Discerning the linear convergence of ADMM for structured convex optimization through the lens of variational analysis

Xiaoming Yuan[*] Shangzhi Zeng[†] Jin Zhang[‡]

First version: August 2018 / second version: November 2018

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, 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 linear convergence of ADMM in the context of convex optimization through the lens of variaitonal analysis. We show that the 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 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. 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 Douglas-Rachford splitting method in the general operator form and the primal-dual hybrid gradient method for saddle-point problems.

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

2010 Mathematics Subject Classification: 90C06, 90C25, 90C52, 49J52, 49J53

[*] Department of Mathematics, The University of Hong Kong, Hong Kong, China. This author was supported by a General Research Fund from Hong Kong Research Grants Council. (xmyuan@hku.hk).

[†] Department of Mathematics, The University of Hong Kong, Hong Kong, China. (zengsz@connect.hku.hk).

[‡] Corresponding author. Department of Mathematics, Hong Kong Baptist University, Hong Kong, China. This author was supported by NSFC No. 11601458 (zhangjin@hkbu.edu.hk).

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 convex functions. 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 [9, 30] 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 [7, 15, 29] for some review papers. Even though the original ADMM is of our core interest, to capture its various variants simultaneously, as [35, 50], we study the so-called proximal version of ADMM with semi-positive-definite regularization terms for updating the primal variables (x, y) and Fortin and Glowinski's larger step size (see [18]) 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
    $x^{k+1} = \arg\min_x \{\theta_1(x) - \langle \lambda^k, Ax + By^k - b\rangle + \frac{\beta}{2}\|Ax + By^k - b\|^2 + \frac{1}{2}\|x - x^k\|_{G_1}^2\},$
    $y^{k+1} = \arg\min_y \{\theta_2(y) - \langle \lambda^k, Ax^{k+1} + By - b\rangle + \frac{\beta}{2}\|Ax^{k+1} + By - b\|^2 + \frac{1}{2}\|y - y^k\|_{G_2}^2\},$   (2)
    $\lambda^{k+1} = \lambda^k - \gamma\beta(Ax^{k+1} + By^{k+1} - b),$
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 [9, 30] 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 [18] 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 [16, 76, 78] 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 [37] is the special case of (2) with $G_1 \succ 0$, $G_2 \succ 0$ and $\gamma = 1$. Note that, as in [35, 50], 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, e.g. [52, 76, 78], for various applications of the linearized ADMM in areas such as statistical learning and computer vision, and [31, 38, 67, 77] for numerical acceleration performance of Fortin and Glowinski's larger step size $\gamma \in (0, \frac{1+\sqrt{5}}{2})$. Furthermore, as found in [21], the original ADMM is equivalent to the application of the general Douglas-Rachford splitting method (DRSM) proposed in [12, 53] to a stationary system to the dual of Problem (1); and as analyzed in [16, 66], 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 [8] for saddle-point problems. Finally, it is worthy mentioning that in some existing literatures such as [22, 69], 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 [28, 68].

Our primary purposes are: (1) discussing the linear convergence of the PADMM-FG (2) through the lens of variational analysis; and (2) showing 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. All the just-mentioned algorithms will be covered by our analysis.

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. [13, 14, 18, 22, 30, 31, 40, 53]. Recently, in [42, 43, 58], 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., [10, 11, 59]. Below we try to summarize some representative scenarios in which the 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. [11, 27].

(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. [11].

(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. [10].

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 example that these scenarios (S1)-(S3) might be too restrictive.

Example 1. *Consider the least absolute shrinkage and selection operator (LASSO) model proposed in [72] and its dual form:*

$$\min_{\mathbb{x}\in\mathbb{R}^m} \tfrac{1}{2}\|\mathbb{A}\mathbb{x}-\mathbb{b}\|_2^2 + \nu\|\mathbb{x}\|_1 \qquad \begin{array}{ll} \min_{\mathbb{y}\in\mathbb{R}^l} & \tfrac{1}{2}\|\mathbb{y}\|_2^2 - \mathbb{b}^T\mathbb{y} \\ s.t. & \|\mathbb{A}^T\mathbb{y}\|_\infty \le \nu, \end{array}$$

where $\mathbb{A} \in \mathbb{R}^{l\times m}$ 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 $\mathbb{A}$ 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 $\mathbb{z} = \mathbb{A}^T\mathbb{y}$ for the $l_\infty-$norm ball constraint, in general $\mathbb{A}^T$ does not meet the full row rank assumption.

[1] For succinctness, several scenarios discussed in [10, 11] 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 $\mathbb{A}$, the well observed linear convergence of ADMM can not 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. Recall that a set-valued map $\Psi : \mathbb{R}^n \rightrightarrows \mathbb{R}^q$ is said to be metrically subregular at $(\bar{u}, \bar{v}) \in \text{gph}\,(\Psi)$ if, for some $\epsilon > 0$, there exists $\kappa \geq 0$ such that

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

where $\text{dist}(d, \mathcal{D}) := \inf\{\|d - d'\| \,\big|\, d' \in \mathcal{D}\}$ for a given subset $\mathcal{D}$ and vector d in the same space, and $\mathbb{B}_{\epsilon}(\bar{u}) := \{u : \|u - \bar{u}\| < \epsilon\}$. A set-valued map $\Phi : \mathbb{R}^q \rightrightarrows \mathbb{R}^n$ is said to be calm (or pseudo upper-Lipschitz continuous, see [65, 80]) 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

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

It is well-known that a set-valued map is calm if and only if its inverse map is metrically subregular.

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

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

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 $\text{Res}(x, y, \lambda)$ as

$$\text{Res}(x, y, \lambda) = \text{dist}\left(0, T_{KKT}(x, y, \lambda)\right) \tag{4}$$

and use $\text{Res}(x, y, \lambda)$ to measure the optimality of the iterate (x, y, λ). In [79], 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 [55], the authors first take advantage of this observation to improve the results in [79]. In particular, let

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

it is shown in [55] 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 [19, page 83]) in terms of the Moreau-Yosida proximal mapping, the following mapping is used in [35, 36]:

$$T^p_{KKT}(x, y, \lambda) = \begin{pmatrix} x - \mathrm{Prox}_f(x + A^T\lambda) \\ y - \mathrm{Prox}_g(y + B^T\lambda) \\ Ax + By - b \end{pmatrix}, \tag{5}$$

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 [35], the linear convergence of PADMM-FG is proved when T^p_{KKT} is metrically subregular. Indeed, it is proved in [55] 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; while it is further justified in [55] 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. [61]. 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 [62, Proposition 1]. As a consequence, the 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

[2] A function $\phi : \mathbb{R}^n \to \mathbb{R}$ is called piecewise linear-quadratic if $dom\,\phi$ 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 $\frac{1}{2}\langle x, \mathbb{A}x\rangle + \langle \mathbb{a}, x\rangle + \mathbb{b}$ for some scalar $\mathbb{b} \in \mathbb{R}$, vector $\mathbb{a} \in \mathbb{R}^n$, and symmetric matrix $\mathbb{A} \in \mathbb{R}^n \times \mathbb{R}^n$. ϕ is a convex piecewise linear-quadratic function if and only if $\partial\phi$ is a polyhedral multifunction.

– $\{(Ax^k, By^k, \lambda^k)\}$ generated by the original ADMM converges linearly, see, e.g., [1, 55];

– $\{(x^k, By^k, \lambda^k)\}$ generated by the linearized ADMM converges linearly, see, e.g., [55, 79];

– $\{(x^k, y^k, \lambda^k)\}$ 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., [35].

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

Below we show some concrete examples to which the ADMM and its variants can be applied, while they are not the cases for which any of the scenarios (S1)-(S4) is effective.

Example 2. *([7, Section 11.2]) Variable selection in regularized logistic regression (RLR):*

$$\min_{x} \quad \sum_{j} \left(\log \left(1 + e^{\mathbb{A}_j^T x} \right) - \mathbb{b}_j \mathbb{A}_j^T x \right) + g(x),$$

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

$$\begin{aligned} \min_{x,y} \quad & \sum_{j} \left(\log \left(1 + e^{\mathbb{A}_j^T x} \right) - \mathbb{b}_j \mathbb{A}_j^T x \right) + g(y) \\ s.t. \quad & x = y. \end{aligned} \tag{6}$$

This setting covers the scenarios where $g(x)$ *is the* l_1 *regularizer, the* $l_\infty-$*norm regularizer, the fused LASSO regularizer (see, e.g., [72]), the octagonal selection and clustering algorithm for regression (OSCAR) (see, e.g., [5]). The definitions of these polyhedral convex regularizers are summarized in Table 1, where* μ, μ_1 *and* μ_2 *are given nonnegative parameters. Obviously, for this example, the strong convexity does not hold unless* $\mathbb{A}_j$ *is of full column rank for each* j.

Table 1: Polyhedral convex regularizers

Regularizers	l_1-norm	$l_\infty-$norm	fused LASSO	OSCAR
$g(x)$	$\mu\|x\|_1$	$\mu\|x\|_\infty$	$\mu_1\|x\|_1 + \mu_2 \sum\limits_i \|x_i - x_{i+1}\|$	$\lambda_1\|x\|_1 + \mu_2 \sum\limits_{i<j} \max\{\|x_i\|, \|x_j\|\}$

Example 3. *([47]) The penalized and constrained (PAC) regression for computing the penalized coefficient paths on high-dimensional generalized linear model*[3]*:*

$$\begin{aligned} \min_{x,y,z} \quad & \sum_j \left(-\log\left(\mathbb{A}_j^T x\right) + \mathbb{b}_j \mathbb{A}_j^T x\right) + g(y) + \delta_{\mathbb{R}_+^{l_2}}(z) \\ s.t. \quad & x = y, \;\; \mathbb{C}x + z = \mathbb{d}, \end{aligned} \tag{7}$$

where $\mathbb{A} \in \mathbb{R}^{l_1 \times m}$, $\mathbb{C} \in \mathbb{R}^{l_2 \times m}$, $\mathbb{b} \in \mathbb{R}_+^{l_1}$ and $\mathbb{d} \in \mathbb{R}^{l_2}$ are predefined matrices and vectors, g is convex polyhedral.

Obviously, assumptions S(1)-S(4) do not hold for this example.

Example 4. *([17, 49] and [7, Sections 11.3]) The $\ell_{1,q}$-norm regularizer with $q \in [1, 2]$:*

$$\begin{aligned} \min_{x,y} \quad & \frac{1}{2}\|\mathbb{A}x - \mathbb{b}\|_2^2 + \sum_{J \in \mathcal{J}} \omega_J \|y_J\|_q \\ s.t. \quad & x = y, \end{aligned} \tag{8}$$

where $\omega_J \geq 0$, $\mathcal{J}$ is a partition of $\{1, \ldots, n\}$ and $\|\cdot\|_q$ denotes the ℓ_q-norm, i.e., $\|x\|_q := \left(\sum_{i=1}^n |x_i|^q\right)^{\frac{1}{q}}$. When $q = 2$, the $\ell_{1,q}$-norm reduces to the group LASSO regularizer which was introduced in [83] in order to allow predefined groups of covariates $\mathcal{J}$ to be selected into or out of a model together. In general g is not a convex piecewise linear-quadratic function unless it degenerates to the l_1 regularizer.

Example 5. *([20, 84]) The sparse-group LASSO model:*

$$\begin{aligned} \min_{x,y} \quad & \frac{1}{2}\|\mathbb{A}x - \mathbb{b}\|_2^2 + \mu\|y\|_1 + \sum_{J \in \mathcal{J}} \omega_J \|y_J\|_2 \\ s.t. \quad & x = y, \end{aligned} \tag{9}$$

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.

The ADMM and its variants have been shown to perform linear convergence for these applications, see, e.g., [7, Sections 11.2,11.3]; and as shown, existing results fail to explain the linear convergence. We are thus motivated to answer the following questions:

> Is it still possible that the PADMM-FG (2) converges linearly in absence of both the strong convexity together with the full rank assumption of the coefficient matrices and the full polyhedricity assumption; and if yes, how to discern the linear convergence?

[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.

We answer these questions affirmatively, and show particularly how to discern the linear convergence of PADMM-FG (2) for a wide range of applications with strong interests 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 nonemptyness 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 $\mathbb{K} \subset dom\,\partial f$, f 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) $\mathcal{R}(L) \cap int(dom\,h) \neq \varnothing$.

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 2, 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 f* $\neq \varnothing$.

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.*

Table 2: Some commonly used loss functions h

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

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 1.1. *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 (6) and PAC model (7). 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

As mentioned, our main purposes are discussing the linear convergence of the PADMM-FG (2) through the lens of variational analysis, and developing new techniques that can be used to discern its linear convergence for an array of concrete applications, including the mentioned RLR, PAC, $\ell_{1,q}$-norm regularized regression and sparse-group LASSO models, which can not be covered by Scenarios (S1)-(S4). 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 (4) 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 3.

- **Discerning the linear convergence of original ADMM**. For the original ADMM where $G_1 = G_2 = 0$ 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 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 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 and B is of full column rank;

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 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 $\{(x^k, y^k, \lambda^k)\}$ converges linearly.

Table 3: Summary of linear convergence for the PADMM-FG (2)

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

In Table 4, we further specify the theory in Table 3 for some concrete applications in statistical learning 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, linearized ADMM and the general PADMM-FG (2) with the conditions $\beta A^TA + G_1 \succ 0$ and $\beta B^TB + G_2 \succ 0$, 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 1.2. *In Table 4, we mark with a thick line box the full polyhedricity case where $f(x) = \frac{1}{2}\|\mathbb{A}x - \mathbb{b}\|_2^2$ 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. [1, 35, 55, 79]. 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).*

Remark 1.3. *Even for the full polyhedricity case, in [55, 79], 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 $\{(x^k, y^k, \lambda^k)\}$*

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

g(y) f(x)	$\mu\Vert y\Vert_1$		$\mu\Vert y\Vert_\infty$		$\mu_1\Vert y\Vert_1$ $+\mu_2\sum_i \vert y_i - y_{i+1}\vert$		$\mu_1\Vert y\Vert_1$ $+\mu_2\sum_{i<j}\max\{\vert y_i\vert, \vert y_j\vert\}$		$\sum_{J\in\mathcal{J}}\omega_J\Vert y_J\Vert_q$ $(1\leq q\leq 2)$		$\mu\Vert y\Vert_1+\sum_{J\in\mathcal{J}}\omega_J\Vert y_J\Vert_2$	
$\frac{1}{2}\Vert\mathbb{A}x-\mathbb{b}\Vert_2^2$		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
$\sum_j\Big(\log\big(1+e^{\mathbb{A}_j^T x}\big)$ $-\mathbb{b}_j\mathbb{A}_j^T x\Big)$		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 $\{(x^k, y^k, \lambda^k)\}$	full column rank B	pADMM $\{(x^k, y^k, \lambda^k)\}$	full column rank B	pADMM $\{(x^k, y^k, \lambda^k)\}$	full column rank B	pADMM $\{(x^k, y^k, \lambda^k)\}$	full column rank B	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
$\sum_j\Big(-\log\big(\mathbb{A}_j^T x\big)$ $+\mathbb{b}_j\mathbb{A}_j^T x\Big)$		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 $\{(x^k, y^k, \lambda^k)\}$	full column rank B	pADMM $\{(x^k, y^k, \lambda^k)\}$	full column rank B	pADMM $\{(x^k, y^k, \lambda^k)\}$	full column rank B	pADMM $\{(x^k, y^k, \lambda^k)\}$	full column rank B	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

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

generated by the original ADMM and linearized ADMM, respectively. In our analysis, for the full polyhedricity case, we only delineate the linear convergence of the original ADMM in terms of the sequence $\{\lambda^k\}$ and the linearized ADMM in terms of the sequence $\{(x^k, \lambda^k)\}$, respectively. It is trivial to deduce that, under similar assumptions of the convergence of the sequence $\{(x^k, y^k, \lambda^k)\}$ as those in [55, 79], 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 linearized ADMM, respectively.

Remark 1.4. *In [35], 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 [35]. 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.*

1.5 Insights

Although the original ADMM and 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 linearized ADMM. Indeed, we shall show that individual treatments on the original ADMM and 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 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 4.1. 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 work [75], 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; and the analysis should be conducted case by case because of their significant differences.

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 usually more challenging to verify some subregularity conditions than to probe such conditions to theoretically ensure the linear convergence. 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., [23, 24, 26, 34, 44, 45, 80, 82]. 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 [1, 35, 51, 73, 74], 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 applications including those listed in Table 4.

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. 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.

1.6 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.

1.7 Notations

Throughout the paper, $\mathbb{R}^n$ denotes an n-dimensional Euclidean space with inner-product $\langle \cdot, \cdot \rangle$. The Euclidean norm is denoted by either $\| \cdot \|$ or $\| \cdot \|_2$. The l_1 norm and l_∞ norm are denoted by $\|x\|_1$ and $\|x\|_\infty$, respectively. $\mathbb{B}_r(x)$ denotes the open ball around x with radius $r > 0$. The open and closed unit ball centered at the origin are denoted by $\mathbb{B}$ and $\overline{\mathbb{B}}$, respectively. For a given subset $\mathcal{D} \subseteq \mathbb{R}^n$, int $\mathcal{D}$ denotes its interior, ri $\mathcal{D}$ denotes its relative interior, bd $\mathcal{D}$ denotes its boundary, $\mathrm{dist}(d, \mathcal{D}) := \inf\{\|d - d'\| \,|\, d' \in \mathcal{D}\}$ denotes the distance from a point d in the same space to $\mathcal{D}$, and $\delta_{\mathcal{D}}(x) := \begin{cases} 0 & \text{if } x \in \mathcal{D} \\ \infty & \text{if } x \notin \mathcal{D} \end{cases}$ denotes the indicator function of $\mathcal{D}$. Let $\Phi : \mathbb{R}^n \rightrightarrows \mathbb{R}^q$ be a set-valued map (multifunction), and its graph is defined by $\mathrm{gph}\,(\Phi) := \{(x, v) \in \mathbb{R}^n \times \mathbb{R}^q \mid v \in \Phi(x)\}$. The inverse mapping of Φ, denoted by Φ^{-1} is defined by $\Phi^{-1}(v) := \{x \in \mathbb{R}^n \mid v \in \Phi(x)\}$. For any matrix $A \in \mathbb{R}^{m\times n}$, $\mathcal{R}(A) \subseteq \mathbb{R}^m$ denotes the range of matrix A, $\mathcal{N}(A) \subseteq \mathbb{R}^n$ denotes the null space of matrix A, and $\|A\| := \max_{x\neq 0} \frac{\|Ax\|}{\|x\|}$. For a convex function ϕ, $\partial\phi$ denotes its subdifferential, i.e., $\partial\phi(x) := \{\xi \mid f(y) \geq f(x) + \langle \xi, y-x\rangle, \quad \forall y\}$ and $dom\,\phi$ denotes the domain of ϕ, i.e. $dom\,\phi := \{x \mid \phi(x) < +\infty\}$ and ϕ^* denotes the conjugate of ϕ, i.e., $\phi^*(\mu) := \sup_x\{\langle\mu, x\rangle - \phi(x)\}$. For any subspace $\mathcal{V}_0$, $\mathcal{V}_0^\perp$ denotes the orthogonal complement of $\mathcal{V}_0$. For given two sets $\mathcal{A}$ and $\mathcal{B}$, we denote $\mathcal{A}+\mathcal{B} := \{c : c = a+b, a\in\mathcal{A}, b\in\mathcal{B}\}$.

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 2.6, using the perturbation analysis techniques, we introduce the set-valued mappings $\mathcal{T}_1$ defined in (11) and $\mathcal{T}_2$ defined in (12) that are tailored for the iterative scheme of the DRSM. In Theorem 2.1 and Corollary 2.2, 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 (14). 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 [60, Proposition 1]. Similarly, we re-characterize the calmness of $\mathcal{T}_2$ with the calmness of the structured $\tilde{\Gamma}_{DR}$ defined in (16). 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.

Remark 2.5. *In [1], 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,*

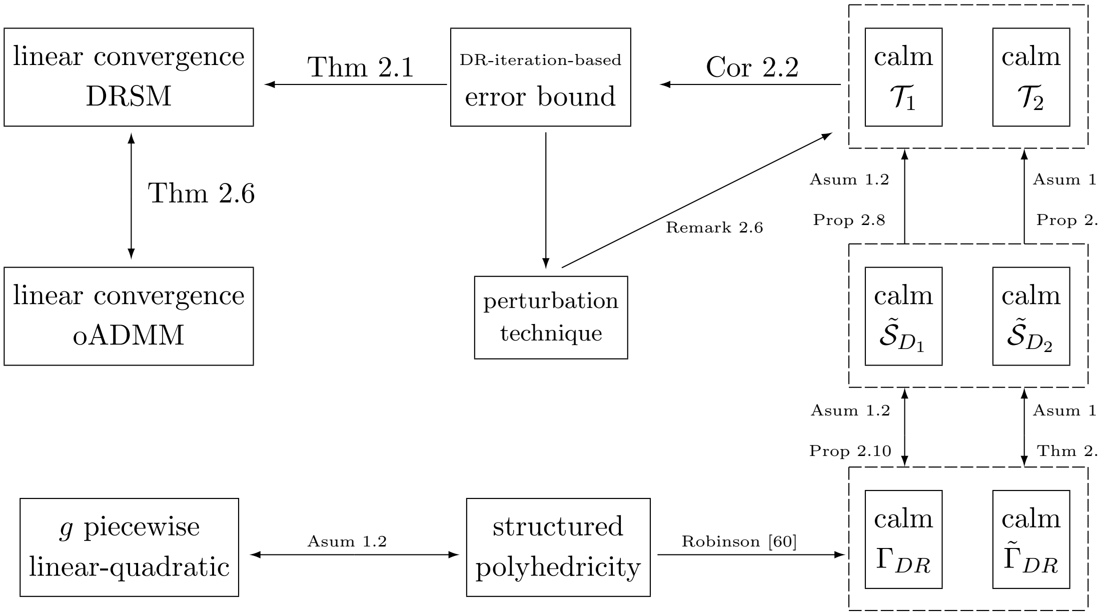

Figure 1: Roadmap to study linear convergence of the original ADMM

by noticing that the DR operator is a composite of two operators, we define the algorithm-tailored perturbed solution set-valued maps (11) *and* (12)*, 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 (6) and the PAC model (7). On the other hand, the linear convergence of DRSM is investigated under the strong monotonicity assumption in [27]. 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 [21], 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 2.1. *Let $\{(x^k, y^k, \lambda^k)\}$ be the sequence generated by the original ADMM. Define $z^k := \lambda^k + \beta By^k$, $u^k := \lambda^k$ 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:*

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

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 [42, 43], 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 [3] as well.

Proposition 2.2. *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 2.3. *[2, Theorem 25.6][3, 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*

$$\begin{aligned} &\|\mathrm{prox}_{\phi_2}(z^{k+1}) - \mathrm{prox}_{\phi_2}(z^*)\|^2 + \|\mathrm{prox}_{\phi_2{}^*}(z^{k+1}) - \mathrm{prox}_{\phi_2{}^*}(z^*)\|^2 \\ \le &\|\mathrm{prox}_{\phi_2}(z^k) - \mathrm{prox}_{\phi_2}(z^*)\|^2 + \|\mathrm{prox}_{\phi_2{}^*}(z^k) - \mathrm{prox}_{\phi_2{}^*}(z^*)\|^2 - \|z^{k+1} - z^k\|^2, \qquad \forall k. \end{aligned} \tag{10}$$

Definition 2.1 (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*

$$\mathrm{dist}\left(\mathrm{prox}_{\phi_2}(z^k), Z\right) + \mathrm{dist}\left(\mathrm{prox}_{\phi_2{}^*}(z^k), W\right) \le \kappa \left\|z^{k+1} - z^k\right\|, \quad \text{for all } k \text{ such that } z^k \in \mathbb{B}(\bar{z}, \epsilon).$$

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 2.3, $\{z^k\}$ converges to some point $\bar{z} \in Fix(T_{DR})$.

Theorem 2.1 (Linear convergence of the DRSM under the DR-iteration-based error bound)**.** *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* $k_0 > 0$ *and* $0 < \rho < 1$ *such that, for all* $k \ge k_0$*, it holds*

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

and thus there exists $C_0 > 0$ *such that*

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

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

and

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

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:

$$\mathcal{T}_1(p) := \left\{\lambda \mid p \in \partial\phi_1(\lambda - p) + \partial\phi_2(\lambda)\right\}, \tag{11}$$

and

$$\mathcal{T}_2(p) := \left\{\mu \mid p \in \partial(\phi_1^* \circ -Id)(\mu - p) + \partial\phi_2^*(\mu)\right\}. \tag{12}$$

Remark 2.6 (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 [3], 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 Corollary 2.2, 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 [75], 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 2.3, $\{z^k\}$ converges to some point $\bar{z} \in Fix(T_{DR})$.

Corollary 2.2 (Linear convergence of DRSM under the calmness of $\mathcal{T}_1$ and $\mathcal{T}_2$)**.** *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 $0 < \rho < 1$, such that, for all $k \geq k_0$, it holds that*

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

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

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

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 2.1. *Let $\mathbb{h}$ 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 $dom\psi^* \subset \mathcal{R}(\mathbb{L}^T) + \mathcal{A}_0^{\perp}$.*

We recall a proposition given in [32, Corollary 4.4].

Proposition 2.4. *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 [64, Theorem 26.1].

Proposition 2.5. *Let ϕ be a lower semicontinuous, convex function, if ϕ is essentially differentiable, then*

$$\mathit{dom}\,\partial\phi = \mathit{int}\,\mathit{dom}\,\phi,$$

and

$$\partial\phi(x) = \nabla\phi(x), \quad \text{when } \; x \in \mathit{int}\,\mathit{dom}\,\phi.$$

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 2.6. *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*

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

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

$$\partial\psi(x) = \mathbb{L}^T\nabla\mathbb{h}(\mathbb{L}x) + \mathcal{N}_{\mathcal{A}}(x), \qquad \mathit{dom}\partial\psi \neq \varnothing,$$

then

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

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 2.2 inspired by Assumption 1.2.

In particular, according to [64, 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 2.2. *Suppose that f meets Assumption 1.2, then $\partial f(x) = \partial\left(h(Lx)\right) + q = L^T\partial h(Lx) + q$.*

Lemma 2.3. *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, vector $\tilde{q}$ and affine space $\mathcal{V}$. Furthermore, we have $\mathit{dom}\partial\phi_1 \neq \varnothing$,

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

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

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

According to [26, Proposition 3], we have the following result whose proof is analogous to that of [26, Proposition 3] and thus omitted.

Proposition 2.7. *Let X and Y be metric spaces, $Q_1 : X \to Y$ and $Q_2 : X \rightrightarrows Y$ be set-valued maps with closed graphs. Further assume that Q_1 is Lipschitz continuous near $\bar{x} \in X$. Then the set-valued map*

$$M_1(x) := Q_1(x) + Q_2(x),$$

is metrically subregular at $(\bar{x}, 0)$ if and only if the set-valued map

$$M_2(x) := (x, -Q_1(x)) - gph\, Q_2$$

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

Proposition 2.8. *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 $(0, \bar{\lambda})$.*

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 2.3, we have the following result.

Lemma 2.4. *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,

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

Similar to Proposition 2.8, Lemma 2.4 inspires the following sufficiency.

Proposition 2.9. *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 2.5. *If Assumptions 1.1 and 1.2 hold for Problem (1), there exist $\bar{t}$, $\bar{g} \in \mathbb{R}^n$ such that*

$$
Z = \{\lambda \in \mathcal{V} \mid K\lambda = \bar{t}, \quad 0 \in \bar{g} + \mathcal{V}_0^{\perp} + \partial\phi_2(\lambda)\}. \tag{13}
$$

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

$$
\Gamma_{DR}(p_1, p_2) := \Gamma_1(p_1) \cap \Gamma_2(p_2) = \{\lambda \in \mathcal{V} \mid p_1 = K\lambda - \bar{t}, \quad p_2 \in \bar{g} + \mathcal{V}_0^{\perp} + \partial\phi_2(\lambda)\}, \tag{14}
$$

where

$$
\Gamma_1(p_1) := \{\lambda \mid p_1 = K\lambda - \bar{t}\}, \qquad \Gamma_2(p_2) := \{\lambda \in \mathcal{V} \mid p_2 \in \bar{g} + \mathcal{V}_0^{\perp} + \partial\phi_2(\lambda)\}. \tag{15}
$$

Since $\Gamma_{DR}(0,0) = Z$, $\Gamma_{DR}(p_1, p_2)$ can be considered as a set-valued map which perturbs Z in (13). 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 2.10. *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* $dist(\lambda, \Gamma_{DR}(0,0)) \leq \kappa_1 dist\left(0, \Gamma_{DR}^{-1}(\lambda)\right), \qquad \forall \lambda \in \mathbb{B}_{\epsilon_1}(\bar{\lambda})$*;*

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

The equivalence in Proposition 2.10 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 2.3. *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* $(\bar{\lambda}, 0)$*.*

Combining Proposition 2.8 and Theorem 2.3, we are able to verify the desired calmness of $\mathcal{T}_1$ under the structured polyhedricity assumption.

Theorem 2.4. *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 2.4, first with Lemma 2.4, 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

$$\tilde{\Gamma}_{DR}(p_1, p_2) := \{\mu \in \hat{\mathcal{V}} \mid p_1 = \hat{K}\mu - \hat{t}, \quad p_2 \in \hat{g} + \hat{\mathcal{V}}_0^{\perp} + \partial\phi_2^*(\mu)\} \tag{16}$$

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 2.10, 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 2.9, we have the desired calmness of $\mathcal{T}_2$.

Theorem 2.5. *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 2.4, Theorem 2.5, and Corollary 2.2. Below, we show the linear convergence of the original ADMM in sense of the dual variable sequence $\{\lambda^k\}$, the KKT residue sequence $\{\mathrm{Res}^k\}$, and the objective function value sequence $\{\mathrm{Val}^k\}$ together with the constraint feasibility sequence $\{\mathrm{Fea}^k\}$, by simply using Corollary 2.2.

Theorem 2.6. *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*

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

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).

As analyzed in [1], 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 [1] can be covered by our analysis.

Corollary 2.7. *In addition to the assumptions in Proposition 2.6, 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 $\Omega_x^* := \{x \mid \exists y, \lambda$ such that $(x, y, \lambda) \in \Omega^*\}$ and $\Omega_y^* := \{y \mid \exists x, \lambda$ such that $(x, y, \lambda) \in \Omega^*\}$.

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. [55, 79]. 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 3.7, the perturbation analysis consideration inspires us to determine the metric subregularity of set-valued map $T(x, \lambda)$ defined in (20) 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 3.4). 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 [60, Proposition 1].

It is worthy 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 [48, Theorem 3.6]. The metric subregularity of T is thereby re-characterized in terms of the calmness of $\hat{\Omega}_x$ defined in (34). The calmness of $\hat{\Omega}_x$ eventually follows directly from [60, Proposition 1].

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

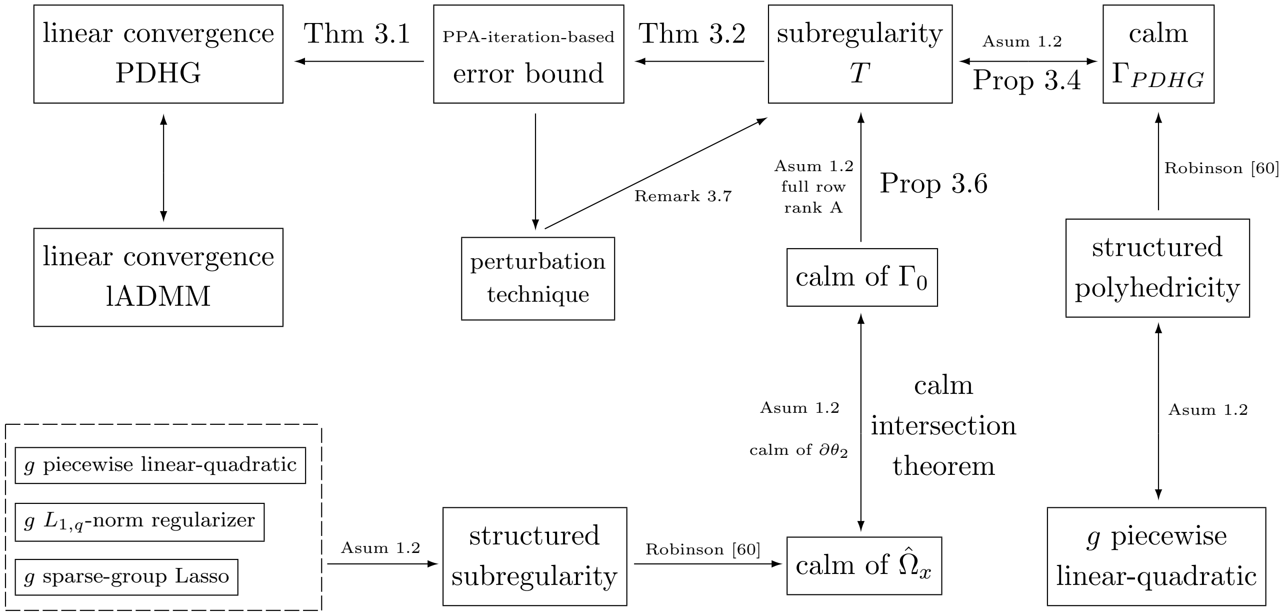

Figure 2: Roadmap to study linear convergence of the linearized ADMM

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):

$$\min_x \max_\lambda \ \theta(x, \lambda) := f(x) - \langle \lambda, Ax \rangle - g^*(B^T\lambda) + \langle b, \lambda \rangle. \tag{17}$$

Let us denote

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

Then (17) can be rewritten into the following compact form

$$\min_x \max_\lambda \ \theta(x, \lambda) := \theta_1(x) + \langle \lambda, \mathcal{K}x \rangle - \theta_2(\lambda). \tag{18}$$

As analyzed in [16, 66], the linearized ADMM applied to Problem (1) turns out to be highly relevant to the application of the PDHG to the saddle-point problem (17). In fact, for the iterative (x^k, y^k, λ^k) generated by 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 $\{(x^k, \lambda^k)\}$ generated by linearized ADMM coincides with the sequence generated by the PDHG applied to (17), i.e.,

$$\begin{cases} x^{k+1} = \operatorname{argmin}_x \ f(x) - \langle 2\lambda^k - \lambda^{k-1}, Ax \rangle + \frac{r}{2}\|x - x^k\|^2, \\ \lambda^{k+1} = \operatorname{argmin}_\lambda \ g^*(B^T\lambda) - \langle b, \lambda \rangle + \langle Ax^{k+1}, \lambda \rangle + \frac{1}{2\beta}\|\lambda - \lambda^k\|^2. \end{cases}$$

At the k-th iteration of the PDHG, it follows from the optimality conditions of its subproblems that

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

which can be further expressed in a more compact form

$$0 \in T(x^{k+1}, \lambda^{k+1}) + \mathcal{M}[(x^{k+1}, \lambda^{k+1}) - (x^k, \lambda^k)], \tag{19}$$

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:

$$\mathcal{M} := \begin{pmatrix} \frac{1}{\tau}I & -\mathcal{K}^T \\ -\mathcal{K} & \frac{1}{\sigma} \end{pmatrix} \quad \text{and} \quad T(x, \lambda) := \begin{pmatrix} \partial\theta_1(x) + \mathcal{K}^T\lambda \\ \partial\theta_2(\lambda) - \mathcal{K}x \end{pmatrix}. \tag{20}$$

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 (18) 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., [6, 16, 39, 41]. Among them is [41] which is the first work showing the close connection between the PDHG and the well-known proximal point algorithm (PPA) proposed in [57, 63], as well as revisiting the PDHG from the contraction perspective for convergence analysis (see Proposition 3.1). 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., [6, 73].

Our approach to studying the linear convergence of the PDHG is motivated by the explanation initiated in [41] 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

$$0 \in T(x, \lambda), \tag{21}$$

where T is defined as in (20). 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 $0 \in \mathbb{T}(x)$ where $\mathbb{T}$ is a maximally monotone operator. For this purpose, let $\{\mathbb{x}^k\}$ be the sequence generated by the PPA. The iterative scheme reads as:

$$0 \in \mathbb{T}(\mathbb{x}^{k+1}) + \mathbb{M}(\mathbb{x}^{k+1} - \mathbb{x}^k), \tag{22}$$

where $\mathbb{M}$ is a positive definite matrix. Based on the convergence analysis of PPA given in the literature, e.g., [33, 63, 70], the sequence $\{\mathbb{x}^k\}$ converges to some point $\bar{\mathbb{x}} \in \mathcal{S}_{\mathbb{T}}$ where $\mathcal{S}_{\mathbb{T}} := \{\mathbb{x} \mid 0 \in \mathbb{T}(\mathbb{x})\}$, and we are going to derive the linear convergence of $\{\mathbb{x}^k\}$ toward $\mathcal{S}_{\mathbb{T}}$ under the following error bound condition. Let the sequence $\{\mathbb{x}^k\}$ be generated by the PPA iterative scheme (22); and it converges to some point $\bar{\mathbb{x}} \in \mathcal{S}_{\mathbb{T}}$.

Definition 3.1 (PPA-iteration-based error bound)**.** *We say that the PPA-iteration-based error bound holds at $\bar{\mathbb{x}}$ if there exist $\epsilon, \kappa > 0$ such that*

$$\operatorname{dist}_{\mathbb{M}}\left(\mathbb{x}^{k+1}, \mathcal{S}_{\mathbb{T}}\right) \leq \kappa \|\mathbb{x}^{k+1} - \mathbb{x}^k\|_{\mathbb{M}}, \quad \textit{for all } k \textit{ such that } \mathbb{x}^k \in \mathbb{B}(\bar{\mathbb{x}}, \epsilon).$$

The following PPA linear convergence relies heavily on [54]. The proof is needed for our further discussion and hence stated here.

Theorem 3.1. *Assume that the PPA-iteration-based error bound holds at $\bar{\mathbb{x}}$, and then the sequence $\{\mathbb{x}^k\}$ converges to $\mathcal{S}_{\mathbb{T}}$ 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*

$$\operatorname{dist}_{\mathbb{M}}\left(\mathbb{x}^{k+1}, \mathcal{S}_{\mathbb{T}}\right) \leq \rho \, dist_{\mathbb{M}}\left(\mathbb{x}^k, \mathcal{S}_{\mathbb{T}}\right). \tag{23}$$

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

$$\operatorname{dist}\left(\mathbb{x}^k, \mathcal{S}_{\mathbb{T}}\right) \leq C_0 \rho^k, \tag{24}$$

and

$$\|\mathbb{x}^{k+1} - \mathbb{x}^k\| \leq C_0 \rho^k. \tag{25}$$

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 3.7 (Perturbation perspective)**.** *Following the perturbation analysis technique in [75], 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

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

Thus we consider the metric subregularity of set-valued mapping $\mathbb{T}$ in Theorem 3.2.

Theorem 3.2. *If $\mathbb{T}$ 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 $\mathcal{S}_{\mathbb{T}}$ linearly. That is, there exist $k_0 > 0$ and $0 < \rho < 1$ such that, for all $k \geq k_0$, it holds that*

$$\operatorname{dist}_{\mathbb{M}}\left(\mathbb{x}^{k+1}, \mathcal{S}_{\mathbb{T}}\right) \leq \rho \operatorname{dist}_{\mathbb{M}}\left(\mathbb{x}^{k}, \mathcal{S}_{\mathbb{T}}\right). \tag{26}$$

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

$$\operatorname{dist}\left(\mathbb{x}^{k}, \mathcal{S}_{\mathbb{T}}\right) \leq C_0 \rho^k, \tag{27}$$

and

$$\|\mathbb{x}^{k+1} - \mathbb{x}^k\| \leq C_0 \rho^k. \tag{28}$$

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 3.1 ([8, 41])**.** *Let $\{(x^k, \lambda^k)\}$ be the sequence generated by the PDHG applied to the saddle-point problem (18). If $\tau\sigma < \frac{1}{\|A^TA\|}$, then the sequence $\{(x^k, \lambda^k)\}$ converges to some point $(\bar{x}, \bar{\lambda}) \in \Omega^*_{x,\lambda}$.*

With the given convergence of the sequence $\{(x^k, \lambda^k)\}$ generated by the PDHG applied to (18), the linear convergence of $\{(x^k, \lambda^k)\}$ can be achieved according to Theorem 3.2, with the consideration that $\{(x^k, \lambda^k)\}$ can also be regarded as the sequence generated by the PPA applied to (21). Note that when $\tau\sigma < \frac{1}{\|A^TA\|}$, $\mathcal{M}$ defined by (20) is positive definite. Then, the desired linear convergence of the PDHG follows immediately from the discussion above.

Theorem 3.3. *Suppose the sequence $\{(x^k, \lambda^k)\}$ generated by PDHG with $\tau\sigma < \frac{1}{\|A^TA\|}$. Then according to Proposition3.1, $\{(x^k, \lambda^k)\}$ converges to some point $(\bar{x}, \bar{\lambda}) \in \Omega^*_{x,\lambda}$. If T defined by (20) is metrically subregular at $(\bar{x}, \bar{\lambda}, 0)$ with modulus κ, then the sequence $\{(x^k, \lambda^k)\}$ 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*

$$\operatorname{dist}_{\mathcal{M}}\left((x^{k+1}, \lambda^{k+1}), \Omega^*_{x,\lambda}\right) \leq \rho\, dist_{\mathcal{M}}\left((x^{k}, \lambda^{k}), \Omega^*_{x,\lambda}\right). \tag{29}$$

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

$$\operatorname{dist}\left((x^{k}, \lambda^{k}), \Omega^*_{x,\lambda}\right) \leq C_0 \rho^k, \tag{30}$$

and

$$\|x^{k+1} - x^k\| + \|\lambda^{k+1} - \lambda^k\| \leq C_0 \rho^k. \tag{31}$$

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 3.2. *[4, Proposition 2.6.1] When a saddle-point of the min-max problem* (18) *exists, the set of saddle-points $\Omega^*_{x,\lambda}$ for* (18) *can be characterized by $X \times \Lambda$ with*

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

and

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

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 3.2 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 3.3. *When Problem (1) meets Assumption 1.2, the saddle-point set $\Omega^*_{x,\lambda}$ can be characterized as*

$$\Omega^*_{x,\lambda} = \{(x,\lambda) \mid Lx = \bar{t},\ \mathcal{K}^T\lambda = -\bar{s},\ 0 \in \partial\theta_2(\lambda) - \mathcal{K}x\}, \tag{32}$$

with some vector $\bar{t} \in \mathbb{R}^l$ and $\bar{s} := L^T\nabla\theta_1(\bar{t}) + q$.

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 (32):

$$\Gamma_{PDHG}(p) := \{(x,\lambda) \mid p_1 = Lx - \bar{t}, p_2 = \bar{s} + \mathcal{K}^T\lambda, p_3 \in \partial\theta_2(\lambda) - \mathcal{K}x\}.$$

Obviously, $\Gamma_{PDHG}(p)$ coincides with $\Omega^*_{x,\lambda}$ when $p = 0$. Similar to [81, Proposition 4.1], we have following equivalence.

Proposition 3.4. *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

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

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

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

3.4.2 Sufficient condition for the metric subregularity of T

Thanks to Proposition 3.3, when A is of full row rank, we can easily obtain another characterization of $\Omega^*_{x,\lambda}$.

Proposition 3.5. *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*

$$\Omega^*_{x,\lambda} = \{(x,\lambda) \mid Lx = \bar{t},\ \lambda = \bar{\lambda},\ 0 \in D - \mathcal{K}x\}, \tag{33}$$

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

We introduce an auxiliary set-valued map associated with characterization of $\Omega^*_{x,\lambda}$ in (33):

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

A useful connection is clarified below.

Proposition 3.6. *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 $(\bar{\lambda}, \mathcal{K}\bar{x})$ and there exist $\kappa_1, \epsilon_1 > 0$ such that*

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

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

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

Now, we are going to present a lemma for studying the metric subregularity of Γ_0.

Lemma 3.1. *Let $\mathbb{E}$ ba a matrix, and $\mathbb{D}$ be a closed subset. If $\mathbb{D} \subseteq \mathcal{R}(\mathbb{E})$, the multifunction $\mathcal{M}(x) := \mathbb{E}x - \mathbb{D}$ is metrically subregular at $(\bar{x}, 0)$ for any $\bar{x} \in \mathcal{M}^{-1}(0) = \{x \mid \mathcal{M}(x) = 0\}$.*

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

$$\Omega_x(p) := \{x \mid p_1 = Lx - \bar{t}, p_2 \in D - \mathcal{K}x\}$$

at $(0, \bar{x})$. For studying the calmness of Ω_x, we recall the calm intersection theorem introduced in [48, Theorem 3.6],

Proposition 3.7 (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

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

First, according to [81], $(\Omega_x^1)^{-1}$ is metrically subregular and pseudo-Lipschitz continuous at any point on its graph. Combining Lemma 3.1, Propositions 3.6 and 3.7, we obtain a sufficient condition for the metric subregularity of T.

Theorem 3.4. *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 $(\bar{\lambda}, \mathcal{K}\bar{x})$ with modulus κ_2 and*

$$\hat{\Omega}_x(p_1) := \{x \mid p_1 = Lx - \bar{t},\ 0 \in D - \mathcal{K}x\} \tag{34}$$

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

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

where

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

In particular,

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

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

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.

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 3.4, together with Theorem 3.4, straightforwardly yields the following criteria for the metric subregularity of T.

Theorem 3.5. *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 .

Before giving applications of the criteria in Theorem 3.5, we need the following lemma.

Lemma 3.2. *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 $B^T\bar{z} = \bar{v}$, and then $B\partial g^* B^T$ is calm at $(\bar{z}, B\bar{y})$ with modulus $\kappa\|B\|^2$;

(3) in addition, when $\mathcal{R}(B^T) \cap ri(dom\, g^*) \neq \varnothing$, 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 $\kappa\|B\|^2$.

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 3.3. *[85] Let g represent the $\ell_{1,q}$-norm regularizer, i.e., $g(x) := \sum_{J\in\mathcal{J}} w_J \left\|x_J\right\|_q$ 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 $x \in \mathbb{B}_\epsilon(\bar{x})$,

$$\operatorname{dist}\left(x, (\partial g)^{-1}(\bar{s})\right) \leq \kappa_g \cdot \operatorname{dist}\left(\bar{s}, \partial g(x)\right), \tag{35}$$

where

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

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 $\mathbb{B}_\epsilon\left(\bar{x}_J\right)$.

Lemma 3.4. *[85] Let g denote the sparse-group LASSO regularizer, i.e., $g(x) := \sum_{J\in\mathcal{J}} w_J \left\|x_J\right\|_2 + \mu \cdot \left\|x\right\|_1$ 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 $x \in \mathbb{B}_\epsilon(\bar{x})$,

$$\operatorname{dist}\left(x, (\partial g)^{-1}(\bar{s})\right) \leq \kappa_g(\lambda) \cdot \operatorname{dist}\left(\bar{s}, \partial g(x)\right). \tag{36}$$

where

$$\kappa_g(\mu) := \max_{J\in\mathcal{J}}\{\kappa_J(\mu)\} \textit{ with } \kappa_J(\mu) = \begin{cases} 1 & \textit{if } w_J = 0, \\ 1 & \textit{if } w_J > 0 \textit{ and } \left\|\mathcal{T}_\mu(\bar{s}_J)\right\|_q < w_J, \\ \kappa_{J,1}(\mu) \cdot \kappa_{J,2}(\mu) \cdot w_J^{-1} & \textit{if } w_J > 0 \textit{ and } \left\|\mathcal{T}_\mu(\bar{s}_J)\right\|_q = w_J, \end{cases}$$

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 3.6. *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 $(\bar{x}, \bar{\lambda}, 0)$.

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 3.6. We next focus on calculating the metric subregularity modulus of T. Noting that $D = \partial\theta_2(\bar{\lambda})$, $\mathcal{K} = -A$, Theorem 3.7 then follows directly from Lemma 3.2 and Theorem 3.4.

Theorem 3.7. *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*

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

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 3.7, 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 (34), 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 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 [81, 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 RLR regression: we consider the RLR problem with l-1 norm regularizer

$$\begin{aligned} \min_{x,y} \quad & \sum_j \left(-\log\left(\mathbb{A}_j^T x\right) + \mathbb{b}_j \mathbb{A}_j^T x\right) + \mu\|y\|_1 \\ s.t. \quad & x = y, \end{aligned} \tag{37}$$

where $\mathbb{A} \in \mathbb{R}^{l_1 \times m}$, 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 [81, 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 [81, Lemma 4, Lemma 5],

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

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

In order to calculate the metric subregularity modulus of T, according to Theorem 3.7, 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:

$$\hat{\Omega}_x(p_1) := \{x \mid p_1 = \mathbb{A}x - \mathbb{A}\bar{x},\ -x = -D\alpha, \alpha \geq 0\} \tag{38}$$

We are in the position to apply Lemma 3.5 taken from [81] to calculate the calmness modulus of $\hat{\Omega}_x$. In fact, Lemma 3.5 can be regarded as a variant of Hoffman's error bound theory.

Lemma 3.5 (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

$$\bar{\theta}(\mathcal{M}) := \sup_{\lambda,\mu,\nu} \left\{ \|\lambda\| \;\middle|\; \begin{array}{l} \|\mathcal{M}(\lambda,\mu,\nu)\| = 1, \nu \geq 0, \\ \text{The corresponding rows of } \mathcal{M} \text{ to } \lambda,\mu,\nu\text{'s} \\ \text{ non-zero elements are linearly independent.} \end{array} \right\}. \tag{39}$$

Applying Lemma 3.5 to $\hat{\Omega}_x$ in (38), we obtain the following result.

Proposition 3.8. *For the RLR problem (37), $\hat{\Omega}_x$ 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

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

and $\bar{\theta}(\mathcal{M})$ is defined as in (39).

Theorem 3.8. *Consider the RLR problem (37). Suppose that $-\log$ is strongly convex on some neighborhood U_j of $\mathbb{A}_j^T\bar{x}$ 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,

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

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

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

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

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 Proposition3.1, $\{(x^k, \lambda^k)\}$ 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 $\{(x^k, \lambda^k)\}$, $\{\mathrm{Res}^k\}$, $\{\mathrm{Val}^k\}$ and $\{\mathrm{Fea}^k\}$.

Theorem 3.9. *If T is metrically subregular at $(\bar{x}, \bar{\lambda}, 0)$ with modulus κ, then the sequence $\{(x^k, \lambda^k)\}$ converges 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

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

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

and

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

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

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

and

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

Let $\{(x^k, y^k, \lambda^k)\}$ be the sequence generated by the linearized ADMM. Theorems 3.6 and 3.9 motivate the following corollary directly.

Corollary 3.10. *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 the sequence $\{(x^k, \lambda^k)\}$ converges to $\Omega^*_{x,\lambda}$ linearly. That is, there exist $k_0 > 0$, $C_0 > 0$ and computable $0 < \rho < 1$ such that, for all $k \ge k_0$, it holds that

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

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

and

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

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

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

4 Linear convergence rate of PADMM-FG

In the literature [55, 35, 79], 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 4.2, 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 $\{(x^k, y^k, \lambda^k)\}$, $\{\text{Res}^k\}$ and $\{(\text{Fea}^k, \text{Val}^k)\}$ can be guaranteed for a wide range of applications including the RLR model (6), the $\ell_{1,q}$-norm regularized regression with $1 \leq q \leq 2$ (8) and sparse-group LASSO (9).

4.1 Linear convergence of PADMM-FG under metric subregularity of T_{KKT}

The linear convergence of PADMM-FG (2) is shown in [35, Theorem 2] when the metric subregularity of T^p_{KKT} defined in (5) 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 [55], we have the following result.

Theorem 4.1. *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*

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

where the explicit expression of ρ which is characterized in terms of c_{KKT} can be found in [35, 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 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 4.2. *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*

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

Theorems 3.5 and 4.2 straightforwardly inspire the following criteria for the metric subregularity of T_{KKT}.

Theorem 4.3. *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 one of the following statements is satisfied:*

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

(2) Problem (1) meets the structured subregularity assumption at $(\bar{x}, \bar{y}, \bar{\lambda})$.

Motivated by the proof in Theorem 3.6, the linear convergence of the general PADMM-FG (2) can be obtained easily.

Theorem 4.4. *Suppose that Assumption 1.2 is satisfied and B is of full column rank. 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 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 $B^T\bar{\lambda} \neq 0$.

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 4.2, 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 assumption of their subdifferential. 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 statistical learning such as the RLR, PAC, $l_{1,q}$-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.

Appendix A. Proof of Proposition 2.1

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

$$\begin{cases} 0 \in \partial g(y^k) - B^T\lambda^k, \\ 0 \in \partial f(x^{k+1}) - A^T(\lambda^k - \beta(Ax^{k+1} + By^k - b)), \\ \lambda^{k+1} = \lambda^k - \beta(Ax^{k+1} + By^{k+1} - b). \end{cases} \tag{40}$$

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

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

Therefore, we have

$$\begin{cases} \lambda^k = (I + \beta \partial \phi_2)^{-1}(\lambda^k + \beta B y^k), \\ \lambda^k - \beta(Ax^{k+1} + By^k - b) = (I + \beta \partial \phi_1)^{-1}(\lambda^k - \beta B y^k), \\ \lambda^{k+1} + \beta B y^{k+1} = \lambda^k + \beta B y^k - \lambda^k + \lambda^k - \beta(Ax^{k+1} + By^k - b). \end{cases} \tag{41}$$

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

$$\begin{cases} u^k = (I + \beta \partial \phi_2)^{-1}(z^k), \\ v^k = (I + \beta \partial \phi_1)^{-1}(2u^k - z^k), \\ z^{k+1} = z^k - u^k + v^k, \end{cases} \tag{42}$$

and the conclusion follows.

Appendix B. Proof of Theorem 2.1

By Propositions 2.2 and 2.3 and the closedness of Z and W, we have

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

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

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

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

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

which implies that

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

where $C_0 = \sqrt{2}\Big(\operatorname{dist}\left(\operatorname{prox}_{\phi_2}(z^{k_0}), Z\right)^2 + \operatorname{dist}\left(\operatorname{prox}_{\phi_2^*}(z^{k_0}), W\right)^2\Big)^{\frac{1}{2}} (1 - \frac{1}{\kappa^2})^{-\frac{k_0}{2}}$ and $\rho = (1 - \frac{1}{\kappa^2})^{\frac{1}{2}}$. Then, it follows directly from (10) that

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

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

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

Appendix C. Proof of Corollary 2.2

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,

$$
z^k - u^k \in \partial\phi_2(u^k), \quad 2u^k - z^k - v^k \in \partial\phi_1(v^k). \tag{43}
$$

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

$$
\operatorname{dist}(u^k, Z) \leq \kappa_1\|u^k - v^k\|, \qquad \textit{when } u^k \in \mathbb{B}_{\epsilon_1}(\bar{\lambda}). \tag{44}
$$

Also, since $u^k = \operatorname{prox}_{\phi_2}(z^k)$, and $\|u^k - \bar{\lambda}\| = \|\operatorname{prox}_{\phi_2}(z^k) - \operatorname{prox}_{\phi_2}(\bar{z})\| \leq \|z^k - \bar{z}\|$, which comes from the nonexpansiveness of $\operatorname{prox}_{\phi_2}$, substituting $u^k - v^k = z^k - z^{k+1}$ in (44) enables us to obtain

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

Again, by (43) 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

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

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

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

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

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

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

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

According to Theorem 2.1, the sequence $\{z^k\}$ converges to $\bar{z}$ linearly.

Appendix D. Proof of Lemma 2.1

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

$$\begin{aligned}
\psi^*(y) &= \sup_x \{\langle y, x\rangle - \mathbb{h}(\mathbb{L}x) - \delta_{\mathcal{A}}(x)\} \\
&= \sup_x \{\langle y_1 + y_2, x\rangle - \mathbb{h}(\mathbb{L}x) - \delta_{\mathcal{A}}(x)\} \\
&\ge \sup_\alpha \{\langle y_1 + y_2, a + \alpha y_2\rangle - \mathbb{h}(\mathbb{L}a + \alpha \mathbb{L} y_2) - \delta_{\mathcal{A}}(a + \alpha y_2)\} \\
&= \sup_\alpha \{\langle y_1 + y_2, a\rangle - \mathbb{h}(\mathbb{L}a) + \alpha \|y_2\|^2 - \mathbb{h}(0)\} = +\infty.
\end{aligned}$$

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

Appendix E. Proof of Proposition 2.6

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 2.4, we have $\tilde{\mathbb{h}}^* \in \mathcal{C}$.

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

$$y = \mathrm{Proj}_{\mathcal{R}(\mathbb{L}^T)}y + (I - \mathrm{Proj}_{\mathcal{R}(\mathbb{L}^T)})y.$$

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 $y \in \mathcal{R}(\mathbb{L}^T) + \mathcal{A}_0^{\perp}$ as

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

where $\tilde{y} := \mathbb{L}^T F y \in \mathcal{R}(\mathbb{L}^T)$ with

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

and

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

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

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

such that $\mathbb{L}x = \mathbb{L}a + U_{\mathbb{L}(\mathcal{A}_0)}z$. Besides, according to Lemma 2.1, we understand that

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

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

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

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 $dom\partial\psi \neq \varnothing$, then there exists $\hat{x}$ such that

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

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

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

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

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

Recall the fact that $\tilde{h}^* \in \mathcal{C}$. By Proposition 2.5 and [64, Corollary 23.5.1], we have

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

which implies

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

To the end, according to [64, Theorem 23.8, Theorem 23.9] and Proposition 2.5, we have

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

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

Appendix F. Proof of Lemma 2.3

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 2.6, there exist $\tilde{h}^* \in \mathcal{C}$ and matrix $\tilde{L}$ such that

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

and thus

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

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

To prove the second argument, noting that Assumption 1.1 holds, by virtue of Lemma 2.2, 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 2.6, we have

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

We may now obverse that any vector $y \in dom\,\partial f^*$ if and only if $\tilde{L}(y-q) \in dom\,\nabla\tilde{h}^*$ and $y-q \in \mathcal{R}(L^T)$. Since $\tilde{h}^*$ is essentially differentiable, thanks to Proposition 2.5, we understand that $y \in dom\,\partial f^*$ if and only if $\tilde{L}(y-q) \in int\ dom\,\tilde{h}^*$ and $y - q \in \mathcal{R}(L^T)$. 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

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

which further implies that

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

According to [64, 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 G. Proof of Proposition 2.8

We define the multifunction

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

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

$$\tilde{\mathcal{T}}_1(p) = \left\{\lambda \in \mathcal{V} \mid 0 \in \mathcal{M}\left(p, \lambda\right)\right\}, \tag{45}$$

where

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

Following the technique presented in [25], 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

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

By [25, 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

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

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 $\bar{\lambda} \in \text{int}(\text{dom}\tilde{\phi}_1)$ and by the locally Lipschitz continuity of $\nabla\tilde{\phi}_1$, there is a constant $L > 0$ 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 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:

$$\text{dist}\left((\lambda,0), H_{\mathcal{M}}(0)\right) \leq \|(\lambda,0) - (\lambda, \mathcal{G}(0,\lambda) - \mathcal{G}(p,\lambda))\| \leq \|\mathcal{G}(0,\lambda) - \mathcal{G}(p,\lambda)\| \leq L\left\|p\right\|,$$

  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 Proposition2.7, $\mathcal{M}_0(\lambda)$ 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 $(0,\bar{\lambda})$.

Appendix H. Proof of Lemma 2.4

By Lemma 2.3, 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

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

Since $\tilde{h}^* \left(\cdot - \tilde{q}\right) \in \mathcal{C}$, by Proposition 2.6, 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 2.3, 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 2.6, 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 I. Proof of Lemma 2.5

Motivated by [56, 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 $t_1 \in \mathrm{dom}\nabla\tilde{h}^*$, 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 J. Proof of Proposition 2.10

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

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

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

$$\begin{aligned} \operatorname{dist}(\lambda, Z) = \operatorname{dist}(\lambda, \Gamma_{DR}(0,0)) &\leq \kappa_1 \operatorname{dist}\left(0, \Gamma_{DR}^{-1}(\lambda)\right) \\ &\leq \kappa_1 \left(\|K\lambda - \bar{t}\| + \|\xi - \nabla\tilde{\phi}_1(\lambda) + \bar{g}\|\right) \\ &\leq \kappa_1 \left(\|K\lambda - \bar{t}\| + \|K^T\nabla\tilde{h}^*(K\lambda - \tilde{q}) - K^T\nabla\tilde{h}^*(\bar{t} - \tilde{q})\| + \|\xi\|\right) \\ &\leq (\kappa_1 + L_{\tilde{h}^*}\|K\|)\|K\lambda - \bar{t}\| + \kappa_1\|\xi\|. \end{aligned} \tag{46}$$

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

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

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

$$\sigma\|K\lambda - \bar{t}\|^2 \leq \langle \nabla\tilde{h}^*(K\lambda - \tilde{q}) - \nabla\tilde{h}^*(\bar{t} - \tilde{q}), K\lambda - \bar{t} \rangle \leq \langle \xi, \lambda - \hat{\lambda} \rangle \leq \|\xi\| \cdot \|\lambda - \hat{\lambda}\| = \|\xi\| \cdot \operatorname{dist}(\lambda, Z). \tag{47}$$

Combining (46) and (47), 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

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

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

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

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

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

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

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

The proof is complete.

Appendix K. Proof of Theorem 2.3

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 2.10, $\tilde{\mathcal{S}}_{D_1}$ is calm at $(\bar{\lambda}, 0)$.

Appendix L. Proof of Theorem 2.6

According to Theorem 2.4, Theorem 2.5, and Corollary 2.2, 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

$$\operatorname{dist}\left(\operatorname{prox}_{\phi_2}(z^k), Z\right) + \operatorname{dist}\left(\operatorname{prox}_{\phi_2{}^*}(z^k), W\right) \leq C_0 \rho^k, \quad \forall k \geq k_0. \tag{48}$$

According to Proposition 2.1, we know that, $\lambda^k = \operatorname{prox}_{\phi_2}(z^k)$. So we get the linear convergence of $\{\lambda^k\}$. Next, according to Proposition 2.1, 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

$$\|Ax^{k+1} + By^k - b\| = \frac{1}{\beta}\|z^{k+1} - z^k\| \leq \frac{C_0}{\beta}\rho^k, \quad \forall k \geq k_0, \tag{49}$$

where the last inequality follows from Corollary 2.2. Then, as shown in Proposition 2.1, 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 (49), we get

$$\|Ax^{k+1} + By^{k+1} - b\| + \|By^{k+1} - By^k\| \leq \frac{C_0}{\beta}\rho^k, \quad \forall k \geq k_0. \tag{50}$$

From (40) in Proposition 2.1, 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 (50), for all $k \geq k_0$, we have

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

which implies the linear convergence of the KKT residue sequence.

Additionally, note that

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

and

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

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

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

Combining above two inequalities, we get

$$\begin{aligned} f(x^*) + g(y^*) &\geq f(x^{k+1}) + g(y^{k+1}) + \langle \lambda^{k+1}, Ax^* + By^* - Ax^{k+1} - By^{k+1} \rangle + \beta \langle By^{k+1} - By^k, Ax^* - Ax^{k+1} \rangle \\ &\geq f(x^{k+1}) + g(y^{k+1}) + \langle \lambda^{k+1}, b - Ax^{k+1} - By^{k+1} \rangle + \beta \langle By^{k+1} - By^k, Ax^* - Ax^{k+1} \rangle, \end{aligned} \tag{51}$$

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

$$f(x^{k+1}) + g(y^{k+1}) \geq f(x^*) + g(y^*) + \langle \lambda^*, Ax^{k+1} + By^{k+1} - b \rangle. \tag{52}$$

Combining (51) and (52), we get

$$|f(x^{k+1}) + g(y^{k+1}) - f(x^*) - g(y^*)| \leq \max\{\|\lambda^{k+1}\|, \|\lambda^*\|\} \|Ax^{k+1} + By^{k+1} - b\| + \beta \|Ax^{k+1} - Ax^*\| \|By^{k+1} - By^k\|. \tag{53}$$

Then, by the non-emptiness of Ω^*, as proved in [40, 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 (53) 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 (50), we obtain the linear convergence with respect to the objective function value of Problem (1) straightforwardly.

Appendix M. Proof of Corollary 2.7

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

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

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 [64, 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

$$0 \in A\partial f^*(A^T\lambda) - b + \mu,$$

$$0 \in B\partial g^*(B^T\lambda) - \mu,$$

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 2.1, we know $\text{prox}_{\phi_2{}^*}(z^k) = By^k$. Following (48) in Proposition 2.6, 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

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

Furthermore, from Proposition 2.6, 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 N. Proof of Theorem 3.1

At each iteration, the PPA iterative scheme (22) reads also as

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

Therefore, for any $\mathbb{x}^* \in \mathcal{S}_{\mathbb{T}}$, it follows from the monotonicity of $\mathbb{T}$ that

$$\|\mathbb{x}^{k+1} - \mathbb{x}^*\|_{\mathbb{M}}^2 \leq \|\mathbb{x}^k - \mathbb{x}^*\|_{\mathbb{M}}^2 - \|\mathbb{x}^{k+1} - \mathbb{x}^k\|_{\mathbb{M}}^2. \tag{54}$$

Since $\mathbb{x}^*$ can be taken arbitrarily in $\mathcal{S}_{\mathbb{T}}$, we immediately have

$$\operatorname{dist}_{\mathbb{M}}^2\left(\mathbb{x}^{k+1}, \mathcal{S}_{\mathbb{T}}\right) \leq \operatorname{dist}_{\mathbb{M}}^2\left(x^k, \mathcal{S}_{\mathbb{T}}\right) - \|\mathbb{x}^{k+1} - \mathbb{x}^k\|_{\mathbb{M}}^2. \tag{55}$$

Because of the PPA-iteration-based error bound, there exist $\epsilon, \kappa > 0$ such that

$$\operatorname{dist}_{\mathbb{M}}\left(\mathbb{x}^{k+1}, \mathcal{S}_{\mathbb{T}}\right) \leq \kappa\|\mathbb{x}^{k+1} - \mathbb{x}^k\|_{\mathbb{M}}, \quad \text{for all } k \text{ such that } \mathbb{x}^k \in \mathbb{B}(\bar{\mathbb{x}}, \epsilon).$$

Given this ϵ, with the by-default given proximity of the sequence $\{\mathbb{x}^k\}$ generated by the PPA to $\bar{\mathbb{x}} \in \mathcal{S}_{\mathbb{T}}$, there exists $k_0 > 0$ such that $\mathbb{x}^k \in \mathbb{B}_\epsilon(\bar{\mathbb{x}})$ for $k \geq k_0$. Therefore, we have

$$\operatorname{dist}_{\mathbb{M}}^2\left(\mathbb{x}^{k+1}, \mathcal{S}_{\mathbb{T}}\right) \leq \frac{\kappa^2}{1+\kappa^2}\operatorname{dist}_{\mathbb{M}}^2\left(\mathbb{x}^k . \mathcal{S}_{\mathbb{T}}\right), \quad \forall k \geq k_0,$$

and thus there exists $C > 0$ such that

$$\operatorname{dist}_{\mathbb{M}}\left(\mathbb{x}^k, \mathcal{S}_{\mathbb{T}}\right) \leq C\rho^k, \quad \forall k \geq k_0,$$

with $\rho = \sqrt{\frac{\kappa^2}{1+\kappa^2}}$. Moreover, according to (55), we get

$$\|\mathbb{x}^{k+1} - \mathbb{x}^k\|_{\mathbb{M}} \leq C\rho^k, \quad \forall k \geq k_0.$$

The desired linear convergence then follows from the positive definiteness of matrix $\mathbb{M}$.

Appendix O. Proof of Theorem 3.2

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

$$\operatorname{dist}\left(x, \mathbb{T}^{-1}(0)\right) \leq \kappa \operatorname{dist}(0, \mathbb{T}(x)), \quad \forall x \in \mathbb{B}_\epsilon(\bar{\mathbb{x}}).$$

Note that $\mathbb{T}^{-1}(0) = \mathcal{S}_{\mathbb{T}}$. The metric subregularity of $\mathbb{T}$ at $(\bar{\mathbb{x}}, 0)$ allows us to estimate the distance from $\mathbb{x}^{k+1}$ to $\mathcal{S}_{\mathbb{T}}$ in terms of scaled optimality residual at $\mathbb{x}^{k+1}$, i.e., for all k such that $\mathbb{x}^k \in \mathbb{B}(\bar{\mathbb{x}}, \epsilon)$,

$$\operatorname{dist}_{\mathbb{M}}\left(\mathbb{x}^{k+1}, \mathcal{S}_{\mathbb{T}}\right) \leq \sqrt{\rho(\mathbb{M})}\operatorname{dist}\left(\mathbb{x}^{k+1}, \mathcal{S}_{\mathbb{T}}\right) \leq \kappa\sqrt{\rho(\mathbb{M})}\|\mathbb{M}(\mathbb{x}^{k+1} - \mathbb{x}^k)\| \leq \kappa\rho(\mathbb{M})\|\mathbb{x}^{k+1} - \mathbb{x}^k\|_{\mathbb{M}}, \tag{56}$$

where $\rho(\mathbb{M})$ represents the spectral radius of matrix $\mathbb{M}$. Thus, the PPA-iteration-based error bound holds at $\bar{\mathbb{x}}$. The conclusion then follows from Theorem 3.1.

Appendix P. Proof of Proposition 3.3

According to Proposition 3.2, we have

$$X = \arg\min_{x} \{\theta_1(x) + \theta_2^*(\mathcal{K}x)\}.$$

Thanks to Assumption 1.2, there exists $\bar{t} \in \mathbb{R}^l$ such that $Lx = \bar{t}$ for all $x \in X$. Moreover, for any $(x, \lambda) \in \Omega^*_{x,\lambda}$, we have

$$0 \in \partial\theta_2(\lambda) - \mathcal{K}x$$

$$0 \in \partial\theta_1(x) + \mathcal{K}^T\lambda = L^T\nabla h(Lx) + q + \mathcal{K}^T\lambda = L^T\nabla h(\bar{t}) + q + \mathcal{K}^T\lambda = \bar{s} + \mathcal{K}^T\lambda,$$

where $\bar{s} := L^T\nabla h(\bar{t}) + q$. Therefore, the following inclusion holds

$$\Omega^*_{x,\lambda} \subseteq \{(x, \lambda) \mid Lx = \bar{t}, \mathcal{K}^T\lambda = -\bar{s}, 0 \in \partial\theta_2(\lambda) - \mathcal{K}x\}.$$

It is easy to obtain the reverse direction. The proof is complete.

Appendix Q. Proof of Proposition 3.4

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

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

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 $(\xi, \eta) \in T(x, \lambda)$

$$\xi = \partial\theta_1(x) + \mathcal{K}^T\lambda = L^T\nabla h(Lx) + q + \mathcal{K}^T\lambda, \tag{57}$$

$$\eta \in \partial\theta_2(\lambda) - \mathcal{K}x, \tag{58}$$

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

$$\begin{aligned} \text{dist}\left((x, \lambda), \Omega^*_{x,\lambda}\right) = \text{dist}\left((x, \lambda), \Gamma_{PDHG}(0)\right) &\leq \kappa_1 \text{dist}\left(0, \Gamma^{-1}_{PDHG}(x, y)\right) \\ &\leq \kappa_1 \left(\|Lx - \bar{t}\| + \|\xi + \bar{s} - L^T\nabla h(Lx) - q\| + \|\eta\|\right) \\ &\leq \kappa_1 \left(\|Lx - \bar{t}\| + \|L\|\|\nabla h(\bar{t}) - \nabla h(Lx)\| + \|\xi\| + \|\eta\|\right) \\ &\leq \kappa_1 \left((1 + \|L\|L_h)\|Lx - \bar{t}\| + \|\xi\| + \|\eta\|\right). \end{aligned} \tag{59}$$

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 $0 \in \partial\theta_2(\hat{\lambda}) - \mathcal{K}\hat{x}$ and $\partial\theta_2$ is monotone, we have

$$\langle \eta + \mathcal{K}x - \mathcal{K}\hat{x}, \lambda - \hat{\lambda}\rangle \geq 0,$$

and subsequently,

$$\langle \eta, \lambda - \hat{\lambda}\rangle \geq -\langle \mathcal{K}^T\lambda - \mathcal{K}^T\hat{\lambda}, x - \hat{x}\rangle.$$

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

$$\begin{aligned} \langle \xi, x - \hat{x}\rangle + \langle \eta, \lambda - \hat{\lambda}\rangle &\geq \langle L^T\nabla h(Lx) - L^T\nabla h(\bar{t}), x - \hat{x}\rangle \\ &= \langle \nabla h(Lx) - \nabla h(L\hat{x}), Lx - L\hat{x}\rangle \\ &\geq \sigma\|Lx - L\hat{x}\|^2 = \sigma\|Lx - \bar{t}\|^2. \end{aligned} \tag{60}$$

Combining (59) and (60), we obtain that

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

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

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

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

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

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

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

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

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

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

$$p_1 = Lx - \bar{t},\ p_2 = \bar{s} + \mathcal{K}^T\lambda,\ p_3 \in \partial\theta_2(\lambda) - \mathcal{K}x.$$

To summarize, we have

$$\begin{aligned} p_2 + L^T\nabla h(Lx) - L^T\nabla h(Lx - p_1) &= \partial\theta_1(x) + \mathcal{K}^T\lambda, \\ p_3 &\in \partial\theta_2(\lambda) - \mathcal{K}x. \end{aligned}$$

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

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

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

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

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

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

Appendix R. Proof of Proposition 3.6

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

$$\mathrm{dist}\left((x,\lambda),\Gamma_0(0)\right) \leq \kappa_1\mathrm{dist}\left(0,\Gamma^{-1}_0\left(x,\lambda\right)\right),\ \forall (x,\lambda)\in\mathbb{B}_{\epsilon_1}(\bar{x},\bar{\lambda}).$$

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 $(\xi,\eta)\in T(x,\lambda)$

$$\begin{aligned}&\xi = L^T\nabla h(Lx) + q + \mathcal{K}^T\lambda,\\ &\eta \in \partial\theta_2(\lambda) - \mathcal{K}x,\end{aligned}$$

since $0\in L^T\nabla h(\bar{t}) + q + \mathcal{K}^T\bar{\lambda}$, and by the local Lipschitz continuity of ∇h, there exists $L_h>0$ such that

$$\|\mathcal{K}^T\lambda - \mathcal{K}^T\bar{\lambda}\| \leq \|L\|\|\nabla h(Lx) - \nabla h(\bar{t})\| \leq L_h\|L\|\|Lx-\bar{t}\|.$$

Moreover, noting that $\mathcal{K}^T$ is of full column rank, the smallest singular value of $\mathcal{K}^T$ is strictly positive, i.e., $\sigma_{\min}(\mathcal{K}^T)>0$. Therefore

$$\|\lambda-\bar{\lambda}\| \leq \frac{1}{\sigma_{\min}(\mathcal{K}^T)}\|\mathcal{K}^T\lambda - \mathcal{K}^T\bar{\lambda}\| \leq \frac{L_h\|L\|}{\sigma_{\min}(\mathcal{K}^T)}\|Lx-\bar{t}\|. \tag{61}$$

According to the calmness of $\partial\theta_2$ at $(\bar{\lambda},\mathcal{K}\bar{x})$, there exist $\epsilon_3,\kappa_3>0$ such that

$$\mathrm{dist}\left(v,\partial\theta_2(\bar{\lambda})\right) \leq \kappa_3\mathrm{dist}\left(\bar{\lambda},(\partial\theta_2)^{-1}(v)\right),\ \forall v\in\mathbb{B}_{\epsilon_3}(\mathcal{K}\bar{x}).$$

We now assume that $(x, \lambda) \in \mathbb{B}_{\epsilon_2}(\bar{x}, \bar{\lambda})$ with $\epsilon_2 := \min\{\epsilon_1, \epsilon_3/(2\|\mathcal{K}\|)\}$ and $\|\eta\| \leq \epsilon_3/2$. Since $\eta \in \partial\theta_2(\lambda) - \mathcal{K}x$ and thus

$$\eta + \mathcal{K}x \in \partial\theta_2(\lambda), \|\eta + \mathcal{K}x - \mathcal{K}\bar{x}\| \leq \|\eta\| + \|\mathcal{K}\|\|x - \bar{x}\| \leq \epsilon_3.$$

Then, by the calmness of $\partial\theta_2$ at $(\bar{\lambda}, \mathcal{K}\bar{x})$, we have

$$\operatorname{dist}(0, D - \mathcal{K}x) \leq \|\eta\| + \operatorname{dist}(\eta + \mathcal{K}x, D) \leq \|\eta\| + \kappa_3\|\lambda - \bar{\lambda}\|. \tag{62}$$

By (61) and (62), we have

$$\begin{aligned}\operatorname{dist}\left((x,\lambda), \Omega^*_{x,\lambda}\right) = \operatorname{dist}((x,\lambda), \Gamma_0(0)) &\leq \kappa_1 \operatorname{dist}\left(0, \Gamma_0^{-1}(x,\lambda)\right) \\ &\leq \kappa_1 \left(\|Lx - \bar{t}\| + \|\lambda - \bar{\lambda}\| + \operatorname{dist}(0, D - \mathcal{K}x)\right) \\ &\leq \kappa_1 \left(\|Lx - \bar{t}\| + \frac{(1+\kappa_3)L_h\|L\|}{\sigma_{\min}(\mathcal{K}^T)}\|Lx - \bar{t}\| + \|\eta\|\right) \\ &\leq \kappa_1 \left((1 + \frac{(1+\kappa_3)L_h\|L\|}{\sigma_{\min}(\mathcal{K}^T)})\|Lx - \bar{t}\| + \|\xi\| + \|\eta\|\right).\end{aligned} \tag{63}$$

Then, similar to the proof of Proposition 3.4, there exists $\sigma > 0$ such that

$$\sigma\|Lx - \bar{t}\|^2 \leq \langle \xi, x - \hat{x} \rangle + \langle \eta, \lambda - \hat{\lambda} \rangle, \tag{64}$$

where $(\hat{x}, \hat{\lambda})$ is the projection of (x, λ) on $\Omega^*_{x,\lambda}$. Upon combining (63) and (64), inspired by the proof of Proposition 3.4, we prove the conclusion with

$$\epsilon_2 = \min\{\epsilon_1, \epsilon_3/(2\|\mathcal{K}\|)\},\ \kappa_2 = \max\{\frac{1}{\|\mathcal{K}\|}, \tilde{\kappa}\},$$

where

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

and

$$c_1 = \kappa_1(\sigma_{\min}(\mathcal{K}^T) + (1+\kappa_3)L_h\|L\|)/(\sqrt{\sigma}\sigma_{\min}(\mathcal{K}^T)),\ c_2 = \sqrt{2}\kappa_1.$$

Appendix S. Proof of Lemma 3.1

Since $\mathbb{D}$ is closed and $\mathbb{D} \subseteq \mathcal{R}(\mathbb{E})$, for any x, there exists $x_{\mathbb{D}}$ such that

$$\operatorname{dist}(0, \mathbb{E}x - \mathbb{D}) = \|\mathbb{E}x - \mathbb{E}x_{\mathbb{D}}\|$$

and $\mathbb{E}x_{\mathbb{D}} \in \mathbb{D}$. Define $\mathbb{F}_x := \{z \mid \mathbb{E}z = \mathbb{E}x_{\mathbb{D}}\}$, according to Hoffman error bound (see, e.g., [46]),

$$\text{dist}\,(x, \mathbb{F}_x) \leq \frac{1}{\tilde{\sigma}_{\min}(\mathbb{E})} \|\mathbb{E}x - \mathbb{E}x_{\mathbb{D}}\|,$$

where $\tilde{\sigma}_{\min}(\mathbb{E})$ denotes the smallest nonzero singular value of $\mathbb{E}$. Since $\mathbb{F}_x \subseteq \mathcal{M}^{-1}(0)$ for any x, we have

$$\text{dist}\left(x, \mathcal{M}^{-1}(0)\right) \leq \text{dist}\,(x, \mathbb{F}_x) \leq \kappa \|\mathbb{E}x - \mathbb{E}x_{\mathbb{D}}\| = \kappa \text{dist}\,(0, \mathbb{E}x - \mathbb{D})\,,$$

which implies the metric subregularity of $\mathcal{M}$.

Appendix T. Proof of Theorem 3.4

The metric subregularity of T follows directly from Lemma 3.1, Propositions 3.6 and 3.7. We next estimate the metric subregularity modulus of T. Firstly, according to the proof of Lemma 3.1, we understand that Ω_x^2 is calm at $(0, \bar{x})$ with modulus $\frac{1}{\tilde{\sigma}_{\min}(\mathcal{K})}$, where $\tilde{\sigma}_{\min}(\mathcal{K})$ denotes the smallest nonzero singular value of $\mathcal{K}$. Inspired by the proof of [81, 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

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

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 3.6, T is metrically subregular at $(\bar{x}, \bar{\lambda}, 0)$ with modulus $\kappa_T = \max\{\frac{1}{\|\mathcal{K}\|}, \bar{\kappa}\}$, where

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

In particular,

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

Appendix U. Proof of Lemma 3.2

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 $\mathcal{R}(B) \cap \mathbb{B}_{\epsilon_1}(B\bar{y}) \subseteq \{By \mid y \in \mathbb{B}_\epsilon(\bar{y})\}$. 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 [64, Theorem 23.8, Theorem 23.9]. The proof is complete.

Appendix V. Proof of Theorem 3.6

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 3.3, ∂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 $0 \in ri(dom\, g^*)$, we surely have $\mathcal{R}(B^T) \cap ri(dom\, g^*) \neq \varnothing$. Then, according to Lemma 3.2, $\partial \theta_2$ is also calm everywhere on its graph. Note that, for any fixed η, from Lemma 3.3, $\partial g^*(\eta) = (\partial g)^{-1}(\eta)$ 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 $g = \delta_{\mathbb{B}}(\cdot)$, $\theta_2 = g^*(B^T y)$ with $g^*(z) = \|z\|_2$. Since $\partial g^* = \partial \|\cdot\|_2$ is calm everywhere on its graph, and $0 \in ri(dom\, g^*)$, $\mathcal{R}(B^T) \cap ri(dom\, g^*) \neq \varnothing$, then by Lemma 3.2, $\partial \theta_2$ is also calm everywhere on its graph. Moreover, since $\partial \|\eta\|_2$ 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 $B^T \bar{\lambda} \neq 0$.

Appendix W. Proof of Theorem 3.9

From Theorem 3.3, we know that 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

$$\mathrm{dist}_{\mathcal{M}}\left((x^{k+1}, \lambda^{k+1}), \Omega^*_{x,\lambda}\right) \leq \rho \mathrm{dist}_{\mathcal{M}}\left((x^k, \lambda^k); \Omega^*_{x,\lambda}\right), \tag{65}$$

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

$$\begin{gathered} \mathrm{dist}\left((x^k, \lambda^k), \Omega^*_{x,\lambda}\right) \leq C_0 \rho^k, \\ \|x^{k+1} - x^k\| + \|\lambda^{k+1} - \lambda^k\| \leq C_0 \rho^k. \end{gathered} \tag{66}$$

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

$$\mathrm{Fea}(x^{k+1}, y^{k+1}, \lambda^{k+1}) = \|Ax^{k+1} + By^{k+1} - b\| = \frac{1}{\beta}\|\lambda^{k+1} - \lambda^k\| \leq \frac{C_0}{\beta}\rho^k. \tag{67}$$

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

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

Thus, by (66), (67) 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 $k \geq k_0 + 1$, it holds that

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

Additionally, similar to the proof in Theorem 2.6, since

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

and

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

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

$$\begin{aligned} f(x^*) + g(y^*) \geq f(x^{k+1}) + g(y^{k+1}) + \langle \lambda^{k+1}, b - Ax^{k+1} - By^{k+1}\rangle + \beta\langle By^{k+1} - By^k, Ax^* - Ax^{k+1}\rangle \\ - r\langle x^{k+1} - x^k, x^* - x^{k+1}\rangle \end{aligned} \tag{68}$$

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

$$f(x^{k+1}) + g(y^{k+1}) \geq f(x^*) + g(y^*) + \langle \lambda^*, Ax^{k+1} + By^{k+1} - b\rangle. \tag{69}$$

Combining (51) and (52), we get

$$\begin{aligned} |f(x^{k+1}) + g(y^{k+1}) - f(x^*) - g(y^*)| &\leq \max\{\|\lambda^{k+1}\|, \|\lambda^*\|\}\|Ax^{k+1} + By^{k+1} - b\| \\ &\quad + \beta\|Ax^{k+1} - Ax^*\|\|By^{k+1} - By^k\| + r\|x^{k+1} - x^*\|\|x^{k+1} - x^k\|. \end{aligned} \tag{70}$$

According to (54) (see also [41]), fixing any $(\bar{x}, \bar{y}, \bar{\lambda}) \in \Omega^*$, for any k, $\{\|(x^k, \lambda^k) - (\bar{x}, \bar{\lambda})\|\}$ 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 (66), we obtain the linear convergence in terms of the objective function value of Problem (1) straightforwardly.

Appendix X. Proof of Theorem 4.2

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

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

$$p_3 \in B\partial g^*(B^T\tilde{\lambda}) - b + Ax,$$

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

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

We now assume that $(x, y, \lambda) \in \mathbb{B}_{\epsilon_1}(\bar{x}, \bar{y}, \bar{\lambda})$ with $\epsilon_1 = \epsilon/2$ and $\|p_2\| \le \epsilon/(2\|B(B^TB)^{-1}\|)= \epsilon\sigma_{\min}(B)/2$, where $\sigma_{\min}(B)$ denotes the smallest nonzero singular value of B. Then, $\|\tilde{p}_2\| \le \epsilon$ and $\|\tilde{\lambda} - \bar{\lambda}\| \le \|\lambda - \bar{\lambda}\| + \|\tilde{p}_2\| \le \epsilon$. Thus, by the metric subregularity of T at $(\bar{x}, \bar{\lambda}, 0)$

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

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

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

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

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

References

[1] Aspelmeier T, Charitha C, Luke DR (2016) Local linear convergence of the ADMM/ Douglas-Rachford algorithms without strong convexity and application to statistical imaging. *SIAM J. Imag. Sci.* 9(2):842-868.

[2] Bauschke HH, Combettes PL (2011) Convex Analysis and Monotone Operator Theory in Hilbert Spaces (Vol. 408). New York: Springer.

[3] Bauschke HH, Moursi WM (2017) On the Douglas-Rachford algorithm. *Math. Program.* 164(1-2): 263-284.

[4] Bertsekas DP, Nedi A, Ozdaglar AE (2003) Convex Analysis and Optimization. Belmont: Athena Scientific.

[5] Bondell HD, Reich BJ (2010) Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics* 64:115-123.

[6] Bonettini S, Valeria R (2012) On the convergence of primal-dual hybrid gradient algorithms for total variation image restoration. *J. Math. Imaging Vis.* 44(3):236-253.

[7] Boyd S, Parikh N, Chu E, Peleato B, Eckstein J (2010) Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learning* 3:1-122.

[8] Chambolle A, Pock T (2011) A first-order primal-dual algorithms for convex problem with applications to imaging. *J. Math. Imaging Vis.* 40:120-145.

[9] Chan TF, Glowinski R (1978) Finite element approximation and iterative solution of a class of mildly non-linear elliptic equations. Technical report, Stanford University.

[10] Davis D, Yin WT (2017) Faster convergence rates of relaxed Peaceman-Rachford and ADMM under regularity assumptions. *Math. Oper. Res.* 42:783-805.

[11] Deng W, Yin WT (2016) On the global and linear convergence of the generalized alternating direction method of multipliers. *J. Sci. Comput.* 66:889-916.

[12] Douglas J, Rachford HH (1956) On the numerical solution of heat conduction problems in two and three space variables. *Trans. Amer. Math. Soc.* 82:421-439.

[13] Eckstein J, Bertsekas DP (1990) An alternating direction method for linear programming. LIDS-P, Cambridge, MA, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology.

[14] Eckstein J, Bertsekas DP (1992) On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Math. Program.* 55:293-318.

[15] Eckstein J, Yao W (2015) Understanding the convergence of the alternating direction method of multipliers: Theoretical and computational perspectives. *Pacific J. Optim.*, 11:619-644.

[16] Esser E, Zhang X, Chan T (2010) A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM J. Imaging Sci.* 3(4):1015-1046.

[17] Fornasier M, Rauhut H (2008) Recovery algorithms for vector-valued data with joint sparsity constraints. *SIAM J. Numer. Anal.* 46(2):577-613.

[18] Fortin M. and Glowinski R (1983) On decomposition-coordination methods using an augmented Lagrangian. In: Fortin, M. and Glowinski, R. (eds.): Augmented Lagrangian Methods: Applications to the Solution of Boundary Problems, North-Holland, Amsterdam, 97-146.

[19] Francisco F, Pang JS (2007) Finite-dimensional Variational Inequalities and Complementarity Problems (Springer-Verlag, New York).

[20] Friedman J, Hastie T, Tibshirani R (2010) A note on the group lasso and a sparse group lasso. Unpublished manuscript, `https://arxiv.org/abs/1001.0736`.

[21] Gabay D (1983) Applications of the method of multipliers to variational inequalities. Fortin M, Glowinski R, eds. *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems* (Elsevier, Amsterdam), 15:299-331.

[22] Gabay D, Mercier B (1976) A dual algorithm for the solution of nonlinear variational problems via finite-element approximations. *Comput. Math. Appli.* 2:17-40.

[23] Gfrerer H (2011) First order and second order characterizations of metric subregularity and calmness of constraint set mappings. *SIAM J. Optim.* 21:1439-1474.

[24] Gfrerer H (2013) On directional metric regularity, subregularity and optimality conditions for nonsmooth mathematical programs. *Set-Valued Anal.* 21(2):151-176.

[25] Gfrerer H, Klatte D (2016) Lipschitz and Holder stability of optimization problems and generalized equations. *Math. Program.* 158(1-2):35-75.

[26] Gfrerer H, Ye JJ (2017) New constraint qualifications for mathematical programs with equilibrium constraints via variational analysis. *SIAM J. Optim.* 27:842-865.

[27] Giselsson P, Boyd S (2017) Linear convergence and metric selection for Douglas-Rachford splitting and ADMM. *IEEE Transactions on Automatic Control* 62(2):532-544.

[28] Glowinski R (1984), Numerical Methods for Nonlinear Variational Problems, Springer-Verlag, New York, Berlin, Heidelberg, Tokyo.

[29] Glowinski R (2014), On alternating direction methods of multipliers: A historical perspective, Modeling, Simulation and Optimization for Science and Technology, W. Fitzgibbon and Y. A. Kuznetsov and P. Neittaanm ki and O. Pironneau(Springer Netherlands):59-82, .

[30] Glowinski R, Marrocco A (1975) Sur I'approximation, par *é*l*é*ments finis d'ordre un, et la r*é*solution, par pénalisation-dualit*é* d'une classe de problèmes de Dirichlet non lin*é*aires. *R.A.I.R.O.* 9(R2):41-76.

[31] Glowinski R, Tallec PL (1989) Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics (SIAM Studies in Applied Mathematics, Philadelphia).

[32] Goebel R, Rockafellar RT (2008) Local strong convexity and local Lipschitz continuity of the gradient of convex functions. *J. Convex Anal.* 15(2):263.

[33] Güler O (1991) On the convergence of the proximal point algorithm for convex minimization, *SIAM J. Optim.* 1:403-419.

[34] Guo L, Ye JJ, Zhang J (2013) Mathematical programs with geometric constraints in Banach spaces: enhanced optimality, exact penalty, and sensitivity. *SIAM J. Optim.* 23:2295-2319.

[35] Han DR, Sun DF, Zhang LW (2017) Linear rate convergence of the alternating direction method of multipliers for convex composite programming. *Math. Oper. Res.* 43(2): 622-637.

[36] Han DR, Yuan XM (2013) Local linear convergence of the alternating direction method of multipliers for quadratic programs. *SIAM J. Numer. Anal.* 51:3446-3457.

[37] He BS, Liao LZ, Han DR, Yang H (2002) A new inexact alternating directions method for mononotone variational inequalities, *Math. Program.*, 92:103-118.

[38] He BS, Xu MH, Yuan XM (2011) Solving large-scale least squares covariance matrix problems by alternating direction methods,. *SIAM J. Matrix Anal. Appl.* 32, 136-152.

[39] He BS, You YF, Yuan XM (2014) On the convergence of primal-dual hybrid gradient algorithm. *SIAM J. Imaging Sci.* 7:2526-2537.

[40] He BS, Yang H (1998) Some convergence properties of a method of multipliers for linearly constrained monotone variational inequalities. *Oper. Res. Let.* 23:151-161.

[41] He BS, Yuan XM (2012) Convergence analysis of primal-dual algorithms for a saddle-point problem: From contraction perspective. *SIAM J. Imaging Sci.* 5(1):119-149.

[42] He BS, Yuan XM (2012) On the O(1/n) convergence rate of the Douglas-Rachford alternating direction method. *SIAM J. Numer. Anal.* 50(2):700-709.

[43] He BS, Yuan XM (2015) On non-ergodic convergence rate of Douglas-Rachford alternating direction method of multipliers. *Numer. Math.* 130(3):567-577.

[44] Henrion R, Jourani A, Outrata J (2002) On the calmness of a class of multifunctions. *SIAM J. Optim.* 13:603-618.

[45] Henrion R, Outrata J (2005) Calmness of constraint systems with applications. *Math. Program.* 104:437-464.

[46] Hoffman AJ (1952) On approximate solutions of systems of linear inequalities. *J. Res. Natl. Bur. Standards* 49:263-265.

[47] James GM, Paulson C, Rusmevichientong P (2013) Penalized and constrained regression. Unpublished manuscript, `http://www-bcf.usc.edu/~gareth/research/Research.html`.

[48] Klatte D, Kummer B (2002) Constrained minima and Lipschitzian penalties in metric spaces. *SIAM J. Optim.* 13:619-633.

[49] Kowalski M (2009) Sparse regression using mixed norms. *Appl. Comput. Harmon. Anal.* 27(3):303-324.

[50] Li M, Sun DF, Toh KC (2016) A majorized ADMM with indefinite proximal terms for linearly constrained convex composite optimization. *SIAM J. Optim.* 26:922-950.

[51] Liang J, Fadili J, Peyré G (2017) Local convergence properties of Douglas-Rachford and alternating direction method of multipliers. *J. Optim. Theory Appl.* 172:874-913.

[52] Lin ZC, Liu R, Li H (2015) Linearized alternating direction method with parallel splitting and adaptive penalty for separable convex programs in machine learning. *Mach. Learn.* 99(2):287-325.

[53] Lions PL, Mercier B (1979) Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.* 16:964-979.

[54] D Leventhal, Metric subregularity and the proximal point method, J. Math. Anal. Appl. 360 (2009) 681-688.

[55] Liu YC, Yuan XM, Zeng SZ, Zhang J (2018) Partial error bound conditions and the linear convergence rate of alternating direction method of multipliers. *SIAM J. Numer. Anal.* 56: 2095-2123.

[56] Luo ZQ, Tseng P (1992) On the linear convergence of descent methods for convex essentially smooth minimization. *SIAM J. Control Optim.* 30:408-425.

[57] Martinet B (1970) Regularization d'inequations variationelles par approximations successives, Revue Francaise d'Informatique et de Recherche Opérationelle, 4: 154-159.

[58] Monteiro RD, Svaiter BF (2013) Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. *SIAM J. Optim.* 23(1):475-507.

[59] Nishihara R, Lessard L, Recht B, Packard A, Jordan MI (2015) A general analysis of the convergence of ADMM. Unpublished manuscript, `https://arxiv.org/abs/1502.02009`.

[60] Robinson SM (1981) Some continuity properties of polyhedral multifunctions, *Math. Programming Stud.* 14:206-214.

[61] Robinson SM (1975) Stability theory for systems of inequalities. Part I: Linear systems, *SIAM J. Numer. Anal.*, 12(5):754-769

[62] Robinson SM (1980) Strongly regular generalized equations. *Math. Oper. Res.* 5:43-62.

[63] Rockafellar RT (1976) Monotone operators and the proximal point algorithm. *SIAM J. Control Optim.* 14(5):877-898.

[64] Rockafellar RT (2015) Convex Analysis (Princeton University Press, Princeton).

[65] Rockafellar RT and Wets R (2009) Variational Analysis, *Springer Science & Business Media.*

[66] Shefi R (2015) Rate of convergence analysis for convex nonsmooth optimization algorithms. Unpublished doctoral dissertation, Tel Aviv University, Israel.

[67] Sun J and Zhang S (2010) A modified alternating direction method for convex quadratically constrained quadratic semidefinite programs, *European J. Oper. Res.* 207:1210-1220.

[68] Tao M and Yuan XM (2018) The generalized proximal point algorithm with step size 2 is not necessarily convergent, *Comput. Optim. Applic.*, 70(3): 827-839.

[69] Tao M and Yuan XM (2018) On Glowinski's open question of alternating direction method of multipliers, *J. Optim. Theory Applic.*, 179(1): 163-196.

[70] Teboulle M (1997) Convergence of proximal-like algorithms, *SIAM J. Optim.*, 7: 1069-1083.

[71] Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)* 267-288.

[72] Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K (2005) Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67:91-108.

[73] Valkonen T (2014) A primal-dual hybrid gradient method for non-linear operators with applications to MRI. *Inverse Probl.* 30:055012.

[74] Valkonen T (2017) Preconditioned proximal point methods and notions of partial subregularity, *https://arxiv.org/abs/1711.05123.*

[75] Wang XF, Ye JJ, Yuan XM, Zeng SZ, Zhang J (2018) Perturbation techniques for convergence analysis of proximal gradient method and other first-order algorithms via variational analysis. *arXiv preprint.*

[76] Wang XF, Yuan XM (2012) The linearized alternating direction method of multipliers for Dantzig selector. *SIAM J. Sci. Comput.* 34:2792-2811.

[77] Wen Z, Goldfarb D and Yin W (2010) Alternating direction augmented Lagrangian methods for semidefinite programming, Math. Prog. Comp. 2, 203-230.

[78] Yang JF, Yuan XM (2013) Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization. *Math. Comp.* 82:301-329.

[79] Yang WH, Han DR (2016) Linear convergence of the alternating direction method of multipliers for a class of convex optimization problems. *SIAM J. Numer. Anal.* 54:625-640.

[80] Ye JJ, Ye XY (1997) Necessary optimality conditions for optimization problems with variational inequality constraints. *Math. Oper. Res.* 22:977-997.

[81] Ye JJ, Yuan XM, Zeng SZ, Zhang J (2018) Variational analysis perspective on linear convergence of some first order methods for nonsmooth convex optimization problems. *optimization-online preprint.*

[82] Ye JJ, Zhang J (2013) Enhanced Karush-Kuhn-Tucker condition and weaker constraint qualifications. *Math. Program.* 139:353–381.

[83] Yuan M, Lin Y (2006) Model selection and estimation in regression with group variables, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 68:49-67.

[84] Zhou H, Sehl ME, Sinsheimer JS, Lange K (2010) Association screening of common and rare genetic variants by penalized regression. *Bioinformatics* 26(19):2375-2382.

[85] Zhu XD, Zhang J, Zeng SZ, Yuan XM (2018) Linear Convergence of R-BCPGM/prox-SVRG under bounded metric subregularity. *Manuscript*