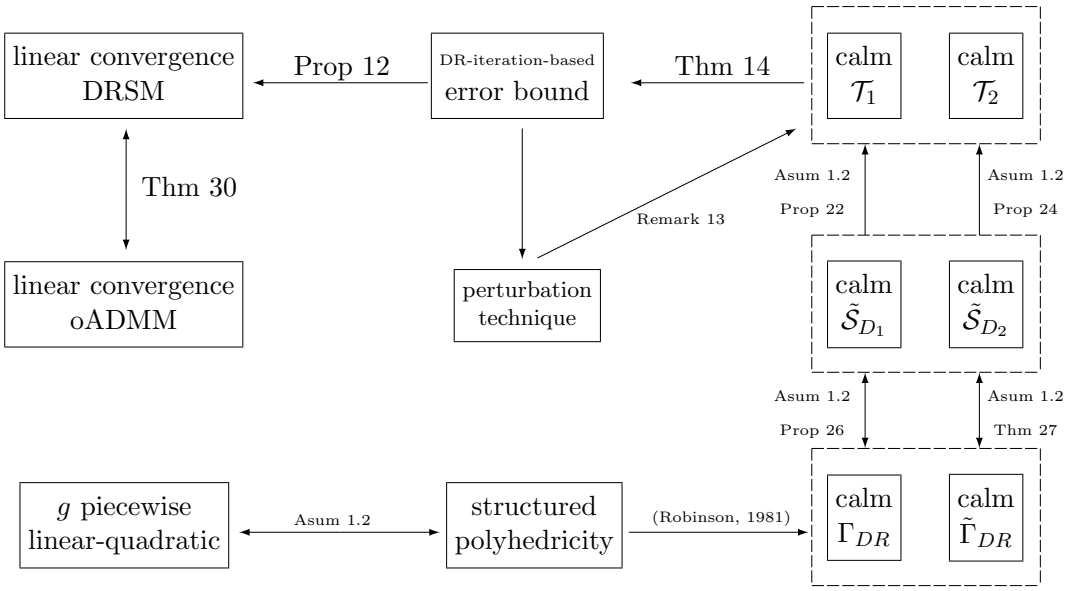

Figure 1: Roadmap to study linear convergence of the original ADMM

Remark 7 *In (Aspelmeier et al., 2016), the linear convergence of DRSM is studied under the metric subregularity of the DR operator T_{DR} (see subsection 2.3.1 for the definition). Hence, the linear convergence rate of the original ADMM for the full polyhedricity case (S4) can be derived as well. In our analysis, by noticing that the DR operator is a composite of two operators, we define the algorithm-tailored perturbed solution set-valued maps* (13) *and* (14)*, through which we can discern the linear convergence of the original ADMM for a much broader spectrum of applications beyond the full polyhedricity case, e.g., the RLR model (7) and the PAC model (8). On the other hand, the linear convergence of DRSM is investigated under the strong monotonicity assumption in (Giselsson and Boyd, 2016). Thus the linear convergence rate of the original ADMM can be recovered under some strong convexity conditions together with some full rank assumptions of the coefficient matrix.*

2.2. Original ADMM on Primal Problem Is Equivalent to DRSM on Dual Problem

It is clear that the dual of Problem (1) can be written as

$$(D) \qquad \min_{\lambda} \ -b^T\lambda + f^*(A^T\lambda) + g^*(B^T\lambda),$$

where f^* and g^* denote the conjugates of the convex functions f and g, respectively. Let

$$\phi_1(\lambda) = f^*(A^T\lambda) - b^T\lambda, \quad \phi_2(\lambda) = g^*(B^T\lambda),$$

the dual problem (D) can be represented as the following inclusion problem:

$$0 \in \partial\phi_1(\lambda) + \partial\phi_2(\lambda).$$

As analyzed in (Gabay, 1983), applying the original ADMM to the primal Problem (1) is equivalent to applying the DRSM to its dual problem. We summarize some prerequisites in the following proposition for further analysis.

Proposition 8 *Let $\{(x^k, y^k, \lambda^k)\}$ be the sequence generated by the original ADMM. Define $z^k := \lambda^k + \beta B y^k$, $u^k := \lambda^k$ and $v^k := \lambda^k - \beta(Ax^{k+1} + By^k - b)$. Then, $\{(u^k, v^k, z^k)\}$ coincides with the sequence generated by the DRSM applied to the dual problem (D), with the following details:*

$$\begin{cases} u^k = (I + \beta\partial\phi_2)^{-1}(z^k), \\ v^k = (I + \beta\partial\phi_1)^{-1}(2u^k - z^k), \\ z^{k+1} = z^k - u^k + v^k. \end{cases}$$

Proof See Appendix A. ■

2.3. Linear Convergence of DRSM

Because of the equivalence shown in the preceding subsection, we just need to discuss the linear convergence of DRSM for solving the dual problem (D) to derive the linear convergence of the original ADMM for Problem (1).

2.3.1. Linear Convergence of DRSM under DR-iteration Based Error Bound

Recall that the iterative scheme of the DRSM applied to the dual problem (D) reads as

$$\begin{cases} u^k = (I + \beta\partial\phi_2)^{-1}z^k, \\ v^k = (I + \beta\partial\phi_1)^{-1}(2u^k - z^k), \\ z^{k+1} = z^k - u^k + v^k. \end{cases}$$

Let

$$T_{DR} := \frac{1}{2}I + \frac{1}{2}(2(I + \beta\partial\phi_1)^{-1} - I)(2(I + \beta\partial\phi_2)^{-1} - I)$$

represent the DR operator, i.e., $z^{k+1} = T_{DR}z^k$. As shown in (He and Yuan, 2012a, 2015), the sequence $\{z^k\}$ converges to a certain point in $Fix(T_{DR})$, where $Fix(T_{DR})$ represents the fixed point set of T_{DR}, i.e., $Fix(T_{DR}) := \{z \mid z = T_{DR}(z)\}$. Without loss of generality, we focus on the case where $\beta = 1$, because the sequence generated by applying the DRSM with $\beta = c > 0$ to $0 \in \partial\phi_1(\lambda) + \partial\phi_2(\lambda)$ is the same as that of the DRSM with $\beta = 1$ to $0 \in \partial(c\phi_1(\lambda)) + \partial(c\phi_2(\lambda))$. Before we present the linear convergence of DRSM, we recall some preliminary results. The following proposition that can be found in (Bauschke and Moursi, 2017) as well.

Proposition 9 *Define that*

$$Z := (\partial\phi_1 + \partial\phi_2)^{-1}(0), \quad W := (\partial(\phi_1^* \circ -Id) + \partial\phi_2^*)^{-1}(0).$$

The following relationships hold:

(1) $Z = \mathrm{prox}_{\phi_2}(Fix(T_{DR}))$, *and* $W = \mathrm{prox}_{\phi_2{}^*}(Fix(T_{DR}))$.

(2) $Fix(T_{DR}) = Z + W$.

Proposition 10 *(Bauschke and Combettes (2011, Theorem 25.6) and Bauschke and Moursi (2017, Theorem 2.7)) Let $\{z^k\}$ be the sequence generated by the DRSM.*

(1) The sequence $\{z^k\}$ converges to some point $\bar{z}$ in $Fix(T_{DR})$.

(2) For any $z^* \in Fix(T_{DR})$, there holds the following estimation

$$\begin{aligned}&\|\mathrm{prox}_{\phi_2}(z^{k+1}) - \mathrm{prox}_{\phi_2}(z^*)\|^2 + \|\mathrm{prox}_{\phi_2{}^*}(z^{k+1}) - \mathrm{prox}_{\phi_2{}^*}(z^*)\|^2 \\ \leq &\|\mathrm{prox}_{\phi_2}(z^k) - \mathrm{prox}_{\phi_2}(z^*)\|^2 + \|\mathrm{prox}_{\phi_2{}^*}(z^k) - \mathrm{prox}_{\phi_2{}^*}(z^*)\|^2 - \|z^{k+1} - z^k\|^2, \quad \forall k.\end{aligned} \tag{12}$$

Definition 11 (DR-iteration-based error bound) *Let the sequence $\{z^k\}$ be generated by the DRSM and $\bar{z} \in Fix(T_{DR})$ be an accumulation point of $\{z^k\}$. We say that the DR-iteration-based error bound holds at $\bar{z}$ if there exist $\epsilon, \kappa > 0$ such that*

$$\mathrm{dist}\left(\mathrm{prox}_{\phi_2}(z^k), Z\right) + \mathrm{dist}\left(\mathrm{prox}_{\phi_2{}^*}(z^k), W\right) \leq \kappa \left\|z^{k+1} - z^k\right\|,$$

for all k such that $z^k \in \mathbb{B}(\bar{z}, \epsilon)$.

Then, it is easy to prove the linear convergence of DRSM under the just-defined error bound condition. Let $\{z^k\}$ be the sequence generated by the DRSM applied to the dual problem (D), according to Proposition 10, $\{z^k\}$ converges to some point $\bar{z} \in Fix(T_{DR})$.

Proposition 12 (Linear convergence of DRSM under DR-iteration-based error bound) *Let the sequence $\{z^k\}$ be generated by the DRSM, and $\bar{z} \in Fix(T_{DR})$ be the limit point of $\{z^k\}$. Suppose that the DR-iteration-based error bound holds at $\bar{z}$. The sequence $\{z^k\}$ converges to $\bar{z}$ linearly, i.e., there exist $k_0(\epsilon) > 0$ and $0 < \rho = \sqrt{1 - \frac{1}{\kappa^2}} < 1$ such that, for all $k \geq k_0(\epsilon)$, it holds*

$$\begin{aligned}&\mathrm{dist}\left(\mathrm{prox}_{\phi_2}(z^{k+1}), Z\right) + \mathrm{dist}\left(\mathrm{prox}_{\phi_2{}^*}(z^{k+1}), W\right) \\ \leq\ &\rho\left(\mathrm{dist}\left(\mathrm{prox}_{\phi_2}(z^k), Z\right) + \mathrm{dist}\left(\mathrm{prox}_{\phi_2{}^*}(z^k), W\right)\right),\end{aligned}$$

and thus there exists $C_0 > 0$ such that

$$\mathrm{dist}\left(\mathrm{prox}_{\phi_2}(z^k), Z\right) + \mathrm{dist}\left(\mathrm{prox}_{\phi_2{}^*}(z^k), W\right) \leq C_0\rho^k, \quad \forall k \geq k_0(\epsilon),$$

$$\|z^{k+1} - z^k\| \leq C_0\rho^k, \quad \forall k \geq k_0(\epsilon),$$

and

$$\mathrm{dist}\left(z^k, Fix(T_{DR})\right) \leq C_0\rho^k, \quad \forall k \geq k_0(\epsilon).$$

Proof By Propositions 9 and 10 and the closedness of Z and W, we have

$$\begin{aligned}&\mathrm{dist}\left(\mathrm{prox}_{\phi_2}(z^{k+1}), Z\right)^2 + \mathrm{dist}\left(\mathrm{prox}_{\phi_2{}^*}(z^{k+1}), W\right)^2 \\ \leq &\mathrm{dist}\left(\mathrm{prox}_{\phi_2}(z^k), Z\right)^2 + \mathrm{dist}\left(\mathrm{prox}_{\phi_2{}^*}(z^k), W\right)^2 - \left\|z^{k+1} - z^k\right\|^2.\end{aligned}$$

Then, since the DR-iteration-based error bound holds at $\bar{z}$, there exist $\epsilon, \kappa > 0$ such that

$$\begin{aligned}&\operatorname{dist}\left(\operatorname{prox}_{\phi_2}(z^{k+1}), Z\right)^2 + \operatorname{dist}\left(\operatorname{prox}_{\phi_2{}^*}(z^{k+1}), W\right)^2\\ \leq&(1-\frac{1}{\kappa^2})\left(\operatorname{dist}\left(\operatorname{prox}_{\phi_2}(z^{k}), Z\right)^2 + \operatorname{dist}\left(\operatorname{prox}_{\phi_2{}^*}(z^{k}), W\right)^2\right) \quad \text{for all } k \text{ such that } z^k \in \mathbb{B}(\bar{z}, \epsilon).\end{aligned}$$

Since $\{z^k\}$ converges to $\bar{z}$, there exists $k_0 > 0$ such that $z^k \in \mathbb{B}(\bar{x}, \epsilon)$ when $k \geq k_0$. Therefore, we have

$$\begin{aligned}&\operatorname{dist}\left(\operatorname{prox}_{\phi_2}(z^{k+1}), Z\right)^2 + \operatorname{dist}\left(\operatorname{prox}_{\phi_2{}^*}(z^{k+1}), W\right)^2\\ \leq&(1-\frac{1}{\kappa^2})\left(\operatorname{dist}\left(\operatorname{prox}_{\phi_2}(z^{k}), Z\right)^2 + \operatorname{dist}\left(\operatorname{prox}_{\phi_2{}^*}(z^{k}), W\right)^2\right) \quad \forall k \geq k_0,\end{aligned}$$

which implies that

$$\begin{aligned}&\operatorname{dist}\left(\operatorname{prox}_{\phi_2}(z^{k}), Z\right) + \operatorname{dist}\left(\operatorname{prox}_{\phi_2{}^*}(z^{k}), W\right)\\ \leq&\sqrt{2}\sqrt{\operatorname{dist}\left(\operatorname{prox}_{\phi_2}(z^{k}), Z\right)^2 + \operatorname{dist}\left(\operatorname{prox}_{\phi_2{}^*}(z^{k}), W\right)^2}\\ \leq&C_0\rho^k, \quad \forall k \geq k_0,\end{aligned}$$

where $C_0 = \sqrt{2}\left(\operatorname{dist}\left(\operatorname{prox}_{\phi_2}(z^{k_0}), Z\right)^2 + \operatorname{dist}\left(\operatorname{prox}_{\phi_2{}^*}(z^{k_0}), W\right)^2\right)^{\frac{1}{2}}(1-\frac{1}{\kappa^2})^{-\frac{k_0}{2}}$ and $\rho = (1-\frac{1}{\kappa^2})^{\frac{1}{2}}$. Then, it follows directly from (12) that

$$\|z^{k+1} - z^k\| \leq \operatorname{dist}\left(\operatorname{prox}_{\phi_2}(z^{k}), Z\right) + \operatorname{dist}\left(\operatorname{prox}_{\phi_2{}^*}(z^{k}), W\right) \leq C_0\rho^k, \quad \forall k \geq k_0.$$

Note that $z^k = \operatorname{prox}_{\phi_2}(z^k) + \operatorname{prox}_{\phi_2{}^*}(z^k)$ and $Fix(T_{DR}) = Z + W$. In conclusion,

$$\operatorname{dist}\left(z^k, Fix(T_{DR})\right) \leq C_0\rho^k, \quad \forall k \geq k_0,$$

which completes our proof. ■

2.3.2. Calmness Conditions to Ensure DR-iteration-based Error Bound

In the last subsection, we show that the DR-iteration-based error bound condition can conceptually ensure the linear convergence of the DRSM. Generally, this condition cannot be checked directly. In this subsection, we show that certain calmness conditions which are independent of the iterative scheme suffice to ensure the DR-iteration-based error bound condition. This means appropriate conditions on the model itself can guarantee the DR-iteration-based error bound condition and hence the linear convergence of the DRSM. To this end, we first define the following two multifunctions:

$$\mathcal{T}_1(p) := \left\{\lambda \;\middle|\; p \in \partial\phi_1(\lambda - p) + \partial\phi_2(\lambda)\right\}, \tag{13}$$

and

$$\mathcal{T}_2(p) := \left\{\mu \;\middle|\; p \in \partial(\phi_1^* \circ -Id)(\mu - p) + \partial\phi_2^*(\mu)\right\}. \tag{14}$$

Remark 13 (Perturbation perspective) *As aforementioned, the set-valued maps $\mathcal{T}_1$ and $\mathcal{T}_2$ are defined from the perturbation perspective. In particular, according to the convergence result given in (Bauschke and Moursi, 2017), we know that the sequences $\{u^k\}$ and $\{z^k-u^k\}$ converge to some points in $Z=\mathcal{T}_1(0)$ and $W=\mathcal{T}_2(0)$, respectively. Additionally, as shown in the proof of Theorem 14, at each iteration k, we have*

$$\begin{aligned} u^k - v^k &\in \partial\phi_1(u^k-(u^k-v^k)) + \partial\phi_2(u^k),\\ u^k - v^k &\in \partial(\phi_1{}^* \circ -Id)(z^k-u^k-(z^k-z^{k+1})) + \partial\phi_2{}^*(z^k-u^k). \end{aligned}$$

Note that the DRSM iterative scheme implies that $z^k-z^{k+1}=u^k-v^k$,

$$\begin{aligned} z^k - z^{k+1} &\in \partial\phi_1(u^k-(u^k-v^k)) + \partial\phi_2(u^k),\\ z^k - z^{k+1} &\in \partial(\phi_1{}^* \circ -Id)(z^k-u^k-(z^k-z^{k+1})) + \partial\phi_2{}^*(z^k-u^k). \end{aligned}$$

Following the perturbation technique introduced in (Wang et al., 2018), if we introduce perturbation p^k to the place where the difference between two consecutive generated points z^k-z^{k+1} appears, $\mathcal{T}_1$ and $\mathcal{T}_2$ are therefore defined,

$$\begin{aligned} u^k &\in \mathcal{T}_1(p^k),\\ z^k-u^k &\in \mathcal{T}_2(p^k). \end{aligned}$$

We next show that the calmness of $\mathcal{T}_1$ and $\mathcal{T}_2$ ensures the DR-iteration-based error bound and hence the linear convergence of the DRSM. Let $\{z^k\}$ be the sequence generated by the DRSM, and according to Proposition 10, $\{z^k\}$ converges to some point $\bar{z}\in Fix(T_{DR})$.

Theorem 14 (Linear convergence of DRSM under the calmness of $\mathcal{T}_1$ and $\mathcal{T}_2$) *Let the sequence $\{z^k\}$ be generated by the DRSM, and $\bar{z}\in Fix(T_{DR})$ be the limit point of $\{z^k\}$. Suppose that $\mathcal{T}_1$ is calm at $(0,\bar{\lambda})$, where $\bar{\lambda}=\mathrm{prox}_{\phi_2}(\bar{z})$, and $\mathcal{T}_2$ is calm at $(0,\bar{\mu})$, where $\bar{\mu}=\mathrm{prox}_{\phi_2{}^*}(\bar{z})$. Then the DR-iteration-based error bound holds at $\bar{z}$ and hence the sequence $\{z^k\}$ converges to $\bar{z}$ linearly. That is, there exist $k_0>0$ and $0<\rho=\sqrt{1-\frac{1}{\kappa^2}}<1$, such that, for all $k\geq k_0$, it holds that*

$$\begin{aligned} &\mathrm{dist}\left(\mathrm{prox}_{\phi_2}(z^{k+1}),Z\right)+\mathrm{dist}\left(\mathrm{prox}_{\phi_2{}^*}(z^{k+1}),W\right)\\ \leq\ &\rho\left(\mathrm{dist}\left(\mathrm{prox}_{\phi_2}(z^{k}),Z\right)+\mathrm{dist}\left(\mathrm{prox}_{\phi_2{}^*}(z^{k}),W\right)\right). \end{aligned}$$

Furthermore, there exists $C_0>0$ such that

$$\mathrm{dist}\left(\mathrm{prox}_{\phi_2}(z^{k}),Z\right)+\mathrm{dist}\left(\mathrm{prox}_{\phi_2{}^*}(z^{k}),W\right)\leq C_0\rho^k,\quad \forall k\geq k_0,$$

$$\|z^{k+1}-z^k\|\leq C_0\rho^k,\quad \forall k\geq k_0,$$

and

$$\mathrm{dist}\left(z^k,Fix(T_{DR})\right)\leq C_0\rho^k,\quad \forall k\geq k_0.$$

Proof See Appendix B. ■

2.4. Verification of the Calmness of $\mathcal{T}_1$ and $\mathcal{T}_2$

It becomes necessary to prove when $\mathcal{T}_1$ and $\mathcal{T}_2$ meet the calmness conditions. For this purpose, taking into consideration the problem structure, we shall investigate sufficient conditions for the calmness of $\mathcal{T}_1$ and $\mathcal{T}_2$. Before we do so, we establish some preliminary results.

Lemma 15 *Let ψ be a proper, lower semicontinuous, convex function in form of $\psi(x) = \mathbb{h}(\mathbb{L}x) + \delta_{\mathcal{A}}(x)$, where $\mathbb{L} \in \mathbb{R}^{m\times n}$ and $\mathcal{A} := a + \mathcal{A}_0$ be an affine space in $\mathbb{R}^n$ with some vector a and subspace $\mathcal{A}_0$ in $\mathbb{R}^n$. Suppose $\psi^*(y)$ is the conjugate function of $\psi(x)$, and then* $\mathrm{dom}\ \psi^* \subset \mathrm{range}\,(\mathbb{L}^T) + \mathcal{A}_0^{\perp}$.

Proof See Appendix C. ■

For the sake of mathematical generality, we employ $\mathbb{L}$ and $\mathcal{A}$ to introduce Lemma 15 as an independent result. When Lemma 15 is applied, in Lemma 21, $\mathbb{L}$ and $\mathcal{A}$ are specified as A^T and $\mathbb{R}^m$, respectively; while in Lemma 23, $\mathbb{L}$ and $\mathcal{A}$ play the same roles as K and $\mathcal{V}$, respectively.

We recall a proposition given in (Goebel and Rockafellar, 2008, Corollary 4.4).

Proposition 16 *Let $\mathcal{C}$ be the class of all proper, lower semicontinuous, convex function ϕ satisfying parts (i) and (ii) of Assumption1.2, i.e., ϕ is essentially differentiable, $\nabla\phi$ is locally Lipschitz continuous and ϕ is essentially locally strongly convex. ϕ^* denotes the convex conjugate function of ϕ, i.e., $\phi^*(x^*) := \sup_x\{\langle x^*, x\rangle - \phi(x)\}$. Then*

$$\phi \in \mathcal{C} \quad \textit{if and only if} \quad \phi^* \in \mathcal{C}.$$

We also need the following proposition given in (Rockafellar, 1970, Theorem 26.1).

Proposition 17 *Let ϕ be a lower semicontinuous, convex function. If ϕ is essentially differentiable, then*

$$\mathrm{dom}\ \partial\phi = \mathrm{int}(\mathrm{dom}\ \phi),$$

and

$$\partial\phi(x) = \nabla\phi(x), \quad \textit{when} \ \ x \in \mathrm{int}(\mathrm{dom}\ \phi).$$

We are now in the position to present a decomposition for the conjugate of a structured convex function, which will play an important role in our analysis.

Proposition 18 *Let $\psi(x) = \mathbb{h}(\mathbb{L}x) + \delta_{\mathcal{A}}(x)$, where $\mathbb{h} \in \mathcal{C}$, $\mathbb{L} \in \mathbb{R}^{m\times n}$ and $\mathcal{A} := a + \mathcal{A}_0$ be an affine space in $\mathbb{R}^n$ with some vector a and subspace $\mathcal{A}_0$ in $\mathbb{R}^n$. Then, the conjugate function ψ^* of ψ can be expressed as*

$$\psi^*(y) = \tilde{\mathbb{h}}^*(\tilde{\mathbb{L}}y) + \langle y, a\rangle + \delta_{\mathrm{range}\,(\mathbb{L}^T)+\mathcal{A}_0^{\perp}}(y), \quad \forall y \in \mathbb{R}^n,$$

where $\tilde{\mathbb{L}}$ is a matrix and $\tilde{\mathbb{h}} \in \mathcal{C}$. In addition, assume that

$$\partial\psi(x) = \mathbb{L}^T\nabla\mathbb{h}(\mathbb{L}x) + \mathcal{N}_{\mathcal{A}}(x), \quad \forall x \in \mathbb{R}^n, \qquad \textit{and} \quad \mathrm{dom}\ \partial\psi \neq \varnothing,$$

then

$$\partial\psi^*(y) = \tilde{\mathbb{L}}^T\nabla\tilde{\mathbb{h}}^*(\tilde{\mathbb{L}}y) + a + \mathcal{N}_{\mathrm{range}\,(\mathbb{L}^T)+\mathcal{A}_0^{\perp}}(y), \quad \forall y \in \mathbb{R}^n.$$

Proof See Appendix D. ■

Remark 19 *As* dom $\mathbb{h}$ *is not necessarily the entire space* $\mathbb{R}^m$*, and some affine space* $\mathcal{A}$ *and* $x \in \mathcal{A}$ *such that* $\partial\psi(x) \neq \mathbb{L}^T\nabla\mathbb{h}(\mathbb{L}x) + \mathcal{N}_{\mathcal{A}}(x)$ *may exist. Thus we need to assume that* $\partial\psi(x) = \mathbb{L}^T\nabla\mathbb{h}(\mathbb{L}x) + \mathcal{N}_{\mathcal{A}}(x), \forall x \in \mathbb{R}^n$ *in order to obtain the exact formula of* $\partial\psi^*$*.* $\mathrm{int}(\mathrm{dom}\ h) \cap \mathbb{L}\mathcal{A} \neq \varnothing$ *is a sufficient condition for such assumption.*

2.4.1. Sufficient Conditions for the Calmness of $\mathcal{T}_1$ and $\mathcal{T}_2$

With the preliminaries we have introduced, we are now able to characterize some sufficient conditions to ensure the calmness of $\mathcal{T}_1$ and $\mathcal{T}_2$, and hence the linear convergence of original ADMM. Before that, we state a basic result in Lemma 20 inspired by Assumption 1.2.

In particular, according to (Rockafellar, 1970, Theorem 23.8, Theorem 23.9), Assumption 1.2 guarantees that the chain rule for the subdifferential expansion of f under structured assumption holds strictly.

Lemma 20 *Suppose that* f *meets Assumption 1.2, then* $\partial f(x) = \partial\left(h(Lx)\right) + q = L^T\partial h(Lx) + q$*.*

Lemma 21 *Assume that* f *satisfies Assumption 1.2. Moreover, Assumption 1.1 holds. Then* $\phi_1(\lambda) = f^*(A^T\lambda) - b^T\lambda$ *admits an alternative form of*

$$\phi_1(\lambda) = \tilde{h}^*\left(K\lambda - \tilde{q}\right) - b^T\lambda + \delta_{\mathcal{V}}(\lambda)$$

with some $\tilde{h} \in \mathcal{C}$*, matrix* $K := \tilde{L}A^T$*, vector* $\tilde{q} := \tilde{L}q$ *and affine space* $\mathcal{V} := \{\lambda \mid A^T\lambda - q \in \mathrm{range}\,(L^T)\}$*. Furthermore, we have* dom $\partial\phi_1 \neq \varnothing$*,*

$$\partial\phi_1(\lambda) = K^T\nabla\tilde{h}^*\left(K\lambda - \tilde{q}\right) - b + \mathcal{N}_{\mathcal{V}}(\lambda).$$

Proof See Appendix E. ■

Since $\mathcal{V}$ is an affine space, there must be some $v \in \mathcal{V}$ and a subspace $\mathcal{V}_0$ such that $\mathcal{V} = v + \mathcal{V}_0$.

By denoting

$$\tilde{\phi}_1(\lambda) := \tilde{h}^*\left(K\lambda - \tilde{q}\right) - b^T\lambda,$$

we can define a perturbed dual solution set multifunction as follows

$$\begin{aligned}\tilde{\mathcal{S}}_{D_1}(p) &:= \left\{\lambda \middle|\ p \in \partial\phi_1(\lambda) + \partial\phi_2(\lambda)\right\} \\ &= \left\{\lambda \in \mathcal{V} \middle|\ p \in \nabla\tilde{\phi}_1(\lambda) + \mathcal{V}_0^{\perp} + \partial\phi_2(\lambda)\right\}.\end{aligned}$$

Note that $\tilde{\mathcal{S}}_{D_1}(0) = Z$. We next investigate sufficient conditions to ensure the calmness of $\mathcal{T}_1$. To this end, let us recall

$$\begin{aligned}\mathcal{T}_1(p) &:= \left\{\lambda \ \middle|\ p \in \phi_1(\lambda - p) + \partial\phi_2(\lambda)\right\}, \\ &= \left\{\lambda - p \in \mathcal{V} \ \middle|\ p \in \nabla\tilde{\phi}_1(\lambda - p) + \mathcal{V}_0^{\perp} + \partial\phi_2(\lambda)\right\}.\end{aligned}$$

Proposition 22 *For any solution $\bar{\lambda}$ to the dual problem (D), the calmness of $\tilde{\mathcal{S}}_{D_1}(p)$ at $(0,\bar{\lambda})$ suffices to ensure the calmness of $\mathcal{T}_1(p)$ at $(0,\bar{\lambda})$.*

Proof We define the multifunction

$$\tilde{\mathcal{T}}_1(p) := \left\{\lambda \in \mathcal{V} \;\middle|\; p \in \nabla\tilde{\phi}_1(\lambda) + \mathcal{V}_0^{\perp} + \partial\phi_2(\lambda + p)\right\}.$$

It is easy to see that

$$\tilde{\mathcal{T}}_1(p) = -p + \mathcal{T}_1(p).$$

Straightforwardly, the calmness of $\tilde{\mathcal{T}}_1(p)$ at $(0,\bar{\lambda})$ is equivalent to the calmness of $\mathcal{T}_1$ at $(0,\bar{\lambda})$.

We next rewrite $\tilde{\mathcal{T}}_1(p)$ as

$$\tilde{\mathcal{T}}_1(p) = \left\{\lambda \in \mathcal{V} \;\middle|\; 0 \in \mathcal{M}(p,\lambda)\right\}, \tag{15}$$

where

$$\mathcal{M}(p,\lambda) := \mathcal{G}(p,\lambda) + 0 \times \mathcal{V}_0^{\perp} + gph\left(\partial\phi_2\right),\ \mathcal{G}(p,\lambda) := \begin{pmatrix} -\lambda - p \\ -p + \nabla\tilde{\phi}_1(\lambda) \end{pmatrix}.$$

Following the technique presented in (Gfrerer and Klatte, 2016), we introduce two multifunctions $H_{\mathcal{M}} : \mathbb{R}^n \rightrightarrows \mathcal{V} \times \mathbb{R}^n$ and $\mathcal{M}_p : \mathcal{V} \rightrightarrows \mathbb{R}^n \times \mathbb{R}^n$ defined, respectively, by

$$H_{\mathcal{M}}(p) := \left\{(\lambda, y) \;\middle|\; \lambda \in \mathcal{V}, y \in \mathcal{M}(p,\lambda)\right\} \quad \text{and} \quad \mathcal{M}_p(\lambda) := \left\{y \;\middle|\; y \in \mathcal{M}(p,\lambda)\right\}.$$

By (Gfrerer and Klatte, 2016, Theorem 3.3), if $\mathcal{M}_0(\lambda) := \mathcal{M}(0,\lambda)$ is metrically subregular at $(\bar{\lambda}, 0)$ and $\mathcal{M}$ has the restricted calmness property with respect to p at $(0,\bar{\lambda},0)$, i.e., if there are reals $\kappa > 0$ and $\epsilon > 0$ such that

$$\operatorname{dist}\left((\lambda,0), H_{\mathcal{M}}(0)\right) \le \kappa \left\|p\right\|, \quad \forall \left\|p\right\| \le \epsilon, \left\|\lambda - \bar{\lambda}\right\| \le \epsilon, (\lambda,0) \in H_{\mathcal{M}}(p),$$

then $\tilde{\mathcal{T}}_1$ is calm at $(0,\bar{\lambda})$ and thus $\mathcal{T}_1$ is calm at $(0,\bar{\lambda})$. Based on this theorem, in order to prove the the calmness of $\mathcal{T}_1$ provided the calmness of $\tilde{\mathcal{S}}_{D_1}$, we only have to justify the metric subregularity of $\mathcal{M}_0(\lambda)$ and the restricted calmness property of $\mathcal{M}$.

- We first show that $\mathcal{M}$ meets the restricted calmness property with respect to p at $(0,\bar{\lambda},0)$. Indeed, because $\bar{\lambda} \in \operatorname{int}(\operatorname{dom}\ \tilde{\phi}_1)$ and by the locally Lipschitz continuity of $\nabla\tilde{\phi}_1$, there is a constant $L_1 > 0$ along with neighborhoods $\mathbb{U}(0)$ of 0 as well as $\mathbb{U}(\bar{\lambda})$ of $\bar{\lambda}$ such that $\mathcal{G}$ is also Lipschitz continuous with modulus L_1 on $\mathbb{U}(0) \times \mathbb{U}(\bar{x})$. Given $(p,x,0)$ where $p \in \mathbb{U}(0), \lambda \in \mathbb{U}(\bar{\lambda})$ and $(\lambda,0) \in H_{\mathcal{M}}(p)$, by definition, $\lambda \in \mathcal{V}$ and $0 \in \mathcal{M}(p,\lambda) = \mathcal{G}(p,\lambda) + 0 \times \mathcal{V}_0^{\perp} + gph(\partial\phi_2)$. As a consequence, $\lambda \in \mathcal{V}$, $\mathcal{G}(0,\lambda) - \mathcal{G}(p,x) \in \mathcal{G}(0,x) + 0 \times \mathcal{V}_0^{\perp} + gph(\partial\phi_2)$ and hence $(\lambda, \mathcal{G}(0,\lambda) - \mathcal{G}(p,\lambda)) \in H_{\mathcal{M}}(0)$. Therefore we have the following inequality:

  $$\operatorname{dist}\left((\lambda,0), H_{\mathcal{M}}(0)\right) \le \left\|(\lambda,0) - (\lambda, \mathcal{G}(0,\lambda) - \mathcal{G}(p,\lambda))\right\| \le \left\|\mathcal{G}(0,\lambda) - \mathcal{G}(p,\lambda)\right\| \le L_1 \left\|p\right\|,$$

  which means that $\mathcal{M}$ has the restricted calmness property with respect to p at $(0,\bar{\lambda},0)$;

- We next show that $\mathcal{M}_0(\lambda) := \mathcal{M}(0,\lambda)$ is metrically subregular at $(\bar{\lambda},0)$ provided that $\mathcal{S}_P$ is *calm* at $(0,\bar{\lambda})$. Indeed, by the the locally Lipschitz continuity of $\nabla\tilde{\phi}_1$ around $\bar{\lambda}$ and (Gfrerer and Ye, 2017, Proposition 3), $\mathcal{M}_0(\lambda)$ is metrically subregular at $(\bar{\lambda},(0,0))$ if and only if $\nabla\tilde{\phi}_1(\lambda)+\mathcal{V}_0^{\perp}+\partial\phi_2(\lambda)$ is metrically subregular relative to $\mathcal{V}$ at $(\bar{\lambda},0)$, which is equivalent to the calmness of $\tilde{\mathcal{S}}_{D_1}$ at $(0,\bar{\lambda})$.

Consequently, $\mathcal{T}_1$ is calm at $(0,\bar{\lambda})$ provided the calmness of $\tilde{\mathcal{S}}_{D_1}$ at $(0,\bar{\lambda})$. ■

Naturally, we shall explore sufficient conditions to ensure the calmness of $\mathcal{T}_2$. For this purpose, let us define the multifunction

$$\tilde{\mathcal{S}}_{D_2}(p) := \left\{\mu \middle|\ p \in \partial(\phi_1^* \circ -Id)(\mu) + \partial\phi_2^*(\mu)\right\}.$$

Note that $\tilde{\mathcal{S}}_{D_2}(0) = W$. Similar to Lemma 21, we have the following result.

Lemma 23 *Assume that f satisfies Assumption 1.2. Moreover, Assumption 1.1 holds. Then $\phi_1{}^*(-\mu)$ admits a form of*

$$\phi_1{}^*(-\mu) = \hat{h}\left(\hat{K}\mu+\hat{q}\right) - \langle v,\mu\rangle + \delta_{\hat{\mathcal{V}}}(\mu) + \langle v,b\rangle,$$

with some $\hat{h}\in\mathcal{C}$, matrix $\hat{K}$, vector $\hat{q}$ and affine space $\hat{\mathcal{V}}$. Furthermore,

$$\partial\left(\phi_1{}^*(-\mu)\right) = \hat{K}^T\nabla\hat{h}\left(\hat{K}\mu+\hat{q}\right) - v + \mathcal{N}_{\hat{\mathcal{V}}}(\mu).$$

Proof See Appendix F. ■

Similar to Proposition 22, Lemma 23 inspires the following sufficiency.

Proposition 24 *For any $\bar{\mu}\in\tilde{\mathcal{S}}_{D_2}(0)$, the calmness of $\tilde{\mathcal{S}}_{D_2}$ at $(0,\bar{\mu})$ is sufficient for the calmness of $\mathcal{T}_2$ at $(0,\bar{\mu})$.*

2.4.2. Verifying Calmness of $\mathcal{T}_1$ and $\mathcal{T}_2$ under Structured Assumptions

As mentioned, we want to find verifiable conditions to discern the linear convergence of the original ADMM. Based on our previous analysis, it is clear that if Problem (1) meets the structured polyhedricity assumption, then both $\tilde{\mathcal{S}}_{D_1}$ and $\tilde{\mathcal{S}}_{D_2}$ are calm, both $\mathcal{T}_1$ and $\mathcal{T}_2$ are also calm, and eventually the linear convergence of the original ADMM can be ensured. To show how to verify the calmness of $\mathcal{T}_1$ and $\mathcal{T}_2$ under structured assumptions, recall that under Assumptions 1.2 and 1.1, it holds that

$$Z = \arg\min_{\lambda}\{\phi_1(\lambda)+\phi_2(\lambda)\} = \left\{\lambda\in\mathcal{V}\middle|\ 0\in K^T\nabla\tilde{h}^*\left(K\lambda-\tilde{q}\right)-b+\mathcal{V}_0^{\perp}+\partial\phi_2(\lambda)\right\}.$$

Lemma 25 *If Assumptions 1.1 and 1.2 hold for Problem (1), there exist $\bar{t},\bar{g}\in\mathbb{R}^n$ such that*

$$Z = \{\lambda\in\mathcal{V}\mid K\lambda=\bar{t},\quad 0\in\bar{g}+\mathcal{V}_0^{\perp}+\partial\phi_2(\lambda)\}. \tag{16}$$

Proof See Appendix G. ■

To facilitate our analysis, we introduce an auxiliary set-valued map:

$$\Gamma_{DR}(p_1,p_2) := \Gamma_1(p_1)\cap\Gamma_2(p_2) = \{\lambda\in\mathcal{V} \mid p_1 = K\lambda-\bar{t},\quad p_2\in\bar{g}+\mathcal{V}_0^{\perp}+\partial\phi_2(\lambda)\},\quad (17)$$

where

$$\Gamma_1(p_1) := \{\lambda \mid p_1 = K\lambda-\bar{t}\},\qquad \Gamma_2(p_2) := \{\lambda\in\mathcal{V}\mid p_2\in\bar{g}+\mathcal{V}_0^{\perp}+\partial\phi_2(\lambda)\}.\quad (18)$$

Since $\Gamma_{DR}(0,0) = Z$, $\Gamma_{DR}(p_1,p_2)$ can be considered as a set-valued map which perturbs Z in (16). The following proposition links the metric subregularity of $\tilde{\mathcal{S}}_{D_1}^{-1}$ and that of Γ_{DR}^{-1}, which thereby allows us to verify the subregularity conditions of Γ_{DR}^{-1} instead of $\tilde{\mathcal{S}}_{D_1}^{-1}$.

Proposition 26 *Suppose that Assumption 1.2 holds for Problem (1). The metric subregularity conditions of Γ_{DR}^{-1} and $\tilde{\mathcal{S}}_{D_1}^{-1}$ are equivalent. Precisely, given $\bar{\lambda}\in\mathcal{S}_D$, the following two statements are equivalent:*

(i) there exist $\kappa_1,\epsilon_1>0$ such that $dist\,(\lambda,\Gamma_{DR}(0,0))\le\kappa_1 dist\left(0,\Gamma_{DR}^{-1}(\lambda)\right),\ \forall\lambda\in\mathbb{B}_{\epsilon_1}(\bar{\lambda})$;

(ii) there exist $\kappa_2,\epsilon_2>0$ such that $dist\left(\lambda,\tilde{\mathcal{S}}_{D_1}(0)\right)\le\kappa_2 dist\left(0,\tilde{\mathcal{S}}_{D_1}^{-1}(\lambda)\right),\ \forall\lambda\in\mathbb{B}_{\epsilon_2}(\bar{\lambda})$.

Proof Given $\bar{\lambda}\in Z=\Gamma_{DR}(0,0)$, suppose that there exist $\kappa_1,\epsilon_1>0$ such that

$$\mathrm{dist}\,(\lambda,\Gamma_{DR}(0,0))\le\kappa_1\mathrm{dist}\left(0,\Gamma_{DR}^{-1}(\lambda)\right),\ \forall\lambda\in\mathbb{B}_{\epsilon_1}(\bar{\lambda})\subset\mathrm{int}(\mathrm{dom}\ \tilde{\phi}_1).$$

For any $\lambda\in\mathbb{B}_{\epsilon_1}(\bar{\lambda})\cap\mathcal{V}$, and any $\xi\in\nabla\tilde{\phi}_1(\lambda)+\mathcal{V}_0^{\perp}+\partial\phi_2(\lambda)$, by the locally Lipschitz continuity of $\nabla\tilde{h}^*$ implied by Assumption 1.2, there exists $L_{\tilde{h}^*}>0$ such that

$$\begin{aligned}\mathrm{dist}\,(\lambda,Z) &= \mathrm{dist}\,(\lambda,\Gamma_{DR}(0,0))\\ &\le \kappa_1\mathrm{dist}\left(0,\Gamma_{DR}^{-1}(\lambda)\right)\\ &\le \kappa_1\left(\|K\lambda-\bar{t}\|+\|\xi-\nabla\tilde{\phi}_1(\lambda)+\bar{g}\|\right)\\ &\le \kappa_1\left(\|K\lambda-\bar{t}\|+\|K^T\nabla\tilde{h}^*(K\lambda-\tilde{q})-K^T\nabla\tilde{h}^*(\bar{t}-\tilde{q})\|+\|\xi\|\right)\\ &\le (\kappa_1+L_{\tilde{h}^*}\|K\|)\|K\lambda-\bar{t}\|+\kappa_1\|\xi\|.\end{aligned}\quad (19)$$

Let $\hat{\lambda}$ be the projection of λ on Z. Since $0\in\bar{g}+\mathcal{V}_0^{\perp}+\partial\phi_2(\hat{\lambda})$, $\lambda-\hat{\lambda}\in\mathcal{V}_0$ and $\partial\phi_2$ is monotone, we have

$$\langle\xi-\nabla\tilde{\phi}_1(\lambda)+\bar{g},\lambda-\hat{\lambda}\rangle\ge 0.$$

Moreover, since $\bar{g}=K^T\nabla\tilde{h}^*(\bar{t}-\tilde{q})-b$, $K\hat{\lambda}=\bar{t}$, and due to the essentially locally strong convexity of $\tilde{h}^*$ around $\bar{t}$ again, there exists $\sigma>0$ such that

$$\sigma\|K\lambda-\bar{t}\|^2\le\langle\nabla\tilde{h}^*(K\lambda-\tilde{q})-\nabla\tilde{h}^*(\bar{t}-\tilde{q}),K\lambda-\bar{t}\rangle\le\langle\xi,\lambda-\hat{\lambda}\rangle\le\|\xi\|\cdot\|\lambda-\hat{\lambda}\|=\|\xi\|\cdot\mathrm{dist}\,(\lambda,Z).\quad (20)$$

Combining (19) and (20), we obtain

$$\operatorname{dist}(\lambda, Z) \leq \frac{\kappa_1 + L_{\tilde{h}^*}\|K\|}{\sqrt{\sigma}} \sqrt{\|\xi\| \operatorname{dist}(\lambda, Z)} + \kappa_1 \|\xi\|,$$

and consequently,

$$\operatorname{dist}(\lambda, Z) \leq \tilde{\kappa}\|\xi\|,$$

where $\tilde{\kappa} = \kappa_1 + 2c^2 + 2c\sqrt{\kappa_1 + c^2} > 0$ and $c = \frac{\kappa_1 + L_{\tilde{h}^*}\|K\|}{2\sqrt{\sigma}}$. Because ξ is arbitrarily chosen in $\nabla\tilde{\phi}_1(\lambda) + \mathcal{V}_0^{\perp} + \partial\phi_2(\lambda)$, we have

$$\operatorname{dist}\left(\lambda, \tilde{\mathcal{S}}_{D_1}(0)\right) = \operatorname{dist}(\lambda, Z) \leq \tilde{\kappa} \operatorname{dist}\left(0, \tilde{\mathcal{S}}_{D_1}^{-1}(\lambda)\right).$$

For $\lambda \in \mathbb{B}_{\epsilon_1}(\bar{\lambda}) \backslash \mathcal{V}$, $\tilde{\mathcal{S}}_{D_1}^{-1}(\lambda) = \varnothing$, the above inequality comes directly. Hence, there exists $\kappa_2 = \tilde{\kappa} > 0$ such that

$$\operatorname{dist}\left(\lambda, \tilde{\mathcal{S}}_{D_1}(0)\right) \leq \kappa_2 \operatorname{dist}\left(0, \tilde{\mathcal{S}}_{D_1}^{-1}(\lambda)\right), \text{ for all } \lambda \in \mathbb{B}_{\epsilon_1}(\bar{\lambda}).$$

Conversely, given $\bar{\lambda} \in Z$, suppose that there exist $\kappa_2, \epsilon_2 > 0$ such that

$$\operatorname{dist}\left(\lambda, \tilde{\mathcal{S}}_{D_1}(0)\right) \leq \kappa_2 \operatorname{dist}\left(0, \tilde{\mathcal{S}}_{D_1}^{-1}(\lambda)\right), \ \forall \lambda \in \mathbb{B}_{\epsilon_2}(\bar{\lambda}) \subset \operatorname{int}(\operatorname{dom} \tilde{\phi}_1).$$

For any fixed $\lambda \in \mathbb{B}_{\epsilon_2}(\bar{\lambda}) \cap \mathcal{V}$, and $(p_1, p_2) \in \Gamma_{DR}^{-1}(\lambda)$, it follows that

$$\begin{aligned} p_1 &= K\lambda - \bar{t}, \\ p_2 &\in K^T \nabla\tilde{h}^*(\bar{t} - \tilde{q}) - b + \mathcal{V}_0^{\perp} + \partial\phi_2(\lambda). \end{aligned}$$

To summarize, it holds that

$$p_2 + K^T \nabla\tilde{h}^*(K\lambda - \tilde{q}) - K^T \nabla\tilde{h}^*(K\lambda - p_1 - \tilde{q}) \in K^T \nabla\tilde{h}^*(K\lambda - \tilde{q}) - b + \mathcal{V}_0^{\perp} + \partial\phi_2(\lambda).$$

By virtue of the locally Lipschitz continuity of $\nabla\tilde{h}^*$, there exists $L_{\tilde{h}^*} > 0$ such that

$$\begin{aligned} \operatorname{dist}(\lambda, Z) = \operatorname{dist}\left(\lambda, \tilde{\mathcal{S}}_{D_1}(0)\right) &\leq \kappa_2 \operatorname{dist}\left(0, \tilde{\mathcal{S}}_{D_1}^{-1}(\lambda)\right) \\ &\leq \kappa_2 \|p_2 + K^T \nabla\tilde{h}^*(K\lambda - \tilde{q}) - K^T \nabla\tilde{h}^*(K\lambda - p_1 - \tilde{q})\| \\ &\leq \kappa_2 L_{\tilde{h}^*} \|K\| \|p_1\| + \kappa_2 \|p_2\|. \end{aligned}$$

Moreover, since (p_1, p_2) can be any element in $\Gamma_{DR}^{-1}(\lambda)$, we have

$$\operatorname{dist}(\lambda, \Gamma_{DR}(0,0)) = \operatorname{dist}(\lambda, Z) \leq \kappa_2 (L_{\tilde{h}^*}\|K\| + 1) \operatorname{dist}\left(0, \Gamma_{DR}^{-1}(\lambda)\right).$$

When $\lambda \in \mathbb{B}_{\epsilon_2}(\bar{\lambda}) \backslash \mathcal{V}$, $\Gamma_{DR}^{-1}(\lambda) = \varnothing$, the above inequality follows directly. Therefore, there exists $\kappa_1 = \kappa_2(L_{\tilde{h}^*}\|K\| + 1) > 0$ such that

$$\operatorname{dist}(\lambda, \Gamma_{DR}(0,0)) \leq \kappa_1 \operatorname{dist}\left(0, \Gamma_{DR}^{-1}(\lambda)\right) \text{ for all } \lambda \in \mathbb{B}_{\epsilon_2}(\bar{\lambda}).$$

The proof is complete. ■

The equivalence in Proposition 26 further yields a sufficient condition for the calmness of $\tilde{\mathcal{S}}_{D_1}$ as shown below; this is the main result of this section.

Theorem 27 *Suppose that Assumptions 1.1 and 1.2 hold and ∂g is a polyhedral multifunction. Given any $\bar{\lambda} \in \tilde{\mathcal{S}}_{D_1}(0)$, then $\tilde{\mathcal{S}}_{D_1}$ is calm at $(\bar{\lambda}, 0)$.*

Proof It is easy to see that both Γ_1 and Γ_2 are polyhedral multifunctions. Taking into consideration the fact that the class of polyhedral set-valued maps is closed under (finite) addition, scalar multiplication, and (finite) composition, we conclude that Γ_{DR} is a polyhedral multifunction and hence clam. By virtue of Proposition 26, $\tilde{\mathcal{S}}_{D_1}$ is calm at $(\bar{\lambda}, 0)$. ■

Combining Proposition 22 and Theorem 27, we are able to verify the desired calmness of $\mathcal{T}_1$ under the structured polyhedricity assumption.

Theorem 28 *Suppose that Problem (1) fulfills the structured polyhedricity assumption. Given any $\bar{\lambda} \in Z$, $\mathcal{T}_1$ is calm at $(0, \bar{\lambda})$.*

The last task in this part is to verify the desired calmness of $\mathcal{T}_2$ under the structured polyhedricity assumption. Indeed, analogous to the discussion for deriving Theorem 28, first with Lemma 23, there exist vector $\hat{t}$, $\hat{g}$ and affine space $\hat{\mathcal{V}} := \hat{v} + \hat{\mathcal{V}}_0$ with subspace $\hat{\mathcal{V}}_0$, such that

$$\tilde{\mathcal{S}}_{D_2}(0) = \tilde{\Gamma}_{DR}(0, 0),$$

with $\tilde{\Gamma}_{DR}$ defined as

$$\tilde{\Gamma}_{DR}(p_1, p_2) := \{\mu \in \hat{\mathcal{V}} \mid p_1 = \hat{K}\mu - \hat{t}, \quad p_2 \in \hat{g} + \hat{\mathcal{V}}_0^{\perp} + \partial\phi_2^*(\mu)\} \tag{21}$$

Then, considering the fact that ∂g is polyhedral multifunction if and only if ∂g^* be polyhedral multifunction, we know that $\partial\phi_2^*$ is polyhedral multifunction when the structured polyhedricity assumption is satisfied and we can conclude that $\tilde{\Gamma}_{DR}$ is calm at any point $(0, \bar{\mu})$ with $\bar{\mu} \in \tilde{\mathcal{S}}_{D_2}(0)$. Then, similar to Proposition 26, we can prove that $\tilde{\mathcal{S}}_{D_2}$ is calm at any point $(0, \bar{\mu}) \in gph\, \tilde{\mathcal{S}}_{D_2}$.

Moreover, together with Proposition 24, we have the desired calmness of $\mathcal{T}_2$.

Theorem 29 *Suppose that Problem (1) fulfills the structured polyhedricity assumption. Given any $\bar{\mu} \in \tilde{\mathcal{S}}_{D_2}(0)$, $\mathcal{T}_2$ is calm at $(0, \bar{\mu})$.*

2.5. Transporting the Linear Convergence from DRSM to Original ADMM

Previous analysis for the linear convergence of the DRSM can be regarded as preparation for the analysis for the original ADMM. In this subsection, we show how to convert the previous analysis to derive the linear convergence of the original ADMM. Recall that the linear convergence of the DRSM through the lens of variational analysis is summarized in Theorem 28, Theorem 29, and Theorem 14. Below, we show the linear convergence of the original ADMM in sense of the dual variable sequence $\{\lambda^k\}$, the KKT residue sequence $\{\text{Res}^k\}$, and the objective function value sequence $\{\text{Val}^k\}$ together with the constraint feasibility sequence $\{\text{Fea}^k\}$, by simply using Theorem 14.

Theorem 30 *Assume that Problem (1) fulfills the structured polyhedricity assumption. Let $\{(x^k, y^k, \lambda^k)\}$ be the sequence generated by the original ADMM. Then, the sequence $\{\lambda^k\}$*

converges to Z linearly, where Z is the solution set of the dual problem (D). That is, there exist $k_0 > 0$, $0 < \rho < 1$ and $C_0 > 0$ such that, for all $k \geq k_0$, it holds that

$$\operatorname{dist}\left(\lambda^k, Z\right) \leq C_0 \rho^k.$$

Furthermore, we have

$$Fea(x^{k+1}, y^{k+1}, \lambda^{k+1}) \leq \frac{C_0}{\beta}\rho^k,$$

and there exist $\tilde{C}_0 > 0, \hat{C}_0 > 0$ such that for all $k \geq k_0 + 1$

$$Res(x^k, y^k, \lambda^k) \leq \tilde{C}_0 \rho^k,$$

and

$$|\, Val(x^k, y^k, \lambda^k) - Val^*| \leq \tilde{C}_0 \rho^k,$$

where Val^* represents the optimal objective value of Problem (1).

Proof According to Theorem 28, Theorem 29, and Theorem 14, when Problem (1) fulfills the structured polyhedricity assumption, there exist $k_0 > 0$, $C_0 > 0$ and $0 < \rho < 1$, such that, for all $k \geq k_0$, it holds that

$$\operatorname{dist}\left(\operatorname{prox}_{\phi_2}(z^k), Z\right) + \operatorname{dist}\left(\operatorname{prox}_{\phi_2{}^*}(z^k), W\right) \leq C_0 \rho^k, \quad \forall k \geq k_0. \tag{22}$$

According to Proposition 8, we know that, $\lambda^k = \operatorname{prox}_{\phi_2}(z^k)$. So we get the linear convergence of $\{\lambda^k\}$.

Next, according to Proposition 8, we have $z^k = \lambda^k + \beta B y^k$; and because of

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

we have

$$\|Ax^{k+1} + By^k - b\| = \frac{1}{\beta}\|z^{k+1} - z^k\| \leq \frac{C_0}{\beta}\rho^k, \quad \forall k \geq k_0, \tag{23}$$

where the last inequality follows from Theorem 14. Then, as shown in Proposition 8, we have $B^T\lambda^k \in \partial g(y^k)$ and $B^T\lambda^{k+1} \in \partial g(y^{k+1})$. By the monotonicity of ∂g, it follows that

$$\langle Ax^{k+1} + By^{k+1} - b, By^k - By^{k+1}\rangle = \frac{1}{\beta}\langle B^T\lambda^{k+1} - B^T\lambda^k, y^{k+1} - y^k\rangle \geq 0, \quad \forall k \geq 1.$$

Combining with (23), we get

$$\|Ax^{k+1} + By^{k+1} - b\| \leq \frac{C_0}{\beta}\rho^k, \quad \forall k \geq k_0. \tag{24}$$

From (68) in Proposition 8, we have

$$T_{KKT}(x^{k+1}, y^{k+1}, \lambda^{k+1}) = \begin{pmatrix} \beta(A^T B y^{k+1} - A^T B y^k) \\ 0 \\ Ax^{k+1} + By^{k+1} - b \end{pmatrix}.$$

Thus, by (24), for all $k \geq k_0$, we have

$$\mathrm{Res}(x^{k+1}, y^{k+1}, \lambda^{k+1}) \leq \beta\|A\|\|By^{k+1} - By^k\| + \|Ax^{k+1} + By^{k+1} - b\| \leq \max(\beta\|A\|, 1)\frac{C_0}{\beta}\rho^k,$$

which implies the linear convergence of the KKT residue sequence.

Additionally, note that

$$\beta(A^TBy^{k+1} - A^TBy^k) + A^T\lambda^{k+1} \in \partial f(x^{k+1})$$

and

$$B^T\lambda^{k+1} \in \partial g(y^{k+1}).$$

For any $(x^*, y^*, \lambda^*) \in \Omega^*$, we have

$$\begin{aligned} f(x^*) &\geq f(x^{k+1}) + \langle \beta(A^TBy^{k+1} - A^TBy^k) + A^T\lambda^{k+1}, x^* - x^{k+1}\rangle, \\ g(y^*) &\geq g(y^{k+1}) + \langle B^T\lambda^{k+1}, y^* - y^{k+1}\rangle. \end{aligned}$$

Combining above two inequalities, we get

$$\begin{aligned} f(x^*) + g(y^*) &\geq f(x^{k+1}) + g(y^{k+1}) + \langle \lambda^{k+1}, Ax^* + By^* - Ax^{k+1} - By^{k+1}\rangle \\ &\quad + \beta\langle By^{k+1} - By^k, Ax^* - Ax^{k+1}\rangle \\ &\geq f(x^{k+1}) + g(y^{k+1}) + \langle \lambda^{k+1}, b - Ax^{k+1} - By^{k+1}\rangle \\ &\quad + \beta\langle By^{k+1} - By^k, Ax^* - Ax^{k+1}\rangle, \end{aligned} \tag{25}$$

where the last inequality follows from $Ax^* + By^* - b = 0$. Similarly, since $A^T\lambda^* \in \partial f(x^*)$ and $B^T\lambda^* \in \partial g(y^*)$, we have

$$f(x^{k+1}) + g(y^{k+1}) \geq f(x^*) + g(y^*) + \langle \lambda^*, Ax^{k+1} + By^{k+1} - b\rangle. \tag{26}$$

Combining (25) and (26), we get

$$\begin{aligned} &|f(x^{k+1}) + g(y^{k+1}) - f(x^*) - g(y^*)| \\ &\leq \max\{\|\lambda^{k+1}\|, \|\lambda^*\|\}\|Ax^{k+1} + By^{k+1} - b\| + \beta\|Ax^{k+1} - Ax^*\|\|By^{k+1} - By^k\|. \end{aligned} \tag{27}$$

Then, by the non-emptiness of Ω^*, as proved in (He and Yang, 1998, Theorem 3), there exists $(\bar{x}, \bar{y}, \bar{\lambda}) \in \Omega^*$ such that $\|\lambda^k - \bar{\lambda}\| \to 0$ and $\|Ax^k - A\bar{x}\| \to 0$. We may take such KKT point $(\bar{x}, \bar{y}, \bar{\lambda}) \in \Omega^*$ in (27) and thus there exists $C_1 > 0$ such that

$$|f(x^{k+1}) + g(y^{k+1}) - f(\bar{x}) - g(\bar{y})| \leq C_1(\|Ax^{k+1} + By^{k+1} - b\| + \|By^{k+1} - By^k\|).$$

According to (24), we obtain the linear convergence with respect to the objective function value of Problem (1) straightforwardly. ■

As analyzed in (Aspelmeier et al., 2016), if Problem (1) meets the full polyhedricity assumption (S4), matrix A is of full column rank, and B is identity matrix, apart from the linear convergence of $\{\lambda^k\}$, the sequences $\{x^k\}$ and $\{y^k\}$ also converge linearly. We next clarify the relationship between W and the KKT solution set Ω^*. This connection helps us establish the linear convergence of $\{x^k\}$ and $\{y^k\}$ under the structured polyhedricity assumption and full rank conditions of A and B as well. Therefore, the linear convergence results in (Aspelmeier et al., 2016) can be covered by our analysis.

Corollary 31 *In addition to the assumptions in Proposition 30, if the matrices A and B are both of full column rank, then we have the linear convergence of the sequences $\{x^k\}$ and $\{y^k\}$. That is, there exist $k_0 > 0$, $0 < \rho < 1$ and $C_0 > 0$, $\tilde{C}_0 > 0$ such that, for all $k \geq k_0$, it holds that*

$$\operatorname{dist}\left(x^k, \Omega_x^*\right) \leq \tilde{C}_0 \rho^k,$$

and

$$\operatorname{dist}\left(y^k, \Omega_y^*\right) \leq C_0 \rho^k,$$

where $\Omega_x^* := \{x \mid \exists y, \lambda \text{ such that } (x, y, \lambda) \in \Omega^*\}$ and $\Omega_y^* := \{y \mid \exists x, \lambda \text{ such that } (x, y, \lambda) \in \Omega^*\}$.

Proof See Appendix H. ■

3. Linear Convergence Rate of Linearized ADMM

In this section, we focus on the linearized ADMM where $G_1 = rI - \beta A^T A$ with $r > \beta \|A^T A\|$, $G_2 = 0$ and $\gamma = 1$ in (2); and discuss its linear convergence in terms of sequences $\{(x^k, \lambda^k)\}$, $\{\mathrm{Res}^k\}$ and $\{\mathrm{Val}^k, \mathrm{Fea}^k\}$, under certain structured assumptions.

3.1. Roadmap of Analysis

As aforementioned, for the full polyhedricity case (S4), the linear convergence of the sequence $\{(x^k, By^k, \lambda^k)\}$ generated by the linearized ADMM can be found in the literature; (see, e.g., Liu et al., 2018; Yang and Han, 2016). Hence, here we investigate other nontrivial cases. As well known, the linearized ADMM is highly relevant to the PDHG via a primal and dual perspective. Their relevance indicates that we can study the linear convergence of the linearized ADMM through the perspective of the PDHG. As illustrated in Remark 34, the perturbation analysis consideration inspires us to determine the metric subregularity of set-valued map $T(x, \lambda)$ defined in (32) for deriving the linear convergence of the PDHG. Taking full advantage of Assumption 1.2, we provide a finer characterization of the metric subregularity of T in terms of the calmness of set-valued map Γ_{PDHG} (see Proposition 40). When g is further assumed to be a convex piecewise linear-quadratic function, i.e., the structured polyhedricity assumption holds, the calmness of Γ_{PDHG} follows directly from Robinson's celebrated result (Robinson, 1981, Proposition 1).

It is worth mentioning that the main difficulty is the situation where g is not piecewise linear-quadratic; for instance, the $\ell_{1,q}$-norm regularizer with $q \in (1, 2]$ and the sparse-group LASSO regularizer. To this end, we further unearth a underlying property, i.e., the calmness of $\partial(g^*(B^T\lambda))$ holds automatically for the $\ell_{1,q}$-norm regularizer with $q \in (1, 2]$ and the sparse-group LASSO regularizer. Recall the calm intersection theorem introduced in (Klatte and Kummer, 2002, Theorem 3.6). The metric subregularity of T is thereby re-characterized in terms of the calmness of $\hat{\Omega}_x$ defined in (50). The calmness of $\hat{\Omega}_x$ eventually follows directly from (Robinson, 1981, Proposition 1).

To present our analysis more clearly, we summarize the roadmap of analysis in this section in Figure 2.

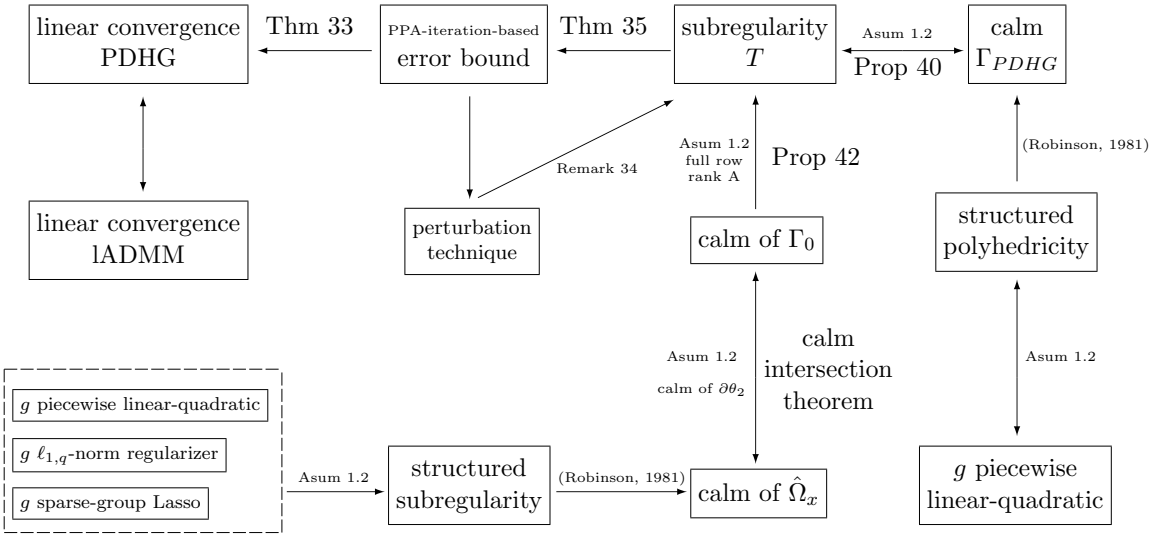

Figure 2: Roadmap to study linear convergence of the linearized ADMM

3.2. Linearized ADMM for Primal Problem Is Equivalent to PDHG for Min-max Problem

Under Assumption 1.1, Problem (1) is equivalent to the following saddle-point problem (min-max problem):

$$\min_x \max_\lambda \ \theta(x,\lambda) := f(x) - \langle \lambda, Ax\rangle - g^*(B^T\lambda) + \langle b, \lambda\rangle. \tag{28}$$

We define $\Omega^*_{x,\lambda}$ as the set of saddle-points to the above min-max problem (28). Let us further denote

$$\theta_1(x) = f(x),\ \theta_2(\lambda) = g^*(B^T\lambda) - \langle b, \lambda\rangle.$$

Then (28) can be rewritten into the following compact form

$$\min_x \max_\lambda \ \theta(x,\lambda) := \theta_1(x) - \langle \lambda, Ax\rangle - \theta_2(\lambda). \tag{29}$$

As analyzed in (Esser et al., 2010; Shefi, 2015), the linearized ADMM applied to Problem (1) turns out to be highly relevant to the application of the PDHG to the saddle-point problem (28). In fact, for the iterative (x^k, y^k, λ^k) generated by the linearized ADMM at the k-th iteration, we have

$$x^{k+1} = \arg\min_x \ f(x) - \langle \lambda^k, Ax\rangle + \langle \beta A^T(Ax^k + By^k - b), x\rangle + \frac{r}{2}\|x - x^k\|^2.$$

Since $\beta(Ax^k + By^k - b) = -\lambda^k + \lambda^{k-1}$, we know that

$$x^{k+1} = \arg\min_x \ f(x) - \langle 2\lambda^k - \lambda^{k-1}, Ax\rangle + \frac{r}{2}\|x - x^k\|^2.$$

Moreover, since

$$y^{k+1} = \arg\min_y \ g(y) - \langle \lambda^k, Ax^{k+1} + By - b\rangle + \frac{\beta}{2}\|Ax^{k+1} + By - b\|^2$$

and $\lambda^{k+1} = \lambda^k - \beta(Ax^{k+1} + By^{k+1} - b)$, we have

$$0 \in \partial g(y^{k+1}) - B^T\lambda^{k+1},$$

which implies

$$0 \in B\partial g^*(B^T\lambda^{k+1}) - By^{k+1}.$$

Furthermore, since $-By^{k+1} = \frac{1}{\beta}(\lambda^{k+1} - \lambda^k) + Ax^{k+1} - b$, we have

$$0 \in B\partial g^*(B^T\lambda^{k+1}) - b + Ax^{k+1} + \frac{1}{\beta}(\lambda^{k+1} - \lambda^k),$$

which implies

$$\lambda^{k+1} = \arg\min_\lambda g^*(B^T\lambda) - \langle b, \lambda\rangle + \langle Ax^{k+1}, \lambda\rangle + \frac{1}{2\beta}\|\lambda - \lambda^k\|^2.$$

Because the solution to the above problem is unique, the iterative scheme for λ^{k+1} is equivalent to that in the linearized ADMM

$$\begin{aligned} y^{k+1} &= \arg\min_y \ g(y) - \langle \lambda^k, Ax^{k+1} + By - b\rangle + \frac{\beta}{2}\|Ax^{k+1} + By - b\|^2 \\ \lambda^{k+1} &= \lambda^k - \beta(Ax^{k+1} + By^{k+1} - b). \end{aligned}$$

In summary, the sequence $\{(x^{k+1}, \lambda^k)\}$ generated by the linearized ADMM coincides with the sequence generated by the PDHG applied to (28), i.e.,

$$\begin{cases} \lambda^k = \operatorname{argmin}_\lambda \ g^*(B^T\lambda) - \langle b, \lambda\rangle + \langle Ax^k, \lambda\rangle + \dfrac{1}{2\tau}\|\lambda - \lambda^{k-1}\|^2, \\ x^{k+1} = \operatorname{argmin}_x \ f(x) - \langle 2\lambda^k - \lambda^{k-1}, Ax\rangle + \dfrac{1}{2\sigma}\|x - x^k\|^2, \end{cases} \tag{30}$$

where $\tau = \beta$ and $\sigma = 1/r$. At the k-th iteration of the PDHG, it follows from the optimality conditions of its subproblems that

$$0 \in \begin{pmatrix} \partial\theta_1(x^{k+1}) - A^T\lambda^k \\ \partial\theta_2(\lambda^k) + Ax^{k+1} \end{pmatrix} + \begin{pmatrix} \frac{1}{\tau}I & -A^T \\ -A & \frac{1}{\sigma} \end{pmatrix} \begin{pmatrix} x^{k+1} - x^k \\ \lambda^k - \lambda^{k-1} \end{pmatrix},$$

which can be further expressed in a more compact form

$$0 \in T(x^{k+1}, \lambda^k) + \mathcal{M}[(x^{k+1}, \lambda^k) - (x^k, \lambda^{k-1})], \tag{31}$$

where the matrix $\mathcal{M} \in \mathbb{R}^{(n_1+m)\times(n_1+m)}$ and the set-valued map $T : \mathbb{R}^{n_1+m} \rightrightarrows \mathbb{R}^{n_1+m}$ are defined, respectively, as:

$$\mathcal{M} := \begin{pmatrix} \frac{1}{\tau}I & -A^T \\ -A & \frac{1}{\sigma}I \end{pmatrix} \quad \text{and} \quad T(x, \lambda) := \begin{pmatrix} \partial\theta_1(x) - A^T\lambda \\ \partial\theta_2(\lambda) + Ax \end{pmatrix}. \tag{32}$$

3.3. Linear Convergence of PDHG under Metric Subregularity of T

In this subsection, we shall derive the linear convergence of the PDHG for solving problem (29) under the metric subregularity of T.

In the literature, there are some results for analyzing the convergence of the PDHG and its variants, (see, e.g., Bonettini and Ruggiero, 2012; Esser et al., 2010; He et al., 2014; He and Yuan, 2012b). Among them is He and Yuan (2012b) which is the first work showing the close connection between the PDHG and the well-known proximal point algorithm (PPA) proposed in (Martinet, 1970; Rockafellar, 1976), as well as revisiting the PDHG from the contraction perspective for convergence analysis (see Proposition 36). Research for PDHG's faster convergence rates, however, still stays in its infancy. In particular, it is known that, if both f and g are strongly convex, then the PDHG converges linearly; (see, e.g., Bonettini and Ruggiero, 2012; Valkonen, 2014).

Our approach to studying the linear convergence of the PDHG is motivated by the explanation initiated in (He and Yuan, 2012b) of that the PDHG can be regarded as an application of the PPA. More specifically, let us consider the application of PPA to the inclusion problem

$$0 \in T(x, \lambda), \tag{33}$$

where T is defined as in (32). We define the saddle-point set as $\Omega^*_{x,\lambda} := \{(x, \lambda) \mid 0 \in T(x, \lambda)\}$.

To proceed, we first establish the linear convergence of the PPA for solving a general generalized equation $0 \in T(x, \lambda)$ where T is a maximally monotone operator defined in (32).

$$0 \in T(\mathbb{x}^{k+1}) + \mathcal{M}(\mathbb{x}^{k+1} - \mathbb{x}^k), \tag{34}$$

where $\mathbb{x}^{k+1} := (x^{k+1}, \lambda^k)$ and $\mathcal{M}$ is a positive definite matrix in form of (32). Based on the convergence analysis of PPA given in the literature, for example, (Güler, 1991; Rockafellar, 1976; Teboulle, 1997), the sequence $\{\mathbb{x}^k\}$ converges to some point $\bar{\mathbb{x}} \in \Omega^*_{x,\lambda}$, and we are going to derive the linear convergence of $\{\mathbb{x}^k\}$ toward $\Omega^*_{x,\lambda}$ under the following error bound condition. Let the sequence $\{\mathbb{x}^k\}$ be generated by the PPA iterative scheme (34); and it converges to some point $\bar{\mathbb{x}} \in \Omega^*_{x,\lambda}$.

Definition 32 (PPA-iteration-based error bound) *Let the sequence $\{\mathbb{x}^k\}$ be generated by the PPA iterative scheme (34), and $\bar{\mathbb{x}} \in \Omega^*_{x,\lambda}$ be an accumulation point of $\{\mathbb{x}^k\}$. We say that the PPA-iteration-based error bound holds at $\bar{\mathbb{x}}$ if there exist $\epsilon, \kappa > 0$ such that*

$$\mathrm{dist}_{\mathcal{M}}\left(\mathbb{x}^{k+1}, \Omega^*_{x,\lambda}\right) \le \kappa \|\mathbb{x}^{k+1} - \mathbb{x}^k\|_{\mathcal{M}}, \quad \textit{for all } k \textit{ such that } \mathbb{x}^k \in \mathbb{B}(\bar{\mathbb{x}}, \epsilon),$$

where $\|d\|_{\mathcal{M}} := \sqrt{d^T \mathcal{M} d}$ and $dist_{\mathcal{M}}(d, \mathcal{D}) := \inf\{\|d - d'\|_{\mathcal{M}} \,\big|\, d' \in \mathcal{D}\}$ for a given subset $\mathcal{D}$ and vector d in the same space.

The following PPA linear convergence relies heavily on (Leventhal, 2009). The proof is needed for our further discussion and hence stated here.

Theorem 33 *Let the sequence $\{\mathbb{x}^k\}$ be generated by the PPA iterative scheme (34), and $\bar{\mathbb{x}} \in \Omega^*_{x,\lambda}$ be the limit point of $\{\mathbb{x}^k\}$. Assume that the PPA-iteration-based error bound holds*

at $\bar{\mathbb{x}}$*, and then the sequence* $\{\mathbb{x}^k\}$ *converges to* $\Omega^*_{x,\lambda}$ *linearly. That is, there exist* $k_0 > 0$ *and* $0 < \rho = \sqrt{\frac{\kappa^2}{1+\kappa^2}} < 1$ *such that, for all* $k \geq k_0$*, it holds that*

$$\mathrm{dist}_{\mathcal{M}}\left(\mathbb{x}^{k+1}, \Omega^*_{x,\lambda}\right) \leq \rho dist_{\mathcal{M}}\left(\mathbb{x}^{k}, \Omega^*_{x,\lambda}\right). \tag{35}$$

Furthermore, there exists $C_0 > 0$ *such that, for all* $k \geq k_0$*, it holds that*

$$\mathrm{dist}\left(\mathbb{x}^{k}, \Omega^*_{x,\lambda}\right) \leq C_0\rho^k, \tag{36}$$

and

$$\|\mathbb{x}^{k+1} - \mathbb{x}^k\| \leq C_0\rho^k. \tag{37}$$

Proof See Appendix I. ■

We have shown that the PPA-iteration-based error bound condition can conceptually ensure the linear convergence of the PPA. We next show that certain metric subregularity conditions which are independent of the iterative scheme suffice to ensure the PPA-iteration-based error bound condition.

Remark 34 (Perturbation perspective) *Following the perturbation analysis technique in (Wang et al., 2018), we introduce perturbation* $\mathbb{p}^k$ *to the place where the difference between two consecutive generated points* $\mathbb{x}^{k+1} - \mathbb{x}^k$ *appears, i.e.,*

$$\mathbb{p}^k = \mathbb{x}^{k+1} - \mathbb{x}^k,$$

which further induces the canonically perturbed system

$$-\mathcal{M}\mathbb{p}^k \in T(\mathbb{x}^{k+1}).$$

Thus we consider the metric subregularity of set-valued mapping T *in Theorem 35.*

Theorem 35 *Let the sequence* $\{\mathbb{x}^k\}$ *be generated by the PPA iterative scheme* (34)*, and* $\bar{\mathbb{x}} \in \Omega^*_{x,\lambda}$ *be the limit point of* $\{\mathbb{x}^k\}$*. If* T *is metrically subregular at* $(\bar{\mathbb{x}}, 0)$*, then the PPA-iteration-based error bound holds at* $\bar{\mathbb{x}}$ *and hence the sequence* $\{\mathbb{x}^k\}$ *converges to* $\Omega^*_{x,\lambda}$ *linearly. That is, there exist* $k_0 > 0$ *and* $0 < \rho < 1$ *such that, for all* $k \geq k_0$*, it holds that*

$$\mathrm{dist}_{\mathcal{M}}\left(\mathbb{x}^{k+1}, \Omega^*_{x,\lambda}\right) \leq \rho\mathrm{dist}_{\mathcal{M}}\left(\mathbb{x}^{k}, \Omega^*_{x,\lambda}\right). \tag{38}$$

Furthermore, there exists $C_0 > 0$ *such that, for all* $k \geq k_0$*, it holds that*

$$\mathrm{dist}\left(\mathbb{x}^{k}, \Omega^*_{x,\lambda}\right) \leq C_0\rho^k, \tag{39}$$

and

$$\|\mathbb{x}^{k+1} - \mathbb{x}^k\| \leq C_0\rho^k. \tag{40}$$

Proof See Appendix J. ■

Our main purpose in this subsection is discussing the linear convergence of the PDHG. As a prerequisite of the analysis to be delineated, the convergence of the PDHG can be given by the following proposition.

Proposition 36 *(Chambolle and Pock, 2011; He and Yuan, 2012b) Let* $\{(x^k, \lambda^{k-1})\}$ *be the sequence generated by the PDHG applied to the saddle-point problem (29) as in* (30)*. If* $\tau\sigma < \frac{1}{\|A^TA\|}$*, then the sequence* $\{(x^k, \lambda^{k-1})\}$ *converges to some point* $(\bar{x}, \bar{\lambda}) \in \Omega^*_{x,\lambda}$.

With the given convergence of the sequence $\{(x^k, \lambda^{k-1})\}$ generated by the PDHG applied to (29), the linear convergence of $\{(x^k, \lambda^{k-1})\}$ can be achieved according to Theorem 35, with the consideration that $\{(x^k, \lambda^{k-1})\}$ can also be regarded as the sequence generated by the PPA applied to (33). Note that when $\tau\sigma < \frac{1}{\|A^TA\|}$, $\mathcal{M}$ defined by (32) is positive definite. Then, the desired linear convergence of the PDHG follows immediately from the discussion above.

Theorem 37 *Suppose the sequence* $\{(x^k, \lambda^{k-1})\}$ *generated by PDHG in* (30) *with* $\tau\sigma < \frac{1}{\|A^TA\|}$*. Then according to Proposition36,* $\{(x^k, \lambda^{k-1})\}$ *converges to some point* $(\bar{x}, \bar{\lambda}) \in \Omega^*_{x,\lambda}$*. If* T *defined by (32) is metrically subregular at* $(\bar{x}, \bar{\lambda}, 0)$ *with modulus* κ*, then the sequence* $\{(x^k, \lambda^{k-1})\}$ *converges to* $\Omega^*_{x,\lambda}$ *linearly. That is, there exist* $k_0 > 0$ *and* $0 < \rho = \sqrt{\frac{\kappa^2}{1+\kappa^2}} < 1$ *such that, for all* $k \geq k_0$*, it holds that*

$$\mathrm{dist}_{\mathcal{M}}\left((x^{k+1}, \lambda^k), \Omega^*_{x,\lambda}\right) \leq \rho dist_{\mathcal{M}}\left((x^k, \lambda^{k-1}), \Omega^*_{x,\lambda}\right). \tag{41}$$

Furthermore, there exists $C_0 > 0$ *such that, for all* $k \geq k_0$*, it holds that*

$$\mathrm{dist}\left((x^{k+1}, \lambda^k), \Omega^*_{x,\lambda}\right) \leq C_0\rho^k, \tag{42}$$

and

$$\|x^{k+1} - x^k\| + \|\lambda^k - \lambda^{k-1}\| \leq C_0\rho^k. \tag{43}$$

3.4. Verification of Metric Subregularity of T

We have shown in the preceding section that the PDHG converges linearly under the metric subregularity of T. Then, we need to answer the question of which T satisfies the metric subregularity. For this purpose, taking into consideration the problem structure, we shall characterize equivalent or sufficient conditions for the metric subregularity of T. The following property is useful for developing our main results.

Proposition 38 *(Bertsekas et al., 2003, Proposition 2.6.1) When a saddle-point of the min-max problem* (29) *exists, the set of saddle-points* $\Omega^*_{x,\lambda}$ *for* (29) *can be characterized by* $X \times \Lambda$ *with*

$$X := \arg\min_x\{\sup_\lambda\ \theta(x, \lambda)\} = \arg\min_x\{\theta_1(x) + \theta_2^*(-Ax)\}$$

and

$$\Lambda := \arg\max_\lambda\{\inf_x\ \theta(x, \lambda)\} = \arg\min_\lambda\{\theta_1^*(A^T\lambda) + \theta_2(\lambda)\}.$$

Furthermore, we have $(x^*, \lambda^*) \in X \times \Lambda$ *if and only if* $0 \in T(x^*, \lambda^*)$.

3.4.1. Equivalent Characterization for the Metric Subregularity of T

In general, Proposition 38 provides a characterization of the saddle-point set. Thanks to the structure of f imposed in Assumption 1.2, we present an alternative characterization of the saddle-point set $\Omega^*_{x,\lambda}$.

Proposition 39 *When Problem (1) meets Assumption 1.2, the saddle-point set $\Omega^*_{x,\lambda}$ can be characterized as*

$$\Omega^*_{x,\lambda} = \{(x,\lambda) \mid Lx = \tilde{t},\ A^T\lambda = \tilde{g},\ 0 \in \partial\theta_2(\lambda) + Ax\}, \tag{44}$$

with some vector $\tilde{t} \in \mathbb{R}^l$ such that $Lx = \tilde{t}$ for all $x \in X$ and $\tilde{g} := L^T\nabla h(\tilde{t}) + q$.

Proof See Appendix K. ■

To facilitate our analysis, we introduce an auxiliary perturbed set-valued map with perturbation $p = (p_1, p_2, p_3)$ associated with the saddle-point-set characterization (44):

$$\Gamma_{PDHG}(p) := \{(x,\lambda) \mid p_1 = Lx - \tilde{t}, p_2 = \tilde{g} - A^T\lambda, p_3 \in \partial\theta_2(\lambda) + Ax\}.$$

Obviously, $\Gamma_{PDHG}(p)$ coincides with $\Omega^*_{x,\lambda}$ when $p = 0$. Similar to (Ye et al., 2018, Proposition 4.1), we have following equivalence.

Proposition 40 *Assume that Assumption 1.2 is satisfied. Then the metric subregularity conditions of Γ^{-1}_{PDHG} and T are equivalent. Precisely, given $(\bar{x},\bar{\lambda}) \in \Omega^*_{x,\lambda}$, the following two statements are equivalent:*

(i) There exist $\kappa_1, \epsilon_1 > 0$ such that

$$\operatorname{dist}\left((x,\lambda), \Gamma_{PDHG}(0)\right) \le \kappa_1 \operatorname{dist}\left(0, \Gamma^{-1}_{PDHG}(x,\lambda)\right), \quad \forall (x,\lambda) \in \mathbb{B}_{\epsilon_1}(\bar{x},\bar{\lambda}).$$

(ii) There exist $\kappa_2, \epsilon_2 > 0$ such that

$$\operatorname{dist}\left((x,\lambda), \Omega^*_{x,\lambda}\right) \le \kappa_2 \operatorname{dist}\left(0, T(x,\lambda)\right), \qquad \forall (x,\lambda) \in \mathbb{B}_{\epsilon_2}(\bar{x},\bar{\lambda}).$$

Proof Given any $(\bar{x},\bar{\lambda}) \in \Omega^*_{x,\lambda}$. Suppose that there exist $\kappa_1, \epsilon_1 > 0$ such that

$$\operatorname{dist}\left((x,\lambda), \Gamma_{PDHG}(0,0)\right) \le \kappa_1 \operatorname{dist}\left(0, \Gamma^{-1}_{PDHG}(x,\lambda)\right),\ \forall (x,\lambda) \in \mathbb{B}_{\epsilon_1}(\bar{x},\bar{\lambda}).$$

Due to the essentially locally strongly convexity of h and the locally Lipschitz continuity of ∇h, without loss of generality, we assume that ϵ_1 is small enough so that ∇h is strongly monotone and Lipschitz continuous on $\{Lx \mid (x,\lambda) \in \mathbb{B}_{\epsilon_1}(\bar{x},\bar{\lambda})\}$. For any $(x,\lambda) \in \mathbb{B}_{\epsilon_1}(\bar{x},\bar{\lambda})$, and any $(\xi,\eta) \in T(x,\lambda)$

$$\xi = \partial\theta_1(x) - A^T\lambda = L^T\nabla h(Lx) + q - A^T\lambda, \tag{45}$$
$$\eta \in \partial\theta_2(\lambda) + Ax, \tag{46}$$

and by the local Lipschitz continuity of ∇h, there exists $L_h > 0$ such that

$$\begin{aligned}\operatorname{dist}\left((x,\lambda),\Omega^*_{x,\lambda}\right) &= \operatorname{dist}\left((x,\lambda),\Gamma_{PDHG}(0)\right)\\ &\le \kappa_1 \operatorname{dist}\left(0,\Gamma^{-1}_{PDHG}(x,\lambda)\right)\\ &\le \kappa_1\left(\|Lx-\tilde{t}\|+\|\xi+\tilde{g}-L^T\nabla h(Lx)-q\|+\|\eta\|\right)\\ &\le \kappa_1\left(\|Lx-\tilde{t}\|+\|L\|\|\nabla h(\tilde{t})-\nabla h(Lx)\|+\|\xi\|+\|\eta\|\right)\\ &\le \kappa_1\left((1+\|L\|L_h)\|Lx-\tilde{t}\|+\|\xi\|+\|\eta\|\right).\end{aligned} \tag{47}$$

Let $(\hat{x},\hat{\lambda})$ be the projection of (x,λ) on $\Omega^*_{x,\lambda}$ and then $(\hat{x},\hat{\lambda})\in\mathbb{B}_{\epsilon_1}(\bar{x},\bar{\lambda})$. Since $0\in\partial\theta_2(\hat{\lambda})+A\hat{x}$ and $\partial\theta_2$ is monotone, we have

$$\langle \eta - Ax + A\hat{x}, \lambda-\hat{\lambda}\rangle \ge 0,$$

and subsequently,

$$\langle \eta, \lambda-\hat{\lambda}\rangle \ge \langle A^T\lambda - A^T\hat{\lambda}, x-\hat{x}\rangle.$$

Moreover, since $\xi = L^T\nabla h(Lx)+q-A^T\lambda$, $0=\tilde{g}-A^T\hat{\lambda}$, thanks to the local strong convexity of h, there exists $\sigma>0$ such that

$$\begin{aligned}\langle \xi, x-\hat{x}\rangle + \langle \eta,\lambda-\hat{\lambda}\rangle &\ge \langle L^T\nabla h(Lx)-L^T\nabla h(\bar{t}), x-\hat{x}\rangle\\ &= \langle \nabla h(Lx)-\nabla h(L\hat{x}), Lx-L\hat{x}\rangle\\ &\ge \sigma\|Lx-L\hat{x}\|^2 = \sigma\|Lx-\bar{t}\|^2.\end{aligned} \tag{48}$$

Combining (47) and (48), we obtain that

$$\operatorname{dist}((x,\lambda),\Omega^*_{x,\lambda}) \le c_1\sqrt{\|(\xi,\eta)\|\cdot\operatorname{dist}((x,\lambda),\Omega^*_{x,\lambda})}+c_2\|(\xi,\eta)\|,$$

with $c_1=\kappa_1(1+\|L\|L_h)/\sqrt{\sigma}, c_2=\sqrt{2}\kappa_1$, and consequently,

$$\operatorname{dist}((x,\lambda),\Omega^*_{x,\lambda}) \le \tilde{\kappa}\|(\xi,\eta)\|, \quad \text{where } \tilde{\kappa}=\left(\frac{c_1+\sqrt{c_1^2+4c_2}}{2}\right)^2>0.$$

Because ξ and η are arbitrarily chosen in $T(x,\lambda)$, we have

$$\operatorname{dist}\left((x,\lambda),T^{-1}(0)\right)=\operatorname{dist}((x,\lambda),\Omega^*_{x,\lambda})\le\tilde{\kappa}\operatorname{dist}\left(0,T(x,\lambda)\right).$$

Hence, there exists $\kappa_2=\tilde{\kappa}>0$ such that

$$\operatorname{dist}\left((x,\lambda),\Omega^*_{x,\lambda}\right)\le\kappa_2\operatorname{dist}\left(0,T\left(x,\lambda\right)\right),\ \forall(x,\lambda)\in\mathbb{B}_{\epsilon_2}(\bar{x},\bar{\lambda}).$$

Conversely, given any $(\bar{x},\bar{\lambda})\in\Omega^*_{x,\lambda}$, suppose that there exist $\kappa_2,\epsilon_2>0$ such that

$$\operatorname{dist}\left((x,\lambda),\Omega^*_{x,\lambda}\right)\le\kappa_2\operatorname{dist}\left(0,T\left(x,\lambda\right)\right),\ \forall(x,\lambda)\in\mathbb{B}_{\epsilon_2}(\bar{x},\bar{\lambda}).$$

For any fixed $(x,\lambda)\in\mathbb{B}_{\epsilon_2}(\bar{x},\bar{\lambda})$, and $(p_1,p_2,p_3)\in\Gamma^{-1}_{PDHG}(x,\lambda)$, it follows that

$$p_1=Lx-\bar{t},\ p_2=\tilde{g}-A^T\lambda,\ p_3\in\partial\theta_2(\lambda)+Ax.$$

To summarize, we have

$$p_2 + L^T\nabla h(Lx) - L^T\nabla h(Lx - p_1) = \partial\theta_1(x) - A^T\lambda,$$
$$p_3 \in \partial\theta_2(\lambda) + Ax.$$

By virtue of the locally Lipschitz continuity of ∇h, there exists $L_h > 0$ such that

$$\begin{aligned}\operatorname{dist}\left((x,\lambda),\Omega^*_{x,\lambda}\right) &\le \kappa_2 \operatorname{dist}\left(0, T\left(x,\lambda\right)\right)\\ &\le \kappa_2\left(\|p_2 + L^T\nabla h(Lx) - L^T\nabla h(Lx - p_1)\| + \|p_3\|\right)\\ &\le \kappa_2 L_h\|L\|\|p_1\| + \kappa_2\|p_2\| + \kappa_2\|p_3\|.\end{aligned}$$

Moreover, since (p_1, p_2, p_3) can be any element in $\Gamma^{-1}_{PDHG}\left(x,\lambda\right)$, we have

$$\operatorname{dist}\left((x,\lambda),\Gamma_{PDHG}(0)\right) = \operatorname{dist}\left((x,\lambda),\Omega^*_{x,\lambda}\right) \le \kappa_2(L_h\|L\| + 2)\operatorname{dist}\left(0,\Gamma^{-1}_{PDHG}\left(x,\lambda\right)\right).$$

Therefore, there exists $\kappa_1 = \kappa_2(L_h\|L\| + 2) > 0$ such that

$$\operatorname{dist}\left((x,\lambda),\Gamma_{PDHG}(0)\right) \le \kappa_1 \operatorname{dist}\left(0,\Gamma^{-1}_{PDHG}(x,\lambda)\right),\ \forall (x,\lambda) \in \mathbb{B}_{\epsilon_2}(\bar{x},\bar{\lambda}).$$

■

3.4.2. Sufficient Condition for the Metric Subregularity of T

Thanks to Proposition 39, when A is of full row rank, we can easily obtain another characterization of $\Omega^*_{x,\lambda}$.

Proposition 41 *Suppose that Assumption 1.2 is satisfied and A is of full row rank. The saddle-point set $\Omega^*_{x,\lambda}$ can be characterized as*

$$\Omega^*_{x,\lambda} = \{(x,\lambda) \mid Lx = \tilde{t},\ \lambda = \bar{\lambda},\ 0 \in D + Ax\}, \tag{49}$$

with $\bar{\lambda} \in \Lambda$, closed set $D := \partial\theta_2(\bar{\lambda})$ and some vector $\tilde{t} \in \mathbb{R}^l$.

We introduce an auxiliary set-valued map associated with characterization of $\Omega^*_{x,\lambda}$ in (49):

$$\Gamma_0(p) := \{(x,\lambda) \mid p_1 = Lx - \tilde{t},\ p_2 = -\bar{\lambda} + \lambda,\ p_3 \in D + Ax\}.$$

A useful connection is clarified below.

Proposition 42 *Suppose that Assumption 1.2 is satisfied and A is of full row rank. Then, given $(\bar{x},\bar{\lambda}) \in \Omega^*_{x,\lambda}$, if $\partial\theta_2$ is calm at $(\bar{\lambda}, -A\bar{x})$ and there exist $\kappa_1, \epsilon_1 > 0$ such that*

$$dist\left((x,\lambda),\Gamma_0(0)\right) \le \kappa_1\, dist\left(0,\Gamma_0^{-1}\left(x,\lambda\right)\right), \quad \forall (x,\lambda) \in \mathbb{B}_{\epsilon_1}(\bar{x},\bar{\lambda}),$$

then there exist $\kappa_2, \epsilon_2 > 0$ such that

$$dist\left((x,\lambda),\Omega^*_{x,\lambda}\right) \le \kappa_2\, dist\left(0, T\left(x,\lambda\right)\right), \qquad \forall (x,y) \in \mathbb{B}_{\epsilon_2}(\bar{x},\bar{\lambda}).$$

Proof See Appendix L. ■

It can be observed easily that the calmness of $\Gamma_0(p)$ at $(0, \bar{x}, \bar{\lambda})$ is equivalent to the calmness of the set-valued map $\Omega_x(p)$ defined by

$$\Omega_x(p) := \{x \mid p_1 = Lx - \tilde{t}, p_2 \in D + Ax\}$$

at $(0, \bar{x})$. For studying the calmness of Ω_x, we recall the calm intersection theorem introduced in (Klatte and Kummer, 2002, Theorem 3.6),

Proposition 43 (Calm intersection theorem) *Let $T_1 : \mathbb{R}^{q_1} \rightrightarrows \mathbb{R}^n$ and $T_2 : \mathbb{R}^{q_2} \rightrightarrows \mathbb{R}^n$ be two set-valued maps. Define set-valued maps:*

$$\begin{aligned} \widetilde{T}(p_1, p_2) &:= T_1(p_1) \cap T_2(p_2), \\ \widehat{T}(p_1) &:= T_1(p_1) \cap T_2(0). \end{aligned}$$

Let $\tilde{x} \in T(0,0)$. Suppose that both set-valued maps T_1 and T_2 are calm at $(0, \tilde{x})$ and T_1^{-1} is pseudo-Lipschitz at $(\tilde{x}, 0)$. Then $\widetilde{T}$ is calm at $(0, 0, \tilde{x})$ if and only if $\widehat{T}$ is calm at $(0, \tilde{x})$.

By expressing $\Omega_x(p)$ as

$$\Omega_x(p) := \Omega_x^1(p_1) \cap \Omega_x^2(p_2),$$

where

$$\Omega_x^1(p_1) := \{x \mid p_1 = Lx - \tilde{t}\} \quad \text{and} \quad \Omega_x^2(p_2) := \{x \mid p_2 \in D + Ax\}.$$

Before studying the metric subregularity of Γ_0, we present a lemma.

Lemma 44 *If $D \subseteq \text{range}\,(A)$, the multifunction $\Omega_x^2(p) := \{x \mid p \in D + Ax\}$ is calm at $(0, \bar{x})$ for any $\bar{x} \in \Omega_x^2(0) = \{x \mid 0 \in D + Ax\}$.*

Proof See Appendix M. ■

First, according to (Ye et al., 2018), $(\Omega_x^1)^{-1}$ is metrically subregular and pseudo-Lipschitz continuous at any point on its graph. Combining Lemma 44, Propositions 42 and 43, we obtain a sufficient condition for the metric subregularity of T.

Theorem 45 *Suppose that Assumption 1.2 is satisfied and A is of full row rank. Given $(\bar{x}, \bar{\lambda}) \in \Omega_{x,\lambda}^*$, if $\partial\theta_2$ is calm at $(\bar{\lambda}, -A\bar{x})$ with modulus κ_2 and*

$$\hat{\Omega}_x(p_1) := \{x \mid p_1 = Lx - \bar{t},\ 0 \in D + Ax\} \tag{50}$$

is calm at $(0, \bar{x})$ with modulus κ, then T is metrically subregular at $(\bar{x}, \bar{\lambda}, 0)$ with modulus

$$\kappa_T = \max\{\frac{1}{\|A\|}, \bar{\kappa}\},$$

where

$$\bar{\kappa} = \left(\frac{c_1 + \sqrt{c_1^2 + 4c_2}}{2}\right)^2 > 0.$$

In particular,

$$c_1 = \kappa_1(\sigma_{\min}(A^T) + (1+\kappa_2)L_h\|L\|)/(\sqrt{\sigma}\sigma_{\min}(A^T)),\ c_2 = \sqrt{2}\kappa_1,$$

$$\kappa_1 = \max\{\hat{\kappa}, 1\},\ \hat{\kappa} = (1+2\kappa\|L\|)\max\{\frac{1}{\tilde{\sigma}_{\min}(L)}, \frac{1}{\tilde{\sigma}_{\min}(A)}\},$$

where σ and L_h are the strong convexity modulus of h and Lipschitz continuity constant of ∇h on $\{Lx \mid (x,\lambda) \in \mathbb{B}_\epsilon(\bar{x},\bar{\lambda})\}$ for some $\epsilon > 0$, respectively, where $\tilde{\sigma}_{\min}(L)$ and $\tilde{\sigma}_{\min}(A)$ denotes the smallest nonzero singular value of L and A, respectively.

Proof The metric subregularity of T follows directly from Lemma 44, Propositions 42 and 43. We next estimate the metric subregularity modulus of T. Firstly, according to the proof of Lemma 44, we understand that Ω_x^2 is calm at $(0,\bar{x})$ with modulus $\frac{1}{\tilde{\sigma}_{\min}(A)}$, where $\tilde{\sigma}_{\min}(A)$ denotes the smallest nonzero singular value of A. Inspired by the proof of (Ye et al., 2018, Theorem 7), according to the calm intersection theorem, we shall estimate the calmness modulus of Ω_x in terms of the calmness modulus of $\hat{\Omega}_x$, i.e., Ω_x is calm at $(0,\bar{x})$ with modulus

$$\hat{\kappa} = (1+2\kappa\|L\|)\max\{\frac{1}{\tilde{\sigma}_{\min}(L)}, \frac{1}{\tilde{\sigma}_{\min}(A)}\}.$$

Immediately, Γ_0 is metrically subregular at $(\bar{x},\bar{u},0)$ with modulus $\kappa_1 = \max\{\hat{\kappa},1\}$. Thanks to the essentially locally strongly convexity of h and the locally Lipschitz continuity of ∇h, without loss of generality, we shall assume that ϵ is small enough so that ∇h is strongly monotone and Lipschitz continuous on $\{Lx \mid (x,\lambda) \in \mathbb{B}_\epsilon(\bar{x},\bar{\lambda})\}$ with modulus σ and L_h, respectively. Thanks to the proof of Propositions 42, T is metrically subregular at $(\bar{x},\bar{\lambda},0)$ with modulus $\kappa_T = \max\{\frac{1}{\|A\|}, \bar{\kappa}\}$, where

$$\bar{\kappa} = \left(\frac{c_1 + \sqrt{c_1^2 + 4c_2}}{2}\right)^2 > 0.$$

In particular, $c_1 = \kappa_1(\sigma_{\min}(A^T) + (1+\kappa_2)L_h\|L\|)/(\sqrt{\sigma}\sigma_{\min}(A^T))$ and $c_2 = \sqrt{2}\kappa_1$. ■

3.4.3. Verifying Metric Subregularity of T under Structured Assumptions

As long as ∂g is a polyhedral multifunction, $\partial\theta_2$ and hence Γ_{PDHG} are polyhedral multifunctions as well. Therefore, Proposition 40, together with Theorem 45, straightforwardly yields the following criteria for the metric subregularity of T.

Theorem 46 *The metric subregularity of T at $(\bar{x},\bar{\lambda},0)$ where $(\bar{x},\bar{\lambda}) \in \Omega^*_{x,\lambda}$ holds if one of the following statements is satisfied:*

(1) Problem (1) meets the structured polyhedricity assumption;

(2) Problem (1) meets the structured subregularity assumption at $(\bar{x},\bar{y},\bar{\lambda})$ which is a KKT point, and A is of full row rank.

Before giving applications of the criteria in Theorem 46, we need the following lemma.

Lemma 47 *Assume that ∂g is metrically subregular for some $(\bar{y}, \bar{v}) \in gph\, \partial g$ with modulus κ. Then*

(1) ∂g^* is calm at $(\bar{v}, \bar{y}) \in gph\, \partial g^*$ with modulus κ;

(2) for any matrix B, let $\bar{z}$ be any vector satisfying $B^T\bar{z} = \bar{v}$, then $B\partial g^* B^T$ is calm at $(\bar{z}, B\bar{y})$ with modulus $\kappa\|B\|^2$;

(3) in addition, when $\mathrm{range}\,(B^T) \cap \mathrm{ri}(\mathrm{dom}\ g^*) \neq \varnothing$, *we have*

$$\partial\theta_2(\lambda) = B\partial g^*(B^T\lambda) - b,$$

and thus $\partial\theta_2$ is calm at $(\bar{z}, B\bar{y} - b)$ with modulus $\kappa\|B\|^2$.

Proof See Appendix N. ■

Let $t \in (0, +\infty)$ be given, we define the multi-function $\varphi_t : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ as

$$\varphi_t(x) := \Big(\mathrm{sign}(x_1)\cdot|x_1|^t; \cdots; \mathrm{sign}(x_n)\cdot|x_n|^t\Big).$$

Lemma 48 *(Zhu et al., 2018) Let g represent the $\ell_{1,q}$-norm regularizer, i.e., $g(x) := \sum_{J\in\mathcal{J}} w_J \|x_J\|_q$ with $q \in [1,2]$ where $\mathcal{J}$ is a non-overlapping partition of the index set $\{1, \cdots, n\}$, $w_J \geq 0$ for $J \in \mathcal{J}$.*

(1) For any fixed $s \in \mathbb{R}^n$, $(\partial g)^{-1}(s)$ is a polyhedral convex set if it is nonempty.

(2) ∂g is metrically subregular at any $(\bar{x}, \bar{s}) \in gph(\partial g)$, namely, there exists $\epsilon > 0$ such that for any $x \in \mathbb{B}_\epsilon(\bar{x})$,

$$\mathrm{dist}\left(x, (\partial g)^{-1}(\bar{s})\right) \leq \kappa_g \cdot \mathrm{dist}\left(\bar{s}, \partial g(x)\right), \tag{51}$$

where

$$\kappa_g := \max_{J\in\mathcal{J}}\{\kappa_J\} \text{ with } \kappa_J = \begin{cases} 1 & \text{if } w_J = 0, \\ 1 & \text{if } w_J > 0 \text{ and } \|\bar{s}_J\|_q < w_J, \\ \kappa_{J,1}\cdot\kappa_{J,2}\cdot w_J^{-1} & \text{if } w_J > 0 \text{ and } \|\bar{s}_J\|_q = w_J, \end{cases}$$

and $\kappa_{J,1}$ denotes the Lipschitz constant of $\varphi_{\frac{q}{p}}(\cdot)$ at $\mathbb{B}_\epsilon\left(\varphi_{\frac{p}{q}}(\bar{x}_J)\right)$, $\kappa_{J,2}$ denotes the supremum of $\|\varphi_{\frac{p}{q}}(\cdot)\|_q$ at $\mathbb{B}_\epsilon(\bar{x}_J)$.

Lemma 49 *(Zhu et al., 2018) Let g denote the sparse-group LASSO regularizer, i.e., $g(x) := \sum_{J\in\mathcal{J}} w_J \|x_J\|_2 + \mu\cdot\|x\|_1$ where $\mathcal{J}$ be a non-overlapping partition of the index set $\{1, \cdots, n\}$, $w_J \geq 0$ for $J \in \mathcal{J}$ and $\mu \geq 0$ be given parameters.*

(1) For any fixed $s \in \mathbb{R}^n$, $(\partial g)^{-1}(s)$ is a polyhedral convex set if it is nonempty.

(2) ∂g is metrically subregular at any $(\bar{x},\bar{s})\in gph(\partial g)$, i.e., there exist $\epsilon>0$ such that for any $x\in\mathbb{B}_\epsilon(\bar{x})$,

$$\operatorname{dist}\left(x,(\partial g)^{-1}(\bar{s})\right)\leq\kappa_g(\lambda)\cdot\operatorname{dist}\left(\bar{s},\partial g(x)\right). \tag{52}$$

where

$$\kappa_g(\mu):=\max_{J\in\mathcal{J}}\{\kappa_J(\mu)\}$$

with

$$\kappa_J(\mu)=\begin{cases}1 & \text{if } w_J=0,\\ 1 & \text{if } w_J>0 \text{ and } \|\mathcal{T}_\mu(\bar{s}_J)\|_q<w_J,\\ \kappa_{J,1}(\mu)\cdot\kappa_{J,2}(\mu)\cdot w_J^{-1} & \text{if } w_J>0 \text{ and } \|\mathcal{T}_\mu(\bar{s}_J)\|_q=w_J,\end{cases}$$

and $\kappa_{J,1}(\mu)$ denotes the Lipschitz constant of $\varphi_{\frac{q}{p}}(\cdot)$ at $\mathbb{B}_\epsilon\left(\varphi_{\frac{p}{q}}(\bar{x}_J)\right)$, $\kappa_{J,2}(\mu)$ denotes the supremum of $\|\varphi_{\frac{p}{q}}(\cdot)\|_q$ at $\mathbb{B}_\epsilon\left(\bar{x}_J\right)$.

Theorem 50 *Suppose that Assumption 1.2 is satisfied. Given $(\bar{x},\bar{\lambda})\in\Omega^*_{x,\lambda}$, we have the desired metric subregularity of T as follows*

(1) when g is a convex piecewise linear-quadratic function, then T is metrically subregular at $(\bar{x},\bar{\lambda},0)$;

(2) when A is of full row rank, and g represents the $\ell_{1,q}$-norm regularizer with $q\in[1,2]$, then T is metrically subregular at $(\bar{x},\bar{\lambda},0)$;

(3) when A is of full row rank, and g represents the sparse-group LASSO regularizer, then T is metrically subregular at $(\bar{x},\bar{\lambda},0)$;

(4) when A is of full row rank, and g represent the indicator function of a ball constraint, i.e., $g=\delta_{\mathbb{B}}(\cdot)$ and $B^T\bar{\lambda}\neq 0$, then T is metrically subregular at $(\bar{x},\bar{\lambda},0)$.

Proof The first assertion coincides with nothing but the structured polyhedricity assumption. We just focus on the others.

For Parts (2) and (3), according to Lemma 48, ∂g is metrically subregular at any point on its graph, since $\partial g^*=(\partial g)^{-1}$, ∂g^* is calm everywhere on its graph. As $\theta_2(\lambda)=g^*(B^T\lambda)$, and $0\in\operatorname{ri}(\operatorname{dom} g^*)$, we surely have $\operatorname{range}(B^T)\cap\operatorname{ri}(\operatorname{dom} g^*)\neq\varnothing$. Then, according to Lemma 47, $\partial\theta_2$ is also calm everywhere on its graph. Note that, for any fixed η, from Lemma 48, $\partial g^*(\eta)=(\partial g)^{-1}(\eta)$ is a convex polyhedral set, straightforwardly $\partial\theta_2(\eta)$ is convex polyhedral. Consequently, the structured subregularity assumption is satisfied everywhere on the KKT solution set automatically.

For Part (4), as $g=\delta_{\mathbb{B}}(\cdot)$, $\theta_2=g^*(B^Ty)$ with $g^*(z)=\|z\|$. Since $\partial g^*=\partial\|\cdot\|$ is calm everywhere on its graph, and $0\in\operatorname{ri}(\operatorname{dom} g^*)$, $\operatorname{range}(B^T)\cap\operatorname{ri}(\operatorname{dom} g^*)\neq\varnothing$, then by Lemma 47, $\partial\theta_2$ is also calm everywhere on its graph. Moreover, since $\partial\|\eta\|$ is a convex polyhedral set whenever $\eta\neq 0$, easily $\partial\theta_2$ is a polyhedral set if $\eta\neq 0$. As a consequence, the structured subregularity assumption is satisfied on the KKT solution $(\bar{x},\bar{y},\bar{\lambda})$ if $B^T\bar{\lambda}\neq 0$. ■

3.5. Calculus of Metric Subregularity Modulus of T

So far we have verified the metric subregularity of T for some popular applications in Theorem 50. We next focus on calculating the metric subregularity modulus of T. Noting that $D = \partial\theta_2(\bar{\lambda})$, Theorem 51 then follows directly from Lemma 47 and Theorem 45.

Theorem 51 *Suppose that Assumption 1.2 is satisfied and A is of full row rank. Given $(\bar{x}, \bar{\lambda}) \in \Omega^*_{x,\lambda}$, if ∂g is metrically subregular at $(\bar{y}, B^T\bar{\lambda})$ with modulus κ_g for some $\bar{y}$ such that $B\bar{y} = b - A\bar{x}$, $\partial\theta_2(\bar{\lambda}) = B\partial g^*(B^T\bar{\lambda}) - b$ and*

$$\hat{\Omega}_x(p_1) := \{x \mid p_1 = Lx - \tilde{t},\ -Ax \in B\partial g^{-1}(B^T\bar{\lambda}) - b\}$$

is calm at $(0, \bar{x})$ with modulus κ, then T is metrically subregular at $(\bar{x}, \bar{\lambda}, 0)$ with modulus

$$\kappa_T = \max\{\frac{1}{\|A\|}, \bar{\kappa}\},$$

where

$$\bar{\kappa} = \left(\frac{c_1 + \sqrt{c_1^2 + 4c_2}}{2}\right)^2 > 0.$$

In particular,

$$c_1 = \kappa_1(\sigma_{\min}(A^T) + (1+\kappa_g)L_h\|L\|)/(\sqrt{\sigma}\sigma_{\min}(A^T)),\ c_2 = \sqrt{2}\kappa_1,$$

$$\kappa_1 = \max\{\hat{\kappa}, 1\},\ \hat{\kappa} = (1 + 2\kappa\|L\|)\max\{\frac{1}{\tilde{\sigma}_{\min}(L)}, \frac{1}{\tilde{\sigma}_{\min}(A)}\},$$

where $\tilde{\sigma}_{\min}(L)$ and $\tilde{\sigma}_{\min}(A)$ denotes the smallest nonzero singular value of L and A, respectively, σ and L_h are the strong convexity modulus of h and Lipschitz continuity constant of ∇h on $\{Lx \mid (x, \lambda) \in \mathbb{B}_\epsilon(\bar{x}, \bar{\lambda})\}$ for some $\epsilon > 0$, respectively.

Thanks to Theorem 51, suppose that Assumption 1.2 is satisfied and A is of full row rank, once we know the metric subregularity modulus of ∂g and the calmness modulus of $\hat{\Omega}_x$, the modulus of the metric subregularity of T can be estimated. The essential difficulty is associated with the estimation of the calmness modulus of $\hat{\Omega}_x$. According to its definition in (50), under Assumption 1.2 and full row rank of A, $\hat{\Omega}_x$ represents a perturbed linear system on a convex polyhedral set for a wide range of applications, including scenarios where g denotes the LASSO, the elastic net, the fused LASSO, the OSCAR, the group LASSO and the sparse-group LASSO. Hence, the calmness modulus of $\hat{\Omega}_x$ is achievable through the Hoffman's error bound theory or its variant (see Ye et al., 2018, Lemma 8).

We next show how to calculate the calmness modulus on specific application problems. We take the variable selection in regularized logistic regression (RLR) as an illustrative example while the extension to other problems is purely technical and hence omitted.

Calculus of the metric subregularity modulus of T for the ℓ_1 RLR: we consider the RLR problem with ℓ_1 norm regularizer

$$\begin{aligned} \min_{x,y} \quad & \sum_j \left(-\log\left(L_j^T x\right) + \mathbb{b}_j L_j^T x\right) + \mu\|y\|_1 \\ s.t. \quad & x = y, \end{aligned} \tag{53}$$

where $L \in \mathbb{R}^{l_1 \times m}$, and $\mathbb{b} \in \mathbb{R}^{l_1}_+$ are predefined matrices and vectors.

Denote that $g(y) = \mu\|y\|_1$, $\mu > 0$. Suppose the reference point we are considering is $(\bar{x}, \bar{\lambda})$. We may let $\bar{y} = \bar{x}$, then according to (Ye et al., 2018, Lemma 4, Lemma 5), $\partial g(y)$ is metrically subregular at $(\bar{y}, -\bar{\lambda})$ with modulus $\kappa_g = \frac{\kappa_{-\bar{\lambda}/\mu}}{\mu}$, where $\kappa_{-\bar{\lambda}/\mu}$ is the metric subregularity modulus of $\partial\|\cdot\|_1$ at $(\bar{y}, -\bar{\lambda}/\mu)$. Therefore, thanks again to (Ye et al., 2018, Lemma 4, Lemma 5),

$$\kappa_g \leq \frac{2\|\bar{y}\|}{\mu(1-\bar{c})},$$

$$\bar{c} = \max_{\{i:|-\bar{\lambda}_i/\mu|<1\}} |-\bar{\lambda}_i/\mu|; \quad \bar{c} = 0 \text{ if } \{i : |-\bar{\lambda}_i/\mu|<1\} = \emptyset.$$

In order to calculate the metric subregularity modulus of T, according to Theorem 51, we are left to estimate the calmness modulus of $\hat{\Omega}_x$. Again under the setting that $g(y) = \mu\|y\|_1$ for some $\mu > 0$, given $\bar{\lambda}$, we shall define index sets

$$\begin{aligned} I_+ &:= \{i \in \{1, \ldots, m\} \mid \bar{\lambda}_i = \mu\}, \\ I_- &:= \{i \in \{1, \ldots, m\} \mid \bar{\lambda}_i = -\mu\}, \\ I_0 &:= \{i \in \{1, \ldots, m\} \mid |\bar{\lambda}_i| < \mu\}. \end{aligned}$$

Moreover, we shall need the following notations.

- $e_i \in \mathbb{R}^m$ denotes the vector whose ith entry is 1 and other entries are zero,
- $D \in \mathbb{R}^{m \times (|I_+|+|I_-|)}$ denotes a matrix whose columns are $\{-e_i\}_{i\in I_+} \cup \{e_i\}_{i\in I_-}$.

$\hat{\Omega}_x$ can be rewritten as a partially perturbed system of linear equality and inequality constraints:

$$\hat{\Omega}_x(p_1) := \{x \mid p_1 = Lx - L\bar{x},\ -x = -D\alpha, \alpha \geq 0\} \tag{54}$$

We are in the position to apply Lemma 52 taken from (Ye et al., 2018) to calculate the calmness modulus of $\hat{\Omega}_x$. In fact, Lemma 52 can be regarded as a variant of Hoffman's error bound theory.

Lemma 52 (Partial error bound over a convex cone) *Let P be a polyhedral set $P := \{x \in \mathbb{R}^n \mid \tilde{A}x = \tilde{b},\ \tilde{K}x + \tilde{c} \in \mathcal{D}\}$, where $\tilde{A}$ is a matrix of size $m \times n$, $\tilde{K}$ is a matrix of size $p \times n$, $\tilde{b} \in \mathbb{R}^m$, $\tilde{c} \in \mathbb{R}^p$, $\mathcal{D} := \{z \mid z = \sum_{i=1}^l \alpha_i d_i, \alpha_i \geq 0\}$, and $\{d_i\}_{i=1}^l \subseteq \mathbb{R}^p$. Then*

$$dist\,(x, P) \leq \bar{\theta}(\mathcal{M}) \left\| \tilde{A}x - \tilde{b} \right\|, \quad \forall x \in \mathcal{D},$$

where $\mathcal{M} := \begin{bmatrix} \tilde{A}^T & -\tilde{K}^T & 0 \\ 0 & \tilde{D}^T & -I \end{bmatrix}$, I and 0 are identity and zero matrices of appropriate order, $\tilde{D} \in \mathbb{R}^{p\times l}$ is the matrix whose columns are $\{d_i\}_{i=1}^l$ and

$$\bar{\theta}(\mathcal{M}) := \sup_{\lambda,\mu,\nu} \left\{ \|\lambda\| \;\middle|\; \begin{array}{l} \|\mathcal{M}(\lambda, \mu, \nu)\| = 1, \nu \geq 0, \\ \textit{The corresponding rows of } \mathcal{M} \textit{ to } \lambda, \mu, \nu\textit{'s} \\ \textit{non-zero elements are linearly independent.} \end{array} \right\}. \tag{55}$$

Applying Lemma 52 to $\hat{\Omega}_x$ in (54), we obtain the following result.

Proposition 53 *For the RLR problem* (53), $\hat{\Omega}_x$ *is globally calm with modulus* $\bar{\theta}(\mathcal{M})$, *i.e.,*

$$dist\left(x, \hat{\Omega}_x(0)\right) \leq \bar{\theta}(\mathcal{M})\, dist\left(0, (\hat{\Omega}_x)^{-1}(x)\right), \quad \forall\, x,$$

where

$$\mathcal{M} := \begin{bmatrix} L^T & I & 0 \\ 0 & -D^T & -I \end{bmatrix},$$

and $\bar{\theta}(\mathcal{M})$ *is defined as in* (55).

Theorem 54 *Consider the RLR problem* (53). *Suppose that* $-\log$ *is strongly convex on some neighborhood* U_j *of* $L_j^T\bar{x}$ *for each* j *with uniform modulus* σ *and* $\nabla(-\log)$ *is Lipschitz continuous on* U_j *for each* j *with uniform constant* L_h, *then* T *is metrically subregular at* $(\bar{x}, \bar{\lambda}, 0)$ *with modulus* $\kappa_T = \max\{1, \bar{\kappa}\}$, *where*

$$\bar{\kappa} = \left(\frac{c_1 + \sqrt{c_1^2 + 4c_2}}{2}\right)^2 > 0.$$

In particular,

$$c_1 = \kappa_1(1 + (1+\kappa_g)L_h\|L\|)/\sqrt{\sigma},\ c_2 = \sqrt{2}\kappa_1,$$

$$\kappa_1 = \max\{\hat{\kappa}, 1\},\ \hat{\kappa} = (1 + 2\bar{\theta}(\mathcal{M})\|L\|)\max\{\frac{1}{\tilde{\sigma}_{\min}(L)}, 1\},$$

where $\tilde{\sigma}_{\min}(L)$ *denotes the smallest nonzero singular value of* L, $\kappa_g = \frac{2\|\bar{x}\|}{\mu(1-\bar{c})}$ *with*

$$\bar{c} = \max_{\{i:|-\bar{\lambda}_i/\mu|<1\}} |-\bar{\lambda}_i/\mu|; \quad \bar{c} = 0 \ \text{ if } \{i : |-\bar{\lambda}_i/\mu|<1\} = \emptyset.$$

3.6. Transporting the Convergence from PDHG to Linearized ADMM with Quantifiable Linear Convergence Rate

Based on the analysis in the preceding subsections for the linear convergence of the PDHG, we are able to convert the result to derive the linear convergence of the linearized ADMM. Let $\{(x^k, y^k, \lambda^k)\}$ be the sequence generated by the linearized ADMM. Then, according to Proposition36, $\{(x^k, \lambda^{k-1})\}$ converges to some point $(\bar{x}, \bar{\lambda}) \in \Omega^*_{x,\lambda}$. We next show the linear convergence of the linearized ADMM in sense of the sequences $\{(x^k, \lambda^{k-1})\}$, $\{(x^k, \lambda^k)\}$, $\{\text{Res}^k\}$, $\{\text{Val}^k\}$ and $\{\text{Fea}^k\}$.

Theorem 55 *If* T *is metrically subregular at* $(\bar{x}, \bar{\lambda}, 0)$ *with modulus* κ, *then both the sequence* $\{(x^k, \lambda^{k-1})\}$ *and* $\{(x^k, \lambda^k)\}$ *converge to* $\Omega^*_{x,\lambda}$ *linearly. That is, there exist* $k_0 > 0$, $C_0 > 0$ *and*

$$0 < \rho = \sqrt{\frac{\kappa^2}{1+\kappa^2}} < 1$$

such that, for all $k \geq k_0$, *it holds that*

$$\text{dist}\left((x^{k+1}, \lambda^k), \Omega^*_{x,\lambda}\right) \leq \rho\text{dist}\left((x^k, \lambda^{k-1}), \Omega^*_{x,\lambda}\right),$$

$$\mathrm{dist}\left((x^k,\lambda^k),\Omega^*_{x,\lambda}\right)\le 2C_0\rho^k,$$

and

$$Fea(x^k,y^k,\lambda^k)=\|Ax^k+By^k-b\|\le\frac{C_0}{\beta}\rho^k.$$

Furthermore, there exist $\tilde{k}_0>0$, $\tilde{C}_0>0$, and $\hat{C}_0>0$ such that, for all $k\ge\tilde{k}_0$, it holds that

$$Res(x^k,y^k,\lambda^k)\le\tilde{C}_0\rho^k,$$

and

$$|\,Val(x^k,y^k,\lambda^k)-Val^*|\le\hat{C}_0\rho^k.$$

Proof From Theorem 37, we know that there exist $k_0>0$ and $0<\rho=\sqrt{\frac{\kappa^2}{1+\kappa^2}}<1$ such that, for all $k\ge k_0$, it holds that

$$\mathrm{dist}_{\mathcal{M}}\left((x^{k+1},\lambda^k),\Omega^*_{x,\lambda}\right)\le\rho\mathrm{dist}_{\mathcal{M}}\left((x^k,\lambda^{k-1});\Omega^*_{x,\lambda}\right),\tag{56}$$

and there exists $C_0>0$ such that, for all $k\ge k_0$, it holds that

$$\begin{aligned}\mathrm{dist}\left((x^k,\lambda^{k-1}),\Omega^*_{x,\lambda}\right)&\le C_0\rho^k,\\ \|x^{k+1}-x^k\|+\|\lambda^k-\lambda^{k-1}\|&\le C_0\rho^k,\end{aligned}\tag{57}$$

and therefore

$$\mathrm{dist}\left((x^k,\lambda^k),\Omega^*_{x,\lambda}\right)\le\mathrm{dist}\left((x^k,\lambda^{k-1}),\Omega^*_{x,\lambda}\right)+\|\lambda^k-\lambda^{k-1}\|\le 2C_0\rho^k.$$

Since $\lambda^{k+1}=\lambda^k-\beta(Ax^{k+1}+By^{k+1}-b)$, we have

$$\mathrm{Fea}(x^{k+1},y^{k+1},\lambda^{k+1})=\|Ax^{k+1}+By^{k+1}-b\|=\frac{1}{\beta}\|\lambda^{k+1}-\lambda^k\|\le\frac{C_0}{\beta}\rho^{k+1}.\tag{58}$$

From the optimality conditions of the subproblems of each iteration generated by the linearized ADMM, we have

$$T_{KKT}(x^{k+1},y^{k+1},\lambda^{k+1})=\begin{pmatrix}\beta(A^TBy^{k+1}-A^TBy^k)-r(x^{k+1}-x^k)\\0\\Ax^{k+1}+By^{k+1}-b\end{pmatrix}.$$

Thus, by (57), (58) and

$$By^{k+1}-By^k=\frac{1}{\beta}(\lambda^k-\lambda^{k+1})+\frac{1}{\beta}(\lambda^k-\lambda^{k-1})+A(x^k-x^{k+1}),$$

we know that, for all $k\ge k_0+1$, it holds that

$$\begin{aligned}\mathrm{Res}(x^{k+1},y^{k+1},\lambda^{k+1})&\le\beta\|A\|\|By^{k+1}-By^k\|+r\|x^{k+1}-x^k\|+\|Ax^{k+1}+By^{k+1}-b\|\\&\le\|A\|(\|\lambda^{k+1}-\lambda^k\|+\|\lambda^k-\lambda^{k-1}\|+\beta\|A\|\|x^{k+1}-x^k\|)\\&\quad+r\|x^{k+1}-x^k\|+\|Ax^{k+1}+By^{k+1}-b\|\\&\le\max(\beta\|A\|^2+r,\rho\|A\|+\|A\|+\rho/\beta)C_0\rho^k.\end{aligned}$$

Additionally, similar to the proof in Theorem 30, since

$$\beta(A^TBy^{k+1} - A^TBy^k) - r(x^{k+1} - x^k) + A^T\lambda^{k+1} \in \partial f(x^{k+1})$$

and

$$B^T\lambda^{k+1} \in \partial g(y^{k+1}),$$

for any $(x^*, y^*, \lambda^*) \in \Omega^*$, we have

$$\begin{aligned} f(x^*) + g(y^*) \geq f(x^{k+1}) + g(y^{k+1}) + \langle \lambda^{k+1}, b - Ax^{k+1} - By^{k+1}\rangle \\ + \beta\langle By^{k+1} - By^k, Ax^* - Ax^{k+1}\rangle - r\langle x^{k+1} - x^k, x^* - x^{k+1}\rangle \end{aligned} \tag{59}$$

Furthermore, since $A^T\lambda^* \in \partial f(x^*)$ and $B^T\lambda^* \in \partial g(y^*)$, we have

$$f(x^{k+1}) + g(y^{k+1}) \geq f(x^*) + g(y^*) + \langle \lambda^*, Ax^{k+1} + By^{k+1} - b\rangle. \tag{60}$$

Combining (25) and (26), we get

$$\begin{aligned} |f(x^{k+1}) + g(y^{k+1}) - f(x^*) - g(y^*)| \leq \max\{\|\lambda^{k+1}\|, \|\lambda^*\|\}\|Ax^{k+1} + By^{k+1} - b\| \\ + \beta\|Ax^{k+1} - Ax^*\|\|By^{k+1} - By^k\| \\ + r\|x^{k+1} - x^*\|\|x^{k+1} - x^k\|. \end{aligned} \tag{61}$$

According to (73) (see also He and Yuan, 2012b), fixing any $(\bar{x}, \bar{y}, \bar{\lambda}) \in \Omega^*$, for any k, $\{\|(x^k, \lambda^{k-1}) - (\bar{x}, \bar{\lambda})\|\}$ is bounded, and so is $\{\|Ax^{k+1} - A\bar{x}\|\}$. Note that

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

Hence, there exists $C_1 > 0$ such that

$$|f(x^{k+1}) + g(y^{k+1}) - f(\bar{x}) - g(\bar{y})| \leq C_1(\|x^{k+1} - x^k\| + \|\lambda^{k+1} - \lambda^k\| + \|\lambda^k - \lambda^{k-1}\|).$$

According to (57), we obtain the linear convergence in terms of the objective function value of Problem (1) straightforwardly. ■

Let $\{(x^k, y^k, \lambda^k)\}$ be the sequence generated by the linearized ADMM. Theorems 50 and 55 motivate the following corollary directly.

Corollary 56 *Suppose Assumption 1.2 is satisfied. If one of the following statements is satisfied:*

(1) g is convex piecewise linear-quadratic function;

(2) A is of full row rank, and g represents the $\ell_{1,q}$-norm regularizer with $q \in [1, 2]$;

(3) A is of full row rank, and g represents the sparse-group LASSO regularizer;

(4) A is of full row rank, and g represent the indicator function of a ball constraint and $B^T\bar{\lambda} \neq 0$;

then both the sequence $\{(x^k, \lambda^{k-1})\}$ *and* $\{(x^k, \lambda^k)\}$ *converge to* $\Omega^*_{x,\lambda}$ *linearly. That is, there exist* $k_0 > 0$, $C_0 > 0$ *and computable* $0 < \rho < 1$ *such that, for all* $k \geq k_0$, *it holds that*

$$\mathrm{dist}\left((x^{k+1}, \lambda^k), \Omega^*_{x,\lambda}\right) \leq \rho \mathrm{dist}\left((x^k, \lambda^{k-1}), \Omega^*_{x,\lambda}\right),$$

$$\mathrm{dist}\left((x^k, \lambda^k), \Omega^*_{x,\lambda}\right) \leq 2C_0\rho^k,$$

and

$$Fea(x^k, y^k, \lambda^k) = \|Ax^k + By^k - b\| \leq \frac{C_0}{\beta}\rho^k.$$

Furthermore, there exist $\tilde{k}_0 > 0$, $\tilde{C}_0 > 0$, *and* $\hat{C}_0 > 0$ *such that for, all* $k \geq \tilde{k}_0$, *it holds that*

$$Res(x^k, y^k, \lambda^k) \leq \tilde{C}_0\rho^k$$

and

$$|\, Val(x^k, y^k, \lambda^k) - Val^*| \leq \hat{C}_0\rho^k.$$

4. Linear Convergence Rate of PADMM-FG

In the literature (Liu et al., 2018; Han et al., 2017; Yang and Han, 2016), linear convergence of the general PADMM-FG (2) is conceptually derived under the metric subregularity of T_{KKT}. It is noticed that essentially only the full polyhedral case (S4), in which the metric subregularity of T_{KKT} is trivially fulfilled, is discussed therein. As mentioned, the essential difficulty is how to verify the desired metric subregularity. In Theorem 60, we show the rather surprising fact that the metric subregularity of T is equivalent to that of T_{KKT} when B is of full column rank. This interesting observation allows us to apply all the established results known for the linearized ADMM to the general PADMM-FG (2). Indeed, by this line of analysis, in this section, we show that the subregularity conditions of T_{KKT} can be verified and thus the linear convergence of the PADMM-FG (2) in sense of $\{(x^k, y^k, \lambda^k)\}$, $\{\mathrm{Res}^k\}$ and $\{(\mathrm{Fea}^k, \mathrm{Val}^k)\}$ can be guaranteed for a wide range of applications including the RLR model (7), the $\ell_{1,q}$-norm regularized regression with $1 \leq q \leq 2$ (9) and sparse-group LASSO (10).

4.1. Linear Convergence of PADMM-FG under Metric Subregularity of T_{KKT}

The linear convergence of PADMM-FG (2) is shown in (Han et al., 2017, Theorem 2) when the metric subregularity of T^p_{KKT} defined in (6) is assumed at the limit point of the sequence. According to the equivalence between the metric subregularity of T^p_{KKT} and T_{KKT} proved in (Liu et al., 2018), we have the following result.

Theorem 57 *When* $\beta A^T A + G_1 \succ 0$ *and* $\beta B^T B + G_2 \succ 0$, *there exists* $(\bar{x}, \bar{y}, \bar{\lambda}) \in \Omega^*$ *with* Ω^* *being the set consisting of KKT points of Problem (1) such that the sequence* $\{(x^k, y^k, \lambda^k)\}$ *generated by PADMM-FG converges to* $(\bar{x}, \bar{y}, \bar{\lambda})$. *If, additionally, the multifunction* T_{KKT} *is metrically subregular at* $(\bar{u}, 0)$ *with modulus* c_{KKT}, *then the sequence* $\{(x^k, y^k, \lambda^k)\}$ *converges to* Ω^* *linearly. That is, there exist* $\tilde{M} \succ 0$, $k_0 > 0$ *and* $0 < \rho < 1$ *such that, for all* $k \geq k_0$, *it holds that*

$$\mathrm{dist}^2_{\tilde{M}}\left((x^{k+1}, y^{k+1}, \lambda^{k+1}), \Omega^*\right) + \|y^{k+1} - y^k\|^2_{G_2} \leq \rho\left[dist^2_{\tilde{M}}\left((x^k, y^k, \lambda^k), \Omega^*\right) + \|y^k - y^{k-1}\|^2_{G_2}\right],$$

where the explicit expression of ρ which is characterized in terms of c_{KKT} can be found in (Han et al., 2017, Theorem 2). Furthermore, there exist $\tilde{k}_0 > 0$, $C_0 > 0$, $\tilde{C}_0 > 0$, and $\hat{C}_0 > 0$ such that, for all $k \geq \tilde{k}_0$, it holds that

$$Fea(x^k, y^k, \lambda^k) \leq C_0 \rho^k,$$

$$Res(x^k, y^k, \lambda^k) \leq \tilde{C}_0 \rho^k,$$

and

$$|Val(x^k, y^k, \lambda^k) - Val^*| \leq \hat{C}_0 \rho^k.$$

4.2. Verification of Metric Subregularity of T_{KKT}

We understand that PADMM-FG (2) converges linearly under the metric subregularity of T_{KKT}. We next answer the question when T_{KKT} satisfies the metric subregularity.

4.2.1. Metric Subregularity of T_{KKT} under Structured Polyhedricity Assumption

We first verify the metric subregularity of T_{KKT} under the structured polyhedricity assumption. In fact, when Problem (1) meets Assumption 1.2, similar to Proposition 39, we can have following characterization of Ω^*,

$$\Omega^* = \{(x, y, \lambda) \mid Lx = \tilde{t},\ 0 = \tilde{g} - A^T\lambda,\ 0 \in \partial g(y) - B^T\lambda,\ 0 = Ax + By - b\}, \tag{62}$$

with some vector $\tilde{t} \in \mathbb{R}^l$ for which $Lx = \tilde{t}$ for all $x \in X$ and $\tilde{g} := L^T\nabla h(\tilde{t}) + q$. We introduce an auxiliary perturbed set-valued map with perturbation $p = (p_1, p_2, p_3, p_4)$ associated with the characterization (62):

$$\Gamma_{KKT}(p) := \{(x, y, \lambda) \mid p_1 = Lx - \tilde{t}, p_2 = \tilde{g} - A^T\lambda, p_3 \in \partial g(y) - B^T\lambda, p_4 = Ax + By - b\}.$$

Obviously, $\Gamma_{KKT}(p)$ coincides with Ω^* when $p = 0$. Highly similar to Proposition 40, we have following equivalence.

Proposition 58 *Assume that Assumption 1.2 is satisfied. Then the metric subregularity conditions of Γ_{KKT}^{-1} and T_{KKT} are equivalent. Precisely, given $(\bar{x}, \bar{y}, \bar{\lambda}) \in \Omega^*$, the following two statements are equivalent:*

(i) There exist $\kappa_1, \epsilon_1 > 0$ such that

$$\mathrm{dist}\left((x, y, \lambda), \Gamma_{KKT}(0)\right) \leq \kappa_1 \mathrm{dist}\left(0, \Gamma_{KKT}^{-1}(x, y, \lambda)\right), \quad \forall (x, y, \lambda) \in \mathbb{B}_{\epsilon_1}(\bar{x}, \bar{y}, \bar{\lambda}).$$

(ii) There exist $\kappa_2, \epsilon_2 > 0$ such that

$$\mathrm{dist}\left((x, y, \lambda), \Omega^*\right) \leq \kappa_2 \mathrm{dist}\left(0, T_{KKT}(x, y, \lambda)\right), \qquad \forall (x, y, \lambda) \in \mathbb{B}_{\epsilon_2}(\bar{x}, \bar{y}, \bar{\lambda}).$$

Proof Given any $(\bar{x}, \bar{y}, \bar{\lambda}) \in \Omega^*$. Suppose that there exist $\kappa_1, \epsilon_1 > 0$ such that

$$\mathrm{dist}\left((x, y, \lambda), \Gamma_{KKT}(0, 0)\right) \leq \kappa_1 \mathrm{dist}\left(0, \Gamma_{KKT}^{-1}(x, y, \lambda)\right),\ \forall (x, y, \lambda) \in \mathbb{B}_{\epsilon_1}(\bar{x}, \bar{\lambda}).$$

Due to the essentially locally strongly convexity of h and the locally Lipschitz continuity of ∇h, without loss of generality, we assume that ϵ_1 is small enough so that ∇h is strongly monotone and Lipschitz continuous on $\{Lx \mid (x,y,\lambda) \in \mathbb{B}_{\epsilon_1}(\bar{x},\bar{y},\bar{\lambda})\}$. For any $(x,y,\lambda) \in \mathbb{B}_{\epsilon_1}(\bar{x},\bar{y},\bar{\lambda})$, and any $(\xi,\eta,\zeta) \in T_{KKT}(x,y,\lambda)$

$$\xi = \partial f(x) - A^T\lambda = L^T\nabla h(Lx) + q - A^T\lambda, \tag{63}$$

$$\eta \in \partial g(y) - B^T\lambda, \tag{64}$$

$$\zeta = Ax + By - b, \tag{65}$$

and by the local Lipschitz continuity of ∇h, there exists $L_h > 0$ such that

$$\begin{aligned}
\operatorname{dist}\left((x,y,\lambda),\Omega^*\right) &= \operatorname{dist}\left((x,y,\lambda),\Gamma_{KKT}(0)\right) \\
&\le \kappa_1 \operatorname{dist}\left(0,\Gamma_{KKT}^{-1}(x,y,\lambda)\right) \\
&\le \kappa_1 \left(\|Lx-\tilde{t}\| + \|\xi+\tilde{g}-L^T\nabla h(Lx)-q\| + \|\eta\| + \|\zeta\|\right) \\
&\le \kappa_1 \left(\|Lx-\tilde{t}\| + \|L\|\|\nabla h(\tilde{t})-\nabla h(Lx)\| + \|\xi\| + \|\eta\| + \|\zeta\|\right) \\
&\le \kappa_1 \left((1+\|L\|L_h)\|Lx-\tilde{t}\| + \|\xi\| + \|\eta\| + \|\zeta\|\right).
\end{aligned} \tag{66}$$

Let $(\hat{x},\hat{y},\hat{\lambda})$ be the projection of (x,y,λ) on Ω^* and then $(\hat{x},\hat{y},\hat{\lambda}) \in \mathbb{B}_{\epsilon_1}(\bar{x},\bar{y},\bar{\lambda})$. Since $0 \in \partial g(\hat{y}) - B^T\hat{\lambda}$ and ∂g is monotone, we have

$$\langle \eta + B^T\lambda - B^T\hat{\lambda}, y - \hat{y}\rangle \ge 0,$$

and combining with (65) that,

$$\begin{aligned}
\langle \eta, y-\hat{y}\rangle &\ge -\langle B^T\lambda - B^T\hat{\lambda}, y-\hat{y}\rangle \\
&= -\langle \lambda-\hat{\lambda}, By - B\hat{y}\rangle \\
&= -\langle \zeta, \lambda-\hat{\lambda}\rangle + \langle \lambda-\hat{\lambda}, Ax - A\hat{x}\rangle.
\end{aligned}$$

Moreover, since $\xi = L^T\nabla h(Lx) + q - A^T\lambda$, $0 = L^T\nabla h(\tilde{t}) + q - A^T\hat{\lambda}$, thanks to the local strong convexity of h, there exists $\sigma > 0$ such that

$$\begin{aligned}
\langle \xi, x-\hat{x}\rangle + \langle \eta, y-\hat{y}\rangle + \langle \zeta, \lambda-\hat{\lambda}\rangle &\ge \langle L^T\nabla h(Lx) - L^T\nabla h(\tilde{t}), x-\hat{x}\rangle \\
&= \langle \nabla h(Lx) - \nabla h(L\hat{x}), Lx - L\hat{x}\rangle \\
&\ge \sigma\|Lx - L\hat{x}\|^2 = \sigma\|Lx-\tilde{t}\|^2.
\end{aligned} \tag{67}$$

Combining (66) and (67), we obtain that

$$\operatorname{dist}((x,y,\lambda),\Omega^*) \le c_1\sqrt{\|(\xi,\eta,\zeta)\| \cdot \operatorname{dist}((x,y,\lambda),\Omega^*)} + c_2\|(\xi,\eta,\zeta)\|,$$

with $c_1 = \kappa_1(1+\|L\|L_h)/\sqrt{\sigma}, c_2 = \sqrt{3}\kappa_1$, and consequently,

$$\operatorname{dist}((x,y,\lambda),\Omega^*) \le \tilde{\kappa}\|(\xi,\eta,\zeta)\|, \quad \text{where } \tilde{\kappa} = \left(\frac{c_1 + \sqrt{c_1^2+4c_2}}{2}\right)^2 > 0.$$

Because (ξ,η,ζ) is arbitrarily chosen in $T_{KKT}(x,y,\lambda)$, we have

$$\operatorname{dist}\left((x,y,\lambda),T_{KKT}^{-1}(0)\right) = \operatorname{dist}((x,y,\lambda),\Omega^*) \le \tilde{\kappa}\operatorname{dist}\left(0,T_{KKT}(x,y,\lambda)\right).$$

Hence, there exists $\kappa_2 = \tilde{\kappa} > 0$ such that

$$\mathrm{dist}\left((x,y,\lambda),\Omega^*\right) \leq \kappa_2 \mathrm{dist}\left(0, T_{KKT}\left(x,y,\lambda\right)\right),\ \forall (x,y,\lambda) \in \mathbb{B}_{\epsilon_1}(\bar{x},\bar{y},\bar{\lambda}).$$

Conversely, given any $(\bar{x},\bar{y},\bar{\lambda}) \in \Omega^*$, suppose that there exist $\kappa_2, \epsilon_2 > 0$ such that

$$\mathrm{dist}\left((x,y,\lambda),\Omega^*\right) \leq \kappa_2 \mathrm{dist}\left(0, T_{KKT}\left(x,y,\lambda\right)\right),\ \forall (x,y,\lambda) \in \mathbb{B}_{\epsilon_2}(\bar{x},\bar{y},\bar{\lambda}).$$

For any fixed $(x,y,\lambda) \in \mathbb{B}_{\epsilon_2}(\bar{x},\bar{y},\bar{\lambda})$, and $(p_1,p_2,p_3,p_4) \in \Gamma^{-1}_{KKT}\left(x,y,\lambda\right)$, it follows that

$$p_1 = Lx - \tilde{t},\ p_2 = \tilde{g} - A^T\lambda,\ p_3 \in \partial g(y) - B^T\lambda,\ p_4 = Ax + By - b.$$

To summarize, we have

$$\begin{aligned} p_2 + L^T\nabla h(Lx) - L^T\nabla h(Lx - p_1) &= \partial f(x) - A^T\lambda, \\ p_3 &\in \partial g(y) - B^T\lambda, \\ p_4 &= Ax + By - b. \end{aligned}$$

By virtue of the locally Lipschitz continuity of ∇h, there exists $L_h > 0$ such that

$$\begin{aligned} \mathrm{dist}\left((x,y,\lambda),\Omega^*\right) &\leq \kappa_2 \mathrm{dist}\left(0, T_{KKT}\left(x,y,\lambda\right)\right) \\ &\leq \kappa_2 \left(\|p_2 + L^T\nabla h(Lx) - L^T\nabla h(Lx - p_1)\| + \|p_3\| + \|p_4\|\right) \\ &\leq \kappa_2 L_h \|L\|\|p_1\| + \kappa_2\|p_2\| + \kappa_2\|p_3\| + \kappa_2\|p_4\|. \end{aligned}$$

Moreover, since (p_1,p_2,p_3,p_4) can be any element in $\Gamma^{-1}_{KKT}\left(x,y,\lambda\right)$, we have

$$\mathrm{dist}\left((x,y,\lambda),\Gamma_{KKT}(0)\right) = \mathrm{dist}\left((x,y,\lambda),\Omega^*\right) \leq \kappa_2(L_h\|L\| + 3)\mathrm{dist}\left(0,\Gamma^{-1}_{KKT}\left(x,y,\lambda\right)\right).$$

Therefore, there exists $\kappa_1 = \kappa_2(L_h\|L\| + 4) > 0$ such that

$$\mathrm{dist}\left((x,y,\lambda),\Gamma_{KKT}(0)\right) \leq \kappa_1 \mathrm{dist}\left(0,\Gamma^{-1}_{KKT}(x,y,\lambda)\right),\ \forall (x,y,\lambda) \in \mathbb{B}_{\epsilon_2}(\bar{x},\bar{y},\bar{\lambda}),$$

which completes the converse direction. ■

As a consequence, when Problem (1) meets the structured polyhedricity assumption, Γ_{KKT} is obvisouly a polyhedral multifunction, we obtain following result directly.

Theorem 59 *The metric subregularity of T_{KKT} at $(\bar{x},\bar{y},\bar{\lambda},0)$ where $(\bar{x},\bar{y},\bar{\lambda}) \in \Omega^*$ holds if Problem (1) meets the structured polyhedricity assumption.*

4.2.2. Metric Subregularity of T_{KKT} under Structured Subregularity Assumption

We next justify the metric subregularity of T_{KKT} under the structured subregularity assumption. To this end, we shall clarify the relationship between the metric subregularity of T_{KKT} and metric subregularity of T. Therefore, this connection, together with the characterization for the metric subregularity of T, will serve as a sufficient condition to justify the metric subregularity of T_{KKT}.

Theorem 60 *For any point $(\bar{x}, \bar{\lambda}) \in T^{-1}(0)$, if there exists $\bar{y} \in \partial g^*(B^T\bar{\lambda})$ such that T_{KKT} is metrically subregular at $(\bar{x}, \bar{y}, \bar{\lambda}, 0)$, then T is metrically subregular at $(\bar{x}, \bar{\lambda}, 0)$. Additionally, if B is of full column rank, for any KKT point $(\bar{x}, \bar{y}, \bar{\lambda}) \in (T_{KKT})^{-1}(0)$ and T is metrically subregular at $(\bar{x}, \bar{\lambda}, 0)$ with modulus κ, then T_{KKT} is metrically subregular at $(\bar{x}, \bar{y}, \bar{\lambda}, 0)$ with modulus*

$$c_{KKT} = \kappa(2 + \frac{\|A\|}{\sigma_{\min}(B)})^2 + \frac{2}{\sigma_{\min}(B)}.$$

Proof For any (x, λ) such that $p \in T(x, \lambda)$, that is,

$$\begin{cases} p_1 \in \partial f(x) - A^T\lambda, \\ p_2 \in B\partial g^*(B^T\lambda) - b + Ax, \end{cases}$$

there exists $y \in \partial g^*(B^T\lambda)$ such that $p_2 = By - b + Ax$. Taking into consideration the fact that $y \in \partial g^*(B^T\lambda)$ if and only if $B^T\lambda \in \partial g(y)$, we have

$$\begin{cases} p_1 \in \partial f(x) - A^T\lambda, \\ 0 \in \partial g(y) - B^T\lambda, \\ p_2 = Ax + By - b. \end{cases}$$

Apparently, T is metrically subregular at $(\bar{x}, \bar{\lambda}, 0)$ provided the metric subregularity of T_{KKT} at $(\bar{x}, \bar{y}, \bar{\lambda}, 0)$.

Suppose B is of full column rank,. We next show that the metric subregularity of T implies that of T_{KKT}. In fact, for (x, y, λ) such that $p \in T_{KKT}(x, y, \lambda)$, i.e.,

$$\begin{cases} p_1 \in \partial f(x) - A^T\lambda, \\ p_2 \in \partial g(y) - B^T\lambda, \\ p_3 = Ax + By - b, \end{cases}$$

and because of $p_2 \in \partial\theta_2(y) - B^T\lambda$, we have

$$0 \in \partial g^*(B^T\lambda + p_2) - y,$$

and hence that

$$0 \in B\partial g^*(B^T\lambda + p_2) - By.$$

Combining with $p_3 = Ax + By - b$, we get

$$p_3 \in B\partial g^*(B^T\lambda + p_2) - b + Ax.$$

Since B is of full column rank, $\tilde{p}_2 := B(B^TB)^{-1}p_2$ is well defined and it satisfies $B^T\tilde{p}_2 = p_2$. Denoting $\tilde{\lambda} = \lambda + \tilde{p}_2$, we have

$$\begin{aligned} p_1 - A^T\tilde{p}_2 &\in \partial f(x) - A^T\tilde{\lambda}, \\ p_3 &\in B\partial g^*(B^T\tilde{\lambda}) - b + Ax, \end{aligned}$$

that is

$$(p_1 - A^T\tilde{p}_2, p_3) \in T(x, \tilde{\lambda}).$$

By virtue of the metric subregularity of T at $(\bar{x}, \bar{\lambda}, 0)$, there exist $\kappa, \epsilon > 0$ such that

$$\text{dist}\left((x,\lambda), \Omega^*_{x,\lambda}\right) \leq \kappa \text{dist}\left(0, T\left(x,\lambda\right)\right), \qquad \forall (x,y) \in \mathbb{B}_\epsilon(\bar{x}, \bar{\lambda}).$$

We now assume that $(x, y, \lambda) \in \mathbb{B}_{\epsilon_1}(\bar{x}, \bar{y}, \bar{\lambda})$ with

$$\epsilon_1 = \epsilon/2, \quad \|p_2\| \leq \epsilon/(2\|B(B^TB)^{-1}\|) = \epsilon\sigma_{\min}(B)/2$$

where $\sigma_{\min}(B)$ denotes the smallest nonzero singular value of B. Then, $\|\tilde{p}_2\| \leq \epsilon$ and $\|\tilde{\lambda} - \bar{\lambda}\| \leq \|\lambda - \bar{\lambda}\| + \|\tilde{p}_2\| \leq \epsilon$. Thus, by the metric subregularity of T at $(\bar{x}, \bar{\lambda}, 0)$

$$\begin{aligned}
\text{dist}((x, \tilde{\lambda}), \Omega^*_{x,\lambda}) &\leq \kappa(\|p_1 - A^T\tilde{p}_2\| + \|p_3\|) \\
&\leq \kappa(\|p_1\| + \|A\|\|\tilde{p}_2\| + \|p_3\|) \\
&\leq \kappa(2 + \|A\|\|B(B^TB)^{-1}\|)\|p\| \\
&= \kappa\left(2 + \frac{\|A\|}{\sigma_{\min}(B)}\right)\|p\|.
\end{aligned}$$

Let (x_0, λ_0) be the projection of $(x, \tilde{\lambda})$ on $\Omega^*_{x,\lambda} := T^{-1}(0)$. Then with the full row rank of B^T, we have

$$0 \in \partial(g^*(B^T\lambda_0)) - b + Ax_0 = B\partial g^*(B^T\lambda_0) - b + Ax_0.$$

Thus, we can find $y_0 \in \partial g^*(B^T\lambda_0)$ such that $0 = By_0 - b + Ax_0$. Noting that $p_3 = Ax + By - b$, and $\sigma_{\min}(B) > 0$ which follows from the full column rank assumption of B, there holds that

$$\|y - y_0\| = \|(B^TB)^{-1}B^T(p_3 - A(x - x_0))\| \leq \frac{1}{\sigma_{\min}(B)}\|p_3\| + \frac{\|A\|}{\sigma_{\min}(B)}\|x - x_0\|.$$

Since $(x_0, y_0, \lambda_0) \in T^{-1}_{KKT}(0)$, we have

$$\begin{aligned}
\text{dist}((x,y,\lambda), T^{-1}_{KKT}(0)) &\leq \|x - x_0\| + \|\tilde{\lambda} - \lambda_0\| + \|\lambda - \tilde{\lambda}\| + \|y - y_0\| \\
&\leq (1 + \frac{\|A\|}{\sigma_{\min}(B)})\|x - x_0\| + \|\tilde{\lambda} - \lambda_0\| + \frac{1}{\sigma_{\min}(B)}\|p_2\| + \frac{1}{\sigma_{\min}(B)}\|p_3\| \\
&\leq (2 + \frac{\|A\|}{\sigma_{\min}(B)})\text{dist}((x, \tilde{\lambda}), \Omega^*_{x,\lambda}) + \frac{1}{\sigma_{\min}(B)}\|p_2\| + \frac{1}{\sigma_{\min}(B)}\|p_3\| \\
&\leq \kappa(2 + \frac{\|A\|}{\sigma_{\min}(B)})^2\|p\| + \frac{1}{\sigma_{\min}(B)}\|p_2\| + \frac{1}{\sigma_{\min}(B)}\|p_3\| \\
&\leq c_{KKT}\|p\|,
\end{aligned}$$

where the second inequality follows from $\tilde{\lambda} = \lambda + \tilde{p}_2$ and

$$c_{KKT} = \kappa(2 + \frac{\|A\|}{\sigma_{\min}(B)})^2 + \frac{2}{\sigma_{\min}(B)}.$$

■

Theorems 46 and 60 straightforwardly inspire the following criteria for the metric subregularity of T_{KKT}.

Theorem 61 *Provided the full column rank of B, the metric subregularity of T_{KKT} at $(\bar{x}, \bar{y}, \bar{\lambda}, 0)$ where $(\bar{x}, \bar{y}, \bar{\lambda}) \in \Omega^*$ holds if A is of full row rank, and Problem (1) meets the structured subregularity assumption at $(\bar{x}, \bar{y}, \bar{\lambda})$.*

Motivated by the proof in Theorem 50, the linear convergence of the general PADMM-FG (2) can be obtained easily.

Theorem 62 *Suppose that Assumption 1.2 is satisfied. Then the metric subregularity of T_{KKT} at $(\bar{x}, \bar{y}, \bar{\lambda}, 0)$ where $(\bar{x}, \bar{y}, \bar{\lambda}) \in \Omega^*$ holds, and hence the sequence $\{(x^k, y^k, \lambda^k)\}$ generated by the PADMM-FG (2) converges to Ω^* linearly if one of the following statements is satisfied:*

(1) g is convex piecewise linear-quadratic function;

(2) A is of full row rank and B is of full column rank, and g represents the $\ell_{1,q}$-norm regularizer with $q \in [1, 2]$;

(3) A is of full row rank and B is of full column rank, and g represents the sparse-group LASSO regularizer;

(4) A is of full row rank and B is of full column rank, and g represent the indicator function of a ball constraint and $B^T\bar{\lambda} \neq 0$.

We are left to calculate the metric subregularity modulus of T_{KKT} on specific applications. In fact, we have presented with illustrate examples how to calculate the metric subregularity modulus of T in Section 3.5. According to Theorem 60, the metric subregularity modulus of T_{KKT} is easily computable as long as the metric subregularity modulus of T is calculated on specific applications.

5. Conclusions

In this paper, we further discuss the linear convergence of the alternating direction method of multipliers (ADMM) and its variants for some structured convex optimization problems, and develop a rather complete methodology to discern the linear convergence for a wide range of concrete applications. Through the lens of variational analysis, we show that the linear convergence of ADMM and its variants can be guaranteed without the strong convexity of objective functions together with the full rank assumption of coefficient matrices, or the full polyhedricity of their subdifferentials. The understanding of linear convergence of the ADMM and its variants is thus substantially enhanced, and the scope of the ADMM with efficient performance in sense of guaranteed linear convergence is essentially broadened. Indeed, for a number of models arising in statistics and machine learning such as the RLR, PAC, $\ell_{1,q}$-norm with $q \in [1, 2]$ and sparse-group LASSO models, current results in the literature fail to explain why the ADMM and its variants perform linear convergence, and rigorous theory is provided for the first time. The gap between empirically observed numerical performance and checkable theoretical conditions is essentially filled in. Our techniques are entirely relied on variational analysis, and they are tailored for both special properties of the models and structures of the iterative schemes under investigation.

Acknowledgments

We extend our gratitude to the acting editor and three anonymous referees for their helpful suggestions and comments. The first author was supported by the General Research Fund 12302318 from Hong Kong Research Grants Council. The third author would like to acknowledge support from the National Science Foundation of China 11971220 and the Natural Science Foundation of Guangdong Province 2019A1515011152. The alphabetical order of the authors indicates the equal contribution to the paper.

Appendix A. Proof of Proposition 8

For the k-th iteration (x^k, y^k, λ^k) generated by the original ADMM, it follows from the optimality condition of the subproblems that

$$\begin{cases} 0 \in \partial g(y^k) - B^T\lambda^k, \\ 0 \in \partial f(x^{k+1}) - A^T(\lambda^k - \beta(Ax^{k+1} + By^k - b)), \\ \lambda^{k+1} = \lambda^k - \beta(Ax^{k+1} + By^{k+1} - b). \end{cases} \tag{68}$$

Since $(\partial f)^{-1} = \partial f^*$ and $(\partial g)^{-1} = \partial g^*$, we have

$$\begin{cases} y^k \in \partial g^*(B^T\lambda^k), \\ x^{k+1} \in \partial f^*(A^T(\lambda^k - \beta(Ax^{k+1} + By^k - b))), \\ \lambda^{k+1} = \lambda^k - \beta(Ax^{k+1} + By^{k+1} - b). \end{cases}$$

Furthermore, it is easy to see that

$$\begin{cases} \lambda^k + \beta By^k \in \lambda^k + \beta B\partial g^*(B^T\lambda^k) \subseteq (I + \beta\partial\phi_2)(\lambda^k), \\ \lambda^k - \beta By^k \in \lambda^k - \beta(Ax^{k+1} + By^k - b) + \beta A\partial f^*(A^T(\lambda^k - \beta(Ax^{k+1} + By^k - b))) - \beta b \\ \qquad\qquad \subseteq (I + \beta\partial\phi_1)(\lambda^k - \beta(Ax^{k+1} + By^k - b)), \\ \lambda^{k+1} + \beta By^{k+1} = \lambda^k + \beta By^k - \lambda^k + \lambda^k - \beta(Ax^{k+1} + By^k - b). \end{cases}$$

Therefore, we have

$$\begin{cases} \lambda^k = (I + \beta\partial\phi_2)^{-1}(\lambda^k + \beta By^k), \\ \lambda^k - \beta(Ax^{k+1} + By^k - b) = (I + \beta\partial\phi_1)^{-1}(\lambda^k - \beta By^k), \\ \lambda^{k+1} + \beta By^{k+1} = \lambda^k + \beta By^k - \lambda^k + \lambda^k - \beta(Ax^{k+1} + By^k - b). \end{cases} \tag{69}$$

Setting $u^k = \lambda^k$, $v^k = \lambda^k - \beta(Ax^{k+1} + By^k - b)$ and $z^k = \lambda^k + \beta By^k$, we get

$$\begin{cases} u^k = (I + \beta\partial\phi_2)^{-1}(z^k), \\ v^k = (I + \beta\partial\phi_1)^{-1}(2u^k - z^k), \\ z^{k+1} = z^k - u^k + v^k, \end{cases} \tag{70}$$

and the conclusion follows.

Appendix B. Proof of Theorem 14

According to the iterative scheme of the DRSM, at each iteration k, we have

$$z^k \in u^k + \partial\phi_2(u^k), \quad 2u^k - z^k \in v^k + \partial\phi_1(v^k),$$

and subsequently,

$$z^k - u^k \in \partial\phi_2(u^k), \quad 2u^k - z^k - v^k \in \partial\phi_1(v^k). \tag{71}$$

Summing these two equations, we have

$$u^k - v^k \in \partial\phi_1(u^k - (u^k - v^k)) + \partial\phi_2(u^k),$$

which implies

$$u^k \in \mathcal{T}_1(u^k - v^k).$$

Since $\mathcal{T}_1$ is calm at $(0, \bar{\lambda})$, there exist $\epsilon_1, \kappa_1 > 0$ such that

$$\mathrm{dist}(u^k, Z) \le \kappa_1 \|u^k - v^k\|, \qquad when\ u^k \in \mathbb{B}_{\epsilon_1}(\bar{\lambda}). \tag{72}$$

Also, since $u^k = \mathrm{prox}_{\phi_2}(z^k)$, and $\|u^k - \bar{\lambda}\| = \|\mathrm{prox}_{\phi_2}(z^k) - \mathrm{prox}_{\phi_2}(\bar{z})\| \le \|z^k - \bar{z}\|$, which comes from the nonexpansiveness of prox_{ϕ_2}, substituting $u^k - v^k = z^k - z^{k+1}$ in (72) enables us to obtain

$$\mathrm{dist}(\mathrm{prox}_{\phi_2}(z^k), Z) \le \kappa_1 \|z^{k+1} - z^k\|, \qquad when\ z^k \in \mathbb{B}_{\epsilon_1}(\bar{z}).$$

Again, by (71) and $u^k - v^k = z^k - z^{k+1}$, we have

$$z^k - u^k \in \partial\phi_2(u^k), \quad u^k - z^{k+1} \in \partial\phi_1(v^k).$$

and then

$$u^k \in \partial{\phi_2}^*(z^k - u^k), \quad v^k \in \partial{\phi_1}^*(u^k - z^{k+1}).$$

Combining the inclusions together, we have

$$z^k - z^{k+1} = u^k - v^k \in \partial({\phi_1}^* \circ -Id)(z^k - u^k - (z^k - z^{k+1})) + \partial{\phi_2}^*(z^k - u^k),$$

which implies

$$z^k - u^k \in \mathcal{T}_2(z^k - z^{k+1}).$$

Then, since $\mathcal{T}_2$ is calm at $(0, \bar{\mu})$, there exist $\epsilon_2, \kappa_2 > 0$ such that

$$\mathrm{dist}(z^k - u^k, W) \le \kappa_2 \|z^k - z^{k+1}\|, \qquad when\ z^k - u^k \in \mathbb{B}_{\epsilon_2}(\bar{\mu}).$$

Since $z^k - u^k = \mathrm{prox}_{{\phi_2}^*}(z^k)$, and

$$\|z^k - u^k - \bar{\mu}\| = \|\mathrm{prox}_{{\phi_2}^*}(z^k) - \mathrm{prox}_{{\phi_2}^*}(\bar{z})\| \le \|z^k - \bar{z}\|,$$

which comes from the nonexpansiveness of $\mathrm{prox}_{{\phi_2}^*}$, we have

$$\mathrm{dist}(\mathrm{prox}_{{\phi_2}^*}(z^k), W) \le \kappa_2 \|z^{k+1} - z^k\|, \qquad when\ z^k \in \mathbb{B}_{\epsilon_2}(\bar{z}).$$

Then, there exist $\epsilon, \kappa > 0$ such that, for all k satisfying $z^k \in \mathbb{B}(\bar{z}, \epsilon)$, we have

$$\mathrm{dist}\left(\mathrm{prox}_{\phi_2}(z^k), Z\right) + \mathrm{dist}\left(\mathrm{prox}_{{\phi_2}^*}(z^k), W\right) \le \kappa \left\| z^{k+1} - z^k \right\|,$$

where $\kappa = \kappa_1 + \kappa_2$. According to Proposition 12, the sequence $\{z^k\}$ converges to $\bar{z}$ linearly.

Appendix C. Proof of Lemma 15

For any $y \notin \text{range}\,(\mathbb{L}^T)+\mathcal{A}_0^\perp$, y can be expressed as $y = y_1+y_2$, where $y_1 \in \text{range}\,(\mathbb{L}^T)+\mathcal{A}_0^\perp$, $y_2 \in \text{null}\,(\mathbb{L}) \cap \mathcal{A}_0$ and $y_2 \neq 0$. Then, we have

$$\begin{aligned}
\psi^*(y) &= \sup_x\{\langle y, x\rangle - \mathbb{h}(\mathbb{L}x) - \delta_{\mathcal{A}}(x)\} \\
&= \sup_x\{\langle y_1 + y_2, x\rangle - \mathbb{h}(\mathbb{L}x) - \delta_{\mathcal{A}}(x)\} \\
&\geq \sup_\alpha\{\langle y_1 + y_2, a + \alpha y_2\rangle - \mathbb{h}(\mathbb{L}a + \alpha\mathbb{L}y_2) - \delta_{\mathcal{A}}(a + \alpha y_2)\} \\
&= \sup_\alpha\{\langle y_1 + y_2, a\rangle - \mathbb{h}(\mathbb{L}a) + \alpha\|y_2\|^2\} = +\infty.
\end{aligned}$$

That is, dom $\psi^* \subset \text{range}\,(\mathbb{L}^T) + \mathcal{A}_0^\perp$.

Appendix D. Proof of Proposition 18

Consider the singular value decomposition of matrix $\mathbb{L}$. Let $\mathbb{L} = U\Sigma V^T$, where $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$ and $\Sigma \in \mathbb{R}^{r\times r}, \Sigma \succ 0$. Let $Q_{\mathcal{A}_0}$ be the matrix whose columns are normal orthogonal basis of $\mathcal{A}_0$. Next, we consider the singular value decomposition of matrix $\mathbb{L}Q_{\mathcal{A}_0}$, i.e.,

$$\mathbb{L}Q_{\mathcal{A}_0} = U_{\mathbb{L}(\mathcal{A}_0)}\Sigma_{\mathbb{L}(\mathcal{A}_0)}V^T_{\mathbb{L}(\mathcal{A}_0)},$$

where $U_{\mathbb{L}(\mathcal{A}_0)} \in \mathbb{R}^{m\times r_1}$, $V_{\mathbb{L}(\mathcal{A}_0)} \in \mathbb{R}^{m_1\times r_1}$ and $\Sigma_{\mathbb{L}(\mathcal{A}_0)} \in \mathbb{R}^{r_1\times r_1}, \Sigma_{\mathbb{L}(\mathcal{A}_0)} \succ 0$.

Define further that $\tilde{\mathbb{h}} : \mathbb{R}^{r_1} \to \mathbb{R}$ as $\tilde{\mathbb{h}}(z) := \mathbb{h}(\mathbb{L}a + U_{\mathbb{L}(\mathcal{A}_0)}z)$. According to the assumptions that $\mathbb{h} \in \mathcal{C}$ and $U_{\mathbb{L}(\mathcal{A}_0)}$ is of full column rank, we observe that $\tilde{\mathbb{h}} \in \mathcal{C}$. Denoting the conjugate of $\tilde{\mathbb{h}}$ as

$$\tilde{\mathbb{h}}^*(z^*) := \sup_{z\in\mathbb{R}^r}\{\langle z^*, z\rangle - \tilde{\mathbb{h}}(z)\},$$

then, by virtue of Proposition 16, we have $\tilde{\mathbb{h}}^* \in \mathcal{C}$.

Next, for each vector y taken from $\text{range}\,(\mathbb{L}^T) + \mathcal{A}_0^\perp$, y admits a decomposition that

$$y = \text{Proj}_{\text{range}\,(\mathbb{L}^T)}y + (I - \text{Proj}_{\text{range}\,(\mathbb{L}^T)})y.$$

Denote $H := [\mathbb{L}^T, A_p]$ where $A_p \in \mathbb{R}^{n\times l}$ is the matrix whose columns are bases of $\mathcal{A}_0^\perp$ and l is the dimension of $\mathcal{A}_0^\perp$. Let $H = U_H\Sigma_H V_H^T$ be the singular value decomposition of matrix H, where $U_H \in \mathbb{R}^{n\times r_2}$, $V_H \in \mathbb{R}^{(m+l)\times r_2}$ and $\Sigma_H \in \mathbb{R}^{r_2\times r_2}$, $\Sigma_H \succ 0$, then $H^\dagger := V_H\Sigma_H^{-1}U_H^T$. Denote further that E_1 is the matrix whose rows are the first m rows of the identity matrix in $\mathbb{R}^{(m+l)\times(m+l)}$ and E_2 is the matrix whose rows are the last l rows of the identity matrix in $\mathbb{R}^{(m+l)\times(m+l)}$. We therefore have the decomposition of $y \in \text{range}\,(\mathbb{L}^T) + \mathcal{A}_0^\perp$ as

$$y = \tilde{y} + \hat{y},$$

where $\tilde{y} := \mathbb{L}^T F y \in \text{range}\,(\mathbb{L}^T)$ with

$$F := U\Sigma^{-1}V^T + E_1H^\dagger(I - \mathbb{L}^TU\Sigma^{-1}V^T),$$

and

$$\hat{y} := A_pE_2H^\dagger(I - \mathbb{L}^TU\Sigma^{-1}V^T)y \in \mathcal{A}_0^\perp.$$

Plugging in this decomposition of y into the conjugate of ψ^*,

$$\begin{aligned}
\psi^*(y) &= \sup_{x\in\mathbb{R}^n}\{\langle y,x\rangle - \mathbb{h}(\mathbb{L}x) - \delta_{\mathcal{A}}(x)\}\\
&= \sup_{x\in\mathbb{R}^n}\{\langle \tilde{y},x\rangle + \langle \hat{y},x\rangle - \mathbb{h}(\mathbb{L}x) - \delta_{\mathcal{A}}(x)\}\\
&= \sup_{x\in\mathbb{R}^n}\{\langle Fy,\mathbb{L}x\rangle + \langle \hat{y},x\rangle - \mathbb{h}(\mathbb{L}x) - \delta_{\mathcal{A}}(x)\}\\
&= \sup_{z\in\mathbb{R}^r}\{\langle \tilde{y},a\rangle + \langle Fy,U_{\mathbb{L}(\mathcal{A}_0)}z\rangle - \mathbb{h}(\mathbb{L}a + U_{\mathbb{L}(\mathcal{A}_0)}z)\}\\
&= \sup_{z\in\mathbb{R}^r}\{\langle y,a\rangle + \langle Fy,U_{\mathbb{L}(\mathcal{A}_0)}z\rangle - \mathbb{h}(\mathbb{L}a + U_{\mathbb{L}(\mathcal{A}_0)}z)\}\\
&= \sup_{z\in\mathbb{R}^r}\{\langle y,a\rangle + \langle U^T_{\mathbb{L}(\mathcal{A}_0)}Fy,z\rangle - \tilde{\mathbb{h}}(z)\}\\
&= \tilde{\mathbb{h}}^*(U^T_{\mathbb{L}(\mathcal{A}_0)}Fy) + \langle y,a\rangle,
\end{aligned}$$

where the fourth equality follows from the fact that for any $x\in\mathcal{A}$, there exists

$$z\in \mathrm{range}\,(\Sigma_{\mathbb{L}(\mathcal{A}_0)}V^T_{\mathbb{L}(\mathcal{A}_0)}) = \mathbb{R}^{r_1}$$

such that $\mathbb{L}x = \mathbb{L}a + U_{\mathbb{L}(\mathcal{A}_0)}z$. Finally, according to Lemma 15, we understand that

$$\mathrm{dom}\ \psi^* \subset \mathrm{range}\,(\mathbb{L}^T) + \mathcal{A}_0^{\perp}.$$

We therefore conclude that, for all $y\in\mathbb{R}^n$, it holds that

$$\psi^*(y) = \sup_{x\in\mathbb{R}^n}\{\langle y,x\rangle - \mathbb{h}(\mathbb{L}x) - \delta_{\mathcal{A}}(x)\} = \tilde{\mathbb{h}}^*(\tilde{\mathbb{L}}y) + \langle y,a\rangle + \delta_{\mathrm{range}\,(\mathbb{L}^T)+\mathcal{A}_0^{\perp}}(y),$$

where $\tilde{\mathbb{L}} := U^T_{\mathbb{L}(\mathcal{A}_0)}F = U^T_{\mathbb{L}(\mathcal{A}_0)}\left(U\Sigma^{-1}V^T + E_1H^{\dagger}(I - \mathbb{L}^TU\Sigma^{-1}V^T)\right)$.

We next prove the second argument. In fact, if $\partial\psi(x) = \mathbb{L}^T\nabla\mathbb{h}(\mathbb{L}x) + \mathcal{N}_{\mathcal{A}}(x)$ and $\mathrm{dom}\ \partial\psi \neq \varnothing$, then there exists $\hat{x}$ such that

$$\partial\psi(\hat{x}) = \mathbb{L}^T\nabla\mathbb{h}(\mathbb{L}\hat{x}) + \mathcal{N}_{\mathcal{A}}(\hat{x}) \neq \varnothing.$$

Thus, $\hat{x}\in\mathcal{A}$ and there exists $\hat{\xi}$ such that

$$\hat{\xi} = \nabla\mathbb{h}(\mathbb{L}\hat{x}).$$

Taking into consideration that

$$U^T_{\mathbb{L}(\mathcal{A}_0)} = \Sigma^{-1}_{\mathbb{L}(\mathcal{A}_0)}V^T_{\mathbb{L}(\mathcal{A}_0)}Q^T_{\mathcal{A}_0}\mathbb{L}^T = \Sigma^{-1}_{\mathbb{L}(\mathcal{A}_0)}V^T_{\mathbb{L}(\mathcal{A}_0)}Q^T_{\mathcal{A}_0}V\Sigma U^T,$$

we have

$$\begin{aligned}
\tilde{\mathbb{L}}\mathbb{L}^T\hat{\xi} &= U^T_{\mathbb{L}(\mathcal{A}_0)}\left(U\Sigma^{-1}V^T + E_1H^{\dagger}(I - \mathbb{L}^TU\Sigma^{-1}V^T)\right)\mathbb{L}^T\hat{\xi}\\
&= U^T_{\mathbb{L}(\mathcal{A}_0)}U\Sigma^{-1}V^T\mathbb{L}^T\hat{\xi} + E_1H^{\dagger}(I - \mathrm{Proj}_{\mathrm{range}\,(\mathbb{L}^T)})\mathbb{L}^T\hat{\xi}\\
&= U^T_{\mathbb{L}(\mathcal{A}_0)}U\Sigma^{-1}V^T\mathbb{L}^T\hat{\xi}\\
&= U^T_{\mathbb{L}(\mathcal{A}_0)}U\Sigma^{-1}V^TV\Sigma U^T\hat{\xi}\\
&= U^T_{\mathbb{L}(\mathcal{A}_0)}\hat{\xi} = U^T_{\mathbb{L}(\mathcal{A}_0)}\nabla\mathbb{h}(\mathbb{L}\hat{x}).
\end{aligned}$$

On the other hand, because $\hat{x} \in \mathcal{A}$, there exists $\hat{z}$ such that $\mathbb{L}\hat{x} = \mathbb{L}a + U_{\mathbb{L}(\mathcal{A}_0)}\hat{z}$, and thus

$$\tilde{\mathbb{L}}\mathbb{L}^T\hat{\xi} = U_{\mathbb{L}(\mathcal{A}_0)}^T \nabla \mathbb{h}(\mathbb{L}\hat{x}) = U_{\mathbb{L}(\mathcal{A}_0)}^T \nabla \mathbb{h}(\mathbb{L}a + U_{\mathbb{L}(\mathcal{A}_0)}\hat{z}) \in \partial \tilde{\mathbb{h}}(\hat{z}).$$

Recall the fact that $\tilde{h}^* \in \mathcal{C}$. By Proposition 17 and (Rockafellar, 1970, Corollary 23.5.1), we have

$$\tilde{\mathbb{L}}\mathbb{L}^T\hat{\xi} \in \operatorname{dom} \partial \tilde{\mathbb{h}}^* = \operatorname{int}(\operatorname{dom} \tilde{\mathbb{h}}^*),$$

which implies

$$\tilde{\mathbb{L}}(\operatorname{range}(\mathbb{L}^T) + \mathcal{A}_0^\perp) \cap \operatorname{int}(\operatorname{dom} \tilde{\mathbb{h}}^*) \neq \varnothing.$$

To the end, according to (Rockafellar, 1970, Theorem 23.8, Theorem 23.9) and Proposition 17, we have

$$\partial \psi^*(y) = \tilde{\mathbb{L}}^T \partial \tilde{\mathbb{h}}^*(\tilde{\mathbb{L}}y) + a + \mathcal{N}_{\operatorname{range}(\mathbb{L}^T)+\mathcal{A}_0^\perp}(y) = \tilde{\mathbb{L}}^T \nabla \tilde{\mathbb{h}}^*(\tilde{\mathbb{L}}y) + a + \mathcal{N}_{\operatorname{range}(\mathbb{L}^T)+\mathcal{A}_0^\perp}(y),$$

and therefore obtain the desired expression for $\partial\psi^*$.

Appendix E. Proof of Lemma 21

According to Assumption 1.2, $f(x) = h(Lx) + \langle q, x\rangle$ with $h \in \mathcal{C}$, then

$$\begin{aligned} f^*(y) &= \sup_x \{\langle y, x\rangle - h(Lx) - \langle q, x\rangle\} \\ &= \sup_x \{\langle y - q, x\rangle - h(Lx)\}. \end{aligned}$$

Then, by Proposition 18, there exist $\tilde{h}^* \in \mathcal{C}$ and matrix $\tilde{L}$ such that

$$f^*(y) = \tilde{h}^*\left(\tilde{L}(y-q)\right) + \delta_{\operatorname{range}(L^T)}(y-q),$$

and thus

$$\phi_1(\lambda) = f^*(A^T\lambda) - b^T\lambda = \tilde{h}^*\left(\tilde{L}A^T\lambda - \tilde{L}q\right) - b^T\lambda + \delta_{\operatorname{range}(L^T)}(A^T\lambda - q).$$

Denoting $K := \tilde{L}A^T$, $\tilde{q} := \tilde{L}q$ and $\mathcal{V} := \{\lambda \mid A^T\lambda - q \in \operatorname{range}(L^T)\}$, we therefore have shown the first assertion.

To prove the second argument, noting that Assumption 1.1 holds, by virtue of Lemma 20, we find that there exist x^* and λ^* satisfying

$$0 \in \partial f(x^*) + A^T\lambda^* = L^T\nabla h(Lx^*) + q + A^T\lambda^*.$$

Then, by Proposition 18, we have

$$\partial f^*(y) = \tilde{L}^T \nabla \tilde{h}^*\left(\tilde{L}(y-q)\right) + \mathcal{N}_{\operatorname{range}(L^T)}(y-q).$$

We may now obverse that any vector $y \in \operatorname{dom} \partial f^*$ if and only if $\tilde{L}(y-q) \in \operatorname{dom} \nabla \tilde{h}^*$ and $y - q \in \operatorname{range}(L^T)$. Since $\tilde{h}^*$ is essentially differentiable, thanks to Proposition 17, we understand that $y \in \operatorname{dom} \partial f^*$ if and only if $\tilde{L}(y-q) \in \operatorname{int}(\operatorname{dom} \tilde{h}^*)$ and $y - q \in \operatorname{range}(L^T)$.

According to the expression of f^*, we immediately know that dom $\partial f^* \subseteq$ ri(dom f^*). Noting that $-A^T\lambda^* \in \partial f(x^*)$, we may conclude that

$$-A^T\lambda^* \in \partial f(x^*) \subseteq \text{dom } \partial f^* \subseteq \text{ri}(\text{dom } f^*),$$

which further implies that

$$\text{range}\,(A^T) \cap \text{ri}(\text{dom } f^*) \neq \varnothing.$$

According to (Rockafellar, 1970, Theorem 23.9), immediately, $\lambda^* \in$ dom $\partial\phi_1$ and hence

$$\partial\phi_1(\lambda) = A\partial f^*(A^T\lambda) - b = K^T\nabla\tilde{h}^*\left(K\lambda - \tilde{q}\right) - b + \mathcal{N}_{\mathcal{V}}(\lambda).$$

Appendix F. Proof of Lemma 23

By Lemma 21, we know that

$$\phi_1(\lambda) = \tilde{h}^*\left(K\lambda - \tilde{q}\right) - b^T\lambda + \delta_{\mathcal{V}}(\lambda)$$

with some $\tilde{h}^* \in \mathcal{C}$, matrix K, vector $\tilde{q}$ and affine space $\mathcal{V}$. Then, it holds that

$$\begin{aligned}\phi_1{}^*(\mu) &= \sup_{\lambda}\{\langle\mu,\lambda\rangle - \tilde{h}^*\left(K\lambda - \tilde{q}\right) + \langle b,\lambda\rangle - \delta_{\mathcal{V}}(\lambda)\}\\ &= \sup_{\lambda}\{\langle\mu + b,\lambda\rangle - \tilde{h}^*\left(K\lambda - \tilde{q}\right) - \delta_{\mathcal{V}}(\lambda)\}.\end{aligned}$$

By introducing $\hat{h}_0^*(\nu) := \tilde{h}^*\left(\nu - \tilde{q}\right) \in \mathcal{C}$ and $\hat{h}^*(\lambda) := \hat{h}_0^*(K\lambda) + \delta_{\mathcal{V}}(\lambda)$, we have $\phi_1{}^*(\mu) = (\hat{h}^*)^*(\mu + b)$. By Proposition 18, there exist $\hat{h} \in \mathcal{C}$, matrix $\hat{L}$ and affine space $\hat{\mathcal{V}}_1$ such that

$$\phi_1{}^*(\mu) = \hat{h}\left(\hat{L}(\mu + b)\right) + \langle v, \mu + b\rangle + \delta_{\hat{\mathcal{V}}_1}(\mu + b),$$

Letting $\hat{K} := -\hat{L}$, $\hat{q} := \hat{L}b$ and $\hat{\mathcal{V}} := \{\mu \mid -\mu + b \in \hat{\mathcal{V}}_1\}$, we obtain the first conclusion.

Furthermore, from Lemma 21, we know that dom $\partial\phi_1 \neq \varnothing$

$$\partial\phi_1(\lambda) = K^T\nabla\tilde{h}^*\left(K\lambda - \tilde{q}\right) - b + \mathcal{N}_{\mathcal{V}}(\lambda).$$

Then, by Proposition 18, we have

$$\partial\left(\phi_1{}^*(-\mu)\right) = \hat{K}^T\nabla\hat{h}\left(\hat{K}\mu + \hat{q}\right) - v + \mathcal{N}_{\hat{\mathcal{V}}}(\mu).$$

Appendix G. Proof of Lemma 25

Motivated by (Luo and Tseng, 1992, Lemma 2.1), we first prove that there exists $\bar{t}$ such that $K\lambda = \bar{t}$ for all $\lambda \in Z$. On the contrary, suppose the existence of $\lambda_1, \lambda_2 \in Z$ such that $t_1 = K\lambda_1, t_2 = K\lambda_2$, and $t_1 \neq t_2$. Let us assume that λ_1 and λ_2 are sufficiently close; otherwise we can replace λ_2 by $\tilde{\lambda}_2 = \alpha\lambda_2 + (1-\alpha)\lambda_1$ with sufficiently small $\alpha > 0$, and $\tilde{t}_2 = K\tilde{\lambda}_2 \neq t_1$. Then, since $\tilde{h}^*$ is essentially locally strongly convex and $t_1 \in$ dom $\nabla\tilde{h}^*$, there exists $\sigma > 0$ such that

$$\phi_1(\lambda_1) + \phi_2(\lambda_1) \geq \phi_1(\lambda_2) + \phi_2(\lambda_2) + \frac{\sigma}{2}\|t_1 - t_2\|^2 > \phi_1(\lambda_2) + \phi_2(\lambda_2),$$

which is a contradiction. The desirable result then follows by taking $\bar{g} := K^T\nabla\tilde{h}^*\left(\bar{t} - \tilde{q}\right) - b$.

Appendix H. Proof of Corollary 31

For any $\mu \in W$, it follows from the definition of W that

$$0 \in -\partial\phi_1^* \circ (-\mu) + \partial\phi_2^*(\mu),$$

and thus there exists $\lambda \in \partial\phi_2^*(\mu)$ such that $\lambda \in \partial\phi_1^* \circ (-\mu)$. Since $\partial\phi_1^* = (\partial\phi_1)^{-1}$ and $\partial\phi_2^* = (\partial\phi_2)^{-1}$, we have $0 \in \partial\phi_1(\lambda) + \partial\phi_2(\lambda)$, i.e., $\lambda \in Z$, and

$$\begin{aligned} 0 &\in \partial\phi_1(\lambda) + \mu, \\ 0 &\in \partial\phi_2(\lambda) - \mu. \end{aligned}$$

Then, since $\phi_1(\lambda) = f^*(A^T\lambda) - b^T\lambda$ and $\phi_2(\lambda) = g^*(B^T\lambda)$, it follows from the full column rank of A and B and (Rockafellar, 1970, Theorem 23.9) that $\partial\phi_1(\lambda) = A\partial f^*(A^T\lambda) - b$ and $\partial\phi_2(\lambda) = B\partial g^*(B^T\lambda)$. Thus, we have

$$\begin{aligned} 0 &\in A\partial f^*(A^T\lambda) - b + \mu, \\ 0 &\in B\partial g^*(B^T\lambda) - \mu, \end{aligned}$$

which implies that there exist $\hat{x} \in \partial f^*(A^T\lambda)$ and $\hat{y} \in \partial g^*(B^T\lambda)$ such that $A\hat{x} - b + \mu = 0$ and $\mu = B\hat{y}$. Next, by the fact that $\partial f^* = (\partial f)^{-1}$, $\partial g^* = (\partial g)^{-1}$, we have

$$\begin{cases} 0 \in \partial f(\hat{x}) - A^T\lambda, \\ 0 \in \partial g(\hat{y}) - B^T\lambda, \\ A\hat{x} + B\hat{y} - b = 0, \end{cases}$$

and thus $(\hat{x}, \hat{y}, \lambda) \in \Omega^*$, $\hat{y} \in \Omega_y^*$. Therefore, we have $W \subseteqq B\Omega_y^* := \{By \mid \exists (x, y, \lambda) \in \Omega^*\}$. Similarly, employing the above argument from the opposite direction, we can also show that $B\Omega_y^* \subseteq W$. In summary, we have $W = B\Omega_y^*$.

We are now ready to prove the linear convergence of the sequences $\{x^k\}$ and $\{y^k\}$. By Proposition 8, we know $\mathrm{prox}_{\phi_2^*}(z^k) = By^k$. Following (22) in Proposition 30, since B is of full column rank, we know that there exist $k_0 > 0$, $0 < \rho < 1$ and $C_0 > 0$ such that, for all $k \geq k_0$, it holds that

$$\mathrm{dist}\left(y^k, \Omega_y^*\right) \leq C_0\rho^k.$$

Furthermore, from Proposition 30, there exist $\tilde{k}_0 \geq k_0 > 0$, $\tilde{C}_0 > 0$ such that, for all $k \geq \tilde{k}_0$, it holds that

$$\|Ax^k + By^k - b\| \leq \tilde{C}_0\rho^k.$$

From the above arguments, we know that for each $k \geq k_0$, there exists $(\hat{x}^k, \hat{y}^k, \hat{\lambda}^k) \in \Omega^*$ such that

$$\|y^k - \hat{y}^k\| \leq C_0\rho^k,$$

and then

$$\|Ax^k - A\hat{x}^k\| \leq \|Ax^k + By^k - b\| + \|By^k - B\hat{y}^k\| \leq (\tilde{C}_0 + C_0\|B\|)\rho^k.$$

According to the full column rank of A, we get the conclusion.

Appendix I. Proof of Theorem 33

At each iteration, the PPA iterative scheme (34) reads also as

$$-\mathcal{M}(\mathbb{x}^{k+1}-\mathbb{x}^k)\in T(\mathbb{x}^{k+1}).$$

Therefore, for any $\mathbb{x}^*\in\Omega^*_{x,\lambda}$, it follows from the monotonicity of T that

$$\|\mathbb{x}^{k+1}-\mathbb{x}^*\|^2_{\mathcal{M}}\le\|\mathbb{x}^k-\mathbb{x}^*\|^2_{\mathcal{M}}-\|\mathbb{x}^{k+1}-\mathbb{x}^k\|^2_{\mathcal{M}}. \tag{73}$$

Since $\mathbb{x}^*$ can be taken arbitrarily in $\Omega^*_{x,\lambda}$, we immediately have

$$\mathrm{dist}^2_{\mathcal{M}}\left(\mathbb{x}^{k+1},\Omega^*_{x,\lambda}\right)\le\mathrm{dist}^2_{\mathcal{M}}\left(x^k,\Omega^*_{x,\lambda}\right)-\|\mathbb{x}^{k+1}-\mathbb{x}^k\|^2_{\mathcal{M}}. \tag{74}$$

Because of the PPA-iteration-based error bound, there exist $\epsilon,\kappa>0$ such that

$$\mathrm{dist}_{\mathcal{M}}\left(\mathbb{x}^{k+1},\Omega^*_{x,\lambda}\right)\le\kappa\|\mathbb{x}^{k+1}-\mathbb{x}^k\|_{\mathcal{M}},\quad\text{for all } k \text{ such that } \mathbb{x}^k\in\mathbb{B}(\bar{\mathbb{x}},\epsilon).$$

Given this ϵ, with the by-default given proximity of the sequence $\{\mathbb{x}^k\}$ generated by the PPA to $\bar{\mathbb{x}}\in\Omega^*_{x,\lambda}$, there exists $k_0>0$ such that $\mathbb{x}^k\in\mathbb{B}_\epsilon(\bar{\mathbb{x}})$ for $k\ge k_0$. Therefore, we have

$$\mathrm{dist}^2_{\mathcal{M}}\left(\mathbb{x}^{k+1},\Omega^*_{x,\lambda}\right)\le\frac{\kappa^2}{1+\kappa^2}\mathrm{dist}^2_{\mathcal{M}}\left(\mathbb{x}^k,\Omega^*_{x,\lambda}\right),\quad\forall k\ge k_0,$$

and thus there exists $C>0$ such that

$$\mathrm{dist}_{\mathcal{M}}\left(\mathbb{x}^k,\Omega^*_{x,\lambda}\right)\le C\rho^k,\quad\forall k\ge k_0,$$

with $\rho=\sqrt{\frac{\kappa^2}{1+\kappa^2}}$. Moreover, according to (74), we get

$$\|\mathbb{x}^{k+1}-\mathbb{x}^k\|_{\mathcal{M}}\le C\rho^k,\quad\forall k\ge k_0.$$

The desired linear convergence then follows from the positive definiteness of matrix $\mathcal{M}$.

Appendix J. Proof of Theorem 35

Because of the metric subregularity of T at $(\bar{\mathbb{x}},0)$, there exist $\kappa>0,\epsilon>0$ such that

$$\mathrm{dist}\left(\bar{\mathbb{x}},T^{-1}(0)\right)\le\kappa\mathrm{dist}(0,T(\bar{\mathbb{x}})),\quad\forall\bar{\mathbb{x}}\in\mathbb{B}_\epsilon(\bar{\mathbb{x}}).$$

Note that $T^{-1}(0)=\Omega^*_{x,\lambda}$. The metric subregularity of T at $(\bar{\mathbb{x}},0)$ allows us to estimate the distance from $\mathbb{x}^{k+1}$ to $\Omega^*_{x,\lambda}$ in terms of scaled optimality residual at $\mathbb{x}^{k+1}$, i.e., for all k such that $\mathbb{x}^k\in\mathbb{B}(\bar{\mathbb{x}},\epsilon)$,

$$\begin{aligned}\mathrm{dist}_{\mathcal{M}}\left(\mathbb{x}^{k+1},\Omega^*_{x,\lambda}\right)&\le\sqrt{\rho(\mathcal{M})}\mathrm{dist}\left(\mathbb{x}^{k+1},\Omega^*_{x,\lambda}\right)\\&\le\kappa\sqrt{\rho(\mathcal{M})}\|\mathcal{M}(\mathbb{x}^{k+1}-\mathbb{x}^k)\|\\&\le\kappa\rho(\mathcal{M})\|\mathbb{x}^{k+1}-\mathbb{x}^k\|_{\mathcal{M}},\end{aligned} \tag{75}$$

where $\rho(\mathcal{M})$ represents the spectral radius of matrix $\mathcal{M}$. Thus, the PPA-iteration-based error bound holds at $\bar{\mathbb{x}}$. The conclusion then follows from Theorem 33.

Appendix K. Proof of Proposition 39

According to Proposition 38, we have

$$X = \arg\min_x\{\theta_1(x) + \theta_2^*(-Ax)\}.$$

Thanks to Assumption 1.2, there exists $\tilde{t} \in \mathbb{R}^l$ such that $Lx = \tilde{t}$ for all $x \in X$. Moreover, for any $(x, \lambda) \in \Omega^*_{x,\lambda}$, we have

$$0 \in \partial\theta_2(\lambda) + Ax$$
$$0 \in \partial\theta_1(x) - A^T\lambda = L^T\nabla h(Lx) + q - A^T\lambda = L^T\nabla h(\tilde{t}) + q - A^T\lambda = \tilde{g} - A^T\lambda,$$

where $\tilde{g} := L^T\nabla h(\tilde{t}) + q$. Therefore, the following inclusion holds

$$\Omega^*_{x,\lambda} \subseteq \{(x, \lambda) \mid Lx = \tilde{t}, A^T\lambda = \tilde{g}, 0 \in \partial\theta_2(\lambda) + Ax\}.$$

It is easy to obtain the reverse direction. The proof is complete.

Appendix L. Proof of Proposition 42

Given any $(\bar{x}, \bar{\lambda}) \in \Omega^*_{x,\lambda}$. Suppose that there exist $\kappa_1, \epsilon_1 > 0$ such that

$$\text{dist}\left((x, \lambda), \Gamma_0(0)\right) \leq \kappa_1 \text{dist}\left(0, \Gamma_0^{-1}(x, \lambda)\right), \ \forall (x, \lambda) \in \mathbb{B}_{\epsilon_1}(\bar{x}, \bar{\lambda}).$$

Due to the essentially locally strongly convexity of h and the locally Lipschitz continuity of ∇h, without loss of generality, we assume that ϵ_1 is small enough so that ∇h is strongly monotone and Lipschitz continuous on $\{Lx \mid (x, \lambda) \in \mathbb{B}_{\epsilon_1}(\bar{x}, \bar{\lambda})\}$. For any $(x, \lambda) \in \mathbb{B}_{\epsilon_1}(\bar{x}, \bar{\lambda})$, and any $(\xi, \eta) \in T(x, \lambda)$

$$\xi = L^T\nabla h(Lx) + q - A^T\lambda,$$
$$\eta \in \partial\theta_2(\lambda) + Ax,$$

since $0 \in L^T\nabla h($
$tildet) + q - A^T\bar{\lambda}$, and by the local Lipschitz continuity of ∇h, there exists $L_h > 0$ such that

$$\|A^T\lambda - A^T\bar{\lambda}\| \leq \|L\|\|\nabla h(Lx) - \nabla h(\tilde{t})\| \leq L_h\|L\|\|Lx - \tilde{t}\|.$$

Moreover, noting that A^T is of full column rank, the smallest singular value of A^T is strictly positive, i.e., $\sigma_{\min}(A^T) > 0$. Therefore

$$\|\lambda - \bar{\lambda}\| \leq \frac{1}{\sigma_{\min}(A^T)}\|A^T\lambda - A^T\bar{\lambda}\| \leq \frac{L_h\|L\|}{\sigma_{\min}(A^T)}\|Lx - \tilde{t}\|. \tag{76}$$

According to the calmness of $\partial\theta_2$ at $(\bar{\lambda}, -A\bar{x})$, there exist $\epsilon_3, \kappa_3 > 0$ such that

$$\text{dist}\left(v, \partial\theta_2(\bar{\lambda})\right) \leq \kappa_3 \text{dist}\left(\bar{\lambda}, (\partial\theta_2)^{-1}(v)\right), \ \forall v \in \mathbb{B}_{\epsilon_3}(-A\bar{x}).$$

We now assume that $(x, \lambda) \in \mathbb{B}_{\epsilon_2}(\bar{x}, \bar{\lambda})$ with $\epsilon_2 := \min\{\epsilon_1, \epsilon_3/(2\|A\|)\}$ and $\|\eta\| \leq \epsilon_3/2$. Since $\eta \in \partial\theta_2(\lambda) + Ax$ and thus

$$\eta - Ax \in \partial\theta_2(\lambda), \|\eta - Ax + A\bar{x}\| \leq \|\eta\| + \|A\|\|x - \bar{x}\| \leq \epsilon_3.$$

Then, by the calmness of $\partial\theta_2$ at $(\bar{\lambda}, -A\bar{x})$, we have

$$\operatorname{dist}(0, D + Ax) \leq \|\eta\| + \operatorname{dist}(\eta - Ax, D) \leq \|\eta\| + \kappa_3\|\lambda - \bar{\lambda}\|. \tag{77}$$

By (76) and (77), we have

$$\begin{aligned}\operatorname{dist}\left((x,\lambda), \Omega^*_{x,\lambda}\right) = \operatorname{dist}\left((x,\lambda), \Gamma_0(0)\right) &\leq \kappa_1 \operatorname{dist}\left(0, \Gamma_0^{-1}(x,\lambda)\right) \\ &\leq \kappa_1 \left(\|Lx - \tilde{t}\| + \|\lambda - \bar{\lambda}\| + \operatorname{dist}(0, D + Ax)\right) \\ &\leq \kappa_1 \left(\|Lx - \tilde{t}\| + \frac{(1+\kappa_3)L_h\|L\|}{\sigma_{\min}(A^T)}\|Lx - \tilde{t}\| + \|\eta\|\right) \\ &\leq \kappa_1 \left((1 + \frac{(1+\kappa_3)L_h\|L\|}{\sigma_{\min}(A^T)})\|Lx - \tilde{t}\| + \|\xi\| + \|\eta\|\right).\end{aligned} \tag{78}$$

Then, similar to the proof of Proposition 40, there exists $\sigma > 0$ such that

$$\sigma\|Lx - \tilde{t}\|^2 \leq \langle \xi, x - \hat{x}\rangle + \langle \eta, \lambda - \hat{\lambda}\rangle, \tag{79}$$

where $(\hat{x}, \hat{\lambda})$ is the projection of (x, λ) on $\Omega^*_{x,\lambda}$. Upon combining (78) and (79), inspired by the proof of Proposition 40, we prove the conclusion with

$$\epsilon_2 = \min\{\epsilon_1, \epsilon_3/(2\|A\|)\},\ \kappa_2 = \max\{\frac{1}{\|A\|}, \tilde{\kappa}\},$$

where

$$\tilde{\kappa} = \left(\frac{c_1 + \sqrt{c_1^2 + 4c_2}}{2}\right)^2 > 0,$$

and

$$c_1 = \kappa_1(\sigma_{\min}(A^T) + (1+\kappa_3)L_h\|L\|)/(\sqrt{\sigma}\sigma_{\min}(A^T)),\ c_2 = \sqrt{2}\kappa_1.$$

Appendix M. Proof of Lemma 44

Since $D := \partial\theta_2(\bar{\lambda})$ is closed and $D \subseteq \operatorname{range}(A)$, for any x, there exists x_D such that

$$\operatorname{dist}(0, D + Ax) = \|Ax - Ax_D\|$$

and $Ax_D \in D$. Define $\mathbb{F}_x := \{z \mid Az = Ax_D\}$, according to Hoffman error bound (see, e.g., Hoffman, 1952),

$$\operatorname{dist}(x, \mathbb{F}_x) \leq \frac{1}{\tilde{\sigma}_{\min}(A)}\|Ax - Ax_D\|,$$

where $\tilde{\sigma}_{\min}(A)$ denotes the smallest nonzero singular value of A. Since $\mathbb{F}_x \subseteq \Omega^2_x(0)$ for any x, we have

$$\operatorname{dist}\left(x, \Omega^2_x(0)\right) \leq \operatorname{dist}(x, \mathbb{F}_x) \leq \frac{1}{\tilde{\sigma}_{\min}(A)}\|Ax - Ax_D\| = \frac{1}{\tilde{\sigma}_{\min}(A)}\operatorname{dist}(0, D + Ax),$$

which implies the calmness of Ω^2_x.

Appendix N. Proof of Lemma 47

The first assertion follows directly from the fact that $(\partial g)^{-1} = \partial g^*$. We focus on the second assertion. Since ∂g^* is known to be calm at $(\bar{v}, \bar{y})$ with modulus κ, there exist $\epsilon > 0$ such that

$$\partial g^*(v) \cap \mathbb{B}_\epsilon(\bar{y}) \subseteq \partial g^*(\bar{v}) + \kappa\|v - \bar{v}\|\mathbb{B}, \qquad v \in \mathbb{B}_\epsilon(\bar{v}).$$

Also, there exists $\epsilon_1 > 0$ such that $\{B^T z \mid z \in \mathbb{B}_{\epsilon_1}(\bar{z})\} \subseteq \mathbb{B}_\epsilon(\bar{v})$ and range $(B) \cap \mathbb{B}_{\epsilon_1}(B\bar{y}) \subseteq \{By \mid y \in \mathbb{B}_\epsilon(\bar{y})\}$. For any $z \in \mathbb{B}_{\epsilon_1}(\bar{z})$, we have

$$\begin{aligned} B\partial g^*(B^T z) \cap \mathbb{B}_{\epsilon_1}(B\bar{y}) &\subseteq B\left(\partial g^*(B^T z) \cap \mathbb{B}_\epsilon(\bar{y})\right) \\ &\subseteq B\left(\partial g^*(\bar{v}) + \kappa\|B^T z - \bar{v}\|\mathbb{B}\right) \\ &\subseteq B\left(\partial g^*(B^T\bar{z}) + \kappa\|B\|\|z - \bar{z}\|\mathbb{B}\right) \\ &\subseteq B\partial g^*(B^T\bar{z}) + \kappa\|B\|^2\|z - \bar{z}\|\mathbb{B}. \end{aligned}$$

The second assertion is then proved.

For the third assertion, it directly follows from (Rockafellar, 1970, Theorem 23.8, Theorem 23.9). The proof is complete.

Appendix O. Symbols and Notation

$\mathbb{R}^n$	The standard n-dimensional Euclidean space
$\langle\cdot,\cdot\rangle$	Scalar product in n-dimensional Euclidean space
$\|\cdot\|$	The Euclidean norm
$\|\cdot\|_1$	ℓ_1 norm
$\|\cdot\|_\infty$	ℓ_∞ norm
$\mathbb{B}$	Open unit ball centered at the origin
$\overline{\mathbb{B}}$	Closed unit ball centered at the origin
$\mathbb{B}_r(x)$	Open ball around x with radius $r > 0$
int $\mathcal{D}$	Interior of set $\mathcal{D}$
ri $\mathcal{D}$	Relative interior of set $\mathcal{D}$
bd $\mathcal{D}$	Boundary of set $\mathcal{D}$
dist$(\cdot, \mathcal{D})$	Distance function to set $\mathcal{D}$
$\delta_{\mathcal{D}}(x)$	Indicator function of set $\mathcal{D}$
$\mathcal{N}_{\mathcal{D}}$	Normal cone operator of set $\mathcal{D}$
$\mathcal{V}_0^\perp$	Orthogonal complement of $\mathcal{V}_0$
$\mathcal{A} + \mathcal{B}$	Minkowski sum given by $\{a + b : a \in \mathcal{A}, b \in \mathcal{B}\}$
range (A)	Range of matrix A
null (A)	Null space of matrix A
$\|A\|$	$\max_{x\neq 0} \frac{\|Ax\|}{\|x\|}$
$\partial\phi$	Subdifferential of ϕ, i.e., $\partial\phi(x) := \{\xi \mid f(y) \geq f(x) + \langle\xi, y - x\rangle,\ \forall y\}$

$\text{dom}\ \phi$	Domain of ϕ, i.e., $\text{dom}\ \phi := \{x \mid \phi(x) < +\infty\}$
ϕ^*	Conjugate of ϕ, i.e., $\phi^*(\mu) := \sup_x\{\langle \mu, x\rangle - \phi(x)\}$
$\text{gph}\,(\Phi)$	Graph of set-valued map (multifunction) Φ, i.e., $\text{gph}\,(\Phi) := \{(x, \upsilon) \in \mathbb{R}^n \times \mathbb{R}^q \mid \upsilon \in \Phi(x)\}$
Φ^{-1}	Inverse mapping of set-valued map (multifunction) Φ, i.e., $\Phi^{-1}(\upsilon) := \{x \in \mathbb{R}^n \mid \upsilon \in \Phi(x)\}$

Specific notations, descriptions and references:		
T_{KKT}	$T_{KKT}(x,y,\lambda) := \begin{pmatrix} \partial f(x) - A^T\lambda \\ \partial g(y) - B^T\lambda \\ Ax + By - b \end{pmatrix}$	(4) in Section 1.1
ϕ_1	$\phi_1(\lambda) := f^*(A^T\lambda) - b^T\lambda$	Section2.2
ϕ_2	$\phi_2(\lambda) := g^*(B^T\lambda)$	Section2.2
$\mathcal{T}_1$	$\mathcal{T}_1(p) := \{\lambda \mid p \in \partial\phi_1(\lambda - p) + \partial\phi_2(\lambda)\}$	(13) in Section2.3.2
$\mathcal{T}_2$	$\mathcal{T}_2(p) := \{\mu \mid p \in \partial(\phi_1^* \circ -Id)(\mu - p) + \partial\phi_2^*(\mu)\}$	(14) in Section2.3.2
$\tilde{\mathcal{S}}_{D_1}$	$\tilde{\mathcal{S}}_{D_1}(p) := \{\lambda \mid p \in \partial\phi_1(\lambda) + \partial\phi_2(\lambda)\}$	Section2.4.1
$\tilde{\mathcal{S}}_{D_2}$	$\tilde{\mathcal{S}}_{D_2}(p) := \{\mu \mid p \in \partial(\phi_1^* \circ -Id)(\mu) + \partial\phi_2^*(\mu)\}$	Section2.4.1
Γ_{DR}	$\Gamma_{DR}(p_1,p_2) := \{\lambda \in \mathcal{V} \mid p_1 = K\lambda - \bar{t}, \quad p_2 \in \bar{g} + \mathcal{V}_0^\perp + \partial\phi_2(\lambda)\}$	(17) in Section 2.4.2
$\tilde{\Gamma}_{DR}$	$\tilde{\Gamma}_{DR}(p_1,p_2) := \{\mu \in \hat{\mathcal{V}} \mid p_1 = \hat{K}\mu - \hat{t}, \quad p_2 \in \hat{g} + \hat{\mathcal{V}}_0^\perp + \partial\phi_2^*(\mu)\}$	(21) in Section 2.4.2
θ_1	$\theta_1(x) := f(x)$	Section 3.2
θ_2	$\theta_2(\lambda) := g^*(B^T\lambda) - \langle b, \lambda\rangle$	Section 3.2
$\mathcal{M}$	$\mathcal{M} := \begin{pmatrix} \frac{1}{\tau}I & -A^T \\ -A & \frac{1}{\sigma}I \end{pmatrix}$	(32) in Section 3.2
T	$T(x,\lambda) := \begin{pmatrix} \partial\theta_1(x) - A^T\lambda \\ \partial\theta_2(\lambda) + Ax \end{pmatrix}$	(32) in Section 3.2
Γ_{PDHG}	$\Gamma_{PDHG}(p) := \{(x,\lambda) \mid p_1 = Lx - \tilde{t}, p_2 = \tilde{g} - A^T\lambda, p_3 \in \partial\theta_2(\lambda) + Ax\}$	Section 3.4.1
Γ_0	$\Gamma_0(p) := \{(x,\lambda) \mid p_1 = Lx - \tilde{t},\ p_2 = -\bar{\lambda} + \lambda,\ p_3 \in D + Ax\}$	Section 3.4.2
Γ_{KKT}	$\Gamma_{KKT}(p) := \{(x,y,\lambda) \mid p_1 = Lx - \tilde{t}, p_2 = \tilde{g} - A^T\lambda, p_3 \in \partial g(y) - B^T\lambda, p_4 = Ax + By - b\}$	Section 4.2

References

Timo Aspelmeier, C. Charitha, and D. Russell Luke. Local linear convergence of the ADMM/Douglas–Rachford algorithms without strong convexity and application to statistical imaging. *SIAM Journal on Imaging Sciences*, 9(2):842–868, 2016.

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, volume 408. Springer, 2011.

Heinz H. Bauschke and Walaa M. Moursi. On the Douglas–Rachford algorithm. *Mathematical Programming*, 164(1-2):263–284, 2017.

Dimitri P. Bertsekas, Angelia Nedić, and Asuman E. Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific optimization and computation series. Athena Scientific, 2003. ISBN 9781886529458.

Howard D. Bondell and Brian J. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, 64(1):115–123, 2008.

Silvia Bonettini and Valeria Ruggiero. On the convergence of primal-dual hybrid gradient algorithms for total variation image restoration. *Journal of Mathematical Imaging and Vision*, 44(3):236–253, 2012.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40 (1):120–145, 2011.

Tony F. Chan and Roland Glowinski. *Finite Element Approximation and Iterative Solution of a Class of Mildly Non-linear Elliptic Equations*. Computer Science Department, Stanford University Stanford, 1978.

Damek Davis and Wotao Yin. Faster convergence rates of relaxed Peaceman–Rachford and ADMM under regularity assumptions. *Mathematics of Operations Research*, 42(3): 783–805, 2017.

Wei Deng and Wotao Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, 66(3):889–916, 2016.

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

Jonathan Eckstein and Dimitri P. Bertsekas. An alternating direction method for linear programming. 1990.

Jonathan Eckstein and Dimitri P. Bertsekas. On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318, 1992.

Jonathan Eckstein and Wang Yao. Understanding the convergence of the alternating direction method of multipliers: Theoretical and computational perspectives. *Pacific Journal of Optimization*, 11(4):619–644, 2015.

Ernie Esser, Xiaoqun Zhang, and Tony F. Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences*, 3(4):1015–1046, 2010.

Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional Variational Inequalities and Complementarity Problems*. Springer Science & Business Media, 2007.

Massimo Fornasier and Holger Rauhut. Recovery algorithms for vector-valued data with joint sparsity constraints. *SIAM Journal on Numerical Analysis*, 46(2):577–613, 2008.

Michel Fortin and Roland Glowinski. Chapter iii on decomposition-coordination methods using an augmented Lagrangian. In *Studies in Mathematics and Its Applications*, volume 15, pages 97–146. Elsevier, 1983.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010.

Daniel Gabay. Chapter ix applications of the method of multipliers to variational inequalities. In *Studies in Mathematics and Its Applications*, volume 15, pages 299–331. Elsevier, 1983.

Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976.

Helmut Gfrerer. First order and second order characterizations of metric subregularity and calmness of constraint set mappings. *SIAM Journal on Optimization*, 21(4):1439–1474, 2011.

Helmut Gfrerer. On directional metric regularity, subregularity and optimality conditions for nonsmooth mathematical programs. *Set-Valued and Variational Analysis*, 21(2):151–176, 2013.

Helmut Gfrerer and Diethard Klatte. Lipschitz and hölder stability of optimization problems and generalized equations. *Mathematical Programming*, 158(1-2):35–75, 2016.

Helmut Gfrerer and Jane J. Ye. New constraint qualifications for mathematical programs with equilibrium constraints via variational analysis. *SIAM Journal on Optimization*, 27 (2):842–865, 2017.

Pontus Giselsson and Stephen Boyd. Linear convergence and metric selection for Douglas–Rachford splitting and admm. *IEEE Transactions on Automatic Control*, 62(2):532–544, 2016.

Roland Glowinski. *Numerical Methods for Nonlinear Variational Problems*. Scientific Computation. Springer Berlin Heidelberg, 2013. ISBN 9783662126134.

Roland Glowinski. *On Alternating Direction Methods of Multipliers: A Historical Perspective*, pages 59–82. Springer Netherlands, Dordrecht, 2014. ISBN 978-94-017-9054-3. doi: 10.1007/978-94-017-9054-3_4.

Roland Glowinski and A Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 9(R2):41–76, 1975.

Ronald Glowinski and Patrick Le Tallec. *Augmented Lagrangian and Operator-splitting Methods in Nonlinear Mechanics*, volume 9. SIAM, 1989.

Rafal Goebel and R. Tyrrell Rockafellar. Local strong convexity and local lipschitz continuity of the gradient of convex functions. *Journal of Convex Analysis*, 15(2):263, 2008.

Osman Güler. On the convergence of the proximal point algorithm for convex minimization. *SIAM Journal on Control and Optimization*, 29(2):403–419, 1991.

Lei Guo, Jane J. Ye, and Jin Zhang. Mathematical programs with geometric constraints in banach spaces: enhanced optimality, exact penalty, and sensitivity. *SIAM Journal on Optimization*, 23(4):2295–2319, 2013.

Deren Han and Xiaoming Yuan. Local linear convergence of the alternating direction method of multipliers for quadratic programs. *SIAM Journal on numerical analysis*, 51(6):3446–3457, 2013.

Deren Han, Defeng Sun, and Liwei Zhang. Linear rate convergence of the alternating direction method of multipliers for convex composite programming. *Mathematics of Operations Research*, 43(2):622–637, 2017.

Bingsheng He and Hai Yang. Some convergence properties of a method of multipliers for linearly constrained monotone variational inequalities. *Operations research letters*, 23 (3-5):151–161, 1998.

Bingsheng He and Xiaoming Yuan. On the o(1/n) convergence rate of the Douglas–Rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2):700–709, 2012a.

Bingsheng He and Xiaoming Yuan. Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective. *SIAM Journal on Imaging Sciences*, 5(1):119–149, 2012b.

Bingsheng He and Xiaoming Yuan. On non-ergodic convergence rate of Douglas–Rachford alternating direction method of multipliers. *Numerische Mathematik*, 130(3):567–577, 2015.

Bingsheng He, Li-Zhi Liao, Deren Han, and Hai Yang. A new inexact alternating directions method for monotone variational inequalities. *Mathematical Programming*, 92(1):103–118, 2002.

Bingsheng He, Minghua Xu, and Xiaoming Yuan. Solving large-scale least squares covariance matrix problems by alternating direction methods. *SIAM Journal on Matrix Analysis and Applications*, 32(1):136, 2011.

Bingsheng He, Yanfei You, and Xiaoming Yuan. On the convergence of primal-dual hybrid gradient algorithm. *SIAM Journal on Imaging Sciences*, 7(4):2526–2537, 2014.

René Henrion and Jiří V Outrata. Calmness of constraint systems with applications. *Mathematical Programming*, 104(2-3):437–464, 2005.

René Henrion, Abderrahim Jourani, and Jiri Outrata. On the calmness of a class of multifunctions. *SIAM Journal on Optimization*, 13(2):603–618, 2002.

Alan J Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49(4), 1952.

Gareth M. James, Courtney Paulson, and Paat Rusmevichientong. Penalized and constrained regression. *Unpublished manuscript, http://www-bcf. usc. edu/˜ gareth/research/Research. html*, 2013.

Diethard Klatte and Bernd Kummer. Constrained minima and lipschitzian penalties in metric spaces. *SIAM Journal on Optimization*, 13(2):619–633, 2002.

Matthieu Kowalski. Sparse regression using mixed norms. *Applied and Computational Harmonic Analysis*, 27(3):303–324, 2009.

D. Leventhal. Metric subregularity and the proximal point method. *Journal of Mathematical Analysis and Applications*, 360(2):681–688, 2009.

Min Li, Defeng Sun, and Kim-Chuan Toh. A majorized admm with indefinite proximal terms for linearly constrained convex composite optimization. *SIAM Journal on Optimization*, 26(2):922–950, 2016.

Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Local convergence properties of Douglas–Rachford and alternating direction method of multipliers. *Journal of Optimization Theory and Applications*, 172(3):874–913, 2017.

Wei Lin, Pixu Shi, Rui Feng, and Hongzhe Li. Variable selection in regression with compositional covariates. *Biometrika*, 101(4):785–797, 2014.

Zhouchen Lin, Risheng Liu, and Huan Li. Linearized alternating direction method with parallel splitting and adaptive penalty for separable convex programs in machine learning. *Machine Learning*, 99(2):287–325, 2015.

Pierre-Louis Lions and Bertrand Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

Yongchao Liu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. Partial error bound conditions and the linear convergence rate of the alternating direction method of multipliers. *SIAM Journal on Numerical Analysis*, 56(4):2095–2123, 2018.

Zhi-Quan Luo and Paul Tseng. On the linear convergence of descent methods for convex essentially smooth minimization. *SIAM Journal on Control and Optimization*, 30(2): 408–425, 1992.

Bernard Martinet. Regularisation, d'inéquations variationelles par approximations successives. *Revue Francaise d'informatique et de Recherche operationelle*, 1970.

Renato DC Monteiro and Benar F Svaiter. Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. *SIAM Journal on Optimization*, 23(1):475–507, 2013.

Robert Nishihara, Laurent Lessard, Benjamin Recht, Andrew Packard, and Michael I. Jordan. A general analysis of the convergence of ADMM. *arXiv preprint arXiv:1502.02009*, 2015.

Stephen M. Robinson. Stability theory for systems of inequalities. part i: Linear systems. *SIAM Journal on Numerical Analysis*, 12(5):754–769, 1975.

Stephen M. Robinson. Strongly regular generalized equations. *Mathematics of Operations Research*, 5(1):43–62, 1980.

Stephen M. Robinson. Some continuity properties of polyhedral multifunctions. In *Mathematical Programming at Oberwolfach*, pages 206–214. Springer, 1981.

R. Tyrrell Rockafellar. *Convex Analysis*, volume 28. Princeton university press, 1970.

R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

R. Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.

R. Shefi. Rate of convergence analysis for convex nonsmooth optimization algorithms. *Unpublished doctoral dissertation, Tel Aviv University, Israel*, 2015.

Pixu Shi, Anru Zhang, and Hongzhe Li. Regression analysis for microbiome compositional data. *The Annals of Applied Statistics*, 10(2):1019–1040, 2016.

Jie Sun and Su Zhang. A modified alternating direction method for convex quadratically constrained quadratic semidefinite programs. *European Journal of Operational Research*, 207(3):1210–1220, 2010.

Min Tao and Xiaoming Yuan. The generalized proximal point algorithm with step size 2 is not necessarily convergent. *Computational Optimization and Applications*, 70(3):827–839, 2018a.

Min Tao and Xiaoming Yuan. On Glowinskis open question on the alternating direction method of multipliers. *Journal of Optimization Theory and Applications*, 179(1):163–196, 2018b.

Marc Teboulle. Convergence of proximal-like algorithms. *SIAM Journal on Optimization*, 7(4):1069–1083, 1997.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

Paul Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Mathematical Programming*, 125:263–295, 2010.

Tuomo Valkonen. A primal–dual hybrid gradient method for nonlinear operators with applications to mri. *Inverse Problems*, 30(5):055012, 2014.

Tuomo Valkonen. Preconditioned proximal point methods and notions of partial subregularity. *arXiv preprint arXiv:1711.05123*, 2017.

Xiangfeng Wang and Xiaoming Yuan. The linearized alternating direction method of multipliers for dantzig selector. *SIAM Journal on Scientific Computing*, 34(5):2792–2811, 2012.

Xiangfeng Wang, Jane J. Ye, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. Perturbation techniques for convergence analysis of proximal gradient method and other first-order algorithms via variational analysis. *arXiv preprint arXiv:1810.10051*, 2018.

Zaiwen Wen, Donald Goldfarb, and Wotao Yin. Alternating direction augmented Lagrangian methods for semidefinite programming. *Mathematical Programming Computation*, 2(3-4):203–230, 2010.

Junfeng Yang and Xiaoming Yuan. Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization. *Mathematics of Computation*, 82(281): 301–329, 2013.

Weihong Yang and Deren Han. Linear convergence of the alternating direction method of multipliers for a class of convex optimization problems. *SIAM journal on Numerical Analysis*, 54(2):625–640, 2016.

Jane J. Ye and Xiangyang Ye. Necessary optimality conditions for optimization problems with variational inequality constraints. *Mathematics of Operations Research*, 22(4):977–997, 1997.

Jane J. Ye and Jin Zhang. Enhanced karush–kuhn–tucker condition and weaker constraint qualifications. *Mathematical Programming*, 139(1-2):353–381, 2013.

Jane J. Ye, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. Variational analysis perspective on linear convergence of some first order methods for nonsmooth convex optimization problems. *optimization-online preprint*, 2018.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Leon Wenliang Zhong and James T. Kwok. Efficient sparse modeling with automatic feature grouping. *IEEE transactions on neural networks and learning systems*, 23(9):1436–1447, 2012.

Hua Zhou, Mary E. Sehl, Janet S. Sinsheimer, and Kenneth Lange. Association screening of common and rare genetic variants by penalized regression. *Bioinformatics*, 26(19):2375, 2010.

Xide Zhu, Jin Zhang, Shangzhi Zeng, and Xiaoming Yuan. Linear convergence of R-BCPGM / prox-SVRG under bounded metric subregularity. *Manuscript*, 2018.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.